PROJECT SUMMARY

Overview:

Project Summary: Making Data Count: Developing a Data Metrics Pilot
The California Digital Library, PLOS, and DataONE will address what metrics will capture the activity surrounding research data in a valid and credible way. We will design, develop, and prototype data metrics based on field research on data practices; test mechanisms of automatic tracking; explore ways in which the raw, dynamic dataset level metrics (DLM) data be delivered to drive data discovery across data types and research questions; and prototype a flexible report for funders, institutions, and researchers to share DLM results.
Background: Researchers have long grappled with the problem of assessing and tracking the investment in and results of scholarship. As new tools such as article-level metrics have enabled new views into the lifecycle of scholarly communication, the practice of evaluation has begun to incorporate such data and rapid change to the established system has followed. The increased use of metrics has contributed to the recognition that individual objects of research output need distinct means of assessment to understand their role as work products. This proposal will develop convention for addressing research data tracking. DLM will provide a clear and growing picture of the activity around, direct, first-hand views of the dissemination of, and reach of research data. These indicators capture the footprint of the data from the moment of deposition in a repository through to its dynamic evolution over time. The results can be used to power data discovery by enabling new filtering and recommendation mechanisms so that researchers can better find the data most relevant to their specific needs. DLM will provide a multidimensional view to accommodate a broad set of funder and institutional reporting needs. DLM are automatically tracked, thus reducing the burden of reporting and potentially increasing consistency.

Intellectual Merit :

Intellectual merit: At the highest level, the intellectual merit of the proposed work is to understand the contribution of data use and sharing to the larger research ecosystem as well as the evolving objectives of tracking the data lifecycle. The project also provides new tools for the assessment of how early and optimal sharing of research output can further the goals of scientific endeavor. The project will further our understanding of the degree in which automatic impact tracking can lead to discovery of data of value to researchers who can re-use it to further their own work.

Broader Impacts :

Broader impacts: The principal impact of the project is to augment existing scholarly cyberinfrastructure, which currently is focused on journal articles and introduce data as a valued scholarly output. Data metrics will create incentives that support data sharing to increase the velocity of information dissemination across a range of disciplines, once the impact of the research is exposed. The project will also spur a host of other broad-ranging transformations. The community engagement entailed in this project will catalyze deeper discussion on the practice of research data use and reuse across practitioners, funders, and institutions. The software outputs will be publicly shared with an open source license so the community can continue to enhance and re-use the technology to collect and make metrics useful. All data collected will be made available for human and machine-consumption. Since the data usage and activity collected in a systematic format across research areas is the first of its kind, it will be of great interest to scholars who study the practice of research as much as those who conduct research in any of the areas covered. The project will identify differences in how metrics for data compare to metrics for journal articles, and this will inform further development of DLM. This proposal lays the foundations for economic models that can be used to determine the reach of research products as a source of gaining insight into researcher impact and aggregate productivity. Providing data metrics will contribute to how the output of a researcher or research effort is assessed, contributing to a broadening definition of achievement and reputation in the scientific ecosystem beyond publication of research articles. In a world increasingly driven by data, the project will provide more insight on how we can ground our institutions and infrastructures to promote responsible data practices in a way that will further the overall public good.

Making Data Count: a Data Metrics Pilot Project Description

1. Background

Data are the infrastructure of science and they serve as the groundwork for scientific pursuits. They form the outputs of research but also are a gateway to new hypotheses, enabling new scientific insights and driving innovation. And yet stakeholders across the scholarly ecosystem, including practitioners, institutions, and funders of scientific research are increasingly concerned about the lack of sharing and reuse of research data [1-3]. This results in missed opportunities to exploit prior investments, impedes reproducibility, and impacts efficiency and fairness in resource allocation [4]. These concerns are not at all unfounded, as two major studies on the existing researcher opinions and behaviors of data sharing glaringly reveal. The PARSE study found that 25% of researchers make their data openly available for everyone, while the work by Carol Tenopir et. al. revealed similar results (46% make some data available online, 6% make all their data available) [5-6]. At the same time, exemplary data sharing practices are not well supported by the existing scholarly infrastructure by and large, including making data products available, sharing reproducibility results, as well as acknowledging the considerable roles in research played by data producers, data centers, and their funders [7-11].

Recent developments across each of these stakeholder groups have signaled widespread change that is beginning to take hold. The recent memorandum from the White House Office of Science and Technology Policy [12] requires that government funding agencies ensure all research outputs that result from work they support be publicly available. In the UK, Research Council Policies [13] require that data be made available and preserved for ten years and that research publications contain a statement on how the underlying materials – such as data, samples or models – can be accessed. More widely, the European Union Horizon 2020 program includes an Open Research Data pilot [14], which will require data sharing of grantees. Funders are interested in bolstering the re-use of data in the projects they support. New incentives, such as the requirement of data management plans by NSF, and benefits, like widely cited meta-analyses [15], are now appearing, all of which suggest the time is ripe for change [16-19].

