Abstract

Unlocking the Record of American Creativity: The Catalog of Copyright Entries

The New York Public Library (NYPL) requests $189,054 in funding from IMLS's National Digital Infrastructures and Initiatives Category over a one-year period, beginning July 1, 2019, in support of the creation of a database of published books from the US Copyright Office's Catalog of Copyright Entries. The proposed project will support NYPL's broader work to increase digital access to public domain books published in the 20th century, and unlock access to thousands of books in the public domain for libraries across the country. Experts in the copyright field will serve on an advisory committee.

The United States Copyright Office has the most complete and accurate collection of copyright records of ownership in the world, comprised of some 70 million records dating from 1870 through 1977. These documents record a large part of the literary, artistic and scientific production of the US and foreign countries. One set of these records, the Catalog of Copyright Entries (CCE), is the published index of the records that are critical to understanding the copyright status and ownership of copyrighted works. Spread over 450,000 pages, the CCE is divided by dates and classes of works, such as books, music, drama, and maps.

Although the CCE has been photographed and is publicly accessible online, using the digital records is cumbersome. No search function exists to reliably search the entire corpus of records; instead, users rely on analog techniques by opening multiple digitized volumes and paging through the records to find the work they are researching. In order to realize the potential impact of the records for digitizers and researchers, NYPL seeks to build a nationally-accessible dataset of the information held in our nation's copyright record for books. This effort coincides with the Library's broader work to increase digital access to books published in the 20th century. By converting these records into machine-readable text, libraries will be able to unlock access to thousands of titles in their collections. They will be able to prioritize digitization resources for collection items that are in the public domain and require no additional copyright clearance. That also means that libraries can advance knowledge by giving their users the raw materials of the public domain to create new information, insights, and understandings.

This project builds upon and advances the work NYPL has completed over the past 18 months. To date, 40,000 pages of registration records for books from the CCE are transcribed and parsed, focusing on records published between 1923 and 1969. Because the USCO maintains a searchable database that contains records from 1978 to the present, this grant would complete the dataset on book registrations from 1923 to the present. Project goals include: transcribing and parsing 23,995 pages of book registration records from the CCE into well-formed data; making the resulting dataset available to all users online; informing the library community about project and data; and gathering requirements for future front-end to the dataset.

In addition to expanding the number of works determined to be in the public domain, opening access to these records will enable new and interesting uses; creating a searchable and accessible database will greatly benefit the scholarly community interested in aspects of the creation, production, and ownership of creative works.

Unlocking the Record of American Creativity: The Catalog of Copyright Entries

The New York Public Library seeks a $189,054, one-year IMLS National Leadership grant in the National Digital Infrastructures and Initiatives Category to enhance the searchability of certain copyright records that will aid in opening access to books published after 1924. By extracting the data from nearly 24,000 pages of copyright records, this project will lay the foundation to make over 800,000 records about books and other materials more accessible to the public. The proposed activities build on work piloted by NYPL over the past 18 months and will complete a dataset of copyright records created between 1924 and today.

The National Need to Efficiently Access Copyright Records - For nearly two decades, The New York Public Library (NYPL or Library) has been part of a leading cohort of libraries, archives, and museums working to provide large-scale digital access to vast and often unique collections. Complementing the Library's goals for opening access to its collections is a focus on making scholarly publications and research materials -- critical resources for researchers, educators, students, and the general public -- widely available in digital formats. Several critical initiatives at the Library support this vision, in particular one project focused on making out-of-print and print-only books available to patrons as electronic books.

As NYPL and other libraries contemplate the digitization of 20th-century works, a core concern is copyright law. Copyright law grants to the creator of a work a set of exclusive rights for a limited time. These rights have certain exceptions and limitations that permit certain uses of a work by others without the permission of the rightsholder. Reliance on these exceptions and limitations have enabled new and interesting uses of library collections. But these exceptions alone are insufficient to advance the missions of libraries and encourage the creation of new knowledge and cultural works. By identifying the collection items that are in the public domain, libraries can make those items available to the public to be reused and repurposed.

Currently, researchers must still come to the physical library to view monographs that are only available in hard copy, or they must request materials via interlibrary loan. This project targets books that are not available electronically to our patrons from the Internet Archive, HathiTrust, or other online platform. We then prioritize this list of out-of-print books based on evidence of use or interest by our patrons. This list of prioritized books is then reviewed to determine the copyright status of each book. We have partnered with the Author's Guild to locate rightsholders of any copyright protected books on this list in order to create an ebook version that can be made available to our patrons. If the book is in the public domain, then we can immediately digitize and distribute electronic copies of the book.

While it should be easy to determine when a book or collection item has entered the public domain, US copyright law adds complexity and uncertainty as to how the digitized work can be used. Works enter the public domain when the period of copyright protection ends. For example, works published in the United States in 1923 entered the public domain on January 1, 2019.

