Parsons, MA, et al. 2019. The History and Future of Data Citation in Practice. *Data Science Journal,* 18: 52, pp. 1–10. DOI: https://doi.org/10.5334/dsj-2019-052

REVIEW

The History and Future of Data Citation in Practice

Mark A. Parsons[1], Ruth E. Duerr[2] and Matthew B. Jones[3]

[1] Rensselaer Polytechnic Institute (RPI), US

[2] Ronin Institute, US

[3] University of California Santa Barbara, US

Corresponding author: Mark A. Parsons (parsom3@rpi.edu)

In this review, we adopt the definition that 'Data citation is a reference to data for the purpose of credit attribution and facilitation of access to the data' (TGDCSP 2013: CIDCR6). Furthermore, access should be enabled for both humans and machines (DCSG 2014). We use this to discuss how data citation has evolved over the last couple of decades and to highlight issues that need more research and attention.

Data citation is not a new concept, but it has changed and evolved considerably since the beginning of the digital age. Basic practice is now established and slowly but increasingly being implemented. Nonetheless, critical issues remain. These issues are primarily because we try to address multiple human and computational concerns with a system originally designed in a non-digital world for more limited use cases. The community is beginning to challenge past assumptions, separate the multiple concerns (credit, access, reference, provenance, impact, etc.), and apply different approaches for different use cases.

Keywords: data citation; FAIR; credit; access; impact; micro-citation; persistent identifiers

I. Introduction

Data citation helps make data sharing both more FAIR — findable, accessible, interoperable, and reusable (Wilkinson et al. 2016) — and fair. Citation helps make data more findable and accessible through current scholarly communication systems. It can aid interoperability through precise reference to data and associated services and aid reusability by providing some context of how data have been created and used. Citation also helps credit the intellectual effort necessary to create a good data set and provides accountability for that data. It recognizes important scientific contributions beyond the written publication and, therefore, makes things fairer for everyone involved in doing good science.

Data citation is not a new concept. It rests on a fundamental principle of the scientific method that demands recognized, verifiable, and credited evidence behind an assertion. Traditionally, this was done within the literature through citation of materials often held in special library collections, such as field or lab logs, printed books, monographs, and maps (Downs et al. 2015). If the data were small enough, they were simply included directly in the publication. Some fields, such as astronomy, had whole journals devoted to publishing data. The digital age and the corresponding growth in the volume and complexity of data changed all that.

In this review, we adopt the definition that 'Data citation is a reference to data for the purpose of credit attribution and facilitation of access to the data (TGDCSP 2013: CIDCR6). Furthermore, access should be enabled for both humans and machines (DCSG 2014). We use this to discuss how data citation has evolved over the last couple of decades and to highlight issues that need more research and attention.

Early work illustrates the desire to build from existing systems and culture (e.g., Altman & King 2007). Silvello (2018) provides an excellent, extensive review of the motivations, principles, and high-level practices of data citation. Borgman (2016) provides a similar review and illustrates the disconnect between bibliographic citation principles and data citation.

We build on this work by focusing on how the two concerns, credit and access, and the two audiences, humans and machines, can create tensions in how data citation is conceived and implemented. We review

relevant literature and current activities while also drawing from our own experience working as data professionals managing data, systems, and communities for decades. We come from the perspective of observational science, where data record phenomena as they occur and cannot be repeated. This makes precise citation more necessary and more challenging. Our experience is primarily in Earth, environmental, and space science, but we believe our recommendations apply broadly.

II. History

The modern concept of data citation emerged in the late 1990s. For example, one of the authors was involved in an effort at that time, where NASA's Earth science archives (the DAACs) agreed to a common approach to data citation. The US Geological Survey also proposed guidelines (Berquist Jr. 1999). But neither of these approaches were broadly adopted even within NASA and USGS. Part of the issue was that journals were still developing standards on how to cite electronic resources in general. In the analog era, the intangible concepts of an article were manifest in a physical object, but as we moved into the digital age, we needed "to think not of one space (the physical, paper space) but of three 'spaces' in which 'the same' articles appear:

- Information space = the work as intangible entity (ideas)
- Cyberspace = digital manifestation (electronic, made of bits)
- 'Paper space' = physical manifestation (cellulose and ink, made of atoms)" (Paskin 2000: 2).

This added more complexity to the concept of citation, and the problem was exacerbated by the impermanence of web references (e.g., Lawrence et al. 2001). The Digital Object Identifier (DOI) first emerged in the year 2000,[1] even though the underlying Handle system was almost as old as the web (Kahn & Wilensky 1995). Overall, in those early digital years, data were rarely cited, and if they were, the mechanisms were erratic and inconsistent.

