Mapping the Current Landscape of Research Library Engagement with Emerging Technologies in Research and Learning:

Advancing Digital Scholarship

By Sarah Lippincott

Edited by Mary Lee Kennedy, Clifford Lynch, and Scout Calvert

July 6, 2020

ASSOCIATION OF RESEARCH LIBRARIES

born-digital
RESEARCH + CONSULTING

cni
Coalition for Networked Information

EDUCAUSE

Table of Contents

Landscape Overview 3

Strategic Opportunities 4

Develop data services that work for big data and small data across disciplines 4

Provide and sustain machine-actionable collections 14

Deliver data science education and consultation 17

Key Takeaways 23

Endnotes 24

This is the fifth installment of a forthcoming report, *Mapping the Current Landscape of Research Library Engagement with Emerging Technologies in Research and Learning*, that will be published in its entirety by late summer 2020.

The following installments are being published as they become available at https://doi.org/10.29242/report.emergingtech2020.landscape:

- Executive Summary [published March 26, 2020]
- Introduction, Methodology, and Cross-Cutting Opportunities [published April 2, 2020]
- Facilitating Information Discovery and Use [published April 14, 2020]
- Stewarding the Scholarly and Cultural Record [published May 27, 2020]
- Advancing Digital Scholarship [published July 6, 2020]
- Furthering Learning and Student Success
- Building and Managing Learning and Collaboration Spaces

Landscape Overview

As researchers and students across disciplines explore the affordances of emerging technologies to support scholarly inquiry, many research libraries have built successful digital scholarship programs that respond to the "evolution of the methods for the conduct of research."[1] This section discusses only a sampling of the ways in which libraries have responded to the need for broad access to tools and expertise that advance digital scholarship, treating only those that have demonstrated the most influence from emerging technologies such as machine learning (ML), containerization, and high-performance computing that are the focus of this report. Notably, this section does not go into depth about libraries' significant contributions to digital humanities support, building and maintaining digital scholarship centers, or the hosting and maintenance of digital platforms that allow scholars to develop their own digital projects. It also does not discuss research library management or hosting of digital scholarship centers that provide faculty and students access to the cutting edge technologies, collaboration spaces, and expertise to explore new and emerging forms of scholarly inquiry and creation. Several of these topics are discussed further in the sections on library spaces.

This section instead frames digital scholarship support in the context of how libraries can and do provide the infrastructure, education, and services for data management, analysis, visualization, and curation. Data underlies all digital scholarship, from massive data sets generated continuously by sensors and networked devices, to large corpora of textual evidence, to painstakingly collected and curated image sets. While many library data services have focused on helping researchers manage and deposit data to comply with funder and publisher requirements, scholars increasingly need infrastructure and services that recognize data as a living asset. As they work with massive, complex, heterogeneous, and mutable data sets, scholars need tools and education for analysis, sharing, and publication. Library data services support the full data life cycle: extracting and generating data, preparing it for analysis, publishing or sharing it, and preserving it over the long term. Many of the experts interviewed for this report indicated that libraries have myriad strategic opportunities related to curating digital data and giving communities the skills, support, and infrastructure they need to use them.

The following sections explore the technological developments that are most directly impacting the library's contributions to the digital data life cycle, including evolving infrastructure requirements to facilitate use and reuse of big and small data, the need for digital collections that act like data, and the demand for data science education and consulting services to support scholars and students in the full range of disciplines.

Strategic Opportunities

Develop data services that work for big data[2] and small data across disciplines

The rise of data as both a scholarly input and output[3] has expanded library roles in facilitating access to data collections as source material, and providing solutions for long-term data stewardship. A report examining the future of the University of Texas Libraries asserted that "data is the currency of science, even if publications are still the currency of tenure. To be able to exchange data, communicate it, mine it, reuse it and review it is essential to scientific productivity, collaboration and to discovery itself."[4]

Academic and research libraries are natural partners in data management activities, and many maintain robust and active research data management services. Librarians have the disciplinary, information management, and technology expertise required to manage data throughout its life cycle. The profile of library data services is being shaped by a number of forces, including the expanding emphasis on data-driven research in humanities and social sciences fields and the need for infrastructure and services that recognize data as a living asset. As they work with complex, heterogeneous, and mutable data sets, scholars need tools and education that facilitate analysis, sharing, and preservation. Emphasis on data use and reuse has profound implications for repository infrastructure, entailing a shift from infrastructure optimized for storage and retrieval to one optimized for analysis and sharing. While a few libraries have made strides in this area, most data repository services remain focused on helping scholars meet federal and funder requirements around data deposit. Research libraries also face challenges as they design data services and infrastructure that are sensitive to discovery and analysis methods that vary widely by discipline. Emerging technologies have created three interrelated opportunities for research libraries to expand and evolve their data services: collecting and licensing data sets for scholarly analysis, developing reuse-driven data repository infrastructure, and supporting reproducible science.

Collect and license data sets for scholarly analysis

Many libraries have expanded their collecting activities to include licensing data sets for mining and analysis, providing curated access to publicly available data, and offering guidance on intellectual property laws relevant to the use and reuse of data. Libraries can leverage their information curation expertise and their relationships with vendors to provide collections of (big) data and facilitate access to proprietary or sensitive data for mining and analysis. At New York University (NYU), for example, "the growth of data science throughout the university has influenced the library's collecting, such as purchasing more vendor-produced data sets, responding to students' need for big data (for example, large social media feeds), and integrating APIs into their collection and discovery environment."[5]

Many libraries have already embraced the role of negotiating and interpreting licenses to allow content mining of library collections.[6] As data licensing and collection activities mature, academic libraries have noted the need to implement the same well-documented, systematic workflows generally in place for collecting other scholarly resources.

