

## RESEARCH PAPER

# A Discussion of Value Metrics for Data Repositories in Earth and Environmental Sciences

Cynthia Parr<sup>1</sup>, Corinna Gries<sup>2</sup>, Margaret O'Brien<sup>3</sup>, Robert R. Downs<sup>4</sup>, Ruth Duerr<sup>5</sup>, Rebecca Koskela<sup>6</sup>, Philip Tarrant<sup>7</sup>, Keith E. Maull<sup>8</sup>, Nancy Hoebelheinrich<sup>9</sup> and Shelley Stall<sup>10</sup>

<sup>1</sup> National Agricultural Library, Agricultural Research Service, USDA, Beltsville, US

<sup>2</sup> Center for Limnology, University of Wisconsin, Madison, US

<sup>3</sup> Marine Science Institute, University of California Santa Barbara, Santa Barbara, US

<sup>4</sup> Center for International Earth Science Information Network (CIESIN), Columbia University, New York, US

<sup>5</sup> Ronin Institute for Independent Scholarship, Boulder, US

<sup>6</sup> DataONE, University of New Mexico, Albuquerque, US

<sup>7</sup> Julie Ann Wrigley Global Institute of Sustainability, Arizona State University, Tempe, US

<sup>8</sup> National Center for Atmospheric Research, Boulder, US

<sup>9</sup> Knowledge Motifs, US

<sup>10</sup> American Geophysical Union, Washington DC, US

Corresponding author: Cynthia Parr ([cynthia.parr@usda.gov](mailto:cynthia.parr@usda.gov))

Despite growing recognition of the importance of public data to the modern economy and to scientific progress, long-term investment in the repositories that manage and disseminate scientific data in easily accessible ways remains elusive. Repositories are asked to demonstrate that there is a net value of their data and services to justify continued funding or attract new funding sources. Here, representatives from a number of environmental and Earth science repositories evaluate approaches for assessing the costs and benefits of publishing scientific data in their repositories, identifying various metrics that repositories typically use to report on the impact and value of their data products and services, plus additional metrics that would be useful but are not typically measured. We rated each metric by (a) the difficulty of implementation by our specific repositories and (b) its importance for value determination. As managers of environmental data repositories, we find that some of the most easily obtainable data-use metrics (such as data downloads and page views) may be less indicative of value than metrics that relate to discoverability and broader use. Other intangible but equally important metrics (e.g., laws or regulations impacted, lives saved, new proposals generated), will require considerable additional research to describe and develop, plus resources to implement at scale. As value can only be determined from the point of view of a stakeholder, it is likely that multiple sets of metrics will be needed, tailored to specific stakeholder needs. Moreover, economically based analyses or the use of specialists in the field are expensive and can happen only as resources permit.

**Keywords:** ROI; data repositories; metric; return on investment; FAIR data; impact; evaluation

## Introduction

Publicly funded research data repositories would like the ability to demonstrate a return on investment (ROI) in their efforts to archive and publish research data. The business world uses data analytics to improve their decision-making, cost reductions, and product or service launches. Estimates of data-generated revenue range from tens of billions to over 160 billion USD in 2018 (Statista, 2018; IDC, 2018) and may be broken down by market size and value added, such as the number of jobs created, cost savings, and efficiency and productivity gains (Günther et al., 2017). Teasing out the contribution of Earth and environmental obser-

vation data to these revenue estimates is difficult, however, it is important to note that some of the most valuable uses of environmental data are in emergency response management (Dubrow, 2018; Pinelli et al., 2018), drought monitoring (Bernknopf, et al., 2019), pollution assessment, agricultural production and groundwater quality (Forney, et al., 2012), fisheries and water management, logistics and trade, and on longer time scales, in real estate development, risk assessment and the insurance industry (Voosen, 2017; Downs, 2018).

Beyond commercial value, are the returns that accrue to the scientific enterprise itself. While the impact of open and accessible data on accelerating certain areas of science is still an active discussion (Gewin, 2016), a number of positive measures are emerging, e.g., increased numbers of publications by data publishers themselves (Milham et al., 2018) and up to a 25% increase in citations when the data were published in an open repository (Colavizza et al., 2019). Even harder to measure and usually not part of monetary analyses are the intangible benefits to society. For example, the ability to predict – based on data – environmental habitability and needed changes in lifestyle are priceless, but not valueless. Societal impacts are often captured anecdotally (e.g. Ramapriyan and Behnke, 2019; NOAA 2019), or in major impact reports (IPCC 2014). Data repositories are essential to the functioning of these activities as well.

Determining a monetary return from funds invested in the research data repositories that house these data remains challenging. In part, this is due to a lack of hard metrics (Dillo et al., 2016). The few comprehensive ROI studies of research data repositories do not distinguish between the impact of publicly available data and related research services; for these together they estimate ROI between 2- and 10-fold (Beagrie and Houghton, 2013). Beagrie and Houghton's (2013) analyses of the value and impact of three data centers in the UK (Archeology, Economics and Atmospheric sciences) examined complex metrics, ranging from value estimations to both users and depositors of data which measured social welfare, work-time savings, and explored non-economic benefits. They found that data centers contribute to significant increases in research production and that the value to users exceeds their investment in data sharing and curation. Qualitatively, academic users reported that having the data preserved for the long-term, with the repositories targeting dissemination, was the most beneficial aspects of depositing data there. Kuwayama and Mabee (2018) described similar results from impact assessments of the socioeconomic benefits of satellite data applications at different decision-making scales, and report on efforts to measure the benefits to human health of data on air quality and harmful algal blooms. They also summarized other analyses of very specific stakeholder groups, e.g., for the value of Landsat mapping to the gold mining sector and of a frost prediction application to tea farmers in Kenya.

