A Standard for the Scholarly Citation of Archaeological Data

28 September, 2017

|  |
| --- |
| Ben Marwick1 & Suzanne E. Pilaar Birch2 |
|  |
| 1 University of Washington, Seattle, Washington, USA |
| 2 University of Georgia, Athens, Georgia, USA |
|  |

# Author note

Correspondence concerning this article should be addressed to Ben Marwick, Department of Anthropology, University of Washington, Seattle, 98115, WA, USA. E-mail: [bmarwick@uw.edu](mailto:bmarwick@uw.edu)

# Abstract

"How do archaeologists share their research data, if at all? We review what data are, according to current influential definitions, and previous work on the benefits, costs and norms of data sharing in the sciences broadly. To understand data sharing in archaeology, we present the results of three pilot studies: requests for data by email; review of data availability in published articles, and analysis of archaeological datasets deposited in repositories. We find that archaeologists are often willing to share, but discipline-wide sharing is patchy and ad hoc. Legislation and mandates are effective at increasing data-sharing, but editorial policies at journals lack adequate enforcement. Although most of data available at repositories are licensed to enable flexible reuse, only a small proportion of the data are stored in structured formats for easy reuse. We present some suggestions for improving the state of date sharing in archaeology, among these is a standard for citing data sets to ensure that researchers making their data publicly available receive appropriate credit."

*Keywords:*

Word count:

# A Standard for the Scholarly Citation of Archaeological Data

per mille sign

plain text: ‰ (doesn't render properly in PDF) HTML: ‰ (renders properly in PDF) LaTeX: $\text{\textperthousand}$ (renders properly in PDF) delta sign

plain text: δ (doesn't render properly in PDF) LaTeX: (renders properly in PDF) plus-minus sign

plain text: ± (doesn't render properly in PDF) HTML: ± (renders properly in PDF) LaTeX: (renders properly in PDF) degree sign

plain text: ° (doesn't render properly in PDF) HTML: ° (renders properly in PDF) LaTeX: $\text{\textdegree}$ (renders properly in PDF)

Data are the building blocks of research, and access to data is necessary to fully understand and extend the work of others (Fienberg et al 1985). Yet archaeological data are often not available for a variety of reasons, including a perceived lack of credit for making data available. The goal of this paper is to present a standard for the scholarly citation of archaeological data, akin to the way we cite publications. The need for this has been motivated by Huggert’s (2017) findings of low levels of data citation and reuse in archaeology, and Clarke’s (2015) analysis that digital data preservation and management systems are under-developed in archaeology, despite our obligations as stewards of cultural heritage. A standard for data citation is important because it gives credit to the authors or compilers of the data, and so accords them visibility and recognition for their scholarly contribution in a way that is valued broadly in the discipline. A citation standard is also important for tracking the use of the data, and generating metrics about the impact of the dataset on the discipline. Understanding the use of resources such as datasets is important for allocating resources within archaeology and identifying novel research directions.

We will start by defining what we mean by ‘data’, and reviewing previous work on researchers’ attitudes to sharing data. We motivate data sharing and citing behaviours by explaining why data sharing is beneficial to archaeology as a discipline, and to individual researchers. Then we will survey some of the data sharing and citing norms that have emerged in other research areas. From these norms we will present a standard suitable for citation of archaeological data. Finally, we will present some basic instructions on how to share archaeological data.

# What are data?

Defining what data are is an ongoing research problem for information scientists (Leonelli 2016; Borgman 2012). Table 1 summarises some of the definitions debated in the philosophy of science literature. A more concrete starting point is this definition from the National Academies of Sciences:

"The term "data" as used in this document is meant to be broadly inclusive. In addition to digital manifestations of literature (including text, sound, still images, moving images, models, games, or simulations), it refers as well to forms of data and databases that generally require the assistance of computational machinery and software in order to be useful, such as various types of laboratory data including spectrographic, genomic sequencing, and electron microscopy data; observational data, such as remote sensing, geospatial, and socioeconomic data; and other forms of data either generated or compiled, by humans or machines." (Uhlir & Cohen, 2011, cited in Borgman 2015)

Table 1 A selection of definitions of research data

|  |  |
| --- | --- |
| Definition of data | Source |
| Anything portable thing that is treated as potential evidence for one or more claims about phenomena | Leonelli 2015 |
| A lack of uniformity in the real world | Floridi 2011 |
| Traces or signals that have been manipulated so that the can be stored and retrieved | Rheinberger 2011 |
| Anchors of knowledge claims that are stable across different contexts | Latour 1999 |
| Marks or traces produced by human interactions with research instruments | Hacking 1992 |
| Evidence for the identification and characterisation of phenomena | Bogen and Woodward 1988 |

