Data Kindreds: Towards a Legal Framework for Expanding Access to Restricted Collections as Data

Temple University and Texas A&M University request a 2-year $126,000 IMLS National Leadership Grant for Libraries to host two virtual National Forums convening specialists in digital humanities research, library data services, special collections, and copyright law, with the goal of defining best practices and guidelines for libraries supporting computational research on contemporary culture. By developing inclusive legal frameworks and protocols for libraries building and sharing access to copyrighted digital collections, *Data Kindreds* pursues IMLS's 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).

Project Justification: The Collections as Data movement in libraries, archives, and museums has expanded the ways cultural heritage materials can be made available for research and teaching; increasingly, libraries and stewards enable teachers, scholars, and journalists to access cultural heritage collections in computational form, allowing them to mine, map, analyze, and visualize these materials in new and unforeseen ways. But perceived legal barriers to procuring copyrighted data at scale has prevented most academic libraries from growing digital collections of contemporary culture and providing computational access to the literary, artistic, and historical materials that represent the most diverse period of cultural production in human history, the twentieth and twenty-first centuries.

Efforts to make contemporary collections available in computational form have inherited – and been thwarted by - the same gaps and misconceptions that accompany other digital collection development building. Institutions continue to default to what Kevin L. Smith described a decade ago as "a form of self-censorship," prioritizing "safe," public domain works in digitization projects. 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."[1] This problem is perpetuated through inadequate training and expertise, not only about the law, but about proper protocols and procedures. In this setting, the scope and span of digitally-accessible collections has come to reflect widespread institutional risk aversion, defaulting to clearly out-of-copyright works, such as those published before 1927. As a result, collections of public-domain work from the nineteenth century and before are widely available online today, while collections of contemporary culture, as digital exhibits or disaggregated datasets, remain very difficult to find.

Yet collections as data work actually provides unique legal affordances for expansive use of copyrighted collections, opportunities that remain underexplored by cultural heritage institutions. The United States Copyright Act provides for "fair use," a legal exception permitting unlicensed use of copyright-protected works under certain circumstances. Preliminary legal analyses, including precedents set by the HathiTrust Digital Library for supporting data analysis of cultural materials, make a strong case for fair use justifying the curation of copyrighted collections as data for scholarly and educational purposes. A thorough legal review, along with the development of legal best practices and guidelines for copyright education and advocacy, can enable a wide range of academic libraries to prioritize developing the protocols and workflows for provisioning community-valued copyrighted collections as data for research and teaching.

Data Kindreds builds on NEH-sponsored work to teach the legal literacies of text and data mining, as well as IMLS- and Mellon-sponsored work to document, theorize, and implement Collections as Data projects.[2] By consulting scholars from those prior conversations, and by putting them in conversation with diverse specialists in this emerging field of library science, *Data Kindreds* aims to formulate an applied framework with practical measures for library and data practitioners seeking to curate copyrighted collections as data to researchers and teachers within and beyond their own institutions. "Kindreds," a nod to the science fiction writer Octavia Butler's novel, *Kindreds*, acknowledges that the technological solution to our problems are legally scalable: by gathering related, or kindred, use cases, legal scholars can make fair use determinations based on shared characteristics and purposes across institutions and collections, empowering library practitioners to then build shared collections of contemporary culture across institutions that can meet the growing computational needs of researchers and teachers working to analyze and visualize the recent past at scale.

Project Work Plan: *Data Kindreds* will convene two forums: 1) an opening forum of relevant experts, as well as participants selected through an application process, working to generate foundational access questions for digitized and

[1] Kevin L. Smith, "Copyright Risk Assessment in Special Collections," *Research Library Issues: A Quarterly Report from ARL, CNI, and SPARC*, no. 279 (June 2012).