Analyses to quantify benefit in economic terms are complex in that they require expertise in several fields typically unrelated to the repository itself (e.g., social sciences, economics, survey statistics). They are expensive to perform and time-consuming, and so happen only rarely; and typically only for repositories with long lifespans and relatively large user communities. Beagrie and Houghton's reports (2013) were commissioned and funded by national agencies over a period of two years, not by the repositories themselves. Moreover, methods for determining the economic value of repositories might necessarily vary dramatically among scientific domains. Thus, it seems worthwhile to adopt a practical approach that can help repositories demonstrate their value efficiently, on short time scales, and within the context of their disciplines.

The Make Data Count (MDC) project is an initiative to design and develop consistent, standardized metrics that measure accesses of individual research datasets (Kratz and Strasser 2015a), an essential step towards realizing comparable metrics of reuse. MDC surveyed scholars and publishers to determine which data-use metrics and approaches would offer the most value to the research community. Data usage or access metrics for research data were an important indicator of impact by researchers and other stakeholders, second only to data citations (Kratz and Strasser 2015b). However, standards were lacking on how usage metrics should be collected and reported, so the MDC project collaborated with COUNTER, a non-profit organization, which provides the Code of Practice for Research Data Usage Metrics (Fenner et al., 2018) so that publishers and vendors can report consistent, credible and comparable usage data for their electronic resources. Here, we build on the MDC work by focusing on indicators of the value of a repository managed as a whole rather than that of individual datasets.

Repositories are experiencing increased expectations, e.g., to meet criteria for making their data holdings "FAIR" (Findable, Accessible, Interoperable and Reusable, Wilkinson et al., 2016, Stall et al., 2018; GO FAIR, 2016), to align with schema.org (Guha et al., 2015), and ensure that content is machine readable. Previously, repository stakeholders were mainly research funding agencies and researchers; that group has now expanded to include publishers of academic journals and international audiences; yet these new

stakeholders typically do not provide the resources required to implement and maintain the capabilities needed. Consequently, additional requirements that may not provide a clear benefit to primary stakeholders are difficult for repositories to embrace.

This paper explores these increasing challenges in assessing the value of repositories. For background, we introduce generally recognized categories of costs and benefits of publishing data in dedicated repositories. We then describe an approach for quantifying the value of data repositories, assembling possible metrics to measure both the costs and the benefits they create and report on an exercise to closely evaluate and prioritize these metrics, with recommendations intended to guide metrics development and refinement. Ultimately, repositories want to be sure they are worth the funding they receive, and a reduced set of consistent, streamlined, and meaningful metrics will help.

### ***Background: the costs and benefits of publishing data in repositories***

We define 'data publishing' as the process of making data accessible in a public repository that provides a defined level of professional services. Net value requires evaluating both costs and benefits, where benefits should go beyond financial considerations to include broader societal benefits.

#### **Costs**

Cost metrics are an important component of any business (Rubin 1991, Phelps 2003), in that they ensure that expenses are understood and contribute to operational decisions and strategic planning. In principle, measuring these is relatively simple. Data repositories may be more similar to libraries or museums, although those have far more costs associated with physical infrastructure than do data repositories (e.g. Lawrence et al., 2001). When funding data publication in a repository, costs can be categorized into four areas that focus on typical aspects of physical infrastructure and personnel (expertise, time, salary), and are usually outlined in operational budgets. (**Table 1**, Curation Cost Exchange, 2018).

#### **Benefits**

Benefits are less straightforward to articulate and translate less easily into financial terms (compared to costs). Although the view that making data publicly available for reuse will benefit science or society has been contested (Lindenmayer and Likens, 2013; Longo and Drazen, 2016), many scientists, professional societies, funding agencies and journal publishers agree on its overall benefits, summarized in **Table 2** (McNutt, 2016, AGU 2013, Baker et al., 2015, Popkin, 2019, Wilkinson et al., 2016, Starr et al., 2015, Piwowar et al., 2011).

#### **Approach**

This work began under a National Science Foundation grant, which brought together data and repository managers interested in pathways for increased, sustainable collaboration and coordination to benefit both research networks and individual data use scenarios. In late 2015, a collaboration area developed within the Federation of Earth Science Information Partners (ESIP, <http://esipfed.org>) for further activities, of which one was to consider frameworks for describing Return on Investment (ROI) in data repositories.

In a series of workshops and teleconferences, thirteen self-identified data curation specialists representing seven environmental data repositories and two data-aggregation facilities (listed under Notes, below) reviewed the literature and current practices for assessing data repository value. They identified and categorized 50 specific metrics (Appendix 1) for measuring the costs and benefits that were applicable to environmental data repositories. As an exercise, each individual scored each metric with respect to its importance to measuring repository value on a scale from 'not valuable, not applicable, or unclear', 'low', 'moderate', to 'high'. Importance was generally understood as how critical the metric was to demonstrating repository value and was judged based on the extensive experience that curation specialists brought to the discussion.

Each repository scored each metric by its ease of implementation categorized as 'metric is already collected', 'metric not collected, but could be collected easily', 'collection will require nominal additional resources', 'metric could not be implemented without new actions, such as research on its methodology, a refined definition or guidelines, significant additional funds or community policies.' Scorings are somewhat subjective and only roughly quantitative, but the activity allowed us to closely consider the metrics for trends and recommend priorities for adoption or further discussion. No attempt was made here to assign monetary values to any metric. This work was not intended to be an independent survey; in that a group of repository representatives both identified the metrics and scored them. However, the group brought both