This is a useful definition because it emphasizes data as digital manifestations, and so suitable for transmission, storage and retrieval using the internet. Physical objects (such as samples, collections, hard copies of notes, etc.) are data also, but they are out of scope of this discussion because of the different requirements for access, archiving, and so on. For many archaeologists, data are the contents of the files we use to generate visualisations, summary tables, and qualitative statements that appear in, and support the claims we make in, our scholarly publications and reports. The categories shown in Table 2 are useful as a rough guide to what data can be, although this table is not comprehensive, and does not represent the full diversity of data used by archaeologists. Ultimately, data are whatever the community decides, as the The National Science Foundation notes in their commentary on Data Management Plans: "What constitutes such data will be determined by the community of interest through the process of peer review and program management. This may include, but is not limited to: data, publications, samples, physical collections, software and models." This is a highly pragmatic approach to defining data, but is also consistent with the relational account of data in philosophy of science scholarship, which holds that what data are is dependent on who uses it, how and for what purpose (Leonelli 2015; Daston and Galison 1992). It implies that the answer to the question of what counts as data is mostly emergent from our professional interactions, such as the materials we request when we do peer review of manuscripts and grant proposals, or when we make an informal request for more information by email.

Pragmatically in our own work, we consider that the main data suitable for sharing as the contents of the computer files (e.g. text, raw structured data, and images) that we analyse or otherwise use to create the figures, tables and quantitative results that we include in our reports and publications. The selection omits materials that could also be considered data according to the definitions noted above (e.g. field notes), but we feel our selection represents the most information-dense and meaningful data records that are useful for others wanting to assess the reliability of our work, or reuse our findings in their own work. These are also the items that have been most frequently requested by others writing to us about our publications. Our citation standard imposes no constraints on what data can be, but we propose a few minimal qualities for how data should be organised that will help ground our discussion. Table 3 summarises some desirable minimum qualities for archaeological data that optimise its discovery and reuse.

Table 2 A selection of types of data encountereed by archaeologists.

|  |  |
| --- | --- |
| Type of data | Description |
| Observational | Includes site descriptions and artefact measurements, which may be associated with specific places and times, or multiple places and times, as well as ethnoarchaeological observations. |
| Computational | Includes the results of simulations and computer models, such as agent-based models or virtual reality models, hypothesis tests, and regression models. |
| Experimental | Includes results from laboratory procedures, such as instrumental analyses and chemical reactions to determine artefact composition, or field experiments such as knapping, trampling and taphonomic studies |
| Records | Includes government, business, public, and private documents |

Table 3 Best practices for preparing data for sharing, from Hart et al. (2016)

|  |  |
| --- | --- |
| Rule | Details |
| Anticipate how your data will be used | Before collecting data, decide what formats and metadata schema will be used, how and where the data will be archived, embargoed, released, and what software others are likely to use to work with your data. |
| Keep raw data raw | Data should be as unprocessed as possible, within practical limitations, to facilitate future analysis with new methods. The line between raw and derived data is not often clear, so the general idea is to provide data that will enable others to confirm your analytical results, assess the validity of your statistical output, or directly compare findings across other studies. |
| Store data in open formats | Data should be in a format that is easily readable by computers, without being dependent on specific software. Examples of open data formats include comma-separated values (CSV) for tabular data, portable network graphics (PNG) for images, and KML (or other Open Geospatial Consortium [OGC] format) for spatial data. Examples of closed formats include Photoshop document (PSD) for bitmap images, and Microsoft Excel (XLS) for tabular data. While day-to-day work may require closed formats, the files that are archived should be in open formats. |
| Data should be structured for analysis | Data should be structured so it is easy to analyse with a computer, and easy to combine with other datasets. For tabular data, each variable or measurement should be a single column, each observation or record should be a single row, and each type of observational unit should be a table. Each cell should contain a single value, not include units, comments or derivative values (such as percentage) in the same cell as the value. Be consistent with capitalisation, abbreviations, naming conventions, and avoid exotic characters such as accents that may not be readable on other systems. Use a good null value such as a blank cell or NA. Don't use spaces in column names. |
| Data should be uniquely identifiable | Data used in a scholarly publication should be archived online at a location that can be uniquely identified by a persistent URL or a or a Globally Unique Identified/Universally Unique Identifier (GUID/UUID). Types of persistent URLs include Digital Object Identifiers (DOIs), Archival Resource Keys (ARKs), or a persistent URL (PURLs). Individual files should have a distinct names, which includes a version identifier. This identifier could be the date in the ISO 8601 standard (YYYY-MM-DD, i.e. from largest time unit to smallest). |
| Provide relevant metadata | Include information about what, when, where, and how of the data were collected. For tabular data, this should include descriptions of the column names, units of measurement, warnings about missing or inconsistent data, and information to check that the data are complete (e.g. the number of rows and columns that are in the table when it was created) |
| Adopt the proper privacy protocols | Take care not to include data that might expose people, places or objects to risk of harm or loss. Negotiating with stakeholders such as Indigenous groups is essential to avoid harm in data sharing. This might include omitting or degrading location data (e.g. so that site locations are not revealed to the public), and removing or anonymising personally identifiable information. Methods for ensuring privacy can be complex and change rapidly, so seeking expert help may be necessary. |
| Use a trustworthy repository | Data should be easy to find and available for a long time. Archive your data in a major well-established repository that provides persistent URLs and has a commitment to sustainability (e.g. participates in the CLOCKSS archive). Do not use your personal website or a project website, which may change or disappear. |
| Use an open licence | Data should have an explicit and well-established license to let others know exactly what they can and cannot do with the data you shared. One of the most widely used and frequently recommended license is CC-0 or public domain. This places no restrictions on data use, promoting greater engagement with the data. |