Solving the issue of data access can only partly solve the challenges researchers face when sharing and reusing data. Once data are made available, researchers will increasingly encounter diverse data and associated content from unknown origins. These data will be located across disparate journals and data archives. The question then becomes, “How will researchers identify data that best meets their needs and supports their scholarship”? In a world driven by data intensive research, quality research is impacted due to a lack of reliable tools to surface relevant research data. Many tools have been proposed or prototyped, but scant progress has been made beyond simple searches in Google Scholar. Common solutions for the identification of relevant content typically call for better automated search routines or improved semantic understanding of the content at hand, but no widespread offering has met this need. These and other innovative approaches will have far greater success if the user community has a basic toolset and operational framework with the flexibility to experiment and develop their own solutions [20]. This advancement can then be dynamically fed back into the system and shared among the community at large. The call to action and critical challenge is clear: we must provide credible technologies to assess research content and flexibly organize it, so that researchers can find and harvest quality information most relevant to them [20].

The research community also lacks a systematic mechanism by which research data that has been produced and shared can be evaluated and given credit for its impact. Evaluation of scholarly output has traditionally focused solely on research article publications, which remain a fundamentally different work product. Data may or may not be attached to a research article and are certainly used in different ways. In addition, research data are highly structured, and thus are often composed of smaller units (such as

individual data tables or spatial layers) and aggregated into derived data sets that combine data from many source data sets. When these subsets or derived products are used, they frequently are assigned their own identifiers and can be cited independently of the original data set. In addition, individual data objects are versioned over time as new data are added or errors are discovered and fixed. Citation practice is still evolving, and there are no established solutions that address this myriad set of complexities specific to data.

Understanding data usage requires a deep understanding of data provenance, which is just emerging in the research field in practical manifestations. Data also extend beyond the far smaller scope of work communicated in the article narrative. For a more comprehensive view of a researcher's contributions, data metrics are integral. Conventional tools that track data impact based on references made in formal article publications are emerging (ex: Thompson Reuters' Web of Science Data Citation Index). But they exclude the broad range of use and reuse possible for datasets and thus suffer from limitations of scope. They also experience significant time lag for the metrics to accumulate due to their dependence on article publication.

To date, there have been no recent systematic efforts to create a comprehensive, automatic data impact-tracking system that can offer meaningful evidence of impact across a wide range of research areas. New services like the Web of Science Data Citation index are proprietary, subscription-based approaches ill-suited for a vibrant research information environment. An open framework is necessary to facilitate data validation and comparison of data as well as enable the development of programs and tools that add value to the raw data in ever more novel ways. The metrics data must be open for articles as well as datasets with reuse permitted to the fullest extent possible. Research information management systems will be able to more effectively capture the administration of research resources and products with a fuller framework, which incorporates the tracking of research data outputs and can be freely accessed; used; and shared.

2. Project Intent & Goals

Out of the current conditions described, the research community has been calling for solutions to data discovery and to more broadly capture the value of the work that is at the core of the researcher's scholarly pursuit. We propose to design and develop metrics that track and measure data use, "data-level metrics" (DLM).

We will investigate the validity and feasibility of the metrics by collecting them and investigating how to best make use of the data harvested to power discovery and reporting of scholarly these scholarly outputs. The research project will explore the following questions from the standpoint of both collecting as well as using the DLM data:

Data-level Metrics (DLM)
Like article-level metrics, DLM are a multi-dimensional suite of indicators, measuring the broad range of activity surrounding the reach and use of data as a research output.

DLM Generation
○ What metrics are valid, possible to collect through automatic tracking, and useful to the community? ○ What are the limitations and risks to the metrics identified in our pilot? Will they create or enhance any existing bias? ○ What are the community requirements for DLM data validity and reliability? Is gaming an issue and if so, how might it be addressed the metrics implementation? ○ What data channels and associated sources (web services) are sustainable for building a long-

term DLM framework?
- What data collection practices are needed to promote standardization of DLM data and ensure cross-comparison of harvested data?

DLM Use (Data discovery, evaluation, and bibliometric analysis)

- Who will use the metrics and in what manner (use cases & limitations)?
- What do the metrics tell us and what are they not able to tell us?
- Do research communities use data differently and what demographic differences are expressed (subject area, geography, institutional affiliation, etc.)?
- How can the metrics be most effectively communicated and shared (aggregations and reporting recommendations)?
- Which of the piloted metrics highlight differences between open data and closed data? Are data in compliance with open access policies and sharing distinguished in the piloted metrics?
- What cyberinfrastructure gaps (metadata holes, system bridges, etc.) exist for DLM to be fully integrated into the larger research information ecosystem (with repository, funder, researcher, institution information)?

The proposed research and prototyping represent fundamental advances in the state of the art, yet also build on existing capabilities, thereby resulting in work that is simultaneously ambitious and tractable. To carry out this research, we will:

- design, develop, and prototype a reference model for data metrics
- test mechanisms of automatic tracking
- explore ways in which the raw DLM data be delivered to drive data discovery across data types and research questions
- prototype a flexible report for funders, institutions, and researchers to share DLM results.

