Temple University

Data Speculations: A National Forum on Library Digital Stewardship for Copyrighted Contemporary Culture

Temple University and Texas A&M University request a 2-year $124,391 IMLS National Leadership Grant for Libraries to host a virtual National Forum convening specialists in digital humanities research, library data services, and copyright and information policy to collectively advance libraries supporting computational research on copyrighted contemporary culture materials, such as library-owned science fiction collections. By convening a dedicated community of practice at a National Forum, hosting public talks, and producing findings and guidelines for wider dissemination, *Data Speculations* will be foundational to the collective work of removing access barriers for sharing copyrighted collections, serving libraries' core mission to make information, including copyrighted data, available to researchers and teachers to advance learning and knowledge. *Data Speculations* supports IMLS goals to *Advance collections stewardship and access*, especially the National Leadership Grant Programs' goals to *Broaden Access* (Objectives 3.2 and 3.3) and *Strengthen Collaboration for the Benefit of Communities* (Objective 5.1).

At a time when multinational publishing houses and database vendors increasingly speculate (quite literally) on the landscape of digitized contemporary culture, charging additional costs to libraries for licensed access to their collections of newspapers and magazines as data, *Data Speculations* seeks to establish an alternate vision of a future where libraries can steward large collections of copyrighted cultural data to directly support their user communities. This National Forum will thereby offer an opportunity for a diverse field of specialists to offer their perspectives to ensure the guidelines we develop can be useful for a wider range of use cases and contexts, with the broader goal of empowering diverse communities to take ownership of their research data and, on their own terms, open it up to new modes of digital analysis and curation. By analyzing needs, developing a proof of concept legal framework, and convening relevant stakeholders and experts, *Data Speculations* will inform the next-stage development and implementation of partnerships, protocols, tools, and best practice guidance to unlock broader collection access.

Project Justification

Humanities research has been revolutionized by widespread digitization and the growth of digital collections in the 21st century. The Collections as Data movement in libraries and museums has further expanded the ways cultural heritage materials can be made available for research and teaching with digital humanities methods.[1] As a part of this movement, digital materials held in restrictive library portals and previously available to view only as individual items are being made accessible in new ways as datasets.[2] Increasingly, libraries and stewards enable teachers, scholars, and journalists to explore cultural heritage collections in computational form, allowing users to mine, map, analyze, and visualize these materials at scale. Collectively referenced as "Text and Data Mining," or TDM, this field of computational analysis encompasses a wide range of technologies and methodologies aimed at exploring cultural media through data analysis and visualization.

However, both the real and perceived legal barriers to working with copyrighted data at scale has prevented libraries and research centers from growing Collections as Data for modern and contemporary culture, restricting computational access to the literary and artistic materials that represent the most diverse period of cultural production in human history, the twentieth and twenty-first centuries.[3] The vast majority of multicultural writing and media created since 1928 is presumed to be under the lock and key of copyright, preventing scholars of contemporary

[1] See the Collections as Data website for history and information about the movement: https://collectionsasdata.github.io/
[2] See Sarah Ames, "Transparency, provenance and collections as data," *The Liber Quarterly* 31 (2021).
[3] Rachael Samberg and Timothy Vollmer, eds., *Building Legal Literacies for Text Data Mining* (University of California, Berkeley) https://doi.org/10.48451/S1159P; Scott Althaus et al., "Building Legal Literacies for Text Data Mining: Institute White Paper" https://berkeley.pressbooks.pub/buildinglltdm/back-matter/white-paper/

culture from collaborating to do digital humanities research with these materials at scale.[4] While it may be within the bounds of "fair use" to assemble collections for algorithmic study, the onus of creating these digital collections cannot reasonably be borne by individual researchers operating in silos. Libraries, on the other hand, are better-positioned to take on the labor of digitization and curation. Even as libraries grow their investments in digitizing materials and providing computational access to contemporary culture collections, however, concerns about copyright and confusion about legal risk thwart progress both within and across cultural institutions, significantly limiting community efforts to steward computational access.

In this context, *Data Speculations* seeks to raise awareness of and mitigate copyright concerns, as well as misunderstandings, that currently prevent libraries from aggregating and sharing copyrighted contemporary culture collections as data in support of digital humanities research. This National Forum will therefore focus on the following question: How can collection stewards and scholars learn to take control of copyrighted datasets for research purposes, building and sharing data between institutions, and confidently stewarding and provisioning controlled access to restricted research data for the study of published popular culture?

To explore this question, *Data Speculations* will take as a primary case study ongoing efforts at Temple University, in collaboration with members of the the SciFi Collection Libraries Consortium (SFCLC), to build and curate a comprehensive dataset of contemporary "speculative fiction" (focused on twentieth-century "science fiction," but inclusive of a wider range of imaginative literature produced since World War II).[5] The full scope of copyrighted materials that would be of interest to scholars of contemporary culture is difficult to encompass and, accordingly, to analyze for legal risk and opportunity. For this reason, this National Forum grounds our discussion in the specific use case of the speculative fiction collections represented across the dozens of libraries in the SFCLC (such as at Temple and Texas A&M Libraries).[6] These collections contain published, copyrighted works that span distinctive periods of the genre, including not just the "Golden Age" (1934-1963) and the "New Wave" (1964-1983), but also more recent multicultural subgenres central to contemporary popular culture, such as feminist and queer science fiction, cyberpunk and climate fiction, Indigenous SF and Afrofuturism, and many more subgenres exploring the margins of social identity, the limits of the physical universe, and the unimagined potentials of human existence.

