IFLA Big Data Special Interest Group
October 2018

A concept data science framework for libraries

A. Introduction

Libraries are challenged to adopt new service models to assist with the transformation of data into information. Libraries need to improve their technological literacy, especially coding and mark-up, to proactively take advantage of the new possibilities presented in the growing domain of library data analytics which provide new insights into existing service models. However, this acknowledgement of being mindful of proactively calibrating one's professional mission does not in itself imply that libraries generate, or manage, big data as it is usually understood in the traditional sense of the word. It is merely an acknowledgement that the data-intensive world in which libraries function necessitates libraries to have an awareness of being "data savvy". This is coupled with the responsibility to adapting ones professional skillset to proactively respond to the changing requirements of the user base. A Data Science framework is an outline and description of how to respond to such developments; a position paper rather than a statement of intent to administer big data per se, which is an activity usually associated with the term "data science" as it is traditionally understood.

B. Data science and libraries

Data science exists on a spectrum and can span work that requires deep statistical and software engineering skills, to work focusing on advocacy, policy development, data management planning, and evidence-based decision making. The term "data science" was first used by academic statisticians to position the discipline with respect to big data, data analysis, and broader trends. These early discussions emphasized mathematical foundations and the new statistical methods made possible by an abundance of data (Cleveland, 2001; Dononho, 2015). From this academic origin, industrial data science emerged driven by new technology and the ability to extract business value from data.

Today data science is seen as the blending of competencies in computer programming, software engineering and statistics, combined with a particular domain expertise. This perspective is visible in the many definitions and discussions of data scientists that emphasize the algorithms, machine learning, and statistical techniques. While these areas of expertise are certainly important, the work of data cleaning and preparation is possibly the most important set of skills for data science. Extracting

value from data is more than just the mechanical application of a classification or clustering algorithm; there is a significant amount of "janitor work" involved in many data-centric processes (*citation IMLS Report comes here*).

Closer to the domain of libraries, a family of data science roles have been identified, which can be characterised by real-world requirements for actual positions as described in two related small-scale studies (Lyon & Mattern, 2017; Lyon, Mattern, Acker & Langmead, 2015). The six roles are: data archivist, data curator, data librarian, data analyst, data engineer, and data journalist (note: an important factor to keep in mind is that while all of these roles have been framed as data science roles, other framing which comes from the corporate sector has tended to describe only data analyst-type roles as data scientists). Data librarians gain familiarity with the datasets, understand technical methods and techniques, and speak multiple disciplinary languages allowing them to work closely with researchers. Data science-based services have the potential for convincing researchers that the library remains a valuable resource in the digital landscape.

Within academia, data science has pervaded almost every discipline and data-savvy skills are used in many professions to increase efficiencies and gain insight. While many libraries are not yet ready to take on a large data science role on campus, there are many opportunities to partner and contribute to an emerging data science support network. In addition to working with scholarly datasets, librarians now have the opportunity to add value to research data through new roles as data curators, providing a range of research data management services, such as reviewing data management plans, preserving datasets for the long-term in institutional repositories, and tracking the re-use of datasets with persistent identifiers through citation and download metrics.

Data science therefore exists more or less on a spectrum, depends on an institution's size and mission, and spans work requiring the deep statistical and software engineering skills, to work that focuses on advocacy, policy, communication, and data management. Being "data savvy" is an essential ingredient of all of these roles.

C. Data

All data is data, and can even be considered as big data depending on the context. In broad terms, data vary by form (qualitative or quantitative), structure (structured, semi-structured or unstructured), source (captured, derived, exhaust, transient), producer (primary, secondary, tertiary), and type (indexical, attribute, metadata). Simply put, when somebody says something is data, then it is data. It is therefore a matter of *when* it is data, and not *what* is data (i.e. data is fluid, not hard like a book). Even false data is data nonetheless. For libraries such a perspective introduces a departure