We will conduct in-depth field research and analyze existing surveys on data sharing attitudes and perceptions to identify core values for data use and reuse to describe existing norms surrounding the use of and data sharing. We will engage the community in an open requirements gathering process, and will then refine and operationalize the results into a bibliometrics framework for an industry-wide data metrics platform. We will explore metrics that can be generalized across broad research areas and communities of practice, including life sciences, physical sciences, and social sciences. The resulting framework and prototype will represent the connections between data and the various channels in which engagement with the data is occurring.

We will test the validity of the DLMs with real data in the wild and explore the extent to which automatic tracking is a viable approach for implementation. Building on a mature Article-Level Metrics application [21], which has been in operation since 2009 and has significant uptake across the scholarly community, we will expand the technology platform to track and collect the data metrics. We will also obtain “usage” activity (times accessed and downloaded) of the datasets in the pilot corpus by leveraging the existing cyberinfrastructure of DataONE and its member repositories. We will also track references to data in both formal and informal scholarly communications, by connecting the application to the various web services which serve as these channels. The pilot implementation will cover around half a million datasets in the DataONE federation of repositories. We will select the corpus based on diversity amongst the following areas: research areas covered, public accessibility of the data, and association with a research article. We will build a reference tool based on the ALM Reports [22], a web-interface that allow users to discover data relevant to their needs as well as report the activity of a collection of datasets in a convenience and compelling visual manner. This will help us to demonstrate the value of DLMs, add value to the high-quality data generated, and test their utility.

We will also conduct bibliometrics analysis on the dynamics of data use and reuse to better understand the nature of what impact means for research data as a scholarly output. For an applicable subset of the corpus, we will investigate the relationship between the activity surrounding a research paper with that of its associated datasets. At the close of the pilot, we will make final recommendations for industry-wide implementation, metrics standardization, and long-term infrastructure needs.

3. Strategic Partnership & Personnel

The partners in this project comprise internationally recognized experts in all areas of expertise required for the project's success. Together, we bring a wealth of essential experience and capacity to the proposed work through existing infrastructure, expertise, access to publications, data from different domains, and relationships with scientific communities, and global data infrastructure projects. The University of California Curation Center at the California Digital Library (CDL) in collaboration with PLOS, and DataONE are well known in their respective communities and widely connected beyond them through strategic alliances (e.g., ESIP, Research Data Alliance, Global Earth Observing System of Systems, International Geosphere Biosphere Program, etc.). With this innovative partnership between scientists, publishers and libraries, the right combination of timing and tool development will be progressive if not transformative. CDL has a deep connection with the researcher communities included in the pilot, as a result of developing data management tools and the research they have conducted on data management practices [24]. PLOS has a long tradition of leading the field of scholarly communications with its history of innovation in publishing. They are the pioneer of article-level metrics, and their technology platform for tracking usage and reach of research articles online is increasingly adopted across the industry. Their experience with harvesting digital activity via a host of web services is coupled with deep knowledge of its application in scholarly communications. DataONE, as a DataNet program, brings its work providing core services and cyberinfrastructure to improve long term accessibility and re-use of NSF-funded research data.

4. Project Structure

Our one-year project consists of five components:

- **Unit One**: we conduct in-depth field research and analyze existing surveys on data sharing attitudes and perceptions (and fears) to identify core values for data use & reuse as well as articulate existing norms surrounding the use of and sharing of data.
- **Unit Two**: we extend the DataONE usage tracking capacity to serve statistics that comply with industry-usage standards.
- **Unit Three**: we take the research gathered in Part One to formulate a set of metrics to test, and extend existing technology that will allow us to begin harvesting data on these metrics under evaluation.
- **Unit Four**: we develop tools for the community to use the metrics in data discovery and assessment reporting as well as integrate the metrics into research papers where possible.
- **Unit Five**: we analyze DLM data collected to evaluate the suitability and viability across multiple research areas.

4.1. Unit One: Data-Level Metrics Field Research

Metrics design is a critical part of ensuring that metrics meet the community's needs, and so field research to understand such needs is paramount. While the results of the pilot will be useful across all fields, the target audience for the DLM pilot will be environmental, oceanographic, and ecological scientists. Requirements gathering will be iterative in nature to provide richer input into the desired data metrics feature set. We will employ a modified ethnographic approach to allow for a holistic understanding of end-user needs. Specific activities include identifying candidates from the UC and DataONE communities; conducting interviews and observing actions and workflows, deriving initial requirements, convening two focus groups at national conferences; and, using results to develop initial

requirements. We will assess what values and benefits stakeholders would like to receive from data metrics and any major features or characteristics that they consider critical. We will create use cases for the different types of scientists to better address the challenges associated with different types of data, researchers.

Assessment of the scientific community's needs will rely on two components:

1. Quantitative assessment comprised of brief surveys (web- or paper-based) that take 5-10 minutes to complete as well as a quick poll (web- or paper-based, or verbally administered) consisting of 1-3 questions that ask the scientist to identifies the most desired characteristic of data metrics.
2. Qualitative assessment, using in-depth interviews (web-based or in person) used to observe scientists' use of data metrics, ask them about their current perceptions and use of metrics, and identify major features of interest for data metrics.

