Community Tracking Indicators for Open and Inclusive Scholarship

1 STATEMENT OF NATIONAL NEED

To make reliable progress toward a socially-desirable scholarly ecosystem the research community requires ongoing, systematic, and trusted measures of inclusivity, equity, durability, and sustainability. Environmental scans such as the Grand Challenges Summit (Altman et al. 2018), supported by the Mellon Foundation, and the ACRL report on Open and Equitable Scholarly Communications (Maron et al. 2019) have drawn attention to the need to measure and integrate equity and inclusion into the scholarly ecosystem. There is convincing evidence, based on point-in-time studies, that scholarly processes and outputs have substantial bias and/or create barriers to inclusion[1] and that more openness in science and scholarly communication is needed. Assessing progress towards a better scholarly ecosystem requires standard, reliable measures of the desired attributes of a better system.

IMLS's analysis of the National Digital Platform for Libraries (NDP) (Owens et al. 2017, 2018) draws attention to the need for systematic measurement and evaluation of the scholarly ecosystem with a focus that reflects library values including diversity as laid out by the American Library Association and IMLS.[2] This need is not being met by the major players that produce statistics on scholarship. For example, the National Center for Science and Engineering Statistics is the primary source of statistics in these fields and it tracks participation in the workforce by gender and minority status, but does not track participation in scholarly communications.

While it is routine to use publisher-produced citation indicators for the 'impact' of scholarly communication, institutional decision making, and research policy, there is currently no comparable public data that summarizes diversity in who is citing, producing or accessing the same communications. Despite recent advances in making scholarly communication more openly available, few systematic measures are available to track, compare, or evaluate diversity and inclusion in open scholarship. As a consequence, both existing and proposed interventions to improve scholarly practices, norms of scholarly communities, and attitudes of scholars are often in dispute; and institutions lack benchmarks for their local communities and policies.

MIT Libraries Center for Research on Equitable and Open Scholarship (CREOS) will develop open, reliable, standardized indicators that will go beyond measures of 'overall impact' to advance the understanding of who is, and who is not, participating in open scholarship. The indicators will support the evaluation of large-scale interventions, benchmark comparisons with individual institutions and disciplines, and monitor the health of the scholarly information ecosystem over time.

2 PROJECT DESIGN

The project is motivated by the research question:

> *Who is underrepresented in open science and open scholarly communications?*

This question provides a necessary foundation for causal analysis and targets interventions in practice. Scholars and practitioners can use the indicators and the integrated data they derive to make decisions about actions and policies in the scholarly community. The project will develop reliable, comparable, standardized indicators that advance the understanding of who is, and who is not, participating in open scholarship.

The research question will be divided into a subset of focus points that are empirically measurable with the

[1] See for example Lee, Carole J., Cassidy R. Sugimoto, Guo Zhang, and Blaise Cronin. "Bias in peer review." Journal of the American Society for Information Science and Technology 64, no. 1 (2013): 2-17.

[2] American Library Association "Core Values of Librarianship" http://www.ala.org/aboutala/governance/policymanual/updatedpolicymanual/section2/40corevalues and IMLDigital Infrastctucutes that Embody Library Principles https://www.imls.gov/sites/default/files/publications/documents/applying-library-values-emerging-technologychapter-5.pdf

current state of available data:

- What is the prevalence of members of different groups in open-scholarship and open-science initiatives, and outputs?
- Where are open-scholarship and open-science outputs that are produced with and by group members used in the scholarly ecosystem?
- How does group prevalence in open-scholarship and science, and the use of open access products, vary within the scholarly ecosystem?

The state of available data about the scholarly ecosystem limits the indicators that can be derived sustainably. The indicators will be useful, reliable, and comparable but not comprehensive. They will complement existing qualitative analysis efforts, and provide a baseline for comparison over time and across projects. This initiative takes an approach to research design that emphasizes replicability, sustainability, scalability, and transparency.

In the first phase of the project, we will fully operationalize two measurable indicators for each question. The indicators aim to provide data that will inform library leaders in developing institutional policy and strategies; that will guide practitioners in developing and assessing open science initiatives intended to advance diversity and inclusion in open science to improve the diversity and inclusion; and that will provide librarians a baseline for describing and understanding demographic and other characteristics and participation trends in local open scholarship programs, practices, and initiatives.

After the initial set of indicators are compiled we will conduct an assessment of their usefulness with stakeholders such as the CREOS advisory committee and the scholarly communications community. Based on this input, the second iteration of indicators will be reviewed to generate additional measurable research questions.

2.1 Data Sources

The metadata describing open access and science is incomplete, scattered, and imperfect.[3] Notwithstanding, there is much that is openly available. *Table 1* describes a core set of data sources that will be used to develop initial indices. Each of these sources is well-established, regularly updated, provides documented APIs, and has committed to an open-license. While no single source is critical, in aggregate the databases capture a range of open outputs (reviewer activity, editorial activity, publications, software), forms of impact and recognition (citations, grants, publication downloads), and contributor characteristics (contributor role, institution, region, gender, ethnicity, career stage). The selection of source databases will evolve throughout the project and additional public (although not necessarily open) sources, such as Microsoft Academic Graph, and Dimensions.ai will be evaluated.

