Abstract

The Harvard Law School Library Innovation Lab, in cooperation with the Berkman Center for Internet & Society and over 130 partner libraries, and with the support of LOCKSS, DPLA, Memento, the Coalition for Networked Information, and others is seeking a National Digital Platform Grant to sustainably scale Perma.cc to combat link rot in all scholarly fields.

Link rot is a serious problem. It affects upwards of 70% of all scholarly articles in law, medicine, science and technology, causing irreversible harm to the digital scholarly record. As our letters of support vividly explain, it affects authors and publishers of all stripes, as well as their readers, and the scholarly world needs a solution.

Perma is a unique, effective solution to the problem of link rot. Over the last two years, we have aggressively tackled the link rot problem in the narrow world of legal scholarship. At present, Perma's library coalition includes 120 academic law libraries collectively supporting almost 500 law journals and faculty members. Perma also is being used by seven state court systems to preserve links in their decisions. We have roughly 7,500 individual users who have preserved over 200,000 sources. Perma is open source, all of its code is on GitHub, and users can and have submitted changes. It is interoperable, based on Memento, LOCKSS, Web Archive format, and other open standards and protocols, and it publishes its own powerful API.

Perma has proven decisively that a broad coalition of libraries paired with cutting-edge web archiving and distributed data storage tools can have an enormous impact on link rot. We and our supporters believe that Perma is poised to expand its service beyond legal scholarship, grow its library coalition and tackle link rot in other fields. We therefore seek this two-year $782,649 grant to scale our proven technology and approach, and to do so in a sustainable way by designing, testing and launching a paid service that will subsidize the free-of-charge service offered to those supported by our academic library partners.

Harvard Law School Library Innovation Lab - Scaling Up Perma.cc

1. Statement of Need

Citation to persistent sources is fundamental to all academic and many professional fields, among them law, science, technology, medicine and journalism. These fields of knowledge rely on networks of citations and on the tireless efforts of librarians working to collect, organize and preserve the cited sources. Citations today, however, increasingly refer to web pages, not just print sources.[1] Because web pages change their content and disappear all the time, citations to them are ineffective at best and, at times, misleading.[2] This problem, known as "link rot" or "reference rot," means that much of our citation-dependent work in these fields is being written on sand.

The problem is pervasive and growing. A study we conducted in 2013 revealed that, in the legal field, 50% of Supreme Court cases and 70% of law review articles with web citations suffer from link rot.[3] More recently, a detailed study confirmed the same, a roughly 70% level of link rot in a broad corpus of science, technology, and medicine articles published between 1997 and 2012.[4] As web citations become more prevalent, the problem threatens to grow unless checked. As confirmed by our observations and our letters of support, link rot undermines scholarship in all of its forms and is one of the most troubling, urgent problems in modern research and scholarship.

Unfortunately, it is not easy today for scholars and other authors to protect their work and readers against link rot, or for libraries to assist their communities in preventing link rot. While absolutely critical, the broad-based preservation efforts of the Internet Archive are focused primarily on maximizing preservation of content independent of the particular needs of those referencing the preserved content. As a result they do not guarantee authors a precise, accurately preserved source to which they can direct readers for easy access should the original become unavailable. The on-demand URL archiving tools that do exist, such as WebCitation.org and Internet Archive's "Save Page Now" tool, can be difficult for authors and readers to use and lack the administrative features that archiving organizations need to manage their collective efforts. Moreover, as single source solutions, these alternatives can be vulnerable in ways that might lead authors to question reliance. Existing tools also are not designed to put libraries front

---

[1] Klein M., Van de Sompel H., Sanderson R., Shankar H., Balakireva L., Zhou K.,& Tobin, R. (2014) Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot. PLoS ONE 9(12): e115253. doi:10.1371/journal.pone.0115253. Archived at http://perma.cc/P4HF-K6FS.

[2] The best-known example is http://ssnat.com, which Justice Alito cited to in his concurring opinion in *Brown v. Entertainment Merchants Assoc.*, 131 S.Ct. 2729, 2749 n.14 (2011). Archived at http://perma.cc/PH9K-HV7K.

[3] Zittrain J., Albert, K. & Lessig, L. (2014) Perma: Scoping and Addressing the Problem of Link and Reference Rot in Legal Citations, *Harvard Law Review Forum 176.* Retrieved from http://harvardlawreview.org/2014/03/perma-scoping-and-addressing-the-problem-of-link-and-reference-rot-in-legal-citations/. Archived at http://perma.cc/NE2Z-TMFK.

[4] See Klein M., Van de Sompel H., Sanderson R., Shankar H., Balakireva L., Zhou K., & Tobin, R. (2014).

and center, where they can use their expertise and standing to educate their communities about link rot and to promote prevention. As a result, these tools lack the institutional support and engagement needed to encourage widespread adoption and to support a large, diverse user base.

Libraries play a key role in the analog world as the trusted repositories of the scholarly record. Over the past two years, Perma has shown that libraries can serve the same role in the digital world. Perma solves the link rot problem by allowing authors to take snapshots of web pages and place them on deposit with libraries. Each Perma record has a unique URL (e.g. https://perma.cc/F37P-2E4V) that readers can follow to view the preserved snapshot. Perma includes many features to support libraries in their role as trusted administrators - or "registrars" - for their communities of journals and scholars using Perma.