Popular culture materials, like science fiction, continue to be produced at an exponential scale, becoming increasingly vital to major works in the scholarly fields of digital cultural study, such as Ted Underwood's *Distant Horizons* (2019).[7] Due to the compounding forces blocking projects to build and share datasets of contemporary culture, however, libraries have tended to focus on provisioning access to public domain materials. As a result, students and scholars of contemporary culture remain dependent on small datasets hardly representative of the enormous number of works published in the twentieth- and twenty-first centuries. Without the capacity to share collections across institutions to provision the specialized datasets required to answer the diverse questions central to academic research, scholars' findings remain narrow in scope, as they struggle to figure out what a comprehensive, or

---

[4] Works published prior to 1964 may be in the public domain due to failure to renew copyright, but establishing this fact requires painstaking item-by-item research. See, e.g., Claire Woodcock, "Librarians Are Finding Thousands Of Books No Longer Protected By Copyright Law," Vice, Feb. 9, 2023, http://bit.ly/3l5SVS4.

[5] For more information on Temple University Libraries' science fiction digitization project, visit the SF Project's Omeka site (https://lcdssgeo.com/omeka-s/s/scifi/page/digitizing-science-fiction). The Scholars Studio also offers information on dataset curation on their website (https://library.temple.edu/webpages/datasets) and related code is available on Github (https://github.com/SF-Nexus). A static website is currently available at https://sfnexus.io. Also see Wermer-Colan's essay, "The New Wave of Digital Collections: Speculating on the Future of Library Curation," *Transactions* of the American Philosophical Society, Vol. 110, No. 3, *The Past, Present, and Future of Libraries* (2022), pp. 211-241: https://www.jstor.org/stable/45420508

[6] For more information about the SFCLC, visit their wiki at http://bit.ly/3l2jKXr. See also Temple University Libraries' Special Collection Research Center houses the Paskow Science Fiction Collection (http://bit.ly/3JCuSDY) and Texas A&M Libraries' Science Fiction and Fantasy Research Collection (https://cushing.library.tamu.edu/collecting/scifi.html).

[7] Ted Underwood, "The Life Cycle of Genres," *Distant Horizons: Digital Evidence and Literary Change* (UChicago, 2019).

representative, set of data might look like in the first place. A growing body of research has demonstrated just how insidious the consequences of these impediments can be, finding that the race, gender, and other biases evident in the openly available texts within the public domain have contributed to and exacerbated biases in the machine-learning models developed for artificial intelligence tools.[8]

To make matters more pressing, many of these works of mass-market culture are often only preserved in the form of fragile paperbacks; they are also likely to be "orphaned" works, owned by decommissioned publishers and out-of-reach author estates. In an environment where "permission culture" still holds some sway, undue obstacles stand in the way of projects looking to digitize and share materials at scale. For these reasons and more, speculative fiction, and Temple's project to build a comprehensive corpus of the genre, offers a fitting foundation for this National Forum. Extrapolating from this case study, furthermore, *Data Speculations* aspires to speculate on the wider range of obstacles and opportunities for libraries and librarians seeking to collect and provide research access to published, copyrighted contemporary culture materials.

Collections as Silos

When dealing with library materials outside the public domain, relevant work in the intersecting fields of copyright law, digital collections development, data curation, and digital scholarship often happens under the radar, quietly, in singular contracts between or within institutions. Collections as data practitioners working with copyrighted materials frequently operate in silos. No generalized best practices exist for institutions looking to build similar structures or to pursue the more ambitious project of building large-scale datasets from dispersed collections across institutional boundaries. For individual libraries and scholarly communities, autonomous control of large-scale datasets remains largely out of reach.

As a result, digitization efforts of copyrighted materials at academic libraries are usually limited to provisioning materials through centralized, membership-only repositories, such as the HathiTrust Digital Library, relying on their Non-Consumptive Use Research Policy for access to the HathiTrust Research Center.[9] Alarmingly, commercial vendors have quickly moved to impose financial or other prohibitions on the sharing of materials for non-consumptive research, developing siloed text and data mining portals that impose further costs and procedural limits on what libraries and researchers can do with the licensed content at scale. As recent research by Cornell librarians Peter McCracken and Emma Raub finds, "Vendors and content providers pursue many different paths to maintain control over the data that are being mined," including the deployment of contractual licensing terms that severely limit text and data mining access when it is at odds with their primary goal of "data monetization."[10]

With the conglomeration of major publishing houses and database companies, data used for cultural analytics projects is increasingly expensive and controlled by siloed repositories, severely limiting researchers' ability to conduct comprehensive and reproducible analyses across databases. *Data Speculation* seeks to revive the goal of library-owned, library-governed digital collections and open-source infrastructure by bringing into relief a key implementation barrier, empowering our "target group" of library data curators to build and curate data for the wide-ranging "beneficiaries" in diverse research communities, including librarians, scholars, and students looking to mine contemporary culture at scale.

---

[8] Amanda Levendowski, "How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem," 93 *Wash. L. Rev.* 579 (2018). Available at https://digitalcommons.law.uw.edu/wlr/vol93/iss2/2.

[9] HathiTrust's Non-Consumptive Use Research Policy provides a powerful precedent for libraries looking to steward their own data. For more information, visit their website: https://www.hathitrust.org/htrc_ncup

[10] Peter McCracken and Emma Raub, "Licensing Challenges Associated With Text and Data Mining: How Do We Get Our Patrons What They Need?" *Journal of Librarianship and Scholarly Communication* (2023). DOI: 10.31274/jlsc.15530

Legal Perceptions

Efforts to make contemporary culture collections available in computational form have inherited – and been thwarted by - the same gaps and misconceptions that accompany other digital collection development by libraries and scholars. Institutions continue to default to what copyright expert Kevin L. Smith described a decade ago as "a form of self-censorship," prioritizing "safe," public domain works in digitization projects. As Smith notes, "it is more and more troubling to realize that decisions [about digitization] are being made not based on scholarly needs or the importance of the material itself, but merely to avoid controversy and risk."[11]