# What are the benefits of data sharing to the individual researcher?

While public and community benefits and costs are well-rehearsed and widely agreed on (Table 4), they offer no, or only weak, tangible incentives to the individual researcher (Rowhani-Farid et al. 2017). Previous work has identified citation advantages as a tangible benefit to individual researchers who share data with publications, although the magnitude of the advantage varies greatly (see Table 5 for a summary). Increased publication productivity also appeared to be associated with data sharing behaviours. Across over 7,000 NSF and NIH awards, Pienta et al. (2010) reported that research projects with archived data produced a median of 10 publications, versus only five for projects without archived data.

Table 4 Summary of public and community benefits and costs of data sharing

|  |  |  |
| --- | --- | --- |
| Reasons to share | Reasons not to share | Source |
| Encouraging scientific advancement | The time it takes to clean up and document for release | Stodden 2010 |
| Being a good community member | The possibility that your data may be used without citation | Stodden 2010 |
| Potential to encourage others to work on the problem | Legal barriers, such as copyright | Stodden 2010 |
| Encouraging sharing and having others share with you | Time to verify privacy or other administrative data concerns | Stodden 2010 |
| The potential to set a standard for the field | The potential loss of future publications using this data | Stodden 2010 |
| Improvement in the caliber of research | Competitors may get an advantage | Stodden 2010 |
| Increase in publicity | Dealing with questions from users about the data | Stodden 2010 |
| Opportunity to get feedback on your work | Technical limitations, ie. webspace platform space constraints | Stodden 2010 |
| Potential for finding collaborators | Whether there is intense competition in the topic | Stodden 2010 |
| Normalizing understanding in a field | Whether you put in a large amount of work building the dataset | Stodden 2010 |
| To reproduce or to verify research | Insufficient perceived reward, such as promotion or subsequent citation | Borgman 2012 |
| To make the results of publicly funded research available to the public | Effort in documenting | Borgman 2012 |
| To enable others to ask new questions of extant data | Concerns for priority, including control of results and sources | Borgman 2012 |
| To advance the state of research and innovation | Intellectual property issues | Borgman 2012 |

Table 5 Summary of individual benefits to data sharing

|  |  |  |
| --- | --- | --- |
| Research area | Citation advantage (%) | Source |
| Genetics | 9 | Piwowar and Vision (2013) |
| Astronomy | 20 | Henneken and Accomazzi (2011) |
| Astrophysics | 28 | Dorch et al., 2015 |
| Palaeocenography | 35 | Sears (2011) |
| Peace Research | 100 | Gleditsch et al., 2003 |

# What do researchers think about sharing data?

Sharing of data is a quality that is fundamental to the definition of science (cf. Merton’s (1942) communalism norm, where intellectual property rights are normally given up in exchange for recognition and esteem) and to data both as an abstract quality (cf. the definitions in Table 1 that mention data as ‘portable’, ‘stable across different domains’) and as a concrete detail to be managed with policy. For example, for over a decade the National Science Foundation (2016) has had a data sharing policy that states "Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants. Grantees are expected to encourage and facilitate such sharing." Surveys of researchers across different domains repeatedly show that people value the ideal of data sharing (Kratz and Strasser 2015). However, data sharing is not widespread in many disciplines, raising questions about why researchers do or do not share their data. Three recent studies have attempted to answer these questions, in the search for strategies to encourage sharing.

