Abstract

The Montana State University (MSU) Library, in partnership with the MSU School of Computing, the University of New Mexico Library and DuraSpace, seeks a $49,998 Planning Grant from the Institute of Museum and Library Services through its National Leadership Grant program under the National Digital Platform project category.

This proposal will develop a sustainability plan for the *Repositories Analytics & Metrics Portal* (RAMP) that keeps its diagnostic capability and its dataset about institutional repositories (IR) performance available to the library community and other researchers. The proposal includes the development of a preliminary IR reporting model, a search engine optimization audit and remediation plan for IR, and a demonstration of the research potential of the dataset through an exploration into whether machine learning can improve the quality of IR content metadata.

RAMP aggregates data that has surfaced content from institutional repositories in Internet search engine results. It can be used to measure IR performance through assessment of content discoverability, visibility, and use; preliminary analysis of RAMP data has identified large performance variances across IR.

The aggregate dataset provides a previously unavailable view across the performance and use of registered open access IR. RAMP is the only service of its kind in the United States, and nearly 40 institutions have registered their repositories in its first 18 months of operation. This proposal seeks funding to plan high-impact solutions that will address IR deficiencies and improve the research value of RAMP's dataset.

Data-Driven Improvement to Institutional Repository Discoverability and Use

The Montana State University (MSU) Library, in partnership with the MSU School of Computing, the University of New Mexico Library and DuraSpace, seeks a $49,998 Planning Grant from the Institute of Museum and Library Services through its National Leadership Grant program under its National Digital Platform project category to develop a sustainability plan for the *Repositories Analytics & Metrics Portal* that will keep its dataset open and available to all researchers. The proposal also includes developing a preliminary institutional repositories (IR) reporting model; a search engine optimization (SEO) audit and remediation plan for IR; and exploring whether machine learning can improve the quality of IR content metadata. The project team expects work conducted in this planning grant to make the case for advanced research projects that will be high-impact and worthy of funding.

Statement of National Need

Nationally Significant Challenge Proposal Addresses

The *Repositories Analytics & Metrics Portal (RAMP)* is a web service developed in 2017 by the MSU Library and its partners through previous IMLS research funding. RAMP is the only service of its kind in the United States, and nearly 40 institutions have registered their repositories in its first 18 months of operation. The aggregate dataset collected by RAMP can measure IR performance through assessment of content discoverability, visibility, and use; preliminary analysis of RAMP data has identified large performance variances across IR, as well as low overall use (see Supportingdoc2.pdf). RAMP reveals IR items that have appeared in the search results of all Google properties, including the position of each item in the results and whether the item enticed the user to click-through to the IR.[1] Because the URL of each item is included in RAMP data, the potential exists to mine a wealth of additional data. This proposal seeks funding to plan high-impact solutions that will address IR deficiencies and improve the research value of RAMP's dataset. There is so much potential research that can be conducted with this dataset that future proposals to IMLS and other funding agencies, such as the National Science Foundation, are likely. The planning grant will lay the groundwork for the project team to clearly demonstrate this potential in future proposals.

IR hold a wealth of publications and other intellectual output from researchers. The open access nature of IR stands in contrast to the aspirations of commercial publishers and fulfills a cherished tenet of library and information science: free access for all. But after two decades of development, there is a perception that IR have fallen short of their goals, leading some in the profession to question the value of IR.[2] Additionally, the recent acquisition by Elsevier of Bepress, one of the largest IR platforms, has contributed to the collective angst felt in the profession, particularly as this purchase is viewed as further encroachment on the larger scholarly workflow environment.[3]

Questions about the value proposition of open access IR exist because the library profession lacks the data necessary to measure and demonstrate their content, use, and impact. There are over 500 IR in the United States and at least 3,000 more across the world,[4] but their distributed and siloed nature makes it difficult to collect aggregate data that report on IR performance or analyze the scholarly record contained in IR.

[1] RAMP website - https://www.montana.edu/disc/projects/ramp/

[2] Lynch, Clifford. "Rethinking Institutional Repositories," *Report of a CNI Executive Roundtable Held April 2-3, 2017*, Coalition for Networked Information, May 2017. https://www.cni.org/wp-content/uploads/2017/05/CNI-rethinking-irs-exec-rndtbl.report.S17.v1.pdf

[3] McKenzie, Lindsay. "Elsevier Expands Footprint in Scholarly Workflow," *Inside Higher Education*, August 3, 2017. https://www.insidehighered.com/news/2017/08/03/elsevier-makes-move-institutional-repositories-acquisition-bepress

