2021

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

Sarah Lippincott

Mary Lee Kennedy

Clifford Lynch

Scout Calvert

Jocelyn Cozzo

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

Final Report

By Sarah Lippincott

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

With significant research for the glossary contributed by Jocelyn Cozzo

2021

ASSOCIATION OF RESEARCH LIBRARIES

born-digital RESEARCH + CONSULTING

cni Coalition for Networked Information

EDUCAUSE

Table of Contents

Chapter 1: Executive Summary

The generation, dissemination, and analysis of digital information is a significant driver, and consequence, of technological change. As data and information stewards in physical and virtual space, research libraries are thoroughly entangled in the challenges presented by the Fourth Industrial Revolution:[1] a societal shift powered not by steam or electricity, but by data, and characterized by a fusion of the physical and digital worlds.[2] Organizing, structuring, preserving, and providing access to growing volumes of the digital data generated and required by research and industry will become a critically important function. As partners with the community of researchers and scholars, research libraries are also recognizing and adapting to the consequences of technological change in the practices of scholarship and scholarly communication.

Technologies that have emerged or become ubiquitous within the last decade have accelerated information production and have catalyzed profound changes in the ways scholars, students, and the general public create and engage with information. The production of an unprecedented volume and diversity of digital artifacts, the proliferation of machine learning (ML) technologies,[3] and the emergence of data as the "world's most valuable resource,"[4] among other trends, present compelling opportunities for research libraries to contribute in new and significant ways to the research and learning enterprise. Librarians are all too familiar with predictions of the research library's demise in an era when researchers have so much information at their fingertips. A growing body of evidence provides a resounding counterpoint: that the skills, experience, and values of librarians, and the persistence of libraries as an institution, will become more important than ever as researchers contend with the data deluge and the ephemerality and fragility of much digital content.

This report identifies strategic opportunities for research libraries to adopt and engage with emerging technologies,[5] with a roughly five-year time horizon. It considers the ways in which research library values and professional expertise inform and shape this engagement, the ways library and library worker roles will be reconceptualized, and the implication of a range of technologies on how the library fulfills its mission. The report builds on a literature review covering the last five years of published scholarship, primarily North American information science literature, and interviews with a dozen library field experts, completed in fall 2019. It begins with a discussion of four cross-cutting opportunities that permeate many or all aspects of research library services. Next, specific opportunities are identified in each of five core research library service areas: facilitating information discovery, stewarding the scholarly and cultural record, advancing digital scholarship, furthering student learning and success, and creating learning and collaboration spaces. Each section identifies key technologies shaping user behaviors and library services, and highlights exemplary initiatives.

Underlying much of the discussion in this report is the idea that "digital transformation is increasingly about change management"[6] —that adoption of or engagement with emerging technologies must be part of a broader strategy for organizational change, for "moving emerging work from the periphery to the core,"[7] and a broader shift in conceptualizing the research library and its services. Above all, libraries are benefitting from the ways in which emerging technologies offer opportunities to center users and move from a centralized and often siloed service model to embedded, collaborative engagement with the research and learning enterprise.

Cross-Cutting Opportunities

Engage with machine learning to improve research, learning, and scholarly communication

Machine learning, the sub-discipline of artificial intelligence (AI)[8] that "uses collections of examples to train software to recognize patterns, and to act on that recognition,"[9] has demonstrated a remarkable ability to match (and outpace) human performance on certain well-constrained but complex tasks, and is already incorporated into a range of common systems and devices. The term AI has taken on a life of its own; it is frequently invoked as an umbrella term for ML, natural language processing (NLP), expert systems, and related technologies that approximate human cognition. The casual use of the term AI often erases the distinction between substantive applications (for example, pattern and image recognition) and speculative and unproven uses (for example, prediction, reasoning, formulating original ideas).[10] In the interests of specificity and precision, this report makes an effort to identify specific technologies (such as ML) where possible, while recognizing that some initiatives invoke AI even when the scope of their activities focuses on a specific sub-technology.

As the near-term applications of ML and related technologies shape the ways in which scholars create and engage with information, students learn and study, and communities interact with their built environments, libraries will be profoundly implicated, given their role as creators, sources, and stewards of information and as educators. Libraries can strategically engage with ML by better understanding its affordances, limitations, and risks, and by distinguishing the genuine accomplishments of ML and related technologies from AI hype. The application of ML to tasks related to classification, prediction, and pattern recognition and generation, make it particularly germane to information discovery. A number of research libraries have initiatives underway that apply ML, computer vision, natural language processing, and other techniques to automate description of large-scale digital collections[11] and enhance discovery, access, and analysis systems.[12] A few are also leading critical discourse and educational efforts on their campuses around the implications, ethics, and future of ML.[13] Research libraries also have opportunities for field-level collaboration. For example, libraries could assemble the large volume of validated and labelled training data that drive ML algorithms in ways that aim to recognize or mitigate bias and that are sensitive to the specific needs of cultural heritage materials.[14]

Bolster services that recognize the centrality of data to the research enterprise

Big or small, textual, numeric, or visual, in support of the humanities, science, or interdisciplinary research, digital data and structured knowledge have become essential and ubiquitous scholarly inputs and first-order outputs.[15] Research libraries play a key role in data generation, dissemination, discovery, analysis, and stewardship and can contribute to realizing the vision of a FAIR (findable, accessible, interoperable, and reusable) data environment that advances open scholarship.[16] Over the next decade, advancing FAIR data will require significant investment, creating myriad opportunities for libraries. Research libraries can contribute to FAIR data by describing structured data; building and providing access to machine-actionable and ML-ready data sets that facilitate computationally driven research; collaborating with domain experts to develop descriptive standards and ontologies that support disciplinary and multi-disciplinary research by humans and machines; and maintaining reuse-driven repository infrastructure.[17] Research libraries are developing services that are attuned to the needs of scholars working with very large data sets as well as the long tail of smaller, heterogeneous, unique, and often labor-intensive data sets that support research across the disciplinary spectrum. In their role as educators, librarians are also well-positioned to cultivate data fluency and the technology skills required for data-centric research methods.[18]

Integrate the library's services and collections with the networked environment

Researchers operate in geographically distributed, interdisciplinary, networked environments. Scholarly communication has also become diversified and disaggregated. The library's role in information management is being reenvisioned: no longer solely a steward of a unified local collection, the library becomes the facilitator of a networked suite of open and extensible tools, resources, and services. Building local research collections will eventually diminish in importance, while curation and facilitated access to information become critical.[19] Research libraries are leveraging emerging technologies to make their services and collections interoperable and more seamlessly integrated into the lives and work of their constituents. For example, research libraries are ensuring that their unique digital collections—including digitized special collections, institutionally published content, and expert profiles—are interoperable with web-scale and federated discovery tools, by creating harvestable, machine-readable metadata, and associating them with persistent identifiers. As research praxis routinely crosses institutional and geographic boundaries, research libraries also have opportunities to act consortially or outside of their local framework to maximize their impact. Research libraries could, for example, develop coordinated models of research data stewardship in which individual institutions assume responsibility for a segment of data (such as data defined by domain or type) based on local strengths and capacity.[20] Conversely, libraries could contribute their expertise to initiatives that are not affiliated with or hosted by their (or any) campus, such as specialized "data communities."[21]

Cultivate privacy awareness and privacy services

Emerging technologies are redefining expectations of privacy and creating tensions around the ethical use of personal data. The ease of constant surveillance is facilitated in physical space by Internet of Things (IoT) technologies that collect continuous streams of data, and in virtual space by the collection of digital analytics by campus and third-party systems. ML tools can process this data with remarkable speed and precision, making genuine data de-identification nearly impossible. As students and scholars come to expect (data-driven) personalized digital services and campuses expect to reap the benefits of large-scale data analytics, libraries will have critical choices to make. Research libraries can play a key role in helping their campus communities develop a nuanced understanding of privacy in physical and digital space. In their own work, libraries can commit to transparent data collection retention and use policies, and conscious, thoughtful management and control of personal information. This includes negotiating vendor agreements that protect reader privacy,[22] offering trade-offs between privacy and personalization where appropriate,[23] and establishing boundaries around their participation in campus-wide data collection efforts.[24]

A genuine commitment to privacy may become one of the library's fundamental distinguishing features;[25] many libraries are working to provide (physical and virtual) spaces that consciously minimize and make transparent the ways in which users may be tracked or their data collected. Libraries have an opportunity to position themselves as leaders in privacy education and privacy-aware approaches to personalization, learning analytics, and the use of tracking technologies on campus. A core component of user-centered library services will be positioning users at the center of discussions about the ethical use of user data and the implementation of tracking devices, algorithmic decision-making tools, and other potentially invasive technologies in libraries.

Facilitating Information Discovery and Use

Invest in user-centered discovery tools

The widespread adoption of web-scale discovery tools, combined with a landscape of information overabundance, has upended "the notion that the library attempts to licence or provide access to all [published] material" and instead has prompted libraries to focus on creating and licensing discovery tools and services that navigate and curate content.[26] Some of the most promising uses of emerging technologies to make search and discovery more user-centered include various kinds of enhanced search, NLP-based automated text-processing tools, recommendation systems, and personal assistants. While libraries may develop homegrown solutions, most of these tools will be commercial products, making them potentially problematic with regard to privacy. Aspirationally, these technologies expand discovery beyond simple search and retrieval, reconceptualizing it as a process of exploration and engagement with networked information.

Reveal hidden digital collections through enhanced description

The acceleration of digitization and born-digital content creation has left libraries facing an ever-increasing backlog of resource description to drive traditional collection discovery and navigation tools and methods. As libraries place increasing value on their unique local collections, they need new ways of making those collections discoverable and navigable to internal and external audiences, both human and machine. A number of academic libraries are experimenting with technologies such as ML algorithms (including facial recognition and image recognition/classification) and natural language processing to automate metadata creation, improve discoverability of visual information, and provide unprecedented access to their rich digitized and born-digital collections.

Expose library collections beyond library systems

As information becomes distributed, diversified, and open, researchers prefer web-scale discovery tools that aggregate resources from a range of sources over siloed library catalogs and digital asset management systems.[27] Research libraries have a number of strategic opportunities to integrate library collections with a range of other open, digital resources, enriching the information available to users on the open web. Research libraries are meeting users where they are by implementing search engine optimization techniques; exposing metadata for harvesting by aggregators, such as the Digital Public Library of America; providing application programming interfaces (APIs) that permit new forms of computational engagement with collections; adopting interoperability standards, such as the International Image Interoperability Framework (IIIF),[28] to facilitate discovery and reuse; and participating in linked open data (LOD) initiatives. The shift towards revealing local collections to external audiences rather than the reverse, a trend Lorcan Dempsey has called the “inside-out library”[29] and one component of what other authors have termed the “library as platform,”[30] is a natural consequence of an open, oversaturated, and networked information landscape.

Stewarding the Scholarly and Cultural Record

Advance open research and publishing practices

By supporting open research practices—including the adoption of open metadata standards, creation of machine-readable publications, and deposit of outputs (including underlying data and code) in open repositories—libraries make research more discoverable, reusable, reproducible, and durable. These practices improve both the quality of scholarship itself and the quality and manageability of the scholarly record. Libraries play a critical role in achieving FAIR (findable, accessible, interoperable, and reusable) research data through their curation, education, and preservation activities.[31] Realizing the vision of FAIR scholarship will be a central challenge for the research community over the next decade.

Reinforce integrity and trust in the scholarly and cultural record

Memory institutions are built on trust: the trust that materials under their stewardship are authentic, immutable, and preserved in perpetuity or de-acquired through a transparent and well-understood process. Emerging technologies pose new challenges for fulfilling the role of trusted steward. The assurance of authenticity, for example, is threatened by the ease of manipulating and altering digital media, and the complexities of determining provenance of digital materials. Deep fakes—counterfeit video, audio, still images, and textual content created using ML—pose a particular challenge. Research libraries have a range of digital forensics tools at their disposal to authenticate digital artifacts and collections at the time of accession and throughout their life cycle. They are also identifying secure pathways—possibly involving distributed ledger technologies (such as blockchain) and public key infrastructure (PKI)—to acquire copies of digital objects from sources they trust, documenting and proving the chain of custody, and any changes that have been made to it along the way.[32] After accessioning, fixity checking continually proves objects and collections do not change over time, due to degradation of the content, or to intentional or accidental manipulation. Underlying all of these processes is the need to maintain security and integrity of computing and storage operations in the face of cyberattacks[33] and natural disasters. Finally, librarians also help their constituents develop the skills needed to assess and critically engage with the integrity and reliability of information.

Preserve the evolving scholarly and cultural record

The expanded scholarly and cultural record has amplified both the technical and social barriers to achieving digital preservation at scale. On the technical front, emerging technologies have led to new types of research and creative outputs that require new approaches to digital preservation, as well as an unprecedented rate of digital content creation. Software, 3-D data, dynamic web content, and the inputs and outputs of ML, among other media, push the limits of established digital preservation practices. The digital cultural and historical record—the massive volumes of digital images and video, news, social media posts, and other web-based content that constitute essential evidence for present and future scholarship—will be incompletely preserved its scale and complexity.[34] Addressing the thorny questions of what can and should be preserved over the long term, will require deep cross-institutional coordination and cooperation.[35] On the social front, the distributed and licensed nature of digital scholarly and cultural content presents legal, administrative, and financial barriers. Even as emerging technologies have destabilized the digital preservation environment, they have also offered new solutions and opportunities. A few libraries—and their collaborators in computer science and information technology departments—are leveraging developments in containerization, distributed ledger technologies (such as blockchain), new storage media, and automation of digital preservation practices through ML to help ensure that the expanded scholarly record remains accessible well into the future.

Advancing Digital Scholarship

Develop data services that work for big data[36] and small data across disciplines

Academic and research libraries are natural partners with others involved 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.[37] 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.

Provide and sustain machine-actionable collections

Some of the most innovative digital scholarship work uses computational processes to derive new insights from vast troves of digital and digitized content held in library collections. Text and data mining have gained traction with many scholars in a range of disciplines as they seek more nuanced methods of discovery and analysis.[38] Machine-actionable collections enable researchers to go beyond simple information retrieval, treating collections (including their metadata, full-text, and relationships) as the input for powerful computational processes. Such initiatives 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."[39] This means not only making digital collections available online, but providing them as structured, machine-actionable data sets. Machine-actionable collections are essential not only for human-driven computational research, but for the development of new ML tools, which rely on large quantities of structured data to become proficient at a task. Libraries can apply their "expertise and practical experience in creating and managing classification systems" to develop ML training sets that serve the needs of cultural heritage institutions.[40]

Deliver data science education and consultation

Data science proficiency has rapidly become a core competency for researchers and students, as scholars in many or most disciplines routinely rely on computational data analysis in their research and learning.[41] Research libraries can cultivate the data science skill sets to sustain and expand these practices. Some research libraries have identified a niche in providing tailored educational offerings for faculty members and students outside of STEM fields, who may lack opportunities within their department or program of study.[42] These informal educational programs can help undergraduate and graduate students in all disciplines become proficient in common data science tools (such as electronic lab notebooks), techniques (such as web scraping), research data management practices, compliance with funder and federal policies, and open science principles.

Furthering Learning and Student Success

Build digital fluency and digital scholarship skill sets

Research libraries provide a range of informal education and consultation to impart the digital skills that contribute to the academic and professional success of undergraduates, graduate students, and early career researchers. These include workshops that teach concrete digital scholarship and coding skills, such as programming languages,[43] software carpentry,[44] and data visualization;[45] research data management and open science practices; and scholarly communications topics such as copyright, identity management, and navigating academic publishing. Longer-term cohort-based educational programs have also become popular. These programs often encourage interdisciplinary engagement with an emerging technology over the course of a semester or longer.[46] A few research libraries have also launched formal programs that fill gaps in the academic curriculum, for example, the Temple University Libraries' interdisciplinary cultural analytics certificate.[47] In addition to digital scholarship skills, research libraries have opportunities to help students critically engage with and optimize their use of a new generation of productivity tools, many powered by ML, that promise to assist users in a range of tasks related to learning and study.

The ease of publishing information and misinformation on the web, the growing sophistication of counterfeit content, and the use of black box algorithms to generate and display information mean that achieving digital fluency[48] also requires that students be able to interpret and evaluate an unprecedented array of new media formats and sources. Students need to understand not only the credibility and reliability of textual media, they need data and algorithmic literacy skills, strategies for distinguishing between genuine and manipulated or fabricated digital content, and an understanding of online data privacy. Libraries are well-positioned to deliver an expanded digital fluency curriculum in partnership with faculty members, campus IT, and other collaborators.

Foster critical engagement with and access to emerging technologies for all students

As third spaces, independent from any campus department, libraries have become hubs of technology access for students in all majors.

Technology-rich learning and information commons, collaboration studios, makerspaces, and labs are now commonplace in libraries. Locating digital scholarship centers within libraries can help to democratize and de-silo access to cutting edge technologies, encouraging cross-disciplinary collaboration and discovery.[49] These spaces provide access to specialized software and hardware for fabrication (such as 3-D printers, computer-aided design and drafting software); visualization (such as high-resolution displays); immersive reality (such as VR headsets); and other digital research and creation methods. When libraries apply their existing expertise as educators to new forms of knowledge production, they can help their communities thoughtfully and productively engage with technology in these spaces. Librarians are equally well-positioned to collaborate with faculty on the pedagogically grounded integration of technologies such as immersive reality and information visualization in the classroom.

Creating and Managing Learning and Collaboration Spaces

Create dynamic, networked spaces that promote new forms of inquiry

While leading-edge technology is often most conspicuous in makerspaces and labs, some of its most transformative potential lies in the seamless and often invisible integration of emerging technologies into the full library-visitor experience. The use of Internet of Things technologies presents a particularly compelling opportunity for library spaces (whether in the library building or embedded in shared spaces around campus) and services to dynamically adapt to user behaviors. The ubiquitous integration of sensors and networked technologies into the library's physical spaces could transform it into "a living-learning lab that senses and studies human dynamics, human-computer interactions, and human-building interactions."[50] Libraries have an opportunity to pioneer inclusive, privacy-aware approaches to this integration of sensing technologies in the public sphere. Creating networked library spaces complements the library's role as data provider and steward, as a node for digital information discovery, and as a promoter of critical engagement with emerging technologies and the changing nature of research and information behavior.

Enhance the user experience in library spaces

Emerging technologies offer a range of opportunities for libraries to make spaces more welcoming, navigable, interactive, comfortable, and productive. Libraries are experimenting with the Internet of Things (IoT), particularly beacon technology, to create self-guided library tours and navigational aids,[51] build augmented reality (AR) exhibits,[52] provide location-specific mobile alerts,[53] help users locate materials in the library stacks,[54] and facilitate access to bookable or restricted spaces or items.[55] Emerging technologies can also be deployed to enhance a sense of community within library spaces. Several speculative apps propose to help users locate and connect with others in a library space who share their interests, allowing them to form study or collaboration groups on the fly.[56] As they engage with beacons, wearables, and location-based apps, libraries are cognizant of implications around privacy and intellectual freedom, and are developing best practices for privacy-aware implementation of IoT technologies in library spaces.[57]

Conclusion

Research libraries can bring values-based decision-making to bear as they find the right balance in their approach to adopting and experimenting with emerging technologies—the balance between agility and sustainability, convenience and privacy, transformation and persistence. As emerging technologies such as machine learning, immersive reality, and the Internet of Things change the ways researchers and students engage with information, libraries have opportunities to advance their contributions to the research and learning enterprise. As adopters of these technologies, research libraries can make information more discoverable, reusable, and durable. As educators, library workers can help their communities critically and productively engage with technology in the service of research and learning.

Endnotes

1. Klaus Schwab, “The Fourth Industrial Revolution: What It Means, How to Respond,” World Economic Forum, January 14, 2016, https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond/.

2. Donna Ellen Frederick, “Libraries, Data and the Fourth Industrial Revolution,” Data Deluge Column, *Library Hi Tech News* 33, no. 5 (July 4, 2016): 9–12, https://doi.org/10.1108/LHTN-05-2016-0025.

3. “ML is a subset of the larger field of artificial intelligence (AI) that ‘focuses on teaching computers how to learn without the need to be programmed for specific tasks,’ note Sujit Pal and Antonio Gulli in *Deep Learning with Keras*. ‘In fact, the key idea behind ML is that it is possible to create algorithms that learn from and make predictions on data.’ “—James Furbush, “Machine Learning: A Quick and Simple Definition, O’Reilly, May 3, 2018, https://www.oreilly.com/ideas/machine-learning-a-quick-and-simple-definition.

4. “The World’s Most Valuable Resource Is No Longer Oil, but Data,” Leaders, *The Economist*, May 6, 2017, https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data.

5. The definition of emerging technologies developed by Rotolo, Hicks, and Martin “identifies five attributes that feature in the emergence of novel technologies. These are: (i) radical novelty, (ii) relatively fast growth, (iii) coherence, (iv) prominent impact, and (v) uncertainty and ambiguity.”—Daniele Rotolo, Diana Hicks, and Ben R. Martin, “What Is an Emerging Technology?,” preprint, submitted February 13, 2015, last revised January 4, 2016, https://arxiv.org/abs/1503.00673.

6. "'None of This Is Really about Technology' — Digital Transformation and Culture Change," *Jisc* (blog), January 21, 2020, https://www.jisc.ac.uk/news/none-of-this-is-really-about-technology-digital-transformation-and-culture-change-21-jan-2020.

7. Thomas Padilla, *Responsible Operations: DataScience, Machine Learning, and AI in Libraries* (Dublin, OH: OCLC, 2019), https://doi.org/10.25333/xk7z-9g97.

8. For the purposes of this paper we use the following definition of AI from the Association for the Advancement of Artificial Intelligence (AAAI): "the scientific understanding of the mechanisms underlying thought and intelligent behavior and their embodiment in machines."—AAAI, "Information about AI from the News, Publications, and Conferences," *AITopics*, accessed February 19, 2020, https://aitopics.org/search.

9. 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/.

10. Ian Bogost, " 'Artificial Intelligence' Has Become Meaningless," *The Atlantic*, March 4, 2017, https://www.theatlantic.com/technology/archive/2017/03/what-is-artificial-intelligence/518547/.

11. Matthew Short, "Text Mining and Subject Analysis for Fiction; or, Using Machine Learning and Information Extraction to Assign Subject Headings to Dime Novels," *Cataloging & Classification Quarterly* 57, no. 5 (2019): 315–36, https://doi.org/10.1080/01639374.2019.1653413; Rachael Goh, "Using Named Entity Recognition for Automatic Indexing" (paper presented at IFLA WLIC 2018: "Transform Libraries, Transform Societies," Kuala Lumpur, Malaysia, 2018), http://library.ifla.org/id/eprint/2214; Martijn

Kleppe et al., *Exploration Possibilities Automated Generation of Metadata* (The Hague: National Library of the Netherlands, August 23, 2019), https://doi.org/10.5281/zenodo.3375192.

12. Nicolas Fiorini et al., “PubMed Labs: An Experimental System for Improving Biomedical Literature Search,” *Database: The Journal of Biological Databases and Curation* 2018 (September 18, 2018), https://doi.org/10.1093/database/bay094; Victoria L. Rubin, Yimin Chen, and Lynne Marie Thorimbert, “Artificially Intelligent Conversational Agents in Libraries,” *Library Hi Tech* 28, no. 4 (2010): 496–522, https://doi.org/10.1108/07378831011096196.

13. Lindsay McKenzie, “A New Home for AI: The Library,” *Inside Higher Ed*, January 17, 2018, https://www.insidehighered.com/news/2018/01/17/rhode-island-hopes-putting-artificial-intelligence-lab-library-will-expand-ais-reach.

14. Padilla, *Responsible Operations*.

15. 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.

16. Barend Mons, “FAIR Science for Social Machines: Let’s Share Metadata Knowlets in the Internet of FAIR Data and Services,” *Data Intelligence* 1, no. 1 (Winter 2019): 22–42, https://doi.org/10.1162/dint_a_00002.

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. 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/.

19. Lorcan Dempsey, “Library Collections in the Life of the User: Two Directions,” *LIBER Quarterly* 26, no. 4 (October 11, 2016): 338–59, https://doi.org/10.18352/lq.10170.

20. Carole Palmer, interview by author, October 30, 2019.

21. 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.

22. Clifford A. Lynch, “Reader Privacy: The New Shape of the Threat,” *Research Library Issues*, no. 297 (2019): 7–14, https://doi.org/10.29242/rli.297.2.

23. Marshall Breeding, “Strengthening Patron Engagement while Protecting Privacy,” *Computers in Libraries* 38, no. 8 (October 2018): 18–20.

24. Kyle M.L. Jones and Dorothea Salo, “Learning Analytics and the Academic Library: Professional Ethics Commitments at a Crossroads,” *College & Research Libraries* 79, no. 3 (April 2018): 304–23, https://doi.org/10.5860/crl.79.3.304.

25. Tony Ageh and Brent Reidy, interview by author, October 30, 2019.

26. Andrew M. Cox, Stephen Pinfield, and Sophie Rutter, “The Intelligent Library: Thought Leaders’ Views on the Likely Impact of Artificial Intelligence on Academic Libraries,” *Library Hi Tech* 37, no.3 (2019): 418–35, https://doi.org/10.1108/LHT-08-2018-0105.

27. David Attis and Colin Koproske, “Thirty Trends Shaping the Future of Academic Libraries,” *Learned Publishing* 26, no. 1 (January 2013): 18–23, https://doi.org/10.1087/20130104; John Akeroyd, “Discovery Systems: Are They Now the Library?,” *Learned Publishing* 30, no. 1 (January 2017): 87–89, https://doi.org/10.1002/leap.1085.

28. Stuart Snydman, Robert Sanderson, and Tom Cramer, “The International Image Interoperability Framework (IIIF): A

Community & Technology Approach for Web-Based Images," in *Archiving Conference*, vol. 2015 (Springfield, VA: Society for Imaging Science and Technology, 2015), 16–21.

29. 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.

30. Roger C. Schonfeld, "Does Discovery Still Happen in the Library? Roles and Strategies for a Shifting Reality," *Ithaka S+R* (blog), September 24, 2014, https://sr.ithaka.org/blog/does-discovery-still-happen-in-the-library-roles-and-strategies-for-a-shifting-reality/; David Weinberger, "Library as Platform," *Library Journal*, September 4, 2012, https://www.libraryjournal.com?detailStory=by-david-weinberger.

31. Mark D. Wilkinson et al., "The FAIR Guiding Principles for Scientific Data Management and Stewardship," *Scientific Data* 3 (2016), https://doi.org/10.1038/sdata.2016.18.

32. Mark Bell et al., "Underscoring Archival Authenticity with Blockchain Technology," *Insights* 32, no. 1 (2019): 21, https://doi.org/10.1629/uksg.470.

33. Glenn D. Tiffert, "Peering Down the Memory Hole: Censorship, Digitization, and the Fragility of Our Knowledge Base," *American Historical Review* 124, no. 2 (April 2019): 550–68, https://doi.org/10.1093/ahr/rhz286.