**Table 1:** Cost categories (adapted from Curation Cost Exchange 2018).

<b>1.1. Cost of Initial Investment</b>	<p>Gathering requirements, preservation planning, development of repository platform (hardware, software licenses) and search and access capabilities, development of policies for data acceptance and retention.</p> <p>Costs can vary widely depending on the scope of the requirements, the suitability of off-the-shelf software, and the time required for initial set up, testing and evolution to full production. Requirements may be imposed by the funder or the interests of the scientific community which influence the repository's design and infrastructure.</p>
<b>1.2. Cost to Publish</b>	<p>Data acquisition, appraisal, quality review, standards-compliant metadata preparation and dissemination, overhead, marketing, user support.</p> <p>The variety of scientific communities needs leads to a variety of curation practices and repository goals, with costs partly depending on the data source. Earth and environmental data naturally ranges from the relatively homogeneous, e.g., from sensors or instruments to highly complex organismal observations and physical samples (biological, chemical, geoscience) under both ambient conditions and from experimental manipulations. Large, mission-centric repositories (e.g., satellite data) have costs generally tied to data collection. Repositories serving many individual data producers rely considerably on their contributors' expertise and time which distributes part of the curation cost to those projects. Repositories that are primarily aggregators (whose goal is to collect a variety of metadata or sources for indexing) rely on a minimum level of metadata standardization from their sources; their costs typically arise from resolving incoherent source data and heterogeneous metadata, with related outreach efforts to improve practices.</p>
<b>1.3. Cost to Add Value</b>	<p>Data dissemination planning, processing, data product development, and quality control of the new data products, overhead.</p> <p>Varies greatly among repositories, but may represent the most visible return, or possibly even an opportunity for commercialization. Some raw data will have already received comprehensive processing to make them further useable. The concept of "Analysis Ready Data" are applied in other domains with value-adding steps by repository to target uses from multiple disciplines, non-research uses (e.g., policy makers, general public, education), or per the demand by such groups for the development of specific data products (Baker and Duerr, 2017). The cost for tasks to add value depends greatly on data types, diversity and envisioned uses.</p>
<b>1.4. Cost to Preserve</b>	<p>Anticipated retention period, facilities system maintenance, enhancements, and migration; staff development and technology upgrade.</p> <p>While tracking existing needs is relatively straightforward, future costs may be more difficult to predict. Preservation costs are greatly influenced by technological change (e.g., new hardware, standards, vocabularies, storage formats), and new requirements and data policies that must be translated into repository operations (Maness, et al., 2017, Baker and Duerr, 2017). Iterative migration necessitates expenses in development, data and metadata conversion and user engagement, sometimes without immediately noticeable changes in service. Moving from supporting primarily data publishing to supporting data which are frequently reused requires new services and possibly, value-added products.</p>

deep and broad experience in repository management, technology implementation, user support, data curation, and in obtaining funding for repository operations. Because this was merely an exercise, we refer to the outputs of that exercise in an associated dataset (Gries et al., dataset: 2018). In this discussion, the metrics tallied in the accompanying dataset are denoted by *italics*.

## Findings

Of the 50 identified metrics, 30 measure the benefit (or value) created by holding datasets and making them accessible from a repository, and 20 measure the direct costs related to curation, publication and preservation of those same holdings.

In total, 35 (70%) of the 50 identified metrics had been implemented by at least one repository. Eleven (22%) were implemented by only one repository, however, for most of those, several repositories stated that they would be able to implement them with no additional resources, indicating that if these metrics were to be included in a set of unified guidelines, they could be addressed quickly.

In general, highly variable responses reflect that major aspects of operation differ greatly among repositories (even within the research domain of Earth and environmental sciences), and that all repositories

**Table 2:** Generally accepted benefits of publishing data in an open repository.

<b>1.1. Avoidance of Data Generation Costs</b>	<p>Data gathering is expensive; offering reusable data avoids the cost of recreation.</p> <p>Data value may be easily estimated as the cost to create; however, the <i>future</i> value of data cannot be predicted and different kinds of data will have different useful lifespans, generally dependent on how easy or expensive data are to create and whether they lose or gain in applicability over time. It may be feasible to recreate experimental data, but it generally is impossible to recreate observational field data.</p>
<b>1.2. Efficiency of Data Management</b>	<p>Infrastructure investments benefit all data producers; central programming functions for data search and access improve discoverability and reduce distribution costs to researchers; efficiency benefits are most obvious in repositories serving a large number of single investigators, though all repositories keep large amounts of data safe by upgrading hardware and software as technology changes, and by managing services, such as unique identifiers (e.g., Digital Object Identifiers, DOI).</p> <p>Data repositories can be compared to specialized analytical laboratories as they employ an expert workforce having specific skills in data curation and preservation that ensure the quality and interoperability of their holdings. Once data have met curation standards, repositories maintain continued usability capabilities and working life beyond the lifespan of the original creator's data storage options by addressing format obsolescence and other issues.</p>
<b>1.3. Long-term Usability and re-use of Data</b>	<p>Implementing sustainable data curation, stewardship, metadata capture, and quality of data and metadata enables meta-analyses, innovative re-use for new science or applications. Lengthening the working life of data creates enduring value by enabling subsequent usage over time. Ongoing stewardship can support new uses and user communities. Properly curated, data can be combined or analyzed with data that will be collected in the future and allows the ability to build upon prior work (Starr et al., 2015).</p>
<b>1.4. Transparency of Scientific Results</b>	<p>Making data publicly available in a repository is an important step toward transparency and reproducibility of research, which in turn assures credibility of scientific results (McNutt et al. 2016) and the ability to build on prior work.</p> <p>Historically, best efforts have been made to preserve publications and the salient data published in them. In modern publishing, data needs to be managed and published as a product in its own right (Downs et al., 2015).</p>
<b>1.5. Value Added Data Products</b>	<p>Some repositories increase data utility via pre-processing, semantic and format standardization, data aggregation and interpretation, and specific tools that support the creation of new data products, uses and audiences beyond the original data users (e.g. general public, policy makers, education and outreach) (Baker et al., 2015).</p>

were already tracking several metrics, although the suites differed. Across both metric types (cost and benefit), on average the repositories had already implemented 13 (range 4–20), with another 15 rated as simple to implement (mean, range 10–29, Gries et al., 2018 dataset). An average of 15 were judged to require significant additional research, extensive resources or outside expertise. Overall, seven (40%) of the 18 metrics ranked most important were not implemented by any of the repositories. More than half of the metrics in two categories were ranked as important but were not implemented. These were metrics related to “Value-added Data Products (Benefits, see **Table 2**)”, and “Cost to Preserve (Costs, see **Table 1**).”

