Recent policy shifts on the part of funding agencies and journal publishers are causing changes in the acknowledgment and citation behaviors of scholars. A growing emphasis on open science and reproducibility is changing how authors cite and acknowledge “research infrastructures”—entities that are used as inputs to or as underlying foundations for scholarly research, including data sets, software packages, computational models, observational platforms, and computing facilities. At the same time, stakeholder interest in quantitative understanding of impact is spurring increased collection and analysis of metrics related to use of research infrastructures. This article reviews work spanning several decades on tracing and assessing the outcomes and impacts from these kinds of research infrastructures. We discuss how research infrastructures are identified and referenced by scholars in the research literature and how those references are being collected and analyzed for the purposes of evaluating impact. Synthesizing common features of a wide range of studies, we identify notable challenges that impede the analysis of impact metrics for research infrastructures and outline key open research questions that can guide future research and applications related to such metrics.

Introduction

Studies of metrics related to scholarly work typically focus on published documents (Nicolaisen, 2007; White & McCain, 1989), website linkages (Thelwall, Vaughan, & Björneborn, 2005), or social media metrics (Bornmann, 2014; Haustein et al., 2014; Priem & Hemminger, 2010; Sud & Thelwall, 2014; Thelwall, Haustein, Larivière, & Sugimoto, 2013). These studies group in different ways under various names, including “bibliometrics,” “scientometrics,” “informetrics,” “webometrics,” and “altmetrics” (Bar-Ilan, 2008; Hood & Wilson, 2001; Milojević & Leydesdorff, 2013; Wilson, 1999). The results of such studies range widely, but are often used to perform comparative analysis of individual scholars, organizations, literatures, countries, and journals, based on particular
characteristics of texts, such as reference and citation patterns, word frequencies and distributions, and hypertext or linguistic linkages (Borgman & Furner, 2002).

Recently, the interest in the collection and analysis of research metrics has extended to encompass more types of resources and infrastructures, specifically entities that are used as inputs to, or underlying infrastructure for, research studies. These entities include data sets, software packages, computational models, observational platforms, and computing facilities. Of these entities, most scholarly attention has been given to understanding metrics related to the use of digital data sets (CODATA-ICSTI, 2013; Costas, Meijer, Zahedi, & Wouters, 2013; Costello, 2009), but interest is expanding to other kinds of resources. Recent policy shifts on the part of funding agencies and journal publishers are affecting changes in acknowledged and citation behaviors in relation to research resources and infrastructures, further shifting norms of digitally enabled research that emphasize notions of “reproducibility” and “open science” (Stodden, Leisch, & Peng, 2014; Willinsky, 2005; Woelfle, Olliaro, & Todd, 2011). The U.S. National Science Foundation (NSF), for example, changed its grant proposal guidelines in 2013 to allow proposal submitters’ biographical sketches to include descriptions of “research products,” in addition to citations of publications. The revised guidelines specifically note that this change allows biosketches to include references to data sets and software (NSF, 2013). Similarly, journal publishers, including Nature,1 PLOS,2 American Geophysical Union,3 and American Meteorological Society,4 have developed policies that mandate or strongly recommend the archiving and citation of research data and other resources within journal articles. In 2012, Thomson Reuters released the Data Citation Index (DCI), which is intended to produce the same kinds of citation indices for data sets and data repositories that have existed for journal articles and other document types in the Web of Science (WoS) databases for many years (Thomson Reuters, 2012).

This article reviews work spanning the past several decades that focuses on assessing and tracing the use of these types in the Web of Science (WoS) databases for many years that have existed for journal articles and other document types in the Web of Science (WoS) databases for many years (Thomson Reuters, 2012).

This article reviews work spanning the past several decades that focuses on assessing and tracing the use of these kinds of research infrastructures. We use the term “research infrastructures” as an overarching grouping of multiple kinds of resources, building on terminology established in international contexts (see OECD, 2008, 2014), particularly as defined by the European Commission (2015):

“Research Infrastructures” are facilities, resources and services that are used by the research communities to conduct research and foster innovation in their fields. They include: major scientific equipment (or sets of instruments), knowledge-based resources such as collections, archives and scientific data, e-infrastructures, such as data and computing systems and communication networks and any other tools that are essential to achieve excellence in research and innovation. They may be “single-sited,” “virtual” and “distributed.” (p. 4)

This review discusses how research infrastructures are identified and referenced in the research literature and how those references are being collected and analyzed for the purposes of impact evaluation. We begin with some background on bibliometric studies more broadly, which provides the framework for this discussion in terms of the behavioral and methodological issues. We then review the literature covering three categories of infrastructure: data sets, software, and platforms and facilities. The data curation community has made perhaps the most progress toward addressing the behavioral aspects of tracing the impact of data collections. However, these recommended practices have not yet obviated the methodological issues in conducting impact studies. Bibliometric and impact studies of scientific software impacts are more limited, but a growing number of efforts are pursuing research in this area. For platforms and facilities, we examine lines of research related to the impact of both space telescopes and high-performance computers. We next examine a number of studies that leveraged “proxy metrics,” such as downloads and usage data, to bypass some of the methodological issues faced by studies based on scholarly outputs. Finally, we conclude by synthesizing the commonalities that span these studies—including analytical goals, metrics gathered, and methodologies—and the common challenges, barriers to progress, and open research questions for those interested in better understanding the scholarly impacts of research infrastructures.

**Background**

Borgman and Furner (2002) discussed how bibliometric studies can be broken down into two types: “studies that use bibliometric methods, and on the other [hand], studies of the use of bibliometric methods” (p. 8). They categorize these two types as “behavioral” and “methodological” studies, respectively, with the first type generally using bibliometric techniques to study the behavior of scholars, and the second type studying the effectiveness, merit, and uses of bibliometric techniques themselves. Studies of metrics related to research infrastructures fit into both of these categories. Many of the studies reviewed here can be categorized as what Borgman and Furner termed “evaluative link analyses,” in that the motivation for developing, using, and applying bibliometric techniques to the study of research infrastructures is to use the results as “indicators or measurements of the level of quality, importance, influence, or performance” (p. 11) of those infrastructures. Managers and producers of research infrastructures want to understand who is using their resources, and for what purpose. Similarly, funders of these infrastructures want better ways to measure their use and utility.

“Methodological” studies within this set of literature differ from the methodological studies reviewed in the traditional bibliometric and informetric literature, because the central methodological challenge for research infrastructure

---

1http://www.nature.com/authors/policies/availability.html.
2http://www.plosone.org/static/policies.action.
3http://publications.agu.org/author-resource-center/publication-policies/data-policy/.
4http://www.ametsoc.org/PubsDataPolicy.
metrics is the lack of consistent and sustainable ways to gather the underlying data. Noncontextual research resources are, at best, cited and referenced inconsistently within the scientific literature. Typically, however, infrastructural resources, such as data sets, software, and facilities, have not been cited or referenced at all in past scientific studies. The robust tools that have supported the development of the bibliometric and informetric disciplines—such as Dialog, the WoS, and Scopus—do not yet exist for the collection and analysis of metrics for most research infrastructures. The aforementioned Thomson Reuters DCI, still in its nascent stages, is not sufficiently populated with citation information to enable robust citation-based analyses of data (see more discussion below in the section on Data Citation). No single tool yet provides indexing that includes citations to software, facilities, or other types of research infrastructures. The WoS’s “Cited Reference Search” capability has limited utility when looking for references to research infrastructure, because such references, if provided at all, have traditionally been given in the methods or acknowledgments sections of published literature, not in reference lists.

Studies of research infrastructure metrics thus face similar, if not greater, challenges as studies of acknowledgments. Acknowledgments require notoriously labor-intensive data-gathering procedures given their nature as unstructured textual statements. However, studies of acknowledgments have shown that scholars exhibit norms, patterns, and trends in how they write acknowledgment statements (Cronin, 1995, 2001). Such studies have also produced a number of acknowledgment typologies. These typologies commonly include one or more categories for acknowledgments of the use of research infrastructures, such as data, facilities, or instruments (see Cronin, 1995; McCain, 1991; Cronin, Shaw, & La Barre, 2004). Acknowledgment studies typically use manual or semiautomated methods for extracting acknowledgment statements from published literature. Giles and Councill (2004) and Thomer and Weber (2014) presented automated methods for extracting acknowledgments from the CiteSeer database and PubMed Central’s Open Access corpus, respectively, but those techniques have not been carried forward into more-general tools for acknowledgment extraction or indexing.

Similar to typologies of acknowledgments, typologies of the purposes of citation also often include categories related to the use of data, tools, or equipment. Such citations have been characterized with a variety of labels (see Bornmann & Daniel, 2008; Cronin, 1984; Garfield, 1996, for examples). Bornmann and Daniel (2008) categorize these kinds of citations as “citations of the methodological type,” which includes citations for the “use of materials, equipment, practical techniques, or tools of cited work; use of analysis methods, procedures, and design of cited work” (p. 66). Recent efforts to develop web-enabled ways of characterizing and publishing citation relationships likewise include categories of citation related to research infrastructures. One notable example is the Citation Typing Ontology, which includes “cites as data source” and “uses data from” as citation relationship types (Peroni & Shotton, 2012; Shotton, 2010).