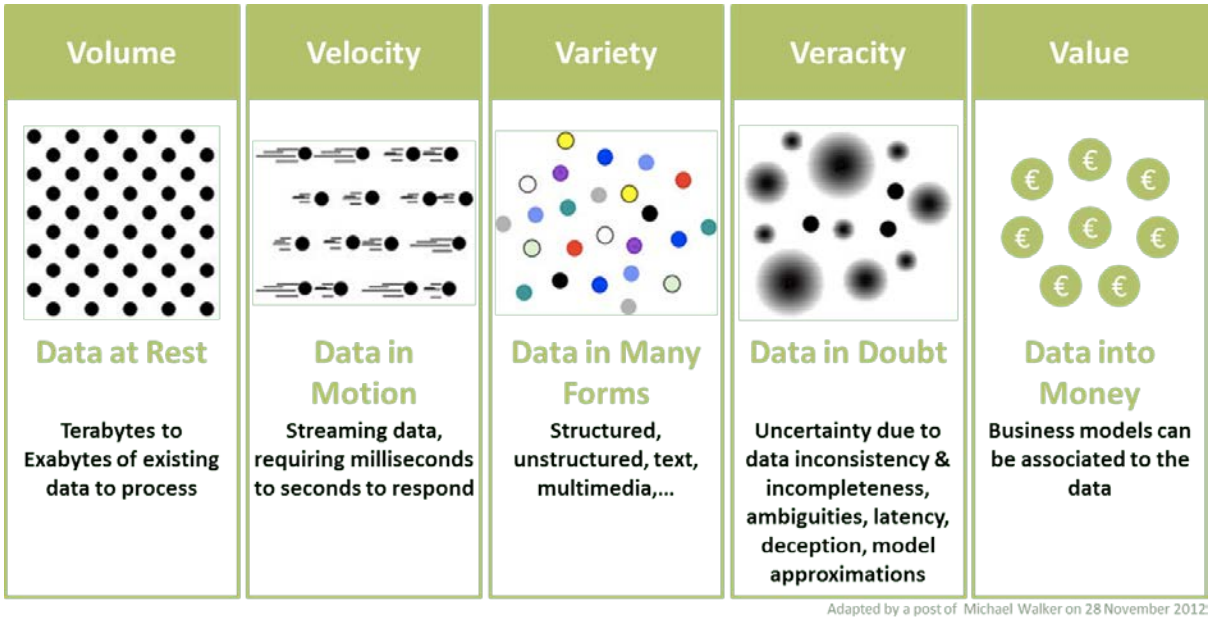
from a more taxonomy-driven approach to classify artefacts as fixed objects of some sort. But similarly it also produces the realisation that there is a strong business case for libraries to analyse their current service catalogues to determine if and where adjustment in service delivery roles and skillsets are needed to adapt to an increasingly data-intensive society in which an information service has to be provisioned.

In a like manner, libraries are no longer just consumers of data but also suppliers and to a lesser extent producers of data. New technologies such as artificial intelligence (AI) rely as heavily on good-quality data as patrons rely on good-quality information to build a good framework of reference and knowledgebase. Data integrity and data ethics are core considerations in determining reliable data sources. The values of privacy, ethics, and equitable access to information have always been core to libraries, uniquely positioning libraries to be new partners to researchers and data sources of high-quality data.

AI is no longer a "future shock" scenario. Likewise to other professions, libraries should start now to plan how AI affects the information profession, and the place to start at is with sourcing and making available reputable data.

D. Big data in libraries

Although definitions vary, big data is typically understood through the 3Vs framework: volume, variety, and velocity. These V-characteristics were first proposed by Doug Laney, now a Gartner consultant, in the context of emerging business conditions (2001). The graphic below adapted by a post of Michael Walker summarises these characteristics.



Volume refers to data set size (typically terabytes and petabytes); variety indicates that big data is unstructured and varied (e.g. text, audio, video and images); and velocity denotes the high frequency at which this data is generated. IBM includes veracity (data uncertainty) as yet another characteristic, and as of late the characteristic of value has been added as the 5th V-characteristic to denote how such data can enrich an organisation's value chain and service delivery offering.

From the above one can assume that if a community's ability to deal with data is overwhelmed, then it is big data. Likewise to other industries big data has become a focus of academic and research libraries due to the rapid evolution of data mining technologies and the proliferation of data sources like mobile devices and social media. In 2013 already McKinsey stated that in the US 15 out of 17 industry sectors hold more data per company than the US Library of Congress itself. Similarly, as far back as 2011 the number of networked devices for the first time already exceeded the number of people on the planet. In that same year the Library of Congress' holdings totalled 235 terabyte (TB) of data, partially qualifying as big data from a volume perspective only. Correspondingly, industry projections suggest that by 2050 the scale of automated machine-to-machine traffic could mean that connected devices will outstrip the number of connected human beings six to one (6:1) (IFLA Trend Report 2013). Vast and expanding data sets acquired by governments and companies through their interactions with Internet users – in conjunction with that generated by scientific research, surveillance and smart object sensors – continuously expand the possibilities for innovative services to be developed whilst simultaneously enabling sophisticated profiling of individuals.