Metrics for costs were generally easier to implement than those for benefits. The metrics most likely to be implemented were related to categories commonly found in budgets: direct costs (e.g., *Hardware*) and for personnel, (both as general *staff positions*, and as the primary *cost of software development*). Most of those metrics were already measured.

With respect to benefits, ease of implementation scores fell into three broad groupings (**Table 3**). Note that the metrics that were most likely to have been implemented or took little effort and resources were those that could be gleaned from the data holdings themselves or from server logs.

Interestingly, a few repositories had already implemented some of the metrics that were ranked important yet difficult (e.g., *use of data in papers*, *reduced storage cost to users*, and *user satisfaction*). Such implementation generally required significant planning, staff time and effort, and confirmed the judgement by more than half of the repositories that implementation would be expensive.



**Table 3:** Summary of findings on ease of implementation of repository benefit metrics. For detailed list and description of metrics see Appendix A. Here the metrics are not necessarily named individually but restated in general terms.

<b>Currently measurable by most repositories</b>	<p><u>Derived from data holdings:</u> Temporal, spatial and subject coverage;</p> <p><u>Value of repository services:</u> number of data submitters and users supported, grants or projects served, workforce development achieved; cost savings for trustworthy data storage and distribution per submitter.</p> <p><u>Support for reuse:</u> completeness of metadata; expressiveness of metadata standard; presence and enforcement of data/metadata quality policies.</p> <p><u>Data reuse:</u> Numbers of downloads, page views, or distinct IP addresses accessing the data, of metadata pages accessed, data products and specific tool accessed; time spent at the site.</p>
<b>Possible in the foreseeable future with research, advanced technology and changed practices</b>	<p><u>Scientific impact:</u> extracted with artificial intelligence technologies from current publications, webpages, blogs, proposals and data management plans, and more reliably based on standardized data citations once practice is established.</p>
<b>Requiring major additional resources and expertise</b>	<p><u>Surveys:</u> interviews to ascertain user satisfaction and perceived impact on research success (research enabled, time saved, new questions developed).</p> <p><u>Economic and societal impact:</u> of data and data products beyond scientific use, or for fraud avoidance.</p>

Among all benefits metrics, the rate of implementation did not correlate positively with being evaluated as ‘important’. E.g., three metrics related to users’ interactions are *finding* and *accessing* data, and actual *downloads*. Interestingly, of these the first two (*finding* and *accessing* data) were deemed far more important than actual *downloads*, which ranked near the bottom in importance – yet number of *downloads* is a frequently implemented metric (likely due to its ease) while feedback on users ability to successfully find data of interest is hard to obtain.

Three metrics, *specific costs of preservation and related infrastructure*, *user support*, and *enabling future access* were the only metrics to consistently receive a high Importance score. However, this group – at best – were measured at only 40% of repositories. Only one repository (NCAR Research Data Archive) has a mechanism for anticipating future costs. Most (seven out of nine) were not yet able to implement this metric for various operational reasons.

Other benefits metrics ranked as important involved the ability to count data *use in publications* or citations and track *impact on societal priorities*. These reflect a repository’s ability to promote efficient data management, to provide for long-term usability of its data holdings, including the generation of new knowledge (proposals, studies), and to create value-added products. However, with the exception of being able to tally possible reuse (e.g., via page or catalog visits), many repositories had no current or planned mechanism to collect these, agreeing that a change in data citation practice, and more research or discussion were needed.

## Discussion

The broad range of perspectives reported here reflects the repositories’ diverse pre-existing priorities, technology choices, user interaction history, and resources, and occasionally, differences in interpretation. Varying degrees of effort are required to implement the different metrics types (e.g. **Table 3**). Budget-related and tangible metrics (e.g., *FTE*) are relatively easy to measure as are the *number of downloads from server logs*. However, the repository managers in this group rated *downloads* and *total page views* as lower in importance than did the Make Data Count project Kratz and Strasser (2015b). This generally reflects the uncertainty that downloads are a reliable correlate of use or impact, given the lapse in time between download, use, decision-making, and knowledge gained, and the difficulties tracking that pathway. These scores also reflect lingering concerns about over-standardization and interpretation. A simple measure like *downloads* ignores the volume of data downloaded; moreover, the repository will have done a better job if users do not need to download data excessively. However, *number of downloads* is typically valued by scientists publishing data and several community efforts are underway to standardize and track the number of dataset downloads (Kratz and Strasser 2015b). Generally, explicit *data citations* were rated as much more important here, which is consistent with Kratz and Strasser (2015b), though this tracks academic use only.

The metrics considered most important to a detailed understanding of value are typically intangible (e.g., the *benefits to future knowledge* or understanding the *impact on policy*, and *cost of ensuring future access*),

and will be a challenge to measure at all, much less measure consistently. The expectations of funding agencies to make data available with minimal conditions, i.e. without requiring account registration or user identification, dramatically reduces opportunities for gathering more detailed customer related metrics. If contact details are not collected at the time of download, e.g., by forms, questionnaires or log-in, the repository has limited ability to follow up with the user regarding the perceived quality of data and metadata, or the value and relevance to the intended purpose. Across all categories would be the need to balance the requirement for free access and privacy-compliant practices with requests from funding agencies or institutions to report on data usage, or even from the user communities themselves (e.g. to generate personalized data use statistics). This is an area of continued discussion with recognised benefits and disadvantages on all sides of the argument.

