Collections as Data: Part to Whole

Summary

The University of Nevada Las Vegas, in collaboration with the University of Iowa seek $750,000 to foster sustainable development of collections as data and the data driven scholarship that aim to make use of them. Over a period of three years *Collections as Data: Part to Whole* investigators and team members aim to organize two cohorts and support them via regrants, training, and consultation. Cohorts would be comprised of project teams jointly led by librarians and disciplinary scholars. Models developed by project teams would support collections as data implementation and holistic reconceptualization of services and roles that support scholarly use. Collections as data produced by project activity would exhibit high research value, bring underrepresented histories to light, represent a diversity of content types, languages, and descriptive practices, and arise from a representative set of institutional contexts. An inclusive approach is taken to model development in order to expand the field of scholarly possibility.

Background

Scholars across the academy are increasingly turning to computational methods such as text and data mining, network analysis, and data visualization to advance their research. Collectively we refer to this work as *data driven scholarship*. Examples of this work include efforts to evaluate the changing prominence of gender in fiction across 170 years of publication, efforts to identify and extract all poetic content from the entirety of American newspaper history, efforts to measure and visualize characteristics of spoken word archives at scale, and efforts to experiment with image analysis as a means to enhance access to digital images.[1] [2] [3] [4] Data driven scholarship depends on data that are computer readable and machine actionable. Unfortunately, most cultural heritage institutions struggle to meet this need because they do not think of their digital collections as this kind of data. Rather, items in digital collections are generally treated as surrogates of physical objects and their organization and the expectations for their use are based on the metaphor of a physical bookshelf, gallery wall, or listening booth. A collections as data orientation suggests different approaches to collection production, documentation, and access. For example, on the production side it is common to produce TIFFs of newspaper images in order to support a high fidelity reading experience. It is less common to consider whether the TIFF format is optimal for experiments where the user of the page is an algorithm rather than the human eye. On the documentation side, it is common to provide descriptive metadata in order to support user evaluation of object provenance. It is less common to document

[1] Underwood, Ted, David Bamman, and Sabrina Lee. "The Transformation of Gender in English-Language Fiction." *Journal of Cultural Analytics*, 2018. https://doi.org/10.22148/16.019.

[2] Lorang, Elizabeth, Leen-Kiat Soh, Maanas Varma Datla, and Spencer Kulwicki. "Developing an Image-Based Classifier for Detecting Poetic Content in Historic Newspaper Collections." *D-Lib Magazine* 21, no. 7/8 (July 2015). https://doi.org/10.1045/july2015-lorang.

[3] Clement, T. and McLaughlin, S. "Measured Applause: Toward a Cultural Analysis of Audio Collections." Cultural Analytics, vol. 1, no. 1, 2016.

[4] Kogan, Gene. "Machine Learning for Artists." Accessed January 30, 2018. https://ml4a.github.io/.

optical character recognition (OCR) quality. On the access side, it is common to provide the ability to download a digitized object. It is less common to enable programmatic download access to a collection at a scale of thousands or even millions of objects. Despite an uptick in data driven scholarship activity, as evidenced by evolving tenure and promotion guidelines advanced by the Modern Language Association and the American Historical Association, library effort allocated to collections as data implementation remains an exception rather than a norm.

In 2016 several authors of this prospectus along with Sarah Potvin, Hannah Frost, and Elizabeth Russey Roke began work on *Always Already Computational: Collections as Data* (IMLS LG-73-16-0096-16).[5] The goal of this grant is to create a framework and set of resources that guide libraries and other cultural heritage organizations in the development, description, and dissemination of collections that are readily amenable to computational analysis. The project began with a national forum at the University of California Santa Barbara that brought together leading researchers, technologists, librarians, archivists, and museum professionals to articulate the key challenges to collections as data development. Subsequently the project team generated the widely circulated *Santa Barbara Statement on Collections as Data*.[6] *The Santa Barbara Statement* distills insights surfaced at the forum into guiding principles for collections as data development. These insights covered topics including but not limited to interoperability, process documentation, centering specific user needs, and ethical considerations that serve to avoid creating collections that replicate contemporary inequities.