Stodden (2010) surveyed registrants at a top machine learning conference (the Neural Information Processing Systems (NIPS) conference) to identify the factors that affect researchers’ decisions to reveal data (Table 4). Stodden found that communitarian factors influence scientists to share their data, but private incentives determine when scientists choose not to share their code and data. In the top three reasons not to share, we see a lack of incentives to invest time in data sharing (however, there are many other time-consuming tasks in research, and they get done because clear reward structures are in place); a lack of faith in data citation (which we hope this paper will help with); and a lack of an efficient copyright process. Stodden’s (2009a, b) ‘Reproducible Research Standard’, summarised in Figure 1, is a convenient and robust copyright schema that is in wide use for many typical research projects.

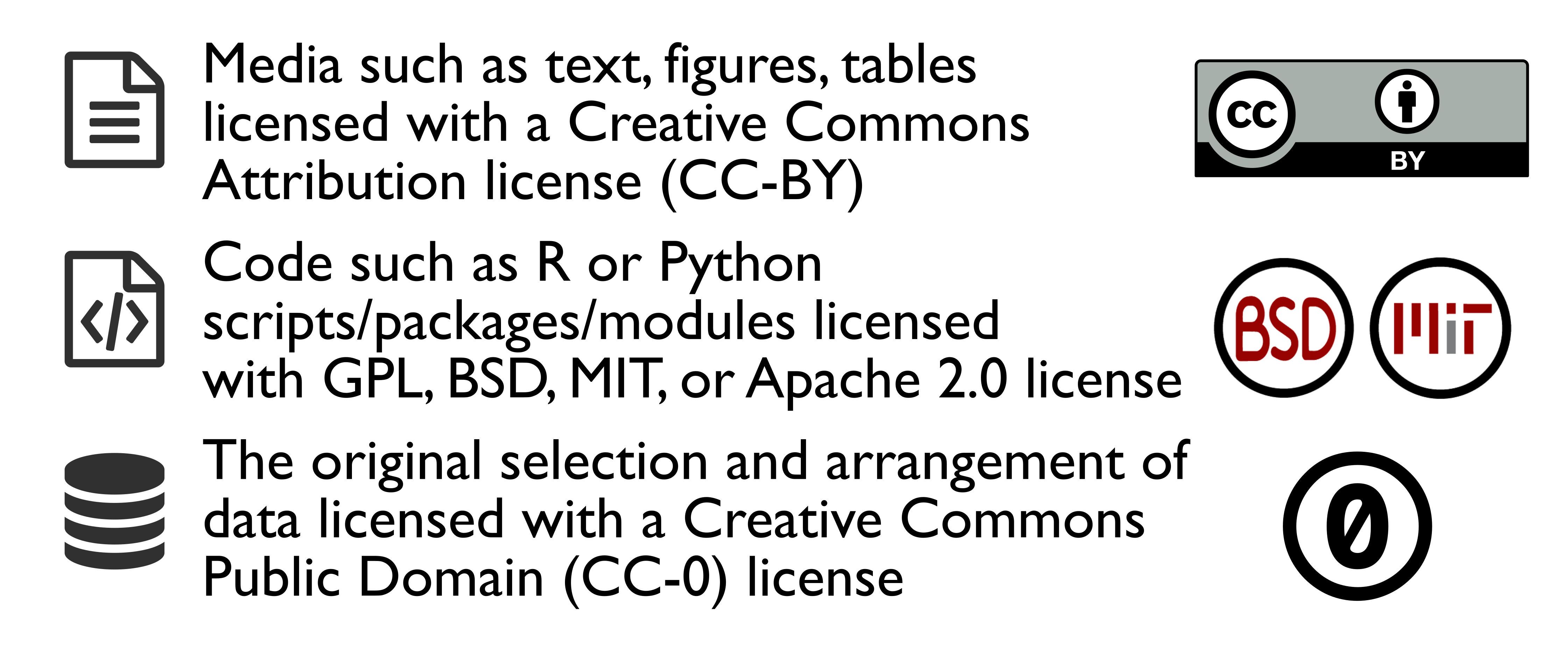


Figure 1 Summary of Stodden’s (2009b) 'Reproducible Research Standard'