34. Oya Y. Rieger, *The State of Digital Preservation in 2018: A Snapshot of Challenges and Gaps*, Issue Brief (New York: Ithaka S+R, October 29, 2018), https://doi.org/10.18665/sr.310626.

35. Carol A. Mandel, *Can We Do More? An Examination of Potential Roles, Contributors, Incentives, and Frameworks to Sustain Large-Scale Digital Preservation* (Arlington, VA: Council on Library and Information Resources, September 2019), https://clir-dlf.app.box.com/s/31tc6nrua3cj8jjwoymee78gl3plzlo2.

36. 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.

37. 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.

38. Christine L. Borgman, "Whose Text, Whose Mining, and to Whose Benefit?," accepted for publication in *Quantitative Social Sciences*, December 3, 2019, https://escholarship.org/uc/item/3682b9j6.

39. Thomas Padilla et al., *Final Report — Always Already Computational: Collections as Data*, May 22, 2019, https://doi.org/10.5281/zenodo.3152935.

40. Michael Ridley, "Training Datasets, Classification, and the LIS Field," *Library AI* (blog), September 26, 2019, https://libraryai.blog.ryerson.ca/2019/09/26/training-datasets-classification-and-the-lis-field/.

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

42. 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.

43. 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.

44. Thea P. Atwood et al., "Joining Together to Build More: The New England Software Carpentry Library Consortium," *Journal of eScience Librarianship* 8, no. 1 (2019): 5, https://doi.org/10.7191/jeslib.2019.1161; see also "Foundations for Research Computing: A University-wide Initiative," Columbia University, accessed February 19, 2020, https://rcfoundations.research.columbia.edu.

45. Lisa M. Federer and Douglas J. Joubert, "Providing Library Support for Interactive Scientific and Biomedical Visualizations with Tableau," *Journal of eScience Librarianship* 7, no. 1 (2018): e1120, https://doi.org/10.7191/jeslib.2018.1120.

46. 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/; also see, for example, "The 99 AI Challenge," University of Toronto Libraries, accessed February 19, 2020, https://onesearch.library.utoronto.ca/ai-challenge.

47. Morgan Zalot, "Temple Libraries Launches Interdisciplinary Cultural Analytics Certificate," *Temple Now*, Temple University, June 3, 2019, https://news.temple.edu/news/2019-06-03/temple-libraries-launches-interdisciplinary-cultural-analytics-certificate.

48. Jennifer Sparrow, "Digital Fluency: Preparing Students to Create Big, Bold Problems," *EDUCAUSE Review*, March 12, 2018, https://er.educause.edu/articles/2018/3/digital-fluency-preparing-students-to-create-big-bold-problems.

49. Joan K. Lippincott and Diane Goldenberg-Hart, *Digital Scholarship Centers: Trends & Good Practice* (Washington, DC: Coalition for Networked Information, 2014), https://www.cni.org/wp-content/uploads/2014/11/CNI-Digitial-Schol.-Centers-report-2014.web_.pdf.

50. Yi Shen, “Intelligent Infrastructure, Ubiquitous Mobility, and Smart Libraries – Innovate for the Future,” *Data Science Journal* 18, no. 1 (March 21, 2019): 11, https://doi.org/10.5334/dsj-2019-011.

51. Jonathan Bradley et al., “Creation of a Library Tour Application for Mobile Equipment Using IBeacon Technology,” April 25, 2016, https://vtechworks.lib.vt.edu/handle/10919/71832.

52. Brandon Patterson, “Talking Portraits in the Library: Building Interactive Exhibits with an Augmented Reality App,” *Code4Lib Journal*, no. 46 (November 5, 2019), https://journal.code4lib.org/articles/14838.

53. Somaly Kim Wu, Marc Bess, and Bob R. Price, “Digitizing Library Outreach: Leveraging Bluetooth Beacons and Mobile Applications to Expand Library Outreach,” *Digitizing the Modern Library and the Transition From Print to Electronic* (IGI Global, 2018), 193–203, https://doi.org/10.4018/978-1-5225-2119-8.ch008; Sidney Eng, “Connection, Not Collection: Using IBeacons to Engage Library Users,” *Information Today*, December 2015, http://www.infotoday.com/cilmag/dec15/Eng--Using-iBeacons-to-Engage-Library-Users.shtml.

54. Valeda Dent et al., “Wayfinding Serendipity: The BKFNDr Mobile App,” *Code4Lib Journal*, no. 42 (November 8, 2018), https://journal.code4lib.org/articles/13811.

55. Hubert C. Y. Chan and Linus Chan, “Smart Library and Smart Campus,” *Journal of Service Science and Management* 11, no. 6 (November 28, 2018): 543–64, https://doi.org/10.4236/jssm.2018.116037.

56. Ian Glover and Kieran McDonald, “Digital Places: Location-Based Digital Practices in Higher Education Using Bluetooth Beacons,” in *EdMedia+ Innovate Learning* (Association for the Advancement of Computing in Education (AACE), 2018), 950–959.

57. Jim Hahn, “Security and Privacy for Location Services and the Internet of Things,” *Library Technology Reports*, 53, no. 1 (2017): 23–28, https://www.journals.ala.org/index.php/ltr/article/view/6178.

Chapter 2: Introduction, Methodology, and Cross-Cutting Opportunities

Introduction

The generation, dissemination, and analysis of digital information is a significant driver, and consequence, of technological change. As data and information stewards in physical and virtual space, research libraries are thoroughly entangled in the challenges presented by the Fourth Industrial Revolution:[1] a societal shift powered not by steam or electricity, but by data, and characterized by a fusion of the physical and digital worlds.[2] Organizing, structuring, preserving, and providing access to growing volumes of the digital data generated and required by research and industry will become a critically important function. As partners with the community of researchers and scholars, research libraries are also recognizing and adapting to the consequences of technological change in the practices of scholarship and scholarly communication.

Technologies that have emerged or become ubiquitous within the last decade have accelerated information production and have catalyzed profound changes in the ways scholars, students, and the general public create and engage with information. The production of an unprecedented volume and diversity of digital artifacts, the proliferation of machine learning (ML) technologies,[3] and the emergence of data as the “world’s most valuable resource,”[4] among other trends, present compelling opportunities for research libraries to contribute in new and significant ways to the research and learning enterprise. Librarians are all too familiar with predictions of the research library’s demise in an era when researchers have so much information at their fingertips. A growing body of evidence provides a resounding counterpoint: that the skills, experience, and values of librarians, and the persistence of libraries as institutions, will become more important than ever as researchers contend with the data deluge and the ephemerality and fragility of much digital content.

This report identifies strategic opportunities for research libraries to adopt and engage with emerging technologies,[5] with a roughly five-year time horizon. It considers the ways in which research library values and professional expertise inform and shape this engagement, the ways library and library worker roles will be reconceptualized, and the implication of a range of technologies on how the library fulfills its mission. The report builds on a literature review covering the last five years of published scholarship—primarily North American information science literature—and interviews with a dozen library field experts, completed in fall 2019. It begins with a discussion of four cross-cutting opportunities that permeate many or all aspects of research library services. Next, specific opportunities are identified in each of five core research library service areas: facilitating information discovery and use, stewarding the scholarly and cultural record, advancing digital scholarship, furthering learning and student success, and building and managing learning and collaboration spaces. Each section identifies key technologies shaping user behaviors and library services, and highlights exemplary initiatives.

Underlying much of the discussion in this report is the idea that "digital transformation is increasingly about change management"[6]—that adoption of or engagement with emerging technologies must be part of a broader strategy for organizational change, for "moving emerging work from the periphery to the core,"[7] and a broader shift in conceptualizing the research library and its services. Above all, libraries are benefitting from the ways in which emerging technologies offer opportunities to center users and move from a centralized and often siloed service model to embedded, collaborative engagement with the research and learning enterprise.

Methodology

The research for this report included a literature review and semi-structured interviews with experts in the library field. The author performed a review of library literature, focusing on publications appearing within the past five years. The literature review included publications that summarized and speculated on current and future technology trends in general, as well as case studies and theoretical treatments of a range of specific technologies and their adoption in the cultural heritage sector. The author conducted semi-structured interviews with a dozen library community experts, including library deans and directors and information science faculty members, in fall 2019. The author asked the interviewees to reflect on the potential impacts of emerging technologies, the most compelling examples of library adoption, pitfalls and challenges of adopting new technologies, and the future of library services in the information age.

This report is structured around five key library roles: facilitating information discovery and use, stewarding the scholarly and cultural record, advancing digital scholarship, furthering learning and student success, and building and managing learning and collaboration spaces. The report addresses both the implications of emerging technologies on the changing needs and behaviors of library constituents, and the adoption of emerging technologies within academic and research libraries. Therefore, each section begins with a brief landscape overview that discusses a number of relevant societal and technological shifts and their implications for aspects of the library mission. Next, each section identifies strategic opportunities for libraries to engage with and adopt emerging technologies to enhance and develop services, form new partnerships, and continue to support the research and learning mission of their institutions. The discussion of each strategic opportunity includes concrete, current examples from academic and research libraries. Each section concludes with a summary of key takeaways. Readers will find that some sections of this report have less extensive coverage. Uneven coverage generally reflects the fact that library engagement with emerging technologies in each of these areas is also uneven, at least as measured in the published literature.

A glossary at the end of the report defines selected terms that may be unfamiliar to the reader, that have ambiguous usage in common speech, or that have a specific meaning within the context of this report.

Cross-Cutting Opportunities

A number of opportunities emerged from the literature and expert interviews that transcend any one area of research library services. These cross-cutting opportunities relate to the technologies that have already seen the most widespread or productive engagement and adoption within research libraries, the societal trends that are shaping research and learning activities most profoundly, and the ways in which both technological and societal shifts intersect with the research library's identity and mission.

Engage with Machine Learning to Improve Research, Learning, and Scholarly Communication

Machine learning, the sub-discipline of artificial intelligence (AI)[8] that "uses collections of examples to train software to recognize patterns, and to act on that recognition,"[9] has demonstrated a remarkable ability to match (and outpace) human performance on certain well-constrained but complex tasks, and is already incorporated into a range of common systems and devices. The term AI has taken on a life of its own; it is frequently invoked as an umbrella term for ML, natural language processing (NLP), expert systems, and related technologies that approximate human cognition. The casual use of the term AI often erases the distinction between substantive applications (for example, pattern and image recognition) and speculative and unproven uses (for example, prediction, reasoning, formulating original ideas).[10] In the interests of specificity and precision, this report makes an effort to identify specific technologies (such as ML) where possible, while recognizing that some initiatives invoke AI even when the scope of their activities focuses on a specific sub-technology.

As the near-term applications of ML and related technologies shape the ways in which scholars create and engage with information, students learn and study, and communities interact with their built environments, research libraries will be profoundly implicated, given their role as creators, sources, and stewards of information and as educators. Many of the experts interviewed for this report identified ML as the most significant emerging technology for research libraries given its implications for the entire research and learning enterprise. This view is consistent with others in the field,[11] and reflected in a flurry of activity in cultural heritage and scholarly communications applications of ML. As ML approaches the peak of inflated expectations,[12] library experiments have proliferated. These tend to be one-off or first-of-a-kind projects that leverage ML in service of digital scholarship (for example, machine-generated metadata, natural language processing of large text corpora), with varying degrees of success. With a few notable exceptions, libraries are not yet systematically engaging with ML in ways that recognize its transformative potential across the full range of academic and research library services and activities. None of the 25 research-intensive libraries surveyed for a 2018 study mentioned ML or AI in their strategic plans.[13]

To move from ad hoc to strategic engagement with ML, libraries can cultivate a nuanced understanding of its affordances, limitations, and risks, and differentiate the genuine accomplishments of ML and related technologies from AI hype. Princeton University computer science professor Arvind Narayanan provides a simple litmus test to distinguish genuinely useful applications of AI and ML from problematic and unproven uses.[14] AI has shown demonstrable success for perception-related tasks (for example, facial recognition, medical diagnosis from images). It is making progress on tasks related to automating judgment (for example, spam detection, grading essays). However, Narayanan describes the premise that AI can be used for predictive analytics, especially predicting social outcomes (such as predicting criminal recidivism or job success), as “fundamentally dubious.” Further, AI tools remain easy to fool and manipulate. They can be easily co-opted by bad actors for purposes never envisioned by their creators;[15] they can be gamed and manipulated for commercial or political gain. ML’s reliance on human judgment and human-assembled training data make it particularly susceptible to problems of bias.[16]

The potential applications of AI and ML to research library workflows are myriad, from describing resources to providing reference services. Strategic investment in ML, informed by the ways emerging technologies have transformed user needs, can help libraries streamline longstanding processes. Perhaps more importantly, it can reinvent the ways in which they carry out their missions. For instance, ML's facility with tasks related to classification and pattern recognition and generation make it particularly germane to information discovery. A number of research libraries have initiatives underway that apply ML, computer vision, natural language processing, and other techniques to automate description of large-scale digital collections[17] and enhance discovery, access, and analysis systems.[18]

Principles of human-centered ML encourage librarians to "design an intelligent information system that respects the sources, engages critical inquiry, fosters imagination, and supports *human* learning and knowledge creation."[19] Human-centered AI does not replace human agency, human creativity, or human judgment. Rather it augments capacity, opens up new avenues of discovery, and enhances human potential by balancing high levels of automation with high levels of human control.[20] Libraries' longstanding interest in human-computer interaction for information retrieval and discovery, and the recent emphasis on user experience design in libraries, provide groundwork for research library involvement in the human-centered ML tools scholars need to create and engage with digital content. The entities such as labels, tags, and metadata generated by ML require infrastructure for preservation, and new approaches to metadata display that thoughtfully and ethically unite machine- and human-generated information.[21]

Finally, several libraries are leading critical discourse and educational efforts on their campuses around the implications, ethics, and future of ML.[22] Research libraries also have opportunities for field-level collaboration. For example, libraries could assemble the large volume of validated and labeled training data that drive ML algorithms in ways that aim to recognize or mitigate bias and that are sensitive to the specific needs of cultural heritage materials.[23] The US national AI strategy includes several points of engagement for libraries, including: understanding and addressing the ethical, legal, and societal implications of AI; developing shared public data sets and environments for AI training and testing; and measuring and evaluating AI technologies through standards and benchmarks.[24]

Bolster Services That Recognize the Centrality of Data to the Research Enterprise

Big or small, textual, numeric, or visual, in support of the humanities, science, or interdisciplinary research, digital data and structured knowledge have become essential and ubiquitous scholarly inputs and first-order outputs.[25] Research libraries play a key role in data generation, dissemination, discovery, analysis, and stewardship and can contribute to realizing the vision of a FAIR (findable, accessible, interoperable, and reusable) data environment that advances open scholarship.[26] Over the next decade, cultivating a FAIR data ecosystem will require significant investment, creating myriad opportunities for libraries. Research libraries can contribute to FAIR data by describing structured data; building and providing access to machine-actionable and ML-ready data sets that facilitate computationally driven research; collaborating with domain experts to develop descriptive standards and ontologies that support disciplinary and multi-disciplinary research by humans and machines; and maintaining reuse-driven repository infrastructure.[27] Research libraries are developing services that are attuned to the needs of scholars working with very large data sets as well as the long tail of smaller, heterogeneous, unique, and often labor-intensive data sets that support research across the disciplinary spectrum. In their role as educators, librarians are also well positioned to cultivate data fluency and the technology skills required for data-driven research methods.[28]

The rise of data as both “scholarly output”[29] and input has expanded research library roles in facilitating access to data collections as source material, and providing solutions for long-term data stewardship.

Libraries recognize that “data is the currency of science,” and that “[t] o be able to exchange data, communicate it, mine it, reuse it and review it is essential to scientific productivity, collaboration and to discovery itself.“[30] Research libraries have responded by licensing data sets for research, providing curated access to publicly available data, offering guidance on intellectual property laws relevant to the use and reuse of data, and providing the infrastructure for use-and-reuse-driven data repositories. Libraries recognize that data stewardship increasingly requires access to the code and computing environments used to produce or analyze data, and are developing solutions to ensure that data is saved with this critical context.

Research libraries are also applying FAIR data principles to one of their most valuable troves of digital information: library digital collections. Making library collections machine-actionable enables new forms of inquiry and gives new life to one of the library’s foundational services: collection stewardship. Some of the most innovative digital scholarship work uses computational processes to derive new insights from vast troves of digital and digitized content held in library collections. Machine-actionable collections enable researchers to go beyond simple information retrieval, treating collections (including their metadata, full-text, and relationships) as the input for powerful computational processes. Initiatives such as the Collections as Data project encourage cultural heritage institutions to thoughtfully develop digital collections (licensed, purchased, and unique) that support “computationally-driven research and teaching.”[31]

The clear and urgent need for data services has led many libraries to hire dedicated data librarians and build data services portfolios and data repositories. Still, a 2017 survey found that around a quarter of R1 universities (doctoral universities with the highest level of research activity) had no dedicated data librarians on staff and that the average number of data librarians at R1 institutions was slightly over two.[32] The next several years may see libraries redefining roles and adding new positions in these areas to meet demand for data services, growing capacity for creating and sustaining machine-actionable collections, and contending with large volumes of data that the library collects and manages.

Integrate the Library's Services and Collections with the Networked Environment

Researchers operate in geographically distributed, interdisciplinary, networked environments. Scholarly communication has also become diversified and disaggregated. The idea that research library services and infrastructure will also become increasingly outwardly focused, interoperable, and collaborative permeated the literature and the discourse of experts interviewed for this report. The formulations of library as platform, inside-out-library, and interoperable library, all allude to this central concept.

Research libraries are leveraging emerging technologies to make their services and collections interoperable and more seamlessly integrated into the lives and work of their constituents. For example, research libraries are ensuring that their unique digital collections—including digitized special collections, institutionally published content, and expert profiles—are interoperable with web-scale and federated discovery tools, by creating harvestable, machine-readable metadata, and associating them with persistent identifiers. The research library's role in information management is being reenvisioned: no longer solely a steward of a unified local collection, the library becomes the facilitator of a networked suite of open and extensible tools, resources, and services. Building local research collections will eventually diminish in importance, while curation and facilitated access to information become critical.[33] As research praxis routinely crosses institutional and geographic boundaries, research libraries also have opportunities to act consortially or outside of their local framework to maximize their impact. Research libraries could, for example, develop coordinated models of research data stewardship in which individual institutions assume responsibility for a segment of data (such as data defined by domain or type) based on local strengths and capacity.[34] Conversely, libraries could contribute their expertise to initiatives that are not affiliated with or hosted by their (or any) campus, such as specialized "data communities."[35]

Cultivate Privacy Awareness and Privacy Services

Emerging technologies are redefining expectations of privacy and creating tensions around the ethical use of personal data. The ease of constant surveillance is facilitated in physical space by Internet of Things (IoT) technologies that collect continuous streams of data, and in virtual space by the collection of digital analytics by campus and third-party systems. ML tools can process this data with remarkable speed and precision, making genuine data de-identification nearly impossible. As students and scholars come to expect (data-driven) personalized digital services and as campuses expect to reap the benefits of large-scale data analytics, libraries will have critical choices to make. Research libraries can play a key role in helping their communities develop a nuanced understanding of privacy in physical and digital space. In their own work, libraries can commit to transparent policies on data collection, retention, and use, as well as conscious, thoughtful management and control of personal information. This includes negotiating vendor agreements that protect reader privacy,[36] offering trade-offs between privacy and personalization where appropriate,[37] and establishing boundaries around library participation in campus-wide data-collection efforts.[38]

A genuine commitment to privacy may become one of the research library's fundamental distinguishing features;[39] many libraries are working to provide (physical and virtual) spaces that consciously minimize and make transparent the ways in which users may be tracked or their data collected. Libraries have an opportunity to position themselves as leaders in privacy education and privacy-aware approaches to personalization, learning analytics, and the use of tracking technologies on campus. A core component of user-centered library services will be positioning users at the center of discussions about the ethical use of user data and the implementation of tracking devices, algorithmic decision-making tools, and other potentially invasive technologies in libraries. At least two libraries—the New York Public Library and the University of Colorado Boulder—have formalized their commitment to privacy by creating a dedicated privacy officer position.[40]

Libraries are also scrutinizing their existing practices to ensure they align with commitments to protecting user data. Libraries' active and passive collection of user data—which may be identifiable, sensitive, and valuable—as well as their role as stewards of trustworthy information, profoundly implicates them in privacy and cybersecurity issues. Despite libraries' best intentions, they may be collecting and retaining data in ways that present risks to users or allow data collection by third-party platforms, which can expose user data to disclosure "by legal means, by hacking, or by human error."[41] Libraries can work internally and with their campus partners to determine their level of tolerance for data collection by external vendors, and negotiate licenses in ways that mitigate these risks.

On campus, libraries have an opportunity to position themselves as leaders in data-governance initiatives (which often have implications for student privacy), and collaborators in campus-wide privacy education and privacy-aware approaches to personalization, learning analytics, and the use of tracking technologies on campus.

Endnotes

1. Klaus Schwab, "The Fourth Industrial Revolution: What It Means, How to Respond," World Economic Forum, January 14, 2016, https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond/.

2. Donna Ellen Frederick, "Libraries, Data and the Fourth Industrial Revolution," Data Deluge Column, *Library Hi Tech News* 33, no. 5 (July 4, 2016): 9–12, https://doi.org/10.1108/LHTN-05-2016-0025.

3. "ML is a subset of the larger field of artificial intelligence (AI) that 'focuses on teaching computers how to learn without the need to

be programmed for specific tasks,' note Sujit Pal and Antonio Gulli in *Deep Learning with Keras*. 'In fact, the key idea behind ML is that it is possible to create algorithms that learn from and make predictions on data.'"—James Furbush, "Machine Learning: A Quick and Simple Definition," O'Reilly, May 3, 2018, https://www.oreilly.com/ideas/machine-learning-a-quick-and-simple-definition.

4. "The World's Most Valuable Resource Is No Longer Oil, but Data," Leaders, *The Economist*, May 6, 2017, https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data.

5. The definition of emerging technologies developed by Rotolo, Hicks, and Martin "identifies five attributes that feature in the emergence of novel technologies. These are: (i) radical novelty, (ii) relatively fast growth, (iii) coherence, (iv) prominent impact, and (v) uncertainty and ambiguity."—Daniele Rotolo, Diana Hicks, and Ben R. Martin, "What Is an Emerging Technology?," preprint, submitted February 13, 2015, last revised January 4, 2016, https://arxiv.org/abs/1503.00673.

6. " 'None of This Is Really about Technology' — Digital Transformation and Culture Change," Jisc, January 21, 2020, https://www.jisc.ac.uk/news/none-of-this-is-really-about-technology-digital-transformation-and-culture-change-21-jan-2020.

7. Thomas Padilla, *Responsible Operations: DataScience, Machine Learning, and AI in Libraries* (Dublin, OH: OCLC Research, 2019), https://doi.org/10.25333/xk7z-9g97.

8. For the purposes of this paper we use the following definition of AI from the Association for the Advancement of Artificial Intelligence (AAAI): "the scientific understanding of the mechanisms underlying thought and intelligent behavior and their embodiment in machines."—AAAI, "Information about AI from the News, Publications, and Conferences," *AITopics*, accessed February 19, 2020, https://aitopics.org/search.

9. 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/.

10. Ian Bogost, “ ‘Artificial Intelligence’ Has Become Meaningless,” *The Atlantic*, March 4, 2017, https://www.theatlantic.com/technology/archive/2017/03/what-is-artificial-intelligence/518547/.

11. Mary Lee Kennedy, *Research Libraries as Catalytic Leaders in a Society in Constant Flux: A Report on the ARL-CNI Fall Forum 2019*, with an introduction by Lorraine J. Haricombe (Washington, DC: Association of Research Libraries and Coalition for Networked Information, January 2020), https://doi.org/10.29242/report.fallforum2019.

12. “Gartner Hype Cycle,” Gartner, accessed March 31, 2020, https://www.gartner.com/en/research/methodologies/gartner-hype-cycle.

13. Amanda Wheatley and Sandy Hervieux, “Need Research? Ask Siri: Evaluating the Impact of Virtual Assistant AI on the Future of the Research Process” (slides presented at NFAIS 2019 AI Conference, Alexandria, VA, May 16, 2019), https://nfais.memberclicks.net/assets/AI2019/Amanda%20and%20Sandy.pdf.

14. Arvind Narayanan, “How to Recognize AI Snake Oil,” Princeton University, accessed March 31, 2020, https://www.cs.princeton.edu/~arvindn/talks/MIT-STS-AI-snakeoil.pdf.

15. Douglas Heaven, “Why Deep-Learning AIs Are So Easy to Fool,” *Nature* 574 (October 9, 2019): 163–66, https://doi.org/10.1038/d41586-019-03013-5.

16. Eileen Jakeway et al., *Machine Learning + Libraries Summit Event Summary* (Washington, DC: Library of Congress, LC Labs Digital Strategy Directorate, February 13, 2020), https://labs.loc.gov/static/labs/meta/ML-Event-Summary-Final-2020-02-13.pdf.

17. Matthew Short, "Text Mining and Subject Analysis for Fiction; or, Using Machine Learning and Information Extraction to Assign Subject Headings to Dime Novels," *Cataloging & Classification Quarterly* 57, no. 5 (2019): 315–36, https://doi.org/10.1080/01639374.2019.1653413; Rachael Goh, "Using Named Entity Recognition for Automatic Indexing" (paper presented at IFLA WLIC 2018: "Transform Libraries, Transform Societies," Kuala Lumpur, Malaysia, 2018), http://library.ifla.org/id/eprint/2214; Martijn Kleppe et al., *Exploration Possibilities: Automated Generation of Metadata* (The Hague: National Library of the Netherlands, August 23, 2019), https://doi.org/10.5281/zenodo.3375192.

18. Nicolas Fiorini et al., "PubMed Labs: An Experimental System for Improving Biomedical Literature Search," *Database: The Journal of Biological Databases and Curation* 2018 (September 18, 2018), https://doi.org/10.1093/database/bay094; Victoria L. Rubin, Yimin Chen, and Lynne Marie Thorimbert, "Artificially Intelligent Conversational Agents in Libraries," *Library Hi Tech* 28, no. 4 (2010): 496–522, https://doi.org/10.1108/07378831011096196.

19. Catherine Nicole Coleman, "Artificial Intelligence and the Library of the Future, Revisited," *Stanford Libraries Digital Library Blog*, November 3, 2017, https://library.stanford.edu/blogs/digital-library-blog/2017/11/artificial-intelligence-and-library-future-revisited.

20. Ben Shneiderman, "Human-Centered Artificial Intelligence: The Landscape of Autonomy" (slides presented at CNI Fall 2019 Membership Meeting, Washington, DC, December 9, 2019), https://www.cni.org/pbs/human-centered-artificial-intelligence-the-landscape-of-autonomy.

21. Jakeway et al., "Machine Learning + Libraries Summit."

22. Lindsay McKenzie, "A New Home for AI: The Library," *Inside Higher Ed*, January 17, 2018, https://www.insidehighered.com/news/2018/01/17/rhode-island-hopes-putting-artificial-intelligence-lab-library-will-expand-ais-reach.