These developments suggest that curated library data is not big data per se. Libraries however could do well by paying greater conceptual attention to data. If doing so, one realises that the matter of data being big or simply its opposite, namely "small" (i.e. that data you can manage locally, using everyday tools) is irrelevant. If analytics depends increasingly on big data, then big data depends on small data to provide meaningful insights. Once librarians have become accustomed to working with complex datasets the general principle and challenges of curating these data are fundamentally the same as those for smaller, traditional datasets. The first step is to look at how to manage small data effectively and then to apply those very same principles to big data.

The V-characteristics of big data should also be of a lesser concern to libraries, since libraries aren't responsible for storing large volumes of data. For libraries "Big Data" starts when methods are applied to data to help foster the knowledge process. Considerations such as whether big data are "findable" and "available" to remain open for analysis should be of greater concern to libraries than the storing and processing of the data. From a library perspective examples of big data are given below, each listed within one of the four classes of big data sources:

1) Archetypal big data sources

Archetypal big data denotes data that is generated through science projects, industry, and government. It is high in volume, velocity, veracity and value, but low in variety. This data is typically generated by automation routines in devices and systems. Data in libraries that are produced in a similar way include:

- Exhaust data, which is produced as a by-product of the main function of a device or system, such as the materials flow self-service unit (self-checkout).
- Patron activity data, such as transactional loans and fine payments, as produced by a centralised library management system.
- Usage pattern data, such as those kept in the activity logs of a discovery system.
- User behaviour data, such as web statistical data which is currently used to track end-user behaviour on a library website, hosted e-journals, and digital repositories.

2) Crowd-sourced big data sources

Crowd-sourced big data denotes data that is volunteered on a public citizenship basis, such as social networks, recommender services, and web commerce systems. It is high in volume and variety, irregular in velocity, mixed in veracity and short-lived in value. Data in libraries that are generated in a similar way are social networks commentary on library Facebook sites or a Twitter channels.

3) Sensor stream big data sources

Sensor stream big data denotes data that is generated in a directed or automated way by surveillance systems, environmental systems and personal wearable devices. It is high in volume, velocity, variety and value, and mixed in veracity. Data in libraries that are produced in a similar way are security camera footage, turnstile activity, and building temperature control meters.

4) Long-tail big data sources

Long-tail big data denotes scholarly data generated by science and research projects, inclusive of citizen science projects. With 2.5 million papers (with supplemental data) published per year, and more than 600 000 Crossref DOIs minted each month, scholarly publishing is indeed big data. It is high in volume and veracity, low in velocity and variety, and from a pure data perspective, unknown in value. Data mining projects such as FutureTDM, OpenMinTeD, WikiData and ContentMine are attempting to uncover the benefits behind the unknown value of big scholarly data.

Big data has significance for academic libraries in their roles as facilitators and supporters of the research process. Besides brokering access to scholarly data through publishers and aggregators,

libraries can also provide access to similar data sources that are under their curatorship. Examples include:

- Bibliographic metadata hosted in aggregated databases, such as OCLC Worldcat.
- Textual data hosted in institutional repositories (e.g. theses, dissertations, research articles).
- Maps, images and videos contained in special collection departments.
- Newspaper collections, which are analysed for named-entity relationships.

E. A multi-faceted framework

Before embarking on a data science program library managers and their teams should address the hard work of defining exactly what set of problems involving data need solving. Without an agreed framework to work from or alignment to institutional goals and policies, a number of fears can intervene to impact progress towards integrating data science into library culture. As a point of departure there are a number of different facets that need to be considered. For the purpose of simplicity these are grouped into three sections, namely Skills, Operations, and Services.

