Abstract

The Institute for Quantitative Social Science (IQSS) at Harvard University proposes a 12-month National Leadership project to analyze the characteristics and activities of ten open source projects in order to develop a set of *open source health indices* or *quantitative standards* that measures the sustainability and long-term viability of open source projects, with a particular focus on open source projects useful to research and academic libraries. The project will be led by Dr. Mercè Crosas, Chief Data Science and Technology Officer at IQSS. The timeframe of this project is October 2018 through September 2019. This project will prepare us for a subsequent two-year study that will test and elaborate on our findings.

Open source software (OSS) is a viable, popular means of connecting the efforts of libraries and museums from disparate locations and contexts in order to realize the IMLS vision of a national digital platform. However, organizations engaging with OSS may find it challenging to evaluate the health and sustainability of a given OSS project. We aim to expand the resources on OSS in libraries by developing a set indices functions that produce a set of scores with which to quantify the health of an open source project. With these indices, we aim to provide aid to research and academic libraries that are: (1) using OSS; (2) integrating with OSS; (3) looking to evaluate OSS for potential use; and (4) developing OSS.

In this *planning* project, we will: (1) engage with partners from ten open source projects to inform our approach; (2) select and review a set of 50 to 100 quantitative metrics that measure the activity of an open source project, and group them into 4-6 categories useful to the research and academic Library community; (3) host an experts workshop (including an advisory board with library expertise) at the halfway point to review our work before collecting the data; and (4) collect the data for the ten open source projects, form a hypothesis for a set of quantitative *open source health indices*, one for each category, that will aim to measure the success of an OSS project, and design the 2-year quantitative study.

Librarians and other decision makers need to be able to easily evaluate open source projects for potential use by using an evidence-based and reproducible approach, such as the one proposed in this project. We envision not only more informed, more effective investments in open source throughout the library community, but also shortened adoption cycles and greater sustainability following adoption. OSS projects themselves will be able to respond to the highest priority concerns, communicate more effectively with their members, and deliver more value to the institutions that use their software.

Understanding What Constitutes a Vibrant Open Source Community

Introduction

Many research and academic libraries build or make use of open source software to provide affordable and sustainable tools for the community and built by the community. But how does an open source software project become sustainable? What does it take to build a growing and vibrant open source community? Which types of open source projects are worth adopting and which are risky investments?

To answer these questions, we are proposing a 12 month planning project to analyze the characteristics and activities of ten open source projects, which we believe to be successful in various forms, in order to develop a set of *open source health indices* that seek to quantitatively measure the sustainability and long-term viability of open source projects. Following this planning phase, we intend to run a larger two-year quantitative study that will test and elaborate on our early findings.

In this planning phase, we will: (1) engage with partners from ten open source projects to inform our approach; (2) select and review a set of 50 to 100 quantitative metrics that measure the activity of an open source project, and group them into 4-6 categories useful to the research and academic Library community; (3) host an experts workshop(including an advisory board with library expertise) at the project's halfway point to review our work before data collection begins; and (4) collect the data for the ten open source projects, form a hypothesis for a set of quantitative *open source health indices*, one for each category, that will aim to measure the success of a project, and design the 2-year quantitative study.

Statement of National Need

Over the past few decades, libraries and other information institutions around the world have increasingly turned to open source software (OSS) for its cost-effectiveness, relatively open licensure, and modifiability to suit institutions' unique needs (Payne & Singh, 2010). IMLS recognizes the importance of the *national digital platform* -- the idea of shared digital infrastructure and collaborative development of digital tools and services across information institutions. OSS represents a viable, popular means of connecting the efforts of research and academic libraries from disparate locations and contexts in order to realize this vision of a national digital platform.

Libraries engaging with OSS may find it challenging to evaluate and improve the health and sustainability of a given open source project. Recent studies like Lyrasis' *It Takes a Village: Open Source Software Sustainability* have provided practical advice to help librarians evaluate and run successful open source projects. Other recent projects such as CHAOSS (https://chaoss.community/) and Tidelift (https://tidelift.com) aim to assist OSS health by providing tools to collect metrics and visualize the activity of a project, or managing tools to implement a sustainable OSS, respectively.

We aim to expand upon and leverage these projects to conduct a quantitative study with the goals of 1) formulating a set of scores or indices useful to research and academic libraries that evaluate the health of OSS and 2) collecting the data to calculate these indices for OSS relevant to these libraries. In particular, we will leverage the OSS activity metrics proposed by the project CHAOSS and use their tools, when appropriate, to collect these metrics.