Thus, citations and acknowledgments for research infrastructures do show up in the scholarly literature, but in inconsistent and hard-to-analyze ways. The following sections discuss trends and research outcomes related to the citation and attribution tracing for data, software, research platforms, and facilities. Because most efforts to track references to research infrastructures focus on one particular type of resource, each type is discussed separately.

Data Citation and Attribution Tracing

The data curation community has been active in recent years to promote “data citation” as a common practice. Multiple international interest groups recently came together to create a consensus set of data citation principles (Data Citation Synthesis Group, 2014). These principles were made public in February 2014 and had been endorsed by over 100 organizations as of December 2015. This data citation principles initiative and the studies that will be reviewed here largely focus on quantitative data, though some similar discussions around research traceability and citation of data are also taking place in relation to qualitative data (Moravcsik, 2010). Data creators and users are being encouraged to create formal citations and references to data, using these persistent identifiers as the key mechanism to help readers find and access the associated data.

In turn, this community has also recognized that standardizing data citation behavior requires consistent methods of identifying data sets. Without citations to uniquely identify data, it is difficult, if not impossible, to develop tools to measure the impact such resources and infrastructures have within the communities they belong to, or to understand the spread of that impact to broader scientific communities. Significant progress has been made in the past few years in developing recommendations, policies, and procedures for creating and promoting citeable identifiers, ranging from general recommendations (Ball & Duke, 2011; Duerr et al., 2011) to recommendations from discipline-specific groups (Federation of Earth Science Information Partners, 2012; Mayernik, Ramamurthy, & Rauber, 2015). Spurred by these initiatives, research organizations, data repositories, and universities are beginning to assign en masse persistent web-accessible identifiers to scientific resources, with most efforts utilizing the DOI system (Brase, Sens, & Lautenschlager, 2015; Paskin 2005).

These efforts to develop recommendations and principles for data citation primarily address solutions to this issue moving forward in time. The first commercial products to index data citations, notably the Thomson Reuters DCI, are largely experiments still in their infancy and have yet to be proven cost-effective or efficient. Early assessments of the DCI showed that the vast majority of included records were from a small number of data archives (Torres-Salinas, Martin-Martin, & Fuente-Gutiérrez, 2014) and showed no recorded citations (Peters, Kraker, Lex, Gumpenberger, &
Gorraiz, in press). At present, these skewed distributions limit the utility of the DCI for research into data citation patterns (Robinson-García, Jiménez-Contreras, & Torres-Salinas, in press). Therefore, the work to understand the past use of research infrastructures has to date focused on citations that have been made in previously published work and has required the use of manual bibliometric data gathering or automated text-mining techniques.

**Data Citation Studies Using Manual Methods**

Manual processes for collecting publications that cite data sets and other research infrastructures are fraught with problems given that they are time-consuming, require multiple tools (e.g., WoS, Google Scholar, and literature databases), and do not scale. Such manual search processes, however, can be effective in producing case studies that demonstrate the value of scientific infrastructures within particular communities. They also provide quantitative measures of how the use of data sets are acknowledged or cited within scientific communities.

A number of small-scale, manually conducted studies have investigated the usage and citation patterns of particular scientific data sets and collections. These studies generally focus on either behavioral analysis of citations or methodologies for impact and productivity metrics (or both behavior and metrics). Kirlaw (2011) studied the usage patterns of 21 life science data repositories from 1976 to 2010, using the Science Citation Index of the WoS and the PubMed database to find publications and the U.S. Patent Office search tools to find patents. Numbers of publications and patents that used these repositories were then compiled, based on mentions of the repository names in article abstracts and bibliographic metadata. Results showed that some repositories are much more widely used than others, with GenBank showing, by far, the largest usage. Kirlaw also used this study to estimate that around 2.5% of all biology-related journal articles and 14.7% of all biology-related patents used data from one or more of these repositories. Alsheikh-Ali, Qureshi, Al-Mallah, and Ioannidis (2011) studied data availability in relation to articles published within the 50 highest impact factor WoS journals. They found that only 47 of 500 articles analyzed (9%) had deposited “full primary raw data” online, despite the fact that more than 60% of the articles were published in journals that contained data-related policies of some kind. Major (2011) analyzed the usage of data from National Aeronautics and Space Administration (NASA) Earth Observing System (EOS) data centers. Relevant publications over the period 2000–2009 were identified by searching the WoS for specific NASA EOS instrument keywords and qualifying keywords that reduced false positives. Major found 5,633 articles published in over 400 different journals that referenced the use of NASA EOS instrument data, with over 55% of those papers referencing data from one instrument, the Moderate Resolution Imaging Spectroradiometer. The number of articles found largely correlated with the volume of data distributed through the NASA EOS data centers. As an indicator of the impact of these data, Major notes that this set of articles represents approximately 37% of the articles published in the WoS Earth science category journals during the analyzed time period. (2014) analyzed citation counts for three data sets archived at the National Oceanographic Data Center. Three sources were used to find relevant publications: the WoS, Google Scholar, and direct searching by journal publishers’ websites. The most widely used data set analyzed within ’s study, the World Ocean Atlas and World Ocean Database, was referenced by more than 8,000 published articles. Though the other two data sets analyzed showed lower usage rates, ’s analysis notes that the estimated citation counts for these three data sets are higher than 99% of all the oceanography journal articles published during the same years. Henderson and Kotz (2015) analyze the use of the Community Resource for Archiving Wireless Data At Dartmouth (CRAWDAD), a data archive for wireless network studies. They used a variety of sources, including the ACM Digital Library, Google Scholar, IEEE Xplore, ScienceDirect, and Scopus, to identify 1,281 articles that mentioned use of CRAWDAD data sets.

Within the social sciences, studying data citation has a much longer history. In the late 1970s, Dodd (1979) published recommendations for the format of references to machine-readable data files (MRDFs). White (1982) followed this up with a small study of how social science data sets were actually being cited. The first sentence of White’s article reads, “[a]n argument by no means new is that social scientists who work with machine-readable data files (MRDF) should cite them in their writings, with formal references set apart from main text, just as they now do books, papers and reports” (p. 467). In White’s study, three sets of data files were examined, using the WoS’s Social Science Citation Index (SSCI). White found that citation of data sets was highly inconsistent and incomplete, and required considerable effort to uncover.

This interest in the citation of social science data largely waned, however, after those initial studies. The social science literature shows little work on this topic between 1985 and 2005, with the exception of Sieber and Trumbo (1995), who studied how social scientists referenced their use of the General Social Survey (GSS), a large and widely used survey data set. Sieber and Trumbo analyzed 198 articles for the presence of several key data citation elements, such as the GSS Principal Investigator (PI), title, and producers of the GSS. References to these elements were found in various locations across the examined articles. In addition, key elements were often totally missing, with 60% of articles failing to mention the PI at all, and 9% failing to mention the name of the GSS itself. More recently, the interest within the social sciences in data citation has again picked up. Mooney (2011) analyzed a sample of 49 journal articles that were based on the use of data sets from the Inter-university Consortium for Political and Social Research (ICPSR) data archive. In this study, 61% of the articles failed to reference ICPSR or the data set used at all. Mooney and Newton
(2012) analyzed 65 articles, randomly sampled roughly equally from the H.W. Wilson Science, Humanities, and Social Science indexes (now part of EBSCO Information Services). Following a similar methodology as Sieber and Trumbo, Mooney and Newton found that mentions of the data set title were the prominent mode of citation and were the only method with more than 50% prevalence. Zenk-Möltgen and Lepthien (2014) analyzed 222 empirical articles published over a period of 2 years in five sociology journals. The journals were selected from the sociology category of the WoS’s SSCI. Of these articles, around half stated that the data used to produce the article were available. In following URLs and identifiers that authors provided for data sets, however, Zenk-Möltgen and Lepthien were only able to confirm that data were truly accessible for around 16% of the sampled articles.

Most of the studies mentioned in this section discuss common methodological challenges related to the variability of references to data resources (i.e., where and how references were provided in the analyzed publications). Scientists and social scientists refer to data sets throughout the texts of articles, and do so in a variety of ways. Starting with White (1982), the inconsistency of references to data sets is a common refrain. These targeted manual studies confirm that many references to such data sets typically occur in the full text of articles, instead of article metadata, abstracts, or references (see Coppin, 2013, for another example). In addition, Belter (2014) discusses how standard referencing recommendations for data seem to have had little effect on the consistency of subsequent data-set citations. Henderson and Kotz (2015) also note that asking users to proactively send information about published articles back to the data archive was highly ineffective.