23. Thomas Padilla, *Responsible Operations: DataScience, Machine Learning, and AI in Libraries* (Dublin, OH: OCLC Research, 2019), https://doi.org/10.25333/xk7z-9g97.

24. Select Committee on Artificial Intelligence of the National Science & Technology Council, *The National Artificial Intelligence Research and Development Strategic Plan: 2019 Update* (Washington, DC: Networking and Information Technology Research and Development Program (NITRD), June 2019), https://www.nitrd.gov/pubs/National-AI-RD-Strategy-2019.pdf.

25. 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.

26. Barend Mons, "FAIR Science for Social Machines: Let's Share Metadata Knowlets in the Internet of FAIR Data and Services," *Data Intelligence* 1, no. 1 (Winter 2019): 22–42, https://doi.org/10.1162/dint_a_00002.

27. 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.

28. 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/.

29. Plantin, Lagoze, and Edwards, "Re-Integrating Scholarly Infrastructure."

30. Michelle Addington and Lorraine Haricombe, *Task Force on the Future of UT Libraries: Final Report* (Austin: University of Texas at Austin, 2019), https://utexas.app.box.com/v/future-of-ut-libraries-report.

31. Thomas Padilla et al., “Final Report --- Always Already Computational: Collections as Data,” May 22, 2019, https://doi.org/10.5281/zenodo.3152935.

32. Rebecca Springer, “Counting Data Librarians,” *Ithaka S+R Blog*, July 29, 2019, https://sr.ithaka.org/blog/counting-data-librarians/.

33. Lorcan Dempsey, “Library Collections in the Life of the User: Two Directions,” *LIBER Quarterly* 26, no. 4 (2016): 338–59, https://doi.org/10.18352/lq.10170.

34. Carole Palmer, interview by author, October 30, 2019.

35. 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.

36. Clifford A. Lynch, “Reader Privacy: The New Shape of the Threat,” *Research Library Issues*, no. 297 (2019): 7–14, https://doi.org/10.29242/rli.297.2.

37. Marshall Breeding, “Strengthening Patron Engagement while Protecting Privacy,” *Computers in Libraries* 38, no. 8 (October 2018): 18–20.

38. Kyle M.L. Jones and Dorothea Salo, “Learning Analytics and the Academic Library: Professional Ethics Commitments at a Crossroads,” *College & Research Libraries* 79, no. 3 (April 2018): 304–23, https://doi.org/10.5860/crl.79.3.304.

39. Tony Ageh and Brent Reidy, interview by author, October 30, 2019.

40. Kennedy, *Research Libraries as Catalytic Leaders*.

41. Lynch, “Reader Privacy.”

Chapter 3: Facilitating Information Discovery and Use

Landscape Overview

The library's role as connector between researchers and information has evolved over hundreds of years. Historically, libraries amassed and disseminated broad and deep collections of print and digital resources to their local communities. To many constituents, this remains the primary perceived function of libraries today. Libraries continue to invest significant portions of their annual budgets to license and purchase information resources, and continue to use collection size as a primary metric of quality and value.[1] Academic libraries are adept at managing discrete publications: negotiating licenses and purchasing agreements, making content "discoverable via institutional systems populated with hand-crafted metadata,"[2] and ensuring long-term preservation. However, this model is being rapidly disrupted and displaced by a "greatly expanded scholarly record—one that is less dependent on papers and articles, and that is increasingly expressed in terms of networks of links and associations among diverse research artifacts."[3] The expanded scholarly record has engendered three interrelated challenges for library discovery and access.

1. **The types of information researchers seek is changing.** Researchers increasingly require access to information resources outside the traditional scope of library collections, from massive data sets, to visualizations, three-dimensional objects, and computer models. Many researchers work outside of and across traditional disciplinary boundaries and require information sources from a range of fields of study. For some researchers, metadata, rather than published content, may be the primary object of study.
2. **What researchers intend to do with that information is changing.** Researchers increasingly expect to mine, process, and analyze content. With knowledge production rapidly outpacing

human processing capacity, researchers will increasingly rely on machines to parse and interpret information. For example, experiments in unsupervised text mining of the scientific literature have demonstrated that the data in the existing published scientific literature contains a wealth of unrecognized discoveries.[4] Only by analyzing this content at scale can scholars identify the overlooked patterns and connections embedded in the scholarly record.

3. **How researchers go about looking for that information is changing.** Researchers increasingly expect search and discovery interfaces that support a range of inputs and outputs. For example, new math-aware search engines allow users to enter mathematical equations as search terms and return results based on similarities in either the structure or meaning of the equation.[5] The Dig That Lick project searches its large-scale corpus of jazz recordings for pattern similarities based on a user's input on a virtual keyboard.[6] In addition to accepting non-textual inputs, researchers increasingly expect searches to return personalized, context-aware results. As search practices vary widely by discipline, scholars desire discovery tools that align with their field's research methods and expectations.

Together, these changes in scholarly expectations signal a future in which the library catalog and other local discovery systems will diminish in value, in favor of web-scale discovery. The library's role in discovery is undoubtedly shifting, a trend accelerated by emerging technologies such as machine learning (ML). One expert interviewed for this report remarked that "the internet has put us [libraries] on a collision course with the world," threatening to disintermediate the library in the discovery process.[7] Some experts have suggested that commercial web-scale search may entirely replace local academic library discovery systems.[8]

Much of the literature on the future of discovery in libraries, along with the expert interviews conducted for this report, provides a resounding counterpoint. Authors and interviewees suggest that the networked environment presents a number of strategic opportunities for libraries, specifically related to helping researchers optimize their use of ML-enhanced search applications, text-mining tools, and other approaches to sifting through the data deluge;[9] making unique digital collections available and discoverable at an unprecedented scale; and meeting users where they are by making unique local resources available in web-scale discovery environments.[10] Key emerging technologies with an impact on discovery include ML, natural language processing (NLP), and computer vision.

The following sections detail these opportunities and highlight examples of academic and research library engagement with the range of emerging technologies that are driving and responding to changes in how scholars discover, use, and create information.

Strategic Opportunities

Invest in user-centered discovery tools

The widespread adoption of web-scale discovery tools, combined with a landscape of information overabundance, may "completely upend the notion that the library attempts to licence or provide access to all [published] material" and instead prompt libraries to focus on licensing (ML-powered) tools and services that navigate and curate content.[11]

An increasing emphasis on user-centered discovery positions the user, rather than the collection, as the organizing principle within a discovery environment.[12] Manifestations of this user focus include expanding functionality beyond "search and retrieval" to enable users to actively engage with, interact with, and supplement library collections.[13] Known-item and exploratory search can be supplemented with "current awareness" tools, that is, mechanisms that help scholars keep up with developments in their field;[14] automated text-processing tools that provide just-in-time article summaries; visualizations of the connections between different resources; the ability to create and curate personal collections that include library-held and external resources; or scholarly profiles that showcase a researcher's work and allow them to set up a personalized feed of newly published research based on their interests.

Some of the most promising uses of emerging technologies to make search and discovery more user-centered include ML-enhanced search, automated text-processing tools, recommendation systems, and virtual assistants. The following sections discuss each in more detail, including several examples of academic library adoption or engagement in each area.

ML-enhanced search

Many academic library search interfaces primarily rely on keyword matching against the full-text of a publication or its metadata record. This approach to information retrieval can be onerous for users, who must experiment with different search terms and combinations, contend with incomplete metadata, and sift through large volumes of search results. As one expert interviewed for this report noted, keyword search makes interdisciplinary research particularly difficult, as it often fails to bring together "parallel conversations."[15]

A range of new search and discovery tools are challenging the centrality of simple keyword search, or enhancing its power through machine learning. The options available to libraries and scholars include several tools tailored to academic literature discovery, including Yewno,[16] Iris.ai,[17] Dimensions,[18] and Semantic Scholar,[19] among others, which rely on NLP and other machine learning to enhance search results.[20] These new tools tout semantic search capabilities, which attempt to return results based on a query's meaning, rather than specific keywords. These and other search tools, which understand the semantic meaning of queries and can build associations between different discipline-specific terms for the same concept, will significantly lower barriers for scholars aiming to discover literature across domains.

Some next-generation discovery tools also aim to produce a more serendipitous search experience, one in which users can discover unlikely sources and unexpected connections. Google's Talk to Books experiment, for example, uses NLP to return potentially relevant book passages based on a user's query.[21] Users are encouraged to ask questions rather than enter search terms (that is, topics or entities). The Talk to Books algorithm then returns search results based on predictions of likely response statements. While Talk to Books does not purport to be a rigorous search tool, it may point to a redefinition of user expectations for information retrieval.

Next-generation search and discovery tools are also improving upon and pushing the boundaries of the traditional search results list. Yewno's underlying technology, for example, produces conceptual units from its vast corpus of literature using a deep learning network to extract and group topics, allowing searchers to explore a complex network of interrelated literature. The biomedical literature search tool PubMed, from the National Library of Medicine (NLM) combines a "state-of-the-art machine-learning algorithm trained on past user search history" with other indicators, such as an article's popularity and publication date, to attempt to deliver the most germane results and sort them by relevance.[22]

Librarians have much to bring to the table in designing, enhancing, and selecting appropriate ML-powered search tools. Librarians' specialized skill sets in managing information could be redirected towards automating processes that remain largely manual. For example, librarians' expertise working with controlled vocabularies and mapping ontologies could be productively applied to training ML models that facilitate interdisciplinary search. Their information literacy and search expertise can help scholars productively select appropriate search tools depending on their goals (for example, a comprehensive literature review versus getting quickly caught up on a topic). Libraries can help ensure that scholars and students understand the limitations and downsides of ML-enhanced search, reminding them that "[b]lindly using any research engine doesn't answer every question automatically."[23]

Perhaps more significantly, libraries can offer their attention to the values of transparency and integrity in the scholarly research process. "Explainable" or "human-centered AI" have emerged as the bywords for transparency and integrity in algorithm-based information tools, and are cited as a crucial feature of the services that libraries acquire, license, or otherwise support.[24] In the context of search and discovery, human-centered AI reveals the "thought process" behind the algorithm, making it clear to the user why they are seeing a certain set of search results, and gives the user some level of control over the algorithm. For example, transparent discovery interfaces might allow users to "adjust the parameters of an algorithm being applied to a collection."[25] One of the experts interviewed for this report underscored the risk of "black box" algorithms to the integrity of the research process, explaining that "once we're in the bot-driven world, it would be trivial for businesses running those bots to tweak algorithms to privilege research from their own publications, and there would be incentives for them to do that."[26]

The promise of ML to enhance discovery goes beyond search tools. Scholars are also turning to a range of emerging technologies that, in the words of one expert interviewed for this report, "distill an overwhelming amount of content into something meaningful and manageable."[27] These include automated text-processing technologies, recommendation systems, and virtual assistants and conversational agents.

Automated text processing

ML tools can generate increasingly accurate content summaries using techniques that are extractive (in which the model abridges text by distinguishing relevant and irrelevant passages) and abstractive (in which the model attempts to interpret and paraphrase content). Google's TensorFlow machine-learning library can perform both types of summarization with high accuracy,[28] and commercial services like Scholarcy have emerged to allow non-computer scientists to take advantage of automated text summarization.[29]

The applications of such tools are clear for scholars striving to keep up with recent publications in their fields. Automated text summarization, perhaps to a greater extent than a human-generated abstract, can help them digest more content at a superficial level and determine which content demands a closer read. The applications for digital libraries are also apparent. At Virginia Tech (VT), for example, the University Libraries and the Digital Library Research Laboratory partnered with a computer science course in fall 2018 to experiment with deep learning models to generate chapter-level summaries for a corpus of VT's electronic theses and dissertations (ETDs).[30] Automated generation of text summaries has the potential to greatly enhance discovery of textual materials in digital libraries and save countless hours of human labor.

Beyond summarization, automated text processing can help researchers discover new meaning and hidden connections in existing texts. For instance, a team of researchers ran a corpus of abstracts in materials science through the Word2vec unsupervised machine-learning algorithm. By associating and clustering related terms, the algorithm replicated existing categories in the domain materials science without human intervention.[31] Next, the researchers successfully trained the algorithm to predict which of a set of materials was most likely to have thermoelectric properties based on its semantic relationships in the corpus. The end goal is to develop a method for scientists to generate hypotheses and glean new insights based on existing literature.

Automated text processing can also be used to make research more accessible to heterogeneous user communities. Researchers at MIT have developed a tool that uses NLP to "read scientific papers and produce a short summary in plain English,"[32] which may be particularly useful to scholars conducting cross-disciplinary research. Get the Research, a project from Impactstory, aims to use NLP to generate plain-language summaries of research for the general public.[33] Machine translation, which has become reliable enough that it can be used for "translating non-English medical studies into English for the systematic reviews that health-care decisions are based on," could be used to make critical research available in the languages of communities that can use it.[34]

Automated processing of scholarly literature will also impact the ways in which research is evaluated. Publishers and publishing-service providers are increasingly exploring the potential of automated text processing to streamline operations, improve discoverability, and add value to their products. Meta Bibliometric Intelligence, for example, uses machine learning to extract likely topics from a submitted manuscript, gauge its relevance to the journal, and predict its impact, all in the name of streamlining editorial workflows and decision-making. An ML-powered tool developed by Scite.ai "automatically detects whether an article's citing papers were written in support or contradiction of the cited article claims."[35] As tools like these demonstrate proficiency, they might be incorporated into researcher evaluation systems, tenure and promotion decisions, and other determinants of scholarly merit. As with most ML tools, this presents both tremendous opportunities and risks. On the one hand, ML tools could provide a more accurate and nuanced understanding of a work's reception in the scholarly community. On the other, they can replicate and amplify biases, be prone to error or manipulation, and further alienate human judgment from critical decisions that affect a scholar's career.

Approaches to machine-**generated** text have also come a long way in recent years. An October 2019 New Yorker article used a predictive text algorithm to co-author an article on the future of writing in a post-AI world;[36] in early 2019 Springer Nature published a proof-of-concept machine-generated book that used abstractive text summarization to peruse a corpus of articles on lithium-ion batteries and produce a general overview of the topic.[37] In the near future, a machine may author the first draft of a researcher's manuscript, automating the rote work of describing materials and methodology. Manuscript Writer,[38] an AI-based tool from the company SciNote, has already proven successful at drafting the introduction, methodology, results, and references sections of a scientific article, liberating the researcher to focus on interpreting the results and writing the discussion section.[39]

Recommendation systems

One strength of ML algorithms is their ability to dynamically adjust and adapt as they receive new inputs. ML enables digital services that tailor themselves to their users; rather than mass produced and generic, ML allows web content to be "customized based on individual users' personas, needs, wishes, and traits—an approach known as mass personalization."[40]

Recommendation systems are one manifestation of mass personalization. ML-powered recommenders can suggest resources based on a user's query or based on the system's understanding of a user's preferences and interests. Such systems have proliferated in the context of e-commerce, streaming media, and social media sites. They seem particularly well suited for library discovery systems, given that researchers are frequently looking for all available content that relates to their research interests. Search platforms for academic literature increasingly incorporate recommendation systems as a complementary discovery tool (for example, Mendeley and Ex Libris's bX Article Recommender). Stand-alone applications, like Meta (backed by the Chan Zuckerberg Initiative)[41] and the recently released Scitrus platform,[42] provide a curated feed of content based on the system's evolving understanding of the user's interests.

While recommendation systems hold promise for streamlining the research process and enhancing serendipitous discovery, they rely on intensive collection and analysis of user data, which can compromise user privacy in ways that are anathema to most libraries. Specifically, recommendation engines, and other discovery systems that rely on personal data, can be perceived as compromising libraries' commitment to open inquiry, which requires the searcher to feel unconstrained by surveillance, and to have agency in the discovery process.[43] Linked data infrastructure, on the other hand, can embed the same types of "meaningful relationships" as recommendation engines, but in a way that "reflects some level of systematic thought and consensus within and among domains of knowledge."[44] Research has shown that students have a complicated relationship with algorithm-driven platforms, including discovery systems, and express a mixture of discomfort and resignation to the idea of being tracked online.[45]

Despite these risks, Clifford Lynch cautions libraries against "taking an absolutist approach to information collection, as opposed to more nuanced, transparent, and opt-in collection of data about user activities and interests," arguing that a refusal to provide convenient and sophisticated search tools may only serve to drive users away.[46] Instead, libraries can develop and advocate for discovery systems that leverage the power and convenience of recommendation engines and other forms of personalization in ways that respect user privacy and facilitate open inquiry. Libraries are already undertaking projects that aim to provide such privacy-aware alternatives. For example, librarians at the University of Illinois at Urbana-Champaign developed an open source plug-in for the VuFind library discovery system that uses anonymized borrowing data to cluster related items and provide recommendations to users. Rather than tracking an individual user's history and habits, the system infers associations based on items checked out in a single transaction.[47] Libraries have also come up with creative recommendation engines that encourage information literacy and robust research skills. At the University of Tsukuba in Japan, for example, the libraries are developing a recommendation engine that will be installed as a browser plug-in on the library's computers and will suggest library materials based on Wikipedia articles the user has accessed.[48] The system uses a convolutional neural network to automatically classify Wikipedia articles and identify related content in the library's collections.

Libraries have an opportunity to contribute approaches to personalization that provide convenience and support information literacy while minimizing and disclosing risks to user privacy, providing transparent opt-in mechanisms, and prioritizing strong cybersecurity practices.

Virtual assistants and conversational user interfaces

The ways that researchers seek information are being shaped by the prevalence of conversational user interfaces and voice-controlled virtual assistants. Virtual assistants have rapidly become ubiquitous in homes and offices, and on the web. Smart devices like phones and speakers come equipped with voice-activated virtual assistants that can perform basic information retrieval tasks, interact with other smart devices like light switches and thermostats, and communicate with other web-based services. Chatbots embedded in websites proactively offer information and assistance. This class of tools, known as virtual assistants, chatbots, or conversational agents, among other terms, gives and receives information in the form of conversational speech, simulating interaction with a human.

Libraries have been experimenting with chatbots since at least the early 2000s.[49] Contemporary chatbots tend to manifest as a pop-up instant-message window in the corner of the library website. Chatbots can answer many fact-based reference questions, and may even be adept at answering more complex queries. A team of liaison librarians at McGill University, for example, has been exploring the effectiveness of commercial voice assistants (Siri, Google Assistant, and Alexa) at providing front-line research assistance.[50] Other libraries are also experimenting with leveraging commercially available virtual assistants to perform library-specific tasks. For example, the University of Oklahoma has developed an Alexa skill that "allows library users to perform a voice search of LibGuides or Primo using vendor APIs."[51]

While virtual assistants do not obviate human-to-human interaction, they can make it easier to provide individualized, point-of-need service to library users at scale; ease the anxiety some students may feel when approaching a librarian or initiating a research task;[52] and function as a digital triage system, automatically directing users to appropriate services and resources. Thus, a proactive virtual assistant invites engagement and provides a gateway for more substantive interactions with human librarians. Jeff Steely, dean of Georgia State University

Library, invoked chatbots as an example of an emerging technology that can make library services more student-centered, advising that "engagement with a chatbot is really about starting the conversation."[53]

Given well-structured and accurate source data, chatbots can rapidly and precisely answer transactional questions about library hours, the status of loans, or the location of a call number range at any time of day or night, from any location. However, they require significant up-front investment, both in developing their functionality and populating them with information. After all, "At its core, a chatbot is a library of answers that are organised to respond to the goals of its user. Poor organisation of the library of responses will negatively impact the responses the chatbot chooses."[54] Chatbots cannot currently approach human proficiency in making inferences, asking clarifying questions, or interpreting ambiguity. At this stage in their maturity, voice-controlled virtual assistants such as Google Assistant, Siri, and Alexa, provide poor user experience, especially beyond very basic queries.[55]

Given their limitations, chatbots are typically offered alongside conventional visual interfaces. That could eventually change. As conversational user interfaces become increasingly sophisticated, they may completely supplant visual interfaces. In this scenario, instead of visiting Google (or a library catalog) and entering a text-based query, a user might instead encounter a proactive chatbot that asks what the user is looking for. The chatbot processes a natural language statement (such as, "three or four references for an article I'm writing on Anglo-Saxon literature, specifically in Wessex") and asks follow-up questions to refine the search (such as, "Do you require only articles or other types of content? Do the articles need to be peer reviewed?").[56] Libraries will have a role refining and maintaining these conversational agents as well as in educating users to optimize their use.

Highlighted Initiative

PubMed
National Library of Medicine
https://pubmed.ncbi.nlm.nih.gov/

PubMed's biomedical literature search tool combines a "state-of-the-art machine-learning algorithm trained on past user search history" with other indicators, such as an article's popularity and publication date, to attempt to deliver the most germane results and sort them by relevance, rather than recency.[57] Text snippets for each search result expose the algorithm's logic and make it easy for researchers to identify the most pertinent articles.

Reveal hidden digital collections through enhanced description

The acceleration of digitization and born-digital content creation has left libraries facing an ever-growing backlog of resource description. As libraries place increasing value on their unique local collections, they need new ways of making those collections discoverable to internal and external audiences, both human and machine. Accurate and comprehensive metadata are essential to the discovery, use, and preservation of digital collections, yet libraries lack the human resources to catalog content at the rate it is being created. Machine-learning approaches to automated metadata generation have shown promising results, opening up new possibilities for libraries to describe digitized collections of text, audio, and still and moving images at scale.

Discovery of textual materials has benefited greatly from advances in optical character recognition (OCR), which enables full-text search. However, structured metadata remains essential to discovery, making it easier for users to systematically identify pertinent items and enabling search aggregators to efficiently harvest and index content. To produce structured metadata at scale for large corpora of digitized texts, libraries are turning to NLP and named-entity recognition (NER) tools. At Northern Illinois University (NIU), the library is using NLP to extract topics from and generate subject headings for a collection of tens of thousands of dime novels.[58] These materials would otherwise require intensive human effort to productively catalog. A similar project is underway at the Koninklijke Bibliotheek, the National Library of the Netherlands, where an NLP algorithm is being trained to apply subject tags to a collection of electronic dissertations.[59] At Singapore's National Library Board (NLB), an experimental initiative utilized NER to populate metadata records across several digital collections.[60] The NLB's NER system extracts the names of places, people, and organizations from a full-text document and compares them against a controlled vocabulary supplied by subject experts. Entities recognized by the system can then be added to an object's metadata record. The project has enriched the metadata of collections that had little to no prior cataloging, and has bolstered cross-collection discovery.

While many efforts focus on text processing, machine learning also has significant implications for processing collections of still and moving images and audio. The British Library Machine Learning Experiment site, launched in 2015 as a test bed for the library's digital research team, is using open source software and public-image recognition APIs to automatically process and tag a collection of over a million public domain images.[61] Japan's National Diet Library (NDL), under the auspices of its Next Digital Library project, has created an illustration search tool to automatically extract images and diagrams from its 30,000 digitized publications, and group similar images across the collection.[62] The Center for Open Data in the Humanities is using a deep-learning-based classification algorithm to extract images, and recognize facial expressions from its collection of digitized Japanese manuscripts.[63] In this instance, the research team chose deep learning (as distinct from machine learning) in order to allow the machine to identify patterns independently.

A collaborative initiative from the Indiana University Bloomington Libraries, the University of Texas at Austin, New York Public Library, and digital consultant AVP, funded by a grant from the Andrew W. Mellon Foundation, also aims to create metadata-generation mechanisms for audiovisual content through an open source Audiovisual Metadata Platform (AMP).[64] To date, the project has piloted the application of "speech-to-text, named entity recognition, video OCR, speaker diarization, and speech/music/silence detection"[65] to a sample collection. Future work will include genre detection and instrument identification for digitized music and object detection for video. The National Library of Norway's Nancy initiative explores several vectors of machine learning for its cultural heritage collections, including a speech-to-text initiative that promises to make thousands of hours of radio broadcasts deeply searchable for the first time.[66]

Machine-learning approaches to metadata generation have been experimental since at least the 1980s, but the computing resources and technical expertise required to implement them presented significant barriers to wide adoption. Improvements in commercially available hardware, containerization technologies, the availability of public APIs and open source code, and the availability of high-speed networking on many university campuses have made it possible to implement machine-learning tools at scale. Using modern tools and computing resources equivalent to a standard laptop computer, a team of researchers indexed the 57 million pages of unstructured digitized text in the Biodiversity Heritage Library in 14 hours, an operation that previously took 45 days.[67]

The growth in available commercial machine-learning services can also lower barriers to entry in this space. Several of the initiatives described in this section rely on commercial cloud-based services for data processing. Amazon and Google both offer machine-learning services, as do dedicated vendors like Clarifai and Machine Box (which provides a containerized machine-learning environment). Microsoft has partnered with the Library of Congress and Israel's Ben-Gurion University of the Negev to apply machine learning to massive troves of digitized manuscripts.[68] The team behind the Audiovisual Metadata Platform (AMP) cautions that commercial machine-learning services lack transparency (using "black-box" algorithms to process data) and that vendor terms of service often require users to proactively opt-out of allowing data reuse.[69] Further, they warn, commercial tools may not be suitable for library use cases without considerable modification.

Indeed, many of the projects referenced above have noted the considerable effort involved in producing machine-generated metadata that matches human accuracy and precision. Significant human intervention is still required in the form of tweaking algorithms, supplying pertinent training data, and performing quality control.[70] The NLB in Singapore undertook multiple rounds of iteration before it was confident in the performance of its NER tool. The University of Utah, which recently received a grant to develop and test a machine-learning tool for its historical image collection, will rely on nearly a half-million digitized images with existing, detailed, human-created metadata as a training corpus.[71] Well-resourced libraries could collectively develop "gold-standard" training data sets that could be broadly shared within the cultural heritage community as a step towards making this technology accessible to institutions of all sizes.[72]

Machine-**assisted** cataloging may be a productive middle ground in the near term. The NIU dime-novel project, for example, will "aggregate unusual keywords into different top-level dime-novel genres, like seafaring, Westerns, and romance," allowing human catalogers to make educated inferences about a novel and complete the catalog record.[73] Western Washington University (WWU) is using a commercial service, Clarifai, for machine-assisted description of photographs and videos in its Islandora digital repository.[74] During the ingest process, images are sent to the Clarifai server for processing. They are returned with a set of suggested tags (and their confidence intervals). Human repository administrators can add or remove suggested tags before publishing the content.

As libraries grapple with the thorny technical challenges of automated resource description, they will also face critical questions about policy and implementation. Poor-quality metadata can undermine researchers' confidence in the search process; overly broad subject tags, for example, could exacerbate rather than mitigate the problem of an overabundance of material. Inaccurate metadata concerning locations, identities, or other factual information could have serious implications for research. Responsible approaches to integrating machine-generated metadata will therefore require clear indications to users. The British Library's machine-learning-powered search interface illustrates one approach: each metadata record includes a set of hand-created metadata fields and a clearly designated set of machine-generated tags with their corresponding confidence interval.