Table 1 Overview of database resources

		ORCID	**DOAJ**	**I40C**	**ROARMAP**	**PLOS Articles**	**OSF.io preprints**
Overview		Largest registry of research identifiers	Largest database of open access journals	Largest open network of citation information	The largest repository of institutional open access policies	Most detailed open article contributor information openly available	Preprint database spans broadest range of fields
Coverage		The US and worldwide					US-centric

[3] See for a review Gregg et al. 2019

What it measures	*Directly Recorded*	Researcher characteristics: name, institution Researcher outputs: publications, grant funding, review activity	Journal characteristics Journal articles	Publication citation network Authorship of articles	Institutional open access policy	Researcher characteristics: contributor roles; institution, downloads, social media mentions.[4]	Authorship of preprints
	Potentially Mineable	Region, gender, race/ethnicity, career stage	Editorial board membership	Researcher gender, race/ethnicity Bibliometrics	Policy related to equity and inclusion	Researcher gender, race/ethnicity	Researcher institutions engagement with preprints
How it links with other sources	*Structured Identifiers*	Institution, journal, publication, researcher	Journal, publication, researcher	Publication, researcher		Institution, journal, publication, researcher	Researcher
	Named Entities	Institution	Institution, researcher	Institution, researcher	Institution	Institution, researcher	Institution, researcher

A substantial part of the project will be developing robust and durable pipeline components for monitoring, collecting, linking, cleaning, and standardization from each source. Once the data has been processed, indicators can be constructed to address a range of questions including:

- What is the prevalence of members of different groups in open-scholarship and open-science initiatives, and outputs?
    - How do patterns of participation differ by participation role (e.g. editor vs. contributor)?
    - How do patterns of participation differ over time?
- Where are open-scholarship and open-science outputs that are produced with and by group members used in the scholarly ecosystem?
    - How do patterns of citation differ based on the group membership of the cited and citing work?
    - Is the diversity of authorship associated with measures of citation impact?
    - Is the diversity of authorship associated with indicators of use (e.g. downloads)?
- How does group prevalence in open scholarship and science vary within the scholarly ecosystem?
    - How do patterns of participation and impact differ by region?
    - How do patterns of participation and impact differ by discipline?
    - How do patterns of participation and impact differ by institutional setting?

Addressing each research question requires operationalizing the question, measurement, and analysis. The design for the project is multi-phased, multimodal, and longitudinal. The design adopts an empirical quantitative mode of analysis that will enable ecosystem-wide tracking and provide practitioners with practical measures that can be incorporated into their analyses and initiatives.

2.2 Work Packages and Methodology

The project encompasses three research work packages: the development of open-data-based participation and inclusion indicators; the development of salience indicators using web and social media mining; and the piloting

[4] In addition PLOS has provided article-level usage metrics through its ALM services. PLOS is transitioning social media mention tracking to the Altmetrics services, which is not open, but does provide free API access for research purposes.

of community-based extensions. These research packages will be complemented with systematic engagement and dissemination initiatives.

The packages will be phased-in over the first half of the project, starting with the development of a core data processing pipeline; then proceed in parallel for the project.[5] By the completion of the funding period, we aim to have automated data production to continue to produce updates of core indicators over an extended period.

> **Package 1**: An automated, repeatable data science pipeline to retrieve, clean, link, and normalize data from a set of open repositories of information. The data will be augmented through automated coding (e.g. application of gazetteer services to estimate region; and of name-matching to estimate contributor gender). Then the data will be run through a cross-sectional analysis to derive population-level statistics and estimate trends.
>
> **Package 2**: Panel-based design (repeated measures of the same units over time) will provide specific and comparable evidence of changes occurring at the individual institutional-subject level. We will create a panel by targeting key institutional stakeholders in the open science and open access fields; then use social media mining and web-mining approaches to extract information about targeted open access and open science initiatives. This approach enables tracking and comparison mentions of open access, open science, and references to specific targeted open science and open data projects. Based on the patterns of communications we will develop indicators of overall salience for the topics.[6]
>
> **Package 3**: Community-requested indicators will extend packages 1 and 2 for additional analyses of data for a fixed period. The process for selection and development of community indicators will be modeled on the approach for incorporating research community pioneered by the TESS project (Loftis & Lupia 2008) and American National election studies (see Aldrich and McGraw 2012, for an overview), that enable the scholarly community to propose additional tracking indicators and one-time or repeated social-media-mining, bibliometric, and institutional measures. This package will enable the scholarly communications community to focus on questions of primary interest while building on staff expertise, and project-developed data pipelines, to collect and analyze data from heterogeneous sources. The project will solicit 2-3 proposals of indicators over the initial grant period.

2.3 An Example: Tracking Authorship in Open Monographs

While there is growing literature that uses bibliometric data to characterize inclusion in science, almost all of the work consists of one-shot analyses of specific dimensions of inclusion in a selected area of scholarship during a limited period. Out of this broader literature, we know of only two projects that aim for a longer-term analysis of scholarly inclusion. Both projects are prototypes, and target a narrow scope of scholarly content: BASE (Summann et al. 2020) aims to use OAI-PMH metadata harvesting to track statistics on the size of collections in institutional repositories worldwide, and reports activity by country. ORION (Stathoulopoulos et al. 2020) is a prototype for interactive visualization patterns of metadata describing publications in the life sciences in Microsoft Academic graph: it provides geographic visualization by gender and region.