As Congress has amended copyright law over time, a patchwork of complicated rules governs how long a work is protected by copyright. In another example, one of these rules required rightsholders of items published in the US between 1924 and 1964 to renew their copyright between the 27th and 28th year after publication. The failure to renew the copyright resulted in that work entering the public domain.

To aid in their copyright status determinations, libraries use the historical records produced by the United States Copyright Office (USCO or Copyright Office).[1] These historical records are vast; there are 70 million historical records in a variety of formats that begin in 1870 when copyright was centralized under the Library of Congress and run through 1978, the beginning of the modern recordation system. Unfortunately, the paper records require specialized knowledge and skills to use.

Our proposed project is focused on a portion of the Catalog of Copyright Entries (CCE), a subset of the 70 million historical records. The CCE is comprised of about 450,000 pages spread between 674 volumes and are printed compilations of brief registration and renewal records. The CCE is divided by classes of works, such as books, periodicals, music, drama, maps, and photographs. Records about books make up more than a third of the 450,000 pages of the CCE. These records were published by the Copyright Office at regular intervals, ranging in length from semi-weekly to semi-annually.

Libraries use the CCE to assist their copyright status determinations and clearance efforts. First, records are accessed for books subject to renewal requirements. Then, registration records are reviewed to determine when a book was first registered with the USCO. If a book was registered before 1924, then the book is likely in the public domain. Finally, if the book is not in the public domain and we want to secure permission from the rightsholder to make uses of the book, we use the CCE to generate leads to find the current rightsholder. For example, if the registration record was filed by someone other than the author, we now have an additional name to use in our research.

In general, the CCE is difficult to search and requires some expertise to use. To find records about an item, we need to know how the item was originally categorized (e.g., book, photograph, print, pamphlet, contribution to a periodical, etc.) and the year the item was registered with the USCO. If a researcher is looking for registration records pertaining to one book, multiple volumes of the CCE will likely have to be consulted in order to find the relevant records.

The USCO contracted with Internet Archive to photograph the CCE and made it available to the public in 2012. Internet Archive (IA) subjected the digital photos to standard optical character recognition (OCR) to generate machine-readable transcription, but the level of transcription accuracy is insufficient for detailed copyright research.[2] In addition, there is no reliable way to

[1] The project director is in regular contact with the USCO and has briefed the Acting Register on this project. The USCO does not have plans to address the scope of work proposed in this proposal.

[2] Very accurate transcription and OCR is critical for a researcher to feel confident that a “no results” query is because the record does not exist in the data; in many cases, inaccurate transcription have resulted in typos that hide the record from the query. Libraries are often trying to prove a negative; namely, that no renewal was filed.

simultaneously search across all 674 digitized volumes. Without accurate transcription, IA's digitized CCE must be used similar to the analog version—the electronic pages must be 'turned' in order to find the records sought.

The CCE represents one of the best records of American creativity and documents a significant part of the literary, musical, artistic and scientific production of the United States from 1891 through 1977. Unlocking the data from these records is critical for libraries determining the copyright status of items in their collections, but will also serve as a rich dataset for researchers. Making these records more computationally accessible will enable the study of creative works in new ways.

Previous and Related Projects - The proposed project builds upon the work of many researchers, scholars, and practitioners; the projects described below have made significant advances in extracting or using data from the CCE.

Project Gutenberg and Distributed Proofreaders (PGDP) is one of the first collaborations to transcribe the CCE. PGDP volunteers transcribe the renewal records for books and publish the transcription online. The proposed project differs in methodology from the PGDP project in that the data transcription process will be undertaken by a contractor with considerable expertise, leaving less room for error. NYPL's project will also transform the transcription into well-formed records using a parsing process more fully described in 'Project Details'.

This project also builds upon the Stanford Copyright Renewal Database (SCRD). The SCRD allows users to search the CCE renewal records for books. The data comes primarily from the output of the PGDP project, but with some important advances. The SCRD project parsed the raw transcription from the PGDP project and created well-formed records. Parsing the data into the appropriate fields allows users of the SCRD to search specific fields and facet results. NYPL's proposed project adopts many of the methodological advances of the SCRD and extends them to registration records.

Finally, this project complements the University of Michigan's IMLS-funded Copyright Review Management System (CRMS). CRMS is the first successful attempt to determine the copyright status of books at scale utilizing data from the CCE. CRMS relies on the SCRD to allow reviewers to find the presence or absence of copyright renewals for books that were digitized primarily by Google as part of the Google Books project. CRMS does not determine the copyright status of books that are not yet in the HathiTrust corpus. Instead of reactively determining the copyright status of books, the proposed project would enable libraries to identify the books that are likely to be in the public domain. This approach would allow libraries to devote their limited resources to digitizing books that can be made widely available.