1. Skills

A data-savvy skills gap exists at all levels of librarianship, from the formal training received in library school, to the practices of mid-career librarians, to the leadership of library directors. To be an effective manager in the data-driven era one will need the vision of where library data services are transitioning to, know how to stay ahead of the data science waves, and have a good understanding of and skills in data science. Addressing the skills gap and helping librarians become data savvy involves both formal and informal training. Embedding a local data-savvy trainer that links training programs with organisational goals and projects could be helpful. Workshop programs such as Library Carpentry are also maturing; the Carpentries sits at the doorstep of libraries with a curriculum already in place, a worldwide network to learn from, and certificate-based instructor training programs.

To address the skills gap three areas stand out that could warrant attention:

1.1. Develop broad data knowledge and digital competency profiles

JISC defines **digital literacies** broadly as "*the capabilities which fit an individual for living, learning and working in a digital society*". It is generally recognised that an individual's ability to be digital fluent is measured across four levels of development, namely access, skills, practice and identity. These areas are expanded in Belshaw's (2014) book "*The essential elements of digital literacies*", in which he lists eight essential elements of digital fluency. There are cultural, cognitive, constructive, communicative, confident, creative, critical and civic. Digital literacy fluency is a

requirement for information literacy and data literacy skills, since it measures all aspects of an individual's interaction with the digital domain in which information and data are to be found. It includes privacy considerations as described in legislation such as the POPI Act and General Data Protection Regulation (GDPR).

Developing in-house digital literacy competency profiles for library professionals is currently done by a number of academic and research libraries worldwide, such as the Library Freedom Project, the Library Intelligence diagnostic tool to assess digital literacy outcomes, the FOSTER Plus Project's digital skills for library staff and researchers working group, and the Open University Library Services' project to establish resources to grow digital capabilities of staff and tutors.

Digital literacy fluency exercises allow library professionals to measure their own readiness to engage with medium-weight data literacy programs such as Library Carpentry and Data Carpentry. These programs equip library professionals with the basic toolset to source, administer, and manipulate data using industry developed toolkits such as statistical packages and software programming languages. Libraries stand to benefit by coordinating and aggregating information about data-savvy training and education initiatives, such as the Data Science in Libraries Learning Registry, as well as the Training and Resources Guide for the World Data System.

1.2. Develop a data literacy framework for curriculum instruction

Although **data literacy** is still going through a gestation period a lot has been written and blogged on the subject for the past ten years. Data literacy is better known in the academic sphere and closely related to information literacy and generally accepted as a crucial ability for librarians involved in supporting data-intensive research. The community of practice for data librarians differ from the one of information literacy, but there are also a number of similarities causing the boundaries between information literacy and data literacy to blur even further. To this end Calzada and Marzal (2013) reviewed the literature of information literacy and data literacy instruction, and proposed a common reference framework for data literacy instruction. The framework builds on trends in data literacy instruction in both data use and data management, and translates competencies identified within the literature into discrete modules that can be used as the foundation for course instruction. The modules are:

i. Understanding data
ii. Finding and/or obtaining data
iii. Reading, interpreting, and evaluating data
iv. Managing data
v. Using data

Likewise, Lisa Federer (UCLA, 2012) started a list of core competencies for data literacy instruction to be applicable across a variety of scientific fields. They are:

i. Understanding the "data life cycle"
ii. Knowing how to write a data management plan (DMP)
iii. Making appropriate choices about file forms and formats
iv. Keeping data organised and discoverable
v. Planning for long-term, secure storage of data
vi. Promoting sharing by publishing datasets and assigning persistent identifiers like DOIs
vii. Awareness of data as scholarly output that should be considered in the context of promotion and tenure

Libraries could do well by proactively investigating time and resources into determining how library professionals could benefit from greater data literacy awareness and advanced skills. In turn, such knowledge and skills can be incorporated into the academic literacy framework of institutions.

1.3. Develop specialist data roles

The matter of the value of formal versus informal education (multi-year degrees or short courses), and the benefits from recruiting individuals with a deep disciplinary knowledge (as evidenced by a Honour or Master degree) compared to individuals with a broad knowledge only (demonstrated by technical certifications and short courses), has given rise to an educational continuum which is labelled "the credentialing tension". The report show there is a growing tension between stewardship and science activities in libraries, with the former being much more established as research data management or research data services than the latter (Cox, Kennan, Lyon & Pinfield, 2017). However, both stewardship and science represent valid and beneficial "data curation" activities, viewed from the perspective of the following definition from the UK Digital Curation Centre: "*curation is adding value to data*".