A report on Exploring the Public Evidence on Open Access Monographs (Altman 2021), provides an example of a more general and reproducible approach envisioned as part of the first work package. This report examines trends in open monograph publishing. The initiative was a pilot using limited effort over three months and only

[5] Our approach to technical implementation is based on the 'tidy' data science framework (Grolemund and Wickham 2017) primarily employing vetted tidy and community of science packages for data manipulation and processing supplemented by open Python libraries for specific data sources. Processing data will employ cloud computing services for scalability: primarily AWS EC2 and Lamba. The project website will be hosted using Github pages and a static website generator (Hugo), and be based on a framework of interactive notebooks using R, Plot.ly and Shiny.

[6] See Epstein and Segal 2000 for a discussion of issue salience concepts, and Aiello et al. 2020 for a discussion of the opportunities and challenges of measurements using social media sources for salience and awareness measurements.

using open-source tools and data. The result is a self-contained, reproducible, open-source report that incorporates interactive data tables and visualization.

The source for the document is available through GitHub and takes the form of a fully replicable analysis simply by re-running the report.

The report, although intended as an exploration, yielded two unanticipated and suggestive findings relevant to inclusion: first that women have been consistently underrepresented as authors of open monographs since 2011 (see Figure 1 below), and second that author-paid book-publishing fees are lower than expected.

Any Female Authors \ Opened Year	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	Totals
false	60.4%	57.5%	58.1%	64.8%	62.4%	63.0%	69.8%	60.0%	68.7%	56.9%	64.4%
true	39.6%	42.5%	41.9%	35.2%	37.6%	37.0%	30.2%	40.0%	31.3%	43.1%	35.6%
Totals	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Figure 1: Distribution of Female Authors in Open Monographs (Source: Altman 2021)

The results, based on the imputation of gender from names, should be considered a very preliminary aggregate estimate, created to promote general discussion, potential issue spotting, and hypothesis generation. Methodology for imputing gender, although used widely in bibliometric studies, is evolving and requires careful evaluation and validation.[7]

The report represents an exploration of only two data sources. Building a robust, comprehensive, comparable, reliable set of indicators requires more development including: developing measures that are standardized across data sources, computing and monitoring a standardized set of data quality indicators, cross-validating the results and generating reliable measures of statistical uncertainty, automatically monitoring data sources for changes, packaging reusable code as public libraries and disseminating them through open archives (e.g. ROpenSci), developing documentation, engaging in training and outreach, and tracking usage of the data, reports, and tools for evaluation. This less robust exploratory analysis does, however, provide a compelling proof-of-concept that open data sources can be used to analyze inclusion in an open and reproducible way and that such an analysis can yield new and important insights..

2.3 Practitioner and Researcher Engagement

The project is designed to engage broader communities in both the design and dissemination phases.[8] In the design phase, we will develop additional indicators based on community input and engagement. As part of dissemination, CREOS will contribute workshop sessions to major library forums and conferences to guide

[7] This method is intended for aggregate analysis and not for individual-level analysis – e.g. the assignment of a pronoun to an author. Further, the reported imputation describes only point estimates and does not reflect uncertainty from several sources including: omissions in the original data sources, heuristic name extraction, and uncertainty in name to gender assignment. Further, the analysis treats gender as a binary category, and thus will structurally omit non-binary gender categories. The classification reported in the table is based on the IPUMS corpus (see Blevins, et al.) As a sensitivity check, we evaluated using two other methods: use of the historical Social Security Administration database yields a higher estimate of participation by at least one female author, but still lower than baseline expectation. Use of the popular 'Kantrowitz' method, which is based on a much smaller corpus yields significantly lower estimates of female author participation. Notwithstanding, the range of estimates does not alter the overall substantive conclusions.

[8] Concerning human subjects, the library community will participate in research design, but there are no human subjects involved in the research itself. The research will use a combination of existing open data; institutional data shared under an open license; and observation of public social media communications. Since the research does not involve interaction or interventions with humans nor the study of private identifiable information it does not constitute human subjects research.

practitioners on how to access and interpret the indicators. Dissemination and engagement will include participation in the following community venues:

- Library leaders who are incorporating the indicators into their institutional policy and strategies at the Coalition for Networked Information (CNI).
- Practitioners developing open science initiatives on tracking participation and inclusiveness as part of FORCE11.
- Librarians using the measures to baseline their institutions and initiatives will be presented at ALA.

MIT Libraries occupies an unusual position within the research and publication community because it is a part of an interdisciplinary research community at MIT and also oversees the MIT Press. This enables effective engagement with the scholarly community for review and extension of these measures, which is critical to curating an evidence base that is broadly regarded as useful, trusted, and transparent.