Given these limitations and the fact that many easily acquired benefit metrics are hard to translate into a comparable value for data (Kratz and Strasser 2015b) or repositories, data citation has been identified by this group and many others (e.g., <https://datacite.org/>, Kratz and Strasser 2015a) as the best metric for measuring 'value of repository' to the science community, if not the larger world. However, data citation is still evolving as a practice (Parsons, MA, et al., 2019, Garza and Fenner 2018, Data Citation Synthesis Group 2014, and references therein), and is neither a direct analog to paper citation nor firmly established within the Earth and environmental science communities. Hence, several represented repositories have resorted to manual linking of datasets to publications based on expert knowledge and manual or semi-automated literature searches. Wider implementation of Scholix will certainly help here (Cousijn, et al., 2019). In the future, we expect that reliable metrics about academic use will be based on standardized data citation practices. Established practices in turn, could form the training datasets for artificial intelligence technologies that more fully measure complex metrics (e.g., *policy impacts*) which extract usage from publications, laws and regulations, webpages, blogs, proposals, and data management plans.

Major additional resources or expertise will be required for socio-economic metrics that rely on user surveys and interviews to assess satisfaction and perceived value and impact; these include metrics such as *research enabled*, *time saved*, *new questions developed*, the *economic impact* of data and data products on society beyond scientific endeavors, and even the repositories' ability to engage the public and its scientific and data management communities. Those types of societal value and impact metrics require expertise generally not found among data curators or repository managers and necessitate targeted resources or funding for survey techniques and economics to simply define, let alone carry out. For example, the methods of Tanner (2012) for measuring impact of digital resources from memory institutions, such as museums and libraries could be adapted for use in this context. Some US federal agencies, e.g., the National Aeronautics and Space Administration (NASA), fund annual customer satisfaction surveys, with results typically driving repository activity over the next year. NASA also collects and has historically sponsored the creation of stories about the use and impact of particular types of data (Ramapriyan and Behnke, 2019). In some instances repositories have been able to obtain funding for advanced products including science analyses to support products for newly identified audiences (Baker et al., 2015).

Some agencies or networks have specifically targeted data synthesis and reuse efforts instead of new data creation, sometimes awarding funds primarily on that basis. These efforts will highlight the importance of initiatives such as "FAIR Data Principles" (Wilkinson, et al., 2016). For example, the Belmont Forum encourages new science to come from existing data concomitantly, even requesting examples of successful research working side-by-side with data management (Belmont Forum, 2018). The Marine Biodiversity Observation Networks specifically target existing long-term research-grade data to model practices for networks of scientists, resource managers and end-users (Wetzel et al., 2015). Repository curation and preservation services make these types of integrated, synthetic research possible.

### **Recommendations and conclusion**

1. Sponsors should invest in research on defining the most important complex benefit metrics, support their implementation, and support evolving repository practices in this new environment. Repositories should be involved in the research components to ensure applicability and feasibility.
2. An initial set of metrics for regular reporting by environmental science repositories should be those that are already measurable and generally useful, with consistent dashboards (such as those noted above or promoted by Make Data Count), but repositories should progressively develop specific metrics to suit their individual stakeholders, while coordinating with similar repositories to avoid duplication of effort.
3. Stakeholders should be aware that many extant ROI calculations from economics-based analyses or specialists are expensive and will happen only when resources permit.

Without the investment of curation and long-term funding of repositories to preserve their data holdings, further research with today's irreplaceable data will not be feasible. Repositories measure what is valuable to their stakeholders and reasonable to collect given their budgets and missions. Future science and societal needs will help determine the value of the long-term investment.

As research data publishing and data repositories continue to mature into an integral part of our scientific research endeavors, we should expect to gain a better understanding of the costs and benefits of publishing and preserving research data. We should also expect to see a rationalization of the repository landscape with refined practices for how and where research data are curated. It may be determined that fewer repositories will better leverage the investments made in infrastructure, or alternatively metadata aggregators will provide an ideal entry point into smaller, discipline-focused repositories. Either way, our goal should be to maximize the availability and usability of the data produced. By achieving this goal we can ensure that we extract the maximum value from our research funding while also improving the transparency and credibility of the conclusions drawn.

## Data Availability Statement

Gries et al., 2018. (see References).

## Notes

Ag Data Commons, <https://data.nal.usda.gov>, <http://doi.org/10.17616/R3G051>;

Environmental Data Initiative, <https://environmentaldatainitiative.org>, <http://doi.org/10.25504/FAIRsharing.xd3wmy>;

DataONE, <https://www.dataone.org/>, <http://doi.org/10.17616/R3101G>;

Interdisciplinary Earth Data Alliance (IEDA) <https://www.iedadata.org/> <http://doi.org/110.25504/FAIRsharing.be9dj8>;

NASA DAAC at National Snow and Ice Data Center (NSIDC), <https://nsidc.org/daac/>;

Exchange for Local Observations and Knowledge of the Arctic (ELOKA), <https://eloka-arctic.org/>;

NASA Socioeconomic Data and Applications Center (SEDAC), <https://sedac.ciesin.columbia.edu/>;

NCAR Research Data Archive (RDA), <https://rda.ucar.edu/>, <http://doi.org/10.17616/R3H01T>;

Laboratory for Atmosphere and Space Physics (LASP), <http://lasp.colorado.edu/home/>

## Additional File

The additional file for this article can be found as follows:

- **Appendix A.** Metrics examined. DOI: <https://doi.org/10.5334/dsj-2019-058.s1>

## Acknowledgements

This work was supported by the National Science Foundation grant DBI-EAGER 1500306. The contributions of Cynthia S. Parr were supported by the by the U.S. Department of Agriculture, Agricultural Research Service (Project # 8260-88888-003-00D). The contributions of Robert R. Downs were supported by the National Aeronautics and Space Administration (NASA) under Contract 80GSFC18C0111 for the Socioeconomic Data and Applications Distributed Active Archive Center (DAAC). The contributions of Rebecca Koskela were supported by The National Science Foundation Grant 1430508. The contributions of Ruth Duerr were supported by the National Science Foundation Grant 1513438. The contributions of Shelley Stall were supported by the American Geophysical Union. USDA is an equal opportunity provider and employer.