Other surveys have found similar results. Tenopir et al. (2011) received 1,329 responses to a survey conducted by the research team of the National Science Foundation-funded DataONE project. They found that when researchers did not share data it was due to reasons that include: insufficient time (54%), a lack of funding (40%), having no place to put the data (24%), a lack of standards (20%), and because their sponsor did not require it (17%). A remarkable observation by Tenopir et al. is that while 75% of respondents say they share data, only 36% of those respondents say their data are easy to access, while only 6% say they make all their data available. This suggests that most data sharing occurs in private directly between individual researchers, rather than in public via open repositories. A similar focus on private, peer-to-peer sharing is evident in interviews conducted by Wallis et al. (2013) with 43 researchers at the Center for Embedded Networked Sensing. They found that the most commonly cited conditions for sharing data were (1) first rights to publish results, (2) proper attribution to the data source, (3) familiarity between sharer and recipient, (4) funding agency expectations, and (5) the amount of effort required to share.

# What are the current norms of data sharing?

The surveys above show that data sharing practices tend to be a private activity, but where public sharing does occur, what forms does it take, and how do these vary across different fields of research? Wouters and Haak (2017) surveyed 1,200 researchers about data sharing behaviours and found a spectrum from ‘intensive data-sharing fields’ to ‘restricted data-sharing fields’. Intensive data-sharing fields often depend on access to global databases, for example human genetics, while restricted data-sharing fields depend mostly on data collected by the individual researcher or their immediate collaborators, such as digital humanities. These differences in disciplinary norms have implications for data ownership (intensive: collectively owned, willing to share; restrictive: owned by the individual, reluctant to share) and the effort required to share (intensive: low effort because data are prepared for sharing as an integral part of the research process, restrictive: high effort because data are prepared for sharing after the research, or apart from it, in response to a request from another researcher). This intensive-restricted spectrum also maps on to Sawyer's (2008) analysis of data sharing in data-rich fields compared to data-poor fields. Data rich fields, such as astronomy, oceanography, and ecology, are have three common characteristics: pooling and sharing data resources is expected; the cumulative nature of work that builds on common data and common methods reduces the likelihood of alternative data sources and other methods being common; and the number of theoretical choices is limited because there is enough data to adequately test competing theories. Data-poor fields have three common characteristics: data are rare, or hard to get, or both; the types of data available often dictate the methods taken; and theories are relatively easy to generate and more difficult to validate. In data-poor fields, researchers often invest substantial time and resources on gathering their own data. The getting of data, and the quality of the data collected, is often a key determinant of the status of a researcher, leading to hoarding of data as an asset to be traded for professional benefit.

Archaeologists can be found all over this spectrum because of the diversity of archaeological research. However, our observations of data sharing in archaeology, and the nature of archaeological research generally, is that it is, for the most part, currently a restricted data-sharing and a data-poor field. There are few standardised and widely used methods for sharing and citing data, a challenge for archaeological data sharing that we hope this paper will help to address.

Even though there is a wide spectrum of data sharing behaviours, overall rates of data sharing by individual researchers are very low in many fields. Wicherts et al. (2006) emailed authors of 141 empirical articles recently published by the American Psychological Association (APA) journals to request a copy of the original data reported in the articles. They received 64 of the 249 datasets they requested, resulting in a data sharing rate of about 26%, despite all the authors of the articles in their study being bound to the APA's Ethical Principles, which state authors must ‘not withhold the data on which their conclusions are based from other competent professionals who seek to verify the substantive claims’ (American Psychological Association, 2001, p. 396). Valpaemel et al. (2015) similarly requested data from 394 papers, published in all issues of four APA journals in 2012 and found that 38% of authors responded with their data. They concluded that although the data sharing rate increased as compared to the study by Wicherts et al., their findings are ‘worrisome’ and indicate a 'poor availability of data'.

Savage and Vickers (2009) requested data from authors of ten papers published in PLOS journals, and received only one reply. Although this is a very small sample, it is notable because PLOS journals require authors to share data as a condition of publication: "PLOS journals require authors to make all data underlying the findings described in their manuscript fully available without restriction, with rare exception." (PLoS Medicine Editorial and Publishing Policies nd), and yet only one out of the ten authors actually complied with this requirement. Alsheikh-Ali et al. (2011) inspected 500 articles published in high-impact science journals in 2009 and found that only 47 papers (9%) deposited full primary raw data online. Similarly, Womak (2015) inspected 4,370 articles published in high-impact science journals in 2014 and found that 13% of articles with original data published make the data available to others.