In recruiting scientists for our assessment, we will query individuals from all domains (earth, environmental, oceanographic, geological, ecological) and various sub-domains. We will interview all levels of professional scientists, from graduate students to principal investigators, and we will include scientists from a spectrum of institutions (academia, museums, non-profit organizations). In order to recruit scientists, we will use several tools including social media (Twitter feed, blogs), emails via listservs, face-to-face interactions at conferences, meetings, and other venues, and seminars and meetings at the UC campuses.

4.2. Unit Two: Data Usage Tracking

We will expand the existing DataONE technology platform to process, anonymize and disseminate usage metrics for data deposited into DataONE repositories so that the service can be integrated into the DLM system. DataONE provides machine-level access to a federated network of member data repositories containing tens of thousands of data sets, thereby providing convenient, central access into a highly distributed network of research data repositories. Usage of these data sets is measured throughout the federation, and these usage statistics are aggregated at the DataONE Coordinating Node (CN). The new DLM platform will interact with the DataONE network in two ways:

1. At the point of acquisition and update of datasets in the DataONE network, bibliographic metadata and persistent identifiers (PIDs) for data will be harvested from the DataONE coordinating node (CN) and registered in the DLM.
2. Subsequently, the DLM will periodically request from the CNs current usage statistics about all DLM-registered datasets to update the key impact metrics established by the project.

Both of these interactions will not only leverage the existing DataONE APIs and serve as tangible use cases, but also drive continual refinements to the design and deployment of this core functionality.

Individual repositories already collect this information but it is infeasible to require clients such as the DLM to issue a large number of parallel queries against individual repository sources; it is much better to automate the centrally-managed aggregation of the statistics at the CN from which it can be retrieved easily. API methods already exist for requesting and delivering associated metadata, including PIDs. As part of this activity, the DataONE team will customize their emerging usage statistics API to meet the needs of the DLM platform. Moreover, the API will enable retrieval of usage data aggregated by space and time.

4.3 Unit Three: Data Activity Aggregation

Simultaneously with the data usage tracking development, we build; test; and promote the open-source data metrics platform in Component Three to aggregate the impact of data sets based on the test metrics as well as test the validity and feasibility of the metrics. The supporting technology for DLM will build off of the existing the technology platform of the growing, open source, community Article-Level Metrics project.

PLOS pioneered Article-Level Metrics (ALM) when in March 2009 it became the first publisher to gather and display information about the usage and reach of published articles onto the articles themselves, so that the entire academic community could assess their value. PLOS uses the ALM application to aggregate relevant data and statistics for research articles including online usage, citations, social bookmarks, social media mentions, blog coverage, and reviews/recommendations. The ALM application is a freely available open source community project for the public to implement, enhance, and build third party applications from it. It also includes an API that makes this data available for anyone to re-use and mash-up. Over the years, the ALM application has continued to mature with an expanded suite of metrics as well as a host of supplementary tools that allow users to search, sort, evaluate, and discover articles on the basis of such metrics. This infrastructure of the platform has also been developed to be more robust and scaled to accommodate larger volumes of article records and extensible for other identifiers.

As such, the open-source ALM application serves as the optimal technology to rapidly build out of the data metrics platform and test the proposed data metrics. We will (1) extend the capability of supporting multiple persistent identifier schemes for datasets, (2) track downloads, conversation, and reverse dependencies by building connections to formal web services from which we can directly harvest (including usage activity for the DataONE data processed and packaged for use), and (3) make them available through the application web interface and API. We will also investigate technical solutions to the lack of standardized channels for activity beyond those captured by formal web services. To address this traceability problem, we will explore emerging tools designed to track alternative mechanisms that spot data use. By establishing new markers (digital breadcrumbs of the data files), which leave their imprint where activity has occurred, we can establish tracking calls to pick up the anonymized signals of activity that do not depend on the capacity for these destinations to collate the data. We will follow the applications of this approach in other implementations found in the technology industry as appropriate and feasible.

Collection of data on a repository-by-repository basis is an onerous effort and may not effectively meet the goals articulated above. But integration with the DataONE network will allow the DLM platform to access multiple repositories in a genuine, dynamic data environment that continues to expand in size while its holdings accumulate more usage and re-use events.

4.4 Unit Four: Data Metrics Integration & Presentation

The raw DLM data will form the basis of the project and the uses cases for which they support. For the data to become information and insight, however, they need to be made usable and sharable. We will develop a reference implementation of a web-based discovery and reporting tool for researchers, funders, and institutions. This tool will showcase the myriad ways in which DLM data can be organized in order to support scholars' needs to find relevant data for their research needs as well as funders and institutions to track the reach of these research outputs. The reference implementation will also demonstrate the value of bundling the data and associated metrics with their corresponding research papers and their associated article-level metrics.