## Competing Interests

The authors represent the repositories that are described in this article.

## References

**AGU.** 2013. Earth and Space Science Data Should Be Credited Preserved Open and Accessible as an Integral Responsibility of Scientists Data Stewards and Sponsoring Institutions. September, 2015. Available at: <https://sciencepolicy.agu.org/files/2013/07/AGU-Data-Position-Statement-Final-2015.pdf> [Last accessed 24 July 2018].

**Baker, KS and Duerr, RE.** 2017. Research and the changing nature of data repositories. Chapter 1 in: Johnston, L (ed.), *Curating Research Data: A Handbook of Current Practice, Volume One: Practical Strategies for Your Digital Repository*, 33. Chicago, IL: Association of College & Research Libraries.



- Baker, KS, Duerr, RE and Parsons, MA.** 2015. Scientific knowledge mobilization: Co-evolution of data products and designated communities. *International Journal of Digital Curation*, 10(2): 110–135. DOI: <https://doi.org/10.2218/ijdc.v10i2.346>
- Beagrie, N and Houghton, J.** 2013. The Value and Impact of Data Sharing and Curation. A synthesis of three recent studies of UK research data centres. *JISC Report*. Available at [http://repository.jisc.ac.uk/5568/1/iDF308\\_-\\_Digital\\_Infrastructure\\_Directions\\_Report,\\_Jan14\\_v1-04.pdf](http://repository.jisc.ac.uk/5568/1/iDF308_-_Digital_Infrastructure_Directions_Report,_Jan14_v1-04.pdf) [Last accessed 02 May 2019].
- Belmont Forum.** 2018. Science-driven e-Infrastructure Innovation (SEI) for the Enhancement of Transnational, Interdisciplinary and Transdisciplinary Data Use in Environmental Change. Available at <https://www.belmontforum.org/news/upcoming-funding-opportunity-science-driven-e-infrastructure-innovation-sei-for-the-enhancement-of-transnational-interdisciplinary-and-transdisciplinary-data-use-in-environmental-change/> [Last accessed 10 May 2019].
- Bernknopf, R, Kuwayama, Y, Gibson, R, Blakely, J, Mabee, B, Clifford, TJ, Quayle, B, Epting, J, Hardy, T and Goodrich, D.** 2019. The Cost-Effectiveness of Satellite Earth Observations to Inform a Post-Wildfire Response. *Resources for the Future*. Available at <https://www.rff.org/publications/working-papers/cost-effectiveness-satellite-earth-observations-inform-post-wildfire-response/>.
- Colavizza, G, Hrynaszkiewicz, I, Staden, I, Whitaker, K and McGillivray, B.** 2019. The citation advantage of linking publications to research data. *arXiv preprint arXiv:1907.02565*. <https://arxiv.org/pdf/1907.02565.pdf>.
- Cousijn, H, Feeney, P, Lowenberg, D, Presani, E and Simons, N.** 2019. Bringing Citations and Usage Metrics Together to Make Data Count. *Data Science Journal*, 18(1): 9. DOI: <https://doi.org/10.5334/dsj-2019-009>
- Curation Cost Exchange.** Available at: <http://www.curationexchange.org/>. [Last accessed 23 May 2019].
- Data Citation Synthesis Group.** 2014. Joint Declaration of Data Citation Principles. In: Martone, M (ed.). San Diego, CA: FORCE11. DOI: <https://doi.org/10.25490/a97f-egy>
- Dillo, I, Hodson, S and de Waard, A.** 2016. 'Income Streams for Data Repositories: Final Report of RDA-WDS Cost Recovery Interest Group'. *Research Data Alliance*.
- Downs, RR.** 2018. Enabling the ReUse of Geospatial Information. In: Kruse, JB, Cromptoets, J and Pearlman, F (eds.), *GeoValue: The Socioeconomic Value of Geospatial Information*, 129–146. Boca Raton: CRC Press. DOI: <https://doi.org/10.1201/9781315154640-9>
- Downs, RR, Duerr, R, Hills, DJ and Ramapriyan, HK.** 2015. Data Stewardship in the Earth Sciences. *D-Lib Magazine*, 21(7/8). DOI: <https://doi.org/10.1045/july2015-downs>
- Dubrow, A.** 2018. Preventing natural hazards from becoming societal disasters. <https://www.tacc.utexas.edu/-/preventing-natural-hazards-from-becoming-societal-disasters> [Last accessed 16 May 2019].
- Fenner, M, Lowenberg, D, Jones, M, Needham, P, Vieglais, D, Abrams, S, Cruse, P and Chodacki, J.** 2018. Code of practice for research data usage metrics release 1. *PeerJ Preprints*, 6: e26505v1. DOI: <https://doi.org/10.7287/peerj.preprints.26505v1>
- Forney, WM, Raunikar, R, Mishra, S and Bernknopf, R.** 2012. "An economic value of remote sensing information: Application to agricultural production and maintaining ground water quality." In *2012 Socio-economic Benefits Workshop: Defining, measuring, and Communicating the Socio-economic Benefits of Geospatial Information*, 1–6. IEEE.
- Garza, K and Fenner, M.** 2018. Glad You Asked: A Snapshot of the Current State of Data Citation. *DataCite Blog*. DOI: <https://doi.org/10.5438/h16y-3d72>
- Gewin, V.** 2016. Data sharing: An open mind on open data. *Nature*, 529: 117–119. DOI: <https://doi.org/10.1038/nj7584-117a>
- GO FAIR.** 2016. GO FAIR Initiative. Available at <https://www.go-fair.org> [Last accessed 10 May 2019].
- Gries, C, Downs, RR, O'Brien, M, Parr, C, Duerr, R, Koskela, R, Tarrant, P, Maull, KE, Stall, S, Wilson, A, Hoebelheinrich, N and Lehnert, K.** 2018. Return on Investment Metrics for Data Repositories in Earth and Environmental Sciences. *Environmental Data Initiative*. Dataset accessed 5/15/2019. DOI: <https://doi.org/10.6073/pasta/d49bec63f51603512efa7e0fd2717203>
- Guha, RV, Brickley, D and McBeth, S.** 2015. Schema.org: Evolution of Structured Data on the Web. <https://queue.acm.org/detail.cfm?id=2857276> (accessed 2019-09-15).
- Günther, WA, Rezazade, MH, Mehrizi, R, Huysman, M and Feldberg, F.** 2017. Debating big data: A literature review on realizing value from big data. *The Journal of Strategic Information Systems*, 26(3): 191–209. DOI: <https://doi.org/10.1016/j.jsis.2017.07.003>