Legal barriers to digitally accessing copyrighted texts are so pervasive and accepted in libraries that they are largely observable as "the dog that didn't bark" - collections not digitized, digital collections that aren't accessible online, projects that are canceled in the planning or risk assessment phases. Thomas Padilla, lead researcher for multiple grant-funded Collections as Data initiatives, locates rights assessments at scale as a "wicked problem" for libraries seeking to provision machine-actionable collections.[12] Research by Lise Jaillant, PI for multiple grant-funded artificial intelligence and cultural heritage network efforts, has shown that barriers posed by copyright extend beyond purely legal concerns, as "risk aversion" causes "many institutions to protect their own reputations and interpret legislation in a very restrictive way."[13] The heightened concern within library institutions about legal obstacles is perhaps best exemplified by Peter B. Hirtle, Emily Hudson, and Andrew T. Kenyon's landmark 2009 *Copyright & Cultural Institutions*: "Drafting and implementing copyright procedures often reveals the uncertainties in the law and demonstrates how difficult it can be to apply abstract legal principles to specific circumstances."[14]

This legal barrier, at once real and perceived, has been exacerbated by a lack of institutional investment in the copyright expertise required for the development of copyrighted digital collections. Surveying dozens of librarians from US academic and research libraries, ranging "from Ivy League colleges to rural satellite campuses," *Code of Best Practices in Fair Use for Academic and Research Libraries* authors Prudence S. Adler, Pat Aufderheide, Brandon Butler, and Peter Jaszi found that respondents "lacked a clear sense of what they and their peers might agree to as appropriate employment of fair use in recurrent situations. As a result, librarians frequently did not use their fair use rights when they could have …. The cost of this uncertainty was amplified because many research and academic librarians routinely act as the de facto arbiters of copyright practice for their institutions and the constituencies they serve."[15] Only the best-resourced libraries possess the in-house copyright expertise, in the form of both legal counsel and trained librarians, necessary to navigate these questions.

Scope of Materials

While legal precedent supporting libraries engaging in large scale exchange of restricted cultural materials as data has been well-established, the challenge presented by copyright for contemporary library practitioners includes complex issues of literacy and education. What remains to be created for individual institutions are the framework and guidelines for generalizable, decentralized implementation of copyrighted culture Collections as Data projects across

---

[11] Kevin L. Smith, "Copyright Risk Management: Principles and Strategies for Large-Scale Digitization Projects in Special Collections," *Research Library Issues: A Quarterly Report from ARL, CNI, and SPARC*, no. 279 (June 2012), https://publications.arl.org/rli279

[12] Thomas Padilla, *Responsible Operations: Data Science, Machine Learning, and AI in Libraries*. OCLC Research, 2009. https://doi.org/10.25333/xk7z-9g97.

[13] Lise Jaillant, "How can we make born-digital and digitized archives more accessible? Identifying obstacles and solutions," *Arch Sci* (2022). https://doi.org/10.1007/s10502-022-09390-7

[14] Peter B. Hirtle, Emily Hudson, and Andrew T. Kenyon, *Copyright and Cultural Institutions: Guidelines for Digitization for U.S. Libraries, Archives, and Museums* (Ithaca: Cornell University Library, 2009).

[15] Prudence S. Adler, Pat Aufderheide, Brandon Butler, and Peter Jaszi, *Code of Best Practices in Fair Use for Academic and Research Libraries* (Association of Research Libraries; Center for Social Media, School of Communication, American University, Washington College of Law, January 2012): http://bit.ly/3JflZhV

a wide-range of library institutions and communities. This National Forum is an effort to re-assert the importance of and provide guidelines for library digital stewardship of copyrighted cultural material to meet the scholarly needs of researchers, students, and librarians who benefit from access to specialized datasets. Temple University's growing digital collection of contemporary science fiction will serve, then, as an extensible proof of concept for this National Forum's exploration and development of guidelines that are widely feasible and implementable. The generalizable character of Temple's SF Project can be well-illustrated by a brief overview of the essential legal literacies for text and data mining.

Experts (including those serving as team members on this project) have articulated a framework and pedagogy for how scholars and libraries can traverse four essential legal literacies when assembling or disseminating digital corpora for computational text analysis: copyright, contracts, privacy, and ethics. Collectively, these legal literacies are referred to as "legal literacies for text data mining," or "LLTDM." To understand how LLTDM can shape copyrighted Collections as Data projects, consider the example of a researcher mining and analyzing the effects of harassing speech in social media posts, and then seeking to share these datasets for research reproducibility. As set forth above, the scholar would need to address four key principles: (i) copyright (e.g., are the posts protected by copyright? Does an exception like fair use enable TDM without seeking permissions from hundreds or thousands of authors?); (ii) contracts (e.g., do social media websites impose terms of use? Do such website agreements override copyright exceptions?); (iii) privacy (e.g., do the posts reveal information that infringes upon federal and state privacy rights of the persons described in the posts? Is republishing data a further privacy violation?); and (iv) ethics (e.g., could downloading and recirculating the content exacerbate harm to the subjects of the posts?).

Certainly, tools and frameworks exist to navigate all four literacies when digitizing and curating collections, including the Responsible Access Workflows developed by project team member Rachael Samberg and the *Codes of Best Practices in Fair Use* developed by project team members Peter Jaszi, Brandon Butler, and others. But—unlike in the social media example—by focusing on datasets of published popular culture materials that are purchased by or gifted to libraries and later digitized, library practitioners can bypass the thornier considerations within three of the four issues: *Contracts or license agreements* that can override or impose additional limitations on copyrighted content typically do not come into play when a library or institution has purchased or been gifted physical materials outright, such as a book or DVD (as opposed to licensing assets through a commercial content provider for controlled digital lending); *Privacy* laws protect against the harmful disclosure of previously unknown private information about actual and living people, and as such they are unlikely to be violated by the digitization and curation of previously published fiction or other popular culture; and *Ethics* are not laws but moral constructs for which consideration remains relevant with published works of speculative fiction, but in a far more navigable fashion than with unpublished diaries, social media posts, or personal papers.