From the mid-2000s, there was a growing consensus on the use of registered, resolvable, authority-based persistent identifiers (PIDs) — not only for papers but also other digital artifacts, notably data. DOIs emerged as the PID of choice for many publishers and data repositories, but there are other popular choices, such as Archival Resource Keys (ARKs)[2] or compact URIs (CURIEs) resolved through metaresolvers like Name to Thing[3] or identifiers.org. Indeed, local identifiers (accession numbers) have been used for centuries for internal management, and even external referencing, especially for biocollections. Moreover, digital entities (e.g., computer files), physical entities (e.g., rock samples), living things (e.g., wildlife), and descriptive entities (e.g., mitosis) have different requirements for identifiers (Guralnick et al. 2015). McMurry et al. (2017) provide a good contemporary review of identifiers and how to use them, and the RDA Data Fabric Interest Group has developed a set of assertions about the nature, creation and implementation of PIDs (Wittenburg et al. 2017).

At the same time, there was an emerging argument that data should be 'published' in a manner akin to scientific literature and cited accordingly (Callaghan et al. 2009; Costello 2009; Klump et al. 2006; Lawrence et al. 2011). Data started getting DOIs in 2004 through a pilot project in Germany, and DataCite was established in December 2009, with the explicit global mission of minting DOIs for data (Klump et al. 2015). Even as the 'publication' paradigm for data was questioned (Parsons & Fox 2013; Schopf 2012), citation was broadly supported by the library and data management community. This support culminated with the broadly endorsed *Joint Declaration of Data Citation Principles* in 2014, which defined the core purposes and general practices of data citation (DCSG 2014). But this was an agreement of the library, data, and information science communities. The research community remained largely unaware, and studies have revealed that data citation remains an infrequent and inconsistent process (Howison & Bullard 2016; Mooney & Newton 2012; Mayernik et al. 2016; Silvello 2018).

Part of the reason for infrequent and inconsistent data citation is that, from the researcher's point of view, making data FAIR implies sharing and reuse and therefore effort by the researcher. Based on our collective experience managing multiple data archives and networks, large scale community data, like satellite imagery, may be reused a lot, but much data from more-localized, research collections may never be shared or reused until the broad community recognizes that aggregating these data globally is necessary

[1] DOI Factsheet at https://www.doi.org/factsheets/DOIKeyFacts.html.
[2] https://n2t.net/e/ark_ids.html.
[3] https://n2t.net.

to make further progress (Parsons et al. 2008; Baker & Yarmey 2009). This takes time and effort. For example, starting in 1995 with the creation of the National Center for Ecological Analysis and Synthesis (NCEAS) (Hackett et al. 2008) and subsequent synthesis centers, sharing and reuse through synthesis became an established norm for disciplines like ecology, evolution, and socio-ecology. Nonetheless, only recently has data citation been common in synthesis papers. It appears a culture of data citation must be preceded by a culture of sharing and reuse; one that values reproducibility, transparency, and credit in practice (Stuart 2017).

III. Current Activity and Issues

In the last few years, we have seen much activity to promote data citation and to define specific guidelines for both data and software citation. The Digital Curation Centre and Earth Science Information Partners (ESIP) have updated their respective, long-standing data citation guidelines (Ball & Duke 2015; EDPSC 2019). The Research Data Alliance (RDA) produced a Recommendation on citing specific subsets of very dynamic data (Rauber et al. 2015). Publishers and repositories are coming together on common citation practices (Cousijn et al. 2018; Fenner et al. 2019). Force11 and ESIP have collaborated on software citation principles and guidelines (ESSCC 2019; Katz & Chue Hong 2018; Smith et al. 2016).

We are optimistic that data (and software) citation is emerging as a norm for observational science. We are especially encouraged by the recent project led by the American Geophysical Union and others on 'Enabling FAIR Data' which has led to many publishers now requiring data citation in their author guidelines (Stall et al. 2018). The new Transparency and Openness Promotion statement shows similar commitment beyond Earth, environmental, and space sciences (Aalbersberg et al. 2018). It appears we are approaching the critical mass for a broad behavioral shift. Nonetheless, multiple issues remain.

Many of the issues are rooted in the fact that we are taking a concept implemented for physically printed literature and human beings and trying to use it to address multiple concerns for both humans and machines. We awkwardly try to have the digital space match the 'paper' space, as Paskin (2000) put it.

A. Specific and verifiable citation

One issue of data citation practice is citing precise subsets of versioned data. This is necessary to meet the 'specific and verifiable' principle of the *Joint Declaration* and is fundamental to the reproducibility use case for citation. Of course, the simplest, logical approach is to assign a new PID if there is any change in the data set, but this can become unwieldy with large, dynamic data such as data streaming from a remote instrument undergoing multiple calibrations and corrections. Furthermore, repositories can have very different approaches to versioning their data and how they recommend citing different versions. They also package data and assign PIDs at very different levels of granularity.