2.4 Dissemination and outputs

Publications. The primary output will be a set of standardized indicators and a series of reports describing the state of open access and open scholarship accompanied by community-selected additional indicators. The project will also publish summary reports and research articles highlighting trends, discoveries, methodology as white papers, and through scholarly journals such as *JASIST, Scientometrics, In the Library with a Lead Pipe, Kula, or PLOS.*

Data products. The indicator set will be derived from an integrated database and generated through a set of automated data pipelines. The database of indicators, standardized reports, replication data, and code will be disseminated under open licenses through the CREOS website and a Github repository, as described in more detail in the digital products plan. Reports, data, and code will be archived through DSPACE@MIT, MIT Libraries Dataverse, and actively developed code will be published through Github. Results from the research (e.g, environmental scan, methodology, trend analysis) will be disseminated through conference presentations, conference workshop sessions, journal publications, and publicly available project documentation. The Research Scientist is active in professional associations, academic societies, and practitioner working-groups and will provide opportunities to share findings and recommendations across domains.

Evaluation and reporting. The project will use a combination of traditional bibliometrics and alt-metrics to assess the use and impact of the outputs. This will include publication-citation counts; data-citation counts; downloads; traditional media and social media coverage; and counts of contributions and contributors to the community modules. The evaluation metrics will also be made available as open data.

2.5 Quality Factors

The usefulness of any set of indicators depends on their temporal regularity, measured accuracy, and comparability with other measures. The project is designed to manage each of the three quality factors independently.

Temporal regularity is required to detect trends in the scholarly ecosystem and as a building block for measuring the effects of different interventions and events. We will employ a dual-level approach to create temporal regularity. The project will target a set of core data sources that are frequently or continuously updated and construct an automated retrieval and linking pipeline so indicators may be efficiently produced at regular intervals. Using this setup can produce intermediate estimates that are synchronized across data sources and are sufficiently frequent (e.g. monthly) to enable adjustment for seasonality in the construction of indicators.

Measured accuracy is required to reliably distinguish systematic differences from statistically random variation. Accuracy will be managed using a total survey error approach that bounds error from each stage of

the estimation process, including errors related to measurement, linkage, coverage, and sampling.[9] While there are sources of error inherent in each data source, the project will employ appropriate uncertainty-aware methods including multiple imputations, and probability weights to produce the indicators; and will accompany each indicator with truthful (bias-corrected) measures of uncertainty.

Comparability is required to coherently combine the indicators with independent measures collected by other projects and research. While comparability is inherently contextual we can promote comparison of the indicators with other independent measures through standardization. The approach will standardize the components measures used to construct the indicators (e.g. standardizing roles to align with the CREDIT taxonomy, and regions to align with US Census coding)[10] and seek community input to review the indicators.

Replicability and reproducibility are necessary both to ensure that the outputs are reliable, and to enable indicators and analyses to be updated efficiently over time. The core reports, indicators, and databases will be versioned and updated regularly for the duration of the grant through the construction of a data processing pipeline. It will retrieve, clean, standardize, and link the data sources and generate summary indicators. The pipeline itself will be fully automated using a combination of R-Tidyverse libraries and python modules and developed using a continuous-integration methodology. Before each official release, a manual quality-assurance review will be used to check the results of the pipeline. Code will be managed through Github under an OSS-approved license (Apache v.2)

2.6 Project Team & Management

CREOS Research Scientist, Micah Altman, PhD., will direct the project and is co-PI. Altman has authored over eighty books and articles in leading outlets and has received numerous awards. He currently serves in leadership roles for several library and stewardship organizations, including the National Digital Stewardship Alliance, Data-Preservation Alliance for Social Science, Force11, and the Qualitative Data Archive.

Chris Bourg, PhD, Director of Libraries at MIT and has oversight of the MIT Press. She is the PI and will provide expert guidance on research design as well as lead the CREOS advisory committee.

CREOS Deputy Director, Sue Kriegsman, will support the project as part of her duties for CREOS.. This includes financial management, reporting, documentation, data management planning, long-term disposition of program materials at MIT, and strategic planning for the program within the broader scope of MIT Libraries. Kriegsman has managed grant programs including Harvard Library Lab and programs at the Berkman Klein Center for Internet & Society at Harvard University.

Covid-19 planning. Because COVID-19 and associated complexities are evolving rapidly we have adapted our operation to fully remote work; communications; and dissemination. Neither the successful execution of this project, nor the timeline for the schedule of completion are dependent on access to physical travel, nor rely on other assumptions regarding COVID-19.

Performance measurement. Performance measures will be collected quarterly. The program manager will be responsible for overall coordination of performance measurements. Efficiency and timeliness will be measured through data collected from our established project management tools. Quality and effectiveness measures will be derived from processing and impact measurements as described in sections 2.4 and 2.5 above.

3 DIVERSITY PLAN

Increased diversity and inclusion in open access and open science is the motivating goal for the project and central to the research questions. The project will contribute to an understanding of diversity and inclusion in open access and open scholarship by establishing baseline tracking participation trends across open science and

[9] See Weisberg 2009; Groves et al. 2011

[10] See Brand et al. 2015

open access ecosystems. The project will standardize, document, and establish comparable measures of inclusion that can be incorporated by other initiatives into their evaluation processes.