In many cases, academic libraries purchase or license data sets only in response to specific requests from faculty members. These data sets may not be formally integrated into the library catalog or made available to other potential users. An internal report reviewing the Virginia Tech Libraries' data licensing workflows identified a number of challenges inherent in this *ad hoc* approach.[7] The report noted that data sets were often delivered via CD, USB drive, or hard drive "due to vendor concerns about the security of proprietary data as well as problems involving the online transfer of very large datasets" but that these media lacked corresponding catalog records, making it difficult to control inventory and facilitate discovery.

Cross-institutional research library initiatives are experimenting with approaches to formalize ongoing access to large-scale data sets for scholarly analysis. In 2019, for example, the Big 10 institutions used their collective purchasing power to license 13 terabytes per year of bibliometric data from Web of Science. The CADRE project[8] processes the raw data into a relational database in the high-performance computing center at Indiana University in order to make it available to constituents on the Big 10 campuses. When complete, "CADRE will feature standardized data formats, data available in multiple formats including relational and graph database formats as well as flat tables and native formats, shared and custom/private computational resources, a space to share and store queries, algorithms, derived data, results of analyses, workflows, and visualizations."[9]

The need for broad access to existing data has only grown as researchers in fields as diverse as life sciences and history explore new, technology-enabled ways of interrogating primary source material. A single big data corpus might be mined almost infinitely by different researchers asking different questions, or used by computer scientists to train ML models. To take advantage of the possibilities enabled by both big and small data sets, "researchers who produce those data must share them, and do so in such a way that the data are interpretable and reusable by others."[10] A growing volume of research suggests that the published scientific literature and existing data sets already contain multitudes of hidden hypotheses, insights, and connections, which can be discovered by applying data mining and ML techniques. One study demonstrated, for example, that confirming the existence of the Higgs Boson, which involved years of experimentation and the construction of a new particle accelerator, could have been accomplished through new analyses of existing data.[11] This premise has gained new significance at the time of this writing, as researchers use ML in myriad ways to fight the COVID-19 pandemic by classifying CT scan images, aiding in vaccine development, and attempting to predict new outbreaks.

Building upon established "distant reading" approaches that use computational models, visualization tools, and other methodologies, humanities scholars are also applying ML tools to extract patterns and relationships from text corpora at a scale unattainable by humans. In addition to producing new avenues of humanistic inquiry, applying ML techniques in the digital humanities provides particularly rich opportunities for critical reflection and action regarding ethical and transparent use of ML.[12]

Develop infrastructure that supports data use and reuse

The demand for infrastructure that supports data sharing and long-term preservation has grown commensurately with funder and publisher data deposit requirements, and evolving research regarding data sharing. Library-maintained data repositories, disciplinary repositories, and general purpose repositories (for example, figshare and Zenodo) have proliferated. However, with several notable exceptions, libraries have invested more in data management *services* than *infrastructure*.[13] In addition to the valuable data management planning and consultative services that libraries routinely provide, scholars also require infrastructure that supports very large, heterogeneous, living, networked, and complex data sets in a range of formats. They desire infrastructure that facilitates (geographically distributed) collaboration, data reuse, and long-term preservation. The research library model of data repositories does not always align with these expectations. The current data repository model tends to support "highly derived, processed data sets that support a paper," while faculty desire "a living organism, a database that is in continuous development."[14]

Emphasis on use and reuse has profound implications for repository infrastructure, entailing a shift from infrastructure optimized for storage and retrieval to one optimized for analysis and sharing. The Virginia Tech Libraries has embraced a use-and-reuse framework as "the driving force behind" its data management infrastructure and services.[15] It has become increasingly difficult to divorce scholarly datasets from the algorithms and computing environments used to create, display, or interpret them. Even with extensive documentation of such "data and their usage context, mummifying live data out of their natural habitats of analysis to be preserved in an isolated vault can significantly diminish their value."[16]

Unlike many data repositories optimized for data archiving, a reuse-driven data repository is designed to support built-in analysis tools and the co-location of data with computing resources and to enable ongoing collaboration, including granular permissions options and access by geographically distributed teams. A use-and-reuse driven repository resembles "a lively workshop equipped with powerful tools to handle big data sets as the raw materials," rather than an attic or warehouse for data storage.[17] This idea is echoed in other metaphors that reconceptualize data as a living asset: the idea of moving from reservoirs to rivers of data[18] and from data stock to data flows.[19]

Built-in visualization tools are becoming a popular feature in data repositories as they facilitate preview before download and a basic level of access for users that lack specialized software. PURR, Purdue University Libraries' research data repository, has incorporated geospatial data visualization tools by adding a GIS server to their repository infrastructure. The web mapping capabilities effectively allow end users to preview a data set and determine its relevance to their research interests before downloading and without requiring the specialized software generally needed to view and manipulate much geospatial data.[20] The University of Virginia (UVA) Library in collaboration with the UVA Institute for Advanced Technology in the Humanities (IATH) has also implemented this approach for 3D data, creating an enhanced interface for digital data sets stored in Dataverse, which "uses the open-source web 3D viewer 3D Heritage Online Presenter (3DHOP) to provide an interactive 3D model for users to explore the data before download."[21]

The built-in tools supported by reuse-driven data repositories might one day include ML models that automatically process data on ingest, leading to new methods of discovery and analysis. The experimental ScienceSearch tool, for example, aims to make searchable a massive collection of largely undescribed micrographs (images captured with the aid of a microscope) from The National Center For Electron Microscopy (NCEM) at the Molecular Foundry at Lawrence Berkeley National Laboratory. The tool runs analysis as data is ingested into the repository, aggregating information from computer vision techniques, text analysis, and extant metadata.[22]

To enable collaboration, reuse-driven data repositories are taking advantage of tools that reduce the computing resources and effort needed to work with distributed teams and decentralized data sets. The iRODS data management software, for example, virtualizes its data storage resources so that users can access data regardless of their geographic location or device.[23] Data virtualization allows users to query across systems, rather than downloading to a single device or copying data between systems.