Together, these manually compiled studies of data citation within the sciences and social sciences illustrate how information about the use of particular data sets and repositories can be found in the published literature. The human capital required to do this manual work, however, is far too scarce to be practical beyond small-scale case studies. Automated processes are thus of considerable interest to enable these traceability efforts to scale with the numbers of data sets being created in the era of digital science.

Data Citation Studies Using Computational Methods

Although manual efforts have been successful at capturing citation information for small-scale case studies, they lack the accuracy, speed, and scale necessary to collect and study large enough corpora to be meaningful, repeatable, and statistically generalizable. The goals of computational approaches to studying data citation are thus to: (i) design and develop computational heuristics for finding citations that generalize to a broader set of citation requirements and (ii) design and develop scalable software platforms that can be robustly applied to finding and analyzing these citations. When formal references are included in a scholarly work, for example, to data sets using unique identifiers (e.g., DOIs), analysis tasks are computationally trivial, given that searching for such DOIs requires little effort in developing collection algorithms. However, the complexity of finding references to data sets increases significantly when persistent identifiers are not cited or, as is commonly the case, when a data set is not cited directly (such as when references are made to an article that describes the data set). With these complexities to overcome, the studies that will be discussed here primarily have a methodological focus, though, in some cases, they do apply their new methods to citation behavior and metric analyses.

A notable study by Pepe, Goodman, Muench, Crosas, and Erdmann (2014) automated the analysis of data-related URLs found in the published astronomy papers of the American Astronomical Society. In lieu of formal citations to data sets, the authors mined these URLs and categorized their link status (whether or not they resolve; and whether or not they were links to scholarly repositories). By further developing heuristics to analyze the life cycle of the URLs, the authors were able to characterize which URLs were still valid, for what period of time postpublication, and thus begin to understand the reuse and sharing of the “cited” data behind such links. Such study underscores the necessity of computational analysis, given that some 13,390 articles and 13,447 filtered URLs were ultimately examined.

Chao (2011, p. 7) developed Perl scripts to automatically parse and analyze data-set metadata records held by NASA’s Global Change Master Directory (GCMD). GCMD metadata were first parsed to find “affiliated publications” that were explicitly listed in the records of individual data sets. Those publications were then queried against Google Scholar to find the quantity of, and metadata for, each “secondary publication”—those publications that cited the “affiliated publication.” Such automation was “successful in attaining a wide range of results at a much quicker rate than manual operation,” yet the author acknowledged a number of automation limitations and barriers, including the complexity of parsing multiple nuanced variations of text citations for constituent metadata (e.g., authors, title, journal, and publication date), the incompleteness of metadata overall, as well as the inconsistency of metadata such as titles and date formats. Here, and in many of these studies, the actual software used to obtain the research results was not provided. Furthermore, some of the issues cited have greatly improved since that study, but there does not yet exist a common software platform that abstracts the components of the process to make such a study repeatable, extensible, and generalizable.

Robust computational approaches to citation identification require linguistic analysis, such as those found in Boland, Ritze, Eckert, and Mathiak (2012). In their research to analyze social science publications for mentions of data sets, pattern induction algorithms were developed to explore whether the context (words and characters) around a given data set seed string, typically the name of a study, could be valuable in recognizing the actual mentions of data sets.
Their results indicated that quality results could be obtained when good seeds were manually chosen, even with sparse data. Furthermore, Ritze and Boland (2013) examine an implementation for integrating the discovery of research data and their publications into an open access repository search system. Building on the work of Boland et al., their study examines automating the search, transformation, and linking of research data to publications. End users of a traditional institutional repository (Mannheim University Library) were presented with research data links when searching for publications, thus reducing the burden of using another system manually to find such data sets. The results of this study indicate that though such automation can be achieved, integrating publication metadata with research data metadata presents many challenges, such as domain-specific mappings and synchronization between repositories (e.g., updating new data sets with publications and vice versa).

Although there are many issues around the technical requirements of finding citations, links between citations and publications are often not clear when specific subsets of data sets are referenced or when only the provider of a data set is referenced. Automating the latter scenario is complex, given that data-set names and data provider names may not be unique within or across collections. Automating the former scenario, however, is discussed in Mathiak and Boland (2015). In their work, they automate the discovery of “underspecified” citations. By examining various citation matching scenarios, for example, whether a citation is specifying a superset or subset of an existing data set, the authors propose first creating a hierarchy of linked relationships among data sets and their variants. The relationships, specified as linked data in an ontology, are then used to establish relationships with corresponding publications.

In the biomedical sciences, using automation to determine data-sharing patterns (Piwowar & Chapman, 2010) and to extract data-set citations has been the source of active research (Haeussler, Gerner, & Bergman, 2011). For example, Neveol, Wilbur, and Lu (2011) apply support vector machines to automatically extract data deposition statements from PubMed central articles. By automating deposition patterns, the authors developed methods to explore the use of such data and trends of data production to improve the information technology (IT) infrastructure needs of deposition repositories. Using sentence classification algorithms, the authors report promising results for finding articles with deposition statements (0.81 F-Measure) and also for understanding the deposition patterns within various databases (e.g., GenBank and Protein Data Bank [PDB]). Kafkas et al. study data citation patterns in the context of structural annotation for metadata enrichment (Kafkas, Kim, & McEntyre, 2013; Kafkas, Kim, Pi, & McEntyre, 2015). Applying an automated text-mining pipeline against articles from Europe PubMed Central, the authors find direct citations to biological database accession numbers. The authors suggest that automatically integrating such data with metadata in other common biological databases could lead to better integration points for researchers and might also lead to better metrics for data citation tracking. Further automation of biological database citation has been explored by Yu et al. (2015), in the context of understanding the links between databases mentioned in PubMed Central articles. By exploring the network of databases, their research shows how automation can yield improved comprehension of the similarity between databases as well as their citing articles. Huang, Rose, and Hsu (2015), also using PubMed Central, examined how citations to the PDB have evolved since 2000, specifically analyzing citations to published articles about the PDB, and in-text mentions of PDB-related URLs. They examined citation cascades and disruptions, which are both measures of how citation patterns shift over time. This technique showed how authors who reference the PDB chose to either cite the first article published about the PDB in 2000 or mention the PDB URL, but rarely do both.

A series of studies by Piwowar et al. demonstrate that automated methods can shed light on behavioral trends related to data-set citation and impact; in particular, how the open availability and access of data related to published literature can, in some domains, provide a citation advantage. Piwowar, Day, and Fridsma’s (2007) study of data sharing in 85 cancer clinical trials found that publications with openly shared data were associated with a 69% increase in citations. In a broader analysis that used a corpus of 10,550 publications describing gene expression microarray data, Piwowar and Vision (2013) showed that publications whose underlying data were archived in openly accessible repositories received 9% more citations than similar studies without publicly available data. Of the studies that did archive data, 20% had been reused in another study. This work built on Piwowar’s (2011) earlier findings, which used automated methods to identify 11,603 publications from 2001 to 2009 and their associated data, which were archived at a rate of just under 25%. Using first-order factors analysis and multivariate regression, Piwowar (2011, n.p.) was able to demonstrate that “authors were most likely to share data if they had prior experience sharing or reusing data, if their study was published in an open access journal or a journal with a relatively strong data sharing policy, or if the study was funded by a large number of NIH grants. Authors of studies on cancer and human subjects were least likely to make their data sets available.”

**Software Citation and Attribution Tracing**

Software developed as a part of scientific research is increasingly recognized as a vital part of a reproducible scholarly record (Stodden, 2010). Whereas many metrics have been developed for measuring the reliability and performance of software applications (e.g., Fenton & Bieman, 2014), relatively few have attempted to measure the impact of software used in science. This is problematic because the sustainable development and maintenance of this type of research infrastructure depends, in part, on having reliable documentation about how software is used and having ways...
to measure and clearly communicate that use (Ahalt et al., 2015; Katz, 2014; Katz et al., 2014). Much like research data, the infrastructure and cultural norms that support traditional metric-based approaches to impact, such as citation analysis, are emergent and not well established (Jackson, 2012). As such, studies of software impact have thus far been exploratory and illustrative of an emerging paradigm of computationally intensive work practices.

Qualitative studies in this group focus mainly on behavioral issues related to software development, archiving, and citation. The omission of methodological discussions about collecting publications and possible metrics reflects the uncertainty and lack of standards around software citation in general. Howison and Herbsleb (2011, 2013) have produced a series of studies focused on sociotechnical aspects of software development and use in scientific research settings. Their work highlights the fact that the iterative, ongoing maintenance required of software development is at odds with a traditional reputation economy in science that rewards published (static) findings with clear authorship roles. Studies that have focused on software citation and acknowledgment in published literature include Howison and Bullard (2015) in biology and Weber and Thomer (2014) in bioinformatics. Using a small sample of 90 biology articles, Howison and Bullard found that although software was often mentioned within the text of an article, less than half (44%) of these mentions took the form of a citation, whereas 31% of these mention the software in name only. Weber and Thomer (2014) looked at acknowledgment and authorship practices in bioinformatics as a domain where software development and implementations play a critical role in knowledge production. Drawing on a large corpus of 1,000 bioinformatics publications, they found that only 13% explicitly acknowledged a software package that was used to generate a research finding, whereas many other publications generally acknowledged the use of software platforms, such as Matlab or SPSS, used to conduct data analysis. Similar to the qualitative studies of software use, Weber and Thomer describe how acknowledgment statements in bioinformatics contain many different forms of labor and indebtedness for software use, noting that over half of the articles they reviewed included personal thanks to individuals who helped develop or implement a piece of research software.