We find the RDA Recommendation on Data Citation of Evolving Data (Rauber et al. 2015) the best approach to citing specific subsets of very dynamic data. The basic idea is to assign and maintain a PID for a specific, time-stamped query of a data set, as well as a PID for the data set as a whole. This means the repository must continue to resolve the query PID and maintain or migrate the technology necessary to resolve the actual query within the data set. The RDA Data Citation WG has conducted multiple implementation workshops and has reports from repositories adopting this approach every six months at RDA Plenaries. The approach is getting broader adoption, but it is not at all a norm across repositories. Many simply do not have the capacity to implement it yet; sustaining these citations means sustaining query systems not just data; and maintaining access to these cited queries through technology cycles can be quite challenging (Stockhause & Lautenschlager 2017). Note, this approach provides reference and access to a precise subset, but it does not necessarily address specific credit concerns for that subset, such as when different authors contribute to a larger collection.

There are other approaches for citing dynamic data recommended by DCC, ESIP, and DataVerse (Ball & Duke 2015; Crosas 2014; EDPSC 2019). These include capturing time slices or snapshots of an ongoing time series, having different PIDs for the data set concept and specific versions or instances, or simply establishing careful documentation practices. These approaches are incomplete in that they tend to be more appropriate for relatively static data and often require human interpretation. For example, it is not practical to continually mint new PIDs for a data set that may update every six seconds (typical for many automated meteorological stations). Similarly, some repositories do not find it appropriate to mint new PIDs for minor changes to a data set (e.g., a typo in the documentation) because it can unnecessarily complicate tracking the use of a data set. They rely on the user to apply their judgement on what is a meaningful change for their application. Computers cannot exercise such judgement.

B. What to cite

Another issue is deciding what constitutes a first-class object in scholarly discourse — the 'importance' principle. Data are only truly useful if they are accompanied with detailed documentation about how they were collected, their uncertainties, and relevant applications. Multiple data journals have emerged to provide 'peer-review' of data and documentation and to publish 'data papers' that provide recognizable credit for data authors or creators. They have different approaches, however, on what is to be cited—the document, the data, or both. Often the paper and the associated data set have different authors. The papers also differ in how they structure the information for humans and machines.

Some organizations are now developing well-structured, machine-readable 'publications' that provide the best services of both publishers and data repositories. A good example is the Whole Tale project which is a collaboration among repositories (e.g., DataOne, Globus, DataVerse), computing providers, and publishers to implement reproducible data papers (Brinckman et al. 2019; Chard et al. 2019). Whole Tale and similar systems like Binder[4] and CodeOcean[5] provide mechanisms to package the data and products of research along with the code and computing environment that produced them, machine-readable provenance about how they were produced, references to published input data, and the scientific narrative that frames the rationale and conclusions for the work, all in a citable and re-executable Research Object (Bechhofer et al. 2010). These complex, hybrid publications span the boundaries between data, software, and publications to enable fully transparent research publication.

C. Tracking use and impact

Part of the purpose of citing data is to provide credit and attribution for the creation of the data set and correspondingly to help determine how and how often a data set is used. Data can have many important uses outside of research publications, though, and people are exploring better ways to track the impact of data. The National Information Standards Organization (NISO) defined a 'Recommended Practice' for alternative assessment metrics, but they primarily emphasize the need for data citation. They observe that 'there currently seems to be a lack of interest in altmetrics for data in the community' (NISO 2016: 16). Peters et al. (2016) also find little use of altmetrics and no correlation between altmetrics and citation. Because of inconsistent citation practices, text mining may be a better way to identify literature-data relationships (Kafkas et al. 2013). Nevertheless, research from Kratz and Strasser (2015b) indicates that citation and data downloads are the measures most valued by researchers. To that end, work through RDA has led to the 'Make Data Count' project (Cousijn et al. 2019; Kratz & Strasser 2015a), which has defined a consistent way to count data downloads through the COUNTER Code of Practice for Research Data (Fenner et al. 2018). These projects are still primarily oriented to the research community.

Much more research is needed on how to assess data use and impact beyond bibliometrics. Groups such as the VALUABLES consortium, are beginning to explore these issues through the lens of economics. For example, Bernknopf, et al. (2016) found that federal agencies could save $7.7 million/year in post-wildfire response if Landsat data were used; while Cooke & Golub (2019) report that a 30% reduction in weather uncertainty impacts on corn and soybean futures due to soil moisture measurements from NASA's Soil Moisture Active Passive satellite had a net worth of $1.44 Billion/year. Even more uncertain are methods of quantifying the impacts of data on public policy.

Providing credit for that impact is also tricky. Many people play critical roles in the creation of even the simplest data set, and their performance is evaluated in different ways. Not everyone is measured by their research paper publication record. One effort to recognize these other contributions is Project CRediT (Contributor Roles Taxonomy), which has defined a taxonomy of contributor roles for research objects and suggests using digital badges that detail what each author did for the work and link to their profiles elsewhere on the Web (Allen et al. 2014). We also recognize the concept of transitive credit, which could be used to recognize the developers of products other than papers (Katz 2014), and to use provenance to understand how upstream data and software enabled advances in research. Ultimately, we must recognize that credit is a human concern requiring context and interpretation that cannot be readily automated. We can work to build attribution into the scientific workflow, but human judgement is still required to assess the relative value of various contributions. Indeed, a study of software attribution found that automating credit mechanisms can lead to perverse metrics and incentives that can falsely represent the value of a contribution (Alliez et al. 2019).