As researchers seek to extract meaning from ever increasing volumes of data through mining and other data processing methods, they need ever greater access to computing power.[24] One expert interviewed for this report cautioned that "research libraries and research computing should not be evolving separately" and cited a need for programmatic partnerships between research libraries and research computing centers to ensure alignment between computing needs and data curation needs. Researchers are applying a number of emerging technologies to build computing capacity and accelerate computing tasks, including multiprocessor systems, graphic processing units (GPUs), and field programmable gate array (FPGA) devices. Experts interviewed for this report also cited the need for co-located storage and computing nodes.[25] Researchers working with massive data sets in geographically distributed teams need access to high-speed networking to facilitate large-scale data transfer, analytics, and storage. In some research communities, "shipping hard drives is still the preferred option to move data when the size reaches a certain threshold" as users confront network speed and processing capacity limits when attempting to access or download large data sets.[26]

Providing the infrastructure for high-speed networking requires cooperation at a national level. The NSF-funded Pacific Research Platform (PRP) represents one attempt at regional coordination, which will give "data-intensive researchers at participating institutions the ability to move data 1,000 times faster compared to speeds on today's inter-campus shared Internet."[27] An NSF-funded follow-up project envisions scaling this approach to develop a National Research Platform (NRP) that would facilitate access to distributed data sets and allow researchers to leverage the computing and storage resources of national supercomputer facilities.

At many institutions, research computing infrastructure is gradually moving from local data centers to the cloud. Businesses and researchers alike are turning to the cloud for access to AI and ML tools, blockchain, and more.[28] Cloud computing facilitates collaboration between distributed teams, provides co-located data storage and processing capacity, and provides solutions for researchers who do not have access to local computing resources. However, it also comes with risks. Data stored in a commercial cloud is no longer fully under a researcher's control. It is vulnerable to breaches, hacks, or catastrophic loss. Depending on the specific services being used, researchers may also be giving permission (knowingly or not) to third parties to access or use their data. Whether they store data in the cloud or in local data centers, libraries that host data repository infrastructure must consider whether they can provide cybersecurity commensurate with the sensitivity of personally identifiable data, especially if it is being actively used.

The future of data-intensive research support and data management will require libraries to work beyond institutional boundaries. In addition to or in lieu of organizing data repositories around institutional affiliation, research libraries may invest in supporting cross-institutional groups of researchers affiliated by discipline or research interest, through infrastructure, curation guidance, intellectual property expertise, and community building.[29] These "data communities" (which often comprise infrastructure alongside informal and formal knowledge sharing and collaboration) might receive financial and human resources from a research library, or might collaborate with librarians as campus ambassadors and curators. While disciplinary and other public data repositories (such as figshare) have demonstrated high deposit rates and engagement, they lack the institutional connections and relationships that campus data curators and research librarians can build. Coordination and collaboration between institution-based data experts and institution-independent data repositories can advance open science practices and FAIR data principles by "ensuring that researchers follow best practices and their outputs are preserved and reusable."[30]

Support reproducible science

Scientific progress depends on research that can be validated, built upon, and repurposed. As more and more scientific research is conducted using computationally intensive methodologies, validating and reproducing results has become infinitely more complicated. Research library data services support reproducible science through educational and awareness efforts that encourage scholars to apply appropriate disciplinary standards; to deposit data in open repositories; and to structure, document, and license data with human and machine reuse in mind. Libraries are also contributing to the development of software and infrastructure that facilitates the creation and preservation of reproducible data sets.

To reproduce results, scientists need access not only to well-documented, openly available data, but also to the code used to process and analyze it. In order to support an open science environment, “access to the computational steps taken to process data and generate findings is as important as access to the data themselves.”[31] The electronic lab notebooks where many data scientists conduct exploratory research do not natively support broad sharing or publication. A notebook’s dependencies on its environment make its behaviors unpredictable when shared with colleagues; the same code may produce different results in a different environment, or fail to compute entirely.[32]

Virtual containers offer one solution to this challenge. Container technology, or containerization, is often described as “a lightweight alternative to virtual machines” that bundles code, software, and an operating system such that users can accurately reproduce computational research. Container technologies like Docker[33] and Singularity[34] have seen widespread adoption as a way to “encapsulate a software environment (e.g. a complex software tool-chain including application-specific settings) into a single portable entity.”[35] Projects such as CiTAR,[36] ReproZip,[37] and Binder[38] aim to make reproducibility via containerization technologies broadly accessible to the academic research community. ReproZip works by “automatically tracing the execution of work and then packaging all dependencies in a single, distributable package” (an RPZ file), and is compatible with a wide range of data analysis tools, scripting and software languages, databases, and electronic lab notebooks like Jupyter.[39] Binder can retrieve Jupyter notebooks hosted in a Git repository, build a container image to serve them, and make that image publicly available to anyone on the web.[40]

Libraries are supporting reproducibility by building and providing access to the tools needed to reproduce computationally intensive research and by creating and redefining staff roles to explicitly include reproducibility support. NYU first created a dedicated position in service of reproducibility in 2017; the University of Florida Libraries recently advertised a similar position. At NYU, the librarian for research data management and reproducibility position is a dual appointment shared by the Division of Libraries and the Center for Data Science (CDS) and is responsible for education and outreach, as well as tool and infrastructure building in support of data services.[41] At the University of Arizona Libraries, "support for reproducibility has taken the form of integrating best practices for data management, promotion of scripting/software to automate workflows, promotion of tools that support reproducible research (e.g., Jupyter notebooks), and advocating for open research practices into workshops and lectures."[42] A University of Texas Libraries report on the future of the research library predicts that librarians will "become embedded partners that enable researchers to do their work in an environment where research data, lab notes and other research process are freely available under terms that enable reuse, redistribution and reproduction of the research and its underlying data and methods" and will become more attuned to discipline-specific research methods.[43] An inaugural "Librarians Building Momentum for Reproducibility" conference in 2020 explored the many facets of library contributions to reproducibility, including incorporating reproducibility education into graduate and undergraduate programs of study, investigating emulation services and other library-managed tools, and applying principles of reproducibility to library science research.[44]