Quantitative studies of research software impact are also just beginning to emerge. Howison et al. observe that part of the difficulty in producing reliable measures of software development and use are attributed to the fact that “software runs ‘on top of’ other software components in a layered architecture, effectively hiding the components from users but gaining services and benefitting from those components” (2015, p. 454). These authors create a four-part framework to describe the current state of research software evaluation, noting that such studies can be categorized based on their relationships to funding, development processes of the software itself, use of software within scientific research, and in relation to traditional scientific impact measurements.

A notable early contribution to quantitatively tracing software usage is Thain, Tannenbaum, and Livny (2006). Although not focused on scientific software per se, the authors note that understanding the impact of software often requires sociological methods of interpretation rather than technological solutions, which exist in forms such as measuring downloads of a package or analyzing the number of clients connecting to a server. Thain et al. also note that tracking the distribution of software is likely to overlook nuances in the impact of different versions, and overall such approaches suffer from biases related to the size of the user pool.

Tools for mapping software contribution networks, measuring software usage, and aggregating various metrics from version control systems such as git are in a nascent, but promising, state of development. Notably, Bogart, Howison, and Herbsleb (2015) have released a novel exploratory tool to understand trends in usage, development, and publication outcomes for R packages by combining Github metrics and data submitted by volunteers running a distribution of R on their personal computers. Piwowar and Priem (2015) have more recently released Depsy, a platform that tracks developer contributions to R and Python packages as well as mentions and citations of the software in published literature.

Although these tools offer the potential to perform more robust, wide-ranging studies of research software impact, the empirical studies described earlier have emphasized that the heterogeneity and pervasiveness of research software translates to relatively few straightforward solutions for impact assessment. The difficulty of generating universally applicable software metrics is further evidenced by the focus of these tools on a small subset of the languages (R and Python) used in computationally intensive research settings. Some additional challenges related to software impact assessment and the opportunities for future research around these issues are described at the conclusion of this article.

**Platform and Facility Attribution Tracing**

The agencies and governments directing investments into national-scale and international research infrastructures—such as large-scale, high-performance computing systems or research instruments—need to measure the impact on science from these infrastructures (O’Neil, 2013). However, as with data collections, assessing and tracing the impact of these shared user facilities and platforms presents challenges at several levels: (i) citation behavior—how the facilities and platforms are cited or acknowledged; (ii) methodological difficulties—how the relevant publications can be assembled into collections for analysis; and (iii) metrics compilation—what bibliometric measures are most relevant to the impact of these infrastructures. At the same time, increasing demand from funding agencies for quantitative measures is rapidly making this line of research inquiry a practical and managerial
challenge, with many ad hoc case studies and approaches described online and in the literature.

High-Performance Computing Platforms and Facilities

The issues related to appropriate citation or acknowledgment of data sets or collections also apply to high-performance computing (HPC) platforms and facilities. Publications are generally deemed the best indicator of scientific impact; however, authors may cite or acknowledge the resource or instrument, the operating facility, or either the resource or facility in a given publication. For example, the EarthScope project operates several Earth observing facilities, including the USArray, which supports continental-scale seismological research. These facilities are referenced in a wide variety of ways (Woodward & Simpson, 2013). Furthermore, the precise wording used in a reference may vary substantially from the recommended wording provided by the facility (Patton, Stahl, Potok, & Wells, 2012). These challenges are not limited to HPC platforms and facilities, and most studies raise this inconsistent citation behavior as a challenge that affects their methodological approaches.

Different studies take a variety of manual, automated, or other approaches to collecting the relevant set of publications. The EarthScope management team devoted staff effort to manually assembling the publications that resulted from use of their research instruments. To make the approach tractable, they focused on only 11 key journals and conducted an exhaustive search of those publications (Woodward & Simpson, 2013). Limiting the search to a fixed number of journals may work well for research infrastructures more narrowly focused on a single domain. However, national-scale computing user facilities face perhaps greater challenges because their user communities may span many different scientific domains and may have thousands of users—that is, potential article authors—in a given year. Such facilities often rely on a variety of user self-reporting approaches to assemble relevant publications: through surveys; regular pleas to mailing lists; per-article or per-project notification by users (e.g., see the processes used by the National Center for Atmospheric Research [NCAR], Extreme Science and Engineering Discovery Environment [XSEDE], and nanoHUB.org); or an embedded part of another required process, such as allocation requests (e.g., NCAR, XSEDE). All such self-reporting options suffer from a degree of uncertainty, and little information exists as to how well such approaches collect the complete or correct set of publications. To better understand the effectiveness of such an approach, NCAR recently conducted a statistical assessment of its annual survey of NCAR supercomputer users. This assessment demonstrated that whereas the number of publications collected by the user survey may vary 10% above or below the true count, the survey was largely effective in collecting an accurate number of publications that result from NCAR supercomputer usage. This assessment also showed that the scale of infrastructure usage, measured by the amount of supercomputing time used, does not reliably indicate publication productivity (Rishel, Hart, & Nychka, 2015).

Because of the uncertainties in user-based reporting and effort associated with manual collection efforts, Oak Ridge National Laboratory (ORNL) staff have developed a workflow and set of algorithms for systematically searching through online literature indexes, such as Google Scholar, to find citations and acknowledgments to the use of ORNL supercomputing facilities. Based on a limited evaluation, this computational approach to finding such acknowledgments has proven to perform better (in terms of the number of relevant publications found) than manual searches with the same goal (Patton, Stahl, Hines, Potok, & Wells, 2013; Patton et al., 2012). Because of the challenges in determining whether an author acknowledged correctly, partially, or not at all, support from a facility or platform, alternative approaches have been used to assemble a superset of the relevant publications for analysis. A superset, in this context, refers to a compilation of publications from known facility users or project leads, whether or not a direct line can be drawn from the usage of a facility to any particular article. Wang et al. (2014) used a combined set of user-reported, targeted publications and an automatically collected superset of publications. Their “mashup” of 142,000 publications included all publications reported to the NSF award database by more than 20,000 users with XSEDE allocations, the assumption being that researchers would be more motivated to provide reports to their funding agency than to the facility operators. Though this superset of publications admittedly included publications not related to the XSEDE facilities, they argued that this superset provided an indirect measure of the impact by the XSEDE user community.

Regardless of how the publications are collected, the bibliometric analyses conducted lean heavily toward the simplistic, with the number of publications serving as the primary metric in several (Patton et al., 2012; Rishel et al., 2015; Woodward & Simpson, 2013). The next level of analysis encompasses a set of descriptive statistics about the collected publication sets (Bollen, Fox, & Singhal, 2011; Patton et al., 2013; also see http://nanoHUB.org/citations), such as journal distribution, publications per author or year, citation counts, or keyword frequencies. Several studies do take the next step and investigate the relationships between HPC allocations and more-advanced bibliometric measures such as h-indices and total citation counts for the researchers in their sample. Using h-indices and citation counts on a superset of user publications as measures of impact, Bollen et al. (2011) found moderate positive correlations between both PIs’ h-indices and PIs’ total citation counts with respect to the size of HPC allocations; however, the relationships all have a high degree of scatter. They proposed a preliminary return on investment (ROI) metric of total cites per unit
allocated for all PIs, but do not recommend a specific strategy for maximizing this metric. Wang et al. (2014) are perhaps unique in describing an end-to-end framework for not only collecting publications, but also building a portal to explore the data set and associated metrics and defining and calculating measures of scientific impact. They used similar PI-oriented base metrics (HPC allocations, citation counts, and h-indices) and replicated the findings of positive, although somewhat weaker, correlations among those metrics using the full data set rather than a sampling of PIs. Ultimately, they proposed a similar ROI metric and concluded that “sponsoring a larger number of smaller scale projects could actually produce a higher scientific impact than a smaller number of very large projects” (p. 8).