[4] https://mybinder.org.
[5] https://codeocean.com.

Another concern related to both credit and understanding impact is to identify and trace the connections between all sorts of research objects (data, literature, software, people, organizations, algorithms, etc). Multiple efforts try to address this. One effort directly related to scholarly publishing is the Scholix (Scholarly Link Exchange) initiative which emerged out of RDA to more formally interconnect data and literature (Burton et al. 2017). The approach is functional and operational, as it builds from established citation hubs like DataCite, CrossRef, and OpenAire. This effort is expanding and collaborating with other related initiatives through a newly proposed 'Open Science Graphs for FAIR Data Interest Group' within RDA.

Other approaches use more decentralized, Web-based mechanisms which may allow more adaptability and extensibility (e.g., Ma et al. 2017; Parsons & Fox 2018). Work in ESIP and RDA explores how we can use schema.org web markup to identify and link data and repositories. ESIP is also exploring another W3C recommendation called Linked Data Notifications[6] — a sort of RSS-style protocol to request and receive notifications about activities, interactions, and new information. The notification itself is an individual entity with its own URI. Another emergent approach is the Digital Object Interface Protocol (Kahn et al. 2018) which assigns a persistent ID to any digital object and allows that object to express what type of object it is and what operations it allows such as various web services or basic management functions.

All these interconnecting technologies are somewhat peripheral to the core purpose of citation, but they highlight how applying machine-actionable PIDs to digital objects can expand the possibilities for knowledge sharing well beyond the bounds of traditional (paper-based) citation. This brings us to the issue and concern of identity itself.

D. Identifying things

People are beginning to rethink how PIDs work. To date, the basic issue of persistence of locators on the web has been addressed by what we might call authority-based identifiers, which separate the identity of an object from its location and are maintained in trusted registries. Klump et al. (2015, 2017) discuss how this approach has evolved and raise issues around the interconnection of identity, institutional commitment, and cost models. They note: 'The focus of the DOI for the data community on paper-like documents and human actors has left some conceptual gaps' (Klump et al. 2015: 133). They argue that we need to explore more advanced features such as identifier templates, more sophisticated content negotiation when resolving identifiers, more machine actionability in general, as well as the social process of maintaining the persistence of an object and its reference.

In other work, data managers are looking to content-based identifiers (i.e., cryptographic-hash-based IDs) to identify exact copies of data, ideally without relying on third parties and external administrative processes. These content-based identifiers can be deployed and resolved in peer-to-peer environments like the InterPlanetary File System (IPFS)[7] and Dat.[8] There is already an established system called Qri[9] (query) which allows users to reference, browse, download, create, fork, and publish data sets with a broad network of peers in IPFS. Furthermore, Dat includes public-key technology to provide assurance on the source of the data and any changes that may have occurred.[10] These approaches are still not well-suited to massive volumes of streaming data, and there is still the issue that different representations of data may be scientifically equivalent but not identical. The hash really needs to include the provenance chain as well as the data set. Nonetheless, these approaches show great promise. In related work, Bolikowski et al. (2015: 281) use the concepts of blockchain and version control systems like Git to propose a distributed system for maintaining long-term resolvability of persistent identifiers. They argue that the 'system should be agnostic with respect to referent type (data sets, source codes, documents, people) and content delivery technology (HTTP, BitTorrent, Tor/Onion)'.

It is important to note that content-based identifiers have quite different properties and applications from authority-based identifiers. Content-based identifiers and blockchain technologies can be useful for tracking provenance, precise data queries, and internal repository management concerns. Authority-based identifiers are being used for ensuring the social requirements necessary to maintain a persistent and managed access location (Di Cosmo et al. 2018), while those governance structures are only now being developed for distributed, content-based identifier systems.

---

[6] Linked Data Notifications: W3C Recommendation 2 May 2017 https://www.w3.org/TR/2017/REC-ldn-20170502/.
[7] https://ipfs.io.
[8] https://dat.foundation.
[9] https://qri.io.
[10] https://datprotocol.github.io/how-dat-works/.

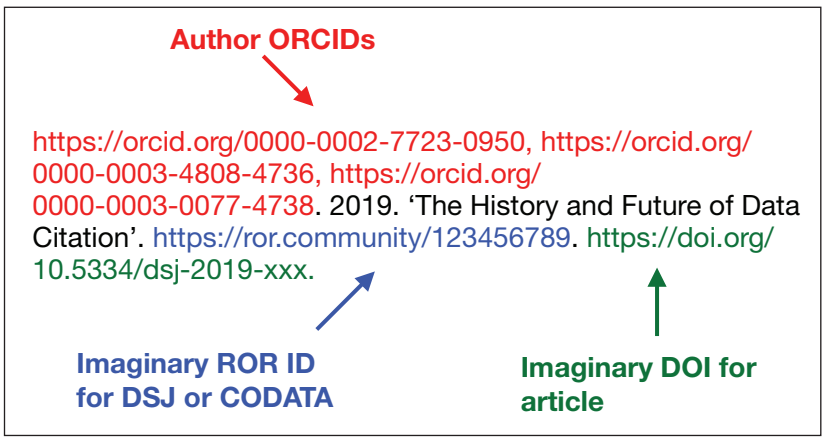

Figure 1: Imaginary citation for this article making full use of PIDs.