- International Data Corporation (IDC).** 2018. Worldwide Semiannual Big Data and Analytics Spending Guide. Available at [https://www.idc.com/getdoc.jsp?containerId=IDC\\_P33195](https://www.idc.com/getdoc.jsp?containerId=IDC_P33195) [Last accessed 16 May 2019].
- IPCC.** 2014. *Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, Field, CB, Barros, VR, Dokken, DJ, Mach, KJ, Mastrandrea, MD, Bilir, TE, Chatterjee, M, Ebi, KL, Estrada, YO, Genova, RC, Girma, B, Kissel, ES, Levy, AN, MacCracken, S, Mastrandrea, PR and White, LL (eds.), 1132. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press. [https://www.ipcc.ch/site/assets/uploads/2018/02/WGIAR5-PartA\\_FINAL.pdf](https://www.ipcc.ch/site/assets/uploads/2018/02/WGIAR5-PartA_FINAL.pdf) (Last accessed 2019-09-24).
- Kratz, JE and Strasser, C.** 2015a. Making data count. *Scientific Data*, 2: 150039. DOI: <https://doi.org/10.1038/sdata.2015.39>
- Kratz, JE and Strasser, C.** 2015b. Researcher perspectives on publication and peer review of data. *PLOS ONE*, 10(4): e0123377. DOI: <https://doi.org/10.1371/journal.pone.0123377>
- Kuwayama, Y and Mabee, B.** 2018. Quantifying the socioeconomic benefits of satellite data applications at different decision-making scales. *AGU Fall Meeting Abstracts*. Available at <https://ui.adsabs.harvard.edu/abs/2018AGUFM.B44A..02K/abstract> [Last accessed 09 May 2019].
- Lawrence, SR, Connaway, LS and Brigham, KH.** 2001. Life Cycle Costs of Library Collections: Creation of Effective Performance and Cost Metrics for Library Resources. *College & Research Libraries* [Online], 62(6): 541–553. DOI: <https://doi.org/10.5860/crl.62.6.541>
- Lindenmayer, D and Likens, GE.** 2013. Benchmarking Open Access Science Against Good Science. *Bulletin of the Ecological Society of America*, 94: 338–340. DOI: <https://doi.org/10.1890/0012-9623-94.4.338>
- Longo, DL and Drazen, JM.** 2016. Data sharing. *New England Journal of Medicine*, 374(3): 276–7. DOI: <https://doi.org/10.1056/NEJMe1516564>
- Maness, J, Duerr, R, Dulock, M, Fetterer, F, Hicks, G, Merredyth, A, Sampson, W and Wallace, A.** 2017. Revealing our melting past: Rescuing historical snow and ice data. *GeoRes J*, 14: 92–97. DOI: <https://doi.org/10.1016/j.grj.2017.10.002>
- McNutt, M, Lehnert, K, Hanson, B, Nosek, BA, Ellison, AM and King, JL.** 2016. Liberating field science samples and data. *Science*, 351: 1024–1026. DOI: <https://doi.org/10.1126/science.aad7048>
- Milham, MP, Cameron Craddock, R, Son, JJ, Fleischmann, M, Clucas, J, Xu, H, Koo, B, Krishnakumar, A, Biswal, BB, Castellanos, FX, Colcombe, S, Di Martino, A, Zuo, X-N and Klein, A.** 2018. Assessment of the impact of shared brain imaging data on the scientific literature. *Nature Communications*, 9: 1–7. DOI: <https://doi.org/10.1038/s41467-018-04976-1>
- NOAA Stories.** 2019. (updated frequently). All Stories|National Oceanographic and Atmospheric Administration. Available at <https://www.noaa.gov/stories/> [Last accessed 24 September, 2019].
- Parsons, MA, Duerr, RE and Jones, MB.** 2019. The History and Future of Data Citation in Practice. *Data Science Journal*, 18(1): 52. DOI: <https://doi.org/10.5334/dsj-2019-052>
- Phelps, R.** 2003. Only rigorous metrics can demonstrate the value of IT to the rest of the business. *Computer Weekly*, 22. December 16 2003.
- Pinelli, JP, Rodriguez, D, Roueche, DB, Gurley, K, Baradaranshoraka, M, Cocke, S, Dong-Wook, S, Lapaiche, L and Gay, R.** 2018. Data management for the development of a flood vulnerability model. In: *Proceedings of European Safety and Reliability Conference*, Trondheim, Norway, 17–21 June 2018, 2781–2788. DOI: <https://doi.org/10.1201/9781351174664>
- Piowar, HA, Vision, TJ and Whitlock, MC.** 2011. Data archiving is a good investment. *Nature*, 473: 285. (19 May 2011). DOI: <https://doi.org/10.1038/473285a>
- Popkin, G.** 2019. Data sharing and how it can benefit your scientific career. *Nature*, 569: 7756. DOI: <https://doi.org/10.1038/d41586-019-01506-x>
- Ramapriyan, H and Behnke, J.** 2019. Importance and Incorporation of User Feedback in Earth Science Data Stewardship. *Data Science Journal*, 18(1). Ubiquity Press. DOI: <https://doi.org/10.5334/dsj-2019-024>
- Rubin, H.** 1991. *Capacity Management Review*, 19(1): 1. (Jan 1991). DOI: [https://doi.org/10.1016/0045-7930\(91\)90013-8](https://doi.org/10.1016/0045-7930(91)90013-8)
- Stall, S, Yarmey, LR, Boehm, R, Helena Cousin, H, Patricia Cruse, P, Joel Cutcher-Gershenfeld, J, Robin Dasler, R, de Waard, A, Duerr, R Elger, K, Fenner, M, Glaves, H, Hanson, B, Hausman, J, Heber, J, Hills, DJ, Hoebelheinrich, N, Hou, S, Kinkade, D, Koskela, R, Martin, R, Lehnert, K, Murphy, F, Nosek, B, Parsons, MA, Petters, J, Plante, R, Robinson, E, Samors, R, Servilla, M, Ulrich, R, Witt, M and Wyborn, L.** 2018. Advancing FAIR Data in Earth, Space, and Environmental Science. *Eos*, 99. Published on 05 November 2018. DOI: <https://doi.org/10.1029/2018EO109301>