Our background in developing open source software, our wealth of quantitative social science research expertise, and our relationships with open source developers across various domains give us a unique perspective by which to conduct this research.

By analyzing a mixture of open source projects from within and without the library world, we aim to provide crucial aid to research and academic libraries that are: 1) using OSS; 2) integrating with OSS; 3) looking to evaluate OSS for potential use; and 4) developing OSS.

Additionally, we aim to glean insights that will prove helpful not just to the library community, but also to the wider open source community.

Project Design

Goals

In accordance with IMLS's goal of fostering a national digital platform, we aim to aid research and academic librarians in making effective and collaborative use of OSS by focusing on the following goals:
1) formulate a set of scores or indices useful to research and academic libraries with which to quantify the health of any given open source project.
2) collect the data to calculate these indices for open source software relevant to research and academic libraries.
3) establish a method by which to collect the data for a larger number of OSS and be able to ask research questions about what makes a project work in the future.

Toward these goals, we will conduct a 12-month preliminary investigation into what aspects of an open source project can be seen as indicators of the project's "health" and group them into 4-6 broad categories useful to research and academic libraries. We will begin by identifying ten OSS projects we think to be successful to partner with for our research. Through close communication with an advisory board with library expertise and our OSS partners, we will determine what "healthy" means for an open source project, in particular the context of how OSS are created or used within libraries.

We will then confirm and expand on the selected quantitative variables that serve as metrics to measure OSS health by hosting an experts workshop featuring our advisory board, representatives from our partner projects, and experts from the wider open source community. Finally, we will gather data from the 10 selected OSS projects, form

a hypothesis for a set of OSS health indices, and design and seek funding for a full 2-year study that will expand this work into a more extensively tested set of OSS health indices and a much larger data collection for OSS created and/or used by research and academic libraries.

This project assumes that it's feasible to identify a set of quantitative variables that can serve as highly generalizable indicators of health for all OSS projects. We will test this assumption by profiling ten OSS projects from varying fields. By pulling five of our partner projects from the library world and five from outside the library world, we ensure that our analysis will not be cloistered in one domain. However, the broader quantitative study proposed in the next phase, which will use 50-100 OSS instead of only the first 10 pilot ones, will be needed to test our hypothesis formulated during the planning phase.

Activities, Metrics, and Timeframe

The project period is 10/1/18 through 9/30/19. The main activities in this project are described below:

1: Form Advisory Board (Month 1)

We have already engaged 4 library representatives to be part of an advisory board that will help us to review and evaluate our activities and outputs (see letters of support). So far we have recruited these four representatives:

- Ceilyn Boyd, Research Data Program Manager, Harvard Library
- Christine Borgman, Distinguished Professor & Presidential Chair in Information Studies, UCLA; Director, UCLA Center for Knowledge Infrastructures
- Suzanne Wones, Associate University Librarian for Digital Strategies and Innovations, Harvard Library
- Wendy Gogel, Manager of Digital Content and Projects, Harvard Library

We plan to include 5-7 more experts from the LIS community. The advisory board will help us identify which OSS projects to profile and will further advise us over the course of our research.

2: Identify and engage with partners (Month 1 - Month 3)

We will begin by identifying a pilot set of ten open source projects that we will engage with and collect data from. We will work closely with these projects, communicating with members of their teams to learn about how they operate and what dimensions of the project have contributed to or taken away from their health and success. Five of our partners will be primarily research and academic library focused, and five will be primarily focused outside of the library world.

We have already identified our tentative five non-library-focused partner projects:

- Dataverse (King 2007; Crosas 2011)
- Globus (Chard, K. et al., 2015)
- Jupyter (Perez & Granger, 2015)

- Red Hat (Young, 1999)
- Open Nebula (Milojičić, Llorente, & Montero, 2011)

We have selected these projects because they are mature, they have been active for at least two years, and they have some connection to our institution, Harvard's Institute for Quantitative Social Science (IQSS). Having a prior connection to these projects will aid us in gathering detailed, accurate information about their activity metrics, the way they run, the state of the project, and the challenges they face. The Dataverse project is developed, managed, maintained, and sustained by IQSS. We already have an existing relationship with the rest of the projects, and plan to establish a formal collaboration within the 12-month planning phase.