Over the course of the next year, investigators engaged with information professionals and scholars at several conferences and meetings including Stanford University's LDCX, the Texas Conference on Digital Libraries, the American Library Association Annual Meeting, the Society of American Archivists Annual Meeting, Digital Humanities 2017, Digital Library Federation 2017, Samvera Connect, the Fall 2017 Coalition for Networked Information meeting, and American Historical Association 2018. Engagement with these communities helped the project team make significant revisions to the statement and supported the development of use cases and personas that document the needs of collections as data producers *and* collections as data users. The project team also solicited a collection of case studies called *Collections as Data Facets*.[7] *Collections as Data Facets* provide examples of how different cultural heritage institutions have developed and provided access to collections as data. Taken collectively, analysis of this work suggests that collections as data implementation models are needed in order to support broader library participation. A move toward models would represent a vital maturation of a field of work that is primarily characterized by examples absent the explicit intention of supporting broader replicability.

In addition to producing the above deliverables, *Always Already Computational* has sparked an international conversation about collections as data. Multiple events have referred to their focus on "collections as data", and the term has been included in state-wide collection strategies, research agendas, and job descriptions. These developments evidence that collections as data is a concept that

[5] Padilla, T., Allen, L., Varner, S., Potvin, S., Frost, H., and Russey Roke, E. *Always Already Computational: Collections as Data* Available at https://collectionsasdata.github.io/

[6] Ibid., Available at https://collectionsasdata.github.io/statement

[7] Ibid., Available at https://collectionsasdata.github.io/facets

resonates with an increasing number of communities.[8] [9] [10] [11] While there is considerable excitement about libraries and other cultural heritage institutions providing new kinds of collections and new kinds of access to them, this conversation has also brought to light significant shortcomings within existing systems for supporting scholarly use of these collections.

A common response to supporting data driven scholarship is to hire a Digital Humanities or Digital Scholarship Librarian or create a digital scholarship department or center. These roles and spaces are often siloed rather than integrated across library roles and services. Lack of integration directly impacts services that support collections as data production, use, and long term preservation. For example, a researcher and a digital scholarship librarian may work together to create corpora for a text analysis project but it is unlikely that the acquisitions department will bring those corpora into the library's collection or that a metadata librarian will describe it. As an additional example, a scholar may want to assign students to perform network analysis on the metadata describing an archival collection. The digital humanities librarian may be tasked with creating that dataset but the job will likely be seen as a one-off and not as a possible use-case for which systems should be designed. It remains the case that libraries often assign responsibility for data driven scholarship to a single position or handful of staff within a department where outside the library data driven scholarship is increasingly an institution-wide pursuit. A more holistic approach to role and service development is needed. *Collections as Data: Part to Whole* draws upon the lessons of *Always Already Computational* and aims to set a field of engagement that equally addresses the challenge of collections as data implementation and the development of roles and services that support their scholarly use.

Rationale

Collections as Data: Part to Whole seeks to make a wider variety of data driven scholarship possible. Target user communities include but are not limited to the Digital Humanities, Digital Art History, Public History, and Computational Social Science. Scholars employing data driven methods and tools typically do not have access to collections that are designed to support computational use. Much of this situation can be attributed to decades of digitization workflows, interface design, and infrastructure development that assumes a user who desires a digital environment that emulates physical interactions with collection objects on an item by item basis. The Hathitrust Research Center and the Library of Congress' *Chronicling America* are useful examples of enabling computational use of large cultural heritage collections, but they may be difficult for other institutions to follow, given the huge size of both institutions. *Always Already Computational* initiated an effort to close this gap by documenting a range of collections as data implementations. *Collections as Data: Part to Whole*

[8] "Collections as Data - Hackathon/Collaborative Workshop", Moore Institute at National University of Ireland Galway, Available at http://mooreinstitute.ie/event/collections-data-hackathon-collaborative-workshop/

[9] "2017/2018 SCLG Plans & Priorities for 2017/2018 Based on the University of California Library Collection: Content for the 21st Century and Beyond," September 29, 2017. http://libraries.universityofcalifornia.edu/groups/files/sclg/docs/SCLG_2017_2018%20Plan.pdf.