**Space-Observing Facilities**

Another group of studies used bibliometric analyses to better understand the impact of space facilities on their scientific communities. This body of literature began in the 1980s with separate lines of study by Abt and Trimble, both accomplished astronomers in their own right. They both studied the scientific productivity and impact of astronomical observatories, using numbers of articles published based on the use of observatories and subsequent citations to those articles as the key measures. Abt’s initial studies showed that smaller telescopes were more cost-effective than larger telescopes, when comparing publications produced and citations received with the monetary cost of constructing and operating telescopes of different sizes (Abt, 1980). Abt’s studies also illustrated that high-impact science was coming out of both publicly owned and privately owned observatories (Abt, 1985). These studies were used to argue within policy circles against the closing of smaller telescopes and facilities (Abt & McCray, 1999). However, Leverington (1996) later showed that high-impact science, as measured by citations received, increasingly were produced by studies that used larger telescopes. In a subsequent study, Abt (2012) again found that the largest telescopes do not produce most of the highly cited astronomical articles, but notes that from an economic perspective, astronomical projects should be done on “the smallest telescopes that can yield the required data” (p. 3) to preserve the larger telescopes for studies that truly need them. The relationship between facility size, costs, and resulting impact was also examined by Benn and Sánchez (2001), who found that the relative citation impact of ground-based optical telescopes was approximately proportional to their optical collecting area (e.g., the telescopes’ mirror diameters) and the telescopes’ production costs.

Building on Abt (1985), Trimble (1995) conducted a follow-on study to compare the productivity and impact of different telescopes. She found that little had changed in the decade since Abt’s study; the relative metrics of individual telescopes had remained fairly stable, and larger and smaller facilities were both producing high-impact articles. Trimble (1996) looked at a larger range of factors in relation to telescope productivity and impact, including the size of the telescope’s mirror, the physical location of the telescope, and the budget size and stability of telescope facilities. As expected, facilities with better physical locations and more robust financial support were positively associated with increases in number of articles published and citations received. Trimble, Zaich, and Bosler (2005, 2006), Trimble and Zaich (2006), and Trimble and Ceja (2008) successively updated the Abt (1985) and Trimble (1995) studies, again to examine changes in metrics roughly a decade later. Their studies examined radio, optical, and space-based facilities as separate categories and showed many of the same characteristics related to relative productivities and impacts of different facilities. New findings, however, showed that mid-sized telescopes seemed to be waning in influence in the mid-2000s, given that larger telescopes and space-based facilities (namely, the Hubble Space Telescope) showed increasing productivity and impact metrics. Trimble (2009) performed a retrospective study of the astronomical literature produced between 1960 and 1964 to examine facility metrics for an earlier time period. Trimble (2010), the last study in this series (thus far), performed a comparative analysis of the data from each time period examined in the past studies, namely, articles published in the 1960s, 1990s, and 2000s. The key takeaway from Trimble’s longitudinal analysis is that the number of “dominant” facilities, as measured by articles produced and citations accrued, has stayed relatively consistent across four decades, but that the dominant facilities themselves have changed. She uses this analysis to suggest that the “half life” of the observatories is around 40 years; the term “half life” is undefined in Trimble’s studies, but her usage is better interpreted as ‘the time period in which the facility is productive,’ not the typical usage meaning the time in which half of the articles and citations are produced. Finally, Trimble et al. (2005), Trimble and Ceja (2008), and Trimble (2010) all note that articles that used the Sloan Digital Sky Survey (SDSS) data collections showed large citation and publication rates in comparison to the other observation facilities examined. Abt (2012) provides a similar finding on the growing significance of the reuse of existing data from large-scale observational surveys, such as the SDSS. Zhang, Vogele, and Chen (2011) also evaluated the impact of the SDSS on the astronomy community, though they did not distinguish between articles that used the SDSS data versus those that might have mentioned the SDSS for other reasons.

Another body of astronomical bibliometric studies centers on methodologies for compiling “telescope bibliographies.” These studies compile bibliographies of all articles derived from use of particular telescopes and use those bibliographies to produce productivity and impact metrics. Meylan, Madrid, and Macchetto (2004) performed such an analysis of the productivity and impact of articles resulting from use of the Hubble Space Telescope. They showed how each of the methods of allocating telescope observing time within the Hubble facility had each led to high-impact research output. In a more detailed article, Apai...
et al. (2010) presented a separate compilation and analysis of a Hubble Space Telescope bibliography. Using a full-text search tool called FUSE, which was created specifically as a tool to compile astronomical telescope bibliographies (Erdmann & Grothkopf, 2010), Apai et al. (2010) identified articles that used Hubble data, based on the presence of Hubble instrument keywords (see also Lagerstrom, 2010 for more details). All potentially relevant articles were then examined manually to determine the related instruments, data sets, and Hubble observing programs. A key outcome of this project was an illustration of how the different processes used to allocate observing time on the Hubble Space Telescope showed different results when measured by articles produced and citations accrued. By these metrics, small allocations of observing time had a higher relative productivity rate. Observation programs that received large allocations of observing time, however, were much more likely to generate extremely high-citation articles and highly used archival data sets.

These Hubble studies are representative of a larger community of interest around bibliometric methodologies for telescope facilities. Rots, Winkelman, and Becker (2012) and Accomazzi, Henneken, Erdmann, and Rots (2012) present similar studies of the bibliometric impact of the Chandra X-ray Observatory, another high-profile, space-based astronomical observing platform. White et al. (2010) used the results of these Hubble and Chandra studies to argue for the value of astronomical telescope data archives as a whole. Crabtree (2014) used a number of telescope bibliographies to compare the relative productivity and impact (measured by publications produced and citations accrued), showing that productivity rates were similar for the eight examined telescopes, but that citation impacts were weighted toward specific high-impact articles and authors. Kim (2011) also performed a comparative analysis of telescope productivity and impact based on publicly available telescope bibliographies. Kim’s study suggested that having more telescopes of the same kind at the same location can lead to higher numbers of resulting articles, attributed to management and operation efficiency gains, and the increased availability of telescope instruments for researchers who utilize a particular facility.

The compilation of observatory or telescope bibliographies has become a well-established procedure, largely involving manual or partially manual methods (Grothkopf & Treumann, 2003). Grothkopf and Lagerstrom (2011) published a set of guidelines for creating such bibliographies, and the larger astronomical community subsequently formulated an agreement titled, “best practices for creating an observatory or telescope bibliography” (IAU Working Group Libraries, 2013). This best practice document provides guidelines for how to determine whether an article should be counted as representing scholarly output of the observatory facility, including how to distinguish between different types of articles, for example, articles that present analyses of observatory-generated data versus those that describe the observatory capabilities of the facilities themselves.

**Proxy Metrics: Downloads and Usage Data**

With impetus from funding agencies, ongoing efforts are exploring a variety of quantitative approaches to understanding the impact of research infrastructures. Because of the behavioral and methodological challenges in quantifying scientific impact by traditional bibliometric avenues, many of these studies seek indirect measures that serve as a proxy for bibliometric data, with mixed results. By making simplifying assumptions, the proxy metrics studies tend to bypass all or some of the behavioral and methodological challenges. For example, metrics relating to the performance and usage of supercomputing facilities (Furlani et al., 2013; Hart, 2011) have been used as proxy measurements for scientific impact. For supercomputing resources, usage metrics provide some evidence of impact, the argument being that some usage is known to produce some impact, typically measured as numbers of publications in incomplete collections; therefore, greater utilization should imply greater impact. Further aspects of supercomputing usage have been examined, which may contribute to understanding of impact. How users intend to use the resources (Katz et al., 2011) or what software elements of the environment are used in practice (Hadri, Fahey, Robinson, & Renaud, 2012) may serve as metrics for measuring how well the system provider’s plans and expectations for proposed use cases match the eventual users’ reality.

Apon et al. (2010) approached the impact question more broadly by comparing the existence of HPC resources at a university campus with the campus’s success in competing for federal research funding. Using federal funding data, institutionally compiled bibliometric data, and the Top 500 list,7 a list of the 500 most powerful supercomputers in the world that has been updated twice yearly since 1993, they conducted statistical analyses that showed a significant correlation between consistent HPC investments and institutional research competitiveness. Though a strong statement of the value of HPC research infrastructures in general, the results are difficult to apply to a specific supercomputing system or to a large-scale shared facility with users from many institutions. Bollen et al. (2011), as noted previously, compared the HPC system usage with the overall impact metrics (h-index, etc.) of the users. Again, they showed a positive relationship between usage levels and a user’s scientific productivity overall, but not a direct relationship between the HPC research infrastructure and specific scientific impact.

For data and software collections, download metrics have been used as proxy measures of usage and thus impact from data archives (Jacobs & Worley, 2009). Though download metrics illustrate how data infrastructures are being accessed over time, such metrics do not guarantee a direct mapping to

---

7www.top500.org.
usage or impact (e.g., data might be downloaded but never used) and are contingent on the specific characteristics of their respective repositories, making them difficult to compare across organizations (Ingwersen & Chavan, 2011; Weber et al., 2013). However, at least for open-source software, Wiggins, Howison, and Crowston (2009) present a method for measuring the size of user communities based on downloads of individual software packages. The measure illustrates how the baseline day-to-day level of downloads of a given software package is usually quite small in comparison to download spikes that follow the release of new versions. Wiggins et al. argue that these spikes associated with version releases are a better illustration of the size of the active user base.