The projects above are not commonly used by libraries (although Dataverse and Jupyter have increasingly been used by academic libraries in the last 2-3 years). We plan to additionally collect activity metrics from open source projects more typically used and/or built by research and academic libraries.

To do this, we will reach out to our proposed advisory board to find 5 more relatively mature open source projects that these institutions use and/or develop, and that they consider to be successful.

3: Define variables that measure the health of an OSS (Month 3 - Month 6)

During this phase, we will determine what metrics we plan to use for our research. These metrics will be defined and collected as quantitative (numerical) variables. We understand that some information about an OSS could be best described using qualitative (categorical) variables. However, our goal is to define a set of indices or scores that quantitatively evaluate several aspects of an open source project. When possible, we will convert a metric originally thought as qualitative to a quantitative one. For example, instead of defining a variable as "list countries that use the OSS", we would define it as "number of countries that use the OSS". Or instead of a question (variable) such as "what type of license does the OSS have?", we might define the variable as "does the OSS has a license", and the values can be 0 (No) or 1 (Yes).

Reviewing Metrics

We plan to leverage extensively the OSS activity metrics proposed by the project CHAOSS. The project includes 130 metrics (see https://github.com/chaoss/metrics/blob/master/activity-metrics-list.md). Examples of these metrics are: age of community, average time of open issues, number of contributing organizations, project lifecycle. In addition to our own experience with open source projects and communities, we will seek input from the advisory board and from the partners from other OSS to review whether these metrics are appropriate and relevant to research and academic libraries for evaluating OSS. After the review, we might decide to remove some metrics from the CHAOSS project, or to modify them to make them quantitative (if they aren't already) and useful for our research, or to add new metrics that are important to libraries but have not been considered in the

CHAOSS project. We expect to end up in the order of 100 metrics (variables) for our study.

We plan to reach out to the CHAOSS governing board to engage in their project and propose a subset of metrics for a type of OSS projects, such as *library OSS metrics*. This would mean that that there would be a list of metrics common to practically all type of OSS projects, and then subsets of metrics where each subset is useful for a particular type of OSS project (for example, library, universities, industry, data science, etc).

Grouping of Metrics into Categories
The CHAOSS project uses 130 metrics to describe 4 categories or aspects of OSS: 1) Diversity-Inclusion; 2) Growth-Maturity-Decline; 3) Risk; 4) Value. Each one of the categories includes a subset of the 130 metrics, and a metric can belong to one or more categories. We will review these categories with the advisory board and OSS partners to evaluate their relevance when focusing on OSS for libraries. Similar to the method above for selecting the appropriate metrics for our research, after the review process we might decide to use or modify these 4 categories, or create new ones that better describe what's important for research and academic libraries. We expect to select 4-6 categories for our study.

Formulation of OSS health Indices
We will formulate a proposed index for each one of the categories defined. We will construct each index as a function of the metrics scores that an OSS project received. The function applied to each index will be determined by us and reviewed by the participants in the experts workshops (see Point 4 below). The function would be tailored to the specific context. For consistency, it would be helpful to have all indices be on the same scale. To ensure that, all metrics marks will be on the scale from 0 to 1 or 0 to 100.

This indices can be understood conceptually similarly to the democracy index defined by the Economist Intelligence Unit (EIU, 2017), which measures the state of democracy of 167 countries based on the weighted average of the experts' answers to 60 questions that evaluate democracy. However, in our case, we won't necessarily apply a weighted average as our index function, as it is used to calculate the democracy index and other similar index. Instead, we plan to review the literature for more appropriate functions for our case. For example, the function proposed by King & Murray (2002) takes into account that for if a variable's value is 0, then the entire index should be 0, which might be more appropriate for some of the indices we plan to formulate.

4: Host an experts workshop (Month 6)

Halfway through the project, we will convene a group containing our advisory board representing research and academic libraries, together with open source experts and selected members from the open source projects participating in the study in order to review the study's design up to that point. At this workshop, we will review the variables we have identified that define a healthy open source community for OSS useful to

libraries, and alter or replace them as recommended. We will also review the categories and proposed function to calculate the corresponding indices. The workshop will take place in month 6 of the project, to ensure that we can incorporate the feedback it generates into our data gathering process and eventually our final study design.

5: Collect data (Month 6 - Month 10)

Following the experts workshop, we will finalize our set of variables to measure the growth, success, and health of an OSS project.