[4] Directory of Open Access Repositories – OpenDOAR - http://www.opendoar.org

Implementation of repository software varies, application of metadata is inconsistent, and the SEO techniques that have proven to increase use are usually non-existent. The fractured IR landscape has failed to leverage the network effect inherent in cloud-based technologies, and as a result, librarians are unable to gather and analyze aggregate data across IR. Even tools used to measure individual IR performance are flawed. Our previous research demonstrated that commonly used web analytics tools are unable to provide IR managers with accurate reporting metrics. For example, the page-tagging web analytics method used by Google Analytics significantly *undercounts* the number of non-HTML file downloads sustained by IR. Conversely, the log file analysis method that is built into several IR platforms dramatically *overcounts* those downloads due to the difficulty of identifying and filtering non-human robot traffic, whose unfettered visitation to repositories has been measured as high as 85% of total IR traffic.[5]

The inability to provide accurate reports on IR use and impact hampers the library profession's ability to improve IR usage or influence provosts, vice presidents for research, and other university administrators. Our research suggests that open access IR do offer significant value and RAMP can provide the data to help address IR deficiencies and report their value proposition.

For the purposes of this proposal, IR are understood as a type of asset or content management system (CMS) designed to promote the scholarship of an institution through open access publication of its affiliated research products. They share a common set of architectural components, consisting of one or more storage layers, databases, discovery layers, and user interfaces. This definition is applicable to the proposed research precisely because it is inclusive of the variety of platforms currently in use as IR, including the top platforms listed in OpenDOAR: DSpace, Fedora, Eprints, Digital Commons, Weko, Drupal, and others.[6]

Description of the Dataset

The RAMP dataset that forms the basis of the proposed research consists of Google Search Console (GSC) Search Analytics data aggregated from participating RAMP repositories. The data are unique in two respects. First, in contrast with Google Analytics (GA) data currently used by many IR to measure usage, GSC data include click-through activity and content downloads executed directly from search engine results pages (SERP). This activity is not currently accounted for within GA. Second, within the United States there are currently no resources, of which we are aware, that provide access to aggregate data in which a single analytical model has been consistently applied to multiple IR providers and platforms. In addition to providing a unique baseline for IR performance metrics, the scope and scale of the dataset offer significant research potential with regard to quantifying and characterizing variances in the discoverability, visibility, and usage of IR content.

Standard fields and variables provided by GSC include the URL of the item that was downloaded, the position of the web page within SERP, the number of times (or "impressions") a page was included within SERP, the number of clicks that occurred on the page, and a calculated "click through" rate which is the ratio of clicks to impressions. Additional optional variables included in the RAMP dataset include the date of the click event, the type of device used, and the country from where the click event originated. Finally, the RAMP dataset includes a calculated field (which we refer to as "citable content"), which is a Boolean value indicating whether a URL points to a content file (PDF, CSV, etc.) or an HTML page.

There are no privacy concerns associated with the RAMP dataset because it contains no personally identifiable information (PII), not even IP addresses.

[5] Patrick OBrien, Kenning Arlitsch, Leila Sterman, Jeff Mixter, Jonathan Wheeler, and Susan Borda. "Undercounting File Downloads from Institutional Repositories," *Journal of Library Administration*, vol. 56, no. 7, 2016. https://doi.org/10.1080/01930826.2016.1216224

[6] Directory of Open Access Repositories – http://www.opendoar.org

Project Beneficiaries

In the short term, this proposal will benefit IR managers and library administrators as it models methods to leverage RAMP data to report IR impact and provide comparable benchmarking data on discoverability, visibility and use across IR. RAMP currently provides participants with accurate use data, which will be used to quantify recommendations for improving repository performance.

In the longer term, the larger library and information science profession will benefit from the ability to report overall impact of IR, and researchers from other disciplines will benefit from myriad possibilities of an open data set that facilitates analysis of the scholarly record contained in IR.

Project's Contribution to Information Science Theory, Scholarship, and/or Practice

This project will contribute to information science in several ways. First, the aggregate dataset provides a previously unavailable view across the performance and use of many open access IR. As a reporting tool, RAMP provides robot-free counts of file downloads sustained by individual IR; it is the only data source for insight on the discoverability and visibility of IR content found with specific search terms used in Google's 3.5 Billion searches each day[7].

Second, the IR performance data will facilitate SEO research and remediation practices that will improve the discoverability and visibility of IR content. RAMP can help conduct very simple diagnoses of IR performance: highlighting potential SEO deficiencies that could easily be remedied and immediately increase discoverability and use. Furthermore, RAMP usage data can be used to identify the metadata and rich snippet descriptions that will result in greater visibility in search results and click-through behavior that results in downloads.

Third, the dataset allows for considerable additional informetrics research. With nearly 40 institutional subscribers to date, RAMP collects a unique and open aggregate dataset that has never before been available to the profession, and which provides the potential for robust analysis of the scholarly record. Researchers could conduct gap analyses of the content contained in IR by understanding what users are searching for versus what they are finding. They could mine additional metadata to determine whether an institution's disciplinary strengths are represented in the IR. They could create a Google NGram-like viewer that would demonstrate the timeliness of research topics over decades. The possibilities are numerous, and the growth and inclusiveness of this unique dataset will make it valuable for many years of potential research.