In the context of bioinformatics research infrastructures, Huang, Rose, and Hsu (2015) compared website access statistics for the PDB and PDB URL mentions in published biomedical literature. They found that the two metrics were highly correlated. Jonkers, Derrick, Lopez-Illescas, and Van den Besselaar (2014) analyzed the relationships between citations of several different databases hosted by the Expert Protein Analysis Server and two different proxy metrics: (i) usage intensity, measured by the number of web visits to each database, and (ii) in-text mentions, both acknowledgments and other unstructured mentions within articles. In analyzing citation and proxy metrics for a handful of infrastructures, they argue that citation counts consistently underestimated the impact of the research infrastructures being studied. The numbers of unique visitors to the examined databases, for example, showed 10 to 30 times higher values than the citation counts that those databases received. In addition, their article echoed many other studies reviewed earlier in highlighting the challenges involved in assembling relevant collections of publications for analysis.

Synthesis

The breadth of studies discussed in this review illustrates the widespread interest in improving the collection of publications supported by research infrastructures and in developing a better understanding of the impacts of data, software, instruments, and other infrastructural resources that facilitate scientific research. The following sections discuss common features of these studies, notable challenges involved in producing and analyzing impact metrics for research infrastructures, and key open research questions that can guide future studies.

Commonalities

The studies reviewed vary widely in their disciplinary focus and granularity. A number of common features of these studies, however, can be articulated, ranging from the consistency in motivations and metrics gathered to some repeated tools, methods, and findings.

Commonalities Related to the Metrics Gathered and Analyzed

Another analytical commonality among the studies reviewed here is that they almost uniformly use similar base measures of productivity and impact. Productivity is measured as the number of publications produced, and impact is measured as the number of citations that those publications receive. These metrics are thus limited to measures of “impacts” on academic research, and many rely wholly on formally published and professionally indexed publications. These metrics would not show the impacts of research infrastructure in educational or policy settings. For example, the data in the PDB\(^8\) are widely used in the biosciences, but, unanticipated by the designers of the repository, are also widely used by schoolchildren (Bourne, 2012). Similarly, only one study reviewed here, by Kirlew (2011), studied citations that data sets received in patents.

8http://www.rcsb.org/.
Common Evaluative Findings

Looking across the reviewed studies, some common findings emerge. First is the measurable and growing impact of open data on scholarly research. As Piwowar, Day, and Fridsma (2007) have shown, articles that are published based on open data have higher citation rates than those that are not. Other studies presented complementary findings. Major (2011), for example, showed that data sets archived by NASA’s EOS contribute to over one third of the articles published in the WoS Earth Science journals. Belder (2014) likewise estimated that the three National Oceanic and Atmospheric Administration data sets analyzed would place in the top 1% of articles in their domain in terms of citations received, if the data sets were actually cited in articles that used them. Apaï et al. (2010) found similarly high impacts in looking at the use of data from the Hubble Space Telescope, and noted that Hubble contributed from 3% to 13% of the highest impact articles in astronomy each year between 1995 and 2010. A number of telescope impact studies previously reviewed also noted the growth in astronomical literature that derives from use of data from the SDSS.

Another finding that crossed multiple studies related to the relative value of large and small allocations of research time at observational or computing facilities. Bollen et al. (2011) and Wang et al. (2014) noted how smaller allocations of supercomputing time showed higher relative productivity rates in terms of publications produced. Apai et al. (2010) described a similar phenomenon when analyzing allocations of observing time on Hubble, namely, that smaller allocations produce proportionally more publications. In addition, Abt’s series of studies notes that smaller telescopes are highly productive for their relative costs. None of these studies discount the value of large allocations or facilities—many note that qualitatively different science may require large-scale use of larger facilities or instruments—but they do provide fodder for policy and management discussions about how to allocate resources to larger versus smaller projects within specific facilities or research domains.

Commonalities Related to the Methodologies and Tools Used for Data Collection

Another aspect that unites these studies is their predilection for methodological novelty in finding, retrieving, and structuring relevant trace data for analysis. Although many use the same kinds of bibliographical data, the tools, scripts, or databases developed are almost always custom made and rarely shared among different groups. Reproducibly gathering the same kinds of trace data and uniformly processing these traces for analysis is critical to performing comparative analysis across sites, domains of study, and types of research infrastructures. Academic publishers have been especially buoyed by the standardization work in altmetrics (Konkkel, Piwowar, & Priem, 2014) in recent years, and many of the same efforts at creating cross-platform, open-source tools will be needed for research infrastructure analysis to mature in both its sophistication of analysis and its reliability for use in science and technology policy making.

Finally, these various studies of research infrastructures are linked by the frustrations with what might be possible given the open availability of trace data for analysis. For instance, the inability to compare usage statistics across different data repositories or different high-performance computing centers greatly prohibits the development of usage-based studies. This is attributed to different centers collecting different usage metrics, and to different software logging different user activities. In the latter case, much of the software that is responsible for logging the downloads, search and browse sessions, or even basic web traffic statistics in these types of research infrastructures could be made uniform through a closer coordination of data and computing centers.

From a frustration around the availability of trace data comes a great potential for private industry collaboration. This is especially salient for bibliographic data found in indices such as Google Scholar that contain much gray literature, white papers, and patent information—all of which may prove to be valuable sources of research infrastructure references, citations, and acknowledgments. (On the value of gray literature for evaluating research findings, see McAuley, Tugwell, & Moher, 2000.)

Notable Challenges and Barriers to Progress

The challenges that investigators face when attempting to compile and analyze impact metrics for research infrastructures range from data collection difficulties and methodological imprecision to a lack of consistent conceptual underpinnings. The main overarching challenge is that the referencing and citing behavior of authors who have used research infrastructures is extremely inconsistent. This inconsistency in referencing practices was noted in previous sections and is a long-lasting problem. The problem is not unique to studies of research infrastructures, as Cronin (1984, 1995) and others have discussed; citation and acknowledgment behavior is inherently imprecise and idiosyncratic. These referencing inconsistencies, however, are amplified considerably for references to research infrastructures because the academic institutions that structure referencing behavior, including citation norms, standards, and tools, are poorly developed. As one example, the distinction between citing the providers of the resource versus citing resources themselves is often glossed over by authors. Authors might provide an acknowledgment to a research organization as a whole in lieu of citing or acknowledging specific services that they used, such as computing or data facilities.

These inconsistencies in referencing behavior lead directly to methodological challenges for the bibliometric analyst. The tools and services for identifying and collecting the relevant publications do not lend themselves to automated data collection. Automated approaches to mining and tracking persistent identifiers to scientific resources face a number of challenges, starting with the lack of full-text article availability for text
mining. Because many references to research infrastructures are given in the body text of documents, searching metadata will not find them consistently. Mining the full text of publications through ad hoc general Internet search engines and web page screen-scraping tools is being investigated (see, e.g., Patton et al., 2012, 2013), but these approaches rely on technologies and platforms that are in flux. For example, extracting information from an ever-shifting bed of article indexer websites (such as Google Scholar), journal websites, and article display formats requires sophisticated parsing of a multitude of formats such as HTML, PDF, JavaScript, etc. Not only are these methods time-consuming to develop, and in some cases may potentially violate the licensing agreements of the publisher, but also they are brittle when new formats are developed or unexpected structural changes are made to existing services. Some of these issues can be offset through structured open application programming interfaces (APIs), yet such open APIs are not widely available for many scientific journals or article indexing services. The services that do provide such APIs are the basis for many of the studies reviewed earlier, such as the APIs for the National Institutes of Health’s (NIH) PubMed Central and the Astronomical Data Service. For researchers studying research infrastructures outside of the medical and astronomical domains, however, manual data collection procedures are necessary. As illustrated by the large number of studies previously reviewed, the case study approach can be quite effective in providing impact metrics for small sets of resources. This approach, however, is not scalable to comparative studies of large numbers of infrastructures.

Finally, conceptual challenges impede the analysis and evaluation of productivity and impact metrics for research infrastructures. Some of these conceptual issues echo longstanding challenges in the bibliometric and informetric literature. Studies of research infrastructure impact metrics show a lack of consistently used and applied metrics that are equivalent across projects and infrastructure types. Numbers of publications produced and citations received are highly variable across domains and infrastructure types, making it difficult to directly compare data collections or facilities. In addition, understanding the validity and applicability of productivity and impact metrics is not trivial. Such metrics could be gamed like any other metric (Davis, 2011; Smeyers & Burbules, 2011) or used for policy decisions without an understanding of the limits inherent to bibliometric techniques (Gläser & Laudel, 2007). Often, for example, publication and citation rates for individual infrastructures are presented without comparison to control groups of equivalent publications that might provide context for the overall population of published studies. Determining what such a control group should be is itself a contingent decision. In their separate series of studies of astronomical observational facilities, Abt and Trimble compared a number of facilities with one another. Other studies compare publications that used a given research infrastructure with articles randomly chosen from journals of the same domain (see Belter, 2014; Major, 2011). Even when looking at one particular resource type, such as data repositories, the metrics of productivity and impact might be highly contingent on the characteristics of each repository, as well as the communities that they were designed to serve (Weber et al., 2013). Without a better conceptual understanding of how such metrics will vary across and within analytical categories, or of what the gaming of such metrics might entail, it will be hard to develop general understanding of the impact of research infrastructures.

