University of Pennsylvania Libraries

**Opening access to mid-20th century serials: Sparks grant proposal abstract**

The University of Pennsylvania Libraries seeks a $25,000 1-year IMLS Sparks grant to provide and demonstrate tools and data for opening access to the large body of scholarly and general serial literature from 1923-1989 that is in the public domain, but not yet freely accessible online.

Libraries hold many serials published after 1922 that are in the public domain, and contain unique and valuable information for researchers. Most have not gone online to date, due to uncertainty and difficulty of establishing that they are in fact free of copyright restrictions. Led by John Mark Ockerbloom, with Joseph Zucca and other collaborators at Penn, we will produce data and procedures that will make it substantially easier for libraries of all sizes to establish the public domain status of serials published in the US between 1923 and 1963. We will provide:

- A complete, searchable inventory of serials with issue and contribution renewals made between 1950 and 1977, and the dates of the first such renewals. (The Copyright Office already provides a searchable database of renewals from 1978 and later).
- Procedures for using this inventory and other data to expeditiously decide whether a portion of a serial (such as a volume or a specific contribution) is in the public domain, similar to the efficient copyright-clearance procedures developed for HathiTrust's Copyright Review Management System (CRMS).
- An example set of post-1922 serial volumes that have been cleared for posting using the data and procedures above.
- A report describing how the information above was developed and cleared.

Each of these products builds on the previous ones described. We intend to produce and publish them under open licenses between October 2017 and September 2018.

The primary audience for this work consists of librarians and other digitizers who have access to copies of these serials, whether physical copies in their own collection or digital surrogates through review systems like CRMS. They will be given confidence and resources to more easily copyright-clear various kinds of serials and make them openly available to readers online. The readers that will have access to the serials are the ultimate beneficiaries of the project.

With reviewed, tested, and documented procedures and data for establishing public domain status of serials from 1923 to 1963, we hope to enable digitization projects across the nation, large and small, to put many more 20th century serial volumes online while they are still in digitizable condition in library and museum collections. The project is inspired by, and builds on the work of, HathiTrust's IMLS-funded CRMS project, which has similarly implemented processes that have enabled hundreds of thousands of post-1922 digitized books to be copyright-cleared and made freely available online. We hope that this project can in turn help make as many post-1922 serial volumes freely available online as well.

## Opening access to mid-20th century serials: Sparks grant proposal narrative

The University of Pennsylvania Libraries seeks a $25,000 1-year IMLS Sparks grant to provide and demonstrate tools and data for opening access to the large body of scholarly and general serial literature from 1923-1989 that is in the public domain, but not yet freely accessible online.

### 1. Statement of national need

Our nation's libraries collectively offer a wealth of informational and cultural resources from the twentieth century in the form of serials such as magazines, newspapers, and trade and research journals. In the past few years, mass digitization programs have made millions of books published prior to 1922 freely available to the world, with well-documented cultural benefits. Projects like HathiTrust and the Internet Archive have brought hundreds of thousands of book volumes from 1922-1963 to the world as well, thanks to established programs (some funded by IMLS) to clear copyrights for these materials. (Copyrights for US-originating publications prior to 1964 had to be renewed in order to stay in effect, and many such copyrights were not renewed.)

Much of our knowledge and culture, however, has not been published in books but in serials. Serials are often the only source for information like

- Research and scholarly articles appearing in journals, which in many fields can stay relevant for decades after publication
- The first (and often only) publications of many works of fiction and other creative literature
- News, essays, opinion pieces, and other journalistic works
- Advertisements documenting national and local culture and commerce
- Trade and specialty journals that shed light on communities, practices, and knowledge outside the general public eye
- Local-interest publications that uniquely document the history and culture of communities across the United States

Serials like these are held by libraries across America. Some are in large libraries, but many small libraries also hold unique serials with local or specialized focus. Many serials are both valuable for research and also at high risk to be lost to posterity, due to a low number of copies held by libraries, deteriorating copies, and deaccessioning by libraries strapped for space. The sooner these serials can be digitized, the more likely they will be available for the world to access.