Data-savvy librarians could occupy a space at the intersection of these competing tensions whilst two different type of specialist data roles in libraries can be conceptualised to give expression to the need for further areas of data specialisation beyond just being "data savvy":

The data librarian

Focuses on **stewardship-oriented** support roles which are centred on policy, data management planning, data literacy training and advocacy (data archiving, data curation). The data librarian acts as facilitator in all stages of scientific research and provides data services

within the full life cycle. In addition, the data librarian advocates for the need for data literacy training and demonstrate proficiency with statistical packages and data cleaning techniques.

The data technologist

Focuses on **science-oriented** support roles which are centred on technical workflows to manipulate and wrangle the data to produce novel insights (data analysis, data engineering). The data technologist has technical skills to complement the data librarian's knowledge with computing skills, facilitates mid-weight training interventions like The Carpentries, and develops data mining techniques to convert library information sources into data collections.

Data specialists roles could possess advanced data science skills which are likely not possessed by other library professionals since it is unrealistic of librarians to gain the same data skills as researchers. Supporting data science is a team sport – at best librarians will over time gain exposure to skills and develop a spectrum of data-savvy skills.

2. Operations

For library directors and operational managers, there is a requirement for timely and insightful management information (or "business intelligence", BI) derived from library data. Data science approaches may shed light on otherwise hard-to-see or misunderstood problems in the library. In order to make informed decisions in annual planning, senior library managers need to have tangible evidence to validate operational decision making, future investments, and staff deployments. This evidence can be gathered from data collection and data analysis, followed by subsequent insight.

There are three disciplinary areas in data science, namely Sources, Analytics and Visualization. From an operations perspective these apply in different ways using different analytical programs. Most data sources applied to operations will be of type archetypal, crowd-sourced and sensor stream. Special mention is made of two analytical programs in libraries:

2.1. Library analytics

Library analytics is a field onto its own and mainly derived from commerce business intelligence techniques. Analytical methods in libraries are useful for library planning, informing business operations and optimising collections. The foundation of library analytics is largely statistical analysis and is mainly build from archetypal data sources. Libraries are approximately 8 – 10 years behind other industries in the uptake of data analytical methods to inform better operational and strategic management decision-making. Interest in library analytics can be seen to have arisen from a confluence of many factors both from within and outside academic libraries. Some reasons include:

- An interest in big data, data science and AI in general.

- Library systems are becoming more open and more capable at analytics (e.g. Alma Analytics).
- Assessment and an increasing demand to show value have become hot trends.
- A rising interest in learning analytics (i.e. understanding why some students are not succeeding, what could contribute to their success, and how and when interventions might be helpful).
- The increasing academic focus on managing research data has provided synergy.

There are increasing levels of capability and impact related to library analytics which could go through a number of levels, for example:

Level 1 – analysis done is library function specific (e.g. collection dashboard for Alma or web dashboard for Google Analytics)

Level 2 – a centralised library wide dashboard is created covering most functional areas in the library

Level 3 – the library "shows value" by running correlation studies

Level 4 – the library ventures into predictive analytics or learning analytics.

Library analytics is visualised in the form of "library dashboards" where data are pulled from diverse systems into one centralized dashboard. There are a number of dashboards systems available – some open source and some proprietary – each requiring configuration of disparate data sources and the subsequent normalization of data. Setting up the dashboards requires initial effort to present the data from heterogeneous systems in one common format, and additionally requires on-going effort to maintain it consistently to provide clean and comparable data that is meaningful to library decision-makers. Library dashboards are particularly useful for visualizing patron behaviour when they interact with library resources; both in digital format (online e-resources) and in print format (circulation statistics).

2.2. Crowdsensing of user interests

Crowdsensing of user interest is based on data contained in crowd-sourced and sensor-stream big data sources. Libraries themselves generate data through their online resources and services, and the social media services they use to promote their programs and amenities. The analysis of library footfall data can shed light on difficult resourcing challenges, for example when and if to staff information desks, how to deploy library shelving, and to identify optimal opening and closing times. Crowdsensing of user interest is accomplished by tracking user behaviour on library systems

and following user interests on social media. Data is integrated to build entity relationships for patron profiling; from this insight recommender services can be developed to enhance service delivery.