The absence of a standardised and sustainable approach to data sharing is starkly demonstrated by Vines et al.’s (2014) study of data availability over time in biology publications. They examined the availability of data from 516 studies of plants or animal morphology between 2 and 22 years old. Vines et al. found that odds of a data set referred to in a publication being available fell by 17% per year. The found that the key factors responsible for the unavailability of data was that the data were lost or on inaccessible storage media.

# Do archaeologists share data?

Here we present the results of three small pilot studies aimed at understanding how archaeologists share their data. First, we emailed a sample of authors of research publications to request access to the data behind their publication. Second, we surveyed a sample of journal articles to see what data was available. Third, we analysed information about archaeological data files in online repositories tracked by DataCite (an organisation that provides persistent identifiers, or DOIs for research data). The small sample sizes of the first two of these pilot studies limit how generalisable the results are, but as the first of their kind for archaeology, they are valuable for proposing hypotheses for larger-scale investigations.

## Reproducibility and open source materials for this study

To enable re-use of our materials and improve reproducibility and transparency according to the principles outlined in Marwick (2017), we include the entire R code used for all the analysis and visualizations contained in this paper in our compendium at <http://dx.doi.org/10.17605/OSF.IO/KSRUZ>. Also in this version-controlled compendium are the raw data for all the tests reported here. All of the figures and quantitative results presented here can be independently reproduced with the code and data in this repository. In our compendium our code is released under the MIT license, our data as CC-0, and our figures as CC-BY, to enable maximum re-use (for more details, see Marwick 2017).

## Email requests for data in a published article

Our sample of authors that we wrote email to focused on *Journal of Archaeological Science* articles on experiments with stone artefacts that were published during 2009 and 2015. We focused on stone artefact experiments because the all research materials in these projects are generated by the researcher, so there are no other stakeholders to consider when determining how to share the data: the researcher has complete control over the data. There are also no risks of damage to people or property in sharing these data because they do not include archaeological sites or human subjects. In contacting the authors, we requested the datasets for use in a graduate seminar with the aim of reproducing the published results and further exploring the data (this stage is still in progress). We promised to contact the authors if we discovered anything new, and not to share the data further without their permission. The data included with this paper is anonymised so the specific authors we contacted are not identifiable.

Our sample focused on *Journal of Archaeological Science* articles on experiments with stone artefacts that were published during 2009 and 2015. We focused on stone artefact experiments because the all research materials in these projects are generated by the researcher, so there are no other stakeholders to consider when determining how to share the data: the researcher has complete control over the data. There are also no risks of damage to people or property in sharing these data because they do not include archaeological sites or human subjects. In contacting the authors, we requested the datasets for use in a graduate seminar with the aim of reproducing the published results and further exploring the data (this stage is still in progress). We promised to contact the authors if we discovered anything new, and not to share the data further without their permission. The data included with this paper is anonymised so the specific authors we contacted are not identifiable.

We sent out emails to the first authors of 23 stone artefact experiment papers, and received 15 replies, resulting in a 70% response rate. We received 5 responses that included data files, giving an overall sharing rate of 20%. Our small sample size limits the robustness of our analysis, but points to some observations that may be worth following up in a larger study. We found no clear relationship of data sharing and date of publication. We do not have any insights into why authors of older papers decline to share, but three of the six responses for papers published in 2014-2015 said they declined to share data because they intend to use these data in a future publication. Other noteworthy reasons provided by respondents for declining to share include the author about to defend their thesis or about to go on vacation. We interpret these to mean that the author does not have time to prepare their data for sharing, perhaps because the data are not already in a well-organised for ready for distribution. We found no effect on the author's student status at time of publication on data sharing behaviours, with roughly equal proportions of students and non-students sharing and not sharing.

Similarly there are no clear patterns data sharing of among different journals. It is remarkable that the authors of the PLOS One paper we contacted declined to share because that refusal violates the policy of that journal at the time of publication of that paper (early 2014). Our finding is not unique - Chambers' (2017) examined 50 PLOS One papers on brain images published in November 2014 - May 2015 and found that only 38% of these papers archived their data in an open repository. Similarly, Nuijten et al. (2017) found that of 462 PLOS One articles that promised data availability, 29% of these did not have any available data. Unfulfilled promises to share data are not unique to PLOS One (Kidwell et al. 2016), but these studies highlight the need to resolve ambiguity among authors about what counts as data, and for editors and peer reviwers to more critically assess whether an author has made their data openly available.