- Starr, J, Castro, E, Crosas, M, Dumontier, M, Downs, RR, Duerr, R, Haak, LL, Haendel, M, Herman, I, Hodson, S, Hourclé, J, Kratz, JE, Lin, J, Nielsen, LH, Nurnberger, A, Proell, S, Rauber, A, Sacchi, S, Smith, A, Taylor, M and Clark, T.** 2015. Achieving human and machine accessibility of cited data in scholarly publications. *PeerJ Computer Science*, 1: e1. DOI: <https://doi.org/10.7717/peerj-cs.1>
- Statista.** 2018. Big data market size revenue forecast worldwide from 2011 to 2027. March 2018. Available at <https://www.statista.com/statistics/254266/global-big-data-market-forecast/> [Last accessed 16 May 2019].
- Tanner, S.** 2012. Measuring the Impact of Digital Resources: The Balanced Value Impact Model. King's College London. October 2012. Available at: [https://www.kdl.kcl.ac.uk/fileadmin/documents/pubs/BalancedValueImpactModel\\_SimonTanner\\_October2012.pdf](https://www.kdl.kcl.ac.uk/fileadmin/documents/pubs/BalancedValueImpactModel_SimonTanner_October2012.pdf) [Last accessed 23 May 2019].
- Voosen, P.** 2017. Q&A: NASA's Science Head Says Investment in Earth Science Is a 'no-Brainer'. *Science*. <https://www.sciencemag.org/news/2017/03/qa-nasa-s-science-head-says-investment-earth-science-no-brainer>. DOI: <https://doi.org/10.1126/science.aal0906>
- Wetzel, FT, Saarenmaa, H, Regan, E, Martin, CS, Mergen, P and Smirnova, L.** 2015. The roles and contributions of Biodiversity Observation Networks (BONs) in better tracking progress to 2020 biodiversity targets. *Biodiversity*, 16: 137–149. DOI: <https://doi.org/10.1080/14888386.2015.1075902>
- Wilkinson, MD, Wilkinson, MD, Dumontier, M, Aalbersberg, IJ, Appleton, G, Axton, M, Baak, A, Blomberg, N, Boiten, J, da Silva Santos, LB, Bourne, PE, Bouwman, J, Brookes, AJ, Clark, T, Crosas, M, Dillo, I, Dumon, O, Edmunds, S, Evelo, CT, Finkers, R, Gonzalez-Beltran, A, Gray, AJG, Groth, P, Goble, C, Grethe, JS, Heringa, J, Hoen, PA, Hooft, R, Kuhn, T, Kok, R, Kok, J, Lusher, SJ, Martone, ME, Mons, A, Packer, AL, Persson, B, Rocca-Serra, P, Roos, M, Rene van Schaik, R, Susanna-Assunta Sansone, SA, Schultes, E, Sengstag, T, Slater, E, Strawn, G, Swertz, MA, Thompson, M, van der Lei, J, van Mulligen, E, Velterop, J, Waagmeester, A, Wittenburg, P, Wolstencroft, K, Zhao, J and Mons, B.** 2016. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Sci. Data*, 3: 160018. DOI: <https://doi.org/10.1038/sdata.2016.18>

**How to cite this article:** Parr, C, Gries, C, O'Brien, M, Downs, RR, Duerr, R, Koskela, R, Tarrant, P, Maull, KE, Hoebelheinrich, N and Stall, S. 2019. A Discussion of Value Metrics for Data Repositories in Earth and Environmental Sciences. *Data Science Journal*, 18: 58, pp.1–11. DOI: <https://doi.org/10.5334/dsj-2019-058>

**Submitted:** 23 May 2019    **Accepted:** 12 November 2019    **Published:** 09 December 2019

**Copyright:** © 2019 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

 *Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

**OPEN ACCESS** 