An additional contribution comes through the specification of an improved IR reporting model based on an environmental scan and analysis of complementary repository and publishing frameworks (see 1.b and 1.c under Sequence of Activities). The nature of the scholarly record has evolved to include research products such as datasets and interactive media along with the traditionally recognized peer-reviewed journal article. Over the past two decades, IR have served to democratize access to sponsored research products by providing publicly accessible, open access copies of peer reviewed articles as well as other content relevant to the evolving scholarly record: grey literature; conference papers and proceedings; research datasets; etc. IR are therefore critical components of an 'open scholarly record' that comprise additional disciplinary and general-purpose data repositories, open educational resource repositories, and open access journals adhering to a variety of open publishing models. Through identification and preliminary analysis of complementary metrics from resources such as SHARE, Unpaywall, the Registry of Research Data Repositories, IRUS-UK and IRUS-USA among others, the proposed research will investigate methods for characterizing the impact of IR content downloads relative to other systems and services across the broader open scholarly ecosystem. In addition to exploring methods for quantifying actual use of IR content via altmetrics (social media mentions of IR content), and discovery of uncited references to IR content within abstracts, methods, and acknowledgment sections of research articles,

[7] Internet Live Stats - http://www.internetlivestats.com/google-search-statistics/

the proposed environmental scan will provide an assessment of high potential opportunities for research cooperation between RAMP and organizations including the Center for Open Science, IRUS UK/US, etc.

Project Design

Goals, Outcomes, and Assumptions

The overall goals of this project are to assure that RAMP will become sustainable, that its dataset continues to grow and remain open to researchers, and to demonstrate the value of the data. Specific outcomes are listed below and are further explained in the section titled ***Sequence of Activities***.

Outcomes from this work plan include:

1. Development of a plan to transition RAMP to DuraSpace management while maintaining its open dataset;
2. Recommendations for an improved IR reporting model that is standardized and comparable across IR;
3. Creation of a pilot SEO audit plan for IR; and
4. Exploration of a process that deploys machine learning and natural language processing to improve the quality of IR content metadata for improved performance and reporting.

It is well-established that most users mediate their research through search engines, and therefore the practice of SEO is fundamental to the use of any digital repository. A repository that cannot be crawled, harvested, and indexed by search engines will almost certainly suffer from low use. Unfortunately, preliminary RAMP data are demonstrating that most digital repositories have not implemented formal SEO programs and are therefore suffering from low use. SEO operates on the successive principles of discoverability, visibility, and usage. The RAMP dataset provides the ability to quickly analyze whether these principles have been successfully employed in the IR. Brief descriptions of each of these principles follow:

Discoverability

Repository content is considered discoverable to search engine users when it has been successfully crawled and harvested, and when it is found in the search engine's index. IR items that have not been indexed are inaccessible via the 3.5 billion Google searches users generate each day.

Visibility

The first page of search engine results captures 70% of all traffic clicks.[8] IR content that ranks on this first page has a far better chance of being viewed.

Usage

The most critical IR metric is the Citable Content Download (CCD). A CCD occurs when a user downloads the full text IR item; it is prerequisite for consumption and potential citation of the item.

Preliminary analysis of RAMP participants indicates that the large variance in IR discoverability, visibility and usage performance may be caused by the failure of many repositories to address even these basic SEO principles. The problems may stem from the most basic technical steps: failure to submit accurate sitemaps to the search engine; improperly configured or missing robots.txt files that prevent crawlers from accessing content; labyrinthine paths to the items that crawlers have difficulty following; inappropriate metadata schema; and heavy use of graphics in the website that are indecipherable to machines.

[8] Google Organic Click-Through Rates in 2014 - https://moz.com/blog/google-organic-click-through-rates-in-2014

Potential Risks

The co-PIs anticipate no risks associated with their proposed work. However, failure of the profession to address the longstanding underperformance of IR carries the risk of irrelevance and elimination of resources to support continued IR development and management.

A question that often arises when the RAMP team has given presentations is whether we will make the RAMP code available as open source. While there is nothing proprietary or commercially valuable about the code itself, the value of RAMP is that it aggregates from many IR, providing an open dataset that the library community has never seen before. Local installations of the RAMP code would compromise this aggregation and the benefits it could bring to us all; local installations would simply perpetuate the siloed nature of the current IR landscape. DuraSpace is a community-owned organization, and under its management the concern about local RAMP installations would be mitigated. RAMP's dataset would grow in value and it would continue to be available for all researchers.

Theory and Practice Informing the Project

The co-PIs on this proposal have a long track record of research and publication about SEO to help improve the performance of IR. RAMP is finally providing data to support their theories and it dramatically improves the ability of IR managers to diagnose and report IR performance through the measurement of file downloads. More compelling for research purposes, however, are the other variables of the RAMP data set, which, for the first time, offer the possibility of analyzing the scholarly record across IR.