Perma was launched in September 2013 by the Harvard Library Innovation Lab under the leadership of Jonathan Zittrain and is well on its way to eradicating link rot in the legal field. In just over two years, its library network has grown to over 120 law libraries. Roughly 500 archiving organizations, including law journals, courts, faculty members, and other entities have adopted Perma. Perma is open source, all of its code is on GitHub, and users can and have submitted changes. It is interoperable, based on Memento, LOCKSS, Web ARChive file format, and other open standards and protocols and publishes its own powerful API. With our high quality captures, group management features, and distributed network of trusted library partners, Perma has quickly become the best available preservation tool for published web citations.

The link rot problem, however, extends far beyond academic legal scholarship and court opinions. It affects everyone who cites or links to web sources with the expectation that they will be accessible to future readers. The problem thus impacts other academic disciplines, university press offices, commercial publishers, professional archivists, research centers and think tanks, government offices, law firms, news organizations and individual authors. Requests for access to Perma have rolled in, unsolicited, from these groups, as have requests from academic libraries outside the legal field. Our supporting documents and letters also demonstrate this demand.

Unfortunately, Perma cannot serve this broader demand in its current form with its existing team and economic model. For the Perma service to scale beyond legal literature, its infrastructure and interface must develop, it must add some additional personnel to the team, and it must pursue economic sustainability through a paid service in which private and commercial usage subsidizes the free-of-charge service

made available to public and academic users. The Scaling Up Perma.cc project, if funded, will make these things possible.

2. Impact

This project will have a broad beneficial impact on authors, publishers, readers and libraries, and on web archiving efforts in general.

The link rot problem affects everyone who cites or links to web sources with the expectation that they will be accessible to future readers. And it is a problem of deep concern to libraries and archivists, who for so many years have worked to ensure preservation of the scholarly record.[5]

Perma has a proven track record of success in solving this problem in one field, a dedicated, experienced team, and a fundamental model that is well-suited to scale. It embraces decentralization and celebrates the role of libraries in handling the front-line work of educating and supporting users and administering the system according to local needs, policies and priorities.

We currently measure our impact in the legal field by tracking several important metrics, including growth in library partners, archiving groups, individual users, links preserved and Perma Links published in-the-wild. We will extend these metrics as we scale, tracking new library partners, new archiving organizations, new individual users, new links preserved, and new links published in-the-wild.

Our target for attracting new library partners during the grant period is to reach 5-10% of the 3,700 existing academic libraries, including 25% of major research libraries.[6] This would add 370-740 libraries as Perma registrars, which is three to six times the number of libraries in our current network.

---

[5] Peoples, L. (2015). Internet Citations in Oklahoma Attorney General Opinions, *Law Library Journal, 107,* 347-375, Retrieved from http://www.aallnet.org/mm/Publications/llj/LLJ-Archives/Vol-107/no-3/2015-17.pdf. Archived at http://perma.cc/MDY7-W2CF; Engle, E. (2015, August 17). Cooking Up a Solution to Link Rot (blog post). Retrieved from https://blogs.loc.gov/digitalpreservation/2015/08/cooking-up-a-solution-to-link-rot/. Archived at http://perma.cc/EDU2-XLLU; Chant, I. (2013, October 2). Perma.cc Aims to Bring Staying Power to Online Legal Citations, *Library Journal,* Retrieved from http://lj.libraryjournal.com/2013/10/academic-libraries/perma-cc-aims-to-bring-staying-power-to-online-legal-citations/. Archived at http://perma.cc/9UU3-ZMQP.

[6] http://www.ala.org/tools/libfactsheets/alalibraryfactsheet01. Archived at http://perma.cc/W2YY-QHXP; http://www.arl.org/membership/list-of-arl-members. Archived at http://perma.cc/3C44-W9KQ.

In addition, as we scale, we will measure and act upon engagement and satisfaction metrics, like active/inactive users, usage frequency, adoption rates within institutions, and satisfaction among library partners. Currently we see fairly high rates of active usage for this type of service, with roughly one-third of all of our users having used Perma in the last three months. We have more room to grow in this area, and it will be critical to emphasize these metrics as we grow to ensure that we and our library partners are directing our efforts toward the right activities. Perma is extremely user-focused, and we actively solicit feedback from all community members, especially the library partners who work directly with end users, through our blog, our newsletter and through partner emails. We have a user support team assigned to receive, triage and respond to inquiries at info@perma.cc concerning technical and other issues, as well as requests to join Perma.

We also have opportunities to use metrics regarding traffic to Perma records, API calls and the like to gain a sense of the broader impact of Perma on readers and the web archiving community. Although measuring link rot itself is challenging, we also hope to develop metrics that will allow us to demonstrate when and to what extent Perma records become necessary as the source URLs degrade.

Our planned approach to developing, testing and launching a paid service is set out in more detail below, but success in that aspect of the project will be determined by standard web service metrics on adoption, active engagement, retention, churn and overall satisfaction (e.g. through NPS scoring). Also critical to validating self-sustainability will be detailed cost and revenue modeling.

The tangible products of this project include a well-functioning, widely available link preservation service, an open source code base (as well as contributions to open source libraries we use), a powerful, open API that can support further development and innovation, a feed of high-value, scholar-curated archived items that will be available via Memento, a new private LOCKSS network focused on preservation of the digital scholarly record, which could have other potential uses in the future, and perhaps most importantly, a broad, effective coalition of loosely affiliated libraries working together to solve a major problem in digital preservation.

3. Project Design

Scaling Up Perma.cc is a two-year project designed to achieve three objectives:

- Grow our library-partner network to include academic libraries of all disciplines;

- Expand our free-of-charge service to academic users from all disciplines, who will be supported by their academic libraries;

- Develop, test and launch a paid service for non-academic users that will allow Perma to sustainably scale to meet the full scope of the link rot problem.