Example: a typical use case would be to analyse user interests through search keywords on library websites combined with determining collection gaps based on search results and "click-throughs" on discovery platforms. These results are integrated with an analysis of trending topics based on user activity on social media and other relevant website platforms. Entity relationships are created for the purpose of user profiling to forecast user behaviour and interest patterns (predictive analysis).

3. Services

Libraries have traditionally serviced their user base by providing access to reliable information sources in an impartial way. In a research and data-intensive environment this role to broker access to reliable information is naturally transgressing into one of facilitating access to reliable data. For the sake of simplicity data services are categorised as either (1) providing discovery, management and access services to institutional data sources that are contained in ready data-format and which are in the process of entering the scholarly domain, or categorised as (2) providing "findability" and "availability" services of scholarly data sources that are commonly accessible in information-format only, and which have already entered the scholarly domain.

Such data services are defined as research data services and "collections as data" services:

3.1. Research data services

Research data services lie at the intersection of stewardship-oriented and science-oriented support roles, and are shaped to provide management services for research data that are entering the scholarly domain. Research data can be defined as data collected, observed, created or collated to analyse and eventually verify research findings (Boston University Libraries, [2016] & EPSRC, 2016). Research data management (RDM) in turn refers to an explicit process of effectively organizing, structuring, storing and caring for research data – during and after research (Ingram, 2016, DCC, 2016c & University of Edinburgh, 2016a:3). Alternatively stated, RDM entails firstly, planning for the manner in which research data will be managed during and after the research process and, secondly, controlling the collection, processing, analysis, sharing, dissemination, curation and reuse of research data.

Furthermore, organisational structures influences the way work is conducted. As academic libraries are extending their services to research data, careful consideration must be given to how

such a service should be staffed and provisioned. To this effect the RD-Alliance has identified a number of organisational structure archetypes applied in various academic and research institutions to provide research data services:

Solo librarian	RDM services responsibilities assigned to a single person such as a research data librarian within the library. The RDM Librarian is a generalist role, working across the data lifecycle
Dedicated working group	An informal approach pulling staff from different library and campus units to develop and implement RDM services
Multifunctional team	RDM services functions are added to an existing library team (e.g. digital initiatives, scholarly communication)
Specialised team	Appointment of a new team for RDM services

Through RDM services academic libraries aim to discover data science requirements at each stage of the research data lifecycle and build and deploy services to support them. Such RDM services are:

Advocacy

The following are examples of advocacy services:

- Increase awareness of research data management and demonstrate the value of research data management services to the academic community
- Promote data sharing and re-use
- Articulate the benefits of data sharing and re-use

Information services

Provide help and give advice to researchers and students on the following RDM related activities by means of consultation and training and providing relevant information via library guides and RDM webpages:

- Creating data management plans (DMPs) and providing access to DMP tools where applicable
- Finding data (e.g. searching data repositories)
- Analyzing data, including the visualization thereof
- Organizing data, including advice on file naming conventions and metadata
- Storing data, including directing researchers to various existing data repositories

- Sharing data (where applicable, advise researchers on platforms to share datasets)
- Citing data

Data repository management

In collaboration with applicable library and institutional stakeholders, provide the following data repository management services:

- Setting up a repository platform
- Prepare and manage guidelines for the use of the repository (which have to be aligned with an institutional RDM policy)
- Manage ingest of datasets
- Manage metadata and description of datasets
- Overall content management of repository

Since libraries do not provide infrastructure services per se, will the setup of a data repository platform necessitate nascent support and backing from an institutional Information Technology division and other relevant campus stakeholders.

3.2. Collections as data services

Digital collections purchased (i.e. subscribed) by libraries and those curated by libraries as well as campus scholarship in the form of documents and data, could be construed as long-tail big data sources. They can be sourced, analysed and visualised to mine and extract the hidden value they contain on a unit level.

Purchased collections

Libraries broker access to electronic information resources on behalf of researchers and students. They are in many ways still the intermediary between the subscription information provider and the scientist / student, and therefore ideally positioned to facilitate the process of clearing access rights, address ethical concerns, and establishing technical gateways to vendor data sets. A number of consultative library services can be created to assist researchers in studying the scholarly domain from a big data perspective. Doing so requires the skills of stewardship-oriented support roles, with the added support of science-oriented roles.