Finally, it is worth noting that various authority-based identifiers have or are being developed for other research objects and entities. These include the Open Researcher and Contributor ID (ORCID) for individual researchers (Haak et al. 2012); the Research Organization Registry (ROR) Community, which is working to develop identifiers for research organizations;[11] and the work of the Persistent Identification of Instruments Working Group of RDA,[12] which is implementing processes to use DataCite DOIs for scientific instruments. These are only a few examples of the growth in the development and application of PIDs. Similar to Linked Open Data approaches (Bechhofer et al. 2011; Bizer et al. 2009), these other-types of PIDs can make more elements of a citation precise and machine-actionable, but again they move us further away from the traditional human-oriented citation. Consider, fancifully, if this article was cited making full use of identifiers (**Figure 1**). It is more precise but also more opaque for the human reader.

Also, PIDs cannot always point precisely to the thing but rather only a representation of the thing (ORCIDs don't reference people, they reference descriptions of people). Access may be precisely defined, but credit and reference are inherently ambiguous to some degree (Hayes & Halpin 2008). The human remains in the loop.

IV. Looking to the Future

Despite recognized definitions (Borgman 2015; DCSG 2014; TGDCSP 2013), data citation remains a complex and evolving issue. On the one hand, the basic principles and process are well established. We know how to cite most data in research publications. We must only accelerate the implementation, and there does appear to be movement in that direction. On the other hand, long-established academic practices and assumptions about what is 'important' in scientific work lead us to bundle many different concerns into the concept of citation. At the same time, the opportunities promised by PIDs lead us to bundle even more concerns into reference schemes. Reconsideration and disaggregation of data citation concerns are overdue.

Data citation is often viewed as a computational or information science problem (Buneman et al. 2016; Silvello 2018), and sometimes more accurately as a social or cultural adaptation problem (Borgman 2015; Klump et al. 2015). But it is a complex socio-technical problem with many nuanced concerns. In short, it is an issue of praxis.

We are inspired by the Force11 effort to identify some of the myriad use cases for software citation (Smith et al. 2016). We feel we need to do the same with data citation: define multiple use cases that 1) de-emphasize the importance of the scientific paper in lieu of more precise assertions and supporting evidence and 2) emphasize the valuable use of data outside traditional scholarly environments. Some of this work has begun in RDA and ESIP. This should help us sort out the different concerns.

We already know that credit is primarily a human concern and access is a machine concern (reference is both), but what does that mean in practice? In health and social sciences, researchers have developed a 'Payback Framework' with a logical model of the complete research process and categories of (social health) payback from research (Donovan and Hanney 2011). Can we extend this and apply it to data by recognizing the reuse and value generated at many different stages? Can machine-actionable badges capture credit better than centralized citation indices?

[11] https://www.ror.community.
[12] https://www.rd-alliance.org/groups/persistent-identification-instruments-wg.

Recognizing access as a machine concern can help us focus on providing data as a service rather than simply as object downloaded by a human. This, in turn, can help us make intelligent choices about what type of identifiers to use for what application. The ID for the human interested in a general description of the data may be different and will behave differently than the ID for the machine.

Going forward, we should accelerate the substantial progress we have made on implementing data citation for the basic scholarly use case. At the same time, we should not overextend the concept nor expand our expectations for what citation can accomplish. It is time to rethink some of our assumptions if we are to make data both FAIR and fair.

Acknowledgements

This article is partially based on work done with community support provided by the Research Data Alliance and the Earth Science Information Partners. Parsons was partially supported by award G-2018-11204 from the AP Sloan Foundation. Duerr was partially supported by National Science Foundation award #1639753. Jones was partially supported by the National Science Foundation (awards #1546024 and #1430508) and the National Center for Ecological Analysis and Synthesis, a Center funded by the University of California, Santa Barbara, and the State of California.

Competing Interests

All the authors have been active in the development of data citation guidelines and implementations. This includes participating in relevant task groups and committees of multiple organizations, including but not limited to CODATA, DataOne, ESIP, Force11, and RDA. They have implemented data citation practice at multiple repositories, including but limited to the Arctic Data Center, the Deep Carbon Observatory Data Portal, the KNB Data Repository, and the National Snow and Ice Data Center. They co-authored data citation and usage standards including DCSG (2014), EDPSC (2019), Fenner et al. (2018), and TGDCSP (2013). MP is the Editor in Chief of the Data Science Journal, but did not oversee the review of this article.