Perhaps more importantly, libraries will face ethical and privacy issues as they apply ML algorithms to their digital collections. Algorithms are prone to adopt and amplify biases, and are only as good as their training data.[75] Facial recognition and NER present even more significant concerns. Thoughtful policies about when and how ML is applied to library collections, and under what conditions it may be removed, can help libraries move forward on solid footing (for example, takedown notices for machine-generated metadata, particularly any metadata derived from facial recognition or NER, which might inappropriately identify living people, perpetuate biases, or expose sensitive information). ML techniques can also be applied to bolster data privacy (for example, using algorithms to automatically identify suspected Social Security numbers or other sensitive information in troves of digitized documents).

At this stage of maturity, automated metadata generation may be particularly advantageous as a "good-enough" tool for describing resources that might otherwise remain uncataloged. Though the quality and precision of machine-generated metadata may not yet match human-created metadata, its potential to describe collections at scale, to provide a minimum level of description for digitized objects that would otherwise remain hidden, represents a watershed moment for cultural heritage organizations. This is an opportunity for reflection on the ethical and privacy implications of machine processing massive volumes of digitized material.

Visual information has proliferated over the past several decades, from mass digitization of historical image collections, to the millions of digital photos and videos uploaded each day from personal electronic devices. Computer-vision technologies, often powered by convolutional neural networks, provide new ways of processing and exploring this deluge of information. Computer vision is an umbrella term that encompasses attempts to computationally replicate the human visual system and automate visual tasks, such as pattern and known-entity recognition.[76] Computer vision is already being used to detect cancer and other illnesses, identify wildlife whose images are caught on trail cameras, guide self-driving vehicles, and inspect food quality, among other experimental uses. Within the cultural heritage sector, computer vision can enable a range of novel approaches to visual-resource description, analysis, and discovery, giving researchers a range of options beyond text-based search (lexical or semantic). Libraries can apply these techniques to their own collections, enhancing broad discovery of visual materials, and support faculty projects that aim to process digital images at scale.

As discussed in the section on automated resource description, ML models have shown promise for identifying objects and known entities in visual materials, retrieving or grouping similar images, and generating topical or thematic metadata. Computer-vision techniques can be applied to digitized still images, moving images, textual documents that contain embedded figures, and even collections of 3D data, which will benefit from shape-based retrieval mechanisms that identify similar objects.[77] A number of notable projects are successfully using computer-vision techniques to engage with library collections.

As part of the Mellon-funded Collections as Data: Part to Whole project, a team at Harvard University and the University of Richmond will implement computer-vision techniques to analyze born-digital ephemera relating to the rise of nationalist and anti-immigrant movements in Europe.[78] The project's goals include "expanding the processing of digital images and subsequent algorithmic discovery of connections across collections." Notably, the project also explicitly aims to “illustrate how distant viewing can offer a paradigm for addressing the social and ethical challenges of using machine learning with images, particularly of sensitive topics.”

At Yale University Library’s Digital Humanities Laboratory, Doug Duhaime, Monica Ong Reed, and Peter Leonard, have used a convolutional neural network to analyze images from the Meserve-Kunhardt Collection of 19th-century photography at the Beinecke Rare Book and Manuscript Library.[79] While the typical end-result of this process would be a text-based caption or description of the image, in this case the researchers were interested in the penultimate level of interpretation, which clusters similar images together. They present the results in a visual interface that allows visual exploration of the photographs in a dynamic website. The related PixPlot tool, also developed at the Yale Digital Humanities Lab, offers an alternative visualization of the entire collection as a dynamic map of content, plotted based upon similarity, which allows pattern recognition at a glance.[80]

At Dartmouth College, researchers are working with a collection of films held by the library and the Internet Archive to develop a tool that allows users to search within moving images just like they would search for keywords in a document. The tool “takes search queries expressed in textual form and automatically translates them into image recognition models that can identify the desired segments in the film.”[81]

In addition to digitized and born-digital special collections content, computer vision also has applications for digging into the published literature. Scientists have used computer vision to analyze diagrams, visualizations, and images embedded in scientific papers, for the purposes of enabling new discovery and engaging in viziometrics research, or the study of the “organization and presentation of visual information in the scientific literature.”[82]

So far, the deep neural networks (DNN) that underlie computer-vision technology remain fragile and easy to fool. Researchers have shown that changing a few select pixels can cause a DNN to interpret an image of a lion as an image of a library, for example.[83] And computer-vision models, like other ML tools, are not optimized for use with cultural heritage materials. In collaboration with other cultural heritage institutions, and possibly with industry, libraries have an opportunity to contribute to building more appropriate training corpora, refining and testing models, and exploring the ethical and policy implications of broadly applying computer vision to their collections.

While the experiments described above are being run on carefully selected corpora by small groups of researchers, this type of functionality may eventually become commonplace in discovery and digital-asset management systems at scale. Libraries have a dual opportunity, supporting innovative, one-of-a-kind projects while generalizing the most promising methodologies and making them broadly available to researchers.

Highlighted Initiatives

Audiovisual Metadata Platform (AMP)
Indiana University Libraries, the University of Texas at Austin, New York Public Library
https://wiki.dlib.indiana.edu/display/AMP

The collaborative AMP initiative, funded by a grant from the Andrew W. Mellon Foundation, aims to create metadata-generation mechanisms for audiovisual content. To date, the project has piloted the application of speech-to-text; named-entity recognition; video OCR; speaker diarization; and speech, music, and silence detection to a sample corpus.

Image Analysis for Archival Discovery (Aida)
University of Nebraska–Lincoln and University of Virginia
http://projectaida.org/

The Aida project explores the application of neural networks to digitized library collections, particularly historic newspapers. The project has demonstrated success in identifying poetry from digitized newspaper images. The team's proof-of-concept suggests that libraries could eventually provide just-in-time, dynamically extracted content from their digitized collections.

Neural Neighbors
Yale University Library Digital Humanities Lab
https://dhlab.yale.edu/projects/neural-neighbors/

The Neural Neighbors project applies machine-vision techniques to a rich collection of 19th-century photographs to identify patterns and similarities, enabling new approaches to visual information discovery and analysis.

Sheeko
The University of Utah
https://sheeko.org/

Sheeko provides a suite of pre-trained ML models for automating image description as well as tools for users to automate the training of their own models.

Expose library collections and services beyond library systems

As information becomes distributed, diversified, and open, many researchers prefer web-scale discovery tools that aggregate resources from a range of sources over siloed library catalogs and digital-asset management systems.[84] Research libraries have a number of strategic opportunities to integrate library collections with a range of other open, digital resources, enriching the information available to users on the open web. Research libraries are meeting users where they are by implementing search engine optimization (SEO) techniques; exposing metadata for harvesting by aggregators, such as the Digital Public Library of America; providing APIs that permit new forms of computational engagement with collections; adopting interoperability standards, such as the International Image Interoperability Framework (IIIF),[85] to facilitate discovery and reuse; and participating in linked open data (LOD) initiatives. The shift towards revealing local collections to external audiences rather than the reverse, a trend Lorcan Dempsey has called the "inside-out library"[86] and one component of what other authors have termed the "library as platform,"[87] is a natural consequence of an open, oversaturated, and networked information landscape. The library's role in content management is being reenvisioned: no longer the steward of a unified collection, the library becomes the facilitator of a networked suite of open and extensible tools, resources, and services. Homegrown and manually maintained discovery systems may become less desirable to maintain as users increasingly turn to web-scale services and as emerging technologies enable more sophisticated discovery mechanisms. The academic library's facilitation services and interactions may supersede its role as a local content collector. Among the core functions of this role is advancing interoperability. Research library collaboration with interoperable repositories of data, preprints, and publications ensures that local troves of knowledge become discoverable at scale. Expertise in metadata and standards development can be contributed to maintaining and enhancing interoperability standards. Librarians' relationships with faculty and students on campus position them well to encourage adoption of persistent identifiers like ORCID IDs that help power interoperable discovery infrastructure, and the use of interoperable metadata schemas in faculty research.

In this vision of academic library services, the library no longer represents a "portal we go through on occasion, but…infrastructure that is as ubiquitous and persistent as the streets and sidewalks of a town."[88] The end users of this infrastructure will increasingly include both humans and machines.[89] A less institutionally driven approach to discovery might include working with vendor-supplied APIs to develop shared discovery layers, contributing to large-scale linked open data initiatives, or collectively developing systems that fill gaps in the discovery ecosystem, such as discovery of open access content. Academic libraries' existing expertise in standards and interoperability will be crucial as they participate in and enhance the "broader scholarly ecosystem, which only works through these frameworks."[90]

Highlighted Initiative

Enslaved: Peoples of the Historic Slave Trade
Matrix, the Center for Digital Humanities and Social Sciences at Michigan State University
http://enslaved.org/

The Enslaved project uses linked data to aggregate materials related to the transatlantic slave trade from a distributed network of library and archives partners. Bringing together disparate resources through linked data creates unprecedented opportunities for scholarly discovery and analysis, and brings light to the histories of underrepresented individuals and issues.[91]

Key Takeaways

1. **Libraries will retain a critical role in information discovery and facilitated access, even as locally acquired collections[92] diminish in importance.** The experts interviewed for this report overwhelmingly asserted that discovery will remain core to the identity and service model of the academic and research library, albeit in different and expanded ways.
2. **ML and NLP technologies will facilitate new forms of search, discovery, and academic inquiry.** At best, these technologies create exciting new modes of inquiry, facilitate cross-disciplinary discovery, and make research more efficient and productive. However, they have the potential to suppress human agency in the research process, amplify biases, and expose users to data-privacy violations.
3. **Library expertise can be effectively redirected towards creating and maintaining computationally ready digital collections that facilitate discovery, analysis, and use.** Libraries' expertise in creating and managing structured data can be effectively utilized to make local collections discoverable in web-scale discovery systems through more widespread adoption of APIs and linked open data. That expertise can also be used to

make digital assets more discoverable through the application of ML tools to resource description. Resources formerly invested in maintaining local catalogs might be repurposed into the purchase, licensing, or development of ML-enhanced search, discovery, and recommendation systems; compiling relevant training data sets for ML models; training virtual research assistants; and enabling other novel approaches to information retrieval and processing.

Endnotes

1. Lorcan Dempsey and Constance Malpas, "Academic Library Futures in a Diversified University System," in *Higher Education in the Era of the Fourth Industrial Revolution*, ed. Nancy W. Gleason, 65–89 (Singapore: Springer, 2018), https://doi.org/10.1007/978-981-13-0194-0_4.

2. Stephen Pinfield, Andrew M. Cox, and Sophie Rutter, *Mapping the Future of Academic Libraries: A Report for SCONUL* (London: SCONUL, 2017), https://sconul.ac.uk/publication/mapping-the-future-of-academic-libraries.

3. Anna Gold, "Cyberinfrastructure, Data, and Libraries, Part 2," *D-Lib Magazine* 13, no. 9–10 (September–October 2007), https://doi.org/10.1045/july20september-gold-pt2.

4. Michael Ridley, "Explainable Artificial Intelligence," *Research Library Issues*, no. 299 (2019): 28–46, https://doi.org/10.29242/rli.299.3.

5. Deanna C. Pineau, "Math-Aware Search Engines: Physics Applications and Overview," preprint, submitted September 8, 2016, http://arxiv.org/abs/1609.03457.

6. Dig That Lick website, accessed April 8, 2020, http://dig-that-lick.eecs.qmul.ac.uk/.

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

8. Roger C. Schonfeld, "Does Discovery Still Happen in the Library? Roles and Strategies for a Shifting Reality," *Ithaka S+R Blog*, September 24, 2014, https://sr.ithaka.org/blog/does-discovery-still-happen-in-the-library-roles-and-strategies-for-a-shifting-reality/.

9. Andrew M. Cox, Stephen Pinfield, and Sophie Rutter, "The Intelligent Library: Thought Leaders' Views on the Likely Impact of Artificial Intelligence on Academic Libraries," *Library Hi Tech* 37, no. 3 (2019): 418–35, https://doi.org/10.1108/LHT-08-2018-0105.

10. 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.

11. Cox, Pinfield, and Rutter, "The Intelligent Library."

12. Gwen Evans and Roger C. Schonfeld, *It's Not What Libraries Hold; It's Who Libraries Serve: Seeking a User-Centered Future for Academic Libraries*, Issue Brief (Columbus, OH, and New York, NY: OhioLINK and Ithaka S+R, January 23, 2020), https://doi.org/10.18665/sr.312608.

13. Marshall Breeding, *The Future of Library Resource Discovery: A White Paper Commissioned by the NISO Discovery to Delivery (D2D) Topic Committee* (Baltimore: National Information Standards Organization, February 2015), https://www.niso.org/publications/future-library-resource-discovery.

14. Schonfeld, "Does Discovery Still Happen in the Library?"

15. Eszter Hargittai, interview by author, November 27, 2019.

16. Yewno website, accessed April 10, 2020, https://www.yewno.com/.

17. Iris.ai website, accessed April 10, 2020, https://iris.ai/.

18. Dimensions website, accessed April 10, 2020, https://www.dimensions.ai/.

19. Semantic Scholar website, accessed April 10, 2020, https://www.semanticscholar.org/.

20. Andy Extance, “How AI Technology Can Tame the Scientific Literature,” *Nature* 561 (September 2018): 273–74, https://doi.org/10.1038/d41586-018-06617-5.

21. Talk to Books website, accessed April 10, 2020, https://books.google.com/talktobooks/.

22. Nicolas Fiorini et al., “PubMed Labs: An Experimental System for Improving Biomedical Literature Search,” *Database: The Journal of Biological Databases and Curation* 2018 (September 18, 2018), https://doi.org/10.1093/database/bay094.

23. Extance, “How AI Technology Can Tame the Scientific Literature.”

24. Ridley, “Explainable Artificial Intelligence.”

25. Thomas Padilla, *Responsible Operations: DataScience, Machine Learning, and AI in Libraries* (Dublin, OH: OCLC Research, 2019), https://doi.org/10.25333/xk7z-9g97.

26. Antelman, interview by author.

27. Keith Webster, interview by author, November 19, 2019.

28. Peter Liu and Xin Pan, “Text Summarization with TensorFlow,” *Google AI Blog*, August 24, 2016, http://ai.googleblog.com/2016/08/text-summarization-with-tensorflow.html.

29. Scholarcy website, accessed April 10, 2020, https://www.scholarcy.com/.

30. Naman Ahuja et al., “Big Data Text Summarization: Using Deep Learning to Summarize Theses and Dissertations,” VTechWorks, December 5, 2018, http://hdl.handle.net/10919/86406.

31. Olexandr Isayev, “Text Mining Facilitates Materials Discovery,” *Nature* 571 (July 2019): 42–43, https://doi.org/10.1038/d41586-019-01978-x.

32. Elliot Jones, Nicolina Kalantery, and Ben Glover, *Research 4.0: Interim Report* (London: Demos, October 2019), https://demos.co.uk/wp-content/uploads/2019/10/Jisc-OCT-2019-2.pdf.

33. Rick Anderson, "Get The Research: Impactstory Announces a New Science-Finding Tool for the General Public," *Scholarly Kitchen*, November 12, 2018, https://scholarlykitchen.sspnet.org/2018/11/12/get-the-research-impactstory-announces-a-new-science-finding-tool-for-the-general-public/.

34. John Seabrook, "The Next Word: Where Will Predictive Text Take Us?," A Reporter at Large, *New Yorker*, October 14, 2019, https://www.newyorker.com/magazine/2019/10/14/can-a-machine-learn-to-write-for-the-new-yorker.

35. Arthur "A.J." Boston, "What Do You Mean? Research in the Age of Machines," *College & Research Libraries News* 80, no. 10 (November 2019): 565–68, https://doi.org/10.5860/crln.80.10.565.

36. Seabrook, "The Next Word."

37. Lettie Y. Conrad, "The Robots Are Writing: Will Machine-Generated Books Accelerate Our Consumption of Scholarly Literature?," *Scholarly Kitchen*, June 25, 2019, https://scholarlykitchen.sspnet.org/2019/06/25/the-robots-are-writing-will-machine-generated-books-accelerate-our-consumption-of-scholarly-literature/.

38. "Manuscript Writer by SciNote," SciNote, accessed April 10, 2020, https://www.scinote.net/manuscript-writer/.

39. Jones, Kalantery, and Glover, *Research 4.0: Interim Report*.

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

41. Meta website, accessed April 10, 2020, https://www.meta.org/.

42. Scitrus website, accessed April 10, 2020, https://www.scitrus.com/.

43. D. Grant Campbell and Scott R. Cowan, "The Paradox of Privacy: Revisiting a Core Library Value in an Age of Big Data and Linked Data," *Library Trends* 64, no. 3 (Winter 2016): 492–511, https://muse.jhu.edu/article/613920.

44. Campbell and Cowan, "The Paradox of Privacy."

45. James F. Hahn, "User Perspectives on Personalized Account-Based Recommender Systems" (paper presented at ACRL 2019 Conference, Cleveland, OH, April 10–13, 2019), http://hdl.handle.net/2142/102364; Alison J. Head, Barbara Fister, and Margy MacMillan, *Information Literacy in the Age of Algorithms: Student Experiences with News and Information, and the Need for Change* (Project Information Literacy Research Institute, January 15, 2020), https://www.projectinfolit.org/uploads/2/7/5/4/27541717/algoreport.pdf.

46. Clifford A. Lynch, "Reader Privacy: The New Shape of the Threat," *Research Library Issues*, no. 297 (2019): 7–14, https://doi.org/10.29242/rli.297.2.

47. Jim Hahn and Courtney McDonald, "Account-Based Recommenders in Open Discovery Environments," *Digital Library Perspectives* 34, no. 1 (2018): 70–76, https://doi.org/10.1108/DLP-07-2017-0022.

48. Keita Tsuji, "Book Recommender System for Wikipedia Article Readers in a University Library," in *2019 8th International Congress on Advanced Applied Informatics* (IIAI-AAI) (IEEE, 2019): 121–26, https://doi.org/10.1109/IIAI-AAI.2019.00034.

49. Victoria L. Rubin, Yimin Chen, and Lynne Marie Thorimbert, "Artificially Intelligent Conversational Agents in Libraries," *Library Hi Tech* 28, no. 4 (2010): 496–522, https://doi.org/10.1108/07378831011096196.

50. Amanda Wheatley and Sandy Hervieux, “Need Research? Ask Siri: Evaluating the Impact of Virtual Assistant AI on the Future of the Research Process” (slides presented at NFAIS 2019 AI Conference, Alexandria, VA, May 16, 2019), https://nfais.memberclicks.net/assets/AI2019/Amanda%20and%20Sandy.pdf.

51. Twila Camp and Tim Smith, “Ready or Not: Here Comes Voice Search” (presented by Carl Grant, CNI Fall 2019 Membership Meeting, Washington, DC, December 9, 2019), https://www.cni.org/topics/information-access-retrieval/ready-or-not-here-comes-voice-search.

52. Indra Ayu Susan Mckie and Bhuva Narayan, “Enhancing the Academic Library Experience with Chatbots: An Exploration of Research and Implications for Practice,” *Journal of the Australian Library and Information Association* 68, no. 3 (2019): 268–77, https://doi.org/10.1080/24750158.2019.1611694.

53. Joan Lippincott et al., “Library Perspectives on the EDUCAUSE 2019 Top 10 IT Issues,” *EDUCAUSE Review*, February 11, 2019, https://er.educause.edu/articles/2019/2/library-perspectives-on-the-educause-2019-top-10-it-issues.

54. Mckie and Narayan, “Enhancing the Academic Library Experience with Chatbots.”

55. Raluca Budiu and Page Laubheimer, “Intelligent Assistants Have Poor Usability: A User Study of Alexa, Google Assistant, and Siri,” *Nielsen Norman Group* (blog), July 22, 2018, https://www.nngroup.com/articles/intelligent-assistant-usability/.

56. Lisa Janicke Hinchliffe, Jason Griffey, Emily King, and Michael Schofield, “Is the Researcher Human? Is the Librarian? Bots, Conversational User Interfaces, and Virtual Research Assistants” (presentation, CNI Spring 2017 Membership Meeting, Albuquerque, New Mexico, April 3, 2017), https://www.cni.org/topics/information-access-retrieval/is-the-researcher-human-is-the-librarian-bots-conversational-user-interfaces-and-virtual-research-assistants.

57. Fiorini et al., "PubMed Labs."

58. Matthew Short, "Text Mining and Subject Analysis for Fiction; or, Using Machine Learning and Information Extraction to Assign Subject Headings to Dime Novels," *Cataloging & Classification Quarterly* 57, no. 5 (2019): 315–36, https://doi.org/10.1080/01639374.2019.1653413.

59. Martijn Kleppe et al., *Exploration Possibilities: Automated Generation of Metadata* (The Hague: National Library of the Netherlands, August 23, 2019), https://doi.org/10.5281/zenodo.3375192.

60. Rachael Goh, "Using Named Entity Recognition for Automatic Indexing" (paper presented at IFLA WLIC 2018: "Transform Libraries, Transform Societies," Kuala Lumpur, Malaysia, 2018), http://library.ifla.org/id/eprint/2214.

61. British Library Machine Learning Experiment website, accessed April 10, 2020, http://blbigdata.herokuapp.com/.

62. Next Digital Library website, accessed April 10, 2020, https://lab.ndl.go.jp/dl/.

63. Asanobu Kitamoto, "Facial Collection (Face Collection)," Center for Open Data in the Humanities, accessed April 9, 2020, http://codh.rois.ac.jp/face/.

64. AMPPD (Audiovisual Metadata Platform Pilot Development) website, last modified November 20, 2019, https://wiki.dlib.indiana.edu/pages/viewpage.action?pageId=531699941.

65. Jon Dunn and Shawn Averkamp, "Commercial ML Tools in Metadata Production" (slides presented at Machine Learning + Libraries Summit, Washington, DC, September 20, 2019), https://wiki.dlib.indiana.edu/display/AMP/AMP+Presentations?preview=%2F549127083%2F549127086%2FAMP+LC+ML%2BLibraries+2019-09-20.pdf.

66. Svein Arne Brygfjeld, “AI: Lessons from the National Library of Norway” (slides presented at SCONUL Summer Conference 2019, Manchester, UK, June 12, 2019), https://www.slideshare.net/secret/x2eTpP3OTUHzYi.

67. Dmitry Mozzherin, Alexander A. Myltsev, and David Patterson, “Finding Scientific Names in Biodiversity Heritage Library, or How to Shrink Big Data,” *Biodiversity Information Science and Standards* 3 (2019), https://doi.org/10.3897/biss.3.35353.

68. Zachary Keyser, “Microsoft Implementing AI in Creating Archive of Ben-Gurion’s Handwritten Works,” Israel News, *Jerusalem Post*, November 6, 2019, https://www.jpost.com/Israel-News/Microsoft-implementing-AI-in-creating-archive-of-Ben-Gurions-handwritten-works-607013.

69. Dunn and Averkamp, “Commercial ML Tools in Metadata Production.”

70. Goh, “Using Named Entity Recognition”; 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/.

71. Valeri Craigle, “Law Libraries Embracing AI,” in *Law Librarianship in the Age of AI*, ed. Ellyssa Kroski (Chicago: American Library Association, 2019), http://dx.doi.org/10.2139/ssrn.3381798.

72. Padilla, *Responsible Operations*; Michael Ridley, “Training Datasets, Classification, and the LIS Field,” *Library AI* (blog), September 26, 2019, https://libraryai.blog.ryerson.ca/2019/09/26/training-datasets-classification-and-the-lis-field/.

73. Short, “Text Mining and Subject Analysis for Fiction.”

74. “IBU (Islandora Batch Uploader),” Western Washington University, accessed April 9, 2020, https://mabel.wwu.edu/ibu.

75. Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (New York: NYU Press, 2018).

76. Thomas S. Huang, “Computer Vision: Evolution and Promise,” in *19th CERN School of Computing: Proceedings*, ed. Carlo E. Vandoni (Geneva : CERN, 1996), 21–25, https://doi.org/10.5170/CERN-1996-008.21.

77. David Koller, Bernard Frischer, and Greg Humphreys, “Research Challenges for Digital Archives of 3D Cultural Heritage Models,” *Journal on Computing and Cultural Heritage* 2, no. 3 (December 2009): 7:1–7:17, https://doi.org/10.1145/1658346.1658347.

78. Collections as Data—Part to Whole, “Announcing Collections as Data Cohort 2,” January 6, 2020, https://collectionsasdata.github.io/part2whole/cohort/.

79. Peter Leonard, “Neural Networks: Machine Vision for the Visual Archive” (presentation at CNI Spring 2018 Membership Meeting, San Diego, CA, March 13, 2018), https://www.cni.org/topics/special-collections/neural-networks-machine-vision-for-the-visual-archive.

80. PixPlot project webpage, Yale University Library Digital Humanities Laboratory, accessed April 9, 2020, https://dhlab.yale.edu/projects/pixplot/.

81. *Digital Humanities 2018: Puentes—Bridges: Book of Abstracts/Libro de resúmenes* (Mexico City: Red de Humanidades Digitales, 2018), https://dh2018.adho.org/abstracts.

82. Po-Shen Lee, Jevin D. West, and Bill Howe, “Viziometrics: Analyzing Visual Information in the Scientific Literature,” *IEEE Transactions on Big Data* 4, no. 1 (March 2018): 117–29, https://doi.org/10.1109/TBDATA.2017.2689038.

83. Douglas Heaven, “Why Deep-Learning AIs Are So Easy to Fool,” *Nature* 574 (October 2019): 163–66, https://doi.org/10.1038/d41586-019-03013-5.

84. David Attis and Colin Koproske, “Thirty Trends Shaping the Future of Academic Libraries,” *Learned Publishing* 26, no. 1 (January 2013): 18–23, https://doi.org/10.1087/20130104; John Akeroyd, “Discovery Systems: Are They Now the Library?,” *Learned Publishing* 30, no. 1 (January 2017): 87–89, https://doi.org/10.1002/leap.1085.

85. Stuart Snydman, Robert Sanderson, and Tom Cramer, “The International Image Interoperability Framework (IIIF): A Community & Technology Approach for Web-Based Images,” in *Archiving Conference*, vol. 2015 (Springfield, VA: Society for Imaging Science and Technology, 2015), 16–21.

86. Dempsey, “Libraries and the Informational Future.”

87. Schonfeld, “Does Discovery Still Happen in the Library?”; David Weinberger, “Library as Platform,” *Library Journal*, September 4, 2012, https://www.libraryjournal.com?detailStory=by-david-weinberger.

88. Weinberger, “Library as Platform.”

89. Chris Bourg, “What Happens to Libraries and Librarians When Machines Can Read All the Books?,” *Feral Librarian* (blog), March 16, 2017, https://chrisbourg.wordpress.com/2017/03/16/what-happens-to-libraries-and-librarians-when-machines-can-read-all-the-books/; Cox, Pinfield, and Rutter, “The Intelligent Library.”

90. Carole Palmer, interview by author, October 30, 2019.

91. Amy Crawford, “A Massive New Database Will Connect Billions of Historic Records to Tell the Full Story of American Slavery,” *Smithsonian Magazine*, January 2020, https://www.smithsonianmag.com/history/massive-new-database-connect-billions-historic-records-tell-full-story-american-slavery-180973721/.