The project implementation and dissemination plans include components to promote diversity and inclusion:

- Recruit data collection 'modules' to develop indicators from the scholarly community with a special focus on modules that characterize participation by historically marginalized communities.
- Disseminate project information with researchers and practitioners engaged in creating open scholarship so future activities can incorporate better measures of inclusion and diversity.
- Engage students in the research project and with its outputs.

CREOS, through this project and others is fully committed to engaging with researchers and learners from underrepresented communities, and to advancing thoughtfully equity and openness in scholarship and knowledge production and use.

The newly formed CREOS advisory committee provides a variety of expertise in quantitative and qualitative research methods, deep connections to communities in cognitive science, the humanities, social sciences, and equity and inclusion. This advisory committee will continue to expand and it is based on an understanding of the needs and perspectives of communities which have been historically and radically underserved by current scholarly communications infrastructure and practices. It includes Leslie Chan at the University of Toronto Scarborough; Erin McKiernan, Community Manager for the Open Funders Research Group (OFRG) at SPARC; Tressie McMillan Cottom, associate professor in the iSchool at the University of North Carolina-Chapel Hill; Safiya Noble, University of California, Los Angeles (UCLA) in the Department of Information Studies where she serves as the Co-Founder and Co-Director of the UCLA Center for Critical Internet Inquiry (C2i2); Anasuya Sengupta, co-founder of Whose Knowledge; Geeta Swamy Associate Vice President for Research, Duke University and the Vice Dean for Scientific Integrity at the School of Medicine. MIT faculty including Roger Levy, Rebecca Saxe, Stephanie Frampton are current collaborators and mentors in a newly launched CREOS postdoctoral research program in equitable and open scholarship (supported by the Mellon Foundation).

4 NATIONAL IMPACT

To date, equity and inclusion in scholarship has been measured by point-in-time studies, generally targeting specific disciplines. While such studies gain much attention, and demonstrate a need for change, they are of limited use for measuring progress and targeting new efforts. This project will address two critical gaps in the understanding of inclusive scholarship and scholarly practice by creating ongoing standardized indicators of equity and inclusion in open science. Communities will be able to use these indicators to reliably and consistently track changes in scholarly practice, to identify and understand mechanisms of change, and to evaluate the impact of new policies, practices, and initiatives. The integration of systematic, comprehensive measures of participation in open access and open science will expand our ability to measure progress in this area and will enable better benchmarking of individual institutions and disciplines.

4.1 Piloting change

Equity and inclusion is a core value of librarianship, and academic libraries and librarians have been leaders in the open access movement for over a decade. We have substantially advanced open access by acting through multiple channels, including advocacy, inclusive collection-development policies, the development of educational resources, and (most recently) library-based publishing initiatives.

Library researchers and institutional stakeholders need reliable and comparable information about how new practices, such as open peer review, have affected patterns of participation in scholarly communication. The proposed set of community tracking indicators addresses these needs by providing ecosystem-level baselines and tracking measures over time, and by standardizing and documenting methodologies for comparison

measurements of diversity and inclusion in open access and open science projects.

4.2 Standardized deliverables to enable adaptation

In the absence of a set of common reliable, standardized measurements and indicators, it is challenging for practitioners and researchers to evaluate scholarly initiatives, or to credibly measure progress towards increased diversity and equity in open scholarship. Although a substantial amount of data on open access publications and activities is available without licensing fees, it is still complicated to interpret and evaluate without specialized skills because creating a baseline measure requires:

- locating multiple data sources;
- interacting with different APIs and protocols;
- converting data across multiple formats;
- linking data with overlapping coverage, aggregated at different levels, and collected at different frequencies;
- and selecting and constructing comparable measures.

The proposed project will address this complexity by releasing standardized indicators, integrating standardized databases needed to generate the indicators, and the software code to link indicators to the integrated databases and sources. As a result, practitioners and researchers will be able to readily access and use summary information; access, adapt, and reuse a set of integrated databases; and adapt the processing pipeline for their data science applications.

4.3 Sustainability

The project is designed to be sustainable with minimal effort and continually increasing value: the project will produce data and measurements for immediate use and the value of baselines and tracking indicators inherently increases as the timespan of collecting data grows. Marginal maintenance costs decrease because of the substantial investment in the early stages of the project to develop a fully automated data pipeline to reduce the effort to update the indicators over time. Stakeholders are invited to help maintain the project in order to contribute to its demonstrated value. The ongoing indicators benefit researchers and practitioners by extending existing community resources, offering data support for new projects, and adding value to outputs of major open science platforms (such as ORCID or COS).

5 REFERENCES

Aldrich, John H., and Kathleen M. McGraw, eds. Improving public opinion surveys: interdisciplinary innovation and the American national election studies. Princeton University Press, 2012.

Aiello, A.E., Renson, A. and Zivich, P.N., Social media–and internet-based disease surveillance for public health. (2020) Annual Review of Public Health, 41, pp.101-118.