References

Aalbersberg, IJ, et al. 2018. Making science transparent by default; introducing the TOP statement. DOI: https://doi.org/10.31219/osf.io/sm78t

Allen, L, et al. 2014. Publishing: Credit where credit is due. *Nature*, 508: 312–313. DOI: https://doi.org/10.1038/508312a

Alliez, P, et al. 2019. Attributing and referencing (research) software: Best practices and outlook from Inria. *Computing in Science & Engineering*. https://arxiv.org/abs/1905.11123.

Altman, M and **King, G.** 2007. A proposed standard for the scholarly citation of quantitative data. *D-Lib Magazine*, 13. http://dlib.org/dlib/march07/altman/03altman.html accessed 2019-07-25.

Baker, KS and **Yarmey, L.** 2009. Data stewardship: Environmental Data Curation and a Web-of-Repositories. *International Journal of Digital Curation*, 4. DOI: https://doi.org/10.2218/ijdc.v4i2.90

Ball, A and **Duke, M.** 2015. How to Cite Datasets and Link to Publications. Edinburgh: Digital Curation Centre. http://www.dcc.ac.uk/resources/how-guides accessed 2019-02-06.

Bechhofer, S, et al. 2010. Research objects: Towards exchange and reuse of digital knowledge. *The Future of the Web for Collaborative Science (FWCS 2010)*. https://eprints.soton.ac.uk/268555/ accessed 2019-07-28.

Bechhofer, S, et al. 2011. Why linked data is not enough for scientists. *Future Generation Computer Systems*. DOI: https://doi.org/10.1016/j.future.2011.08.004

Bernknopf, R, et al. 2016. The cost-effectiveness of satellite Earth observations to inform a post-wildfire response. *Working Paper*, 19–16. https://media.rff.org/documents/Valuables_Wildfires.pdf accessed 2019-07-28.

Berquist, CR, Jr. 1999. Digital map production and publication by geological survey organizations: A proposal for authorship and citation guidelines. *U.S. Geological Survey Open-File Report*, 99–386. https://pubs.usgs.gov/of/1999/of99-386/berquist.html accessed 2019-03-02.

Bizer, C, Heath, T and **Berners-Lee, T.** 2009. Linked data – the story so far. *International Journal on Semantic Web and Information Systems*, 5: 1–22. DOI: https://doi.org/10.4018/jswis.2009081901

Bolikowski, L, Nowiński, A and **Sylwestrzak, W.** 2015. A system for distributed minting and management of persistent identifiers. *International Journal of Digital Curation*, 10: 280–286. DOI: https://doi.org/10.2218/ijdc.v10i1.368

Borgman, C. 2015. *Big Data, Little Data, No Data*. Boston: MIT Press. DOI: https://doi.org/10.7551/mitpress/9963.001.0001

Borgman, C. 2016. Data citation as a bibliometric oxymoron. In: *Theories of Informetrics and Scholarly Communication*, Sugimoto, CR (ed.), 93–115. Berlin & Boston: Walter de Gruyter GmbH & Co KG. https://escholarship.org/content/qt8w36p9zf/qt8w36p9zf.pdf accessed 2019-07-26.

Brinckman, A, et al. 2019. Computing environments for reproducibility: Capturing the "whole tale". *Future Generation Computer Systems*, 94: 854–867. DOI: https://doi.org/10.1016/j.future.2017.12.029

Buneman, P, Davidson, S and **Frew, J.** 2016. Why data citation is a computational problem. *Commun ACM*, 59: 50–57. DOI: https://doi.org/10.1145/2893181

Burton, A, et al. 2017. The Scholix framework for interoperability in data-literature information exchange. *D-Lib Magazine*, 23. DOI: https://doi.org/10.1045/january2017-burton

Callaghan, S, et al. 2009. Overlay journals and data publishing in the meteorological sciences. *Ariadne*. http://www.ariadne.ac.uk/issue60/callaghan-et-al/ accessed 2011-11-27.

Chard, K, et al. 2019. Implementing computational reproducibility in the Whole Tale environment. *Proceedings of the 2nd International Workshop on Practical Reproducible Evaluation of Computer Systems – P-RECS '19*. DOI: https://doi.org/10.1145/3322790.3330594

Cooke, R and **Golub, A.** 2019. Market-based methods for monetizing uncertainty reduction: A case study. *Working Paper*, 19–15. https://media.rff.org/documents/WP_Cooke_Golub_4.pdf accessed 2019-07-28.