To achieve these goals we plan to undertake activities in three categories, as detailed below, and as further reflected on our Schedule of Completion:

A. Technical Development

A significant portion of our effort over the grant period will be dedicated to technology development in three main areas: (1) improving Perma's existing capabilities and enhancing tools for use by larger communities of users; (2) implementing Perma's distributed storage network using LOCKSS; and (3) integrating payment and subscription handling to support a paid service.

We will continue improving our web capture systems, with priority on efforts to improve capture of dynamically generated web content and more complicated file formats. We also will build authentication tools for libraries, which will automate the sign-up process for users already vetted by their internal identity management systems. Following modern web application practice, we will implement affordable, scalable frontend servers that will permit our capacity to flex as needed as usage and traffic require. We will improve administrative tools used by library and journal administrators to manage accounts and affiliations and to monitor Perma usage across their organizations. We also will complete our integration with Internet Archive, by which we will automatically mirror preserved records to Internet Archive as an additional point of redundancy. All of this development will support increased scale and assist our library partners in supporting large communities of users.

We will complete development and testing of our private LOCKSS network and begin deploying it to partner libraries with the support of the LOCKSS team, from whom we will be purchasing a support license. Scaling Perma depends on distributing the storage architecture across multiple libraries to ensure robustness and sufficient redundancy. We already have a working software prototype that is ready for testing with external partners. We also have prototype for a lightweight, affordable hardware appliance that could be used on a plug-and-play basis by libraries, including public libraries, who wish to contribute to the Perma effort but have limited technical expertise and resources.

To support a paid service, we also will integrate with third-party payment services, such as Stripe, and add tiered subscriptions to Perma's internal architecture.

B. Library Outreach and Community-Building

We also will devote major effort during the grant period toward expanding our coalition of library registrars, especially among libraries outside the legal field. Libraries at Princeton University, Xavier University, University of Wyoming, Western Washington University, Skidmore College and Sheridan College have already joined. Our letters of support include letters from Cornell University Library and the University of California, Davis promising to join our network if this grant proposal is funded and to support Perma for non-law materials in the sciences, social sciences and humanities.

To help guide us in expanding the library coalition, we plan to convene an advisory board of representatives from university libraries. And in addition to the technical enhancements described above, we will refine our library onboarding process, create educational and training materials for new library partners and do more to educate libraries about the link rot problem and their role in helping to solve it. Following the model that has served us well in building our law library network, we will also conduct direct outreach to other academic libraries and invite them to join Perma so that their communities of scholars can take advantage of Perma's unrestricted, free-of-charge services. In addition, when faculty, students or other academic users contact us to request unrestricted Perma access, we will continue to direct them to their local libraries for assistance in gaining access to Perma.

C. Research, Testing and Public Launch of Paid Service for Sustainability

The third major prong of our project will be to research, test and publicly launch a paid service that will enable Perma to be self-sustaining.

More details about our sustainability plan are provided below, but we will begin the grant period by convening a separate advisory panel to guide us in this pursuit. We also will conduct customer and market development research in five specific fields that, based on our experience to date, represent promising use cases and markets for a paid service: (1) law; (2) news; (3) publishing; (4) policy research; and (5) individual authors. Our letters of support from Fastcase, MIT Press, Ohio State University Press, BWB Publishing, Vera Institute of Justice, Mackinac Center for Public Policy, the Shorenstein Center, and the Trust for Public Land illustrate this demand, as do the unsolicited

inquiries we have received from major law firms, news organizations, publishing firms, individual authors and a wide variety of others.[7]

Based on this early research into potential use cases, we will conduct private paid beta tests with limited numbers of users from each group. The tests will be designed to evaluate assumptions about market, costs and pricing, to gather feedback about the Perma product, to develop the Perma technology roadmap as necessary to meet these new use cases, and to learn what marketing messages and tactics will be most effective in promoting the paid service. During the testing period, we will iterate based on the concrete feedback we gather.

As guided by the results of our research and beta testing, we will develop and refine detailed business, pricing and cost models to inform forecasts about sustainability. Near the end of the grant period, based on the insight gathered through the period, we will plan to publicly launch a paid service on terms that will put Perma on sound, sustainable footing for the future.

4. Diversity Plan

Because Perma is a hosted service requiring no technical infrastructure or expertise from its library-partners or their community members, it will be as readily accessible to scholars working in rural (but connected) areas as to those in high density academic environments. In addition, while additional technical work needs to be done, the Library of Congress has used Perma extensively through its Global Legal Research Directorate to preserve foreign language materials online, and this demonstrates the potential for Perma as a resource for those working primarily in other languages.

5. Project Resources: Personnel, Time, Budget

Scaling Up Perma is proposed as a two-year, $1,605,775 project, with $782,649 in IMLS grant funds coupled with $823,126 in cost-share provided by Harvard Law School.

Two years is a realistic timeframe for undertaking the planned project activities and achieving project goals based on the fact that Perma is already a well-functioning technology platform, with a proven model for decentralized support and administration by our library network. Perma first went live in September 2013, and in a little over two years has achieved substantial growth and stability. Our schedule of completion details how we will use the project time. By way of overview, we will spend the first year of the project mainly on technical development to serve broader, larger classes of users, as

[7] See supplemental document entitled “List of Inquires”

well as on outreach and community building, and on research, design and initial testing of our paid service. The second year of the project will be focused on deploying our distributed storage architecture, iterating our paid service based on initial private beta testing, refining business and economic models, and preparing for public launch of the paid service.