NYPL's project shares the objectives of the University of Pennsylvania Libraries' IMLS-funded work documenting the renewal records for serial publications. Both projects use the CCE to achieve greater clarity of the public domain; but UPenn's focus is on renewal records for serials, while NYPL's focus is on registration records for books. UPenn's project will interpret data from the CCE to produce a list of dates on which the last renewal could be found, while NYPL's project will produce transcribed and parsed data that others will be able to utilize.

Finally, this project builds upon and advances the scholarship of copyright records. Before CRMS, the scholarship around the number of books in the public domain relied on noisy and incomplete data from WorldCat. As portions of the CCE have been extracted and used by projects like CRMS, scholars have attempted to estimate the size of the public domain by analyzing the rate at which copyrights were renewed by rightsholders.[3] The majority of these projects ultimately demonstrate the need for a complete dataset and offer a tantalizing glimpse of how the data produced can be used to further scholarship. This project will give researchers the largest, most accurate dataset extracted from the CCE to date and will advance scholarship.

This proposed project will have a significant national impact on the capacity of libraries and archives to provide access to digitized collections. By leveraging emerging transcription and parsing techniques, this project reflects IMLS's focus on expanding digital cultural heritage capacities, as noted in the *National Digital Infrastructures and Initiatives: A Report on the 2017 NDP at Three Forum*. Within this focus area, the report highlights efforts to enable computational use of collections. An essential requirement for computational use of collections is to extract data from digital images in a reliable and accurate way. This proposed project will fund the data extraction from nearly 24,000 pages of records and will complete the dataset about book registrations from 1923 to the modern records that began in 1978. That means researchers interested in American publishing will have a rich dataset of over 2 million records at the conclusion of this project.

Project Design: Goals and Outcomes - Overall outcomes for our work with the CCE are multifaceted. Libraries and other cultural heritage institutions should be able to conduct simple, comprehensive and reliable searches over the public records of the USCO so that digitizers can make informed decisions about digitization projects. The searches should produce relevant data more efficiently than the processes available today. Another goal is for researchers to be able to conduct deep analysis of the history of American creativity represented in these records. This kind of research today relies on data that is noisy and difficult to work with. Policymakers would also benefit from the kinds of empirical research enabled through this project.

This project builds upon and advances the work we have completed over the past 18 months. To date, we have transcribed and parsed 40,000 pages of registration records for books from the CCE. The work to date has focused on records published between 1923 and 1969. Because the USCO maintains a searchable database that contains records from 1978 to the present, this grant would complete the dataset on book registrations from 1923 to the present.

[3] For example, see Wilkins, J. (2011) "Bibliographic Indeterminacy and the Scale of Problems and Opportunities of 'Rights' in Digital Collection Building," Ruminations (CLIR), http://www.clir.org/pubs/ruminations/index.html/01wilkin/wilkin.html and Carlstone, J., Stein, A., Norman, M., & Wilkin, J. (2018) "Copyright Renewal of U.S. Books Published in 1932: Re-analyzing Ringer's Study to Determine a More Accurate Renewal Rate for Books." College & Research Libraries, 79(5), 697. https://doi.org/10.5860/crl.79.5.697.

The goals of this project are to:

- Transcribe and parse 23,995 pages of book registration records from the CCE into well-formed data,
- Make the resulting dataset available to all users online,
- Inform the library community about project and data, and
- Gather requirements for future front-end to the dataset.

Transcribe and Parse CCE Pages: We will issue a Request for Process seeking bids from vendors to transcribe and parse 23,995 pages of book registration records. These nearly 24,000 pages of records were published by the United States Copyright Office between 1970 and 1977 in *Part 1: Books and Pamphlets of the CCE*. The vendor will use the digital photographs taken by the Internet Archive in its work with the USCO to make the records more accessible.

The accuracy of the transcription is essential to creating a reliable dataset. The RFP will require the transcription for all data to be 99.9 percent accurate. That level of accuracy means that 999 characters out of every 1,000 will have a high confidence score. Each record in the CCE includes a registration number (e.g., “A149259”) that is an important identifier created by the USCO. Because the registration number is used as a way to uniquely identify a record, we will require the vendor achieve a higher level of accuracy of 99.99 percent. That means 9,999 transcribed characters out of every 10,000 will have a high confidence score. To achieve these high levels of accuracy, the vendor will likely need to employ a mix of optical character recognition technology and human review. Without this mix of OCR and human review, OCR alone has been estimated to achieve accuracy rates between 75 and 95 percent, far below the accuracy required to create reliable data. Our targeted accuracy levels strike an efficient balance between accuracy and costs; achieving an even higher level of accuracy would require more human intervention, putting the project costs out of reach.