Our team has conducted an environmental scan to determine whether similar efforts exist, and there is one in the United Kingdom that has existed since 2012, called the Institutional Repository Usage Statistics (IRUS-UK) project.[9] Funded by Jisc, IRUS-UK has approximately 130 institutional participants in the UK, but in contrast to RAMP it utilizes the logfile analysis method of gathering download statistics; staff work diligently to try to filter robot traffic. The RAMP project team has developed a good working relationship with IRUS-UK staff, and Montana State University is participating in IRUS-UK's pilot expansion, which is known as IRUS-USA and is managed by the Digital Library Federation. Participation in this pilot has provided data to begin making comparisons between the two methods. Early analysis of data demonstrates that RAMP counts file downloads conservatively, while IRUS-UK data tends to show a larger number. It is too early to determine the exact reasons, but this comparison represents another interesting aspect to this research.

MSU Library successfully hosted the international *Open Repositories* conference in June 2018. The proposal to host the conference included an accounting of the strength of our research on open access repositories and may have influenced the selection of MSU as this year's host.

Sequence of Activities

As this is a one-year planning grant, tasks described in the Sequence of Work will be conducted on pilot or sampling bases, enough to generate data to propose full-scale solutions later.

We propose that IR activity reporting requires both <u>technical remediation</u> to address basic SEO deficiencies and <u>metadata remediation</u> that improves IR content visibility once the repositories are successfully harvested and indexed. The work plan proposes these topics and actions:

1. Management and reporting:
    a. The project team will work with DuraSpace staff to develop a plan for RAMP infrastructure to transition to DuraSpace management. Details of the plan will include further development of RAMP features, a communication plan for existing subscribers, a marketing plan to attract new subscribers, a commitment to keeping the dataset open and available to all researchers, and a digital preservation plan for the dataset;

[9] IRUS-UK - http://irus.mimas.ac.uk

   b. Project personnel will develop and implement a survey to collect data from IR managers with the goal of collecting IR reporting requirements. Survey data will be analyzed to identify priorities and key themes among survey respondents, and a standardized report will be suggested;
   c. Complete an environmental scan to identify sources of complementary item-level metadata and repository metrics that can be combined with RAMP data to improve the research value and analytic potential of the dataset. Potential datasets include metadata from the Center for Open Science SHARE project, statistics from IRUS-UK and IRUS-USA participants, and Crossref citation metadata from Unpaywall, among others. Comparison and combination of these data with the RAMP dataset will allow researchers to better characterize the contribution and impact of RAMP-registered IR within the broader open scholarly record.
2. SEO auditing, monitoring and remediation plan:
   a. Design and pilot an IR SEO Audit for the top and bottom-performing RAMP subscribers;
   b. Perform a SEO gap analysis of the top and bottom performing IR;
   c. Develop a SEO remediation plan for bottom performers;
      i. Provide guidance and support to 2 IR managers that are in bottom 33%, with preference given to IR from smaller institutions;
   d. Evaluate impact of remediation plan on discoverability, visibility and usage metrics;
3. Develop a research plan to apply advances in computer science that improve the quality and consistency of IR content metadata for performance reporting and evaluation across IR:
   a. Generate a dataset by combining RAMP analytics data with item level metadata from the SHARE API[10] or other metadata sources identified in 1.c. The goal of this pilot study is to evaluate the effort needed to deliver the reporting requirements collected in 1.b; and prepare the dataset needed for the machine learning structured metadata pilot study below (3.b).
   b. Conduct a pilot study using supervised machine learning models to predict disambiguated structured and ontological metadata for a given item. The process of cataloguing by a human annotator is practically difficult and tedious in nature, and some cataloguers tend to use custom-made labels that are not dereferenceable with a controlled vocabulary, which reduces the quality and usability of the metadata across IR. As a solution to these issues, we will explore the feasibility of using machine learning and natural language processing for automatically predicting disambiguated structured and ontological metadata, such as, LCSH (Library of Congress Subject Headings)[11], Wikidata[12] Items, or author ORCID ID (Open Researcher and Contributor ID) [13] identifiers for a given IR item. The goal of this study is to develop models that can automatically predict the correct structured metadata for an unseen record using RAMP data enhanced with item-level metadata features (i.e., institution, department, abstract, authors, etc.) extracted from the datasets identified in the environmental scan referenced in 1.c above.