The data generated will be easily used by human and machine consumers. To this end, we will develop a web-interface with an intuitive visual language system to functionally engage the DLM content and illuminate the way in which research is being used and digested by the wider world. Consumers will be able to generate reports that are designed to articulate the power of the research data published, the diverse modes of proliferation, and their impact. Furthermore, the tool will also advance the value of DLM data for the discovery process and improve the research experience. Users will be able to use DLM to better filter and sort large sets of results, thus tackling the problem of "filter failure" at the root of information overload. Making the DLM data available with the datasets will enable researchers to more efficiently and accurately uncover related datasets based on a broader spectrum of intellectual "crumb trails" left by other researchers.

For machine consumption, we will invest in a broad set of DLM APIs. The bulk of the APIs will be accommodating to natural language queries, each aimed at providing a specific view of the data, while we will also provide a more powerful, open API for complex queries to the raw DLM data. We will show how metadata across data archives and publishers can be linked together and normalized with author and funder identifiers where available. A small subset of the datasets in the pilot will be selected for their association with a PLOS publication. Together, the APIs and the web-interface tool will provide an expansive portrait of how research activity develops from a grant award.

4.5 Unit Five: Bibliometric analysis

The DLM collected will provide a clear and growing picture of the activity around, direct, first-hand views of the dissemination of, and reach of research data. We will conduct analysis on the dynamics of data use and reuse to better understand the nature of what impact means for research data as a scholarly output. To enable broad bibliometric analysis, the test set will include a diverse set of data types, which cuts across subject areas, repositories, access permissions, and association with research articles (open or subscription-based). Once sufficient data has been collected, we will fully examine the validity and reliability of the pilot metrics, allowing us also to inquire more fully into what the metrics tell us and what are they not able to tell us. We will also begin to explore the data usage behaviors of the research communities represented and delve into the demographic differences are expressed (subject area, geography, institutional affiliation, etc.) and highlight differences that are expressed in the DLM between open data and closed data. Finally, we will investigate the relationship between the activity surrounding a research paper with that of its associated datasets for an applicable subset of the corpus.

We expect that full-scale analyses, which connect researchers, data repositories, institutions, and funders attached to the research outputs may be difficult based on limited metadata availability. We will expose these gaps and provide cyberinfrastucture recommendations for developing a sustainable and robust research information ecosystem, which can serve both research and organizational needs across stakeholders.

5. Personnel roles and qualifications

Patricia Cruse, Director, University of California Curation Center (UC3), CDL, will serve as PI and will be responsible for all aspects of the project, set project priorities, build relationships with key project stakeholders (PLOS, UC researchers and the DataONE community), guide and review results, and ensure successful completion of the project. Cruse is a founding board member of the global DataCite consortium, primary coordinator for the DMPTool project for generating data management plans per NSF and other agency requirements, and is on the DataONE leadership team.

Carly Strasser, Research Data Specialist, UC3, CDL, will manage community interaction, requirements gathering, user testing, and will lead all the interaction with the scientific community and the DataONE team. Strasser will also work collaboratively with the team to disseminate results. Strasser has previously served in this role for the DataUp project [25] and is currently serving as project manager for the UC-wide Dash project.

Jennifer Lin, Senior Product Manager, PLOS, will serve in all aspects of the project, driving project priorities, building relationships as the project manager for the effort. In addition, she will oversee the development of the work throughout its five phases from metrics research and requirements gathering to the buildout of the application and reference tools. She currently manages the Article-Level Metrics program and the publisher's data program.

Martin Fenner, ALM Technical Lead, PLOS, will serve as the technical lead for the data metrics application. He is currently the technical lead for the Article-Level Metrics application, which will serve as the basis of the DLM. He has been a member in the EU-funded ORCID DataCite Integration Project (ODIN), and he serves on the board of the Dryad Digital Repository, a member of DataONE.

Application Developer, National Center Environmental Analysis and Synthesis (NCEAS) (TBN) An application developer will be hired by NCEAS. This developer will develop the data usage statistics API, and will work on extending the statistics API for use of the DLM application in a DataONE environment. The application developer will work with **Matthew Jones**, DataONE and NCEAS, University of California Santa Barbara, to ensure that the development work of extending the statistics API for use in the DLM application can be used by DataONE.

6. Suitability for EAGER

Experimental and High-payoff:

The project we are proposing is a highly experimental project with high payoff potential. Researcher resistance to data sharing and reuse is a core factor and a number of liabilities arise out of it. While the practice of sharing one's own data is widely considered "beneficial," the level of adherence is not complete [6]. Related to the lack of available data, the application of others' research data is not a widespread practice. Instances of this in the wild may be limited, and may affect how widely we can establish formal channels with existing web services to aggregate such activity. It is unclear what the influence of the current lack of standardization in data sharing and reuse will have on our aims of building a more comprehensive set of evidence for impact.