Reflecting the overall project design and goals, the proposed two-year project budget calls for investment in technical development, outreach and community-building, and the research and launch of a paid service. Our budget form and narrative supply more detail, but we have allocated roughly 45% of the budget to technical personnel and resources, roughly 25% of the budget to outreach, communications and marketing personnel and activities, and the remaining 30% to project management, administration and oversight, including indirect costs.

The key personnel who will carry out these activities include:

- *Jonathan Zittrain* (5% time for two years) - Director of the Harvard Law School Library
- *Kim Dulin* (35% time for two years) - Librarian, Director of the Innovation Lab
- *Matt Phillips* (55% time for two years) - Lead Developer
- *Adam Ziegler* (35% time for two years) - Project Manager
- *Jack Cushman* (50% time for two years) - Developer, Innovation Fellow and Project Evangelist
- *Claire DeMarco* (30% time for two years) - Librarian, Library-Partner Lead

6. Communications Plan

Our communications plan flows naturally from our project design and largely will reflect a continuation and expansion of activities already under way. At the beginning of the project, we will convene advisory boards who can help guide our library community-building efforts and potential paid markets. We will participate actively in library conferences over the project period, as well as web archiving events like IIPC which we have joined previously.

We will continue to use our blog to publish and share product information, educational information about link rot, and to highlight work done by our colleagues and library partners to study and combat link rot and
to help their communities use Perma.[8] We also will continue to update, refine and expand our user (https://perma.cc/docs) and developer documentation (https://perma.cc/developer-docs). We communicate actively with our library-partners through email and host a library-partners email listserve. Our info@perma.cc email account is monitored by a team that filters, re-assigns and responds to inquiries and feedback. We also will continue to communicate our development progress, bugs and feature requests in the open on Github, and we will continue to publish our usage statistics on the Perma site.

We will continue to pursue opportunities to speak and write about the link rot problem and how Perma can help. In recent months, members of our team have spoken about link rot and Perma at the ALA Midwinter Meeting, the Georgetown Link Rot Symposium, the Law Repository Symposium at William & Mary, Columbia University's Web Archiving Collaboration conference, the New England Library Association Annual Conference, and the Law Library and Legal Information Section's luncheon at the American Association of Law Schools Annual Meeting. We will have an active social media presence and content marketing operation, and we have budgeted a portion of the requested funds for marketing and communications personnel to expand this effort. And we will continue to be aggressive in reaching out directly to potential library partners, groups of potential users (such as professional associations) and partners in efforts to stop link rot.