Altman, Micah. Exploring the Public Evidence on Open Access Monographs. (2021) CREOS White Paper. < https://hdl.handle.net/1721.1/129690 >

Altman, Micah, et al. "A Grand Challenges-Based Research Agenda for Scholarly Communication and Information Science." (2018). Pubpub. < https://doi.org/10.21428/62b3421f >

Blevins, Cameron, and Lincoln Mullen. “Jane, John... Leslie? A Historical Method for Algorithmic Gender Prediction.” (2015) DHQ: Digital Humanities Quarterly 9 (3).

Brand, Amy, Liz Allen, Micah Altman, Marjorie Hlava, and Jo Scott. "Beyond authorship: attribution, contribution, collaboration, and credit." Learned Publishing 28, no. 2 (2015): 151-155.

Epstein, Lee, and Jeffrey A. Segal. "Measuring issue salience." *AJPS* (2000): 66-83.

Cronin, Blaise, and Cassidy R. Sugimoto, eds. Beyond bibliometrics: Harnessing multidimensional indicators of

scholarly impact. MIT Press, 2014.

Cronin, Blaise, and Cassidy R. Sugimoto, eds., Scholarly Metrics Under the Microscope: From Citation Analysis to Academic Auditing (Medford, NJ: Information Today, 2015).

Gregg, Will, Christopher Erdmann, Laura Paglione, Juliane Schneider, and Clare Dean. "A literature review of scholarly communications metadata." Research Ideas and Outcomes 5 (2019): e38698.

Groves, Robert M., Floyd J. Fowler Jr, Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau. Survey methodology. 2nd ed. John Wiley & Sons, 2011.

Fortunato, Santo, Carl T. Bergstrom, Katy Börner, James A. Evans, Dirk Helbing, Staša Milojević, Alexander M. Petersen et al. "Science of science." Science 359, no. 6379 (2018).

Grolemund, G. and Wickham, H., R for data science. O'Reilly & Sons. 2017.

Larivière, Vincent, Chaoqun Ni, Yves Gingras, Blaise Cronin, and Cassidy R. Sugimoto. "Bibliometrics: Global gender disparities in science." Nature News 504, no. 7479 (2013): 211.

Lee, Carole J., Cassidy R. Sugimoto, Guo Zhang, and Blaise Cronin. "Bias in peer review." Journal of the American Society for Information Science and Technology 64, no. 1 (2013): 2-17.

Loftis KV., Lupia A. Using the Internet to Create Research Opportunities: The New Virtual Communities of TESS and the American National Election Studies. PS: Political Science & Politics. 2008 Jul;41(3):547-50.

Nancy Maron, Rebecca Kennison, Nathan Hall, Yasmeen Shorish, Kara Malenfant. Creating a More Inclusive Future for Scholarly Communications: ACRL's New Research Agenda for Scholarly Communications and the Research Environment. ELPUB 2019, Jun 2019, Marseille, France.

Owens T, Sands AE, Reynolds E, Neal J, Mayeaux S, Marx M. (2018) Digital Infrastructures that Embody Library Principles: The IMLS National Digital Platform as a Framework for Digital Library Tools and Services.

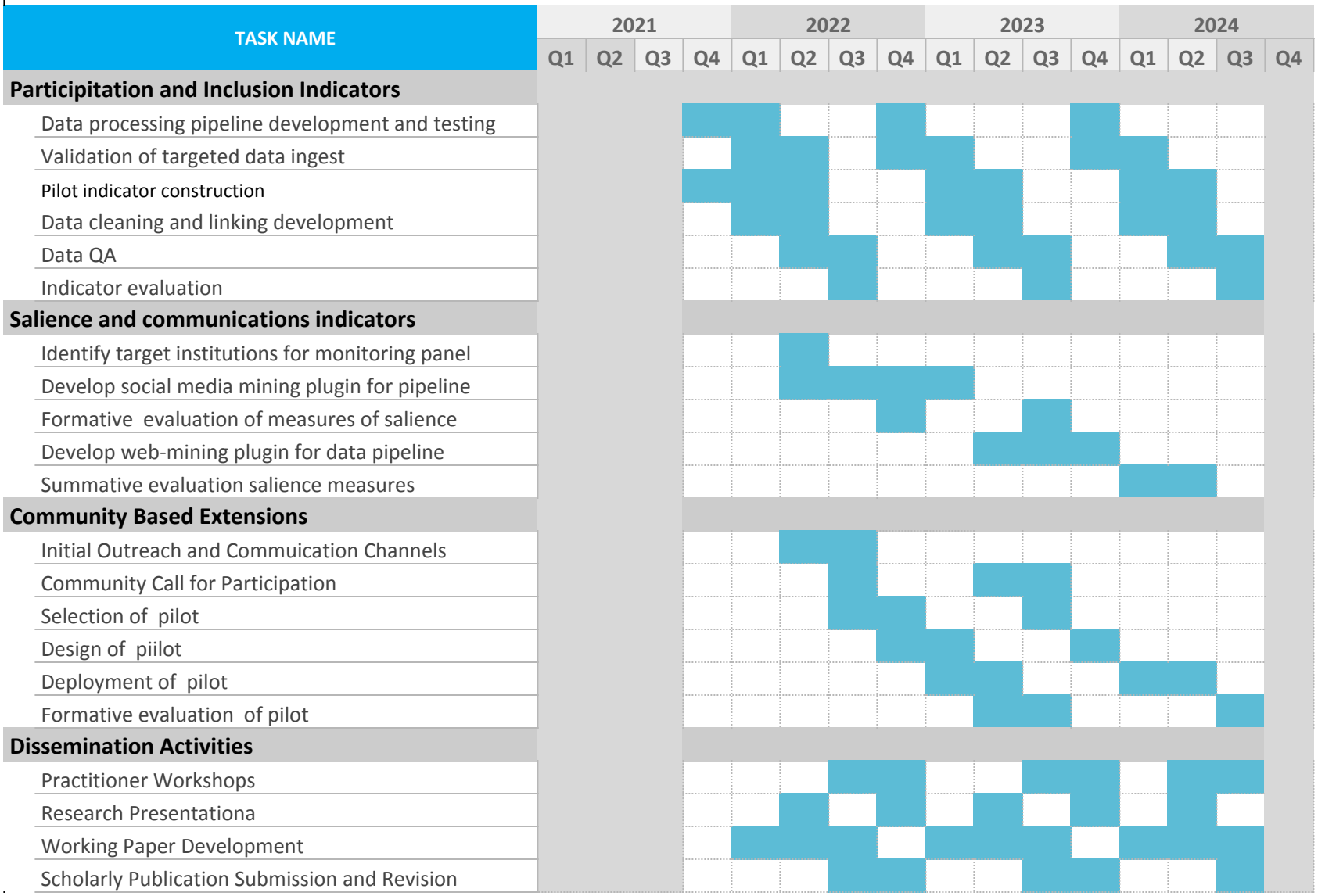
Owens, T., Sands, A.E., Reynolds, E., Neal J., and Mayeaux, S. (2017). The First Three Years of IMLS Investments to Enhance the National Digital Platform for Libraries. Washington, D.C.: Institute of Museum and Library Services, Office of Library Services.