The transcription process will also record the location of each character in the image of the page. This location data aids in parsing the records and may eventually be used in a future database front-end for this data. The location data may enable a front-end to display the transcribed and parsed data alongside the corresponding image of the page so that a user can validate that the data accurately reflects the CCE.

After the data has been transcribed, the vendor will parse the records into defined fields. These fields were developed during our work on the initial 40,000 pages of book registration records, during which the list and structure of data fields was continually refined, resulting in our Document Type Definition (DTD). The vendor will use our existing DTD to parse the data into its elements and attributes. By parsing the data, a user will be able to facet their searches on specific fields in a future front-end. For example, a user attempting to find a work authored by Albert Einstein may want to facet their search to the author field to avoid seeing records of books with Einstein’s name in the title.

After transcription and parsing is complete, the vendor will deliver their data extraction. For the transcription work, the vendor will deliver files in ALTO format, a standardized XML format to store layout and content information. For the parsing work, the vendor will deliver files in an XML format that is structured according to our DTD.

Open Dataset: The CCE is a public record and belongs to the nation; NYPL is committed to making the data produced by this project freely available and accessible without restriction. After the data delivered by the vendor is validated, we will post the raw data onto the project's GitHub repository (https://github.com/NYPL/catalog_of_copyright_entries_project). All of the project data will be available online through Github, which means the data files produced by our vendor will be posted openly, along with metadata describing each file. Making this data available without restriction also reflects our goal of enabling new and interesting uses of the project data.

Increasing Community Awareness: To inform the library community about this project and the CCE data, the project lead will put forward conference proposals at approximately five conferences as well as at additional meetings. This outreach is more fully described on page 9 below.

Gathering Requirements for Future Front-End: The proposed project will also take an important step towards making the data available in a more user-friendly way. Using both physical and virtual meetings, input from potential users will be collected and documented to develop the requirements and assess the costs for a searchable front-end interface for this dataset. Target audiences for these meetings are data technologists, copyright experts, and researchers who are likely to use the CCE data.

By collecting these requirements from likely users, we will be able to prioritize feature development when funding for a front-end has been secured. This will allow us to build a clear roadmap for development of the front-end. We will post the requirements we have gathered on our project page and invite feedback. All project documentation created by NYPL, including the list of requirements for a front-end to the database, will be made available under a Creative Commons Attribution 4.0 International license. This license will permit anyone to easily reuse and adapt project documentation.

Project Assumptions and Risks - As with any large project, there are assumptions that underlie this proposal and risks associated with taking on this work. There are specific risks related to working with a vendor. Having already piloted the data extraction process, we have fine-tuned our requirements to transcribe and parse the data. Our experience has also informed our expectations for the vendor costs; we have learned to mitigate the risk that bids from vendors will be higher than expected by using multiple pricing models to arrive at an expected cost.

We are prepared to address unforeseen risks by recruiting an advisory committee of leaders in the field. The advisory committee has deep experience, both academic and practical, in working with the CCE. Should a question arise that challenges the methodology for parsing data into fields, we feel confident that our advisors can help resolve the question in a creative and informed way.

Finally, legal risk is mitigated because while this project facilitates access to the data from the CCE, it does not make determinations as to copyright status of individual items. We will take steps to inform users of how the data was created and that it is impossible to guarantee that every character is perfectly transcribed. This way, libraries are empowered to make their own informed decisions.

Project Sequence - This is a one-year project, expected to commence on July 1, 2019. The grant will begin with the issuance of an RFP around the transcription and parsing work; this work is expected to begin in the third month and end in the ninth. Data will return from the vendor throughout months four through nine; data will be posted to the GitHub repository as it is delivered. Community outreach and engagement via conference attendance will take place throughout the grant, as will regular communication with the advisory committee. The project lead and advisory committee will identify the target audience for the requirements gathering phase of the project at the project's mid-year and seek to convene a meeting of this audience, likely at the University of Michigan, in months nine or ten. Requirements for the front-end will be published and disseminated broadly by the end of the grant.

Resources and Staffing - Since submitting the preliminary proposal, NYPL has determined that a well-conceived front-end for this project must be informed by feedback from the larger library community. Consequently, the travel budget request has increased by $10,000 in order to accommodate this feedback-collection process. We also adjusted our selection of CCE pages for this proposal as a result of our analysis of the results of the pilot. The selected tranche is more complex and, therefore, more expensive to convert. An additional $14,000 is requested to support this work.

Greg Cram, Associate Director for Copyright and Information Policy at NYPL will serve as the Project Director. Greg is responsible for developing and implementing policies and practices around the use of the Library's collections, both online and in physical spaces. He will spend approximately 18 percent of his time managing this project, including RFP process and vendor relationship, engaging with the advisory committee and larger external audience around the project, and supervising Kiowa Hammons, Rights Clearance Coordinator, who will work with Greg to validate the data produced by the vendor and aid in the coordination of meetings related to the project.