[10] Burrows, Toby, Deb Verhoeven, and Christopher McAvaney. "Cultural Heritage & Library Collections as Data and Their Role in Digital Humanities Infrastructure." *EResearch in Humanities and Social Sciences*, n.d. https://conference.eresearch.edu.au/2017/08/cultural-heritage-library-collections-as-data-and-their-role-in-digital-humanities-infrastructure/.

[11] Weber, Chela Scott. "Research and Learning Agenda for Archives, Special, and Distinctive Collections and Research Libraries." OCLC Research, 2017. https://doi.org/10.25333/C3C34F.

would build on this work by supporting the development of broadly viable collections as data implementation models.

Fully realizing the scholarly possibilities that collections as data provide additionally depend on ready availability of sustainable roles and services that support their use. To date, evolving scholarly needs have led libraries to create various combinations of digital and/or data oriented services. These specialized services variably aim to support scholars in the humanities, arts, social sciences, sciences, and professional schools. The data driven nature of work exhibited by these research communities are interesting in their own right. They become more consequential for the design and support of library collections, roles, and scholarly services where they are considered as evidence that anticipates data driven methods and tools being a core aspect of engaging questions across any discipline or professional path. In this line of thinking the notion of all digital or data oriented services centered in a single functional unit or even distributed across two functional units quickly becomes unsustainable. A collections as data paradigm suggests that an orientation is possible that looks beyond traditionally rendered "front of house" and "back of house" dichotomies, divisions between "functional" and "subject" specialization, and divisions between "technical" and "public services.

Collections as Data: Part to Whole brings the question of how to implement collections as data together with the question of how to develop roles and services that optimally support their scholarly use. A national effort that includes a wide variety of institutions, researchers, and projects is required in order to develop broadly viable models that enable substantive progress on these questions. An emphasis on diversity in this work is vital given a significantly heterogeneous scholarly, institutional, and collection landscape. Given this reality, *Collections as Data: Part to Whole* seeks to form a national cohort, containing diverse projects, facilitated by an institutionally decentralized team of investigators. Drawing from the *Santa Barbara Statement on Collections as Data*, the work of the cohort would **(1)** support ethically grounded creation of collections as data that help collection creators avoid reinscribing bias in the cultural canon and harming marginalized communities, **(2)** provide access to collections as data in ways that align with the needs of multiple users, including users who have a range of technical expectations for types of access, and those who have ethical concerns about the scope of access, **(3)** develop organizational frameworks that support productive partnerships between libraries, departments, labs, research centers, university information technology, local cultural heritage organizations, university presses and more, and **(4)** demonstrate commitments to utilization of open source technologies that interoperate with a broader open scholarly communication infrastructure. *Collections as Data: Part to Whole* may also surface tool gaps and skill gaps that could contribute to The Andrew W. Mellon Foundation's scholarly communications portfolio moving forward.

Collections as Data: Part to Whole investigators and team members are well positioned to steward a national effort that addresses these challenges. They are committed to advancing next generation library engagement with scholarly inquiry regardless of institutional bounds. Qualifying evidence of prior and current investigator and team member engagement in this vein includes but is not limited to *Always Already Computational: Collections as Data (IMLS LG-73-16-0096-16) - a national effort to develop awareness and initials step toward development of computationally amenable collections, Data Refuge - an international effort to save climate data and advocate for climate science*, and

DocSouth Data - an exemplary collections as data effort.[12] [13] [14] Based on these experiences, the investigators and team members believe that transitioning collections as data from a community wide advocacy effort to a community wide implementation effort requires a catalyst that no single library or university is readily suited to provide. The Andrew W. Mellon Foundation, with its commitment to promoting, "the common good by supporting the creation, dissemination, use, and preservation of original sources, interpretive scholarship in the humanities, and other scholarly and artistic materials", is an optimal partner to an effort that seeks to advance scholarly communities on a national scale.