Although copyright renewal rates are lower for serial issues and contributions than they are for books, post-1922 public domain serials have not gone online as quickly as post-1922 public domain books. Those books have gone online in bulk thanks to programs like HathiTrust's Copyright Review Management System, which provides easy-to-follow checklists for researching copyrights of books, and draws on easily-searched data sources like Stanford's Determinator book renewal database. Similar

resources have not been made available to date for researching serial copyrights. While serial copyright clearance is often more complicated than book copyright clearance (since a typical serial issue consists of multiple works of authorship while a typical book consists of one), data sources that are indexed per-serial can make it easier for many serial issues and the works included in them to be checked for copyright in a small, manageable number of searches.

In this project, we will develop tools and procedures that libraries of all sizes can use to more easily clear copyrights for serials and make them available to the public online. Specifically, we will provide

1. An open, searchable online **data** source listing all serials with issue or contribution copyright renewals made between 1950 and 1977, and the dates of the issue with the first renewal for any given serial. Combined with the Copyright Office's online database of registrations and renewals from 1978 onward, librarians will quickly be able to determine whether there were any renewals for material originally appearing in a serial of interest, and when these renewals began. Our data source will also accept additional information from the community for clearing serials.
2. A set of suggested **procedures**, based in part on the procedures developed for clearing books in HathiTrust's CRMS, that librarians without special legal expertise can use to determine whether serial content is in the public domain and eligible for digitizing and openly posting online.
3. An **example set** of post-1922 serial issues that have been copyright-cleared and digitized using the data and procedures that we develop, along with documentation on how we cleared them. Producing the sample set will test the data and procedures we produce, and demonstrate how other libraries can similarly clear and digitize serials in their collections.

By enabling the copyright clearance of many American serials from the mid-20th century, this project should accelerate the spread of American culture, heritage and scholarship across the country and the world. It forms an important part of the infrastructure needed to build national digital libraries that take advantage of the full richness of the American public domain.

## 2. Project Design

As stated above, the goal of the project is to produce data and procedures to more easily determine public domain status of 20th century American serials, in order to make it easier for libraries to digitize and publish them online without copyright infringement. In this section, we discuss our three main deliverables in more detail.

**Data:** We will publish online, under an open license, a complete inventory of serials that have issue and contribution copyright renewals in 1950-1977 volumes of the

Catalog of Copyright Entries (CCE) published by the Library of Congress's Copyright Office. For each of these serials, we will also provide the date of the issue with the earliest issue and contribution renewal mentioned in the Catalog, if any. The data we will provide will be based on data that is already online in the form of page-image scans of the CCE. We will produce a summary of this data in a form that is automatically searchable, and indexed by serial, so that librarians can easily find out whether a serial has any copyright renewals associated with it, and the earliest issue that has such renewals. Most serials published in the US between 1923 and 1963 have no renewals associated with them, and many that do have renewals do not have renewals for all issues. With the serial-indexed searchable data we will provide, librarians will be able to verify the lack of renewals for many serials with one or two quick searches, rather than having to take much more time to visually examine many pages of the Catalog of Copyright Entries.

John Mark Ockerbloom, principal investigator for this project, has already compiled some of the data that librarians would search. In 2006, he completed an inventory of serials with issue renewals made between 1950 and 1977. Since then, he and an intern have slowly compiled an inventory of serials with contribution renewals made between 1950 and 1965. (These inventories can be found online at http://onlinebooks.library.upenn.edu/cce/firstperiod.html.) Funding for this project would enable us to complete the inventory through 1977 in less than a year, and develop, test, and publicize procedures for using it to quickly clear copyrights for serials.

Two library interns will initially compile the data for the inventory. Each will look at the same CCE volumes and note serial titles with renewals. Having two independent sets of eyes on each CCE volume will minimize the chance of missing a serial accidentally, and also let us evaluate the quality of each intern's work. For each volume, the two interns' work will be compared and then incorporated into the project's master serial inventory by a professional librarian. We will also keep track of the work hours and completion dates for each successive year of our inventory, to ensure that we stay on track for completing the inventory through 1977 on time and on budget.