Copyright experts at other institutions will serve as the project's advisory committee, including:

- Robert Brauneis, Professor of Law and Co-Director of the Intellectual Property Law Program, George Washington University School of Law
- Mike Furlough, Executive Director, HathiTrust
- Melissa Smith Levine, Director, Copyright Office, University of Michigan
- John Mark Ockerbloom, Digital Library Strategist & Metadata Architect at the University of Pennsylvania
- Zvi Rosen, Visiting Scholar and Professorial Lecturer in Law at George Washington University School of Law

Group members regularly use, consult, or are consulted about the CCE and will lend their considerable knowledge and prior experience to the project. Letters of commitment from this group are appended to the proposal as supporting documents.

The bulk of financial resources requested for this project will be spent on contracted transcription and parsing of the CCE records ($155,000). Funds will also be used to convene a meeting of likely users from the library and research communities to develop requirements for a future front-end ($9,513). Additional travel funds will support outreach and promotion of the data created by this project ($7,766). Staff time funded by this project covers the review of vendor-created data to ensure that accuracy requirements are being met ($11,648); funding will also support meeting materials ($2,000) and indirect costs ($3,127). The Project Director will contribute approximately 18 percent of his time to the project.

Target Audience and Metrics of Success - In this initial stage, the target audience for the data produced by this project will be libraries and other cultural heritage institutions digitizing their collections. Libraries and other organizations digitizing their collections are making copyright status determinations and data from the CCE will aid their efforts. In order for our users to benefit from the data as soon as possible, the raw data will be available on the project's GitHub repository as soon it is delivered and validated.

As the raw data requires experience in manipulating large datasets, this project will lay the foundation for libraries that lack that experience to benefit from this data. By meeting with libraries and other digitizers to discuss how they would access this data, we will collect the requirements that will guide the future building of a front-end for the data. The requirements will be posted online, allowing other potential users to provide feedback on the prioritization of features. NYPL will also be a consumer of the data, so we have made some assumptions about how libraries would want to use the data. For example, we are parsing the data to aid faceted searching because we anticipate libraries expressing this as a requirement for a front-end.

Beyond libraries and digitizers, this project has been designed with researchers in mind, who will likely want to compare records created over time. Unfortunately, the format of and information contained in the CCE changed over decades of publication. To address this issue, we have adopted a data schema in the DTD that will smooth out these changes. That means the tag for the "author" attribute in the 1930 CCE volume will be the same tag as the "author" attribute in the 1977 CCE volume. By parsing records into consistent data fields, researchers will be able to analyze the data without having to crosswalk data schemas for hundreds of volumes of the CCE.

Success for this project will be measured against the four goals described on page 5. If we transcribe and parse the book registration records from the 1970-1977 CCE at the level of accuracy described above, that goal will have been achieved. We will complete our goal of having the dataset open without restriction by posting the raw data online and not adding any restrictions on the use of the data. If we identify and convene likely library users, then collect their front-end requirements and make them available online, we will have achieved our goal to make progress on an eventual front-end.

A longer-term indicator of this project's success will be use of the data by digitizers. To measure our progress for this key indicator, we will ask users of the data to inform us how the extracted data is advancing their projects. Because we want the data to be used without restriction, users will not be required to provide their answers as a condition of access to the data. We expect that the promotion of this data at conference presentations by the project lead will increase awareness of the data, which will increase use of the data.

Sharing Project Findings with the Library Community - Project deliverables will be robustly disseminated through a variety of activities in order to raise awareness of the project. The Project Director will attend national conferences and intends to deliver presentations and poster sessions about the project. These conferences include the American Library Association's Annual Conference, the Public Library Association's Annual Conference, the Coalition for Networked Information Membership meeting, the Digital Library Federation Forum, the Kraemer Copyright Conference, and the annual University Information Policy Officer Meeting. These conferences were selected to target copyright specialists, libraries engaged in digitization of collections, and public library professionals. We will also use blog posts and social media to communicate updates and project findings to the library community.

National Impact - This project will have immediate impacts upon its conclusion and will transform library practices around the digitization of books. Instead of selecting the book before conducting copyright research, libraries will be able to produce a list of books that are likely to be in the public domain, then make digitization selections. By comparing the renewal records from the SCRD with the registration data produced by this project, libraries will, for the first time, be able to see all books that were registered but not renewed. This approach is more efficient for libraries interested in maximizing access to their digitized resources because it saves searches for books that turn out to be still protected by copyright.

This project will also provide fodder for researchers interested in the American publishing. By completing the book registration CCE from 1970 to 1977, researchers will have book registrations from 1923 to current day available to review. That will enable researchers to answer questions about the output of the American publishing industry and how it might have changed over time. Researchers will finally be able to get empirical answers to questions that they have been asking that will inform public policy.