Project Design and Schedule of Activity

Design

Over a period of three years *Collections as Data: Part to Whole* investigators and team members propose to form two national cohorts, supported by regrants, training, and consultation. The University of Nevada Las Vegas would host cohort institutes and summative forums and administer cohort regranting. Thomas Padilla, Hannah Scates Kettler, Laurie Allen, and Stewart Varner would collectively design and coordinate the Call for Proposals and the selection of teams as well as plan and facilitate the institutes and the forums [for a draft CFP, see appendix 3]. They will also jointly assess and guide team projects and deliverables, with support from the advisory committee. The Call for Proposals asks teams to describe the sustainability, replicability, innovation, research value, ethical engagement, and engagement with complementary collections, standards, and initiatives of the projects. There will be two regranting cycles: Cohort A would be initiated in year one and Cohort B would be initiated in year two. A maximum of 12 projects would be supported in total. Regrants range from $30,000 to $80,000 per project. Example uses for funds include: buying out staff time, supporting team training, allowing for room rental and catering for meetings, and supporting conference travel to present on project activity.

Project teams in each *Collections as Data: Part to Whole* cohort would be jointly led by a librarian with senior administrative responsibilities (e.g. Associate University Librarian, Director, Head), a project lead (e.g. Director, Head, Coordinator, Librarian), and a disciplinary scholar. This leadership design choice is meant to help smooth the path to project implementation on the library side and enhance disciplinary impact on the scholarly side. Having a disciplinary scholar jointly head the project is intended to help ensure that project collections, roles, and services align with disciplinary need. Project team composition must be drawn from multiple units within the library and across the institution. Potential team members beyond the library and disciplinary faculty include but are not limited to labs, research centers, university information technology, local cultural heritage organizations, and university presses. Team compositions that reflect commitment to creative forms of partnership in pursuit of collections as data goals are strongly encouraged.

[12] Padilla, T., Allen, L., Varner, S., Potvin, S., Frost, H., and Russey Roke, E. *Always Already Computational: Collections as Data* Available at https://collectionsasdata.github.io/

[13] Data Refuge Available at https://www.datarefuge.org/

[14] Varner, S. "DocSouth Data: Open Access Data for Digital Humanities." In K. Smith (Ed) *Open Access and the Future of Academic Libraries*. Lanham, MD: Rowman and Littlefield Publishing Group.

The leads from each project team join across two meetings to form a cohort of mutually supportive professionals and scholars undertaking complementary work. The experience of the *Always Already Computational: Collections as Data* project shows that efforts to do innovative or experimental work can be challenging. Communities of practice, such as these cohorts, aim to provide vital support as challenges are pursued. Each cohort would participate in a Team Lead Institute and a summative forum [for draft agendas, see appendix 4]. The Team Lead Institute would be structured to support project teams learning from each other. The Team Lead Institute would give teams an opportunity to ask questions, get new ideas and, strengthen their project plans. Additionally, invited experts would facilitate developmental activities with team leadership that aims to increase the effectiveness of their projects. Experts would also foreground the ethical issues that data driven scholarship predicated on collections as data presents. Apart from the Team Lead Institute, project teams would also have an opportunity to consult with members of the advisory board at key stages throughout their projects. At the end of the project activities, each cohort would come together again to release a new collection as data, a collections as data implementation model, and a roles and services model. Summative forums for each cohort would be livestreamed and promoted via social media. Summative forums would serve as a public opportunity to invite broader community participation in cohort projects. They would also provide an opportunity to evaluate strengths, weaknesses, opportunities, and threats in collections as data work in combination with a consideration of how to pursue collections as data post regrant activity in the context of larger social, professional, and scholarly networks.

Higher level administrative support is vital to the success of project team proposals. For that reason, the call for proposals would require institutions to demonstrate robust commitments from library and departmental leadership to ensure that all members of the project team have the time, resources, and institutional support to take on proposed activity [for a draft CFP, see appendix 3]. For example, a statement of support would be required from the head of the library addressing what concrete measures they will take to make sure the project team has the resources it needs to complete the project. This may include proposed temporary revisions to job descriptions and the establishment of official working groups with representation at leadership council meetings. Compliance with commitments would be evaluated during the Team Lead Institute, a mid-stage project report, and a summative forum.

