Unlocking the Record of American Creativity

Summary - The New York Public Library (NYPL) requests $162,000 in funding (total project costs totaling $188,000) from IMLS's National Digital Infrastructure program 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.

National Need - 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. Researchers must know which of the 660 volumes to search within for records about a particular work, or risk missing the record for which they are searching. In order to realize the potential impact of the records for 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, aesthetics, insights, and understandings.

Extracting the data from the CCE is critical to building a searchable and accessible database. Between 1923 and 1964, copyright owners had to actively renew their copyright; failure to renew meant that the work fell into the public domain. The Library's proposed project will focus on the registration records for books published during this time period, so that books in the public domain can be easily identified. The project will compare the registration records of books published in the U.S. with the renewal records already transcribed and parsed by Stanford University in order to identify books that may be in the public domain and which, as such, can be made available immediately upon digitization, online, to anyone in the United States in full text. 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.

Project Design - NYPL proposes to create a dataset that enables accurate searching of these historical records so that users will be able to retrieve all of the records related to a copyrighted work. In July 2018, NYPL completed a pilot project in which the Library transcribed and parsed 10,000 pages of the CCE, focusing on a sampling of book registration records published between 1923 and 1964. The pilot helped determine the necessary costs to build the dataset and refined the process to extract the records more efficiently. Using the data produced during this pilot phase, NYPL estimates that between 25 percent and 50 percent of scholarly books published between 1923 and 1964 may already be public domain.

The CCE has already been imaged by the Internet Archive and is freely accessible and available online in a variety of formats. To extract data from these images, NYPL will engage a contractor to transcribe the text using OCR and by hand to a high level of accuracy and record the coordinate data for each character. Because the records are often brief and include unique identification numbers, accuracy of the transcription is important. When issuing an RFP for the contractor, the Library will test the ability of potential contractors to return results with a 99.9 percent degree of accuracy.

The record text will then be parsed by the contractor into specific fields to facilitate faceted searching of the data. CCE records are typically in the form of a block of text that lacks clear labeling; this means a user must know the implicit structure of each record to understand which words are the author, title, date of publication or other labels. By parsing the data, a user will be able to facet their searches on specific fields. 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. NYPL is committed to making the data produced by this project freely available and without restriction. All of the project data will be available online through Github, which means the data files produced by NYPL's contractor will be posted openly, along with metadata describing each file. Making this data available without restriction also reflects NYPL's goal of enabling new and interesting uses of the project data.

Impact and Outcomes **-** It is increasingly important that libraries across the country are able to leverage public domain materials on behalf of their users. NYPL is committed to improving digital access to the broad universe of published books, and regularly works to clear copyright for digital items in its collections including books, images, and recordings. Published books in particular are critical resources for researchers, educators, students, and the general public, however they are currently largely inaccessible in digital format. This project will help identify tens of thousands of works for which copyright was never renewed which can now be made available online without additional cost or negotiation. Because many of these works have already been digitized, they can be made available immediately. This project will also support the ability of libraries nationwide to conduct a similar risk analysis in determining whether materials they want to reproduce are still under copyright. Project deliverables include:

- Transcribed and parsed data files for 24,475 pages of book records, ensuring that the most important tranche of entries will be available
- Raw data produced through the grant will be available through Github, alongside data extracted during the pilot phase and future phases of the CCE data extraction project

In addition to helping libraries identify books in the public domain, converting the CCE into well-structured data will provide a rich dataset for researchers. The broad coverage of the CCE provides the nation with a bibliographic record of unique scope. The project team has received positive feedback that having this data available in well-structured form will provide fertile ground for research in the creation, production, and ownership of creative works. For example, a linguist might employ sentiment analysis to the data produced by this project to research whether periods of global conflict can be discerned using solely the titles of books.

Staff and Partners **-** Project lead is Greg Cram, who has served as NYPL's Associate Director of Copyright and Information Policy since 2011. As such, he is responsible for developing and implementing policies and practices around the use of the Library's collections, both online and in physical spaces. He partners regularly with experts at other institutions around copyright issues, including HathiTrust, the University of Michigan, the University of Pennsylvania, and George Washington University Law School, all of whom will form the advisory committee for this project, and lend their knowledge and prior experience with the CCE.

Estimated Project Budget **–** The estimated project budget is $188,500, of which $162,000 is requested through the IMLS National Leadership program. Cost share includes $26,500 in staff salaries, fringe benefits, and indirect costs in support of the project; these costs include approximately 18% of the project lead's time. IMLS funding will support vendor digitization ($141,000), administrative support ($2,000), and travel and outreach ($7,000), as well as indirect costs (rate currently under negotiation, but estimated at $12,000).