Highlighted Initiatives

Collaborative Archive & Data Research Environment (CADRE)
Indiana University Libraries
https://cadre.iu.edu/
The CADRE project processes raw data from Web of Science and other major datasets into a relational database in the high-performance computing center at Indiana University in order to

make it available to constituents on the Big 10 campuses. When complete, “CADRE will feature standardized data formats, data available in multiple formats including relational and graph database formats as well as flat tables and native formats, shared and custom/private computational resources, a space to share and store queries, algorithms, derived data, results of analyses, workflows, and visualizations.”[45]

ReproZip
New York University
https://www.reprozip.org/
The ReproZip software package being developed at New York University (NYU) facilitates reproducible research by packaging the files and dependencies necessary to replicate results. ReproZip is compatible with a wide range of data analysis tools, scripting and software languages, databases, and electronic lab notebooks like Jupyter. The team behind ReproZip includes NYU Libraries’ librarian for research data management and reproducibility.

Provide and sustain machine-actionable collections

Data scientists, humanists, and social scientists are increasingly looking to library collections as data sources for creating and uncovering new knowledge. The potential advantages of library collections for computational research are manifold: they often contain high-quality human-generated metadata, some are open access and may have fewer restrictions on use for data mining, and many are already structured using standards that are machine-readable. Initiatives such as the Collections as Data project encourage cultural heritage institutions to thoughtfully develop digital collections (licensed, purchased, and unique) and allied services (for example, workshops, consultations, digital platforms) that support “computationally-driven research and teaching.”[46] Research libraries can further contribute to building machine-readable collections by developing and implementing processes to extract data from text or other media, clean it, and supply it in a database or other format suitable for analysis.[47]

A 2018 report from the National Academies described a speculative future in which "researchers have immediate access to the most recent publications and have the freedom to search archives of papers, including preprints, research software code, and other open publications, as well as databases of research results, including digital information related to physical specimens, all without charge or other barriers. Researchers use the latest database and text mining tools to explore these resources, to identify new concepts embedded in the research, and to identify where novel contributions can be made."[48] This vision is predicated on the availability of machine-actionable collections, a premise that has significant legal, technical, and policy implications for libraries. Beyond the sciences, the deep reading methods that have long characterized academic inquiry in the humanities and social sciences are also being supplemented by approaches that require access to "amalgamated collections in order to conduct various forms of computational research."[49]

Digitized and born-digital special collections hold particular promise for researchers as unique assets that can lead to data-driven insights about specific places and communities. Using a Collections as Data framework, libraries can add further value to these unique and valuable materials by making them machine-readable. For example, librarians in University of Utah's Digital Library and Digital Matters programs have explored the feasibility of applying computational analysis to digitized special collections materials relevant to their community, such as mapping the history of black Mormons.[50]

The technical affordances of machine-actionable library collections make them ideal not only for human-driven computational research, but for the development of AI and ML. AI and ML tools rely on large quantities of structured data to become proficient at a task, and in the near future, machines and AI training algorithms may become major users of library collections. A recent post from the IFLA Library Policy and Advocacy Blog noted that library collections "contain the richest imaginable resource" for developing ML technologies given that ML tools learn by "looking at existing materials and drawing new connections and conclusions."[51] The same post contends that ML "opens up some truly exciting possibilities to do more with works already in collections (as long as they are digitised, open access, and ideally have the right metadata to be used across institutions)." These caveats underscore the continuing relevance of librarians' roles in collection development, curation, advocacy, and standards development. The post's author cautioned that progress in ML may be constrained by the resources required to prepare data for machine-learning applications, which can require vastly greater effort than the machine learning work itself.[52]

However, the pitfalls of training AI on library collections are many. The authors of "The Santa Barbara Statement on Collections as Data" note that "the scale of some collections may also obfuscate what is hidden or missing in the histories they are perceived to represent. Cultural heritage institutions must be mindful of these absences and plan to work against their repetition."[53] Much like controversial practices such as predictive policing that attempt to predict crime and recidivism through computational analyses of historical criminal data, big data analyses of digitized library collections have the potential to unearth new "insights" that reproduce and even amplify cultural biases and historical racism. The statement encourages librarians to "critically engage with bias in collection and description, archival silences, and assumptions about collection use" when developing machine-actionable collections for use.[54]

Delivering machine-actionable collections presents socio-technical challenges along with political and cultural ones. The technical processes necessary to create structured data also operate in a complex legal framework of negotiating terms of access with publishers and special collections and archives to allow data mining to take place. Borgman notes that despite the broad success of the open access (OA) movement in providing free access to scholarly information, the reader of OA texts is still presumed by OA publishers to be "a human user who is capable of reading a web page, searching for content, and selecting individual items for download…Robots may or may not be allowed to search open access databases."[55] Forward-looking OA advocacy must engage with the rights of non-human readers as part of a free and open scholarly landscape.

Highlighted Initiatives

Always Already Computational: Collections as Data
https://collectionsasdata.github.io/
The first phase of the Collections as Data project "documented, iterated on, and shared current and potential approaches to developing cultural heritage collections that support computationally-driven research and teaching." The next phase, Collections as Data: Part to Whole, is funded by the Mellon Foundation and "aims to foster the development of broadly viable models that support implementation *and* use of collections as data."