Copyright and access

The saying "*The right to read is the right to mine, anyone who has lawful access to read the literature with their eyes should be able to do so with a machine.*" refers to the growing

international trend to campaign for the introduction of copyright exceptions in the legislation of nations that govern how data may be accessed. These words may have become the tagline for innovative text and data mining (TDM) projects, but they also reflect a crucial component of the definition of Open Access (OA). While OA is improving the ability for researchers to read papers, it is not in a similar manner addressing the need for researchers to mine them. The facts that are contained in scholarly articles are what make them so useful and valuable. Researchers recognize that the digital environment gives them the opportunity to use these articles, and to make sense of these facts in entirely new ways. They want, and need, the ability to fully use these articles – to freely download and search, text mine, data mine, compute on and crawl them as data – in order to advance their work, to discover, to innovate. Digital articles are, after all, simply small-scale aggregations of digital data.

Yet increasingly, troubling signs are visible that many commercial publishers are unwilling to support users who want to actually **use** the content in scholarly articles and not simply **read** the content in an analogue fashion. In an era when many commercial publishers insist on selling academic institutions access to digital articles only in large bundles – touting the benefits of these bundles as "databases" – restricting the rights of users to fully use the databases is viewed as unacceptable.

Content mining is the way that modern technology locates digital information. Because digitized scientific information comes from hundreds of thousands of different sources in today's globally connected scientific community, and because current data sets can be measured in terabytes, it is often no longer possible to simply read a scholarly summary in order to make scientifically significant use of such information. A researcher must be able to copy information, recombine it with other data and otherwise "re-use" it so as to produce truly helpful results. Not only is data mining a deductive tool to analyse research data, it is how search engines operate to allow discovery of content. To prevent data mining is therefore to force scientists into blind alleys and silos where only limited knowledge is accessible.

As stated, libraries act as intermediaries between publishers, aggregators and the researchers to whom they provide information services. They are well positioned to facilitate access to not only scholarly information sources, but also to scholarly data sources; either through additional subscription contract clauses or campaigning for copyright exceptions on a legislative level. Libraries could do well to identify campus stakeholders (e.g. research departments) and collaborate with leadership institutes (e.g. Big Data Centres) to identify external datasets, and facilitate access to them without having researchers to pass through the customary web-frontend access gateways. This speaks to the need to institute "findability" and "availability" services for

big data sources, which will allow researchers to do data analysis on a unit level, rather than on a representative information level.

Example: in a real-world example the HathiTrust project uses data mining tools to interpret vast volumes of digitized text without violating copyright laws. Computational analysis and metadata is leveraged to collect, connect, and visualise data acquired from large-scale digitized texts (*citation for NMC Horizon Report: 2017 Library Edition, comes here*).

Ethical use

The promise of big data is also accompanied by the ethical challenges of patron privacy and confidentiality during collection, analysis and usage. Related to this is the issue of ethical use in data ownership and proper citation. Researchers may find it hard to ascertain who the "owners" of big data sources are – a necessity when it comes to properly acknowledging the owner and ascertaining the legal rights over re-use. Researchers need to make best efforts to find out the legal implications of re-publishing data, or parts of data, taking into account the license under which it has been published, copyright laws and where fair dealing may or may not apply. For example, owners of social media data like Twitter and Facebook do have re-use terms in their small print, but these can change over time, so should always be consulted for each instance of data publication.

Data sharing places responsibilities on researchers who plan to re-use existing data to provide full and appropriate acknowledgement via citation. Proper citation is important because it helps maximise the impact of research. Yet, big data, being larger and/or more complex than traditional datasets, present challenges in this regard. When citing data, common assumptions, such as the assumption that all links are persistent, must be dispelled. One way to ensure that *research* data that is being cited can be located, is to use a Digital Object Identifier (DOI). This linking mechanism ensures that even if the location of the data changes the DOI will always link to the data that were used. However, big data sources such as social media are unlikely to have formal DOIs, making the process of citation unclear and challenging.

Researchers and students using data in their research are navigating issues and making ethical decisions in ways that are not necessarily taught in their disciplines. Many have only their peers to turn to for difficult questions that could have long-term impacts on their research or reputation. There is an opportunity for librarians to find their role in supporting researchers by navigating emerging ethical issues in their research and especially coordinating efforts with those units within universities that have responsibility for ethics and research integrity. Libraries are also well

positioned to apply their broad knowledge on persistent linking citation mechanisms to try and come up with useful methods to cite (big) data in a lasting and sustainable way.