Costello, MJ. 2009. Motivating online publication of data. *Bioscience*, 59: 418–427. DOI: https://doi.org/10.1525/bio.2009.59.5.9

Cousijn, H, et al. 2018. A data citation roadmap for scientific publishers. *Sci Data*, 5: 180259. DOI: https://doi.org/10.1038/sdata.2018.259

Cousijn, H, et al. 2019. Bringing citations and usage metrics together to make data count. *Data Science Journal*, 18. DOI: https://doi.org/10.5334/dsj-2019-009

Crosas, M. 2014. The evolution of data citation: From principles to implementation. *IASSIST Quarterly*, 37: 62. DOI: https://doi.org/10.29173/iq504

DCSG – Data Citation Synthesis Group. 2014. *Joint Declaration of Data Citation Principles*. DOI: https://doi.org/10.25490/a97f-egyk

Di Cosmo, R, Gruenpeter, M and **Zacchiroli, S.** 2018. Identifiers for digital objects: The case of software source code preservation. *Open Science Framework*. DOI: https://doi.org/10.17605/OSF.IO/KDE56

Donovan, C and **Hanney, S.** 2011. The payback framework explained. *Research Evaluation*, 20: 181–183. DOI: https://doi.org/10.3152/095820211X13118583635756

Downs, RR, Duerr, R, Hills, DJ and **Ramapriyan, HK.** 2015. Data stewardship in the Earth sciences. *D-Lib Magazine*, 21. DOI: https://doi.org/10.1045/july2015-downs

EDPSC – ESIP Data Preservation and Stewardship Committee. 2019. *Data Citation Guidelines for Earth Science Data, Version 2*. Earth Science Information Partners. DOI: https://doi.org/10.6084/m9.figshare.8441816.v1

ESSCC – ESIP Software and Services Citation Cluster. 2019. *Software and Services Citation Guidelines and Examples. Ver. 1*. Earth Science Information Partners. DOI: https://doi.org/10.6084/m9.figshare.7640426

Fenner, M, et al. 2019. A data citation roadmap for scholarly data repositories. *Scientific Data*, 6(1): 28. DOI: https://doi.org/10.1038/s41597-019-0031-8

Fenner, M, et al. 2018. Code of practice for research data usage metrics release 1. DOI: https://doi.org/10.7287/peerj.preprints.26505v1

Guralnick, RP, et al. 2015. Community next steps for making globally unique identifiers work for biocollections data. *Zookeys*, 133–154. DOI: https://doi.org/10.3897/zookeys.494.9352

Haak, LL, et al. 2012. Orcid: A system to uniquely identify researchers. *Learned Publishing*, 25: 259–264. DOI: https://doi.org/10.1087/20120404

Hackett, EJ, et al. 2008. Ecology transformed: The national center for ecological analysis and synthesis and the changing patterns of ecological research. In: *Scientific Collaboration on the Internet*, 277–296. The MIT Press. DOI: https://doi.org/10.7551/mitpress/9780262151207.003.0016

Hayes, PJ and **Halpin, H.** 2008. In defense of ambiguity. *International Journal on Semantic Web and Information Systems*, 4: 1–18. DOI: https://doi.org/10.4018/jswis.2008040101

Howison, J and **Bullard, J.** 2016. Software in the scientific literature: Problems with seeing, finding, and using software mentioned in the biology literature. *Journal of the Association for Information Science and Technology*, 67: 2137–2155. DOI: https://doi.org/10.1002/asi.23538

Kafkas, Ş, Kim, JH and **McEntyre, JR.** 2013. Database citation in full text biomedical articles. *PLoS One*, 8: e63184. DOI: https://doi.org/10.1371/journal.pone.0063184

Kahn, R and **Wilensky, R.** 1995. A framework for distributed digital object services. http://handle.net/cnri.dlib/tn95-01 accessed 2019-07-20.

Kahn, RE, et al. 2018. *Digital Object Interface Protocol Specification, Ver. 2.0*. DONA. https://www.dona.net/sites/default/files/2018-11/DOIPv2Spec_1.pdf accessed 2019-07-25.

Katz, DS. 2014. Transitive credit as a means to address social and technological concerns stemming from citation and attribution of digital products. *Journal of Open Research Software*, 2: e20. DOI: https://doi.org/10.5334/jors.be

Katz, DS and **Chue Hong, NP.** 2018. Software citation in theory and practice. *Arxiv preprint*. https://arxiv.org/pdf/1807.08149.pdf accessed 2018-12-06.

Klump, J, et al. 2006. Data publication in the open access initiative. *Data Science Journal*, 5: 79–83. DOI: https://doi.org/10.2481/dsj.5.79

Klump, J, Huber, R and **Diepenbroek, M.** 2015. DOI for geoscience data-how early practices shape present perceptions. *Earth Science Informatics*, 1–14. DOI: https://doi.org/10.1007/s12145-015-0231-5

Klump, J, Murphy, F, Weigel, T and **Parsons, MA.** 2017. 20 years of persistent identifiers–applications and future directions. *Data Science Journal*, 16. DOI: https://doi.org/10.5334/dsj-2017-052

Kratz, JE and **Strasser, C.** 2015a. Comment: Making data count. *Sci Data*, 2: 150039. DOI: https://doi.org/10.1038/sdata.2015.39

Kratz, JE and **Strasser, C.** 2015b. Researcher perspectives on publication and peer review of data. *PLoS One*, 10: e0117619. DOI: https://doi.org/10.1371/journal.pone.0117619