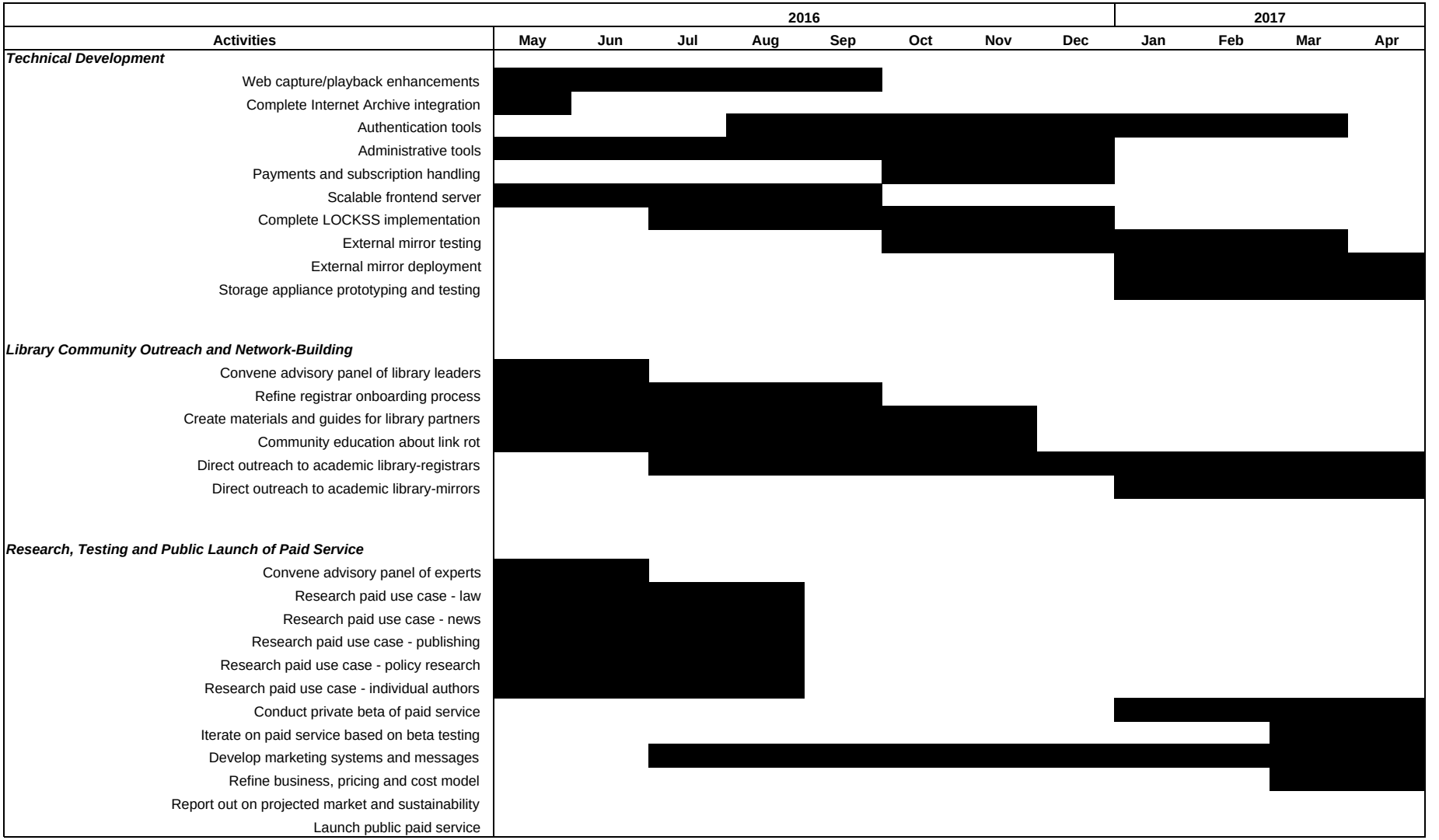
7. Sustainability

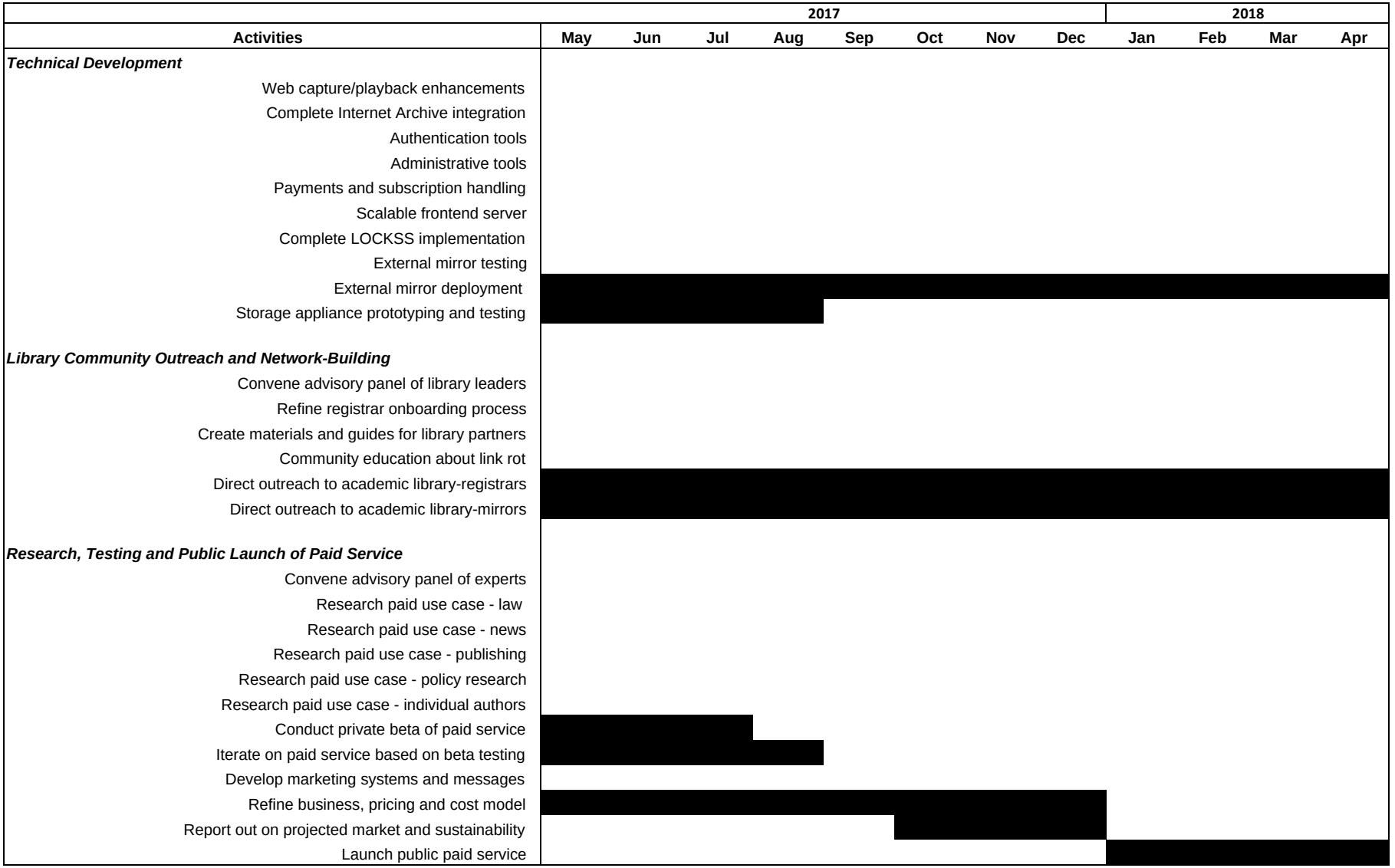
Sustainability is a key component of this project, and much of our effort will be dedicated to designing, testing and launching a paid service that will allow Perma to sustainably scale and eventually achieve economic self-sustainability.

Evidence of demand for a paid service is strong, particularly in five potential markets: (1) law (law firms and legal publishing); (2) publishing; (3) news; (4) policy research; and (5) individual authorship. Our letters of support demonstrate some of this demand, as do the unsolicited inquiries we have received from numerous law firms, publishers, news outlets, research centers and think-tanks and individuals.

[8] For example, in the legal field, colleagues at law libraries at Boston College, Washington University, Villanova, Ohio State University, Indiana University, UCLA and elsewhere have developed their own online resources for journals and/or faculty using Perma.