92. Lorcan Dempsey, “Library Collections in the Life of the User: Two Directions,” *LIBER Quarterly* 26, no. 4 (October 11, 2016): 338–59, https://doi.org/10.18352/lq.10170.

Chapter 4: Stewarding the Scholarly and Cultural Record

Landscape Overview

Libraries bear responsibility not only for providing immediate access to broad and deep research collections, but for the long-term preservation of the scholarly record and the documentary evidence that comprises society's digital cultural heritage. The practices of information stewardship are being challenged by an expanded scholarly and cultural record that is "mutable and dynamic,"[1] unwieldy in its size and complexity, inextricably networked (that is, dependent on other components for context and interpretation), and ephemeral.[2] Many digital outputs are created within closed systems using proprietary technologies that further complicate content harvesting and preservation. Digital formats also pose new challenges for libraries in ensuring authenticity of digital content. Memory institutions are built on trust: the trust that materials under their stewardship are authentic, immutable, and preserved in perpetuity or deaccessioned through a transparent and well-understood process.

The complexity of digital stewardship, and the inversion of value brought about by the networked environment, make preservation of local collections all the more critical. Unique holdings, rather than mass-distributed scholarly resources, are becoming the research library's most valuable assets; libraries have a key role in stewarding this "hyperlocal digital memory."[3]

Stewarding the digital record requires new approaches to managing, "in a transparent and authentic way, support and context for the massively increasing volume of digital content at levels of rapid upward scalability."[4] All of these characteristics of the digital record—its diversity, scale, ephemerality, disaggregation of scholarly communications, and restrictive licensing of digital content—complicate this challenge. They require that memory institutions engage in proactive, upstream, capture processes, rather than the retroactive collecting that has characterized archival and collection development work for centuries.[5]

Yet, while funding and cooperation around mass digitization of physical artifacts has been robust over the last two decades, a similar approach has yet to crystallize for born-digital materials. A proactive approach to the preservation of the born-digital record requires technical, social, and legal solutions. Several of the experts interviewed for this report indicated a pressing need for coordinated, cross-institutional collaboration in order to adequately preserve the digital scholarly and cultural record.

The following sections explore several of the emerging technologies that pose new challenges and offer new solutions to managing digital content throughout its life cycle. These sections address the library's role in advancing open research and publishing practices, reinforcing integrity and trust in the scholarly and cultural record, and preserving the evolving scholarly and cultural record.

Strategic Opportunities

Advance open research and publishing practices

Long-term preservation is in some ways contingent on, or at least the beneficiary of, advances in open scholarship. By supporting open research practices—including the adoption of open metadata standards, creation of machine-readable publications, and depositing outputs (including underlying data and code) in open repositories—libraries make research more discoverable, reusable, reproducible, and durable. Libraries themselves have established open access publishing programs, leveraging new and existing technology infrastructure to develop, host, and distribute scholarly and creative works.[6] Libraries also play a critical role in achieving FAIR (findable, accessible, interoperable, and reusable) research data through their curation, education, and preservation activities.[7] Realizing the vision of FAIR scholarship will be a central challenge for the research community over the next decade.[8] Supporting and engaging in open research and publishing practices improves both the quality of scholarship itself and the quality and manageability of the scholarly record.

The ease of publishing digital content has engendered a shift away from a federated scholarly record produced by established journal and monograph publishers and distributed through libraries. Decentralization of the scholarly record into an assortment of institutional repositories, disciplinary repositories, social sharing sites, small web-only publications, personal blogs, and other channels, creates the need for a more resource-centric approach to dissemination, discovery, evaluation, and preservation of scholarship. Resource-centric scholarly communications relies on making research outputs "discretely exposed, portable, networked, and pluggable in a common way, presenting a rich content layer that serves as the foundation for the development of value added services, like peer-review, social networking, recommender systems, usage measures, and so on."[9] In an environment of "network-enabled literature," content filtering, currently enabled through peer review of individual papers for particular journals, will be superseded by "powerful, online filters" that "distil communities' impact judgements algorithmically, replacing the peer-review and journal systems."[10] The application of machine learning (ML) in scholarly communications processes could accelerate this trend, potentially replacing traditional publishing processes with "a set of decentralized, interoperable services that are built on a core infrastructure of open data and evolving standards."[11]

Many of the experts interviewed for this report cited research libraries' contributions to advocating for and facilitating the use of unique, persistent identifiers as key to enabling this new model of open scholarship and scholarly communications, calling such identifiers "crucial" and "paramount" at every stage of the research life cycle. Unique, persistent identifiers for research outputs can help define provenance, enable discovery, and ensure researchers receive appropriate credit for their work, among other important uses.[12] They are also imperative to support a shift toward a more "researchercentric model of scholarly communication," in which individual scholars themselves become a key organizing principle.[13] This shift, which is evident in new tools that facilitate discovery, collaboration, impact assessment, and other scholarly communications activities, depends on the ability of individual researchers to assert and define their unique digital identity and associate it with their intellectual outputs, their collaborators, affiliations, credentials, and other information. A 2018 survey of scientific researchers found that many are "actively engaged in defining their online identity to assert links to their work and communicate their research beyond conventional channels."[14] The authors cited ORCID[15] as the most widely adopted researcher identifier. Research libraries can contribute to addressing ongoing challenges related to the adoption and utility of persistent identifiers. For example, identifier registries remain siloed and limited in their scope: major services such as ORCID, CrossRef, and DataCite focus on one segment of the identifier landscape (researchers, articles, and data sets, respectively) and do not adequately cover the entities that comprise the scholarly communications network.[16] With their expertise in standards and discovery systems, and their relationships with the research community, research librarians are well-positioned to collaborate with identity registries to promote interoperability, encourage common practices, and move towards a more networked scholarly communications system.

Highlighted Initiatives

Next Generation Repositories
Confederation of Open Access Repositories (COAR)
https://ngr.coar-repositories.org/

The COAR Next Generation Repositories Working Group aims to achieve interoperability between research repositories by "making the resource, rather than the repository, the focus of services and infrastructure." The group's technical vision centers on encouraging and enabling widespread adoption of unique identifiers to support dissemination and discovery of scholarship, and enable collaboration at scale.

TOME (Toward an Open Monograph Ecosystem)
Association of American Universities (AAU), Association of Research Libraries (ARL), and Association of University Presses (AUPresses)
https://www.openmonographs.org/

A joint initiative of AAU, ARL, and AUPresses, the TOME project is coordinating the production of open access digital monographs in support of a robust and sustainable scholarly publishing ecosystem. The project distributes its outputs through multiple open repositories.

Reinforce integrity and trust in the scholarly and cultural record

Memory institutions are built on trust: the trust that materials under their stewardship are authentic, immutable, and preserved in perpetuity or deaccessioned through a transparent and well-understood process. Emerging technologies pose new challenges to fulfilling the role of trusted steward. The assurance of authenticity, for example, is threatened by the ease of manipulating and altering digital media, and the complexities of determining provenance of digital materials. Deep fakes—counterfeit video, audio, still images, and textual content created using ML—pose a particular challenge. Research libraries have a range of digital forensics tools at their disposal to authenticate digital artifacts and collections at the time of accession and throughout their life cycle. They are also identifying secure pathways—possibly involving distributed ledger technologies (such as blockchain) and public key infrastructure (PKI)—to acquire copies of digital objects from sources they trust, documenting and proving the chain of custody, and any changes that have been made to it along the way.[17] After accessioning, fixity checking continually proves objects and collections do not change over time, due to degradation of the content, or to intentional or accidental manipulation. Underlying all of these processes is the need to maintain security and integrity of computing and storage operations in the face of cyberattacks and natural disasters. In their roles as educators, librarians can also help their constituents develop the skills needed to assess and critically engage with the integrity and reliability of information.

Fraudulent or altered content could enter the historical record and be presented as reality either because its inauthenticity was not detected at the time of accessioning, or because bad actors were able to introduce it by hacking into a records management system.[18] Even regular curatorial practices introduce opportunities for content alteration. For example, the practice of offering access copies of digital archival materials in non-original formats (for example, providing an MP4 video file in lieu of an original in an obsolete or proprietary format) significantly improves the usability of digital archival content, but also creates an opening for nefarious or incidental changes during the format conversion process.[19] If these changes go undetected and undocumented, they could have serious implications for research integrity.

ML makes such manipulation of content by bad actors attainable at scale. Individuals, corporations, and governments can engage in ever more sophisticated forms of information control, taking advantage of the curation algorithms that serve digital content, thereby "recursively intermediating our realities according to evolving internal logics that we cannot see."[20] Bad actors may also be motivated and increasingly able to "tamper dynamically with the historical record."[21]

As the gravity and imminence of threats to the integrity of the historical record become increasingly apparent, librarians, archivists, and their collaborators are exploring new methods to ensure and reinforce trust in cultural heritage institutions as stewards. Ideally, workflows and technological protocols document an immutable chain of custody, providing an assurance of authenticity throughout a digital object's life cycle. Emerging technologies can also be applied to authenticating digital records, that is, tracking their provenance and chain of custody (for example, using distributed ledger systems) and comparing suspected fakes with a library of authenticated content to identify common elements that may have been co-opted.[22]

Projects including ARCHANGEL, from the National Archives of the UK and the InterPARES TrustChain project, for example, are exploring the application of distributed ledger technologies (such as blockchain) as a tool for ensuring the integrity of digital archival records. The ARCHANGEL project aims to use "blockchain to record checksums (cryptographic hashes) and other metadata derived from either scanned physical records or born-digital records to allow verification of their integrity over decade- or century-long time spans" and to preserve those hashes in a distributed peer-to-peer network.[23] The project is also experimenting with the use of deep neural networks to refine the process of ensuring the integrity of records while allowing broader access. For example, the project is using research from the University of Surrey Computer Vision Centre "to create a hash which is invariant to changes in format, but changes more drastically if the file is manipulated in other ways."[24] This means that a video file converted to a more accessible format could be verified as an authentic version of the original, while one with frames removed would be flagged as altered. The commercial service ARTiFACTS[25] provides blockchain-based registration of a scholar's intellectual outputs, allowing them to manage their intellectual property prior to publication and validate the origins of outputs attributed to them. ORCID, which provides unique, persistent identifiers for scholars, recently announced integration with ARTiFACTS, making it easier for scholars to link their scholarly identity to their research outputs.

A number of scholars have problematized the use of distributed ledgers for ensuring archival integrity and have pointed out the discrepancies between blockchain's theoretical advantages and the reality of implementation. The promise of blockchain as a comprehensive digital preservation solution may be exaggerated. At this time, blockchain technology has only demonstrated success in addressing one component of digital preservation: ensuring the integrity of metadata records.[26] The digital objects themselves are not integrated into blockchain's distributed network and must undergo a separate preservation process. Other authors have argued that blockchain comes up short even in accomplishing its core goal of ensuring authenticity. The premise that blockchain's distributed peer-to-peer networks ensure their neutrality, has received scrutiny, given that their operation generally depends on a core group of developers.[27] Longstanding methods of ensuring authenticity—entangling hashes and the protocol underlying the LOCKSS system—exceed blockchain's capabilities to provide "tamper-resistant storage against a powerful adversary."[28]

While blockchain may not "fundamentally alter archival practices," it may have a place as one element of the digital archivist's technology toolkit.[29] Blockchain will not replace other methods of ensuring provenance; will not obviate the need for migration, emulation, and other core approaches to content preservation; and will not eliminate the possibility of accidental or malicious corruption of digital records. However, it may become a useful underlying technology in records management systems and one method among many for ensuring the integrity of the scholarly and cultural record.

In addition to technologies that securely document provenance, collections stewards also need tools to detect altered or manipulated content in order to make strategic curatorial decisions: either refusing to accession the object or ensuring it is appropriately described. ML-powered tools can help effectively identify subtle indicators of faked media. For example, researchers have successfully used video analysis algorithms to analyze eye-blinking and detect heartbeats in order to identify fake videos.[30] These techniques are precarious, as the creators of deep fakes continuously enhance their processes to elude detection.

To maintain their status as trusted sources of information, memory institutions will need to deeply engage with current societal debates on the nature of trust and trusted systems. Decentralization and distribution, such as blockchain's distributed ledger, have emerged as new and explicitly anti-institutional methods of establishing provenance and authenticity. Blockchain's original use case as a cryptocurrency system, for example, was developed out of a distrust of traditional banks and financial institutions for financial transactions. Memory institutions have long relied on more traditional notions of institutional trust, a form of trust that is rapidly eroding along with trust in governments and many other institutions. Memory institutions face a formidable challenge moving forward: maintaining their current status as authoritative keepers of the historical record, while also embracing emerging technologies that distribute, decentralize, and open up digital trust relationships.

Highlighted Initiative

ARCHANGEL

UK National Archives

ARCHANGEL, from the National Archives of the UK and the InterPARES TrustChain project, is exploring the application of distributed ledger technologies (such as blockchain) as a tool for ensuring the integrity of digital archival records. The ARCHANGEL project aims to use “blockchain to record checksums (cryptographic hashes) and other metadata derived from either scanned physical records or born-digital records to allow verification of their integrity over decade- or century-long time spans” and to preserve those hashes in a distributed peer-to-peer network.[31]

Preserve the evolving scholarly and cultural record

A complex and expanding digital record has amplified the technical, social, and legal barriers to achieving digital preservation at scale. Over the last several decades, research libraries and their collaborators have made impressive headway in core digital preservation methodologies such as normalization, refreshing, migration, and emulation.[32] Yet, longstanding challenges have persisted even as new ones emerge. On the technical front, software, 3D data, dynamic web content, and massive data sets, among other media, push the limits of established digital preservation practices. The sheer volume of digital information produced each year means only a fraction can be reasonably preserved.

On the social and legal fronts, the increasingly distributed and licensed nature of scholarly content presents legal and administrative barriers. In addition to copyright challenges posed by digital materials, much “substantive digital content” resides within “proprietary social media systems and news platforms, potentially requiring agreements with a complex array of private entities to acquire or rescue content for preservation."[33] Content that resides within proprietary platforms is also particularly at risk of being irrevocably lost, as evidenced by numerous examples of abrupt service shutdowns that allowed little time for users or other entities to migrate data.[34] As libraries and archives contend with ever-growing quantities of digital information, the financial and human resources required to perform digital preservation at scale present a growing challenge. Even as digital storage costs continue to decline dramatically, they remain prohibitive for institutions preserving petabytes of data. Making archived data instantaneously retrievable, a core goal of many digital preservation efforts in academic libraries, exacerbates these costs. Increasingly, the environmental impact of storing digital information is coming under scrutiny.[35]

Even as emerging technologies have destabilized the digital preservation environment, they also offer new solutions and opportunities. Libraries and their collaborators are following developments in containerization, distributed ledger technologies (blockchain), new storage media, and automation of digital preservation practices through ML to help ensure that the expanded scholarly record remains accessible well into the future.

The expanding range of file types and formats that require preservation—from software and code, to three-dimensional data, to dynamic websites—presents a daunting challenge. As contemporary academic research moves away from static, immutable, end products, and towards dynamic and diverse networks of outputs, the assets that require digital preservation grow exponentially. Libraries are not only preserving a journal article, for example, but (multiple versions of) data and code that informed its results; comments, annotations, and reactions from the scholarly community; articles that reproduced, validated, or built upon the original scholarship; and more. Data in particular is rapidly moving from a static research product to a continuous flow of information. Libraries lack sufficient tools and protocols to manage and preserve these streams of networked information.

In order to ensure that the diversity of digital content remains usable over the long term, software preservation is an essential component of any digital preservation program. At the British Library, for example, the digital library program aims to preserve any and all software needed to access the digital objects in its collection, including "software required to open the file directly on current institutional computing technology; the migration and rendering software for such a preservation strategy; and emulators, base operating systems, and any other dependencies necessary to render the digital objects in question."[36]

Several collaborative initiatives aim to make software preservation attainable for libraries of all sizes, including the Emulation-as-a-Service project,[37] led by Yale University Library and supported by the Software Preservation Network, and ReproZip,[38] a software that allows users to capture digital content along with the environment in which it was produced, creating a preservation-ready package. By "capturing computing environments in which research takes place, [ReproZip] could be used to preserve software down to the operating system on which it runs."[39]

As researchers use 3D scanning and virtual reality (VR) tools to capture archaeological sites and artifacts, with the goal of preserving the world's cultural heritage, they "are doing little to conserve their own digital products."[40] The data underlying 3D and VR models is complex and varied, and often requires specific software for reuse and interpretation. Algorithmically generated 3D data (such as data produced through 3D scanning) is particularly difficult to "decouple from the technology used to create" it.[41] A lack of tools, standards, and other resources for 3D and VR data curation means that "in many cases scholars are still reduced to creating screenshots or video documentation of their VR/AR experiences, at least for archival purposes."[42] In the past several years, this problem has received considerable attention. The editors of a 2019 Council on Library and Information Resources (CLIR) report on VR and

3D data in libraries urged the community to "consider 3D/VR as scholarly products in their own right, rather than as illustrations or supplemental material," and therefore worthy of attention to the full suite of data life cycle services, including long-term preservation.[43] A number of libraries are actively engaging in this work.

The University of Virginia (UVA) Library has developed an approach it describes as 3D cultural heritage informatics (LIB3DCHI), which encompasses the "full scope of 3D data curation through the collection, processing, archiving, and distribution of data and its derivatives to the scholarly community."[44] The UVA Library emphasizes access and use, and has implemented Web3D technologies to help conveniently distribute 3D content and data through web browser interfaces. Responding to an "absence of standards and best practices for producing, managing, and preserving 3D and VR content," the collaborative VR Preservation Project[45] will explore metadata standards, infrastructure, and other requirements for preservation to complement the library's active programs supporting VR content creation and use. ML techniques are also being explored as a way to deal with the complexity and breadth of VR data preservation. For example, the game company Electronic Arts (EA), is using ML and AI tools to automate the process of recording every possible interaction with its VR environments in order to capture a comprehensive archival version rather than a single representative experience.[46] As libraries build their own VR and 3D content and advise their communities on best practices, an emphasis on adopting web-native and open 3D formats, where possible, will facilitate use, sharing, and preservation.[47]

One of the most perplexing issues for digital preservationists, web archiving, has only grown more challenging as static websites are replaced by dynamic and personalized feeds of information. The cultural heritage community has not developed sufficient capacity to capture web content, including "contemporary source materials like news, blogs, and online discussion forums," that are "vital to original scholarly research in the humanities and social sciences as they capture viewpoints and new trends and reflect how scientific discourses evolve."[48] As scholars increasingly seek to produce web-based outputs such as digital humanities projects and interactive visualizations in addition to or in lieu of more traditional publications, research libraries are experiencing high demand for digital preservation services. As long as these digital outputs remain fragile and ephemeral, they face significant obstacles to being considered equivalent to more durable forms of scholarship in tenure and promotion considerations.

Typical web crawlers, while able to operate at scale, lack the capacity to harvest dynamic information, instead gathering static snapshots.[49] Technologies like Webrecorder,[50] which allows users to capture live interactions with websites, offer one method of logging a representation of a website for long-term preservation, though these intensive methods break down at scale.[51] For scalable solutions to web archiving, researchers are exploring a number of options, including the potential application of a new web packaging standard introduced by Google in 2019[52] and the use of human-mediated web capture frameworks that can apply a set of heuristics defining elements to be captured to an entire class of web publications rather than individual websites.[53]

Finally, even as storage grows cheaper and more efficient, research libraries face an exponentially mounting volume of digital information; storage capacity remains a fundamental challenge for institutions aspiring to achieve large-scale digital preservation. The use of local and cloud servers for digital hosting and preservation seems likely to remain ubiquitous in the cultural heritage community as emerging storage options offer only marginal improvements over current technologies, or are far from ready for widespread adoption. Many emerging technologies are also unsuitable for providing instantaneous access, making them incompatible with the goals of many library digital preservation programs. Gains are being made as "engineers continue to eke out further performance and capacity gains from hard drives and flash storage—and researchers are developing next-generation technologies such as DNA storage, crystal etching techniques, and molecular storage that could hold massive amounts of data on a small object for hundreds of thousands of years or longer."[54]

Even as emerging technologies begin to provide solutions for automating the digital curation life cycle, it remains an expensive process that entails significant human intervention and judgment.[55] Curation needs depend on the nature of the collection: the format and characteristics of its contents, its intended uses and audiences, its sensitivity and cultural context. Collections that contain ethnographic materials or collections pertaining to or of marginalized communities require culturally appropriate curation methods that align with the values and interests of those communities.[56] Digital curation is therefore an active and collaborative process that requires interdisciplinary expertise and resists large-scale automation.

Highlighted Initiatives

Emulation-as-a-Service Infrastructure (EaaSI)
Yale University Library
https://www.softwarepreservationnetwork.org/projects/emulation-as-a-service-infrastructure/

EaaSI is building a network of institutional partners to build capacity for emulation beyond what any individual institution can offer. The program aims to offer third-party emulation services for memory institutions that allow them to provide access to digital media in an interactive (and where appropriate, secure and restricted) format via a standard web browser.[57]

VR Preservation Project
University of Oklahoma Libraries
http://vrpreservation.oucreate.com/

The University of Oklahoma Libraries aim to develop a set of common standards and best practices for the archiving and preservation of VR-related data. Led by Zack Lischer-Katz, CLIR Postdoctoral Fellow in Virtual Reality Preservation and Archiving for the Sciences, the two-year project will focus on developing both the guidelines and technologies necessary for VR content and software preservation.

Key Takeaways

1. **The growth of dynamic, networked, interactive information presents new challenges for digital preservation.** The scale, diversity, and complexity of digital artifacts complicates efforts to effectively steward the scholarly record and digital cultural heritage. The prevalence of dynamic digital formats such as VR, the curation of the web by inscrutable algorithms, and the siloing of digital content in proprietary formats and platforms, create obstacles for achieving large-scale digital preservation. Preservation of born-digital content depends not only on appropriate technologies to capture and curate it, but on the upstream practices that make content discoverable and harvestable. Both open standards and open licensing therefore are imperative to enabling collection and stewardship of scholarly information.[58]
2. **The ability to easily manipulate digital archival materials threatens trust in memory institutions.** Malicious actors, including individuals, corporations, and governments, have more methods than ever before to attempt to rewrite history through the creation of deep fakes, exploiting file format changes, and hacking digital archives. Memory organizations rely on trust from communities they serve that the information they provide accurately reflects the historical record, and this trust cannot easily be regained after it has been lost.
3. **Emerging technologies present both new solutions and challenges for long-term digital preservation.** Containerization technologies, advances in emulation, distributed ledgers, and ML tools all provide promising new approaches to long-term digital preservation. However, many digital preservation efforts are rooted in a historical, print-centric model of retroactive collecting and need to transition to coordinated and proactive upstream processes.
4. **Digital preservation at scale requires collaboration.** Many individual institutions are engaging in innovative digital preservation initiatives. However, achieving trustworthy, representative digital preservation at scale requires that these

technologies become part of a coordinated, cross-institutional, or even national approach to digital preservation. This coordinated approach must leverage institutional strengths and capacity, and also requires that research libraries continue to advocate for the adoption of the open standards, technologies, and protocols that make digital content available for harvesting and curation. Combining open standards and technologies with collaborative governance will allow for a more comprehensive approach to preserving the digital historical record.

Endnotes

1. Brian Lavoie et al., *The Evolving Scholarly Record* (Dublin, OH: OCLC Research, 2014), https://doi.org/10.25333/C3763V.
2. Carol A. Mandel, *Can We Do More? An Examination of Potential Roles, Contributors, Incentives, and Frameworks to Sustain Large-Scale Digital Preservation* (Arlington, VA: Council on Library and Information Resources, September 2019), https://clir-dlf.app.box.com/s/31tc6nrua3cj8jjwoymee78gl3plzlo2.
3. Jason Griffey, interview by author, November 14, 2019.
4. NDSA Coordinating Committee and NDSA Working Group co-chairs, *2015 National Agenda for Digital Stewardship* (National Digital Stewardship Alliance, September 2014), http://hdl.loc.gov/loc.gdc/lcpub.2013655119.1.
5. Mandel, *Can We Do More?*
6. *Library Publishing Directory 2020* (Atlanta, GA: Library Publishing Coalition, 2020), https://librarypublishing.org/wp-content/uploads/2020/02/library_publishing_directory_2020.pdf.
7. Mark D. Wilkinson et al., "The FAIR Guiding Principles for Scientific Data Management and Stewardship," *Scientific Data* 3 (2016), https://doi.org/10.1038/sdata.2016.18.

8. Directorate-General for Research and Innovation (European Commission) et al., *Future of Scholarly Publishing and Scholarly Communication: Report of the Expert Group to the European Commission*, Copyright, Fair Use, Scholarly Communication, etc. 97 (Luxembourg: Publications Office of the European Union, January 2019; Lincoln, NE: University of Nebraska–Lincoln Libraries, January 2019), https://digitalcommons.unl.edu/scholcom/97.

9. Confederation of Open Access Repositories, Eloy Rodrigues, and Kathleen Shearer, *Next Generation Repositories: Behaviours and Technical Recommendations of the COAR Next Generation Repositories Working Group*, Copyright, Fair Use, Scholarly Communication, etc. 64 (Lincoln, NE: University of Nebraska–Lincoln Libraries, November 28, 2017), https://digitalcommons.unl.edu/scholcom/64.

10. Jason Priem, "Beyond the Paper: The Journal and Article Are Being Superseded by Algorithms That Filter, Rate and Disseminate Scholarship as It Happens," *Nature* 495, no. 7442 (March 28, 2013): 437–41, https://doi.org/10.1038/495437a.

11. Priem, "Beyond the Paper."

12. Angela Dappert et al., "Connecting the Persistent Identifier Ecosystem: Building the Technical and Human Infrastructure for Open Research," *Data Science Journal* 16 (June 15, 2017): 28, https://doi.org/10.5334/dsj-2017-028.

13. Roger C. Schonfeld, "Who Is Competing to Own Researcher Identity?," *Scholarly Kitchen*, January 6, 2020, https://scholarlykitchen.sspnet.org/2020/01/06/competing-researcher-identity/.

14. Michela Bello and Fernando Galindo-Rueda, "Charting the Digital Transformation of Science: Findings from the 2018 OECD International Survey of Scientific Authors (ISSA2)," OECD Science, Technology and Industry Working Papers no. 2020/03 (Paris: OECD Publishing, April 9, 2020), https://doi.org/10.1787/1b06c47c-en.

15. ORCID website, accessed May 19, 2020, https://orcid.org/.

16. Dappert et al., "Connecting the Persistent Identifier Ecosystem."

17. Mark Bell et al., "Underscoring Archival Authenticity with Blockchain Technology," *Insights* 32, no. 1 (2019): 21, https://doi.org/10.1629/uksg.470.

18. Melanie Ehrenkranz, "How Archivists Could Stop Deepfakes from Rewriting History," News, *Gizmodo*, October 16, 2018, https://gizmodo.com/how-archivists-could-stop-deepfakes-from-rewriting-hist-1829666009.

19. Bell et al., "Underscoring Archival Authenticity with Blockchain Technology."

20. Glenn D. Tiffert, "Peering Down the Memory Hole: Censorship, Digitization, and the Fragility of Our Knowledge Base," *American Historical Review* 124, no. 2 (April 2019): 550–68, https://doi.org/10.1093/ahr/rhz286.