Although this project proposal is focused on book registrations, outside of the project period, we aspire to extract data from all 674 volumes of the CCE. At full scale and with a user-friendly front-end, extracting all of the registration and renewal records for works of all types from the CCE will dramatically change library digitization practices. Libraries will be able to determine the copyright status of published collection items of all types in a simple and quick process. No longer will libraries need to have expertise in using the records to ensure they are looking in the correct section in the correct volume for the correct year. Instead, they'll be able to search all 674 volumes in a single search transaction. That means more collection items will be identified as public domain works, greatly expanding the number of digitized items available to the public to reuse and repurpose.

As more fully described in the Digital Product Form, the data from the CCE produced by this project will be available without restriction in an openly-accessible repository. The data is not protected by copyright and is a public record, so the data can be used by anyone. We will not add restrictions or conditions on the use of the data.

We will also produce and make available documentation about the data structure and our extraction process so that the cultural heritage community can adapt the data for their own purposes. They could also use the documentation to conduct their own CCE data extraction project by replicating our methodology so that the two datasets could be merged without significant data massaging or cleaning.

Sustaining the Impact of Transformed Records - NYPL is committed to advancing knowledge by making our collections broadly available and accessible. The proposed project is a key component toward achieving that goal, and the resulting dataset will have a strong, field-wide impact long after the grant period has ended. The Library is committed to maintaining public access to the data. GitHub serves as the repository for the project data, in part, because of the low hosting costs involved. By making the data openly available for download and use, libraries will be able to leverage the data for their own purposes.

For example, those interested in linked data may use the data to match OCLC numbers with copyright registration numbers, further enhancing records in their catalogs. In addition to making the raw data available, this project will create a model for the conversion of the rest of the CCE, outside of books. NYPL will work with the project advisory committee to build partnerships that identify new users and uses of the project data.

NYPL also expects that a strong user base will form around this data because of the national trend of increased digitization of collections. Cultural heritage organizations that want to digitize their collections need easy and efficient access to copyright records to make informed digitization selections. The data from the CCE is also critical to cultural heritage institutions that are increasingly adopting standardized rights statements as a way to communicate the copyright status of digitized items to users. This project will help organizations achieve their goals to encourage the creation of new information, new aesthetics, and new insights and understandings to advance the progress of knowledge.

Proposed Schedule of Completion with IMLS Funds

	2019						2020					
	July	Aug	Sept	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun
Transcription and Parsing												
Create and Issue RFP		■										
Award Contract			■									
Vendor Transcribes, Parses and Delivers Data			■	■	■	■	■	■	■			
Posting Data												
Data posted to GitHub repository				■	■	■	■	■	■			
Community Outreach												
Announce grant award & describe progress to date	■											
Advisory Committee conference calls			■			■			■			■
Announce completion of parsing and transcription of tranche										■		
Present at library, museum, archives, and copyright conferences about	■	■	■	■	■	■	■	■	■	■	■	■
Requirements Gathering												
Identify members of target audience for front-end interface; send invites						■						
Convene meeting of target audience to gather reqs for database									■			
Publicly solicit feedback for requirements										■	■	
Publish requirements for front-end												■

DIGITAL PRODUCT FORM

Introduction

The Institute of Museum and Library Services (IMLS) is committed to expanding public access to federally funded digital products (e.g., digital content, resources, assets, software, and datasets). The products you create with IMLS funding require careful stewardship to protect and enhance their value, and they should be freely and readily available for use and re-use by libraries, archives, museums, and the public. Because technology is dynamic and because we do not want to inhibit innovation, we do not want to prescribe set standards and practices that could become quickly outdated. Instead, we ask that you answer questions that address specific aspects of creating and managing digital products. Like all components of your IMLS application, your answers will be used by IMLS staff and by expert peer reviewers to evaluate your application, and they will be important in determining whether your project will be funded.

Instructions

All applications must include a Digital Product Form.

□ Please check here if you have reviewed Parts I, II, III, and IV below and you have determined that your proposal does NOT involve the creation of digital products (i.e., digital content, resources, assets, software, or datasets). You must still submit this Digital Product Form with your proposal even if you check this box, because this Digital Product Form is a Required Document.

If you ARE creating digital products, you must provide answers to the questions in Part I. In addition, you must also complete at least one of the subsequent sections. If you intend to create or collect digital content, resources, or assets, complete Part II. If you intend to develop software, complete Part III. If you intend to create a dataset, complete Part IV.

Part I: Intellectual Property Rights and Permissions

A.1 What will be the intellectual property status of the digital products (content, resources, assets, software, or datasets) you intend to create? Who will hold the copyright(s)? How will you explain property rights and permissions to potential users (for example, by assigning a non-restrictive license such as BSD, GNU, MIT, or Creative Commons to the product)? Explain and justify your licensing selections.