Stakeholder Input and Consensus Building

RAMP was developed during the "Measuring Up"[14] grant previously funded by IMLS. Published research from that grant includes an analysis of two surveys of the IR reporting practices of stakeholders.[15] The surveys

[10] SHARE API - http://share-research.readthedocs.io/en/latest/
[11] LCSH - http://id.loc.gov/authorities/subjects.html
[12] Wikidata - https://www.wikidata.org
[13] ORCID - https://orcid.org/
[14] Arlitsch, Kenning, Patrick OBrien, Martha Kyrillidou et al (2014). "Measuring Up: Assessing Accuracy of Reported Use and Impact of Digital Repositories," Funded IMLS grant proposal narrative. https://scholarworks.montana.edu/xmlui/handle/1/8924
[15] Baughman, Sue, Gary Roebuck and Kenning Arlitsch (2018). "Reporting Practices of Institutional Repositories: Analysis of Responses from Two Surveys," *Journal of Library Administration*, 58:1, pp. 65-80. https://doi.org/10.1080/01930826.2017.1399705

gathered responses from 82 library directors and 71 IR managers affiliated with 244 institutions that were members of one or more of the following professional organizations: Association of Research Libraries; Coalition for Networked Information; Digital Library Federation; or OCLC Research Library Partnership. Responding to questions about how they assess use of their IR, both groups placed a high value on file downloads as an impact metric. The responses also confirmed that most libraries depend on tools such as Google Analytics and logfile analytics, which the research team had determined supply inaccurate counts of these metrics.[16] The stakeholder input gathered in these surveys led the team to develop the RAMP web service to provide accurate download measurements.

Stakeholder input will continue in this planning grant, as we solicit input from IR managers to create a standardized reporting model (as described in 1.b in the Sequence of Work, above).

Project Audience

The audience for this project includes Library and Computer Science stakeholders. From the library discipline: IR managers, deans and directors, and researchers who are interested in SEO and in analyzing the scholarly record held in repositories. To computer science researchers in data science, RAMP represents the largest and most diverse corpus containing user search and usage of full-text research and could serve as a key component for a national Big Data Infrastructure.

Project Team

The project team includes individuals with the expertise and experience to carry out the proposed work. **Dr. Kenning Arlitsch** is dean of the MSU Library, and he and **Patrick OBrien,** a trained economist with significant expertise in search engine optimization and marketing have collaborated on SEO and Semantic Web research since 2010. Together, they have had two major grants (*Getting Found: Search Engine Optimization for Digital Repositories*[17] and *Measuring Up: Assessing Accuracy of Reported Use and Impact of Digital Repositories*.[18]) and produced more than a dozen related publications and at least that many national-level presentations. **Dr. Indika Kahanda** is a professor of computer science. He currently works on developing computational methods for problems involving large-scale biomedical literature data. **Dr. Justin Shanks** is digital scholarship librarian and interim department head of Digital Library Initiatives at MSU Library and director of Data Infrastructure and Scholarly Communication (DISC) at MSU, a joint effort with University Information Technology; **Jonathan Wheeler** is the Data Curation Librarian at the University of New Mexico and a developer of the RAMP application. Mr. Wheeler possesses expert knowledge of the RAMP architecture and underlying data models to facilitate data retrieval, access, and analysis. **Andrew Woods** is the Technical Lead for Fedora at DuraSpace. He will serve as the liaison with DuraSpace as we develop the business plan for transition of RAMP management. Finally, a computer science graduate student will intern as a **Machine Learning Research Assistant.** She will assist the project team with developing machine learning techniques for library metadata prediction.

Time, Personnel, and Financial Resources Needed to Complete the Project

The planning grant will be completed twelve months after the start date. Personnel will be drawn from the MSU Library, the MSU School of Computing, the University of New Mexico Library, and DuraSpace.

Project Evaluation and Performance Measurement

Evaluation of the transition plan will center on development of a formal business plan. The SEO auditing, monitoring and remediation plan primary evaluation will be determined by improvements in discoverability, visibility and usage of the bottom-performing IR selected for implementing the SEO remediation plan.

[16] Patrick OBrien, Kenning Arlitsch, Leila Sterman, Jeff Mixter, Jonathan Wheeler, and Susan Borda. "Undercounting File Downloads from Institutional Repositories," *Journal of Library Administration*, vol. 56, no. 7, 2016. https://doi.org/10.1080/01930826.2016.1216224

[17] IMLS Award Number: LG-07-11-0345-11 (https://www.imls.gov/grants/awarded/lg-07-11-0345-11)

[18] IMLS Award Number: LG-06-14-0090-14 (https://www.imls.gov/grants/awarded/lg-06-14-0090-14)

Performance evaluation of computational models for predicting disambiguated structured and ontological metadata (3.b) will be carried out using a cross-validation procedure, which is one of the most popular techniques for generating unbiased evaluations of prediction models in machine learning. The main performance metrics used will be the area under the receiver operating characteristics curve[19] (AUROC) and F1-score[20]. In addition to evaluating models computationally, we plan to evaluate the performance of the metadata term prediction models on a previously unannotated set of items using human curators.