For this reason, the scope of materials within the purview of this National Forum are composed entirely of *published* popular culture materials, ranging in media format from literature to cinema, sold on the public marketplace, and restricted from researcher access only by copyright. While issues of contracts and license agreements, privacy and ethics, still require concerted consideration for any project, libraries looking to curate copyrighted popular culture, such as science fiction, will predominantly need to deal with the obstacle of copyright. Despite widespread public perception, copyright actually presents the easiest of the four problems for libraries and researchers to address, thanks to the United States Copyright Act's recognition of "fair use," the legal right to make unlicensed use of copyrighted works under certain circumstances. Indeed, a large and growing body of legal precedent, including cases about search engines and commercial text analysis tools, as well as scholarly uses like the HathiTrust Digital Library and its support for data analysis of cultural materials, makes clear that fair use enables the curation of copyrighted

contemporary culture collections as data for scholarly and educational purposes.[16] The task at hand, then, for *Data Speculations*, is to develop frameworks and guidelines on the scope of fair use, grow consensus among specialists and practitioners, and share approaches for communities of practice looking to diversify and expand access to Collections as Data with copyrighted contemporary culture.

Moving Beyond Silos

Data Speculations builds on a number of recent and ongoing funded projects that have established a foundational legal understanding for text and data mining and documented, theorized, and guided implementation of Collections as Data. *Data Speculations* team members and proposed forum participants have led and contributed to these efforts. The *Building Legal Literacies for Text and Data Mining Institute* (National Endowment for the Humanities) emphasized the user roles of scholars and digital humanities librarians, guiding interpretation of legal and ethical considerations for TDM in existing corpora.[17] *Data Speculations* team member **Rachael Samberg** directed the Building Legal Literacies Institute; **Brandon Butler** served as project team and faculty member; **Sarah Potvin** attended. Produced under the projects, *Always Already Computational: Collections as Data* (IMLS; **Sarah Potvin** served as a Co-Investigator) and *Collections as Data: Part to Whole* (Mellon), Collections as Data forums and guidelines have been comprehensive and inclusive in their approach and audiences, reaching wide-ranging scholars, designers, administrators, and stewards while emphasizing ethical approaches and adaptive, context-specific implementations.[18] *Data Speculations* will build upon this work to target a distinctive institutional and inter-institutional need for legal guidance for the curators and stewards of humanities research datasets.

Data Speculations further develops work by team members and forum participants at their institutions, consortia, and networks, including thematically aligned scholarship and experience with designing, convening, and disseminating best practices. These include the previously-mentioned Responsible Access Workflows developed for University of California-Berkeley by **Rachael Samberg** with the goal of guiding library digitization efforts through legal literacies for text and data mining, as well as the codes of best practices in fair use facilitated by **Peter Jaszi** and **Brandon Butler** (including codes addressing the activities of academic libraries and collections containing orphan works). This National Forum responds to findings from all these wide-ranging projects, such as the need to grow institutional capacity for Collections as Data work, to consider questions of scalability, and to appreciate the scholarly limitations of ad hoc or boutique Collections as Data produced within a single institution.

Project Work Plan

The project will begin on August 1, 2023 and conclude July 31st, 2025. The project will develop over two main phases: 1) an iterative process through which the Project Team of copyright law specialists will develop a report on Temple's science fiction project, as well as preliminary documentation on the necessary **legal frameworks and guidelines** for libraries looking to broaden access to their restricted collections as data; and 2) hosting a **National Forum** with legal scholars, specialists, and stakeholders to discuss case studies and their implications, with the goal of sharing and integrating feedback on the framework and guidelines for curating restricted forms of digitized cultural heritage across diverse institutions. The bulk of the work for this National Forum will involve researching, drafting, developing, and disseminating guidelines for libraries looking to provide research access to copyrighted published culture. To this end, based on their unique experience and specializations in relevant fields of copyrighted

[16] HathiTrust's agreement with Google set a precedent for Collections as Data projects with copyrighted material, and the U.S. Supreme Court decision in Google, LLC v. Oracle Am., Inc. foregrounded teaching and research in the fair use calculus.

[17] For related work by relevant scholars on literacy for copyrighted data and text data mining, see Rachael G. Samberg and Cody Hennesy, "Law and Literacy in Non-Consumptive Text Mining: Guiding Researchers Through the Landscape of Computational Text Analysis." In *Copyright Conversations: Rights Literacy in a Digital World*, ed. Sara R. Benson (Association of College & Research Libraries, 2019). Retrieved from https://escholarship.org/uc/item/55j0h74g

[18] Collections as Data materials can be found on their website: https://collectionsasdata.github.io/

Temple University

Collections as Data, the PI and Co-PI will regularly draft and request feedback on materials from project team members as well as the National Forum participants. We also anticipate providing an opportunity for feedback, via open review, from a broader audience.

PI **Alex Wermer-Colan** (Interim Academic Director and Digital Scholarship Coordinator, Temple University Libraries, Loretta C. Duckworth Scholar's Studio) has worked for over five years at Temple Libraries on developing datasets, legal protocols, and technical infrastructure for sharing their copyrighted collections as data (including science fiction) with other libraries, researchers, and repositories like HathiTrust Digital Library. Wermer-Colan's testimony before the Library of Congress's Copyright Office was vital in securing an Exemption to the Digital Millennium Copyright Act for Text and Data Mining. Wermer-Colan is Co-PI on a new Mellon grant-funded project to support research on contemporary culture under this exemption, and his expansive, first-hand knowledge of copyrighted data curation and cultural analytics methods will provide a firm foundation for this National Forum's orientation and focus.[19] Co-PI **Sarah Potvin** (Associate Professor and Digital Scholarship Specialist, Texas A&M University) brings experience as an *Always Already Computational: Collections as Data* project team member and *Building Legal Literacies in Text and Data Mining* participant. A co-founder of the *dh+lib* project, Potvin is active in Digital Humanities and digital libraries professional communities. Potvin has been working with Wermer-Colan on expanding Temple's science fiction project to include Texas A&M since 2019.