Project Personnel

Thomas Padilla, Digital Research Services Librarian at the University of Nevada Las Vegas, would serve as Principal Investigator. Hannah Scates Kettler, Digital Humanities Librarian at the University of Iowa would serve as Co-Principal Investigator. Laurie Allen, Director of Digital Scholarship at the University of Pennsylvania and Stewart Varner, Managing Director of the Price Lab at the University of Pennsylvania would serve as team members. The PI would rotate delegation of leadership for components of project activity across the team. The PI and team would collectively partner on selecting projects to receive regrant funds based on response to a call for proposals. The PI and team would also partner on design of the regrant opportunity, assessment of regranted projects, design of team lead institutes and summative forums and production of project reports and public presentation of the project.

Advisory Group

Dan Cohen, Vice Provost for Information Collaboration, Dean of the Libraries, and Professor of History at Northeastern University; Greg Eow, Associate Director for Collections, Massachusetts

Institute of Technology Libraries; Karen Estlund, Associate Dean for Technology and Digital Strategies, Penn State University Libraries; Trevor Munoz, Interim Director of the Maryland Institute for Technology and Assistant Dean for Digital Humanities Research, University of Maryland College Park; Barbara Rockenbach, Associate University Librarian for Research and Learning, Columbia University Libraries; Erin O'Meara, Digital Preservation Services Manager, Artefactual Systems. All members of the advisory board have confirmed their commitment to consulting with cohort project teams, per the scope of work described in this proposal – see Activity. For this labor, they will each receive a total of $1200 honoraria.

The following rationale supported advisory group member selection: **(1)** Dan Cohen provides the expertise of a Historian and a senior administrator that has held the lead administrative post at a research center (The Center for History and New Media at George Mason University), a cultural heritage collaborative (DPLA), and an academic library (Northeastern University Libraries), **(2)** Greg Eow provides the expertise of a senior library administrator responsible for a scholarly communications and collections, **(3)** Karen Estlund provides the expertise of a senior library administrator responsible for library technology and digital scholarship, **(4)** Trevor Munoz provides the expertise of a senior library administrator as well as a senior research center administrator responsible for Digital Humanities service and research, **(5)** Barbara Rockenbach provides the expertise of a senior library administrator responsible for teaching and learning services, and **(6)** Erin O'Meara provides the expertise of a senior library and archival administrator responsible for the development of digital preservation solutions.

Activity

[for schedule of completion, see appendix 2]

2018

Collections as Data: Part to Whole team (Padilla, Scates Kettler, Allen, Varner) would formalize a call for proposals (CFP). The CFP narrative and evaluation criteria would be determined in collaboration with the advisory board. Post finalization, a CFP release and promotion period would be initiated. The CFP would be open from August - October. From November - December the *Collections as Data: Part to Whole* team would make regrant project selections. The selection process would accommodate consultation with the *Collections as Data: Part to Whole* team where a project could benefit from rightsizing. For example, a proposed project may underestimate the amount of funds required to achieve planned work. In that event the *Collections as Data: Part to Whole* team will evaluate whether the project may need additional funds to accomplish goals. A November team planning meeting would set the groundwork for executing on this process and also provide for planning the Cohort A Team Lead Institute.

2019

Collections as Data: Part to Whole team would make Cohort A regrant awards in January, initiating Cohort A activity. The advisory board would consult with Cohort A project leads in the month of February. Consults with subsets of the advisory group are meant to address project team strategy on executing their project plans. Individual advisory board members would consult a maximum of 2 projects per cohort over a three year period. Each consultation period would occur within a one month span. Consult assignments would be determined by the *Collections as Data: Part to Whole* team. The Cohort A Team Lead Institute would be held at UNLV in the month of April. Following the Team Lead Institute the advisory board offers a second period of consultation to Cohort A project leads. In July, Cohort A project teams report on progress to *Collections as Data: Part to Whole* team. Progress reports account for funds spent on project activities, progress on goals, barriers to goals, and any proposed changes to project activity. By the end of August the *Collections as Data: Part to Whole* team would have evaluated progress and respond to project teams accordingly.