21. Tiffert, "Peering Down the Memory Hole."

22. Jacquelyn Burkell and Chandell Gosse, "Nothing New Here: Emphasizing the Social and Cultural Context of Deepfakes," *First Monday* 24, no. 12 (December 2, 2019), https://doi.org/10.5210/fm.v24i12.10287.

23. Bell et al., "Underscoring Archival Authenticity with Blockchain Technology."

24. Bell et al.

25. ARTiFACTS website, accessed May 19, 2020, https://artifacts.ai/.

26. Angela Woodall and Sharon Ringel, "Blockchain Archival Discourse: Trust and the Imaginaries of Digital Preservation," *New Media & Society*, November 22, 2019, https://doi.org/10.1177/1461444819888756.

27. Victoria L. Lemieux, “Blockchain Recordkeeping: A SWOT Analysis,” *Information Management* 51, no. 6 (2017): 20–27.

28. David Rosenthal, “Blockchain Solves Preservation!,” *DSHR’s Blog*, September 13, 2018, https://blog.dshr.org/2018/09/blockchain-solves-preservation.html.

29. Victoria Louise Lemieux, “Trusting Records: Is Blockchain Technology the Answer?,” *Records Management Journal* 26, no. 2 (2016): 110–39, https://doi.org/10.1108/RMJ-12-2015-0042.

30. Burkell and Gosse, “Nothing New Here.”

31. Bell et al., “Underscoring Archival Authenticity with Blockchain Technology.”

32. Oya Y. Rieger, *The State of Digital Preservation in 2018: A Snapshot of Challenges and Gaps*, Issue Brief (New York: Ithaka S+R, October 29, 2018), https://doi.org/10.18665/sr.310626.

33. Mandel, *Can We Do More?*

34. Richard Ovenden, “Libraries’ Role in Preserving Digital Information,” *Carnegie Reporter*, October 30, 2019, https://www.carnegie.org/news/articles/libraries-role-in-preserving-digital-information/.

35. Morgan David and Yuji Sekiguchi, “The Role of Optical Storage Technologies in Future Digital Archives,” in *Sustainable Audiovisual Collections through Collaboration: Proceedings of the 2016 Joint Technical Symposium* (Bloomington, IN: Indiana University Press, 2017), 170–76.

36. Peter May, Maureen Pennock, and David A. Russo, “The Integrated Preservation Suite: Scaled and Automated Preservation Planning for Highly Diverse Digital Collections” (paper presented at iPRES 2019: 16th International Conference on Digital Preservation, Amsterdam, The Netherlands, 2019), 10, https://osf.io/smq5w/.

37. Emulation-as-a-Service website, accessed May 19, 2020, https://www.softwarepreservationnetwork.org/projects/emulation-as-a-service-infrastructure/.

38. ReproZip website, accessed May 19, 2020, https://www.reprozip.org/.

39. Vicky Steeves, Rémi Rampin, and Fernando Chirigati, "Using ReproZip for Reproducibility and Library Services," *IASSIST Quarterly* 42, no. 1 (2018): 14, https://doi.org/10.29173/iq18.

40. David Koller, Bernard Frischer, and Greg Humphreys, "Research Challenges for Digital Archives of 3D Cultural Heritage Models," *Journal on Computing and Cultural Heritage* 2, no. 3 (December 2009): 7:1–7:17, https://doi.org/10.1145/1658346.1658347.

41. Jennifer Moore and Hannah Scates Kettler, "Who Cares about 3D Preservation?," *IASSIST Quarterly* 42, no. 1 (2018): 15, https://doi.org/10.29173/iq20.

42. 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/.

43. Zack Lischer-Katz et al., "3D/VR Creation and Curation: An Emerging Field of Inquiry," 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/.

44. 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/.

45. “3D Collection Strategies,” Virginia Tech University Libraries, accessed May 19, 2020, https://lib.vt.edu/research-teaching/lib3dvr.html; see also VR Preservation Project website, accessed May 19, 2020, http://vrpreservation.oucreate.com/.

46. John Tilbury, “Digital Preservation Futures: Looking ahead to 2030…,” *Preservica Blog*, March 20, 2019, https://preservica.com/blog/digital-preservation-futures-looking-ahead-to-2030.

47. Szabo, “Collaborative and Lab-Based Approaches to 3D and VR/AR in the Humanities.”

48. Rieger, *The State of Digital Preservation in 2018*.

49. Sawood Alam et al., “Supporting Web Archiving via Web Packaging,” preprint, submitted June 17, 2019, http://arxiv.org/abs/1906.07104.

50. Webrecorder website, accessed May 19, 2020, https://webrecorder.io/.

51. Martin Klein et al., “The Memento Tracer Framework: Balancing Quality and Scalability for Web Archiving,” in *Digital Libraries for Open Knowledge*, ed. Antoine Doucet et al., Lecture Notes in Computer Science, vol. 11799 (Cham, Switzerland: Springer, 2019), 163–76, https://doi.org/10.1007/978-3-030-30760-8_15.

52. Alam et al., “Supporting Web Archiving via Web Packaging.”

53. Klein et al., “The Memento Tracer Framework.”

54. Samuel Greengard, “The Future of Data Storage,” *Communications of the ACM* 62, no. 4 (April 2019): 12, https://doi.org/10.1145/3311723.

55. Carole Palmer, interview by author, October 30, 2019.

56. Uttaran Dutta, “Digital Preservation of Indigenous Culture and Narratives from the Global South: In Search of an Approach,” *Humanities* 8, no. 2 (2019): 68, https://doi.org/10.3390/h8020068.

57. Euan Cochrane, “Making Things EaaSIer: Overview from EaaSI’s PI,” Software Preservation Network, February 12, 2019, https://www.softwarepreservationnetwork.org/blog/making-things-eaasier-overview-from-eaasis-pi/.

58. Cameron Neylon, “More Than Just Access: Delivering on a Network-Enabled Literature,” *PLOS Biology* 10, no. 10 (October 23, 2012): e1001417, https://doi.org/10.1371/journal.pbio.1001417.

Chapter 5: Advancing Digital Scholarship

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

1. **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.
2. **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.
3. **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.

4. **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*.

Chapter 6: Furthering Learning and Student Success

Landscape Overview

Students intersect with a wider range of technologies over the course of their academic careers than ever before. From using electronic lab notebooks in data science courses, to exploring virtual recreations of archaeological sites, to participating in next-generation learning management and analytics systems, students' academic lives are filled with new technologies and new media. These exciting pedagogical opportunities require a range of new digital competencies. Students not only need access to technology, but they need the education to use it in informed and ethical ways. Libraries are natural partners in this process. As third spaces on campus, libraries can democratize access to software and hardware that students may not have through their program of study. Through existing digital fluency programs, libraries can help students understand the implications of using new digital tools and services, and help them critically engage with new media.

Research libraries provide a range of informal education and consultation to impart the digital skills that contribute to the academic and professional success of undergraduates, graduate students, and early career researchers. These include workshops that teach concrete digital scholarship and coding skills, such as programming languages,[1] software carpentry,[2] and data visualization;[3] research data management and open science practices; and scholarly communications topics such as copyright, identity management, and navigating academic publishing. Longer-term cohort-based educational programs have also become popular. These programs often encourage interdisciplinary engagement with an emerging technology over the course of a semester or longer.[4] A few research libraries have also launched formal programs that fill gaps in the academic curriculum, for example, the Temple University Libraries' interdisciplinary cultural analytics certificate.[5]
In addition to digital scholarship skills, research libraries have opportunities to help students critically engage with and optimize their use of a new generation of productivity tools, many powered by

machine learning (ML), that promise to assist users in a range of tasks related to learning and study.

The ease of publishing information and misinformation on the web, the growing sophistication of counterfeit content, and the use of black-box algorithms to generate and display information mean that achieving digital fluency[6] also requires that students be able to interpret and evaluate an unprecedented array of new media formats and sources. Students not only need to understand the credibility and reliability of textual media, but they also need data and algorithmic literacy skills, strategies for distinguishing between genuine and manipulated or fabricated digital content, and an understanding of online data privacy. Libraries are well-positioned to deliver an expanded digital fluency curriculum in partnership with faculty members, campus IT, and other collaborators. At the campus level, libraries also have a role in advocating for transparent, privacy-aware approaches to learning analytics as institutions increasingly collect sensitive student data at scale for the purposes of evaluating individual students and improving aggregate outcomes.

The following sections highlight some of the most influential technologies related to the learning enterprise, through the lens of the library's involvement in promoting digital fluency, participating in next generation digital learning environments (NGDLEs) and learning analytics initiatives, and supporting a range of new study and productivity tools.

Strategic Opportunities

Build digital fluency and digital scholarship skill sets

Librarians have long held a key role as educators, specifically contributing to information literacy by helping students identify relevant, reliable content. Historically, this has meant imparting strategies for discovering and evaluating suitable resources for their research and learning. The library's role in promoting information literacy has dramatically changed as search behavior has shifted away from the library catalog to web-scale discovery systems. At the same time, the definition of information literacy has significantly expanded alongside the proliferation of digital media. The ease of publishing information and misinformation on the web, the growing sophistication of counterfeit content, and the use of black-box algorithms to generate and display information means that achieving information literacy now requires students to interpret and evaluate an unprecedented array of new media formats and sources. Students not only need to understand the credibility and reliability of textual media, but they also need data and algorithmic literacy skills, strategies for distinguishing between genuine and manipulated or fabricated digital content, and an understanding of online data privacy.

The Pew Research Center has identified algorithmic literacy as a key societal challenge and cited a comment from one expert who predicted that without purposeful intervention through education, "there will be a class of people who can use algorithms and a class used by algorithms."[7] Whether or not students are aware, algorithms have come to shape their daily experience on the web, with significant implications for digital information discovery. Students routinely utilize "systems that predict, recommend, and speculate about [their] interests" based on their search history, social media engagement, and a host of other esoteric variables, processed through proprietary and opaque algorithms over which they have no control.[8]

Yet many students are unaware of the decisions, motives, and biases underlying search engines, news feeds, and other sources of digital information. Search is often considered a "neutral" activity, and students may take as a given that the content delivered by the search algorithm is the most objectively relevant to their needs.[9] Students are unlikely to receive guidance from their professors that addresses algorithmic literacy, and they may feel ill-prepared to critically engage with algorithmic platforms or resigned to having a lack of control in their digital interactions.[10] This lack of critical engagement with information discovery online has both micro and macro implications. On the level of a single search interaction, an opaque relevancy algorithm will likely influence a student's decision to use one information resource over another. In the grander scheme, "the immersion of algorithmic culture into everyday life has the potential to shift how decision making is enacted and agency is performed, in addition to what knowledges and ways of knowing are privileged."[11]

Librarians have a dual role in algorithmic literacy: raising awareness of and encouraging students to think critically about the black-box algorithms underlying information tools, and providing students with search strategies and systems that give them agency in the discovery process. Despite significant attention paid to many new digital fluency skills, algorithmic literacy has not yet permeated library school curricula, the ACRL Framework for Information Literacy, and other professional channels.[12] In 2017, the Institute of Museum and Library Services awarded Montana State University a grant that aims to improve algorithmic literacy among librarians, equipping them to better serve their communities. The project's deliverable, an open curriculum on algorithmic literacy, aims to address this gap.

A handful of other consortial and field-level initiatives also aim to establish libraries as leaders in algorithmic literacy. For example, the AI for All initiative, a pan-Canadian project led by Ryerson University Library, Toronto Public Library, and the Canadian Federation of Library Associations, will "design, deliver, evaluate, and sustain an algorithmic literacy program in Canadian public libraries that provides a variety of pedagogical approaches to understanding the key aspects of artificial intelligence (algorithms) and how they affect and empower individuals and society."[13] Project Information Literacy will release a new report in the coming months that aims to provide librarians with a better understanding of the effects of algorithms in the lives of students.[14]

Alongside the algorithms that mediate their digital experiences, students also face an increasingly complex media landscape. They contend with a proliferation of unreliable information, facilitated by the ease of self-publishing on the web, and, increasingly, with fabricated or altered content that can be difficult to identify. ML tools and generative adversarial networks (GANs) have made it increasingly simple to create altered or completely fabricated content online, from auto-generated text to "deep fake" videos, which are "the product of artificial intelligence or machine-learning applications that merge, combine, replace and superimpose images and video clips" to create a fake product that appears alarmingly authentic.[15] Image and video manipulation are not new, and while individuals coming of age in the digital era have developed a degree of healthy skepticism about the authenticity of visual media online, the sophistication of ML-powered tools enables the creation of fake content with unprecedented speed and perceived legitimacy. This environment leaves students ill equipped to distinguish between genuine and digitally manipulated content, and to determine its origins.[16]

Identifying fabricated and manipulated content is both a technological and social issue. Numerous technological approaches have been developed and deployed to determine the authenticity of video and images online. With each advance, deep fake creators develop new strategies for eluding detection.[17]Regardless of the effectiveness of these technical tools, students require nontechnical strategies for identifying and engaging with altered and fabricated content. Librarians can equip students with strategies for not only spotting suspicious content, but also for asking critical questions about how and why it might have been created, to what ends, and for whose benefit. In order to impart digital visual literacy to students, "It is not so much that we promote paranoia around the content, but alternatively prepare users to engage with technology going forward. We must avoid asserting the products of technology exist in isolation and instead ask how the products got to us in the first place."[18]

Finally, the scope of digital fluency now includes an understanding of how an individual's data is gathered and used on the web. Libraries are considering data privacy a core aspect of information literacy and incorporating it into their teaching.[19] Libraries can educate their communities about online data collection practices, assist students in understanding privacy policies and terms of service, and help their communities become more savvy digital consumers. These efforts may be particularly necessary given the growing number of commercial e-learning platforms that students may be required to use in their coursework.[20] Libraries can help students understand how such platforms collect and use their data, giving them the tools to advocate for their interests.

New approaches to information and digital fluency emphasize students' role as creators, not just consumers, of digital media. In addition to helping students develop skills in critically using and evaluating algorithmic systems, interpreting data, and spotting deep fakes, libraries are increasingly thinking about how students can become ethical creators of digital media. Bryn Mawr College's Digital Competencies Program, "a tool for students to use to reflect on the digital skills and critical perspectives they develop while in college," is managed by the college's Library and Information Technology Services and places design thinking and "critical making" alongside evaluating digital information sources and data literacy skills.[21]

A new generation of productivity tools, many powered by ML, promises to assist users in a range of tasks related to learning and study. As key resources for information literacy on their campuses, librarians have a role in helping students effectively, ethically, and responsibly select and use these emerging productivity apps. Academic libraries commonly host workshops, online resources, and individual consultation services to help their communities optimize their use of citation management, collaborative authoring, and personal digital information management tools, among others. In the coming decade, students are likely to adopt a growing number of new tools that promise to make learning and research easier, faster, and more productive. These may include voice to text transcription services such as Otter,[22] which uses ML algorithms to not only transcribe audio, but also to identify speakers and extract topics; Beautiful.ai,[23] which helps users create polished slide decks using an ML algorithm; Scholarcy,[24] which automatically summarizes text; or Trevor,[25] which uses AI in the service of task and time management.

These tools have tremendous potential benefits. Voice-to-text transcription, for example, could assist students with note-taking and qualitative research activities, and may be particularly helpful for students with hearing or learning disabilities. Automated text summarization tools could allow students to more easily identify content relevant to a particular assignment or area of interest. On the other hand, these tools entail myriad concerns around user privacy, plagiarism and cheating, and misuse. For example, like all ML-powered services, voice transcription has the potential to compromise user data and privacy. Otter's terms of service explicitly state that the app uses segments of voice recordings and transcriptions for its training corpus. Recordings are uploaded to a cloud server, risking exposure in the event of a hack or human error. Automatic text summarization, translation, and generation apps could make it easy and tempting for students to cut corners on writing assignments.

Library support for productivity tools, through workshops, web-based resources, or other channels, could help community members learn about new ways of streamlining or enhancing their research and study, while also encouraging them to think critically about the implications of using these tools, from understanding terms of use and data privacy, to thinking through how they relate to plagiarism and other ethical concerns. Libraries are also taking a seat at the table in campus-wide discussions about institutional adoption of and policies related to the use of these tools.

Highlighted Initiatives

Information Literacy in the Age of Algorithms report
Project Information Literacy
https://www.projectinfolit.org/algo_study.html

Project Information Literacy recently conducted focus groups with students and faculty at eight universities and colleges to understand "how college students conceptualize the ever-changing online information landscape, and navigate volatile and popular platforms that increasingly employ algorithms to shape and filter content." While students understand and resent that their personal information is being used to shape their online experiences, this topic is "rarely mentioned in the classroom, even in courses emphasizing critical thinking and information literacy."

Privacy Services
Cornell University Library
https://www.library.cornell.edu/services/privacy

Recognizing the centrality of supporting intellectual freedom to the library's mission, Cornell University Library recently unveiled a bundled suite of privacy services for students and faculty. Services include general digital privacy literacy workshops and consultations to help students and faculty identify and mitigate risks to their privacy while engaging in academic and personal activities online, as well as specialized privacy consultations for researchers engaging in particularly sensitive work or in contexts that expose them to increased risk.

Digital Competencies
Bryn Mawr College Library and IT
https://www.brynmawr.edu/digitalcompetencies

The Digital Competencies program at Bryn Mawr is managed by a blended Library/IT organization and blends concepts from the ACRL Framework for Information Literacy for Higher Education with "digital survival skills" and concepts for ethical digital media creation for students. Faculty members have incorporated digital competencies into their courses, and students are also encouraged to use them to "reflect on their skills, build skills based on their interests, and practice articulating their competencies to different audiences," including future employers.

Integrate with campus-wide platforms and initiatives that advance learning

Next generation digital learning environments are changing the way students engage with their instructors, advisors, peers, course materials, and the library. According to the EDUCAUSE Learning Initiative (ELI), the core features of an NGDLE include "interoperability and integration; personalization; analytics, advising, and learning assessment; collaboration; and accessibility and universal design."[26] NDGLEs comprise a modular network of "pedagogical tools and applications all connected by means of open standards," rather than a single overarching platform.[27] NGDLEs may encompass a learning management system as one component in a broader, dynamic infrastructure.

Yet, unlike learning management systems (LMSs)—which play a relatively passive role as host for digital course materials, discussions, and grades—NDGLEs incorporate adaptive learning and automated advising, risk-detection and predictive analytics, and other technology-enabled tools to actively evaluate and influence student success. For example, University of Notre Dame has implemented the Apereo Open Learning Record Warehouse as a dashboard for compiling student data from a variety of sources into visualizations that can be used to holistically track student progress, using Sakai as an LMS.[28]

Libraries have typically engaged with the LMS by providing links out to library resources, including general search tools and guidance, tailored subject guides, and contact information for subject specialists. Involvement with the LMS has often required significant investment, either in manually maintaining up-to-date resources for the range of individual courses using the system, or in developing dedicated widgets or portals that can function within the LMS environment.[29] The NGDLE gives libraries an opportunity to not just embed static resources into an external system, but to become a node that dynamically integrates and promotes relevant information and resources at the point of need.

Personalization is one of the core features of an NGDLE. The structure of an NDGLE is defined not only by the institution, but the user. Adaptive learning technologies will dynamically adjust content based on an individual learner's needs and progress, built-in recommendation engines will suggest relevant resources based on a student's courses, and ML-enhanced advising will provide students with individualized guidance throughout their education. Edtech vendors are now building AI into LMS systems, using learner data to study behavioral practices—such as learning styles, emotions, gestures and electro-dermal activation, speech, and online learner behavior types—and deliver personalized content that adapts to "prior learning experiences and performances; self-expressed student preferences in modes of delivery; analytical prediction of likelihood of success for the individual student through different modes of delivery; and much more."[30] In the future, this personalization might include curated library resources relevant to a student's classes or their specific research interests and suggestions for relevant library consulting services or workshops.

Learning analytics (LA) encompasses a range of data collection and analysis activities that "help educators discover, diagnose, and predict challenges to learning and learner success" and design interventions that improve student outcomes.[31] The infrastructure that enables these activities, commonly referred to as integrated planning and advising for student success (iPASS) systems, aggregates data from a range of sources: grades and engagement levels from learning management systems, analytics from electronic learning materials platforms, demographic data from student information systems, and participation in clubs and events from extracurricular involvement systems. Yet, data about engagement with library resources and activities are rarely included.[32]

Learning analytics systems have arisen from a confluence of challenges: increased scrutiny of higher education budgets, intractable student retention issues, and growing student debt loads, among others. The underlying motivation for higher education institutions is to understand which factors contribute to student retention and satisfaction, and which indicate an increased risk of academic failure. Equipped with this information, institutions can address macro- and micro-level challenges, from identifying ways to reduce the cost of education to providing early interventions that help a struggling student succeed in a course.[33]

Learning analytics "focus on leveraging human judgment," providing distilled information to human stakeholders—professors, advisors, administrators—to be combined with observation, dialogue, and interpretation.[34] Analytics represent one piece of a larger puzzle that helps universities understand a student's progress, identify whether and in what ways they are at risk of negative outcomes, and plan the most successful interventions.

iPASS systems enable this type of assessment through the use of both descriptive and predictive analytics. Descriptive analytics quantify a student's behavior (for example, how many hours they interacted with a platform or learning materials), while predictive analytics enable early warning systems to identify students who appear at risk of academic failure. Predictive analytics have come under particular scrutiny for their potential for misuse. One expert interviewed for this report described them as potentially transformative but "fraught with peril."[35]

Within this context, libraries have also come under increasing pressure to quantify their contributions to student success and to contribute data about student interactions with the library to analytics systems that generate a data picture of a student's academic life. Longstanding proxies for library impact such as collections usage, numbers of instruction sessions and consultations, and foot traffic to library buildings are being replaced or complemented by metrics that aim to understand the role of the library's activities on student outcomes. Studies focused on quantifying the library's contribution to student success have proliferated over the past decade. A meta-analysis of student success studies in libraries identified a 570-percent increase in such studies between 2013 and 2014.[36] Responses to a recent ARL

SPEC survey indicated broad uptake of learning analytics activities in libraries. Over 80 percent of respondents reported engaging in "library assessment projects that utilize educational and institutional data, data analysis methods, and share similar goals of other non-library learning analytics work."[37] These activities generally include collecting and analyzing reference, instruction, and circulation data, occasionally in combination with data provided by other campus units.

Despite this trend, academic libraries are not yet systematically participating in or contributing to *campus-wide* learning analytics efforts.[38] One exception, among others, is the DePaul University Library, which collaborated on the development of the campus iPASS system. Among other functions, the system allows faculty and advisers to seamlessly refer students to a librarian for research assistance.[39]

The absence of broad participation in campus-wide initiatives has a number of causes. Central among them is a lack of understanding, within the library and externally, about the relevance of library data to campus-level initiatives. Only half of respondents to the ARL SPEC survey felt that library data was "very important" to learning analytics initiatives at their institution.[40] Outside of the library, administrators used to thinking of the library as a collections-focused entity may not fully grasp its important contributions to student learning.

Data interoperability presents another barrier. Library data may require considerable cleaning and reconciliation to integrate with iPASS systems and with other campus data sets. For libraries opting to participate in iPASS systems, adopting interoperability standards and working with other institutional stakeholders is key to ensuring that library data counts.

Concerns about privacy have also hindered widespread participation. Half of respondents to the SPEC survey identified privacy concerns as a reason they limit which data they share across campus units. Learning analytics are susceptible to the same pitfalls as any big-data practice. Unlike traditional research practice, in which "actors seek consent for data gathering beforehand and use the data as means toward explicitly agreed-upon and respected ends," the affordances of big data encourage actors to collect massive volumes of information without an explicit purpose, and often unbeknownst to the individuals whose data is being collected.[41]

A meta-analysis of 54 studies that utilized library learning analytics data identified inadequate or undefined data security, retention, anonymization, informed consent, and other practices, and a general lack of attention to privacy issues among such studies.[42] Fewer than half of respondents to the ARL SPEC survey "reported having a records-management schedule or policy that controls the retention of learning analytics data."[43]

The impact of learning analytics systems on student success remains unclear. A number of institutions have reported evidence of concrete improvement in retention.[44] However, a literature review of 252 studies of learning analytics system implementations found "little evidence in terms of improving students' learning outcomes," with only 23 of the 252 studies the researchers reviewed presenting evidence of such an effect.[45] A greater proportion of studies (35 percent) found that learning analytics systems had a positive impact on student retention and completion rates.

Whether or not they directly impact student outcomes, LA systems can provide valuable information that helps libraries and other campus units improve services. LA systems can help libraries understand both general patterns (such as which library-related activities correlate with a student's grade point average) and answer specific questions (such as at what time of the semester a library intervention might have the most impact on a student's final grade). Identifying these patterns can lead libraries to further investigate patterns through qualitative research methods, indicate opportunities to pilot new approaches to service development and implementation, and inform activities that improve the library user experience. Librarians seeking to establish definitive, causative relationships between librarian interactions and student learning and success are unlikely to find quick and easy answers through engagement in learning analytics. Current learning analytics systems are built on correlations, not causations.

Legitimate concerns about potential adverse effects of LA initiatives, from the risk of data reidentification to the misuse of predictive analytics, have led some libraries to dismiss participation as inherently antithetical to library values. Other libraries have explored whether ethical and productive approaches to collecting and using student information are possible given additional investment and oversight, a commitment to transparency and informed consent, precautions against data reidentification, and attention to minimizing adverse effects. The library can bring this perspective to bear in campus conversations. European institutions, bound by the General Data Protection Regulation (GDPR), provide models for implementing these values. Jisc's *Code of Practice for Learning Analytics*, which enumerates the "responsibilities of educational institutions to ensure that learning analytics is carried out responsibly, appropriately and effectively" is a robust resource for libraries looking to influence LA initiatives on their campuses.[46]

If libraries opt out of campuswide or internal LA initiatives, they risk missing out on beneficial insights that can lead to concrete service improvements. They also risk downplaying the library's contributions to student learning and success in the eyes of campus administrators. A more productive approach may be to take a seat at the table, principles in hand.

Highlighted Initiatives

Library Learning Analytics Project
University of Michigan
https://libraryanalytics.org/

The IMLS-funded Library Learning Analytics Project aims to develop extensible best practices for library data "collection, storage and analysis" using University of Michigan student data as a testbed. One of the project's early deliverables is a privacy guide for libraries seeking to ethically collect and use student data.

Student Dashboard
Nottingham Trent University
https://www4.ntu.ac.uk/current_students/studying/student_dashboard/index.html

NTU's Student Dashboard reveals key academic engagement metrics to students in a visual dashboard, including library use, e-book usage, LMS logins, and card swipes into academic buildings. Students can then compare their engagement with an anonymized aggregate of peers in the same course. Exposing this data directly to students enables them to better understand the connections between their own academic engagement and success.