Open Research Questions

When collected systematically and analyzed carefully, productivity and impact metrics provide significant benefits to their respective stakeholder and user communities. However, the research frameworks to ask questions about how to methodically and consistently analyze the scientific impact of data sets, facilities, software, instruments, and so forth are lacking. This section outlines some open research questions that will be important in moving these research frameworks forward.

Do Citations and References to Research Infrastructures Increase When Persistent Identifiers Are Assigned? What Are the Characteristics of Such Citations and References? Do Users Create More Consistent Citations and References When Persistent Identifiers Are Assigned to Resources?

The first set of questions revolves around better characterizing the benefits of persistent identifier assignment to research infrastructures. Because assigning and using persistent identifiers (e.g., DOIs) for scientific resources such as instruments, data, and software is a relatively new development, very few assessments have been conducted that systematically examine the effects of such identifiers assigned to these resources. It has not been confirmed, for example, that assigning persistent identifiers to data or software leads to an increase in the number of citations to the identified resources, or an increase in the consistency with which such resources are cited or acknowledged by researchers who use them. Similarly, it has not been confirmed that assigning persistent identifiers to scientific resources leads to an increase in the citation or acknowledgment of those resources in comparison to other resources of the same kind that have not been assigned persistent identifiers. A related open question is what impact the assignment of persistent identifiers will have on the discoverability and retrievability of these resources over time. For instance, Klein et al. (2014) have shown that one in five scholarly articles suffer from “reference rot,” where the web context around these publications cannot be revisited. Broader use of DOIs for research infrastructures is intended to slow the pace of this content drift by decoupling the reference identifier from the actual location of the resource being referenced.
What Computational Platforms and Algorithms Are Needed to Expand the Capacity and Success of Research Related to Scientific Impact of Research Infrastructures?

The second area of open research relates to the tools and methodological approaches that are needed to support research infrastructure impact studies. Existing citation indices are not sufficient to support such studies themselves. Furthermore, as mentioned earlier, openly accessible APIs that would support automated analysis of publications do not exist for the majority of scholarly publishers or indexes that contain valuable bibliographical data. Developing the tools and capabilities to perform robust and repeatable impact studies is essential for the analysis of research infrastructures. Platforms that might support the reliable comparison of research infrastructures, for instance, through usage-based analysis, are also needed. Recent funding initiatives such as the NIH’s Data Discovery Index and further plans for a Software Discovery Index represent an infrastructural evolution for how data and software are formally described and accessed by end users. Along with these new modes of access, discovery, and use come the opportunity to intervene in, and improve upon, the evaluation of the types of metrics described throughout this article.

What Metrics Can Be Commonly or Practically Used, Have Validity, and Can Be Commonly Reported for Research Infrastructures? Can New Measuring Schemes (such as “Transitive Credit” Presented by Katz, 2014) Lead to More-Nuanced Views of the Impact of Individuals and Particular Research Infrastructures?

Another key topic of research is the formalization of impact metrics that can facilitate comparative studies within and across research areas and infrastructure types. Different metrics enable different kinds of analyses, comparisons, and findings. Bollen, Van de Sompel, Hagberg, and Chute (2009), for example, showed how metrics of “usage” (as measured by article downloads) cluster together differently than metrics of citation, indicating that the two sets of measures lead to different conclusions about the relative prestige and popularity of individual publications and authors. Research infrastructures present additional complexity to this diversity of potential measures. Data and software repositories might analyze download size, frequency, and diversity. Research facilities often have other specific usage metrics, such as supercomputer or telescope allocation time. Determining effective combinations of these metrics is an ongoing research task. This also raises the question of how the collaborative nature of infrastructural resources—collectively produced and often maintained, managed, and sustained by large teams—can be effectively described at an institutional, rather than individual, level. Many of the complications in identifying and hence rewarding the various creators of a research infrastructure also complicate the use of traditional bibliometric and scientometric evaluative techniques. Even in fields like high-energy physics, where single-publication author lists may balloon toward the thousands, the accurate accounting of contributions to a research publication are imprecise at best. The need to recognize collective groups of people and institutions is becoming an imperative for metrics-based impact studies. In turn, this requires innovation in the distribution and assignment of credit around these types of resources.

What Can Bibliometric Studies of Research Infrastructures Contribute to Theoretical Knowledge of Citation Purposes?

The purposes of citations and acknowledgments to research infrastructures do not fall cleanly into either the “citation as reward” or “citation as persuasion” categories that have been discussed for traditional citations (Davis, 2009; Luukkonen, 1997; Nicolaisen, 2007). Citing a data set, for example, indicates the value of the data, thereby providing a reward to the data provider; in many cases, the citation serves as a quid pro quo for the data provider’s support, provided at no direct financial cost to the researcher. But citing a data set might also be a persuasive statement, illustrating how data of sufficiently high visibility or quality underlie a particular scientific result. Citations of software, observational platforms, or computational facilities likely show similar mixes of purposes. The studies reviewed here rarely touch on these theoretical issues of citation purposes, because the infrastructure managers that often perform these studies are more focused on metrics gathering for practical arguments about infrastructure value. For instance, looking at successive “release articles” related to a specific oceanographic and atmospheric data set, Weber (2015) and Weber, Mayernik, and Worley (2014) showed how around 80% of the citations to the release articles were data-usage related. This example illustrates how some published articles can be highly cited as proxies for data sets.

What Is the Tractability of Citations as the Appropriate Means for Assessing Impact for Research Infrastructures?

Conceptual questions are emerging about the tractability and practicality of collecting impact metrics for research infrastructures. This article limits the study of research infrastructures to studies focused on data, software, and facilities. All of these infrastructures are themselves complex and could be analyzed at finer resolutions. How far down the research chain should metrics be gathered? For example, when looking at software citations, should researchers be asked to cite all tools that they use, including general-purpose tools such as MatLab and Excel? Should these details extend down to the operating system? The repeatability and reproducibility of some computational processes do have dependencies on these lowest levels of the technological chain (Easterbrook, 2014). Similar questions exist in the biomedical fields, where there is interest in increasing the identification and traceability of laboratory and experimental resources, such as antibodies, tissue samples, and chemical reagents (Bandrowski et al., 2015; Singh Chawla, 2015). The notion of scientific reproducibility with respect to
research infrastructures and resources represents a related, but separate, avenue of investigation (see, e.g., Baker et al., 2014; James et al., 2014).

Opposite this trend toward defining the appropriate citable unit for research infrastructures, particularly software, is the creation of bundles or containers of the infrastructure components that support a research finding: Jupyter notebooks, Docker containers, and virtual machines are examples of innovative technologies that combine data, software, and computing environments for the sake of reproducibility. Trying to determine how individual components within these bundles of scholarly objects are teased apart, identified, referenced, and ultimately counted for reward seems at odds with their intended purpose of creating authoritative dissemination packages.

Scholarly impact analysis, especially as it relates to research infrastructures, must coevolve with the entities being measured, but whether one leads or follows the other remains a substantial open question. Infrastructures change over time: Data collections grow and change scope, software tools come and go, and facilities modify their services and operating procedures. Indeed, designing for flexibility and change is one of the main approaches to sustainable infrastructure development, though such flexibility may need to respond to a variety of factors (Ribes & Polk, 2014). The use of citation metrics in studies of specific infrastructures tend to gloss over temporal factors. How should “authors” be designated for long-lived resources, such as software packages, that may have had contributions by many people over different time ranges? Fixing an author list for such resources may reduce the incentive for new individuals to contribute to their ongoing management and improvement (Howison & Herbsleb, 2013). Studies that investigate infrastructural evolution over time tend to be qualitative in nature, but show how changes to research infrastructures have affected the pursuit of science itself (Pollock & Williams, 2010; Ribes & Polk, 2015). Quantitative metrics that account for temporal evolutions would provide a way of understanding such changes.

Conclusion

The need to understand the impact of investments in research infrastructure and validate the benefits of such services is common to every organization, regardless of the type of infrastructure. As the range of work reviewed in this article illustrates, the importance of having quantitative metrics for tracing the impact of research infrastructures is widely recognized. Increasingly, research funders and research organizations are recognizing that traditional assessments of research impact have missed broad swaths of important activities, including the benefits associated with the collection, management, and preservation of digital resources, such as data and software, and the provision of research facilities and services, such as computational facilities and observational platforms. Metrics are desired for understanding the scientific, educational, and economic impacts of these infrastructures on their stakeholder communities (Beagrie & Houghton, 2014; Ludvig, 2012). This review informs these broader efforts to understand how to assess the use of such infrastructures.