As described in detail above, our project is designed to validate this demand, develop a concrete understanding of the size and dynamics of these markets, and hone our product, value proposition and marketing messages to fit the needs of these users. We also will be closely studying issues of incremental costs and revenues and, relatedly, pricing. At the end of this project, we expect to be in a position to publicly launch a paid service that will subsidize the free-of-charge service provided to public and academic users of Perma, thus sustainably scaling Perma.cc to combat link rot in all scholarly fields.


2016
2017
Activities
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Jan
Feb
Mar
Apr
Technical Development
Web capture/playback enhancements
Complete Internet Archive integration
Authentication tools
Administrative tools
Payments and subscription handling
Scalable frontend server
Complete LOCKSS implementation
External mirror testing
External mirror deployment
Storage appliance prototyping and testing
Library Community Outreach and Network-Building
Convene advisory panel of library leaders
Refine registrar onboarding process
Create materials and guides for library partners
Community education about link rot
Direct outreach to academic library-registrars
Direct outreach to academic library-mirrors
Research, Testing and Public Launch of Paid Service
Convene advisory panel of experts
Research paid use case - law
Research paid use case - news
Research paid use case - publishing
Research paid use case - policy research
Research paid use case - individual authors
Conduct private beta of paid service
Iterate on paid service based on beta testing
Develop marketing systems and messages
Refine business, pricing and cost model
Report out on projected market and sustainability
Launch public paid service


2017
2018
Activities
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Jan
Feb
Mar
Apr
Technical Development
Web capture/playback enhancements
Complete Internet Archive integration
Authentication tools
Administrative tools
Payments and subscription handling
Scalable frontend server
Complete LOCKSS implementation
External mirror testing
External mirror deployment
Storage appliance prototyping and testing
Library Community Outreach and Network-Building
Convene advisory panel of library leaders
Refine registrar onboarding process
Create materials and guides for library partners
Community education about link rot
Direct outreach to academic library-registrars
Direct outreach to academic library-mirrors
Research, Testing and Public Launch of Paid Service
Convene advisory panel of experts
Research paid use case - law
Research paid use case - news
Research paid use case - publishing
Research paid use case - policy research
Research paid use case - individual authors
Conduct private beta of paid service
Iterate on paid service based on beta testing
Develop marketing systems and messages
Refine business, pricing and cost model
Report out on projected market and sustainability
Launch public paid service

DIGITAL STEWARDSHIP SUPPLEMENTARY INFORMATION FORM

Introduction
The Institute of Museum and Library Services (IMLS) is committed to expanding public access to federally funded research, data, software, and other digital products. The assets 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 is not always straightforward. Because technology is dynamic and because we do not want to inhibit innovation, we do not want to prescribe set standards and best practices that could become quickly outdated. Instead, we ask that you answer a series of questions that address specific aspects of creating and managing digital assets. 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
If you propose to create any type of digital product as part of your project, complete this form. We define digital products very broadly. If you are developing anything through the use of information technology (e.g., digital collections, web resources, metadata, software, or data), you should complete this form.

Please indicate which of the following digital products you will create or collect during your project
(Check all that apply):

	Every proposal creating a digital product should complete …	Part I
	If your project will create or collect …	**Then you should complete …**
☑	Digital content	Part II
☑	Software (systems, tools, apps, etc.)	Part III
☑	Dataset	Part IV

PART I.

A. Intellectual Property Rights and Permissions

We expect applicants to make federally funded work products widely available and usable through strategies such as publishing in open-access journals, depositing works in institutional or discipline-based repositories, and using non-restrictive licenses such as a Creative Commons license.

A.1 What will be the intellectual property status of the content, software, or datasets you intend to create? Who will hold the copyright? Will you assign a Creative Commons license (http://us.creativecommons.org) to the content? If so, which license will it be? If it is software, what open source license will you use (e.g., BSD, GNU, MIT)? Explain and justify your licensing selections.

Harvard University will hold copyright. All software will be released under an MIT license, which is selected to allow for the broadest possible use by others.

A.2 What ownership rights will your organization assert over the new digital content, software, or datasets and what conditions will you impose on access and use? Explain any terms of access and conditions of use, why they are justifiable, and how you will notify potential users about relevant terms or conditions.

All software will be publicly available on GitHub without limitation.

Access to http://perma.cc (the version of the software operated by Harvard at our expense) will be limited according to the business model described in our submission package, and the terms and conditions published at http://perma.cc and updated from time to time with notice to users.

A.3 Will you create any content or products which may involve privacy concerns, require obtaining permissions or rights, or raise any cultural sensitivities? If so, please describe the issues and how you plan to address them.

Perma.cc hosts content captured from the public internet at user direction. As a neutral content host, we cannot control in advance what content is hosted. We will disable access to content at the direction of the submitting user or their organization, or if the content is reported to us and found to violate our terms.

Part II: Projects Creating or Collecting Digital Content

A. Creating New Digital Content

A.1 Describe the digital content you will create and/or collect, the quantities of each type, and format you will use.

We capture content from the public internet at the direction of users, in the form of web archive (WARC) files containing original web assets as well as browser screenshots. We expect to capture hundreds of thousands of such files at an average size of 2 megabytes each.

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

We use custom Python software to capture web archives. The software will be hosted by Amazon Web Services and Heroku.

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

- WARC web archives -- conformant with ISO 28500:2009
- PNG browser screenshots at browser width 1024px

B. Digital Workflow and Asset Maintenance/Preservation

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

Content is captured at the direction of Perma users and reviewed by them. Each Perma page offers a form for users to report capture errors, which are then evaluated by our developers.