[2] The Building Legal Literacies for Text and Data Mining Institute (NEH) guided scholars and librarians through legal and ethical considerations for text data mining. The Collections as Data forums and guidelines, produced under Always Already Computational (IMLS) and Part to Whole (Mellon) projects, were comprehensive and inclusive in their approach to audiences working with digital collections. Other cooperative efforts, such as AEOLIAN (Artificial Intelligence for Cultural Organizations; Arts and Humanities Research Council), have forged international networks to expand access to digital cultural heritage with artificial intelligence.

born-digital cultural materials under copyright; and (2) a concluding forum of legal scholars and stakeholders to produce guidelines for sharing copyrighted materials as data. The project will begin on August 1, 2023 and conclude July 31st, 2025, and will take place over three stages: 1) an **environmental scan**, including interviews, identifying a series of use cases and personas, culminating in the **opening forum** on the needs of librarians and scholars; 2) an iterative process of developing with the group of advisors and specialists a **legal framework, protocols, and guidelines** for libraries and archives looking to broaden access to their restricted collections as data; and 3) hosting a **concluding forum** with legal scholars, specialists, and stakeholders to discuss the results and plan how to share and receive feedback on the framework and guidelines for curating restricted forms of digitized cultural heritage across diverse institutions. Each stage will take 6 months of planning and work, and will be designed to ensure deliverables are robust, accessible, and useful. In the final 6 months of the two-year grant cycle, we will finalize our deliverables and explore further areas of need and opportunity.

PI Wermer-Colan has worked for many years at Temple University Libraries on developing copyrighted datasets, legal protocols, and technical infrastructure for sharing their copyrighted collections as data to researchers and libraries; he has presented on Temple University's project to digitize and share science fiction literature as data at regional, national, and international conferences in the digital humanities. Co-PI Potvin brings previous experience as an Always Already Computational project team member and dh+lib co-founder; she has been working with Wermer-Colan on expanding Temple's science fiction project to include Texas A&M since 2019. *Data Kindreds* has secured the commitment of three national legal advisors, representing extensive, unique expertise in fair use, copyright analysis and education, text and data mining, and community standards - Brandon Butler (University of Virginia); Rachael Samberg (University of California, Berkeley); and Peter Jaszi (American University). We will additionally invite three advisors representing library stewards and researchers from diverse backgrounds and institutions who can testify to the obstacles and opportunities for making contemporary cultural materials available as data, especially for marginalized collections, communities and institutions. We also plan to coordinate closely with Glen Layne-Worthey of the HathiTrust Research Center, since HathiTrust operates under a complementary, but distinctively centralized membership model for provisioning collections as data, one driven by the commitment to creating common guidelines and methods for sharing restricted data.

The project team and advisory board anticipate 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 such conferences as the Modern Languages Association, the Alliance of Digital Humanities Organizations, the Association for Computers in the Humanities, the Society of American Archivists, the American Historical Association, as well as venues for the study of contemporary culture, such as the Science Fiction Research Association and the Association for the Study of the Arts of the Present. We will seek to publish with such journals as *PMLA,* the *Journal of Cultural Analytics*, and the *Journal of the Copyright Society of the USA*.

Diversity Plan: *Data Kindreds* will convene specialists from geographically-distributed public institutions across the country, including HSIs and HBCUs. We will prioritize inviting and including ethnically diverse forum participants and advisory team members from a range of institutions and develop and maintain a Code of Conduct for all project participants. 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. Our work also intersects with related efforts to expand access to collections as data for people with disabilities. All project outputs will be openly, digitally disseminated.

Project Results: In the service of developing a legal and technical framework for the exchange and curation of copyrighted digital collections, *Data Kindreds* will produce and disseminate research documentation, including legal and technical guidelines. Advisory board members 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. As a major deliverable of this grant, the project team and advisory board anticipate producing a *Code of Best Practices for Copyrighted Collection Data*. A static website hosted by Temple University Libraries will provide access to works-in-progress and final products in open access format. The project team will present at conferences and publish in journals, seeking out opportunities to engage practitioners at institutions with marginalized collections and revising published guidelines to better serve the diverse needs of librarians and scholars working within and across institutions.

Budget Summary: The requested project budget of $126,000 includes project staff (PI and Co-PI) salaries (15%); two graduate student assistants (15%); national forum participant and advisory board stipends (40%); equipment and virtual hosting fees (3%); travel costs to present at conferences (7%); and administrative overhead (20%).