Collections as Data: Part to Whole team would begin the work of initiating Cohort B. In light of Cohort A experience, the CFP narrative and evaluation criteria would be reconsidered in collaboration with the advisory board. For example, Cohort A submissions may present a bias toward romance language collections which could suggest a need to specifically encourage diversity of language families in the Cohort B CFP. Post finalization, a CFP release and promotion period would be initiated. The CFP would be open from August - October. From November – December the *Collections as Data: Part to Whole* team make regrant project selections. The selection process would accommodate consultation with the *Collections as Data: Part to Whole* team where a project could benefit from rightsizing.

2020

Cohort A would engage in its public facing summative forum [for draft agenda, see appendix 4]. The summative forum accommodates participation by project team members, other than or in addition to, project team leads. In contrast to the Team Lead Institute, the summative forum is primarily geared toward public dissemination of project results. Post forum, Cohort A project teams would work in collaboration with the *Collections as Data: Part to Whole* team to formalize projects into broadly viable models. Models would contain clearly and openly documented workflows, technical stacks, roles, and services generalized to the extent that other institutions can readily use them to guide local practice. In April, Cohort A would publicly release a new collections as data and models derived from project activity.

Collections as Data: Part to Whole team would make Cohort B regrant awards in January, initiating Cohort B activity. The advisory board would consult with Cohort B project leads in the month of February. Consults with subsets of the advisory group are meant to address project team strategy on executing their project plans. Individual advisory board members would consult a maximum of 2 projects per cohort over a three year period. Each consultation period would occur within a one month span. Consult assignments would be determined by the *Collections as Data: Part to Whole* team. The Cohort B Team Lead Institute would be held at UNLV in the month of April. Following the Team Lead Institute the advisory board offers a second period of consultation to Cohort B project leads. In July, Cohort B project teams report on progress to the *Collections as Data: Part to Whole* team. Progress reports account for funds spent on project activities, progress on goals, barriers to goals, and any proposed changes to project activity. By the end of August the *Collections as Data: Part to Whole* team have evaluated progress and respond to project teams accordingly.

2021

Cohort B would engage in its public facing summative forum. The summative forum accommodates participation by project team members, other than or in addition to, project team leads. In contrast to the Team Lead Institute, the summative forum is primarily geared toward public dissemination of project results. Post forum, Cohort B project teams would work in collaboration with the *Collections as Data: Part to Whole* team formalize projects into broadly viable models. Models would contain clearly and openly documented workflows, technical stacks, roles, and services generalized to the extent that other institutions can readily use them to guide local practice In April, Cohort B would publicly release a new collections as data and models derived from project activity. In addition to the above activity, the *Collections as Data: Part to Whole* team would develop a final report to be submitted to Mellon. The final report would be submitted by end of June.

Expected Outcomes and Benefits

Each team that receives regranted funds would be responsible for producing three specific outcomes, and will take part in supporting the fourth outcome:

1. Collections as Data
2. Implementation Model(s)
3. Roles and Services Model(s)
4. Collections as Data Cohort

Collections as data

Outcomes

Collections as Data: Part to Whole would create a minimum of 12 new collections as data that exhibit high research value. Research value determinations are evidenced by regrant project leads. Factors bearing on a research value determination include but are not limited to **(1)** capacity to address clearly defined disciplinary, multidisciplinary, and/or interdisciplinary research agendas **(2)** capacity to address gaps in knowledge pertaining to how to develop and provision access to collections as data - e.g. generation of OCR data from multilingual collections recorded in atypical scripts, enabling computational study of recorded sound, provisioning user facing solutions that enable generation of statistically representative subsets of collections, facilitating computational analysis of web archives and social media data and **(3)** capacity to bring underrepresented histories to light. Significant resources have been expended on collections that document canonical versions of History. Collections as data produced in the course of this project would not add to those efforts, rather they would seek to illuminate absences and/or underdeveloped parts of the historical record. Collections as data would represent a diversity of content types, languages, and descriptive practices. In order for progress to be made with collections as data development, processes must be proven across multilingual collections and data types in a manner that extends existing cultural heritage descriptive practices. Collections as data would arise from a representative set of institutional contexts. In order for collections as data implementation to be broadly viable, work must be done to evidence the capacity of a range of institutions to take part - e.g. small institutions and large institutions, well-resourced institutions and less well-resourced institutions, single institutions and consortial collaborations.