Project evaluation and performance measurement for the proposed environmental scan and analysis of complementary metadata and repository metrics will be assessed according to the completion of the following goals and their corresponding measurements. First, the completeness of the environmental scan as a representative proportion of relevant metadata registries and repository statistics will be assessed through a review of the literature on open scholarship as well as citations and citation metrics of corresponding datasets. Second, correspondence between RAMP data and a subset of identified datasets will be determined using globally unique identifiers including DOI, Handles, and Arc IDs. For items in the RAMP dataset which lack globally unique identifiers, some attempt will be made to use URLs to cross reference RAMP data with other datasets.

Dissemination of Project Deliverables

By the end of the 12-month grant period, the team will have produced a plan to transition management of RAMP to DuraSpace. The team will also produce at least one white paper to chronicle and evaluate results of the project work; the white paper will be subdivided into specific project areas and submitted for publication as several articles after the grant has closed. In addition, the team will submit a project update presentation proposal to the *Coalition for Networked Information* meeting, the *Open Repositories* meeting, and will present at the *DuraSpace Summit*.

Diversity Plan

The MSU Library and MSU School of Computing are institutionally committed to diversity, equity and inclusion, as exemplified by both entities' mission, vision, and values statements:

- *We seek out diverse perspectives, as they challenge us, help us learn, and broaden our worldview. We work to build spaces and services that are equitable and inclusive to everyone in our community. We value collegiality and build a culture of care within the Library." (MSU Library).*[21]
- *We provide a collegial, inclusive, equitable environment that enables diverse faculty, staff, and students to achieve excellence in our mission. (MSU School of Computing).*[22]

Furthermore, DuraSpace's Code of Conduct does "*not tolerate harassment or other exclusionary behavior in any form*" and states its community includes "*a diverse group of collaborators*."[23]

In alignment with these values, our project will contribute to: (1) increased female participation in computer science; and (2) improved discoverability and access to scholarly work. According to the National Center for Education Statistics, women make up only ~19% of computer science (CS) undergraduates.[24] A major contributing factor to this proportionally low figure is that women tend to "have this idea that CS does not

[19] https://en.wikipedia.org/wiki/Receiver_operating_characteristic

[20] https://en.wikipedia.org/wiki/F1_score

[21] The Montana State University Library. "Mission, Vision, and Values." 2018, http://www.lib.montana.edu/about/mission-vision/index.html

[22] The Montana State University School of Computing. "General Information." 2018, https://www.cs.montana.edu/about.html

[23] DuraSpace Code of Conduct - http://www.duraspace.org/about/policies/code-of-conduct/

[24] The U.S. Department of Education, National Center for Education Statistics. "Bachelor's degrees conferred to females by postsecondary institutions, by race/ethnicity and field of study: 2014-15 and 2015-16." *Digest of Education Statistics,* Aug. 2017. https://nces.ed.gov/programs/digest/d17/tables/dt17_322.50.asp

contribute to the social good, and they want to help people."[25] We believe that the proposed project is demonstrative of computer science helping the social good, and that the associated internship will provide its recipient a unique opportunity to gain experience in this area. In an effort to challenge this misconception amongst women in computer science, we will collaborate with the faculty and staff of MSU's School of Computing to intentionally and explicitly seek-out female candidates for this position.

We acknowledge that our existing project team includes no women. However, during the 2018 Open Repositories conference, which was hosted by MSU Library in Bozeman, we held a RAMP meeting and identified several women who indicated interest to become more involved with the project. Our intent is to form a long-term advisory council during the planning grant year with their help.

Open access IR serve researchers across the world, but they are particularly useful to researchers who do not have institutional affiliations that would provide them with access to subscription-based commercial research journals. This describes many researchers in United States whose libraries have been forced to cut commercial journal subscriptions to satisfy budget reductions, as well as researchers or aspiring researchers in developing countries. IR offer the potential for all these researchers to read and cite publications and grey literature, but only if those items are easily discoverable in open access repositories. Our project seeks to improve discoverability and visibility of IR content for a diverse audience.

National Impact

How the Project Will Lead to Systemic Change Within the Community and at the National Level

Our proposed work plan's IR reporting model, SEO audit, and recommendation on the optimal amount and type of metadata per item could help any IR whose managers seek to demonstrate their impact.

This work plan will lead to the development of RAMP reporting tools that will give repository managers the following insights and capabilities:

1. The behavior of users as they seek scholarly information;
2. Discoverability of open access academic publications and whether there is duplication or gaps between what users are searching and what they're finding;
3. Visibility of IR content in search engines and whether it varies across organization or platform;
4. Whether repository metadata are aligned with user search terms; and
5. The ability to cost-effectively generate item metadata that improves knowledge discovery and reporting comparability across all IR.
    a. Computer assisted machine learning will eliminate the time and effort required to train librarians on the various web discoverability, linked data, and SEO skills necessary for maximizing the discoverability, visibility and use achieved by the highest performing IR. Moreover, based upon our own experiences that once the skills are acquired the cost and resources to generate and maintain the data and machine user interfaces (UX) at scale is costly and well beyond most libraries resources.