Metadata

Big data also present metadata challenges. Variables may not be labelled, or the variables may not match the documentation (if there is any). The existence of good metadata is paramount; it is needed to explain the content of data, its provenance and context. Researchers need to know what the data is. It allows them to "match" multiple disparate sources. The need for this becomes increasingly pressing as the number of potentially useful data sources increases in a Big Data world. Libraries could adapt their metadata services for curating collections to organising metadata for data sharing to be more productive.

Example: in a real-world example Shell Australia's technical librarians worked with colleagues in geosciences, information technology, and data management to ensure efficient management of Shell's growing volume of geoscientific data. Their support services included identifying metadata fields, developing controlled vocabularies and naming conventions, defining required search parameters, and developing workflow procedures.

Curated collections

Collections under libraries' curatorship can be analysed on a data unit level to serve as useful data sources for research projects, such as Digital Humanities collaboration projects. Doing so requires the skills of science-oriented support roles, but with the added support of stewardship-oriented support roles. The considerations of copyright and access as well as ethical use that apply to purchased collections would apply to a lesser or greater extent to curated collections as well.

Libraries could do well by doing a data audit of their current special collections to determine which collections are potential candidates for further enrichment or growth through processes such as text analysis and the automatic extraction of (time-based and location-based) information into named entity relationships. Methodologies such as the Data Asset Framework can be applied to existing holdings to identify the internal legacy data which could be of most value for further scholarly research. Data contained in special collections and which is currently unsearchable, can be analysed using big data techniques to allow for image and video searching, voice-to-text conversion, named entity relationships (to logically group people, dates, organisations, etc) and cross-language discovery.

The ability to mine and analyse unstructured data is key to an organisation's competitive advantage in general; an advantage that in an academic and research setting is useful when

collaborating on Digital Humanities projects. The same techniques used in data mining projects, such as natural language processing, could yield fruitful results in text analytics projects. Furthermore, time and location are two of the most fundamental ways in which we organise things. The automatic extraction of geo- and time-based references from the full-text can yield more data than is done through manual tagging techniques.

Example:

A real-world example of such an initiative is the Google Books Ngram Viewer that draws graphs from analysed texts, depicting how phrases have occurred in scholarly corpus over time. Open source equivalents to the Ngram viewer exists, such as Bookworm from Culturomics.

From the above it is evident there is an overlap in the type of services for "collections as data" and those for research data. This confirms that the principles for administering big data are more or less the same as for administering small data. The principles remain the same, they are just applied to different data sources that have different characteristics. Different levels of science-oriented support is also needed between these two classes of data services.

F. In conclusion

Data is one of many economic drivers in the 4th Industrial Revolution, a new currency. It is the fundamental building block of machine learning and the software instruction set for AI. This places libraries in the important position to curate reputable data likewise to how they have before made reputable information available. In doing so, libraries should ideally focus on the methods to foster the knowledge process, rather than concerning themselves with having to deal with the hard characteristics of data, such as the big data V-characteristics. A holistic approach is therefore needed to embed the practice of "data savviness" into the DNA of librarianship. It will require a broad look at all data issues as they present themselves within contemporary data-intensive research practices, and within libraries itself.

--oOo--

References

Boston University Libraries. [2016]. *Research data management*. [Online]. Available from: http://www.bu.edu/datamanagement/background/whatisdata/.

Calzada Prado J and Marzal MA (2013) Incorporating data literacy into information literacy programs: Core competencies and contents. *Libri: International Journal of Libraries & Information Services* 63(2): 123 – 134.

DCC, 2016c. *Glossary.* [Online]. Available from: http://www.dcc.ac.uk/digital-curation/glossary.

EPSRC. 2016. *Scope and benefits*. [Online]. Available from: https://www.epsrc.ac.uk/about/standards/researchdata/scope/ .

Ingram, C. 2016. *How and why you should manage your research data: a guide for researchers*. [Online]. Available from: https://www.jisc.ac.uk/guides/how-and-why-you-should-manage-your-research-data .

University of Edinburgh. 2016a. *A guide to the research data service*. [Online]. Available from: http://www.ed.ac.uk/files/atoms/files/rds_booklet_may2016.pdf.