The project team includes three nationally-recognized legal advisors, with extensive, unique expertise in fair use, copyright analysis and education, text and data mining, and community standards. **Brandon Butler** (Director of Information Policy at the University of Virginia Library) provides national guidance, education, and advocacy on intellectual property and related issues. He helped develop HathiTrust Research Center's non-consumptive use policy and, with **Peter Jaszi**, the *Code of Best Practices in Fair Use for Academic and Research Libraries*. Previously, Butler taught copyright and supervised student attorneys in American University's IP Law Clinic at American University. **Rachael Samberg** (Scholarly Communication Officer and Program Director at UC Berkeley Library) teaches across the country on copyright and information policy, and is a national presenter for ACRL's Scholarly Communication Roadshow. As was described, Samberg led the *Building Legal Literacies for Text and Data Mining Institute* (NEH) and currently leads *Legal Literacies for Text Data Mining - Cross Border* (NEH). **Peter Jaszi** (Professor Emeritus at American University Washington College of Law) co-directed the Program on Information Justice and Intellectual Property at Washington College of Law. The author of three monographs and more than a dozen articles and chapters on intellectual property, copyright, and fair use, Jaszi has frequently testified in front of House and Senate subcommittees and held leadership positions in the Copyright Society of the U.S.A. and its journal, and the Intellectual Property Section of the American Association of Law Schools.

The first year of the National Forum grant cycle will focus on gathering information about the science fiction use case through surveys and interviews, developing a preliminary report, and planning the forum. The National Forum will take place at the beginning of the second grant year, in the fall of 2024, and will serve to provide feedback and areas of growth for the guidelines and resources in the report, ensuring our deliverables are robust, accessible, and useful to a wide audience. In the second year of the two-year grant cycle, we will finalize our deliverables, explore further areas of need and opportunity, and present our findings at conferences and through various publications.

Phase	**Dates**	**Activities**	**Deliverables & Dissemination**
Year 1. Diagnostic Work: Research, Planning, and Environmental Scan			
Phase 1. Scope and Clearly	August 2023 -	● Meet monthly with Project Team - AWC, SP, RS, BB, PJ	● Finalized detailed description of use case with project team

[19] For the ruling on DMCA Section 1201, see the Association of Research Libraries' "Text and Data Mining Exemption to Digital Millennium Copyright Act Would Advance Knowledge of Diverse Works" (2021): http://bit.ly/3l5uBjs

Define Core Use Cases and Problems; Initiate Surveying and Interviews	January 2024	• Create Project Website - AWC • Hire grad student assistant (GSA) - AWC • Finalize research design (including survey plan for IRB) - AWC, SP, RS, BB, PJ • Pursue approval for multi-institutional IRB for interviews - AWC, SP, RS, BB, PJ • Initiate environmental scan - AWC, SP • Identify and begin surveying science fiction collections as data stewards, researchers, and stakeholders - AWC, SP, RS, BB, PJ	• Disseminate detailed description to committed National Forum participants for initial review and feedback • Scholarly article, published in Open Access journal or openly available as preprint, that examines science fiction collections as data
Phase 2. Draft report and proposed guidelines; plan for National Forum	February 2024- Jul 2024	• Meet monthly with Project Team - AWC, SP, RS, BB, PJ • Complete environmental scan - AWC, SP, RS, BB, PJ, GSA • Draft initial report - AWC, SP, RS, BB, PJ • Develop call for participants in the National Forum - AWC, SP, GSA • Identify and invite additional specialists to the National Forum - AWC, SP, GSA • Create agenda for National Forum, delineating publicly-accessible talks and participants meetings - AWC, SP • Identify partners such as CLIR/DLF/ARL to publish project results - AWC, SP	• Scholarly article, published in Open Access journal or openly available as preprint, that analyzes interview data from Phase 1 • Call for participants in National Forum, distributed to national/international listservs • Disseminate agenda for National Forum and open registration for publicly-accessible talks • Presentation at Digital Humanities, Open Repositories, Texas Conference on Digital Libraries
Year 2. Prescriptive Year: National Forum, Iteration, and Dissemination			
Phase 3. Host Virtual National Forum; Gather feedback on report	August 2024- January 2025	• Meet monthly with Project Team - AWC, SP, RS, BB, PJ • Host National Forum - AWC, SP • Conduct surveys of the wider range of copyrighted collections as data - AWC, SP • Gather participant feedback from National Forum and DLF Forums - AWC, SP, GSA • Develop expanded set of use cases - AWC, SP, RS, BB, PJ, GSA	• National Forum • Presentation or workshop at Digital Library Federation • Targeted dissemination of draft guidelines
Phase 4. Finalize and publish guidelines; Disseminate materials and plan for future implementation	February 2025- July-2025	• Meet monthly with Project Team - AWC, SP, RS, BB, PJ • Open review of guidelines - AWC, SP • Finalize and publish guidelines with publishing partners - AWC, SP, RS, BB, PJ • Finalize and publish findings and all project materials - AWC, SP, RS, BB, PJ	• Finalized guidelines for restricted data curation • Present at Keystone DH and ACH • Scholarly article (OA) on state of the field for legally provisioning science fiction collections as data

Data Speculations will convene approximately twenty specialists, including library stewards and researchers from diverse backgrounds and institutions who can attest to the obstacles and opportunities for making contemporary cultural materials available as data, especially for marginalized communities, collections, and institutions. National Forum participants will include both invited experts, such as internationally-recognized legal scholar **Matthew Sag** (Professor of Law and Artificial Intelligence, Machine Learning, and Data Science, Emory University), as well as additional specialists who will apply to our open call for participation with their relevant use cases for copyrighted corpora. We are coordinating closely with **Glen Layne-Worthey** (Associate Director for Research Support Services, HathiTrust Research Center), who provides a valuable perspective on HathiTrust's complementary, but centralized membership model for provisioning collections as data, one driven by the commitment to creating common guidelines and methods for institutions looking to expand access to restricted humanities data. Additional invited experts who have expressed support for this National Forum and indicated their willingness to participate include **Thomas Padilla** (Deputy Director, Internet Archive; *Collections as Data* PI), **Melissa Levine (**Library Lead Copyright Officer, University of Michigan), **Purdom Lindblad** (Assistant Director of Innovation and Learning,

Maryland Institute for Technology in the Humanities), **Michelle Dalmau (**Head, Digital Collections Services, University Libraries; Co-Director, Institute for Digital Arts & Humanities, Indiana University), **Devin Becker** (Associate Dean, Research & Instruction, University of Idaho Library), and **Nicole Lemire-Garlic** (Faculty in Justice Management, School of Social Research, University of Nevada, Reno).