***Woman's Exponent* Modeling the Corpus Tool**
University of Utah Marriott Library
https://exhibits.lib.utah.edu/s/womanexponent/page/modeling-the-corpus
Librarians at University of Utah have digitized the entire run of *Woman's Exponent*, a Salt Lake City–based newspaper focusing on Mormon women, and developed data-mining tools for web-based inquiry, as well as provided downloadable access to the corpus.

Deliver data science education and consultation

In the past decade, data science has moved from a niche field to ubiquitous, and from the domain of a small group of researchers in STEM fields to omnipresent across many domains. At the same time, the big-data era has created new challenges for researchers across the disciplinary spectrum, whose "capability to generate or manipulate data through e-science experiments has far surpassed their ability to manage, organize, or make their data easily accessible."[56] Researchers can now passively generate terabytes of complex data through the use of networked sensors, mining and scraping techniques, and other methods. A National Academies report asserts that “many, if not most, areas of science now involve computational analysis of often very large data sets;”[57] and researchers in humanities and social sciences fields are also turning to data-intensive methods to open new avenues of inquiry. As data science programs proliferate, even undergraduate students will routinely need access to resources for big-data analytics.

As the “ubiquitous availability of sensing technologies, the [w]eb, and the [c]loud” have democratized access to vast quantities of data, researchers often lack the necessary “experience and expertise to effectively extract values from the large data sets.”[58] Working with big data is challenging not only because of its volume, but “its exhaustivity and variety, timeliness and dynamism, messiness and uncertainty, high relationality, and the fact that much of what is generated has no specific question in mind or is a by-product of another activity.”[59] Big-data analysis therefore relies heavily on AI (specifically convolutional neural networks and recurrent neural networks) to analyze data and detect patterns, allowing researchers to gain insights from the data without requiring a formal hypothesis or even notion of what they might be looking for.

This increasing emphasis on data- and computationally intensive research methods creates opportunities for libraries to contribute to the education, tools, infrastructure, and communities that sustain and expand these practices. Given the complexity of big-data analysis and the specialized skills it requires, educational and consulting services are essential across the disciplinary spectrum. Libraries have an opportunity to support both experienced researchers working on cutting-edge projects and novice researchers and students taking their first steps into data science. A number of libraries have launched educational and consulting programs in support of data science tools—hosting one-off workshops and workshops series, interest groups, semester-long collaboration programs, conferences, and other community-building activities—and are positioning themselves as hubs for faculty and student engagement around e-research.

Many libraries have identified a niche in tailoring their educational offerings to faculty members and students outside of STEM fields, who may lack opportunities within their department or program of study. A core goal of data science services at the UC Berkeley Libraries, for example, is to "demystify data science for the campus community, building new pipelines into the field from all directions."[60] Bringing the affordances of big-data analytics to research communities in the humanities and social sciences allows scholars in those fields to explore new avenues of inquiry and also breaks down perceptions of data science as objective and fact-based, as opposed to the subjective and speculative methods of the social sciences and humanities. Libraries can encourage their communities to think critically about data science as it "reframes key questions about the constitution of knowledge, the processes of research, how we should engage with information, and the nature and the categorization of reality," and "risks reinscribing established divisions in the long running debates about scientific method and the legitimacy of social science and humanistic inquiry."[61]

To bring data science to scholars and students across disciplines, a number of libraries have launched educational programs that comprise workshops and non-credit courses. At Georgia Tech, for example, several librarians are collaborating to offer non-credit courses in 3D modeling, programming languages, web scraping, and other data science and digital scholarship methodologies, along with data literacy courses targeted at students in non-data-intensive majors. Columbia University Libraries offer a Foundations for Research Computing course that "provides informal training for Columbia University graduate students and postdoctoral scholars to develop fundamental skills for harnessing computation" and aims to build a community of researchers using computationally intensive methods. At the University of Arizona Libraries, librarians have adapted their digital scholarship workshops over time to better meet the needs of their audience. The librarians found that workshops that aimed to teach programming languages using a conceptual approach "left many participants wondering how to apply what they learned to their own work."[62]

This realization led the libraries to create topic-specific workshops, still appropriate for novices, that make a clearer connection with participants' research goals.

Other libraries are developing lab-based models, inviting collaborative teams to work through data science and digital scholarship challenges. The 99 AI Challenge[63] sponsored by the University of Toronto Libraries, for example, will bring together 99 students, staff, faculty, and other community members with no technical background to collectively learn about and critically engage with AI technologies. The project-focused or lab model encourages deeper engagement and can forge long-term partnerships. It can also help libraries provide responsible, sustainable support for emerging technology projects by inviting "partners from libraries and information technology organizations to help create generalizable solutions and best practices that fit the scholarly questions at the heart of the lab's mission."[64]

Libraries face many challenges in hiring expert data scientists, yet data science education and consulting services must be powered by skilled librarians. At the University of Arizona Libraries, in-house data science specialist Jeffrey Oliver collaborates with other librarians in the data management program and provides "bioinformatic support to life science researchers, especially in data analysis and visualization." While Oliver acknowledges that "the library cannot offer a concierge data analyst service to every researcher on campus," the program plays a critical role in connecting researchers with appropriate resources within the library and externally, providing a basic level of education and guidance, and developing long-term research partnerships.[65] Upskilling existing staff provides a good alternative when hiring for data skills is not feasible. Librarians' traditional skills in information management can be complemented by training programs, such as North Carolina State University's currently inactive Data Science and Visualization Institute for Librarians, to provide librarians the ability to develop new skills in data science.[66] However, in some cases, librarians' professional development can be hampered by managers who may not understand the need for staff to develop data science skills, or "how to vertically and horizontally integrate data-centric practices into their organizations and envision the diverse contexts, opportunities, and benefits in applying data science methods."[67]

Libraries' technical contributions to data science support include providing infrastructure such as data repositories and clouds with co-located computing resources (as discussed in the previous section), as well as supporting the software and tools commonly used by data scientists, such as electronic lab notebooks. In many data science courses, instructors need new approaches for "providing an interactive, online environment where students can run code via the cloud without requiring them to download anything onto their machine."[68] Containerization technologies (such as Docker) provide one promising option. Course materials for a data science course developed in a Docker container will work consistently across a range of devices and platforms, allowing students to interact with dynamic, code-driven instructional materials without worrying about the effect of their operating system.