The proposal is likely too risky for existing NSF programs as sources of automatically-available data are not yet established and accepted to be bearers of scholarly value. The community, especially funding committees, needs preliminary data on dataset usage and other activity in order to accept the validity of these measures. The pilot aims to initiate the conversation by serving as the proving ground with an initial set of observations. Furthermore, the proposal is positioned outside of the domain of general NSF grants with its involvement in infrastructure building. While the project is a pilot, the open source DLM application will be ready for production use. The data aggregated can be linked to research entities across systems in the research information ecosystem (researcher, funder, institution, data repository, scholarly indices, etc.). The reference implementation of a discovery and reporting tool will demonstrate how these linkages can deliver an expansive portrait of how research activity develops from a grant award or research position as well as other novel insights on the impact of research.

Widespread benefit:

The principal impact of the project proposed is to augment the existing scholarly cyberinfrastructure singularly focused on the research article and introduce data as a valued scholarly output into the framework. The DLM service will allow anyone to get a full sense of how data are being used and discussed by displaying usage metrics aggregated across the entire Web. For example, it will support all researchers in presenting meaningful impact evidence as their work is evaluated. Also, research sponsors, data producers, and promotion and tenure committees can use the metrics tools to track the productivity of projects and investigators, or explore the diversity of audiences that have bookmarked, discussed, downloaded, and reused datasets. In the comprehensive ecosystem of identifiable, trackable research data, these metrics tools might become an essential to data-rich science.

In service of this paradigm shift, the project will also spur a host of other broad-ranging transformations. Data metrics will create incentives that support data sharing and usage to increase the velocity of information dissemination across a wide range of disciplines. The community engagement entailed in this project will catalyze deeper discussion and reflection on the practice of research data use and re-use across practitioners, funders, and institutions. The software outputs of the pilot will be publicly shared so that the community can continue to enhance and re-use with an open source license. The data collected will all be made fully available for human and machine-consumption. The aggregation of data usage and activity collected in a systematic format across research areas is the first of its kind. As such, this data will be of great interest to scholars who study the practice of research as much as those who conduct research in any of the areas covered. Finally, our proposal lays the foundations for economic models that

can be used to determine the reach of research products as a source of gaining insight into researcher impact and aggregate productivity. Tracking data citation is ideal for determining the impact of dollars spent on infrastructure, support, time, and research for institutions, funders, and the scientists themselves. In a world increasingly driven by data at scale, the project will provide more insight on how we can ground our institutions and infrastructures in an ethical framework that provides for the responsible use of data metrics to further the overall public good.

7. Timeline

Mo1-3	• Begin to engage scientific community in conversation data usage and activity and communicate the aims of the pilot • Conduct field research on the normative values & practices in data use/reuse • Build public layer for dataset usage in DataONE and API to connect to DLM application • Identify and prepare archived datasets in DataONE • Scale existing base application (ALM) • Deliver field report & data metrics requirements document
Mo4-6	• Define preliminary set of metrics that reflect input from community research • Test DataONE usage pipeline across participating member nodes • Technical design of DLM architecture
Mo7-9	• Extend existing application to collect data metrics • Expand interactions between data metrics with additional external web services • Collect DLM data on DataONE research set
Mo 10-12	• Continue to refine and expand DLM • Develop interactive discovery and reporting tool for DLM data • Bibliometrics analysis of DLM data • Enhance DLM documentation and installation and configuration procedures to enhance reusability • Final report on successes, challenges, and opportunities for open, automated tracking of data use in the pilot with recommendations for next steps

8. Results from prior research

Patricia Cruse, Carly Strasser: NSF Grant No. OCI 0830944 (W. Michener PI). "DataNetONE (Observation Network for Earth)". Budget total: $15,257,190. Period: 08/01/2009 to 07/31/2014.
The DataONE project has met its target milestone of creating an open network (dataone.org) of data archives along with a interfaces, software tools, and a community of scientists, librarians, and students that are all key to establishing data citation as a routine practice. This secure and scalable network has (a) scheme-agnostic persistent identifiers, (b) replication of data and metadata, (c) search and discovery, and (d) federated identity and access control. In addition to a thriving User Group (DUG), DataONE also has a published preservation policy and a host of training events, internships, and working groups covering areas such as education, community engagement, metadata, semantics, and sustainability.

References

[1] Nelson, B. 2009. Data sharing: Empty archives. *Nature* 461: 160-163.

[2] Tenopir, C, S Allard, K Douglass, AU Aydinoglu, L Wu, E Read, M Manoff, and M Frame. 2011. Data Sharing by Scientists: Practices and Perceptions. *PLoS One* 6: e21101

[3] LeClere, F. 2010. Commentary: Too Many Researchers are Reluctant to Share Their Data. The Chronicle of Higher Education, 3 August.

[4] *Nature* Editorial, 2009. Data's shameful neglect. *Nature* 461:145.

[5] PARSE Insight (2009) Available: http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf. Accessed 2010 Oct 12.

[6] Tenopir, C, S Allard, K Douglass, AU Aydinoglu, L Wu, E Read, M Manoff, and M Frame. 2011. Data Sharing by Scientists: Practices and Perceptions. *PLoS One* 6: e21101

[7] Cook, R. 2008. Editorial: Citations to Published Data Sets. *FLUXNET Newsletter* 4:1-2.