We will deposit an open-licensed snapshot of the inventory into Penn's digital repository at the end of the project. In addition, the inventory on the Web will be open to further expansion and contributions from the community. For example, we might note a particular serial that was found to not have any issue or contribution renewals; or we might add information about the specific periodical renewals that were made for a particular serial after the first one.

**Procedures:** HathiTrust's IMLS-funded CRMS system showed how books published after 1922 could be copyright-cleared in bulk, with multiple reviewers using a checklist to determine whether a book's status was public domain, not public domain, or not easily determinable. If two reviewers agreed on a book's public domain status, the book would be noted as public domain with access opened on HathiTrust; if they disagreed, or found the status not easily determinable, the book

would be noted as needing more expert review before opening. The procedures we intend to develop would work in much the same way. They would not depend on the use of the reviewing software developed for CRMS, or any other particular software, but we intend them to be compatible with the software and basic workflow used by CRMS.

CRMS involves a certain tradeoff between efficiency and certainty, and we expect that our procedures will involve similar tradeoffs. Serials produced outside the US, for instance, will be outside the scope of our clearance procedures, due to the complex legal and bibliographic questions involved in copyright restorations of foreign works. We expect our procedures will be best suited to giving a definite determination for serials that publish original, non-syndicated content. Many scholarly journals and other periodicals of current research interest fall into this category. Some types of serials may also be clearable in part under these criteria—for example, the news sections of local newspapers, which may be the portions of the greatest historical interest, may be public domain even if their syndicated comics sections are not.

We will consult with legal experts to review the legal issues and risks of the procedures we develop, and with librarians experienced in copyright clearance to evaluate the practicality of the procedures.

**Example set:** After we have drafted procedures and published the data described above, we will select a set of American serial volumes published after 1922, review the volumes and their contents to check verify public domain status, and make openly available digital copies of the material we determine to be in the public domain. Our digital copies will be published on Penn's web site and also placed on a national digitized books platform such as HathiTrust and/or Internet Archive. The metadata will be harvestable by projects like the Digital Public Library of America.

While digitization is not the primary focus of this project, we do find it important to test our data and procedures on actual serial content, and trust our decisions enough to make this content freely available online once we determine that it is in the public domain. Since we cannot assume that other projects such as HathiTrust will be willing to open their content prior to our project's completion, we are prepared to digitize volumes ourselves at our own expense. We will either use Penn's existing digitization facilities, or contract the digitization to an outside agency, and then publish the volumes online ourselves. The volumes we digitize should be of interest in their own right to present-day researchers, and also allow libraries to do their own verification of our procedures for clearing copyright.

We intend to clear and digitize at least 30 volumes from a variety of serials, and may do more as local budget and collection priorities allow. We will work with bibliographers in the Penn Libraries to determine suitable example volumes. We will focus mainly on research journals and on serials of local interest, two types of serials that we expect many libraries across America will have special interest in digitizing, and that we hope will be relatively straightforward to clear. We may also

consult with local special collections libraries in the Philadelphia area to see if there are serials in their holdings that they are interested in clearing and digitizing. Serials produced by or for underserved groups or communities may be particularly attractive to provide online, as they tend to have lower renewal rates, and digitization can greatly increase their audience.

As with any project, there are **risks** involved in our plans. The data we compile on copyrights may be inaccurate. We hope to minimize inaccuracies with our plans to have Catalog of Copyright Entries volumes reviewed independently by two people, and to alleviate the effect of inaccuracies by encouraging concerned libraries to double-check renewals in the original Catalog of Copyright Entries volumes, which will be reachable by a link from our inventory of renewed serials. We will also solicit corrections and additional information from libraries and other digitizers interested in serials renewals.

The time required to produce data and procedures may be greater than we expect, making it difficult to stay within the project's budget or timeline. Our experience with inventorying contribution renewals from 1950-1965, however, gives us a good idea of how long it takes our interns to complete inventories for a particular year, so our projected time should be close to the time actually required. (We are taking into account the larger number of serial renewals in later years, based on the page count of renewals in various years of the Catalog of Copyright Entries.) The Penn Libraries are also willing to allocate in-kind labor for preparing procedures and reports beyond the budget of the grant, if required.