Faculty members and students in STEM fields, including rapidly growing data science programs, increasingly require considerable computing resources for their coursework. Students may be expected to access and analyze big data, utilize software that requires computing resources beyond the capacity of a typical laptop, or develop and test code. This type of computationally intensive instruction relies on "significant cloud-based and local computational resources to enable ambitious instructional projects," including statistics, engineering, and math software, as well as high-performance computing clusters and big data processing power.[69] The Dataspace, a new high-performance computing space in North Carolina State University's Hunt Library, provides students "access to the tools and training needed to develop critical data science skills", including reservable data workstations with high-capacity storage, processing power, and specialized software, as well as workshops and services targeted at students and faculty.[70]

Many of the experts interviewed for this report identified recruiting or upskilling library workers with data science skills as an imperative, but particularly challenging, aspect of building data and data science services. While some data science skills align well with librarians' strengths, it is unlikely that most libraries will be able to employ teams of in-house data scientists. Intense demand for professionals with data science skills and experience make it difficult for libraries to compete with the salaries and perks available in the corporate world, and "the incentive structures for mid-career librarians can be misaligned or opposed to the development of technical skills."[71]

Highlighted Initiatives

Data Science and Visualization Institute for Librarians (DSVIL)
NC State University Libraries
https://www.lib.ncsu.edu/data-science-and-visualization-institute
Although currently inactive, DSVIL has addressed the current skills gap in data science for librarians by offering a series of one-week intensive trainings on software tools and skills relevant to data analysis, visualization, sharing, and reuse.

Institute for Data Intensive Engineering and Science (IDIES)
Johns Hopkins University
http://idies.jhu.edu/
IDIES, a partnership of the Sheridan Libraries at Johns Hopkins University (JHU) with the schools of public health, business, arts and sciences, medicine, and engineering, seeks to create a complete suite of services, data sets, and education opportunities around data science for faculty, staff, and student members of the JHU community.

99 AI Challenge
University of Toronto Libraries
https://onesearch.library.utoronto.ca/ai-challenge
The 99 AI Challenge sponsored by the University of Toronto Libraries is bringing together 99 students, staff, faculty, and other community

members with no technical background to collectively learn about and critically engage with AI technologies.

Key Takeaways

- **Data is a living, networked asset.** Library data services have long focused on infrastructure, education, and advocacy to support data archiving. Emerging technologies and shifting researcher expectations are engendering a shift towards data services that center data use and reuse. A use- and reuse-driven approach to data services implies development of infrastructure that natively supports data analysis and active collaboration; use of software and workflows that package research data sets alongside the code and operating systems necessary to interpret them and reproduce results; and continuing advocacy for licensing terms that explicitly support data reuse, repurposing, and mining.
- **Research libraries add value to their digital scholarly and special collections by making them machine-readable and actionable.** Research libraries are preparing for a future in which human and machine users derive insight from digital collections through data mining and analysis. Investments in machine-actionability further bolster the value of unique digitized and born-digital collections, some of the research library's most valuable resources.
- **Research libraries foster critical engagement with data.** Library-led workshops and educational programming can bring critical perspectives to bear on technologies often considered "neutral." Bringing the affordances of big-data analytics to research communities in the humanities and social sciences allows scholars in those fields to explore new avenues of inquiry and also breaks down perceptions of data science as objective and fact-based, as opposed to the subjective and speculative methods of the social sciences and humanities.
- **Research librarians and managers need administrative support to re-skill and develop data science skills.** As they expand data

services, research libraries will face a shortage of skilled data and data science professionals to fill high-demand roles. Data science skills are in short supply. Research libraries will face intense competition from industry for professionals with data science education and experience. Re-skilling the existing workforce may prove challenging as research librarians balance new competencies with existing responsibilities.

Endnotes

1. Carole Palmer, interview by author, November 20, 2019.

2. There are many definitions of big data. This report may be helpful to the reader: *NIST Big Data Interoperability Framework: Volume 1, Definitions*, NIST Special Publication 1500-1 (Washington, DC: US Department of Commerce, National Institute of Standards and Technology, September 16, 2015), https://bigdatawg.nist.gov/_uploadfiles/NIST.SP.1500-1.pdf.

3. Jean-Christophe Plantin, Carl Lagoze, and Paul N. Edwards, "Re-Integrating Scholarly Infrastructure: The Ambiguous Role of Data Sharing Platforms," *Big Data & Society* 5, no. 1 (January–June 2018): 1–14, https://doi.org/10.1177/2053951718756683.

4. Michelle Addington and Lorraine Haricombe, *Task Force on the Future of UT Libraries: Final Report* (Austin: University of Texas at Austin, 2019), https://provost.utexas.edu/future-university-texas-libraries-task-force.

5. Jennifer Muilenburg and Judy Ruttenberg, "New Collaboration for New Education: Libraries in the Moore-Sloan Data Science Environments," *Research Library Issues*, no. 298 (2019): 16–27, https://doi.org/10.29242/rli.298.3.

6. Marco Caspers and Lucie Guibault, *Baseline Report of Policies and Barriers of TDM in Europe* (Vienna: FutureTDM, 2016), https://www.futuretdm.eu/wp-content/uploads/FutureTDM_D3.3-

Baseline-Report-of-Policies-and-Barriers-of-TDM-in-Europe-1.pdf.

7. Philip Young et al., “Library Support for Text and Data Mining: A Report for the University Libraries at Virginia Tech,” June 22, 2017, http://hdl.handle.net/10919/78466.