We see similar rates of sharing in response to email requests among archaeologists to what have been described in other disciplines (Table 6, Tenopir et al. 2015; Vanpaemel et al. 2015). However, our result may overestimate the true rate because our sample did not include datasets that authors might not be able or willing to share because they contain sensitive location or cultural information. Analysis of the qualitative responses suggests two key reasons why authors are reluctant to share: they are fearful that they will lose the opportunity to produce further publications from those data (a typical concern of researchers in data-poor fields), and that their data are not in a form suitable for sharing and the task of organising those data is not a high priority. These reasons not to share have also been documented in other research areas, noted above. The absence of patterns in the date of the paper, student status of the author and other variables may relate to more to our small sample size than a real absence of relationships in data sharing behaviour.

Table 6 Summary of a selection of previous studies of data sharing in various fields, from Spencer (2010). In the study by Peng et al., the authors did not attempt to contact other researchers, but noted the availability of data as described in the published article.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Subject area | Requested datasets | Available datasets | Source journal(s) | Study |
| Cancer | 64 | 42 (66%) | Various | Kyzas et al. (2005) |
| Medicine, Cancer | 85 | 41 (48%) | Various | Piwowar et al. (2007) |
| Psychology | 141 | 64 (45%) | Journal of Personality and Social Psychology, Developmental Psychology, Journal of Consulting and Clinical Psychology, Journal of Experimental Psychology: Learning, Memory, and Cognition. | Wicherts et al. (2006) |
| Economics | 15 | 6 (40%) | American Economic Review, the International Journal of Industrial Organization, and the Journal of International Economics | McCullough and Vinod (2005) |
| Economics | 193 | 69 (36%) | Jounal of Money, Credit and Banking | McCullough et al. (2006) |
| Medicine | 10 | 1 (10%) | PLoS Medicine, PLoS Clinical Trials | Savage and Vickers (2009) |
| Medicine | 29 | 1 (3%) | British Medical Journal | Reidpath and Allotey (2002) |
| Epidemiology | 69 | 0 (0%) | American Journal of Epidemiology, Journal of the American Medical Association | Peng et al. (2006) |