The central digital product created by this grant will be transcribed and parsed records of the Catalog of Copyright Entries. The CCE was published by the United States Copyright Office and is, therefore, in the public domain in the United States under 17 U.S.C. §105. All other documentation created by NYPL, including requirements for a front-end to the database for converted CCE records, will be made available under a Creative Commons Attribution 4.0 International license. This license will permit anyone to reuse and adapt the material. It will also ensure the source of the documentation is clearly indicated to verify the content. Because the products described in this paragraph are intended to be complete documents, the attribution requirement is not burdensome like it is for data that is meant to be interoperable.

A.2 What ownership rights will your organization assert over the new digital products and what conditions will you impose on access and use? Explain and justify any terms of access and conditions of use and detail how you will notify potential users about relevant terms or conditions.

NYPL is not creating a copyrightable derivative of the CCE, so NYPL would not own any rights in the transcribed and parsed data. NYPL will not impose any contractual restrictions on the use of the converted CCE records because NYPL believes these public records belong to the country. NYPL will label the data using the “No Copyright - United States” RightsStatements.org statement and will include a short description of why this statement was used in the project repository.

A. 3 If you will create any products that may involve privacy concerns, require obtaining permissions or rights, or raise any cultural sensitivities, describe the issues and how you plan to address them.

NYPL does not expect that any such issues to arise in the project. The CCE is a public record that was widely published and held in many libraries. As stated above, the CCE is in the public domain in the United States and does not require obtaining permissions or rights to convert or use.

Part II: Projects Creating or Collecting Digital Content, Resources, or Assets

A. Creating or Collecting New Digital Content, Resources, or Assets

A.1 Describe the digital content, resources, or assets you will create or collect, the quantities of each type, and the format(s) you will use.

Formal project documentation will be published as PDFs. Other project communication will be published as blog posts on NYPL's website and on the project's Github repository that is openly available online.

For the CCE conversion, NYPL will produce two XML files for each CCE year that is converted. One file will contain the transcription and will be delivered in ALTO format, a standardized XML format to store layout and content information. The other XML file will contain the parsing work and be structured according to our DTD. We estimate that fourteen XML files will be created reflecting the seven years of records, but the vendor may choose to break up a year into parts, multiplying the number of files produced.

A.2 List the equipment, software, and supplies that you will use to create the content, resources, or assets, or the name of the service provider that will perform the work.

Formal project documentation will be created using Microsoft Word and Google Documents and published as PDFs.

We will engage a contractor to convert the CCE records into XML files conforming to a DTD created for this project. The contractor will be selected after the completion of a Request for Proposal. The XML files will be made available on the project's Github repository, as will project documentation.

A.3 List all the digital file formats (e.g., XML, TIFF, MPEG) you plan to use, along with the relevant information about the appropriate quality standards (e.g., resolution, sampling rate, or pixel dimensions).

Formal product documentation will be published as PDFs that meet or exceed NYPL's accessibility standards.

The converted CCE records will be made available as XML files, as described in A.1. The contractor will be required to achieve an accuracy level of 99.9 percent for the records and a 99.99 percent accuracy for the registration number that appears within each record. The contractor will also be required to parse the data in accordance with project's DTD.

B. Workflow and Asset Maintenance/Preservation

B.1 Describe your quality control plan. How will you monitor and evaluate your workflow and products?

The contractor will be required to meet certain accuracy levels and ensure that they meet those levels. NYPL will monitor the accuracy of the transcription by manually spot-checking transcription and analyzing the data for errors like duplicative registration numbers. NYPL will also validate the XML files related to parsing to make sure they are valid.

B.2 Describe your plan for preserving and maintaining digital assets during and after the award period of performance. Your plan may address storage systems, shared repositories, technical documentation, migration planning, and commitment of organizational funding for these purposes. Please note: You may charge the federal award before closeout for the costs of publication or sharing of research results if the costs are not incurred during the period of performance of the federal award (see 2 C.F.R. § 200.461).

The project files and converted CCE data will be made available using a Github repository.

C. Metadata

C.1 Describe how you will produce any and all technical, descriptive, administrative, or preservation metadata. Specify which standards you will use for the metadata structure (e.g., MARC, Dublin Core, Encoded Archival Description, PBCore, PREMIS) and metadata content (e.g., thesauri).

As described above, two XML files will be created for each volume or part of the CCE: 1) one in ALTO format preserving the results of the initial OCR process and connecting specific entries to coordinates in the scanned pages; 2) one in a custom DTD created for this project. The format of the CCE volumes is unique but simple, and the resulting DTD is relatively trivial and easy to understand. The current DTD is available in the project Github repository.

C.2 Explain your strategy for preserving and maintaining metadata created or collected during and after the award period of performance.