This community of practitioners will help disseminate an open call for participants and review proposals. The Project Team will commit concerted time and effort to recruiting library practitioners and DH specialists with diverse use cases for copyrighted Collections as Data and wide-ranging perspectives on copyright law and ethics, including but not limited to representatives of the Association of College & Research Libraries, the Association of Research Libraries, and CLIR/DLF, as well as practitioners at diverse libraries and institutions across the country (including R1s, public institutions, small liberal arts colleges, HBCUs and HSIs). Hosted online, the *Data Speculations* National Forum will include free and publicly-accessible virtual talks and panels, as well as structured exercises and meetings designed for participants. Participants who give talks, contribute to panels, and/or attend structured meetings will receive stipends. Additional perspectives will be incorporated via interviews, surveys, and an open review of draft guidelines.

The project team anticipates travel to professional conferences to present and solicit feedback on findings-in-progress and to grow a community of practice. Given our goal of connecting to collection stewards and digital humanities researchers, we expect to present at national and international conferences of the Digital Library Federation, the Association for Computers in the Humanities, and Open Repositories, regional conferences such as the Texas Conference on Digital Libraries and Keystone DH, as well as venues for the study of contemporary culture, such as the Popular Culture Association and the Science Fiction Research Association. We will seek to publish with such journals as *portal: Libraries and the Academy, dh+lib*, *PMLA,* the *Journal of Cultural Analytics*, and the *Journal of the Copyright Society of the USA*.

Diversity Plan

The implications of developing a legal framework for communities of practice to build and share popular culture datasets are vast, and this National Forum seeks to explore this question from a diverse set of perspectives with a commitment to the ethics and privacy concerns raised by data-based research. While our focus is on enabling libraries and scholars to grow datasets of 'copyrighted' literature for their own purposes, we intend to explore the full scope of what material can come under copyright, with special attention to what unique circumstances, both risks and affordances, are opened up by providing large-scale research access to popular culture materials as diverse as mass-market science-fiction books and Disney movies. Library and scholarly communities need the ability to grow datasets for their own work, but they also should be empowered to protect that data from harmful use by those outside their communities. Inspired by calls to steward data not just according to principles of FAIR (Findable, Accessible, Interoperable, Reusable), but also CARE (Collective Benefit, Authority to Control, Responsibility, Ethics), as this National Forum investigates new ways of expanding access to commodified contemporary culture materials for digital humanities research, we will seek out and tackle vital questions at the intersection of FAIR and CARE, such as the potential for copyright to protect marginalized communities from harm.[20]

To this end, we will convene specialists from geographically-distributed public institutions across the country, including HSIs (beyond Texas A&M) and HBCUs. We aim to reach a diverse range of institutions via surveying, open calls, and interviews.[21] We will prioritize inviting and including racially and ethnically diverse forum

[20] For more on CARE Principles for Indigenous Data Governance, see their website: https://www.gida-global.org/care

[21] We are attentive to the fact that HSI designation relies on undergraduate enrollment demographics, which do not always reflect the demographics of university-affiliated librarians and researchers. Vargas, Villa-Palomino, and Davis (2020), examining Latinx faculty representation in Hispanic-Serving Institutions, find: "Analyses of all Title V funded HSIs from 2009–2016 (N = 167)

participants from a range of institutions, use cases and perspectives on copyright law and ethics. We will also develop and maintain a Code of Conduct for all project participants. Our call for additional forum applicants, as well as our open review of the documents produced through the forum, will be distributed across major listservs for library science and digital humanities scholarship.

This National Forum will thereby develop approaches and protocols that can enable a diverse array of institutions and practitioners to expand access to contemporary culture collections, ensuring the exchange and growth of datasets representing marginalized perspectives in contemporary culture, such as Kansas University Libraries's ongoing Black Book Interactive Project.[22] Our work also intersects with related efforts to reduce barriers presented by copyright law to making digital collections fully accessible for people with disabilities, with an emphasis on an expansive perspective on fair-use. Project outputs will be openly, digitally disseminated.

Project Results

This National Forum endeavors to convene specialists and advisors on legal frameworks for text data mining and restricted data curation to scope out the infrastructure necessary for academic libraries to build copyrighted datasets that can be made available to researchers in the most diverse and accessible formats permissible by the law, thereby presenting an alternative to the predominant model of subscription-based, proprietary databases for primary source research and text data mining. By consulting scholars from prior conversations on this evolving field, and by putting them in conversation with diverse library practitioners of copyrighted data curation, *Data Speculations* will result in the development of an applied framework with practical measures for libraries seeking to expand research access to copyrighted collections as data within and beyond their own institutions.

In the service of developing a legal and technical framework for the library exchange and curation of copyrighted digital collections of contemporary culture, this National Forum will produce and disseminate documentation, including a report on the science fiction project and related use cases, as well as preliminary guidelines for libraries looking to familiarize themselves with the risks and affordances of this work. Our Project Team of legal specialists have previously authored *Codes of Best Practices in Fair Use,* relying on the model established by the Center for Media and Social Impact, and tailored to particular use cases and audiences. We anticipate that *Data Speculations*, by investigating needs and producing a report on our generalizable findings, will seed a subsequent funding proposal for the production of a *Code of Best Practices for Copyrighted Collection Data*.