B.2 Describe your plan for preserving and maintaining digital assets during and after the award period of performance (e.g., storage systems, shared repositories, technical documentation, migration planning, 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 CFR 200.461).

Perma's source code is stored in the git version control system both on local computers and publicly on github.com.

Perma's archives will be stored in a Private LOCKSS Network (PLN) hosted by multiple institutions. LOCKSS complies with the highest standards in distributed archival of digital objects (e.g. the OAIS Reference Model) and existing PLNs have passed FIPS and TRAC audits.

C. Metadata

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

Each WARC archive contains metadata encoded according to the WARC ISO standard, specifying the provenance of assets contained within the WARC.

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

Because metadata is encoded within the WARC files, it is automatically preserved along with the underlying data.

C.3 Explain what metadata sharing and/or other strategies you will use to facilitate widespread discovery and use of digital content created during your project (e.g., an API (Application Programming Interface), contributions to the Digital Public Library of America (DPLA) or other digital platform, or other support to allow batch queries and retrieval of metadata).

Perma.cc will expose its archives through an API, through an RFC7089 Memento HTTP endpoint, and through upload to the Internet Archive.

D. Access and Use

D.1 Describe how you will make the digital content 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).

Perma archives will be openly available online and publicly accessible via standard web browsers.

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

In addition to the existing Perma.cc website, the Harvard Library Innovation Lab is responsible for projects such as:

Free the Law: http://librarylab.law.harvard.edu/projects/free-the-law
The Nuremberg Trials Project: http://nuremberg.law.harvard.edu/
StackLife: http://stacklife.harvard.edu/
Haystacks: http://haystacks.law.harvard.edu/
H2O: http://h2o.law.harvard.edu/

Part III. Projects Creating Software (systems, tools, apps, etc.)

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) this software will serve.

Our software is a database-driven web application that captures links from the public internet at the direction of users. It allows users to organize and share links they have captured with members of their organization, and to cite the content of those links. It distributes the archived content to libraries and plays it back for readers. The primary audiences are academic publications, legal practitioners and courts, and general-interest publishers.

A.2 List other existing software that wholly or partially perform the same functions, and explain how the tool or system you will create is different.

Similar systems include web.archive.org, archive.is, and webcitation.org. Our system is primarily distinguished by (a) our long-term archival plan based on highly redundant storage; (b) high-quality, targeted web captures; (c) polished collaboration tools tailored for publishers and courts; and (d) effective community outreach, support, and stewardship supported by over 100 library partners.

B. Technical Information

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

We use the Python programming language and the Django web application framework because they are productive and widely used in the web development industry. We also use standard frontend web technologies such as HTML and Javascript.

B.2 Describe how the intended software will extend or interoperate with other existing software.

We use a variety of existing Python libraries for web archival and playback, such as warcprox, Selenium, and PyWB. We will use LOCKSS for distributed digital archival.

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

The software depends on a SQL database (MySQL 5), a wsgi web server (gunicorn and nginx), an in-memory cache (memcache), and a message queue (RabbitMQ). The system dependencies each use standard interfaces and the choice of software package for each role is configurable through our Django configuration file.

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

We maintain developer documentation as text files in the root of our open source repository. New developers use the documentation to set up their development environment and are encouraged to submit revisions if the documentation is incomplete.

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

Library Innovation Lab open source projects are hosted at https://github.com/harvard-lil . We host some 20 repositories (including perma) with updates within the last year.

C. Access and Use

C.1 We expect applicants seeking federal funds for software to develop and release these products under an open-source license to maximize access and promote reuse. What ownership rights will your organization assert over the software created, and what conditions will you impose on the access and use of this product? 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 any prohibitive terms or conditions of use or access, explain why these terms or conditions are justifiable, and explain how you will notify potential users of the software or system.

All software will be released under an MIT license, which is selected to allow for the broadest possible use by others.

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

Our hosted version of the Perma software will be available to users at http://perma.cc. The product will be free for academic users and courts, as well as to individuals with a limited use per month. The source code will remain freely available and open source.

C.3 Identify where you will be publicly depositing source code for the software developed:

Name of publicly accessible source code repository: perma
URL: https://github.com/harvard-lil/perma

Part IV. Projects Creating a Dataset

1. Summarize the intended purpose of this data, the type of data to be collected or generated, the method for collection or generation, the approximate dates or frequency when the data will be generated or collected, and the intended use of the data collected.

Perma will not create a "dataset" as such. However, its primary use is to capture and host content at the request of individual users. It therefore collects data such as: the user submitting a public URL; the date; and the content of the URL. Users are directly aware of the data collected and published through their normal use of the service.

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?

Perma does not require approval by an IRB.

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).

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.

5. What 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).

6. What documentation (e.g., data documentation, codebooks, etc.) 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?

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

8. Identify where you will be publicly depositing dataset(s):

   Name of repository:
   URL:

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

Original Preliminary Proposal

Scaling Up Perma.cc: Ensuring the Integrity of the Digital Scholarly Record
Harvard Library Innovation Lab - IMLS Preliminary Proposal

Executive summary: Link rot - the tendency for URL citations to break after publication - is a critical issue facing libraries as the custodians of published scholarship. Over the past two years, the Harvard Library Innovation Lab, with the help of partner law libraries, has proven that we can prevent link rot in the legal field. We and our partner libraries, in cooperation with the Berkman Center for Internet & Society, now request $950,000 from IMLS to expand to all academic publishing, grow our partner network to include academic libraries of all disciplines, and to explore a viable commercial model.