Democratize access to emerging technologies in library spaces

Technology-rich learning and information commons, collaboration studios, makerspaces, and labs are now commonplace in libraries. These spaces provide access to specialized software and hardware for fabrication (such as 3D printers, computer-aided design and drafting software); visualization (such as high-resolution displays); immersive reality (such as virtual reality [VR] headsets); and other digital research and creation methods. The success of these projects depends largely on their ability to bring together sophisticated equipment and software with a range of support services that help users fully exploit these tools and connect them to broader learning outcomes. Equipping a lab with state-of-the-art hardware and software will not on its own create the conditions necessary for students to create, innovate, and learn. When libraries apply their existing expertise as educators to new forms of knowledge production, they can help their communities thoughtfully and productively engage with technology.

Locating digital scholarship centers within libraries may also help to democratize and de-silo access to cutting edge technologies, encouraging cross-disciplinary collaboration and discovery.[47] While the hardware and software available in a library makerspace may be available to subsets of students through their department or college, in many cases the library is the only place that provides access to the entire campus community, regardless of affiliation. Asked about the rationale for building a new AI-focused lab at the University of Rhode Island Libraries, dean of libraries Karim Boughida explained, “When you have an AI lab in a specific college, the impression is that access is only for students of that college. Even if students are told they can use the space, there may be a percentage that may feel unwelcome, or that it is ‘not for me.’ In the library it will be different.”[48]

Many digital scholarship centers help build research communities of practice within the library building by offering semester-long fellowships to faculty and graduate students, hosting longer-term projects or interest groups, and creating durable research outputs that highlight collaboration between librarians, technologists, and disciplinary experts. These longer-term projects complement one-off workshops and events to create programs that are responsive to rapidly evolving needs and interests. The combination of access to software and hardware, collaboration space, and technology expertise has proved compelling to faculty members, bringing them back into the library building.

While many library makerspaces, digital scholarship centers, and labs support a wide range of technologies, libraries are paying particular attention to immersive reality and data science support.

Immersive reality studios

The presence of immersive reality technologies in libraries has grown significantly as academic institutions recognize the pedagogical and research applications of augmented, virtual, and mixed reality (AR, VR, and MxR). Often collectively referred to as immersive reality, these technologies “enable faculty and students to engage with highly detailed 3D data—from cultural heritage artifacts to scientific simulations—in new ways.”[49] Immersive reality can enhance learning experiences by allowing students and scholars to manipulate “rare, fragile, endangered, or microscopic”[50] resources or engage with remote, inaccessible, fictive, or ancient environments.[51] MxR may hold particular pedagogical potential because of its ability to blend "virtually reconstructed cultural content" with "physical cultural heritage elements at their natural location."[52] Libraries and cultural heritage institutions seem particularly well positioned to take the lead in pedagogical applications of MxR given their dual roles as educators and as stewards of cultural and historical artifacts.

The release of affordable, consumer-grade VR headsets and other technologies required to create and experience immersive reality environments has reduced barriers to entry and led to a boom in interest among libraries. While many academic libraries now have small collections of VR headsets for lending, only a few have started building full-fledged programs for immersive reality support. The University of Oklahoma (OU) Libraries's The Edge studio, public VR spaces at the University of Virginia (UVA) Library, and the TRAIL collaboration space at the University of Washington (UW) Health Science Library provide three noteworthy examples of immersive reality spaces as collaborative endeavors with explicit links to the undergraduate curriculum (in the case of OU and UVA) and faculty research (at UW).

At The Edge, a library-based makerspace, the OU Libraries have installed several VR terminals consisting of "a moveable chair-on-rails, coupled with a high-end gaming PC and an Oculus Rift HMD (head mounted display)."[53] The libraries have worked with classes in multiple disciplines to develop custom learning software and deploy it in the undergraduate curriculum. A recent course collaboration brought together students from three university campuses to collectively explore a VR environment simulating a remote cave otherwise inaccessible to the public.[54] The use of VR also allowed students to adjust lighting, zoom, and explore the environment in other ways that would be difficult in the physical world. The OU Libraries have found that incorporating VR into select courses had "significant positive impact on self-efficacy along dimensions related to completion of spatial tasks," an indication that VR can support learning outcomes, particularly in spatially oriented fields such as architecture.[55]

At UVA, a VR lab in the library invites students to engage spatially with research topics. Using the Unity VR platform, "a research topic is represented spatially by creating 'rooms' in a virtual museum that relate to the arguments in a paper. The details of the argument are expressed by images, text, audio, or video objects placed in a room much like objects in a museum exhibition."[56]

At UW Health Science Libraries, the TRAIL collaboration space originated as a general purpose translational research lab backed by the library' clinical information program and IT services.[57] When a faculty member reached out to the library with a specific request to test VR on the existing data wall, the library took the opportunity to consider how the space could accommodate VR experimentation on a larger scale. The library has generalized its planning process into a comprehensive toolkit for VR spaces design in libraries. The toolkit addresses both technical design considerations and theoretical concerns, ranging from the minimum and maximum room scale specifications based on the types of VR headsets employed to how library VR spaces in health science libraries can effectively protect patient privacy.

The growth of immersive reality spaces and services in libraries is yet another indication of the library's burgeoning role in "experimentation and knowledge production," and a promising avenue for libraries to demonstrate continued relevance as "both as the custodian and curator of all forms of research and educational data, and as a catalyst for innovation in scholarship and pedagogy."[58] Immersive reality initiatives, which require close partnerships between technologists and disciplinary experts, further reinforce the library's role as a hub for cross-disciplinary collaboration.

Data science centers

Data science programs have seen dramatic growth over the past several years, as universities hurry to keep pace with student interest and industry demand for skilled data scientists. The highly interdisciplinary nature of data science as a field requires new models of support services. Data science courses and programs are often established outside of existing departments[59] and draw in learners from a range of academic backgrounds and majors beyond computer science."[60]

In order to provide cross-disciplinary opportunities for students to deepen and apply their data science skills, some campuses are creating dedicated spaces equipped with the appropriate software, hardware and associated programming. For example, the Moore-Sloan Data Science Environments (MSDSE) project, initiated in 2015, sponsored the development of three data science environments" (DSEs) at New York University (NYU), the University of Washington and the University of California-Berkeley.

All three DSEs were established outside of existing departments; two of the campuses (NYU and UW) selected the library to host the new space.[61] Libraries were considered ideal sites given their commitment to interdisciplinarity and openness, two core characteristics of data science research, and the perception of libraries as neutral or third spaces without ties to specific departments or programs on campus, or external parties such as corporate research sponsors. Positioning data science centers in libraries or other neutral spaces, rather than within professional degree programs whose goals are primarily to prepare students for the job market, may result in different focuses and priorities.

The data science centers established through the MSDSE project, for example, all developed a focus on the ethical implications of data science and its contributions to the public good, even though this was not an explicit goal at the outset.[62] At URI, which recently established a first-of-its-kind library-based AI lab, the mission, according to chief technology officer for University Libraries Bohyun Kim, is "to help students and faculty learn about and navigate all of the discussions and issues around AI. The goal is a lot broader than just pure scientific research."[63]

Labs situated in libraries can also contribute to de-siloing data science support services and making them more inclusive of all skill levels and majors. While formal instruction in data science is often targeted to students in STEM fields, many labs explicitly strive to offer programming appropriate for students from a range of disciplinary backgrounds. The AI lab at URI, for example, will offer instruction for all skill levels in "robotics, natural language processing, smart cities, smart homes, the internet of things, and big data."[64]

Highlighted Initiatives

Artificial Intelligence Lab
University of Rhode Island Libraries
https://web.uri.edu/ai/

The URI Libraries' AI Lab provides all students access to tools such as high-performance computing for developing machine learning applications, along with services such as robotics and AI workshops. The lab team includes librarians along with faculty from humanities and STEM disciplines and has enhanced campus learning by serving as a site for a diverse range of URI courses, from the Wearable Internet of Things to Intro to Philosophy.

The Edge
University of Oklahoma Libraries
https://libraries.ou.edu/content/edge

The University of Oklahoma Libraries support the use of VR and visualization throughout their curriculum and faculty research through multiple spaces on campus, including The Edge. The Edge combines makerspace technologies such as 3D printing and microcontrollers with VR workstations and headsets for VR experiences and creation. The Oklahoma Virtual Academic Laboratory (OVAL) project has been used by faculty and students to collaboratively explore immersive virtual environments.

Translational Research & Information Lab (TRAIL)
University of Washington Health Sciences Library
https://hsl.uw.edu/trail/

The TRAIL space at University of Washington Health Sciences Library provides a suite of technologies and services to students, researchers, and physicians so they can incorporate VR, visualization, virtual computing environments, and data analysis into their practice. In 2018, the HSL received an IMLS grant to “design and build a Virtual Reality (VR) and Augmented Reality (AR) program and studio for surgical care teams to simulate cardiac surgery in a library environment”, which also led to the release of Virtual Reality in Academic Health Sciences Libraries: A Primer, which provides detailed guidance on best practices for creating a library VR space, including room requirements, headset and software options, and other specifications.

Key Takeaways

1. **Librarians can leverage existing skills in search and protecting patron privacy to promote new digital literacies.** As librarians teach students to navigate increasingly complex and opaque search interfaces, they have the opportunity to promote algorithmic literacy and help students ask questions about how unseen algorithms shape the results. Librarians have long cared deeply about patron privacy and intellectual freedom, and can leverage this knowledge to develop privacy-as-a-service workshops to educate students on managing and protecting their identity and personal information online.
2. **Libraries will help students evaluate and responsibly create digital content in an environment of malicious Twitter bots and deep fakes.** Libraries must continue to help students develop skills in evaluating sources, which will entail continuing engagement with constantly evolving new media. Even as the technological medium changes, the same questions of authorship, reliability, and who benefits from false or misleading information will apply. Deeper learning opportunities can come about for students who create digital content, whether in a library makerspace or in a librarian-led workshop. Librarians could promote thoughtful engagement with new technologies by leading workshops on creating Twitter bots so students can understand how emerging technologies can be used and misused.

3. **Libraries must engage with campus learning analytics initiatives or risk being left out of the conversation.** Many campuses are engaged in broad initiatives to measure and predict student success using a wide variety of data sources, but libraries are often reluctant to participate because they believe library data isn't relevant or are concerned about student privacy. By having a seat at the learning analytics table, librarians can show administrators how they play a crucial role in teaching, learning, and student success while advocating for privacy-aware student data practices on campus.

Endnotes

1. 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.

2. Thea P. Atwood et al., "Joining Together to Build More: The New England Software Carpentry Library Consortium," *Journal of eScience Librarianship* 8, no. 1 (2019): e1161, https://doi.org/10.7191/jeslib.2019.1161; see also "Foundations for Research Computing: A University-wide Initiative," Columbia University, accessed February 19, 2020, https://rcfoundations.research.columbia.edu.

3. Lisa M. Federer and Douglas J. Joubert, "Providing Library Support for Interactive Scientific and Biomedical Visualizations with Tableau," *Journal of eScience Librarianship* 7, no. 1 (2018): e1120, https://doi.org/10.7191/jeslib.2018.1120.

4. 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/; also see, for example, "The 99 AI Challenge," University of Toronto Libraries, accessed February 19, 2020, https://onesearch.library.utoronto.ca/ai-challenge.

5. Morgan Zalot, “Temple Libraries Launches Interdisciplinary Cultural Analytics Certificate,” *Temple Now*, June 3, 2019, https://news.temple.edu/news/2019-06-03/temple-libraries-launches-interdisciplinary-cultural-analytics-certificate.

6. Jennifer Sparrow, “Digital Fluency: Preparing Students to Create Big, Bold Problems,” *EDUCAUSE Review* 53, no. 2 (March/April 2018): 54, https://er.educause.edu/articles/2018/3/digital-fluency-preparing-students-to-create-big-bold-problems.

7. Lee Rainie and Janna Anderson, “Code-Dependent: Pros and Cons of the Algorithm Age,” Pew Research Center, February 8, 2017, https://www.pewresearch.org/internet/2017/02/08/code-dependent-pros-and-cons-of-the-algorithm-age/.

8. Jason A. Clark, “Building Competencies Around Algorithmic Awareness” (presentation, Code4Lib, Washington, DC, February 15, 2018), Clark’s website, https://www.lib.montana.edu/~jason/talks/algorithmic-awareness-talk-code4lib2018.pdf.

9. Olof Sundin, “Critical Algorithm Literacies: An Emerging Framework” (presentation, ECREA Digital Culture and Communication Section Conference, Brighton, UK, 2017), https://portal.research.lu.se/ws/files/35242070/Extended_abstract_Sundin.pdf.

10. Alison J. Head, Barbara Fister, and Margy MacMillan, *Information Literacy in the Age of Algorithms: Student Experiences with News and Information, and the Need for Change* (Project Information Literacy Research Institute, January 15, 2020), https://projectinfolit.org/pil-public-v1/wp-content/uploads/2020/08/algoreport.pdf.

11. Annemaree Lloyd, “Chasing Frankenstein’s Monster: Information Literacy in the Black Box Society,” *Journal of Documentation* 75, no. 6 (2019): 1475–85, https://doi.org/10.1108/JD-02-2019-0035.

12. Jason Clark, Lisa Janicke Hinchliffe, and Scott Young, " 'RE:Search'—Unpacking the Algorithms That Shape Our UX," Institute of Museum and Library Services, 2017, https://www.imls.gov/sites/default/files/grants/re-72-17-0103-17/proposals/re-72-17-0103-17-full-proposal-documents.pdf.

13. AI for All website, accessed October 6, 2020, https://aiforall.ca/.

14. Head, Fister, and MacMillan, *Information Literacy in the Age of Algorithms*.

15. Marie-Helen Maras and Alex Alexandrou, "Determining Authenticity of Video Evidence in the Age of Artificial Intelligence and in the Wake of Deepfake Videos," *The International Journal of Evidence & Proof* 23, no. 3 (2019): 255–262, https://doi.org/10.1177/1365712718807226.

16. Charlie Harper, "Machine Learning and the Library or: How I Learned to Stop Worrying and Love My Robot Overlords," *Code4Lib Journal*, no. 41 (2018), https://journal.code4lib.org/articles/13671.

17. Jacquelyn Burkell and Chandell Gosse, "Nothing New Here: Emphasizing the Social and Cultural Context of Deepfakes," *First Monday* 24, no. 12 (2019), https://doi.org/10.5210/fm.v24i12.10287.

18. Travis L. Wagner and Ashley Blewer, " 'The Word Real Is No Longer Real': Deepfakes, Gender, and the Challenges of AI-Altered Video," *Open Information Science* 3, no. 1 (2019): 32–46, https://doi.org/10.1515/opis-2019-0003.

19. Harper, "Machine Learning."

20. Clifford A. Lynch, "Reader Privacy: The New Shape of the Threat," *Research Library Issues*, no. 297 (2019): 7–14, https://doi.org/10.29242/rli.297.2.

21. "Digital Competencies: Building 21st Century Skills in a Small Liberal Arts College Setting," Bryn Mawr College, accessed June 10, 2020, https://www.brynmawr.edu/digitalcompetencies.

22. Otter website, accessed October 6, 2020, https://otter.ai/login.

23. Beautiful.ai website, accessed October 6, 2020, https://www.beautiful.ai/.

24. Scholarcy website, accessed October 6, 2020, https://www.scholarcy.com/.

25. Trevor website, accessed October 6, 2020, https://www.trevorai.com/.

26. *7 Things You Should Know About NGDLE*, (EDUCAUSE, December 9, 2015), https://library.educause.edu/resources/2015/12/7-things-you-should-know-about-ngdle.

27. Megan Oakleaf and Malcolm Brown, "The Academic Library and the Promise of NGDLE," *EDUCAUSE Review*, August 14, 2017, https://er.educause.edu/articles/2017/8/the-academic-library-and-the-promise-of-ngdle.

28. Pat Miller and Xiaojing Duan, "NGDLE Learning Analytics: Gaining a 360-Degree View of Learning," *EDUCAUSE Review Transforming Higher Ed Blog*, January 30, 2018, https://er.educause.edu/blogs/2018/1/ngdle-learning-analytics-gaining-a-360-degree-view-of-learning.

29. Meredith Farkas, "Libraries in the Learning Management System," *ALA and ACRL Instruction Section Tips and Trends* (Summer 2015), https://acrl.ala.org/IS/wp-content/uploads/2014/05/summer2015.pdf.

30. Ray Schroeder, "Adaptive Learning to Personalized Learning," *Inside HigherEd Blog*, September 4, 2019, https://www.insidehighered.com/digital-learning/blogs/online-trending-now/adaptive-learning-personalized-learning.

31. Megan Oakleaf, *Library Integration in Institutional Learning Analytics*, (EDUCAUSE, 2018), https://library.educause.edu/resources/2018/11/library-integration-in-institutional-learning-analytics.

32. Oakleaf, *Library Integration*.

33. Sakina Alhadad et al., *The Predictive Learning Analytics Revolution: Leveraging Learning Data for Student Success* (EDUCAUSE Center for Analysis and Research, October 6, 2016), https://library.educause.edu/resources/2015/10/the-predictive-learning-analytics-revolution-leveraging-learning-data-for-student-success.

34. Olga Viberg et al., “The Current Landscape of Learning Analytics in Higher Education,” *Computers in Human Behavior* 89 (December 2018): 98–110, https://doi.org/10.1016/j.chb.2018.07.027.

35. Megan Oakleaf, interview by author, November 12, 2019.

36. Kyle M. L. Jones, “‘Just Because You Can Doesn’t Mean You Should’: Practitioner Perceptions of Learning Analytics Ethics,” *portal: Libraries and the Academy* 19, no. 3 (2019), preprint, posted May 14, 2019, https://papers.ssrn.com/abstract=3372591.

37. Michael R. Perry et al., *SPEC Kit 360: Learning Analytics* (Washington, D.C.: Association of Research Libraries, September 4, 2018), https://publications.arl.org/Learning-Analytics-SPEC-Kit-360/.

38. Oakleaf, *Library Integration*.

39. Oakleaf and Brown, “The Academic Library.”

40. Perry et al., *SPEC Kit 360*.

41. Kyle M.L. Jones and Dorothea Salo, “Learning Analytics and the Academic Library: Professional Ethics Commitments at a Crossroads,” *College & Research Libraries* 79, no. 3 (April 2018): 304–323, https://doi.org/10.5860/crl.79.3.304.

42. Kristin A. Briney, “Data Management Practices in Academic Library Learning Analytics: A Critical Review,” *Journal of Librarianship and Scholarly Communication* 7 (2019): eP2268, https://doi.org/10.7710/2162-3309.2268.

43. Perry et al., *SPEC Kit 360*.

44. Paul Dosal, "Culture, Care, and Predictive Analytics at the University of South Florida," *EDUCAUSE Review*, December 9, 2019, https://er.educause.edu/articles/2019/12/culture-care-and-predictive-analytics-at-the-university-of-south-florida; Phil Richards and Joel Mullan, "How Universities Can Use Learning Analytics to Boost Fair Access and Retention," *Jisc Blog*, April 11, 2017, https://web.archive.org/web/20180123092707/https://www.jisc.ac.uk/blog/how-universities-can-use-learning-analytics-to-boost-fair-access-and-retention-11-apr-2017.

45. Viberg et al., "The Current Landscape."

46. Niall Sclater and Paul Bailey, "Code of Practice for Learning Analytics," Jisc, June 4, 2015, https://www.jisc.ac.uk/guides/code-of-practice-for-learning-analytics.

47. Joan Lippincott and Diane Goldenberg-Hart, *Digital Scholarship Centers: Trends & Good Practice* (Washington, DC: Coalition for Networked Information 2014), https://www.cni.org/wp-content/uploads/2014/11/CNI-Digitial-Schol.-Centers-report-2014.web_.pdf.

48. Lindsay McKenzie, "A New Home for AI: The Library," *Inside Higher Ed*, January 17, 2018, https://www.insidehighered.com/news/2018/01/17/rhode-island-hopes-putting-artificial-intelligence-lab-library-will-expand-ais-reach.

49. Zack Lischer-Katz et al., "Introduction. 3D/VR Creation and Curation: An Emerging Field of Inquiry," 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/.

50. Matt Cook, Julie Griffin, and Robert McDonald, "Developing Library Strategy for 3D and VR Collections," https://events.educause.edu/annual-conference/2018/agenda/developing-library-strategy-for-3d-and-virtual-reality-collections.

51. Szabo, “Collaborative and Lab-Based Approaches.”

52. Mafkereseb Kassahun Bekele and Erik Champion, “A Comparison of Immersive Realities and Interaction Methods: Cultural Learning in Virtual Heritage,” *Frontiers in Robotics and AI* 6 (2019), https://doi.org/10.3389/frobt.2019.00091.

53. Zack Lischer-Katz, Matt Cook, and Kristal Boulden, “Evaluating the Impact of a Virtual Reality Workstation in an Academic Library: Methodology and Preliminary Findings,” *Proceedings of the Association for Information Science and Technology* 55, no. 1 (2018): 300–308, https://doi.org/10.1002/pra2.2018.14505501033.

54. Dian Schaffhauser, “Multi-Campus VR Session Tours Remote Cave Art,” *Campus Technology,* October 9, 2017, https://campustechnology.com/articles/2017/10/09/multi-campus-vr-session-tours-remote-cave-art.aspx.

55. Lischer-Katz, Cook, and Boulden, “Evaluating the Impact.”

56. 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/.

57. Michael T. Moore et al., “Virtual Reality in Academic Health Sciences Libraries: A Primer,” 2018. https://digital.lib.washington.edu/researchworks/handle/1773/42765.

58. Lischer-Katz et al., “3D/VR Creation and Curation.”

59. 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.

60. Isha Salian, “Universities Rush to Add Data Science Majors as Demand Explodes,” *San Francisco Chronicle*, September 5, 2017,

https://www.sfchronicle.com/business/article/Universities-rush-to-add-data-science-majors-as-12170047.php.

61. Muilenburg and Ruttenberg, "New Collaboration for New Education."

62. Ibid.

63. Matt Enis, "University of Rhode Island Opens AI Lab in Library," *Library Journal*, September 26, 2018, https://www.libraryjournal.com?detailStory=180926URIlibraryAIlab.

64. McKenzie, "A New Home for AI."

Chapter 7: Building and Managing Learning and Collaboration Spaces

Landscape Overview

As libraries adopt off-site and compact storage options and grow their collections of digital content, the amount of space required for physical collections in library buildings has dramatically diminished. Spaces that were historically "configured around collections and their use" are being reconceived as flexible, interactive environments that connect users to the people and technologies that support learning, research, and creativity.[1] The impact of emerging technologies on library spaces is evident in the growing prevalence of makerspaces, studios, and labs outfitted with specialized equipment, and a movement towards thoughtfully integrating technology into all aspects of the library visitor experience.

Technologies such as high-resolution LED displays utilized in public spaces can showcase the library's involvement in the full "content lifecycle (creation, access, management, curation) for both e-content and analog content."[2] Tablets and touch-screen kiosks can display real-time information and facilitate room booking, event registration, circulation, and other activities. And as the broader focus of public spaces planning has shifted towards designing user experiences—that is, creating environments that respond and adapt to user needs, provide convenience and satisfaction, and empower users to reach their goals—libraries are considering how technology can productively shape user interactions with the full range of library spaces and services. Thoughtful integration of technology in library spaces has the potential to "reverse the library experience from one in which we expect the user to learn the library—how to navigate it both physically and virtually—to one in which the library 'learns' the user and adapts itself to the user's needs."[3]

Descriptions of libraries as "living labs"[4] and aspirations to transform buildings from "containers" into "living organisms"[5] signal a vision of library *spaces* as adaptable, communicative, experimental collaborators in knowledge creation.

Thoughtful integration of technology in the library building can support a range of user needs, from active collaboration to reflection and focused study. Research libraries "can and should accommodate multiple forms of knowledge-seeking—and better yet, and most critically for the continued vibrancy of the institution, forge connections between the old and new."[6] The following sections explore the ways in which libraries are addressing this challenge in their space planning and programming, specifically addressing the effects of the Internet of Things (IoT), immersive reality, and artificial intelligence on how libraries conceptualize and create the learning and collaboration environments of the future.

Strategic Opportunities

Transform the library building into a living lab

While leading-edge technology is often most conspicuous in makerspaces and labs, some of the most transformative potential lies in the seamless and often invisible integration of emerging technologies into the full library visitor experience. The use of IoT technologies presents a particularly compelling opportunity for library spaces and services to dynamically adapt to user behaviors. The "ubiquitous use and integration of networking, sensing, and tracking technologies in physical environments" could transform the academic library into "a living-learning lab that senses and studies human dynamics, human-computer interactions, and human-building interactions."[7] The data generated by large-scale implementations of sensors and networked devices could become a dynamic data set for the entire community to mine. Libraries have an opportunity to pioneer inclusive, privacy-aware approaches to this integration of sensing technologies in the public sphere.

While the notion of flexibility in library space design has largely come to connote movable furniture, technology enables much broader and more transformative ideas of flexibility.[8] The use of tablets, smart devices, and custom applications can turn static spaces into personalizable environments. The pop-up Alterspace project from the Harvard Library Lab, for example, allows users to select from a series of preset lighting and sound environments designed to enhance specific activities, such as focused learning, meditation, or creativity. Users can tweak the presets to create their optimal study environment.[9] Experimental spaces like the Alterspace inspire visions of entire library buildings outfitted with sensors that continuously monitor temperature, traffic flow, occupancy, light levels, and other metrics; and technologies that give users control over and insight into their environment. The data generated by a large-scale implementation of sensors could allow libraries to better understand users, improve spaces and services, and engage the community in designing ideal environments in real time.

Advances in "computationally-enabled devices and building architectures" are transforming the way people navigate and engage with their university campuses.[10] These technologies are lauded for making the student experience "seamless, simple, and streamlined."[11] Specifically, IoT technologies are being used to provide students with individual access to campus facilities and events, easy payment at campus dining, seamless connection to campus printers or other devices, and just-in-time, location-based information.

From virtual assistants (think Amazon's Alexa device) in each student's dorm room to Bluetooth beacons that record student attendance, college campuses are becoming sites of increased surveillance. While IoT and other smart spaces technologies may make students' lives more convenient and productive, they permit (and rely on) data-intensive monitoring and evaluation of students, generating significant concerns about privacy, bias, and the ethics of continuous data collection.

Data collected from IoT technologies around campus—such as an individual's visits to certain academic buildings like the library, their class attendance, or their participation in campus events—can be aggregated with other metrics—like grades and test scores—and demographic information to measure (or even predict) a student's success.[12] While often well-intentioned, this approach to student monitoring has alarmed privacy advocates and generated serious concerns about how the collection and use of student data could harm students, especially those from already marginalized and underrepresented populations.