A static website hosted and maintained by Temple University Libraries' Loretta C. Duckworth Scholars Studio will provide access to drafts and final products in open access format. We also anticipate collaborating with the Council on Library and Information Resources to identify a sustainable place for publishing our report for a general audience. The project team will also present at conferences and publish in journals, seeking out opportunities to engage practitioners at institutions with marginalized collections, while also broadcasting our findings to a wider audience of digital humanists, cultural studies scholars, digital collection librarians, and copyright experts.

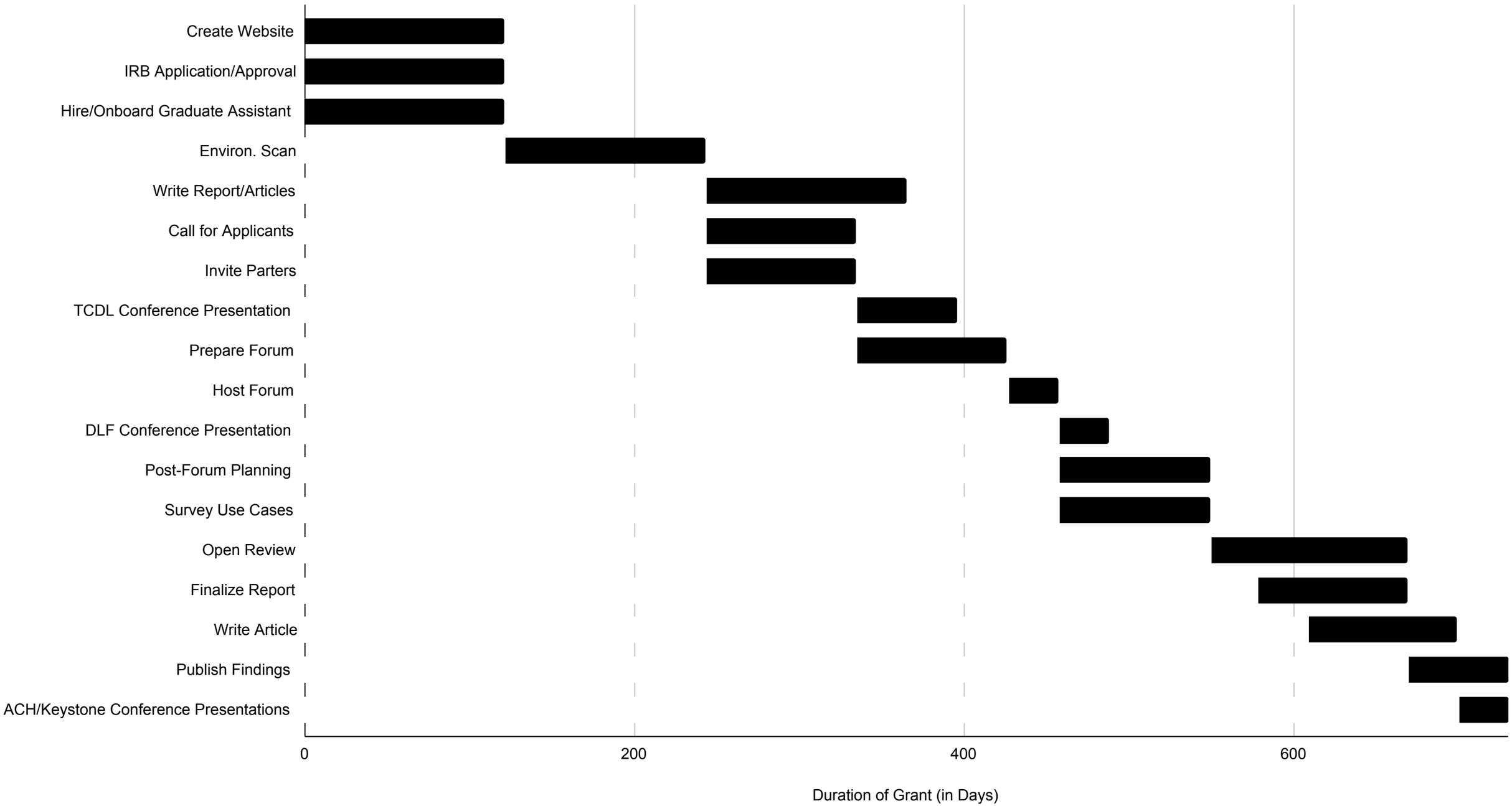
---

indicate that the average Latinx student-to-Latinx faculty ratio is 146:1, whereas the corollary White student-to-White faculty ratio is 10:1." Bonilla-Silva and Peoples (2022) argue in their article on Historically White Colleges and Universities: "our claim is not purely based on demography. Although numbers matter, the way that racial power and history has shaped these institutions matters more." Vargas N, Villa-Palomino J, Davis E, "Latinx faculty representation and resource allocation at Hispanic Serving Institutions." *Race Ethnicity and Education*. 2000;23(1):39-54. doi: 10.1080/13613324.2019.1679749; Bonilla-Silva E, Peoples CE. "Historically White Colleges and Universities: The Unbearable Whiteness of (Most) Colleges and Universities in America." *American Behavioral Scientist*. 2022;66 (11).

[22] The Black Book Interactive Project (https://bbip.ku.edu/) is an excellent example of relevant work ensuring the availability of representative datasets for scholars. We hope to partner with their team and similar projects.