Problem statement and field wide need: "Link rot" refers to the simple fact that websites, unlike books, can change their content or disappear after an author cites them. A study we conducted revealed that 50% of Supreme Court cases and 70% of law review articles with web citations suffer from link rot.[1] The same 70% level of link rot was recently confirmed in a broad study of science, technology, and medicine articles.[2] Often these "rotted" citations refer to material that cannot be cited in print form and cannot be reconstructed if lost. Academic scholarship, which depends on a network of citations going back hundreds of years, is now being written on sand.

Work already completed: Libraries played a key role in the analog world as the trusted repositories of the scholarly record. Over the past two years, we have proved that we can serve the same role in the digital world. A broad coalition of libraries, paired with cutting-edge web archiving and distributed data storage tools, has been uniquely effective in combating link rot.

Our website, Perma.cc, solves the link rot problem by allowing authors to take snapshots of web pages and place them on deposit with university libraries. Each Perma record has a unique URL (e.g. https://perma.cc/F37P-2E4V) that readers can follow to view the preserved snapshot. With our high-quality captures, group management features, and distributed network of trusted library partners, Perma has quickly become the best available preservation tool for published web citations.

Since launching in September 2013, and as of September 2015, Perma has been adopted by over 100 academic law libraries, over 175 academic law journals, numerous law faculty members, several law-school affiliated research centers and by the top courts in six states. The Bluebook legal citation manual, the citation style guide for the legal world, now recommends Perma as a "reliable web archiving service" and advocates its use. Perma has helped over 6300 individual users preserve over 135,000 individual web sources. Perma earned the 2015 Webby Award in the Law category, and has been featured in the *New York Times*, the *New Yorker*, *Fast Company*, and *NPR*. Unlike other fields of scholarship, and over a very short period of time, the link rot problem in the field of law is well on its way to being solved.

Project director & partners: Perma's success is due to the marriage of a technically sophisticated development team with a wide network of libraries that promote and support the service with users. The project is directed by Professor Jonathan Zittrain, faculty director of the Harvard Law Library. Also on the core team are Kim Dulin, Matt Phillips, Adam Ziegler, Jack Cushman, Annie Cain, Claire DeMarco and others.

---

[1] Jonathan Zittrain, Kendra Albert & Lawrence Lessig, *Perma: Scoping and Addressing the Problem of Link and Reference Rot in Legal Citations*, 127 Harv. L. Rev. F. 176 (2014).

[2] Seven out of ten STM articles containing web citations suffer from reference rot: Klein M, Van de Sompel H, Sanderson R, Shankar H, Balakireva L, Zhou K, et al. (2014) Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot. PLoS ONE 9(12): e115253. doi:10.1371/journal.pone.0115253.

Beyond Harvard, Perma works with a network of 128 partner organizations known as "registrars," including 104 academic law libraries, the Library of Congress, and the Digital Public Library of America. We collaborate and share technology with organizations such as the Internet Archive, Amber (a Berkman Center archiving project), Memento, and LOCKSS. And we have received technical assistance and in-kind support from private organizations such as Cloudflare and Dyn.

Proposed work plan: Perma runs on an "agile startup" model: we release features, incorporate user feedback, and refine our goals on a weekly schedule. Throughout the grant period we will continue this successful development approach while gradually adding academic library partners in all fields of study as well as commercial users. The focus will be to *sustainably scale* what we have created, arresting the link rot problem throughout academic publishing while building a financial model for further growth.

On the academic side, we will begin adding non-law academic registrars, a number of whom have already approached us. Supporting these new partners will require improvements to the Perma infrastructure in a variety of ways, such as: expansion of the Perma mirror network; improvements to our web capture systems; authentication tools for libraries with large numbers of members; development of affordable, scalable frontend servers; and improvements to administrative tools. These examples are not fixed, but will respond on a weekly basis to feedback from our broadened user base; the result will be a Perma that is cheap and reliable to run while supporting that broad base.

On the commercial side, we will begin offering Perma services to law firms and private publishers (again, a number of whom have already approached us) on a paid basis. This effort will allow us to continue sustainably broadening Perma access well beyond the grant period.

Potential impact, performance goals and outcomes: The potential outcome of IMLS support is a robust, sustainable service, maintained by libraries and sustained by commercial users, that substantially defeats the link rot problem now threatening all fields of scholarship. We measure our progress by Perma adoption among libraries and publications within a field; for example, Perma is now supported by almost 50% of American law school libraries. Over the grant period we will pursue a goal of partnering with at least 5-10% of the 3,793 academic libraries in the U.S.

Relevance to agency priorities: Perma not only exemplifies the goals of the National Digital Platform, it proves their value. It is *open source,* all of its code is on GitHub, and users can and have submitted changes. It is *interoperable*, based on Memento, LOCKSS, Web Archive format, and other open standards and protocols, and publishes its own powerful API that is already in use by CanaryWatch and others. It is *user-centered* in its design process and impact, nimbly responding to user needs, and offering individual users the power to deposit documents that are important to *them* in illustrious libraries. It combines the *automated* with the *human*, putting sophisticated capture and storage tools at the fingertips of human authors. And it combines the *global* with the *local,* it depends on the expertise and evangelism of local librarians and their contributions of distributed data archives, coordinated by a modern, high-availability web service. Above all, Perma is already a success story for library R&D, and one that is poised to go much farther, faster with IMLS support.

Budget: The estimated two-year budget totals $950,000 with an equal cost-share contribution of $950,000 by Harvard Law School. That budget would be allocated among technical personnel, outreach and community relations personnel, project management and oversight, legal and accounting, travel and community-building events, and equipment, software and network infrastructure.