At this point we will begin gathering data on each variable we identified for the 10 selected pilot OSS. More than one method will be used for the data collection:

a) For metrics that can be collected in an automated way (for example, numbers that can be obtained from GitHub through their Application Programming Interface or API), we will leverage whenever possible the tools developed by the CHAOSS project or code the needed scripts to extract the information. When developing our own scripts, we will use a quality assurance process to test the scripts to ensure that they provide the correct results.
b) For metrics that cannot be easily collected automatically through an API or existing tool, we will collect the data via two independent methods (triangulation): 1) using a web-based survey to obtain the variables' values that will be submitted to OSS partners (using Qualtrics to define and distribute the survey), and 2) using the documentation and website provided by the OSS to collect the data manually. In some cases, we will add an additional method - interview - to obtain or clarify information from the OSS partners. The information gathered through these methods will be used to fill in the quantitative values of the metrics (variables) that cannot be collected automatically. This process will require for some variables a coding step, and therefore we will employ intercoder reliability to validate the results.

6: Calculate OSS health indices, review results, and propose 2-year study design (Month 10 - Month 12)

Using the collected data, we will calculate the proposed indices for the 10 pilot OSS projects, review the results, and improve the indices' formula as needed. We also plan to review the results with the OSS partners and get their feedback.

We will also use this period to finalize a detailed design for the follow-up study. This longer-term study will include a similar quantitative approach to evaluate approximately 100 OSS used by or potentially useful to research and academic libraries, collect the data, and calculate the set of indices. This larger study will allow us to:
- validate the usefulness and accuracy of the indices' values
- find natural categories with which to group the metric in a way that provides useful indicators of OSS health

- conduct a longitudinal study collecting the data for some OSS projects from 5 years from now, and each consecutive year to evaluate whether is possible to predict the success or failure of a project
- provide a report of OSS useful to libraries with a consistent evaluation that can be reproduced by applying a set of well-defined methods.

Outputs and Outcomes

The outputs of this planning project will include:

1. A data documentation (codebook) describing all the quantitative variables used to measure open-source projects
2. A dataset with the data collected for the ten selected open source projects for each variable. This will be a tabular dataset with 10 rows (the 10 pilot OSS projects) and approximately 100 columns (the quantitative variables collected for each project).
3. A hypothesis on how to calculate a set of *open source health indices,* one for each category, including a paper documenting the methods used, and the output values of these indices for the 10 OSS projects evaluated in this pilot phase.
4. A detailed design for a two-year follow-up study to confirm and expand upon our work.

Our expected outcomes will include assisting the research and academic library community in evaluating OSS projects developed by others but useful to libraries, and improving OSS projects developed within the library community. Eventually our project will help to answer questions such as “How important it is to focus on *this vs that* to improve *this* open source project?” or “Can I predict whether *this* OSS is in decline?” using an evidence-based approach.

Risks

Working across a wide variety of open source projects in the library and non-library space, it may be challenging to create a comprehensive set of OSS indices that accurately measures the health of every project, due to differences in project activities and communities served. If this turns out to be the case, we plan to create a typology of OSS projects, grouping projects of similar types together and creating tailored indices for each specific group.

Other risks will be mitigated through the team’s previous experience in open source development and the organization’s deep experience in quantitative study design and research.

Resource Requirements

We request a total of $50,000 for a 12-month planning grant. After this planning phase, to help us conduct the full study, we intend to request funding from IMLS and from the Sloan Foundation.

This 12-month planning grant will be led by PI Mercè Crosas with two research assistants (working 3.0 calendar months combined), and with the support of the Dataverse team at Harvard's Institute for Quantitative Social Science. The Dataverse team develops and supports a software platform and extended tools to create repositories for research datasets, which are accessed by researchers and library personnel at academic and research institutions around the world. The Dataverse team has experience developing open source software, creating and growing a community of contributors from more than 20 countries, and fostering active communication with developers and users.

We seek additional funding from the Sloan Foundation to complement and enhance the work proposed here. The Sloan Foundation has already provided preliminary agreement to extend the funds, should the project be awarded by the IMLS. The funds would be used to 1) extend the study to focus not only on library OSS, but also more general OSS developed by or useful to universities; 2) extend the experts workshop to a larger group, allowing us to bring people from outside the local area; 3) expand the team working on this project to allow for additional accuracy and validation of data collection and extend the number of pilot projects to include OSS widely used by universities; and 4) conduct preliminary analysis of the collected data and further test the proposed health indices before extending the study to a much larger number of open source projects.