Data Speculations: A National Forum

Schedule of Completion



Digital Products Plan

Data Speculations: A National Forum on Library Digital Stewardship for Copyrighted Contemporary Culture will not produce software or code. The primary anticipated outputs of *Data Speculations* are digital, textual documents. This documentation will overview the challenges and opportunities today for growing copyrighted culture collections as data, including a white paper on the state of the science fiction project, the various case studies explored through the National Forum, and the conceptual frameworks and general guidelines we develop for library practitioners managing the workflows, protocols, and technical infrastructure involved with procuring controlled access for researchers to restricted data. In addition to these foundational documents, we will also produce scholarly articles and presentations exploring the implications and impact of this work for the future of library data services in the humanities.

The project team is committed to ensuring that all content produced over the course of the grant is **accessible**, **open**, **persistent**, and **findable**. All output of the project will be openly licensed or made available in openly licensed forms (e.g., articles and presentations may appear in closed venues, but the project team will produce and post preprints, postprints, or other accessible versions).

The National Forum documentation will be created and hosted by Temple University Libraries on a static website. The project website, maintained by the PI and Co-PI, will serve as a site of dissemination during the duration of the project. The website and all posted projects will be designed to comply with Web Content Accessibility Guidelines.

Digital assets will be deposited to multiple persistent, indexed repositories for preservation and access, including Zenodo and project team members' institutional repositories. In addition, we will collaborate with organizations like the Digital Library Federation and the Council on Library and Information Resources to identify open-access platforms for publishing findings from the National Forum and building communities of practice around these issues.

The only exception to *Data Speculations'* commitment to **accessible**, **open**, **persistent**, and **findable** digital products are instances where research ethics preclude public disseminations. Interview and survey data gathered in the first phase of the project may include confidential or sensitive information. While the project team, which possesses extensive experience in analyzing concerns related to intellectual property, ethics, and privacy, anticipates provisioning access via anonymization and selective restriction, all procedures are subject to institutional review board approval. At this time, we do not anticipate any cultural sensitivities or privacy concerns for any digital products produced through *Data Speculations*.

Organizational Profile: Temple University Libraries and University Press

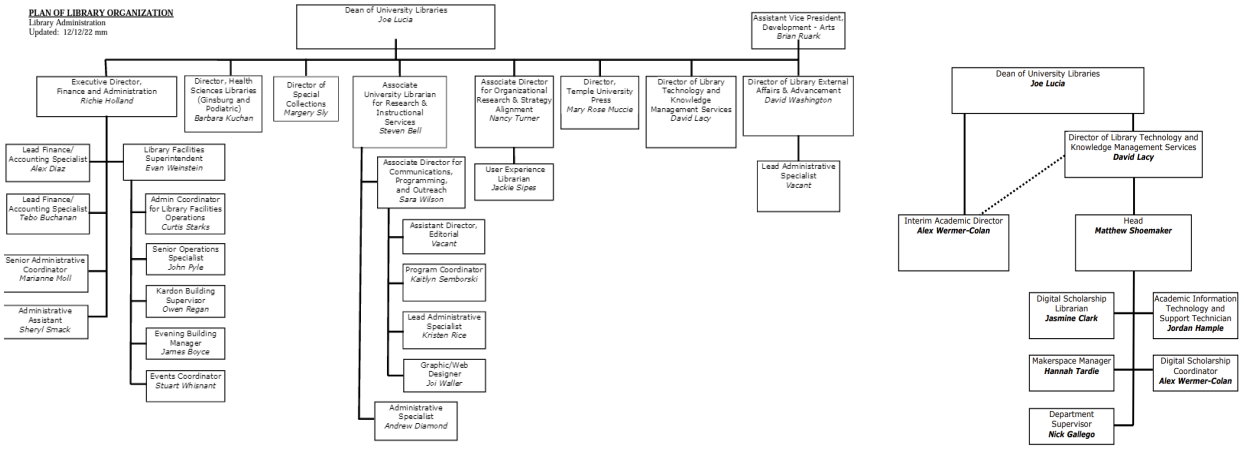
Mission: Connecting people and ideas to enhance learning, research, clinical practice, and creativity.

Key Functions: Temple University Libraries & the University Press nourish and sustain the academic enterprise through:

- Facilities, staff, collections, and services that support study, instruction, inquiry, and scholarship;
- Instructional and service engagement with the learning environment to aid students in the exploration and discovery of new ideas and the development of informed critical thinking;
- Provision of materials, tools, and expertise to amplify the productivity and extend the reach of students, scholars, researchers and clinicians;
- Dissemination of ideas and culture through publishing, events, outreach, and collaborative programming with academic and community partners.

See Temple University Libraries | Mission & Key Functions (Revised and Reaffirmed September 2017 by TULUP Library Administration): https://library.temple.edu/webpages/mission-key-functions

Organizational Structure: The Libraries are part of Temple University. This org chart is on the Libraries' website:



Service area: Temple University Libraries serves the Temple community and beyond, including close to 40,000 students; over 2,000 full-time faculty; and researchers and visitors on Main, Center City, and Health Sciences Center campuses in Philadelphia and on our Ambler and Harrisburg campuses. We are committed to providing research and learning services, offering open access to our facilities and information resources, and fostering innovation and experimentation.

History of the organization: Temple's first library consisted of donated books from students and faculty, located in a house in Northern Philadelphia. Since then, we have evolved, and are not comprised of a law library, two health sciences libraries, and the 2019 Charles Library, a centerpiece on Temple's Main campus. In addition to the Special Collections Research Center, the Charles A. Blockson African American collection is world-renown. A brief timeline of the history is provided here: Temple University Libraries | Temple University Libraries History: https://library.temple.edu/webpages/temple-university-libraries-history

The unit with primary responsibility for this project is the Loretta C. Duckworth Scholars Studio, part of the Charles Library on the University's main campus. The Scholars Studio is a space for collaborative work in areas of research and technology, including digital humanities, digital arts, making, big data, and interactive media and gaming. LCDSS supports researchers and teachers, providing technical equipment and software for a variety of activities including textual analysis, working with big data, working in and creating 3D spaces, geospatial technology, games and visualizations.

Visit the LCDSS website for more information: https://library.temple.edu/lcdss