[8] Costello, MJ. 2009. Motivating Online Publication of Data. *BioScience* 59: 418-427.

[9] Hey, T, S Tansley, and K Tolle (Eds.). 2009. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond, WA: Microsoft Research.

[10] Parsons, MA, R Duerr, and JB Minster. 2010. Data citation and peer-review. *Eos, Transactions of the American Geophysical Union* 91(34): 297-98. DOI: 10.1029/2010EO340001.

[11] Whitlock, MC. 2011. Data archiving in ecology and evolution: best practices. *Trends in Ecology & Evolution* 26: 61-65.

[12] White House Office Open Data Policy. http://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf. Accessed June 1, 2014.

[13] Research Councils UK Policy on Access to Research Outputs. http://roarmap.eprints.org/671/1/RCUK%20_Policy_on_Access_to_Research_Outputs.pdf. Accessed June 1, 2014.

[14] Horizon 2020 Open Research Data Pilot. http://europa.eu/rapid/press-release_IP-13-1257_en.htm. Accessed June 1, 2014.

[15] Hampton, SE and JN Parker. 2011. Collaboration and Productivity in Scientific Synthesis. BioScience 61: 900-910. DOI: 10.1525/bio.2011.61.11.9

[16] Alsheikh-Ali AA, W Qureshi, MH Al-Mallah, and JPA Ioannidis. 2011. Public Availability of Published Research Data in High-Impact Journals. *PLoS One* 6: e24357. DOI:10.1371/journal.pone.0024357

[17] Parr, CS and MP Cummings. 2005. Data sharing in ecology and evolution. *Trends in Ecology & Evolution* 20: 362-363.

[18] Data Replication and Reproducibility: *Science* special issue, 2 December 2011.

[19] Smith, VS. 2009. Data publication: towards a database of everything. *BMC Research Notes* 2:113.

[20] Enriquez, V, SW Judson, NM Weber, S Allard, RB Cook, HA Piwowar, RJ Sandusky, TJ Vision, and B Wilson. 2010. Data citation in the wild. 6th International Digital Curation Conference.

[21] Article-Level Metrics website. http://articlemetrics.github.io/. Accessed June 1, 2014.

[22] ALM Reports website. http://almreports.plos.org/. Accessed June 1, 2014.

[23] Kratz, J and C Strasser. 2014. Data publication consensus and controversies. *F1000Research*, 3:94. DOI: 10.12688/f1000research.4518

[24] Kratz, J and C Strasser. Data publication practices and perceptions. In preparation for submission to PLOS ONE.

[25] Strasser C, J Kunze, S Abrams, and P Cruse. 2014. DataUp: A tool to help researchers describe and share tabular data. *F1000Research* 3:6. DOI: 10.12688/f1000research.3-6.v1

Data Management Plan: Making Data Count: Developing a Data Metrics Pilot

1. Types of data produced

Five kinds of data will be produced or exploited in the course of this project.

The first is social science data in the form of quantitative and qualitative survey results collected as part of Unit One. Given that these surveys will involve human subjects, we will ensure our research is compliant with Institutional Review Board policies and receive prior approval before conducting the surveys. These data will be collected via paper and electronic means. Paper survey results will be digitized by hand, and originals will be kept for the duration of the project.

The second is dataset usage metrics harvested from the DataONE network through mechanisms developed in Unit Two. Individual DataONE member nodes collect raw web log usage statistics. Since this data potentially could be used to identify individual data consumers it is anonymized before being aggregated by the DataIONE coordinating nodes that will be harvested by the DLM tool.

The third consists of software produced or modified as part of the project. The new DLM software created by PLOS will be maintained in a community Github repository during and after the project. Extensions to DataONE code will be maintained in the public DataONE Subversion repository. Fenner will be responsible for code related to the DLM tools.

The fourth is dataset citation data harvested by the DLM tool from traditional and alternative (e.g., blogs, Twitter, etc.) channels of scholarly communication as part of Unit Three activities.

The fifth consists of all other research products (e.g., additional end-user tools developed in Unit Four, analysis results from Unit Five, community feedback, and project status communications) that will be made publicly available via either peer-reviewed journal articles or the project's website, which will be maintained by California by Strasser.

2. Data and metadata standards

The DLM software produced by the project will conform to accepted community best practices, including version control (using git) with tagging of major releases, a permissive open-source license (Apache 2), public availability in a community repository (Github), inline comments, reference and tutorial documentation with download and installation instructions that is available from within the software and from a community website, and extensive test coverage.

As part of this project the DataONE development team will evaluate the utility and level of effort necessary to report data usage statistics in a form that is consistent with COUNTER standards. This would permit meaningful comparisons between metrics for data usage and other types of online resources, including traditional serial and monographic publications.