Evaluation and Performance Measurement

This research will be carried out by a mature team working under Agile project management. Our team conducts daily "stand-up" meetings where we exchange status updates and facilitate cooperation among team members. We track the status of long-term projects using GitHub issues which we aggregate into a timeline using Waffle.io. Regular "backlog grooming" and "sprint planning" meetings get the whole team together to create a detailed work plan for each two-week "sprint" period. While this kind of project management is most common in software development, we have made effective use of it for large-scale UX research projects as well. We will be using this project management strategy to track our progress toward achieving the goals of this project.

To evaluate our performance and ensure we are on track, we will frequently refer to our advisory board of librarians for advice. Additionally, halfway through the project our experts workshop will intensely scrutinize our work up to that point and give us guidance for the second half of the project. When collecting data, we will rigorously evaluate the values we collect by cross-checking the inputs to ensure accuracy in coding. Following this planning project, we plan to conduct a two-year study that will evaluate the planning project's hypothesis by using the OSS health index we will have developed to profile further open source projects, determining how the health index holds up to broader use.

Dissemination

At the end of this planning project, we will deposit our dataset and data documentation (codebook) into the public data repository Harvard Dataverse, where it will be fully accessible to the public without any restriction.

We will also create a publicly accessible website with information about the project and the proposed OSS health indices. We will create this website using OpenScholar, Harvard's academic website publishing platform. We aim to write a paper about the methods used to establish the proposed OSS health indices, but we do not anticipate the paper to be ready for publication during the period of this 12-months project.

Diversity Plan

As a free, community supported option, open source software is critical for groups building infrastructure supporting underserved and diverse communities. The licensing options that allow for adoption without cost, the ability to change the code to better serve a specific population or community, and the lack of recurring subscription charges make open source a valuable option for underserved populations lacking the resources to run commercial software.

Further, by providing easily accessible metrics of project health, we will help decision makers in those communities spend less time and resources on evaluations related to software adoption, allowing them to instead spend time on serving constituents. Variables studied will likely include areas related to ease of adoption, access to free support, costs to run the service, and other areas that may be important to such stakeholders.

National Impact

Driven by librarians and the software that enables them to do their jobs, we will create a rubric for judging the health of open source projects, drawing from our experience in open source development, quantitative research, and investigating and learning from success models beyond the LIS community. The metrics we will gather can be used by a diverse set of organizations to serve their communities, with a clear focus on research and academic libraries. Librarians and other decision makers need to be able to easily evaluate open source projects for potential use, streamlining the investigation and evaluation process. We see not only more informed, more effective investments in open source throughout the library community, but shortened adoption cycles and greater sustainability following adoption.

Beyond those adopting the software, the open source projects themselves will also see significant benefit, extending the impact further. By receiving feedback about the health of the project and the respective community, projects will be able to respond to the highest priority concerns, communicate more effectively with their members, and deliver more value to the institutions that use their software.

Outside of LIS fields, open source projects of all types, from industry and private entities, can expect to see benefit from the study's outputs and will be able to participate in and inform any future work. The research project that this planning grant enables will result in scholarly output that will serve as a useful reference for any institutions looking to evaluate or run an open-source project. It will be particularly tailored to a librarian audience, but we will ensure that our results and analysis are generalizable across domains, increasing the potential impact of the work.

All research outputs will be made available on the project website and on the Harvard Dataverse online repository. Other groups will be able to replicate the study or use the study's data for further research. Providing full access to data is a core value of IQSS, and making the entire output public is a default for this study as well as others.

Success will be determined by completing data collection, creating a dataset reviewed by members of the advisory team, publishing the data in Harvard Dataverse, and creating a hypothesis for the health indices. Each of these items will serve as milestones across the project during the 12 month planning phase.

To test and improve the OSS health indices we intend to run a future two-year study that will look at approximately 100 additional open source projects.