How the Deliverables Will Benefit Multiple Institutions and Audiences

Currently, there are approximately 40 institutions from five countries that have registered their repositories with RAMP. The sustainability plan and greater visibility that DuraSpace will bring to the project will almost certainly increase the number of RAMP subscribers. Greater participation will facilitate the creation of comparable and standardized metrics, and it will increase the size and depth of the RAMP data set, leading in turn to a greater understanding of the strengths and deficiencies of the scholarly record.

[25] Vu, Shana. "Cracking the code: Why aren't more women majoring in computer science?" *UCLA Newsroom*, 26 June 2017, http://newsroom.ucla.edu/stories/cracking-the-code:-why-aren-t-more-women-majoring-in-computer-science

How the Project's Performance Goal and Performance Measure Data Will be Collected and Analyzed

Data for the literature review to assess the environmental scan and combination of complementary datasets will be collected from sources such as the Web of Science Data Citation Index and the journal impact factor ratings to identify high value datasets and publications. Other sources might include citation metrics provided by PLOS ONE, PeerJ, and other open access publishers. Data to verify the selection and mapping of globally unique identifiers across datasets will be collected from identifier registries and authorities such as DataCite and Crossref. In cases where URLs are used to identify common items among datasets, accuracy and currency of URLs will be determined using methods such as WGET or Curl harvest of associated content.

Benchmarks for Project's Performance and How They Will be Measured

Each deliverable will be evaluated with a series of questions that will test whether or not the deliverable fulfils its intended purpose. These evaluative questions are listed below for each deliverable:

1. **Sustainability plan for RAMP as an open source project:**
   a. Is a transition plan in place?
   b. Is a communication plan in place?
   c. Is a marketing plan in place?
   d. Has the RAMP dataset been made publicly available under Creative Commons licensing?
2. **Improved IR reporting model that is standardized and comparable across IR:**
   a. Were sources of complementary item level metadata and repository metrics which can be combined with RAMP identified?
   b. Were these data compared and combined with the RAMP dataset?
   c. Was an IR reporting model based on the above components developed and disseminated to IR managers?
3. **SEO audit plan for IR:**
   a. Has the pilot IR SEO Audit on RAMP subscribers been completed?
   b. Are the results of the audit published and publicly accessible to IR managers?
   c. What percentage of the SEO Remediation Plan was implemented by the selected bottom performing IR?
   d. What was the impact of the SEO Remediation Plan on the discoverability, visibility and usage of the bottom IR?
4. **A recommendation for deploying machine learning and natural language processing to improve the quality of IR content metadata for improved performance and reporting:**
   a. Was the feasibility of using machine learning and natural language processing for automatically predicting disambiguated structured and ontological metadata for a given IR item determined?
   b. Was a dataset created which combines RAMP data with item level metadata collected from sources such as SHARE or directly from participating repositories (via OAI-PMH or other APIs)?
   c. Was a formal recommendation made available through a white paper or other type of publication?
   d. Was the formal recommendation shared with the relevant stakeholders?

The results of this evaluation will be included with the final project report submitted to IMLS at the end of the grant reporting period.

Activities:		Year 1										
		Oct	Nov	Dec	Jan	March	April	May	Jun	Jul	Aug	Sep
Management and Reporting	Collect and prioritize IR manager requirements for standardized IR reporting.	■	■	■	■	■						
	Develop project plan for incorporating SHARE metadata and complementary metrics into RAMP analytics.			■	■	■	■	■				
	Develop plan for RAMP to transition to DuraSpace management.						■	■	■	■	■	
SEO Auditing, Monitoring and Remediation Plan	Design and pilot an IR SEO Audit on the top 3 and bottom IR performers.	■	■	■	■							
	Develop an SEO remediation plan for the bottom performers.			■	■	■						
	Provide guidance and support to 2 IR managers from the bottom 33% of RAMP subscribers.					■	■	■	■	■	■	■
Develop Research Plan to Improve the Quality, Consistency, and Usability of IR Content Metadata Across IR	Conduct a pilot study that combines SHARE and other metadata sources with RAMP analytics data to evaluate the importance of each metadata field for IR reporting requirements.	■	■	■	■	■	■	■	■	■		
	Conduct a pilot study using supervised machine learning models to predict disambiguated structured and ontological metadata for a given item.				■	■	■	■	■	■	■	■

DIGITAL PRODUCT FORM

Introduction

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

Instructions

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

PART I: Intellectual Property Rights and Permissions

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

All digital products created by this project will be assigned Creative Commons Attribution (CC BY-NC) licenses. CC BY-NC is a minimally restrictive license, allowing for maximum reuse, remixing, and redistribution, but prohibits the use of material for commercial purposes (https://creativecommons.org/licenses/by-nc/2.0/).