Stathoulopoulos, Kostas, Zac Ioannidis, and Lilia Villafuerte. "Orion: An interactive information retrieval system for scientific knowledge discovery." Talk presented at AKBC 2020

Summann F, Czerniak A, Schirrwagen J, Pieper D. Data Science Tools for Monitoring the Global Repository Eco-System and its Lines of Evolution. Publications. 2020 Jun;8(2):35.

Weisberg, Herbert F. The total survey error approach: A guide to the new science of survey research. University of Chicago Press, 2009.

Schedule of Completion



DIGITAL PRODUCT FORM

INTRODUCTION

The Institute of Museum and Library Services (IMLS) is committed to expanding public access to digital products that are created using federal funds. This includes (1) digitized and born-digital content, resources, or assets; (2) software; and (3) research data (see below for more specific examples). Excluded are preliminary analyses, drafts of papers, plans for future research, peer-review assessments, and communications with colleagues.

The digital products you create with IMLS funding require effective stewardship to protect and enhance their value, and they should be freely and readily available for use and reuse by libraries, archives, museums, and the public. 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

If you propose to create digital products in the course of your IMLS-funded project, you must first provide answers to the questions in **SECTION I: INTELLECTUAL PROPERTY RIGHTS AND PERMISSIONS.** Then consider which of the following types of digital products you will create in your project, and complete each section of the form that is applicable.

SECTION II: DIGITAL CONTENT, RESOURCES, OR ASSETS

Complete this section if your project will create digital content, resources, or assets. These include both digitized and born-digital products created by individuals, project teams, or through community gatherings during your project. Examples include, but are not limited to, still images, audio files, moving images, microfilm, object inventories, object catalogs, artworks, books, posters, curricula, field books, maps, notebooks, scientific labels, metadata schema, charts, tables, drawings, workflows, and teacher toolkits. Your project may involve making these materials available through public or access-controlled websites, kiosks, or live or recorded programs.

SECTION III: SOFTWARE

Complete this section if your project will create software, including any source code, algorithms, applications, and digital tools plus the accompanying documentation created by you during your project.

SECTION IV: RESEARCH DATA

Complete this section if your project will create research data, including recorded factual information and supporting documentation, commonly accepted as relevant to validating research findings and to supporting scholarly publications.

SECTION I: INTELLECTUAL PROPERTY RIGHTS AND PERMISSIONS

A.1 We expect applicants seeking federal funds for developing or creating digital products to release these files under open-source licenses to maximize access and promote reuse. What will be the intellectual property status of the digital products (i.e., digital content, resources, or assets; software; research data) you intend to create? What ownership rights will your organization assert over the files you intend to create, and what conditions will you impose on their access and use? Who will hold the copyright(s)? Explain and justify your licensing selections. Identify and explain the license under which you will release the files (e.g., a non-restrictive license such as BSD, GNU, MIT, Creative Commons licenses; RightsStatements.org statements). 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.

●All research articles will be available immediately upon publication in MIT's open repository. All CREOS personnel are expected to sign MIT's "opt-in" open access license, granting MIT non-exclusive rights to openly disseminate research articles written as part of CREOS work.
●Any software produced by the program will be made available under an Open Software Initiative approved open license: Apache v.2 software license is the default. However, compatibility with previous licenses used for core components, may require the use of a GPLv3
●All data used in publications will be formally cited, following best practices,
●All data collected by the project and necessary for replication of any publication will be available in a public archive.