There is inherent legal risk in any copyright determination process, and large-scale public domain clearance processes will sometimes produce errors. (We have corrected erroneous determinations of other projects ourselves from time to time.) Based on the track record of other noncommercial projects like HathiTrust and Project Gutenberg, both of which have dealt with multiple takedown requests, our liability risks appear to be low if we have well-documented careful procedures that we follow in good faith, we do not attempt to commercially exploit the material we digitize, and we have clearly indicated contacts and rapid response for takedown requests.

We will have university counsel review our procedures to mitigate the risk of improper copyright determinations. There is a counterbalancing risk of practicality if our procedures end up too risk-averse or costly to be feasible for most serial copyright clearances. We are, however, familiar with HathiTrust's copyright clearance procedures for books, and believe that we can write similar procedures with similar degrees of certainty for serials. Measuring the clearance times required for our production set will help us determine and report on clearance costs for various types of serials, and report to the community on the types of serials and procedures that are likely to be cost-effective for various risk tolerances.

There is uncertainty about the cost of publishing volumes in our example set. The cost of clearance depends on the complexity of the procedures we adopt, and the

amount of time and level of expertise required to carry them out. The cost of digitization can vary depending on who does the digitization. Our own digitization operation can handle a wide variety of materials, including fragile and tightly bound materials, but has higher per-page costs than some outside organizations (which might not be able to handle as wide a range of materials as we can internally). We have committed resources for an example set of at least 30 volumes, however, and may potentially do more if costs are lower than expected or there is increased interest in digitizing additional public domain serials for our digital collections.

**Staffing:** John Mark Ockerbloom, Digital Library Planner at the University of Pennsylvania Libraries and founder and editor of The Online Books Page, will direct the project. Joe Zucca, Director of Strategic Initiatives and Library Technology Services at the University of Pennsylvania Libraries, will coordinate needed IT infrastructure and staffing. Metadata librarians and interns (such as Alison Miner, who has worked with John on early stages of the survey) will compile information on copyright-renewed periodical contributions from the Catalog of Copyright Entries. Bibliographers led by Dick Griscom, Director of Collections and Liaison Services, will select and coordinate digitization of sample serial volumes. Robert Firestone and Robert Terrell, attorneys at the University of Pennsylvania Office of General Counsel, will review our copyright workflow procedures. We will also consult with copyright experts at HathiTrust and other organizations with experience in copyright and digitization.

The **financial resources** required for this project go mostly to labor, with some funds provided for travel for consultations and dissemination. Labor funded by this grant includes intern and professional staff time to complete a basic periodicals contributions data inventory through 1977, as well as professional staff time to develop a copyright clearance procedure for serials, and refine the presentation of inventory data for ease of use. Funds are also allocated, and will be measured, for the labor involved in clearing copyrights in our example set. The University of Pennsylvania will cover digitization costs, general infrastructure and overhead costs (including consultation with the Office of General Counsel), as well as any additional staff time needed for this project's deliverables. A more detailed allocation for the funds we request can be found in our Budget Form and Budget Justification document.

The results of our project will be **disseminated** broadly to an audience of librarians and other digitizers. The inventory of serials and their first renewals will be published on the Penn Libraries website and released CC0. We will publish our clearance procedures and other project reports online with CC-BY licenses. We also intend to discuss our work with colleagues at local libraries in Philadelphia, and at other academic libraries involved with digital projects, to get feedback on usability of our data and practicality of our clearance procedures. Some of our project's budget will go to presenting our work at relevant conferences. Our publications can remain on our websites, and our digitizations in our repositories, past the end of the project with no significant maintenance cost.

## 3. National Impact

If our project delivers what we specify in the previous section, we should reduce the cost, risk, and uncertainty of providing online access to important historic American serials well into the 20th century. Our work will help libraries of all sizes across America to identify serials they have collected for their local communities and make them available to their local online patron and to the entire nation. IMLS funding will enable this work to be completed within one year, much more rapidly than our self-funded efforts have progressed to date, and aid in broadening the community input, collaboration, and promotion for the work. The sooner we can make our data and procedures available to libraries and other digitizers, the sooner serials can be made available to the public online, and the more likely they are to be saved from obscurity and loss.