8. “CADRE,” Indiana University Network Science Institute, accessed June 23, 2020, https://iuni.iu.edu/resources/datasets/cadre.

9. “Collaborative Archive & Data Research Environment,” Indiana University Network Science Institute, accessed May 26, 2020, https://iuni.iu.edu/projects/cadre.

10. Rob Kitchin, “Big Data, New Epistemologies and Paradigm Shifts,” *Big Data & Society* 1, no. 1 (April–June 2014), https://doi.org/10.1177/2053951714528481.

11. Gonzalo P. Rodrigo et al., “ScienceSearch: Enabling Search through Automatic Metadata Generation,” in *2018 IEEE 14th International Conference on E-Science (e-Science)* (IEEE, 2018): 93–104, https://doi.org/10.1109/eScience.2018.00025.

12. Caroline Bassett et al., “Critical Digital Humanities and Machine Learning,” in *Digital Humanities 2017: Conference Abstracts* (Montréal: McGill University and Université de Montréal, 2017), 36–40, https://dh2017.adho.org/abstracts/DH2017-abstracts.pdf.

13. Ixchel M. Faniel and Lynn Silipigni Connaway, “Librarians’ Perspectives on the Factors Influencing Research Data Management Programs,” *College & Research Libraries* 79, no. 1 (January 2018): 100–19, https://doi.org/10.5860/crl.79.1.100.

14. Kristin Antelman, interview by author, November 15, 2019.

15. Zhiwu Xie and Edward A. Fox, “Advancing Library Cyberinfrastructure for Big Data Sharing and Reuse,” *Information Services & Use* 37, no. 3 (2017): 319–23, https://doi.org/10.3233/ISU-170853.

16. Xie and Fox, “Advancing Library Cyberinfrastructure.”

17. Zhiwu Xie et al., “Towards Use and Reuse Driven Big Data Management,” in *JCDL ‘15: Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries* (New York: Association for Computing Machinery, 2015), 65–74, https://doi.org/10.1145/2756406.2756924.

18. Lorcan Dempsey, “Libraries and the Informational Future: Some Notes,” *Information Services & Use* 32, no. 3–4 (2012): 203–14, https://doi.org/10.3233/ISU-2012-0670.

19. David Kremers, Kristin Antelman, and Stephen Davison, “From Stock to Flows” (slides presented at CNI Fall 2017 Membership Meeting, Washington, DC, December 12, 2017), https://www.cni.org/topics/digital-curation/from-stock-to-flows.

20. Yue Li, Nicole Kong, and Stanislav Pejša, “Designing the Cyberinfrastructure for Spatial Data Curation, Visualization, and Sharing,” *IASSIST Quarterly* 41, no. 1–4 (December 10, 2017), https://doi.org/10.29173/iq11.

21. Will Rourk, “3D Cultural Heritage Informatics: Applications to 3D Data Curation,” in *3D/VR in the Academic Library: Emerging Practices and Trends*, ed. Jennifer Grayburn et al. (Arlington, VA: Council on Library and Information Resources, February 2019), https://www.clir.org/pubs/reports/pub176/.

22. Rodrigo et al., “ScienceSearch.”

23. iRODS website, accessed May 26, 2020, https://irods.org/.

24. P. Škoda, B. Medved Rogina, and V. Sruk, “FPGA Implementations of Data Mining Algorithms,” in *2012 Proceedings of the 35th International Convention MIPRO* (IEEE, 2012), 362–67, https://bib.irb.hr/datoteka/625836.dc-vis_022.pdf.

25. Škoda et al., “FPGA Implementations of Data Mining Algorithms”.

26. Xie et al., “Towards Use and Reuse Driven Big Data Management.”

27. Richard Moore, *The Second National Research Platform Workshop: Toward a National Big Data Superhighway*, ed. Tom DeFanti and Maxine Brown (National Research Platform, September 20, 2018), http://pacificresearchplatform.org/images/reports/2NRP_Workshop_Report_final-small-9-20-18.pdf.

28. Nitin Mittal,Dave Kuder, and Samir Hans, "AI-Fueled Organizations," in *Tech Trends 2019: Beyond the Digital Frontier*, ed. Bill Briggs and Scott Buchholz (Deloitte Development, 2019), 21–39, https://www2.deloitte.com/be/en/pages/technology/articles/tech-trends-2019-beyond-the-digital-frontier.html.

29. Danielle Cooper and Rebecca Springer, *Data Communities: A New Model for Supporting STEM Data Sharing*, Issue Brief (New York: Ithaka S+R, May 13, 2019), https://doi.org/10.18665/sr.311396.

30. John Chodacki, Daniella Lowenberg, and Elizabeth Hull, "Advancing Data Publishing: The Future of Dryad," abstract IN52B-07 (presentation at American Geophysical Union Fall Meeting, Washington, DC, December 14, 2018), https://ui.adsabs.harvard.edu/abs/2018AGUFMIN52B..07C/abstract.

31. Victoria Stodden et al., "Enhancing Reproducibility for Computational Methods," *Science* 354, no. 6317 (December 9, 2016): 1240–41, https://doi.org/10.1126/science.aah6168.

32. William Benton and Sophie Watson, "Why Data Scientists Love Kubernetes," *Opensource.com*, January 4, 2019, https://opensource.com/article/19/1/why-data-scientists-love-kubernetes.

33. Docker website, accessed June 23, 2020, https://www.docker.com/.

34. "Singularity Examples," Sylabs, accessed June 23, 2020, https://sylabs.io/docs/.

35. Klaus Rechert, "Preserving Containers—Introducing CiTAR Part 2," *Open Preservation Foundation Blog*, January 23, 2019, https://openpreservation.org/blog/2019/01/23/preserving-containers-introducing-citar-part-2/.