Benefits

Collections as data that achieve these goals contribute to the development of a more diverse pool of national collections as data resources. With greater diversity comes increased ability for institutions to do collections as data work. Increased institutional ability directly impacts the range possible scholarship predicated on those collections.

Implementation Models

Outcomes

Collections as Data: Part to Whole would support creation of a minimum of 12 models that document combination of workflows, tools, and systems that enable collections as data production and access. Promising implementation models exhibit the following features: variation in institutional infrastructure and staffing to support collection implementation; experimentation with repository centric and/or non repository centric solutions; and commitments to utilization of open source technologies that interoperate with a broader open scholarly communication infrastructure.

Benefits

The primary benefit of implementation models will be their ability to support collections as data implementation at other institutions. By generating multiple models that speak to a variety of situations, *Collections as Data: Part to Whole* investigators imagine that these models would either be able to be adopted whole-cloth or guide institutions in the development of their own processes.

Roles and Services Models

Outcomes

Collections as Data: Part to Whole would support the creation of a minimum of 12 models that document paths to developing roles and services that optimally support scholarly use of collections as data. Role and service models would be diversified according to institutional resources as they relate to infrastructure and staffing to support collections as data use, as they relate to local mission - research, pedagogy, community engagement, and a range of experimental collaborations across multiple functional units within libraries, departments, labs, research centers, university information technology, local cultural heritage organizations, university presses and more.

Benefits

Because our work thus far has indicated that few if any sustainable models for roles and services exist, these models will be vitally important to institutions who are interested in supporting scholarly use of collections as data. By selecting project cohorts, in part, on the basis of institutional diversity, *Collections as Data: Part to Whole* would produce a variety of models and services to match a variety of institutions. Diversity among institutional types is crucial to ensuring that roles and services models can ably support progress at the widest possible set of institutions.

Collections as Data Cohort

Outcomes

Collections as Data: Part to Whole would support 36 people from at least 12 institutions in developing the skills and understanding the requirements of supporting collections as data development and use. Each of the scholars, administrators and project leaders would have developed a new set of skills to meet the needs of their projects, and would carry those skills into their future work. They would also be connected to one another, and would have drawn on the experiences of their colleagues across cohorts. They would also have met and connected with leaders in the field through the advisory committee, and through their participation in professional development activities. The work of the people in these institutions would be more deeply connected to the scholarly needs at their institutions.

Benefits

The professional network that would result from this project would be well positioned to propagate the ideas and practices of collections as data. At the same time, the network would form a support system for team members as they continue their work. Finally, this process of developing processes and practices in partnership with researchers and in collaboration with teams at other institutions is a potentially powerful model of leadership from within.

Sustainability

Sustainability of project resources would be addressed via deposit in Zenodo, a CERN backed data repository. Materials to be deposited into the Zenodo *Collections as Data: Part to Whole* collection include the *Collections as Data: Part to Whole* website - initially hosted on github as well as workflows, scripts, and documents produced by investigators and *Collections as Data: Part to Whole* cohort participants. Zenodo provides a retention period tied to the lifetime of the repository - currently spanning the next 20 years. In case of repository closure, Zenodo would make best efforts to integrate all content into suitable alternative institutional and/or subject based repositories.

Sustainability of institutional changes brought about by foundation funding would be assured via project emphasis on the development of broadly viable solutions. Each potential *Collections as Data: Part to Whole* project would be evaluated according to a predetermined set of sustainability criteria [for criteria, see appendix 6]. Sustainability for *Collections as Data: Part to Whole* projects would be further strengthened via consultation with the *Collections as Data: Part to Whole* investigators and advisory group at predetermined points in the project lifecycle.

Reporting

Thomas Padilla (PI) would provide the Foundation with interim and final reports according to predefined criteria, according to key project milestones and success metrics, in alignment with the schedule specified by the Foundation's award letter.