Lawrence, B, et al. 2011. Citation and peer review of data: Moving towards formal data publication. *International Journal of Digital Curation*, 6. DOI: https://doi.org/10.2218/ijdc.v6i2.205

Lawrence, S, et al. 2001. Persistence of web references in scientific research. *Computer*, 34: 26–31. DOI: https://doi.org/10.1109/2.901164

Ma, X, et al. 2017. Weaving a knowledge network for deep carbon science. *Frontiers in Earth Science*, 5. DOI: https://doi.org/10.3389/feart.2017.00036

Mayernik, MS, Phillips, J and **Nienhouse, E.** 2016. Linking publications and data: Challenges, trends, and opportunities. *D-Lib Magazine*, 22. DOI: https://doi.org/10.1045/may2016-mayernik

McMurry, JA, et al. 2017. Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data. *PLoS Biol*, 15: e2001414. DOI: https://doi.org/10.1371/journal.pbio.2001414

Mooney, H and **Newton, MP.** 2012. The anatomy of a data citation: Discovery, reuse, and credit. *Journal of Librarianship & Scholarly Communication*, 1: 1–16. https://jlsc-pub.org/articles/abstract/10.7710/2162-3309.1035/ accessed 2017-10-03. DOI: https://doi.org/10.7710/2162-3309.1035

NISO. 2016. Outputs of the NISO alternative assessment metrics project: A recommended practice of the National Information Standards Organization. NISO RP-25-2016. https://www.niso.org/publications/rp-25-2016-altmetrics accessed 2019-07-22.

Parsons, MA, et al. 2008. Managing permafrost data: Past approaches and future directions. *Permafrost Ninth International Conference 29 June–3 July 2008 Proceedings*, 1369–1374. DOI: https://doi.org/10.5281/zenodo.3519368

Parsons, MA and **Fox, PA.** 2013. Is data publication the right metaphor? *Data Science Journal*, 12. DOI: https://doi.org/10.2481/dsj.WDS-042

Parsons, MA and **Fox, PA.** 2018. Power and persistent identifiers. *International Data Week 2018*. DOI: https://doi.org/10.5281/zenodo.1495321

Paskin, N. 2000. E-citations: Actionable identifiers and scholarly referencing. *Learned Publishing*, 13: 159–166. DOI: https://doi.org/10.1087/09531510050145308

Peters, I, et al. 2016. Research data explored: An extended analysis of citations and altmetrics. *Scientometrics*, 107: 723–744. DOI: https://doi.org/10.1007/s11192-016-1887-4

Rauber, A, Asmi, A, van Uytvanck, D and **Proell, S.** 2015. *Data Citation of Evolving Data: Recommendations of the Working Group on Data Citation (WGDC)*. Research Data Alliance. Accessed 2019-07-14. DOI: https://doi.org/10.15497/RDA00016

Schopf, JM. 2012. Treating data like software: A case for production quality data. *Proceedings of the Joint Conference on Digital Libraries*, 11–14 June 2012. Washington DC. DOI: https://doi.org/10.1145/2232817.2232846

Silvello, G. 2018. Theory and practice of data citation. *Journal of the Association for Information Science and Technology*, 69: 6–20. DOI: https://doi.org/10.1002/asi.23917

Smith, AM, Katz, DS, Niemeyer, KE, FORCE11 and **SCWG.** 2016. Software citation principles. *PeerJ Computer Science*, 2: e86. DOI: https://doi.org/10.7717/peerj-cs.86

Stall, S, et al. 2018. Advancing FAIR data in Earth, space, and environmental science. *Eos*, 99. DOI: https://doi.org/10.1029/2018EO109301

Stockhause, M and **Lautenschlager, M.** 2017. CMIP6 data citation of evolving data. *Data Science Journal*, 16. DOI: https://doi.org/10.5334/dsj-2017-030

Stuart, D. 2017. Data bibliometrics: Metrics before norms. *Online Information Review*, 41: 428–435. DOI: https://doi.org/10.1108/OIR-01-2017-0008

TGDCSP – Task Group on Data Citation Standards and Practices, CODATA-ICSTI. 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. DOI: https://doi.org/10.2481/dsj.OSOM13-043

Wilkinson, MD, et al. 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3: 160018. DOI: https://doi.org/10.1038/sdata.2016.18

Wittenburg, P, Hellström, M, Zwölf, CM, Abroshan, H, et al. (eds.) 2017. *Persistent identifiers: Consolidated assertions*. Research Data Alliance. DOI: https://doi.org/10.15497/RDA00027

How to cite this article: Parsons, MA, Duerr, RE and Jones, MB. 2019. The History and Future of Data Citation in Practice. *Data Science Journal,* 18: 52, pp. 1–10. DOI: https://doi.org/10.5334/dsj-2019-052

Submitted: 30 July 2019 **Accepted:** 04 October 2019 **Published:** 01 November 2019

]u[*Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press. OPEN ACCESS