36. Rechert, “Preserving Containers.”

37. ReproZip website, accessed June 23, 2020, https://www.reprozip.org/.

38. Binder website, accessed June 23, 2020, https://mybinder.org/.

39. Vicky Steeves, Rémi Rampin, and Fernando Chirigati, “Using ReproZip for Reproducibility and Library Services,” *IASSIST Quarterly* 42, no. 1 (2018), https://doi.org/10.29173/iq18.

40. Benton and Watson, “Why Data Scientists Love Kubernetes.”

41. Vicky Steeves, “Reproducibility Librarianship,” *Collaborative Librarianship* 9, no. 2 (2017): 80–89, https://digitalcommons.du.edu/collaborativelibrarianship/vol9/iss2/4/.

42. Jeffrey C. Oliver et al., “Data Science Support at the Academic Library,” *Journal of Library Administration* 59, no. 3 (2019): 241–57, https://doi.org/10.1080/01930826.2019.1583015.

43. Addington and Haricombe, *Task Force on the Future of UT Libraries*.

44. “Librarians Building Momentum for Reproducibility,” Vicky Steeves’s GitLab site, accessed June 23, 2020, https://vickysteeves.gitlab.io/librarians-reproducibility/.

45. “Collaborative Archive & Data Research Environment,” Indiana University Network Science Institute.

46. Always Already Computational website, accessed May 26, 2020, https://collectionsasdata.github.io/.

47. MacKenzie Smith, interview by author, November 13, 2019.

48. National Academies of Sciences, Engineering, and Medicine, *Open Science by Design: Realizing a Vision for 21st Century Research* (Washington, DC: National Academies Press, 2018), 4, https://doi.org/10.17226/25116.

49. Oya Y. Rieger, *What's a Collection Anyway?*, Issue Brief (New York: Ithaka S+R, June 6, 2019), https://doi.org/10.18665/sr.311525.

50. Rachel Wittmann et al., "From Digital Library to Open Datasets: Embracing a 'Collections as Data' Framework," *Information Technology and Libraries* 38, no. 4 (December 2019): 49–61, https://doi.org/10.6017/ital.v38i4.11101.

51. "The Robots Are Coming? Libraries and Artificial Intelligence," *IFLA Library Policy and Advocacy Blog*, July 24, 2018, http://blogs.ifla.org/lpa/2018/07/24/the-robots-are-coming-libraries-and-artificial-intelligence/.

52. Clifford A. Lynch, "Machine Learning, Archives and Special Collections: A High Level View," *ICA Blog*, International Council on Archives, October 2, 2019, https://blog-ica.org/2019/10/02/machine-learning-archives-and-special-collections-a-high-level-view/.

53. Thomas Padilla et al., "The Santa Barbara Statement on Collections as Data," May 20, 2019, https://doi.org/10.5281/ZENODO.3066209.

54. Padilla et al., "The Santa Barbara Statement."

55. Christine L. Borgman, "Whose Text, Whose Mining, and to Whose Benefit?," accepted for publication in *Quantitative Social Sciences*, December 3, 2020, https://escholarship.org/uc/item/3682b9j6.

56. Jake R. Carlson and Jeremy R. Garritano, "E-Science, Cyberinfrastructure and the Changing Face of Scholarship: Organizing for New Models of Research Support at the Purdue University Libraries," in *The Expert Library: Staffing, Sustaining, and Advancing the Academic Library in the 21st Century*, ed. Scott Walter and Karen Williams (Chicago: Association of College and Research Libraries, 2010), 234–69, https://docs.lib.purdue.edu/lib_research/137/.

57. National Academies of Sciences, Engineering, and Medicine, *Open Science by Design*.

58. Xie and Fox, “Advancing Library Cyberinfrastructure.”

59. Kitchin, “Big Data, New Epistemologies and Paradigm Shifts.”

60. Muilenburg and Ruttenberg, “New Collaboration for New Education.”

61. danah boyd and Kate Crawford, “Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon,” *Information, Communication & Society* 15, no. 5 (2012): 662–79, https://doi.org/10.1080/1369118X.2012.678878.

62. Oliver et al., “Data Science Support.”

63. “The 99 AI Challenge,” University of Toronto Libraries, accessed June 23, 2020, https://onesearch.library.utoronto.ca/ai-challenge.

64. Victoria Szabo, “Collaborative and Lab-Based Approaches to 3D and VR/AR in the Humanities,” in *3D/VR in the Academic Library: Emerging Practices and Trends*, ed. Jennifer Grayburn et al. (Arlington, VA: Council on Library and Information Resources, February 2019), https://www.clir.org/pubs/reports/pub176/.

65. Oliver et al., “Data Science Support.”

66. “Data Science and Visualization Institute,” NC State University Libraries, accessed May 26, 2020, https://www.lib.ncsu.edu/data-science-and-visualization-institute.

67. Matt Burton et al., *Shifting to Data Savvy: The Future of Data Science In Libraries* (Pittsburgh: University of Pittsburgh, 2018), http://d-scholarship.pitt.edu/33891/.

68. Chris Holdgraf et al., “Portable Learning Environments for Hands-on Computational Instruction: Using Container- and Cloud-based Technology to Teach Data Science,” in *PEARC17: Proceedings of the Practice and Experience in Advanced Research Computing 2017 on Sustainability, Success and Impact* (New York: Association for Computing Machinery, 2017), https://doi.org/10.1145/3093338.3093370.

69. “2020 Strategic Technologies Glossary,” EDUCAUSE, accessed May 26, 2020, https://www.educause.edu/research-and-publications/research/top-10-it-issues-technologies-and-trends/technologies-survey-glossary.

70. “Dataspace,” NC State University Libraries, accessed May 26, 2020, https://www.lib.ncsu.edu/spaces/dataspace.

71. Burton et al., *Shifting to Data Savvy*.