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.

All work will be open and under a CC license.

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.

There will not be any privacy concerns with the data produced by this project because all data will be in the aggregate and not at the individual level.

SECTION II: 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 the format(s) you will use.

Data will be collected from openly available repositories and resources, and through mining of public websitges and social media sources. Collected data will be cleaned, standardized, and released in open formats -- primarily CSV, and JSON, accompanied by R scripts for ingestion.

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

Each individual on the project will have an computer for their use. We don't anticipate a need for outside vendors or other resources.

A.3 List all the digital file formats (e.g., XML, TIFF, MPEG, OBJ, DOC, PDF) you plan to use. If digitizing content, describe the quality standards (e.g., resolution, sampling rate, pixel dimensions) you will use for the files you will create.

documents: .odt, pdf, txt, html (self contained interactive notebooks)
scripts: Rmarkdown, python
data: csv, json

See the proposal for discussion of qualty criteria.

Workflow and Asset Maintenance/Preservation

B.1 Describe your quality control plan. How will you monitor and evaluate your workflow and products?

The proposal describes 4 quality conrtol criteria in detail. The automated data pipeline includes data cleaning and quality assessment. Official data releases incorporates additional human expert review.

B.2 Describe your plan for preserving and maintaining digital assets during and after the award period. Your plan should address storage systems, shared repositories, technical documentation, migration planning, and commitment of organizational funding for these purposes. Please note: You may charge the federal award before closeout for the costs of publication or sharing of research results if the costs are not incurred during the period of performance of the federal award (see 2 C.F.R. § 200.461).

The program will produce data sets, white papers, presentations, and guides about how to use and interpret the research outputs. CREOS will make the research products openly available in a trusted repository, and will commit to the long-term preservation of those research outputs (e.g. articles, data, code, survey instruments, etc.).

Metadata

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

DSpace metadata for outputs will be captured as Dublin Core.
Data disseminated in dtaverse provided OAI-PMH API to dublic core and DDI metadata.

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

All digital products will be part of the MIT Libraries preservation system.

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).

DSPace and Dataverse provide discovery and harvesting API's

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, delivery enabled by IIIF specifications).

All outputs will be openly available online and the metadata will be indexed by online search tools.

D.2. Provide the name(s) and URL(s) (Universal Resource Locator), DOI (Digital Object Identifier), or other persistent identifier for any examples of previous digital content, resources, or assets your organization has created.

SECTION III: SOFTWARE

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 or similar functions, and explain how the software you intend to create is different, and justify why those differences are significant and necessary.

N/A

Technical Information

B.1 List the programming languages, platforms, frameworks, 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), URL(s), and/or code repository locations for examples of any previous software your organization has created.

N/A

Access and Use

C.1 Describe how you will make the software and source code available to the public and/or its intended users.

Reports, data, and code will be archived through DSPACE@MIT and the Harvard Dataverse (respectively), and actively developed code will be published through Github. Additionally, results from our research (e.g, environmental scan, methodology, trend analysis) will be disseminated through conference presentations, journal publications, and publicly available project documentation.

C.2 Identify where you will deposit the source code for the software you intend to develop:

Name of publicly accessible source code repository:

URL:

SECTION IV: RESEARCH DATA

As part of the federal government's commitment to increase access to federally funded research data, Section IV represents the Data Management Plan (DMP) for research proposals and should reflect data management, dissemination, and preservation best practices in the applicant's area of research appropriate to the data that the project will generate.

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

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?

N/A

A.3 Will you collect any sensitive information? This may include personally identifiable information (PII), confidential information (e.g., trade secrets), or proprietary information. If so, detail the specific steps you will take to protect the information while you prepare it for public release (e.g., anonymizing individual identifiers, data aggregation). If the data will not be released publicly, explain why the data cannot be shared due to the protection of privacy, confidentiality, security, intellectual property, and other rights or requirements.

N/A

A.4 What technical (hardware and/or software) requirements or dependencies would be necessary for understanding retrieving, displaying, processing, or otherwise reusing the data?

Standard web browsers will be able to read and retrieve all of the online data and research information.

A.5 What documentation (e.g., consent agreements, data documentation, codebooks, metadata, and analytical and procedural information) will you capture or create along with the data? Where will the documentation be stored and in what format(s)? How will you permanently associate and manage the documentation with the data it describes to enable future reuse?

A.6 What is your plan for managing, disseminating, and preserving data after the completion of the award-funded project?

Reports, data, and code will be archived through DSPACE@MIT and the Harvard Dataverse (respectively), and actively developed code will be published through Github. Additionally, results from our research (e.g, environmental scan, methodology, trend analysis) will be disseminated through conference presentations, journal publications, and publicly available project documentation.

A.7 Identify where you will deposit the data:

Name of repository:

URL:

A.8 When and how frequently will you review this data management plan? How will the implementation be monitored?

The data management plan will be part of the overall plan for the repositories themselves.