Continuous surveillance and the use of black-box algorithms to analyze data introduces opportunities for bias and misuse. Much has been written on the potential consequences of over-reliance on predictive models and AI in making decisions that could impact an individual's future. People of color and other marginalized groups are especially at risk of losing out in this environment. A recent study published in *Nature* found "rampant racism in decision-making software" widely used in hospitals, leading to poorer health care outcomes for people of color.[13]

There are also risks that user data could be compromised by human error or malicious actors, potentially exposing identifying or sensitive information, or providing third parties with access to a treasure trove of mineable data. Beacons technologies, for example, do not collect user data and "typically do not connect to the Internet without an additional layer of software that can interpret their signals."[14] However, those additional software layers can be used to collect and transmit information about a user's location, activities, or identity. Libraries have a particularly vested interest in ensuring user privacy, given their commitment to intellectual freedom. The use of sensors, even those that do not transmit data in compromising ways, could create an environment where users feel surveilled and therefore inhibited, potentially affecting "how they view the library and what information they seek out from library resources."[15]

There is little evidence that most research libraries have widely adopted IoT technologies in their buildings. Where they have been implemented, they are generally focused on making the user experience more convenient and on making spaces comfortable for both users and collections. At Concordia University's Webster Library, for example, librarians developed a prototype system to measure and display noise levels in various areas of the library, allowing users to "choose the area with the right amount of noise for their purposes."[16] Although the prototype had not been deployed at scale as of the publication date, it is an example of an IoT-based technology that does not rely on invasive surveillance. The system does not record or process sound; it merely measures decibel levels. It makes no attempt to track or identify individual users or their behaviors. At the root, the system enhances, rather than compromises, a user's autonomy within the library space by allowing them to make an informed decision about appropriate study environments depending on their mood or intended activity.

Highlighted Initiatives

Alterspace
Harvard metaLAB and Library Innovation Lab
https://alterspace.github.io/

Harvard's Library Innovation Lab, embedded in the Law School Library, develops experimental projects that engage with the future of libraries. Their Alterspace project allows library users to control various aspects of their physical environment, including "light, color, sound and space" to give them the ability to optimize the space for specific activities, such as study, meditation, or creativity. Alterspace is an open-source project with code released on GitHub that can be reused or modified by other libraries.

Enhance the user experience in library spaces

Poor wayfinding in libraries has long preoccupied librarians, who strive to give visitors better tools to navigate warren-like stacks and intimidating service points. Enter Hugh, the robot librarian at Aberystwyth University, who can “search the catalog, identify a book’s shelf location, and lead a patron to it.”[17] Hugh is touted as a way to make the visitor experience more pleasant while freeing librarians to focus on more complex visitor needs as Hugh handles routine interactions.

While robot librarians remain a novelty, libraries are experimenting with a range of other emerging technologies to support wayfinding and just-in-time visitor services. Beacon technologies, which communicate with mobile devices via Bluetooth low-energy proximity sensing, hold particular promise. The move to 5G networks will accelerate the use of networked devices as data transmission speeds increase. One of the earliest proposed uses of beacons was to support wayfinding within buildings, particularly for those individuals with sight or other impairments that prevent them from benefiting from visual signage and navigation aids.[18]

Beacons can be used in conjunction with specially designed apps to create interactive maps that guide users through the library building with turn-by-turn directions and present students with just-in-time, location-aware information.[19] This could include information that makes visiting the library building more convenient (for example, alerts that direct users to unoccupied seating or during busy periods like the Waitz app deployed at UCSD and UC Santa Barbara[20]); more pleasant (for example, push notifications that remind users when they are entering a designated quiet area); more welcoming (for example, invitations to join library workshops or events as visitors enter the building); or more productive (for example, location-based recommendations systems that suggest nearby books of interest).[21]

A number of libraries have experimented with beacon technology to create self-guided library tours and navigational aids;[22] build augmented reality (AR) exhibits;[23] provide location-specific mobile alerts;[24] and help users locate materials in the library stacks.[25] An app developed at the University of Illinois at Urbana-Champaign, for example, can direct a user to a book in the stacks while providing real-time recommendations based on the user's location and the popularity of nearby items using circulation data.[26] Wearable devices could even provide real-time translation to help users identify materials in their non-native language in the stacks.[27] IoT technologies can also be used to give students access to restricted or reservable spaces (such as bookable study rooms)[28] or physical materials (such as smart lockers that hold course reserves for students in a given class).

Emerging technologies can also be used to enhance a sense of community within library spaces. One recent project uses beacons to create virtual micro-communities or zones within a large, flexible makerspace.[29] Several researchers have proposed hypothetical apps that use beacons to help users connect with one another around shared interests or goals.[30] An article on using beacon technology in study spaces asks readers to imagine "walking into a library commons and receiving recommendations on your phone about locations to sit based on the similarity of the research others are conducting nearby."[31] A similar project proposes an app that would "promote the portfolios, research work, etc. of people in the immediate vicinity by temporarily 'attaching' links to beacons," helping to "build a sense of collegiality as a diverse community of learners, researchers and practitioners."[32]

It is easy to see beacon technologies as simultaneously convenient and intrusive. While some users may appreciate location-based assistance and information, others may find it creepy or bothersome. Frequent alerts may be counterproductive in an environment designed to encourage focused study. Clear opt-in policies (and/or use of beacons exclusively in the context of a voluntarily downloaded app) are therefore advisable. General library privacy policies will require revision and expansion to address the many new ways in which user data may be collected and used.

Highlighted Initiatives

Waitz Find A Seat app
UC San Diego Library
https://libraries.ucsd.edu/visit/study-spaces/index.html

UC San Diego Library has created a study spaces app that shows students real-time space availability based on anonymized WiFi and Bluetooth traffic, in partnership with a startup, Waitz. Waitz sensors are installed throughout the library, and collect anonymized web traffic data to display the busyness of various study spaces to students.

Spaces planning and assessment

While many libraries have found foot traffic to their buildings remains as robust as ever, especially after space renovations that establish new learning and information commons,[33] they face increasing pressure to demonstrate the specific value and impact of their spaces. New tools can help libraries gather and interpret metrics well beyond gate counts and circulation statistics. Smart devices, machine learning, and other technologies have the potential to give libraries insight into library usage patterns that can help them plan for future space and service improvements.

Over a dozen articles in the library literature describe IoT-based approaches to spaces assessment.[34] Data from beacons and sensors, thermal imaging cameras, and other networked devices can provide real-time data about traffic flow (for example, how many visitors browse the stacks versus head straight for the learning commons) and space usage (for example, the number of occupied seats in various zones of the library, busy and slow times).

The Measure the Future Project developed a toolkit for using webcams and a computer vision algorithm to assess space usage.[35] The webcam identifies and tracks visitors to see where they congregate and how they move through a space, generating usage heat maps that librarians can use to understand what kinds of spaces are popular, address overcrowding, or learn about user behavior. The use of thermal cameras mitigates privacy concerns, making it significantly more difficult to identify individual users. Further, the cameras will not record activity when fewer than three individuals are in the frame.

Continuous data collection (think hundreds of sensors running 24 hours a day) will rapidly overwhelm traditional methods of data analysis. Libraries will need machine learning tools to sift through massive troves of sensor data to identify patterns and actionable insights. To fully leverage the data they collect, librarians will need data dashboards that support real-time monitoring and that aggregate data from a range of sources. At the University of Rochester, librarian Lauren Di Monte and data scientist Nilesh Patil are using machine learning to study traffic patterns in the library building. The team set out to determine how many people who entered the library had come to use library spaces and services and how many were just passing through to access other buildings or areas of campus. The team developed a recurrent neural network model and trained an algorithm on data gathered from bidirectional gate counters. The model was then used to predict traffic based on previous patterns.[36]

While these new assessment tools offer exciting opportunities, they also come with limitations and risks. Few libraries have implemented networked monitoring devices at scale because equipping an entire building with sufficient beacons and other sensors to generate useful data remains expensive, and thoughtfully outfitting an entire library building to collect meaningful data takes intensive planning. As data analysts constantly caution, poor data collection methods lead to misleading or inaccurate conclusions.

Finally, data generated by sensors and other passive collection mechanisms will require complementary qualitative research to provide context. For example, using sensors to measure sound volume in a library space ”does not reveal what people actually hear, nor how people value or use sound.”[37] Emerging technologies represent an exciting addition to, rather than a replacement for, existing methods of space planning and evaluation.

Highlighted Initiatives

Measure the Future Project
http://jasongriffey.net/mtf/

The Measure the Future Project, funded by the Institute for Museum and Library Services, has created an open-source hardware and software toolkit that libraries can use to monitor and assess space usage. The project solves for patron privacy by using thermal cameras that make it difficult to identify individual users. The first phase of the project launched with pilots in a mix of public and academic libraries, including SUNY Potsdam and the New York Public Library.

Smart Commons project
Virginia Tech University Libraries
https://github.com/VTUL/smart-commons

Virginia Tech University Libraries' Smart Commons project has taken a different approach to collecting space usage data without compromising patron privacy. WiFi-connected motion sensors are attached to the bottom of individual chairs in the library Learning Commons, allowing for granular data collection on the number of seats occupied at any given time. The hardware plans and source code have been released on GitHub so other libraries can recreate the project.

Key Takeaways

1. **Libraries are thinking beyond the makerspace in considering emerging technologies in their spaces.** While many libraries have now built technology-rich makerspaces, VR/AR spaces, and digital media labs, transforming libraries into smart buildings can also mean infusing technology into the entire building and user experience, from sensors that anonymously monitor space usage to networked devices that allow users to customize their own study environments. Rather than drawing an artificial distinction between "hi-tech" and "traditional" library spaces, librarians are considering how emerging technologies can inform all aspects of space planning and design.

2. **Libraries can leverage their historical commitment to patron privacy in designing user experiences that incorporate sensing technologies.** One notable commonality in the highlighted initiatives included in this section is they all incorporate privacy-aware approaches to collecting data about spaces, whether through anonymized WiFi data, thermal cameras that don't identify faces, or use of motion sensors. Although no longer an emerging technology, infrared beam door counters became ubiquitous in libraries over the past 30 years because they provided a convenient and low-cost way for libraries to track visitors without collecting identifiable user data. As the emerging projects described in this section become more mature and easier to implement, we can similarly expect widespread adoption by libraries.
3. **Develop library apps and tools with sustainability in mind.** Readers will note that many of the projects described in academic literature and featured in this section are no longer active. While some of this can be attributed to the nature of pilot projects that were not necessarily intended to continue, other projects have ended due to a staff member departing or grant funding running out. To mitigate against this tendency, libraries should take the same approach to apps and sensing projects that they do with digital content, and plan for sustainability. On a positive note, many of the projects included in this section have released their code on GitHub, so even if a project becomes inactive, another institution would be able to pick the project up later.
4. **Sensing technologies can empower users by giving them agency in library spaces.** Sensors, beacons, and microcontrollers can improve the user's experience of library spaces by helping them find the least crowded or noisy places to study in real time, be guided to finding books in the stacks, and give them control over their physical study environment. Emerging technologies "have the capacity to reverse the library experience from one in which we expect the user to learn the library—how to navigate it both physically and virtually—to one in which the library "learns" the user and adapts itself to the user's needs."[38]

Endnotes

1. Lorcan Dempsey and Constance Malpas, “Academic Library Futures in a Diversified University System,” in *Higher Education in the Era of the Fourth Industrial Revolution*, ed. Nancy W. Gleason (Singapore: Palgrave Macmillan, 2018), 65–89, https://doi.org/10.1007/978-981-13-0194-0_4.

2. Joan Lippincott, “The Link to Content in 21st-Century Libraries,” *EDUCAUSE Review* 53, no. 1 (2018): 64-65, https://er.educause.edu/articles/2018/1/the-link-to-content-in-21st-century-libraries.

3. Steven J. Bell, “Staying True to the Core: Designing the Future Academic Library Experience,” *portal: Libraries and the Academy* 14, no. 3 (2014): 369–382, http://dx.doi.org/10.1353/pla.2014.0021.

4. Yi Shen, “Intelligent Infrastructure, Ubiquitous Mobility, and Smart Libraries–Innovate for the Future,” *Data Science Journal* 18, no. 11 (March 21, 2019): 11, https://doi.org/10.5334/dsj-2019-011.

5. Jonathan Bradley, Patrick Tomlin, and Brian Mathews, “Building Intelligent Infrastructures: Steps toward Designing IoT-Enabled Library Facilities,” *Library Technology Reports* 54, no. 1 (2018): 23–27, https://doi.org/10.5860/ltr.54n1.

6. Dan Cohen, “Libraries Contain Multitudes,” *Humane Ingenuity*, October 8, 2019, https://buttondown.email/dancohen/archive/humane-ingenuity-5-libraries-contain-multitudes/.

7. Shen, “Intelligent Infrastructure.”

8. Cohen, “Libraries Contain Multitudes.”

9. Reena Karasin, “From Public to Personalized: Alterspace Turns Libraries into Rooms of Requirement,” *Scout Cambridge*, July 19, 2019, https://scoutcambridge.com/from-public-to-personalized-alterspace-turns-libraries-into-rooms-of-requirement/.

10. Shen, “Intelligent Infrastructure.”

11. Itai Asseo et al., “The Internet of Things: Riding the Wave in Higher Education,” *EDUCAUSE Review* 51, no. 4 (2016): 11-31, https://er.educause.edu/articles/2016/6/the-internet-of-things-riding-the-wave-in-higher-education.

12. Asseo et al., “Internet of Things.”

13. Heidi Ledford, “Millions of Black People Affected by Racial Bias in Health-Care Algorithms,” *Nature*, no. 574 (October 31, 2019): 608-609, https://doi.org/10.1038/d41586-019-03228-6.

14. Valeda Dent et al., “Wayfinding Serendipity: The BKFNDr Mobile App,” *Code4Lib Journal*, no. 42 (November 8, 2018), https://journal.code4lib.org/articles/13811.

15. Matthew B. Hoy, “Smart Buildings: An Introduction to the Library of the Future,” *Medical Reference Services Quarterly* 35, no. 3 (2016): 326–31, https://doi.org/10.1080/02763869.2016.1189787.

16. Janice Yu Chen Kung, “Raspberry Pi and Arduino Prototype: Measuring and Displaying Noise Levels to Enhance User Experience in an Academic Library,” *Library Technology Reports* 54, no. 1 (2018): 18–22, https://doi.org/10.5860/ltr.54n1.

17. Steven Bell, “Promise and Peril of AI for Academic Librarians,” *Library Journal*, April 14, 2016, https://www.libraryjournal.com?detailStory=promise-and-peril-of-ai-for-academic-librarians-from-the-bell-tower.

18. Ian Glover and Kieran McDonald, “Digital Places: Location-Based Digital Practices in Higher Education Using Bluetooth Beacons,” in *Proceedings of EdMedia: World Conference on Educational Media and Technology*, ed. Theo Bastiaens (Association for the Advancement of Computing in Education (AACE), 2018): 950–959, http://www.learntechlib.org/p/184298/.

19. Glover & McDonald, “Digital Places.”

20. Waitz website, accessed October 30, 2020, https://waitz.io/.

21. Jim Hahn, “Mobile Augmented Reality Applications for Library Services,” *New Library World* 113, no. 9/10 (September 29, 2012): 429–38, https://doi.org/10.1108/03074801211273902.

22. Jonathan Bradley et al., “Creation of a Library Tour Application for Mobile Equipment Using IBeacon Technology,” *Code4Lib Journal*, no. 32 (April 25, 2016), https://journal.code4lib.org/articles/11338.

23. Brandon Patterson, “Talking Portraits in the Library: Building Interactive Exhibits with an Augmented Reality App,” *Code4Lib Journal*, no. 46 (November 5, 2019), https://journal.code4lib.org/articles/14838.

24. Sidney Eng, “Connection, Not Collection: Using iBeacons to Engage Library Users,” *Information Today* 35, no. 10 (December 2015), http://www.infotoday.com/cilmag/dec15/Eng--Using-iBeacons-to-Engage-Library-Users.shtml; Somaly Kim Wu, Marc Bess, and Bob R. Price, “Digitizing Library Outreach: Leveraging Bluetooth Beacons and Mobile Applications to Expand Library Outreach,” in *Digitizing the Modern Library and the Transition From Print to Electronic*, ed. Raj Kumar Bhardwaj (Hershey, PA: IGI Global, 2018), 193–203, https://doi.org/10.4018/978-1-5225-2119-8.ch008.

25. Dent et al., “Wayfinding Serendipity.”

26. Jim Hahn, “The Internet of Things: Mobile Technology and Location Services in Libraries” *Library Technology Reports* 53, no. 1 (2017), https://doi.org/10.5860/ltr.53n1.

27. Ayyoub Ajmi and Michael J. Robak, “Wearable Technologies in Academic Libraries: Fact, Fiction and the Future,” in *Mobile Technology and Academic Libraries: Innovative Services for Research and Learning*, ed. Robin Canuel and Chad Crichton (Chicago, IL: Association of College & Research Libraries, 2017), https://mospace.umsystem.edu/xmlui/handle/10355/60599.

28. Hubert C.Y. Chan and Linus Chan, “Smart Library and Smart Campus,” *Journal of Service Science and Management* 11, no. 6 (November 28, 2018): 543–64, https://doi.org/10.4236/jssm.2018.116037.

29. Glover & McDonald, “Digital Places.”

30. Bradley et al., “Building Intelligent Infrastructures”; Glover & McDonald, “Digital Places”; Hahn, “Internet of Things.”

31. Bradley et al., “Building Intelligent Infrastructures.”

32. Glover & McDonald, “Digital Places.”

33. DeeAnn Allison et al., “Academic Library as Learning Space and as Collection: A Learning Commons’ Effects on Collections and Related Resources and Services,” *Journal of Academic Librarianship*, no. 45 (2019): 305-314, https://digitalcommons.unl.edu/libraryscience/384.

34. Jason Griffey, “How to Measure the Future,” *Library Technology Reports* 54, no. 1 (2018): 11–17, https://doi.org/10.5860/ltr.54n1.

35. Griffey, “How to Measure.”

36. Lauren Di Monte and Nilesh Patil, “Deep Learning for Libraries” (presentation, Code4Lib, Washington, DC, February 14, 2018), https://2018.code4lib.org/talks/deep-learning-for-libraries.

37. Andrew M. Cox, “Learning Bodies: Sensory Experience in the Information Commons,” *Library & Information Science Research* 41, no. 1 (January 2019): 58–66, https://doi.org/10.1016/j.lisr.2019.02.002.

38. Steven J. Bell, “Staying True to the Core: Designing the Future Academic Library Experience,” *portal: Libraries and the Academy* 14, no. 3 (2014): 369–382, https://doi.org/10.1353/pla.2014.0021.

Conclusion

The emerging technologies explored in this report have prompted libraries to adapt their historical roles as trusted stewards, educators, and curators to suit an academic environment and a society driven by digital data, marked by distributed collaboration, and contending with the challenges of misinformation, white supremacy culture, and a global pandemic.

Research libraries' historical role as trusted stewards of collections takes on new urgency as they ensure the provenance, authenticity, and long-term preservation of increasingly complex digital assets in a societal context where digital misinformation has become ubiquitous. Libraries' long-standing emphasis on protecting user privacy has led them to become advocates for the judicious and ethical use of campus learning analytics. Traditional models of information access are being reimagined: research libraries are building and maintaining computationally ready digital collections and building borderless collections that incorporate open, owned, and licensed content. And in the tradition of building information literacy, research libraries are fostering critical engagement with new forms of digital information and misinformation and enabling their stakeholders to produce new forms of scholarly and creative work.

The research and interviews for this report were primarily conducted in the spring and fall of 2019, before the COVID-19 pandemic abruptly reshaped the higher education landscape. The pandemic has forced rapid innovation and accelerated existing trends in libraries in online/blended learning, facilitating easy access to e-content and data, and helping students build new information fluencies to combat the proliferation of disinformation. After nearly a year of learning fully or partly online for faculty and students at residential colleges and universities, there will be no return to the pre-pandemic status quo for libraries. The "new normal" for library users will be online or blended-first, and users will expect collections and services to operate seamlessly in these hybrid channels even as the library returns to operating a physical space.

The pandemic has highlighted the urgency of providing timely, barrier-free access to information; enabling distributed research and learning; and advocating for digital privacy. The research library is well-positioned to meet the challenges of this increasingly open, distributed, and digital data-centric academy, combining library workers' expertise in education, curation, and preservation with their position as a trusted institution.

Research libraries can bring values-based decision-making to bear as they find the right balance in their approach to adopting and experimenting with emerging technologies—the balance between agility and sustainability, convenience and privacy, transformation and persistence. As emerging technologies such as machine learning, immersive reality, and the Internet of Things change the ways researchers and students engage with information, libraries have opportunities to advance their contributions to the research and learning enterprise. As adopters of these technologies, research libraries can make information more discoverable, reusable, and durable. As educators, library workers can help their communities critically and productively engage with technology in the service of research and learning. By thoughtfully adopting and responding to emerging technologies, research libraries assert their continued and multifaceted value as campus hubs for research and learning.

Glossary

artificial intelligence (AI). The theory or development of computers or other machines to perform tasks that exhibit intelligent behavior, such as visual perception, speech recognition, decision-making, and language translation[1]

big data. Data characterized by huge volume, rapid generation, diversity, and scope, typically to the extent that its manipulation and management present significant logistical challenges; (also) the branch of computing involving such data

blockchain or distributed ledger technology. A type of database of replicated, shared, and synchronized digital data geographically spread across multiple sites, countries, or institutions. Records are stored in blocks, or one after the other in a continuous ledger[2]

computer vision. An umbrella term that encompasses attempts to computationally replicate the human visual system and automate visual tasks such as pattern and known-entity recognition[3]

containerization. "A standard unit of software that packages up code and all its dependencies so the application runs quickly and reliably from one computing environment to another"[4]

data mining. The process or practice of examining large collections of data in order to generate new information, typically using specialized computer software[5]

data science. An inter- and cross-disciplinary field concerned with concepts and topics in statistics, data mining, machine learning, and broad data analytics[6]

deep fakes. The product of merging or combining images, audio, or video, using artificial intelligence or machine learning techniques, to create a fake product that appears authentic[7]

high-performance computing (HPC). Processor-intensive applications that rely on computational clusters and federations of scattered clusters[8]

immersive reality. A collective term for augmented, virtual, and mixed reality (AR, VR, and MxR), which create the perception of physical interaction with virtual environments[9]

Internet of Things (IoT). The extension of the internet into the physical world embedding computing devices on physical items, giving them network connectivity and allowing them to send and receive data[10]

learning analytics. An educational application of data collection and analysis of online activities aimed at learner profiling to discover, diagnose, and predict learner behavior and to design interventions that improve student outcomes[11]

learning management system (LMS). An integrated set of online applications providing access to digital course materials, discussions, grades, and other features in support of education, particularly in colleges and universities[12]

machine learning (ML). A computing system that learns from experience by reviewing large sets of information, creating models based on this data, making predictions, and refining its algorithm on the basis of newly acquired data[13]

natural language processing (NLP). The combination of artificial intelligence with linguistics to process and analyze language-based data[14]

next generation digital learning environment (NGDLE). A learning environment consisting of learning tools and components that adhere to common standards, intended to directly support learning. The NGDLE addresses five dimensions: interoperability and integration; personalization; analytics, advising, and learning assessment; collaboration; and accessibility and universal design.[15]

predictive analytics. The collection and analysis of data from online activities used to predict future behavior or outcomes[16]

reproducibility. The ability to replicate or repeat methods and conditions to yield consistent results[17]

Jocelyn Cozzo, born-digital, contributed significant research to this glossary.

Endnotes

1. Adapted from: *OED Online*, s.v. "artificial intelligence (n.)," accessed December 2019, and Joan M. Reitz, *ABC-CLIO Online Dictionary for Library and Information Science (ODLIS)*, s.v. "artificial intelligence (AI)," accessed January 31, 2020, https://products.abc-clio.com/ODLIS/odlis_a.aspx.

2. Adapted from: UK Government Office for Science, *Distributed Ledger Technology: Beyond Block Chain* (January 2016), https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/492972/gs-16-1-distributed-ledger-technology.pdf.

3. Adapted from: T.S. Huang, Computer Vision: Evolution and Promise," in *1996 CERN School of Computing Proceedings*, ed. C.E. Vandoni (Geneva: CERN—European Organization for Nuclear Research, 1996), 21–25, http://dx.doi.org/10.5170/CERN-1996-008.21.

4. "What Is a Container?" Docker, accessed January 31, 2020, https://www.docker.com/resources/what-container.

5. *OED Online*, s.v. "data mining (n.)," accessed December 2019.

6. Adapted from: Longbing Cao, "Data Science: Challenges and Directions," *Communications of the ACM* 60, no. 8 (2017): 59–68,

https://doi.org/10.1145/3015456.

7. Adapted from: Bobby Chesney and Danielle Citron, "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security," *California Law Review* 107, no. 6 (2019), https://doi.org/10.15779/Z38RV0D15J and Marie-Helen Maras and Alex Alexandrou, "Determining Authenticity of Video Evidence in the Age of Artificial Intelligence and in the Wake of Deepfake Videos," *International Journal of Evidence & Proof* 23, no. 3 (2019): 255–62, https://doi.org/10.1177/1365712718807226.

8. Adapted from: Carlos Arangol, Rémy Dernat, and John Sanabrial, "Performance Evaluation of Container-Based Virtualization for High Performance Computing Environments," *Revista UIS Ingenierías* 18, no. 4 (2019): 31–42, https://doi.org/10.18273/revuin.v18n4-2019003.

9. Adapted from: Mafkereseb Kassahun Bekele and Erik Champion, "A Comparison of Immersive Realities and Interaction Methods: Cultural Learning in Virtual Heritage," *Frontiers in Robotics and AI* 6, article 91 (2019), https://doi.org/10.3389/frobt.2019.00091.

10. Adapted from: Friedemann Mattern and Christian Flörkemeier, "From the Internet of Computers to the Internet of Things," *Informatik-Spektrum* 33 (2010): 107–121, https://doi.org/10.1007/s00287-010-0417-7.

11. Adapted from: Megan Oakleaf, *Library Integration in Institutional Learning Analytics* (Syracuse: Syracuse University Press, 2018), https://library.educause.edu/resources/2018/11/library-integration-in-institutional-learning-analytics.

12. Adapted from: Reitz, *ABC-CLIO Online Dictionary for Library and Information Science*, s.v. "learning management system (LMS)."

13. Adapted from: Thomas Finley, "The Democratization of Artificial Intelligence: One Library's Approach," *Information Technology and Libraries* 38, no. 1 (2019): 8–13, https://doi.org/10.6017/ital.

v38i1.10974 and *OED Online*, s.v. "machine learning (n.)," accessed December 2019.

14. Adapted from: Bradley Beth and Noureddine Elouazizi, *7 Things You Should Know about Natural Language Processing*, (EDUCAUSE, March 6, 2018), https://library.educause.edu/resources/2018/3/7-things-you-should-know-about-natural-language-processing.

15. Adapted from: *7 Things You Should Know about NGDLE*, (EDUCAUSE, December 9, 2015), https://library.educause.edu/resources/2015/12/7-things-you-should-know-about-ngdle.

16. Adapted from: Oakleaf, *Library Integration in Institutional Learning Analytics*.

17. *OED Online*, s.v. "reproducibility (n.)," accessed December 2019.

ASSOCIATION
OF RESEARCH
LIBRARIES

Association of Research Libraries
21 Dupont Circle, NW
Suite 800
Washington, DC 20036
T 202.296.2296
F 202.872.0884

ARL.org
pubs@arl.org