Our performance goals for this project are to broaden access and expand use of the Nation's content and collections. This project is intended to largely meet those goals indirectly, though the example set we will providing online will be a small direct contribution to broadening access to the content of certain serials. The best measure of our success on those goals will be the extent to which the data, procedures, and examples we provide spur other projects to start digitizing and opening public access to post-1922 serials.

We will note in our project report any initiatives we hear about that take advantage of our work prior to the project's conclusion. One project that already has done so is the Media History Digital Library, which has posted mid-20th-century serials and books related to film, broadcasting, and recorded sound at http://mediahistoryproject.org/. David Pierce, founder of the project, has told us in email (quoted here with his permission) "I've used your periodical renewals information, and it was a useful starting point when I searched all renewals for media-related publications through 1977."

We expect that providing a more complete set of serial copyright renewal data, along with instructions and example of how it can be used, will similarly accelerate the pace of making post-1922 serials online at many other institutions. We are not sure how many projects will have taken advantage of our work by the end of the grant period. However, if our work helps large projects like HathiTrust start opening access in bulk to post-1922 serials, or encourages local and special libraries to make their unique 20th-century serials available to the world, our investments will have been clearly worthwhile.

**Opening access to mid-20th century serials: Schedule of completion**

This is a 1-year project, which we expect to commence on October 1, 2017 and complete by September 30, 2018. Below is a table of expected dates for major activities.

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Complete renewals data set | ■ | ■ | ■ | ■ | ■ | ■ | | | | | | |
| Produce serials copyright clearance procedure | | | | ■ | ■ | ■ | ■ | | | | | |
| Select serials to clear and digitize | | | | ■ | ■ | ■ | ■ | | | | | |
| Clear and digitize selected serials | | | | | | | | ■ | ■ | ■ | ■ | |
| Write report | | | | | | | | | | | ■ | ■ |
| Publish work products in repositories | | | | | | | | | | | | ■ |
| | Oct | Nov | Dec | Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep |

## DIGITAL PRODUCT FORM

**Introduction**

The Institute of Museum and Library Services (IMLS) is committed to expanding public access to federally funded digital products (i.e., 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. However, applying these principles to the development and management of digital products can be challenging. 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**

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 data we produce on first copyright renewals for serials will be released CC0, meaning that it is free of copyright restrictions. The data is essentially factual information, which is not copyrightable, and the CC0 dedication both makes this clear, and removes barriers that might otherwise exist against reusing, distributing, and adapting it.

The prose documents that we produce under the project (specifically, the copyright clearance procedures and checklist, the report on our project's findings, and other documentation) will be released CC-BY, with attribution requested for the individual authors and for the University of Pennsylvania. This license both makes the checklists and reports easily reusable, and also makes sure the source of those reports is clearly indicated, to allow for credit and verification of the content. Because the products described in this paragraph are intended to be coherent integrated documents, the attribution requirement is not burdensome for their reuse like it might be for data.

The digitizations that we produce for serials in our example set will be released with a "no known restrictions on publication" statement. We will have determined as part of the project that the periodicals we include are in the public domain in the US (with at most small portions, such as brief quotations, that fall under de-minimis or obviously fair-use exemptions). The digital surrogates for the periodicals are intended as exact reproductions, which do not entitle us to claim new intellectual property rights, even if we wanted to claim them.

**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.

Answered above.

**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.

We do not expect any such issues to arise in the project, which is based on publications and data that have already been widely published, and that do not have copyright restrictions.

## 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 format you will use.

As part of an example test and demonstration set, we will digitize at least 30 serial volumes published after 1922 that we have determined to be in the public domain. This is not the main focus of the project, and we are not requesting funds for the digitization itself. However, since we intend to have the volumes of acceptable quality for reading and for evaluating our work, we include some details of our plans below.