The increasing interest in these developments presents opportunities for bibliometric experts to provide methodological expertise and deep understanding of publication and citation networks. Developing solutions to the open research questions discussed earlier will require the knowledge and skills of the academic bibliometric community. Research infrastructure managers have extensive knowledge of their user communities, but are often limited to simple methodological approaches to collecting and counting impact metrics. Vendors of bibliometric services also have a significant opportunity to create tools and services that would simplify the process of conducting research infrastructure impact studies. The challenges that exist in tracing attribution to scientific infrastructures through the research literature are considerable, and no one tool or method is likely to provide an overarching solution. The better the problems impeding progress in this area can be bounded, the better solutions can be evaluated, tools can be shared, and the likelihood for generalizable solutions to emerge increases.

Acknowledgments

The writing of this article was supported by the NSF, grant no. 1448480, “EAGER: Tracing the Use of Research Resources Using Persistent Citable Identifiers,” P.I. Matthew S. Mayernik, Co-PI Keith Maull. Nicholas Weber was supported by the RDA Data Share fellowship with funding from the Alfred P. Sloan Foundation (grant no. G-2014-13746). The National Center for Atmospheric Research is sponsored by the NSF. Thanks to Matt Ramey, Melissa Rishel, and the two anonymous reviewers for comments on earlier versions of the manuscript.

References

Abt, H.A. (1980). The cost-effectiveness in terms of publications and citations of various optical telescopes at the Kitt Peak National Observatory. Publications of the Astronomical Society of the Pacific, 92(547), 249–254. http://doi.org/10.1086/130658
Abt, H.A. (1985). An assessment of research done at the national optical observatories. Publications of the Astronomical Society of the Pacific, 97(597), 1050–1052. http://www.jstor.org/stable/40678593
Abt, H.A. (2012). Scientific efficiency of ground-based telescopes. Astronomical Journal, 144(4), 91. http://doi.org/10.1088/0004-6256/144/4/91
Abt, H.A., & McCray, P. (1999). Interview of Helmut Abt by Patrick McCray on 1999 October 28 and 29. American Institute of Physics, College Park, MD: Niels Bohr Library & Archives. Retrieved from www.aip.org/history-programs/niels-bohr-library/oral-histories/23364
Accomazzi, A., Henneken, E., Erdmann, C., & Rots, A. (2012). Telescope bibliographies: An essential component of archival data management and operations. In A.B. Peck, R.L. Seaman, & F. Comeron (Eds.), Proceedings of SPIE 8448, Observatory Operations: Strategies, Processes, and Systems IV, 84480K. Retrieved from http://doi.org/10.1117/12.927262
Ahalt, S., Carsey, T., Couch, A., Hooper, R., Ibanez, L., Idaszak, R., . . . & Robinson, E. (2015). NSF Workshop on Supporting Scientific
Discovery through Norms and Practices for Software and Data Citation and Attribution. Report from the National Science Foundation-and Sloan Foundation-funded workshop held January 29–30, 2015 in Arlington, VA. https://softwaredat citation.org/Workshop%20Report/SoftwareDataCitation_workshop_report_2015_April_20_with_logod.pdf

Alsheikh-Ali, A.A., Qureshi, W., Al-Mallah, M.H., & Ioannidis, J.P.A. (2011). Public availability of published research data in high-impact journals. PLoS One, 6(9), e24357. http://doi.org/10.1371/journal.pone.0024357

Apai, D., Lagerstrom, J., Reid, I.N., Levy, K.L., Fraser, E., Nota, A., & Henneken, E. (2010). Lessons from a high-impact observatory: The Hubble Space Telescope’s Science Productivity between 1998 and 2008. Publications of the Astronomical Society of the Pacific, 122(893), 808–826. http://doi.org/10.1086/656851

Baker, A., Xu, H., Dennis, J.M., Levy, M.N., Nychka, D., Mickelson, S., … Wegener, A. (2014). A methodology for evaluating the impact of data compression on climate simulation data. Proceedings of the 23rd International Symposium on High-Performance Parallel and Distributed Computing (pp. 203–214). June 23–27, Vancouver, British Columbia, Canada. http://doi.org/10.1145/2600212.2600217

Bello, C.W. (2014). Measuring the value of research data: A citation analysis of oceanographic data sets. PLoS One, 9(3), e92590. http://doi.org/10.1371/journal.pone.0092590

Benn, C.R., & Sánchez, S.F. (2001). Scientific impact of large telescopes. Publications of the Astronomical Society of the Pacific, 113(781), 385–396. http://doi.org/10.1086/319125

Borgman, C., Howison, J., & Herbsleb, J.D. (2015). Mapping the network of scientific software. Washington, DC: The Science of Team Science.

Boland, K., Ritze, D., Eckert, K., & Mathiak, B. (2012). Identifying ODIs names to scientific data and a five year history of data-cite. D-Lib Magazine, 21(1–2). http://doi.org/10.1045/january2015-brase

Bourne, P. (2012). Towards data attribution and citation in the life sciences. In P.E. Uhlir (Ed.) (Rapporteur), For attribution–developing data attribution and citation practices and standards: Summary of an International Workshop (pp. 43–48). Washington, DC: The National Academies Press. http://www.nap.edu/catalog.php?record_id=13564

Brase, J., Sens, I., & Lautenschlager, M. (2015). The tenth anniversary of assigning DOI names to scientific data and a five year history of data-cite. D-Lib Magazine, 21(1–2). http://doi.org/10.1045/january2015-brase

Branco, T.C. (2011). Disciplinary reach: Investigating the impact of data-set reuse in the earth sciences. Proceedings of the American Society for Information Science and Technology, 48, 1–8. http://doi.org/10.1002/meet.2011.14504801125

CODATA-ICSTI Task Group on Data Citation Standards and Practices. (2013). Out of cite, out of mind: The current state of practice, policy, and technology for the citation of data. Data Science Journal, 12, CIDCR1–CIDCR75. http://doi.org/10.2481/dsj.OSOM13-043

Coppen, A. (2013). Finding science and engineering specific data set usage or funding acknowledgements. Issues in Science and Technology Librarianship, 73. http://doi.org/10.5062/F4CV4FP0

Costas, R., Meijer, I., Zahedi, Z., & Wouters, P. (2013). The Value of research data—Metrics for datasets from a cultural and technical point of view. A Knowledge Exchange Report. Retrieved from http://www. knowledge-exchange.info/datametrics

Costello, M.J. (2009). Motivating online publication of data. BioScience, 59(5), 418–427. http://www.jstor.org/stable/10.1525/bio.2009.59.5.9

Crabtree, D.R. (2014). A bibliometric analysis of observatory publications for the period 2008–2012. In A.B. Peck, C.R. Benn, & R.L. Seaman (Eds.), Proceedings of SPIE 9149. Observatory Operations: Strategies, Processes, and Systems V, 91490A. http://doi.org/10.1117/12.2054058

Cronin, B. (1984). The citation process: The role and significance of citations in scientific communication. London: Taylor Graham.

Cronin, B. (1995). The scholar’s courtesy: The role of acknowledgement in the primary communication process. London: Taylor Graham.

Cronin, B. (2001). Acknowledgement trends in the research literature of information science. Journal of Documentation, 57(3), 427–433. http://doi.org/10.1108/EUM000000007089

Cronin, B., Shaw, D., & La Barre, K. (2004). Visible, less visible, and invisible work: Patterns of collaboration in twentieth century chemistry. Journal of the American Society for Information Science and Technology, 52(2), 160–168. http://doi.org/10.1002/asi.10353

Data Citation Synthesis Group. (2014). Joint declaration of data citation principles. In Martone, M. (Ed.), San Diego, CA: FORCE11. https://www.force11.org/datacitation/

Davis, P.M. (2009). Reward or persuasion? The battle to define the meaning of a citation. Learned Publishing, 22(1), 5–11. http://doi.org/10.1087/095351510X137812

Davis, P. (2011). Gaming the impact factor puts journal in time-out. The Scholarly Kitchen, Oct 17, 2011. Retrieved from http://scholarly kitchen.sspnet.org/2011/10/gaming-the-impact-factor-puts-journal-in-time-out/

Dodd, S.A. (1979). Bibliographic references for numeric social science data files: Suggested guidelines. Journal of the American Society for Information Science, 30(2), 77–82. http://doi.org/10.1002/asi.4630300203

Duerr, R., Downs, R., Tilmes, C., Barkstrom, B., Lenhardt, W., Glassy, J., … Slaughter, P. (2011). On the utility of identification schemes for digital earth science data: An assessment and recommendations. Earth Science Informatics, 4(3), 139–160. http://doi.org/10.1007/s12145-011-0083-6

Easterbrook, S.M. (2014). Open code for open science? Nature Geoscience, 7, 779–781. http://doi.org/10.1038/ngeo2283

Erdding, C., & Grothkopf, U. (2010). Next generation bibliometrics and the evolution of the ESO telescope bibliography. In E. Isaksson, J. Lagerstrom, A. HOLL, & N. Bawdekar (Eds.), Library and information services in astronomy VI: 21st century astronomy librarianship, from new ideas to action (pp. 81–88). San Francisco, CA: Astronomical Society of the Pacific.