Schedule of Completion

Key Tasks	Duration (months)											
	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12
Form advisory board with LIS expertise												
Identify and engage with OSS partners												
Define metrics (quantitative variables) that measure an OSS												
Host an experts workshop												
Collect data for pilot OSS projects												
Estimate OSS Health Indices, review results, design two-year follow-up study												

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

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

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

Part I: Intellectual Property Rights and Permissions

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

The datasets and codebooks produced as part of this project will be published using a Creative Commons CC0 license in order to maximise openness and reusability. They will be made available through the Harvard Dataverse repository (https://dataverse.harvard.edu/), a publicly open data repository used by researchers world-wide to publish, cite, and archive research data.

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.

The products of this project will be fully open, with no restrictions on access or use.

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.*

This project will collect information about open source projects, and not about individuals running these projects. Therefore, no IRB approval will be needed.

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.*

The outputs of this project will include:

- a website containing information about the project and an open source health indices and the description of the proposed functions for each index
- A dataset in CSV format.
- A data documentation or codebook as a PDF
- A description to the methods to generate an OSS health indices as a PDF
- Additional information about the project in HTML

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.*

- Google Drive and Google Docs
- GitHub
- Dataverse
- Adobe Acrobat
- Open Scholar

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).*

- PDF
- CSV

- JSON
- HTML

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).*

Our advisory board of librarians will review our work periodically. Additionally, halfway through the project our experts workshop will scrutinize our work up to that point and give us guidance for the second half of the project. When collecting data, we will rigorously evaluate the values we collect by cross-checking the inputs to ensure accuracy in coding inn cases that is needed.

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).*

During the award period of performance, we will work using Google Drive to share our documentation and data sets to collaboratively collect the data and write and review the documents. We will use GitHub (under https://github.com/iqss) to track the project tasks, and link to the documents and data in Google drive, while the project is in progress. The tasks in GitHub will be labeled and documented appropriately to follow their progress.

After the award period of performance, we will store outputs from the project in Harvard Dataverse, an online repository with a robust preservation policy, viewable at https://dataverse.org/best-practices/harvard-dataverse-preservation-policy.

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 project outputs (that is, the full study with datasets and codebooks) will follow the standards supported by the Dataverse repository. These standard include: Dublin Core, Data Documentation Initiative, and Schema.org.

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

Metadata will be preserved with the outputs of the project as described in B.2 above.

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).

Metadata from the project outputs will be publicly available on the Harvard Dataverse online repository, both through the web-based user interface and through an API.

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).*

Information about the project will be published on a publicly available website using the Open Scholar platform. This will link, when appropriate, to the tasks in GitHub and the in-progress documents in Google Drive.

Our datasets and accompanying information will be deposited to the online repository Harvard Dataverse, where it will be publicly accessible.

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.*

- https://doi.org/10.7910/DVN/Y3WOOE
- https://doi.org/10.7910/DVN/CZYY1N
- https://doi.org/10.7910/DVN/HKHGWZ
- https://doi.org/10.7910/DVN/1EMHTK

Part III. Projects Developing Software
A. General Information

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

N/A

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

N/A

B. Technical Information

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

N/A

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

N/A

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

N/A

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

N/A

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

N/A

C. Access and Use

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

N/A

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

N/A

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

Name of publicly accessible source code repository: N/A

URL: N/A

Part IV: Projects Creating Datasets

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

The data will be collected during month 6 through month 10 of the project. Once all the (50-100) metrics are defined and reviewed, the data for 10 OSS projects will be collected in two main ways: 1) automatically by using scripts, 2) manually through gathering information from the documentation available in the website from each project or by surveying the individual projects. Details on methods are described in the project proposal.

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 information gathered will be about open source projects and not about individuals. No IRB approval is needed.

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.*

N/A

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

This is a quantitative study. The metrics collected will be given a quantitative value (if a metric is initially thought as a qualitative variable initially, it will be converted to a quantitative variable.). More information about methods is shared in the project description.

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?*

Our data documentation or codebook will be created alongside our dataset. It will be permanently stored with the dataset in the Harvard Dataverse online repository in a PDF format, and might include JSON to describe each variable.

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

Data from this project will be deposited into the Harvard Dataverse online repository, where it will be archived and made publicly available.

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

Name of repository: Harvard Dataverse

URL: https://dataverse.harvard.edu/

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

We plan to review this plan with the advisory board in month 2, in the workshop in month 6, and after data collection in month 10. At the end of the project, we will share the data collected and data documentation with the advisory board and with IMLS, by sharing the persistent DOI urls created by the Harvard Dataverse repositories, and the dataset and documentation will have a proper citation to be referenced in the future.