**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.

We have not yet decided whether we will do the digitization in-house or hire an outside agency. If we do it in house, it will be through our Schoenberg Center for Electronic Text & Image (SCETI). The Center follows the NISO standards as articulated in A Framework for Building Good Digital Collections, 3rd edition (http://www.niso.org/publications/rp/framework3.pdf). The following represents specific standards adopted by SCETI in their capture, storage, and presentation of digital images.
Photography specifications:

- Archival Masters: No less than 300 PPI 24-bit TIFF image; 300 PPI 24-bit LZW color image
- Standards: Library of Congress. Building Digital Collections: A Technical Overview

https://memory.loc.gov/ammem/about/techIn.html
Photography & Processing Equipment

- Cameras: For the purpose of this project two Phase One P45 digital cameras will be used. The P45s are state-of-the-art, manual-focus 39-megapixel cameras.
- Lenses: Schneider-Krueznach lenses.
- Lighting: cameras auto flash features; as well as ProPhoto Acute D4 heads and "beauty dish" reflectors

Image Delivery Format: Images will be made available in OPenn and delivered to APS in TIFF format for the high-resolution masters.

We may instead hire an outside agency to do the work for lower cost than in-house, but with similarly high resolution and standard image and metadata formats. One such agency under consideration is the Internet Archive's scanning center services. Their services and standards are described at https://archive.org/scanning .

**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).

(See above.)

### B. Workflow and Asset Maintenance/Preservation

**B.1** Describe your quality control plan (i.e., how you will monitor and evaluate your workflow and products).

SCETI has long-standing quality control procedures described at their website. If we use an outside digitizer we will ensure that they have acceptable quality control as well.

**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 digital serial volumes will be openly licensed, and maintained in Penn's institutional repository as well as in nationally

scoped, long-standing digital text repositories.

**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).

Metadata for the serials scanned will be kept in our online catalog (which currently uses MARC records). MARC will also be used for records fed to HathiTrust as per their requirements. (We will also provide their required technical metadata for volumes they ingest, as we would also for the Internet Archive.) For copies provided to the Internet Archive and/or the Digital Public Library of America, we will provide Dublin Core or a superset for bibliographic data.

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

Metadata stored in our catalog, our digital repositories, and national-scale digital book repositories will be preserved as assets of the Libraries and the repositories.

**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).

We intend to contribute the metadata (and the content) to at least one external digital platform (as noted above). As with other repository and catalog records, we will also make bibliographic metadata available via OAI-PMH.

**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).

Volumes will be openly published online, both at Penn and at other national digital content collections such as HathiTrust or the Internet Archive.

**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.

Schoenberg Center for Electronic Text and Image: http://sceti.library.upenn.edu/
OPenn: http://openn.library.upenn.edu/

## Part III. Projects Developing Software

(Not applicable.)

## 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.

We will be compiling and publishing searchable data on periodicals that had issue or contribution copyright renewals, based on data published in the Copyright Office's Catalog of Copyright Entries.

**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?

The data we will publish does not require approval by an internal review panel or institutional review board.

**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).

No.

**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.

Not applicable.

**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).

The data will be compiled by multiple library interns supervised and cross-checked by a librarian.

The data will be human-readable and machine-searchable, with documentation; we may also provide it in machine-processable formats.

**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?

The documentation and the HTML page with data set will be published on the Web and linked to each other with URLs.

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

The data set will be published on the Web at a standard location, with a URL we intend to keep stable, and available to web archivers like the Wayback Machine. (http://onlinebooks.library.upenn.edu/cce/firstperiod.html will be the canonical location, has been a valid URL for the last 10 years, and has over 180 Wayback snapshots to date.) Also, a snapshot of the data set and documentation at the time of project completion will be deposited in the University of Pennsylvania's Scholarly Commons Repository. Both will be published with open licenses (CC0 and CC-BY, as described in Part I), allowing others to download and archive duplicate copies.

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

Name of repository: Scholarly Commons @ Penn

URL: http://repository.upenn.edu/

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

We will review the plan at the end of this project. We intend to keep the data at the canonical Web location available and growing as individuals contribute additional data, as described in the Narrative. (Our principal investigator has maintained that URL for the last 10 years, intends to keep doing so, and has insured that the web site is well-replicated in web archives in the event he ceases to maintain it.)