The project data will be hosted on Github for the foreseeable future and in whatever distributed copies or clones are made by project staff or the public. NYPL will be exploring other options to preserve the data, independent of Github.

C.3 Explain what metadata sharing and/or other strategies you will use to facilitate widespread discovery and use of the digital content, resources, or assets created during your project (e.g., an API [Application Programming Interface], contributions to a digital platform, or other ways you might enable batch queries and retrieval of metadata).

The project data will be publicly available on Github which is the most widely used platform for sharing code and, increasingly, data.

D. Access and Use

D.1 Describe how you will make the digital content, resources, or assets available to the public. Include details such as the delivery strategy (e.g., openly available online, available to specified audiences) and underlying hardware/software platforms and infrastructure (e.g., specific digital repository software or leased services, accessibility via standard web browsers, requirements for special software tools in order to use the content).

The project files and converted CCE data will be made available on the project's Github repository that is openly available online. The XML files that contain the CCE data do not require specialized software tools to access or use because XML is an open standard.

D.2 Provide the name(s) and URL(s) (Uniform Resource Locator) for any examples of previous digital content, resources, or assets your organization has created.

The XML files created from the pilot discussed in the narrative are available at https://github.com/NYPL/catalog_of_copyright_entries_project/tree/master/xml. The project's Github repository is available at https://github.com/NYPL/catalog_of_copyright_entries_project.

Beyond this project, NYPL has created myriad digital resources. Other digital resources that NYPL has created include:
Digital Collections http://digitalcollections.nypl.org/
Simply e: The Library E-Reader App http://www.librarysimplified.org/
What's On The Menu http://menus.nypl.org/

Part III. Projects Developing Software

A. General Information

A.1 Describe the software you intend to create, including a summary of the major functions it will perform and the intended primary audience(s) it will serve.

Not applicable. Although this project collects requirements for software, it is beyond the scope of this proposal to create software.

A.2 List other existing software that wholly or partially performs the same functions, and explain how the software you intend to create is different, and justify why those differences are significant and necessary.

N/A

B. Technical Information

B.1 List the programming languages, platforms, software, or other applications you will use to create your software and explain why you chose them.

N/A

B.2 Describe how the software you intend to create will extend or interoperate with relevant existing software.

N/A

B.3 Describe any underlying additional software or system dependencies necessary to run the software you intend to create.

N/A

B.4 Describe the processes you will use for development, documentation, and for maintaining and updating documentation for users of the software.

N/A

B.5 Provide the name(s) and URL(s) for examples of any previous software your organization has created.

N/A

C. Access and Use

C.1 We expect applicants seeking federal funds for software to develop and release these products under open-source licenses to maximize access and promote reuse. What ownership rights will your organization assert over the software you intend to create, and what conditions will you impose on its access and use? Identify and explain the license under which you will release source code for the software you develop (e.g., BSD, GNU, or MIT software licenses). Explain and justify any prohibitive terms or conditions of use or access and detail how you will notify potential users about relevant terms and conditions.

N/A

C.2 Describe how you will make the software and source code available to the public and/or its intended users.

N/A

C.3 Identify where you will deposit the source code for the software you intend to develop:

Name of publicly accessible source code repository:

N/A

URL:

N/A

Part IV: Projects Creating Datasets

A.1 Identify the type of data you plan to collect or generate, and the purpose or intended use to which you expect it to be put. Describe the method(s) you will use and the approximate dates or intervals at which you will collect or generate it.

Not applicable. This project converts existing records, it does not create sets of new data.

A.2 Does the proposed data collection or research activity require approval by any internal review panel or institutional review board (IRB)? If so, has the proposed research activity been approved? If not, what is your plan for securing approval?

N/A

A.3 Will you collect any personally identifiable information (PII), confidential information (e.g., trade secrets), or proprietary information? If so, detail the specific steps you will take to protect such information while you prepare the data files for public release (e.g., data anonymization, data suppression PII, or synthetic data).

N/A

A.4 If you will collect additional documentation, such as consent agreements, along with the data, describe plans for preserving the documentation and ensuring that its relationship to the collected data is maintained.

N/A

A.5 What methods will you use to collect or generate the data? Provide details about any technical requirements or dependencies that would be necessary for understanding, retrieving, displaying, or processing the dataset(s).

N/A

A.6 What documentation (e.g., data documentation, codebooks) will you capture or create along with the dataset(s)? Where will the documentation be stored and in what format(s)? How will you permanently associate and manage the documentation with the dataset(s) it describes?

N/A

A.7 What is your plan for archiving, managing, and disseminating data after the completion of the award-funded project?

N/A

A.8 Identify where you will deposit the dataset(s):

Name of repository:

N/A

URL:

N/AN/A

A.9 When and how frequently will you review this data management plan? How will the implementation be monitored?

N/A