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 raw dataset created under this grant will be made openly available to all researchers who wish to wish to use it for noncommercial purposes.

In accordance with the Montana Board of Regents of Higher Education Policy 401.3 - Copyrights, the "...Works which are produced by an employee in connection with an approved and sponsored research project are treated in accordance with the agreement negotiated with the sponsor..." (http://mus.edu/borpol/bor400/401-3.pdf). As a result, ownership rights will be negotiated between MSU and IMLS as part of the grant agreement.

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.

The raw RAMP dataset does not include any personally identifiable information (PII). It leverages Google Search Console and Google Analytics data, and Google does not pass PII on to subscribers to these services. Further data mining that may be conducted by the project team will only have access to metadata that is already publicly available through open access repositories or aggregations thereof.

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.

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.

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).

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).

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).

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).

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

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).

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).

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.

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.

We will not be creating an application, rather, we will be writing original code to conduct machine learning/natural language processing and analyze comparable datasets. The major functions will consist of: data ingest processes, normalization routines, statistical analysis, and the development/processing of test sets for training machine learning algorithms.

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.

This section does not apply to our project because we are not developing an application.

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.

We will use Elasticsearch, Python, R, spreadsheet and database applications such as MySQL, Postgres, and Excel.

These programs were selected because they are widely-used, open source applications under active development with excellent documentation and a robust user community. Python and R have extant modules that support the proposed activities under this project.

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

Because we will use extant software, this question is not applicable to our digital products. Original code will use the platforms mentioned in B.1, and so should be widely usable.

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

We do not know of any "additional software or system dependencies necessary to run the software [we] ntend to create."

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

Scripts will be documented in-line according to the best practices established by the relevant software user communities. Additional documentation will be written in Markdown format and co-located with the code.

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

Repository Analytics & Metrics Portal (http://ramp.montana.edu/)

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.

The source code developed to model machine learning, natural language processing, and comparative analysis with other datasets will be licensed under Apache 2.0, MIT, or other commonly used open source licensing scheme.

Access and use to RAMP will be guaranteed by DuraSpace under an open source governance model.

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

The code will be published via GitHub or similar community-based code repository.

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

Name of publicly accessible source code repository: GitHub

URL: https://github.com/

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 dataset consists of click data harvested daily for participating repositories from Google Search Console (GSC) standard fields and variables provided by GSC include the URL of the web page in which the click occurred, the position of the web page within SERP, the number of times (or "impressions") a page is included within the Search Engine Results Pages (SERP), the number of clicks which occurred on the page, and a calculated "click through" rate which is the ratio of clicks to impressions. Additional optional variables included in the RAMP dataset include the date of the click event, the type of device used, and the country from where the click event originated. Finally, the RAMP dataset includes a calculated field, "citable content," which is a boolean value indicating whether a URL points to a content file (PDF, CSV, etc.) or an HTML page. On download from GSC, page URLs are parsed to determine whether they reference HTML pages or content files, and only content files are assigned a true or "yes" value in the "citable content" field. Calculation of this field is another value addition provided by the RAMP service and represented within the dataset. Following download from GSC and determination of citable content, data are batch loaded into an Elasticsearch index. This index and its outputs represent the tangible "dataset" that will be analyzed per the proposed research.

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?

We believe the research meets the threshold for exemption as per Code of Federal Regulations, Part 46, section 101 (b) (4):

"Research involving the collection or study of existing data, documents, records, pathological specimens, or diagnostic specimens, if these sources are publicly available, or if the information is recorded by the investigator in such a manner that the subjects cannot be identified, directly or through identifiers linked to the subjects."

Investigators will submit a "Request for Designation of Research as Exempt" form to the MSU IRB to secure approval. MSU's Institutional Review Board Identification Number is: 00000799.

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).

We will not collect any personally identifiable information (PII), confidential information, or proprietary information.

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

Not applicable (see response to A.3)

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).

Secondary datasets will be collected over standard network protocols (i.e. http and https) using common tools like Curl, Wget, and the Python Requests library.

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?

As a secondary dataset, RAMP data variables are documented as per Google Search Console's Standard Query Parameters (https://developers.google.com/webmaster-tools/search-console-api-original/v3/parameters).

Copies of, or reference to, this and other open dataset documentation will be co-located with the code described in Section III C.

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

The dataset and associated software code generated by this project will be archived, managed, and disseminated via ScholarWorks, Figshare, the Open Science Framework, or similar repository capable of providing preservation functions.

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

Name of repository: ScholarWorks

URL: https://scholarworks.montana.edu/

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

The data management plan will be reviewed quarterly and its implementation will be monitored by Project Director Kenning Arlitsch.