3. Policies for access and sharing

All five major data categories, as enumerated in Section 1, will be publicly available for review, evaluation, and use as they are generated during the project and after its completion. Announcements about software and data availability will be made using a variety of channels (e.g., blogs, Twitter, email lists) targeting all interested stakeholder communities.

4. Policies for re-use, redistribution

Similarly, all software products resulting from this project will be re-usable and redistributable both during the project and after its completion. The only restriction placed on redistribution of the software is that the copyright and license statement be kept intact as required by the Apache open source license. This software is expected to be of interest to publishers, data centers and repositories, individual researchers, and institutional administrators.

5. Plans for archiving & preservation

In addition to managed in the community based GitHub repository, all major versions of the DLM software will be archived in CDL's Merritt repository with persistent identifiers supplied by the EZID service. Research data and records will be maintained for as long as they are of continuing value to the researchers and project collaborators.

The Merritt Repository Service from the University of California Curation Center (UC3) has capabilities to manage, archive, and share digital content, and provides persistent URLs, search interfaces, and tools for long-term data management. Merritt relies on a highly fault tolerant micro-services architecture with significant redundancy of all computational and storage components. Currently managing over 15 TB of digital resources, Merritt has not experienced any data loss over its five years of production operation.

The NSF-funded DataONE project includes a significant strand of activity that is investigating a number of options for ensuring ongoing sustainability and ongoing organizational, financial, and technical continuity. The software underlying the complete DataONE infrastructure is managed in a public Subversion code repository and is supported by a large and diverse developer community. The DataONE infrastructure relies on a highly distributed fault tolerant architecture. Although all of the individual nodes on the DataONE network have undergone many periodic system upgrades and software refresh cycles, the global architecture has worked properly in all cases, automatically failing over to redundant system capacity on other servers. There has been no interruption of end-user accessibility to the DataONE network since it first went into production operation.

Project Summary: Making Data Count: Developing a Data Metrics Pilot

The California Digital Library will work with PLOS and DataONE to address what metrics will capture the activity surrounding research data in a valid and credible way. We will design, develop, and prototype data metrics based on field research on data practices; test mechanisms of automatic tracking; explore ways in which the raw, dynamic dataset level metrics (DLM) data be delivered to drive data discovery across data types and research questions; and prototype a flexible report for funders, institutions, and researchers to share DLM results.

Background:

Stakeholders across the research ecosystem have perennially grappled with the problem of assessing and tracking the investment in and results of scholarship. As new tools such as article-level metrics have enabled new views into the lifecycle of scholarly communication, the practice of evaluation has begun to incorporate such data and rapid change to the established system has followed. The increasing use of such metrics has contributed to the recognition that individual objects of research output need distinct means of assessment to understand their role as work products. This proposal will develop convention for addressing research data tracking.

DLM will provide a clear and growing picture of the activity around, direct, first-hand views of the dissemination of, and reach of research data. These indicators capture the footprint of the data from the moment of deposition in a repository through to its dynamic evolution over time. The results can be used to power data discovery by enabling new filtering and recommendation mechanisms so that researchers can better find the data most relevant to their specific needs. As a suite of metrics, DLM provide a multidimensional view to accommodating a broad set of funder and institutional reporting needs. More importantly, DLM are automatically tracked, thus reducing the burden of reporting and potentially increasing consistency.

Intellectual merit: At the highest level, the intellectual merit of the proposed work is to understand the contribution of data use and sharing to the larger research ecosystem as well as the evolving objectives of tracking the data lifecycle. The project also provides new tools for the assessment of how early and optimal sharing of research output can further the goals of scientific endeavor. The project will further our understanding of the degree in which automatic impact tracking can lead to discovery of data of value to researchers who can re-use it to further their own work.

Broader impacts: The principal impact of the project proposed is to augment the existing scholarly cyberinfrastructure, which currently is focused on the journal article and introduce data as a valued scholarly output into the framework. Data metrics will create incentives that support data sharing and usage to increase the velocity of information dissemination across a wide range of disciplines, once the impact of the research is exposed.

The project will also spur a host of other broad-ranging transformations. The community engagement entailed in this project will catalyze deeper discussion and reflection on the practice of research data use and reuse across practitioners, funders, and institutions. The software outputs of the pilot will be publicly shared with an open source license so that the community can continue to enhance and re-use the technology to collect and make metrics useful. All data collected will be made fully available for human and machine-consumption. Since the data usage and activity collected in a systematic format across research areas is the first of its kind, it will be of great interest to scholars who study the practice of research as much as those who conduct research in any of the areas covered. The project will identify differences in how metrics for data compare to metrics for journal articles, and this will inform further development of DLM.

This proposal lays the foundations for economic models that can be used to determine the reach of research products as a source of gaining insight into researcher impact and aggregate productivity. Finally, providing metrics on datasets will also contribute to how the output of a researcher or research effort is assessed, contributing to a broadening definition of achievement and reputation in the scientific ecosystem beyond publication of research articles. In a world increasingly driven by data at scale, the project will provide more insight on how we can ground our institutions and infrastructures to promote responsible data practices in a way that will further the overall public good.