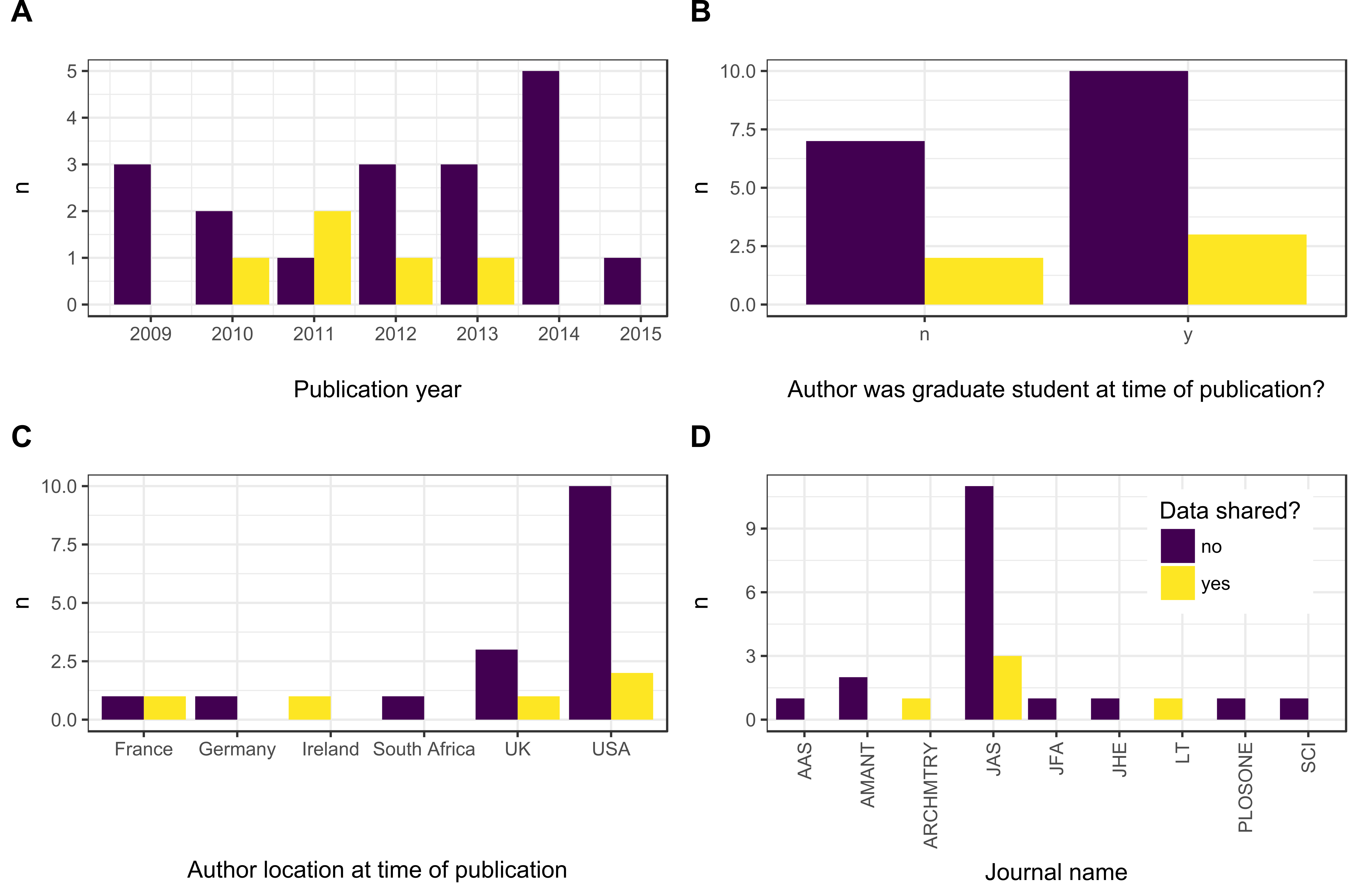


Figure 2 Summary of results of email requests for data. A. Responses by year of journal article publication, B. Student status of first author, C. Country location of first author, D. Journal (AAS = Archaeological and Anthropological Sciences, AMANT = American Antiquity, ARCHMTRY = Archaeometry, JAS = Journal of Archaeological Science, JFA = Journal of Field Archaeology, JHE = Journal of Human Evolution, LT = Lithic Technology, SCI = Science.

## Survey of articles published in the *Journal of Archaeological Science*

In addition to our email survey, we also conducted a pilot study of a random sample of 34 articles published during 2016-2017 in the *Journal of Archaeological Science* to investigate data sharing behaviours. In this sample we found openly available raw data in 18 papers (53%), even though only 7 articles (21%) include a data availability statement. For papers where data was available, 9 papers (50%) have data contained in supplementary files published with the paper, 6 (33%) have all the raw data in tables in the text of the paper, and 3 (17%) have data in an online repository. The most frequently shared type of data is compositional (e.g. element concentrations, n = 9, 28%), followed by radiocarbon and luminescence age data (where the dates are the primary object of analysis, n = 5, 16%) and DNA sequences (n = 4, 12%). Generally, compositional data is presented in tables in the text, dating data is presented in supplementary files, and DNA data are in repositories (as is typical for DNA data). Data shared in supplementary files and online repositories were most often in Microsoft Excel format (n = 38, 38% of the files in supplements files and repositories), followed by CSV (comma separated variables, an open source plain text spreadsheet format, n = 4, 31%), and tables embedded in Microsoft Word or PDF documents (n = 4, 31%). When a scripting language such as R, Python, MATLAB or OxCal was used as the analysis software for a paper (n = 7, 54% of the papers where would identify the software), only 2 papers also provided the script with the paper.

The *Journal of Archaeological Science* has had a 'data disclosure' policy since at least 2013 (Editors, 2013) which states that “All data relating to the article must be made available in Supplementary files or deposited in external repositories and linked to within the article”. The data availability rate of 53% reported in our pilot study reflects either weak enforcement of this policy or an incomplete understanding among editors and authors of how to interpret and implement this policy. The prominence of compositional and dating data among the shared data types suggests a least effort strategy by authors sharing data that do not require extensive cleaning and tidying (because the data are generated by instruments rather than entered by hand, such as artefact attribute data). The size of compositional and dating datasets means that they mostly fit easily into tables in the text or supplementary files, and do not require much effort to organise (e.g. very little metadata is required for others to use these data). The small proportion of papers in our sample using or sharing a scripting language suggests that unlike other fields (Eglen et al. 2017), archaeologists do not appear to recognise code for data analysis as a primary research object to share with the research community.

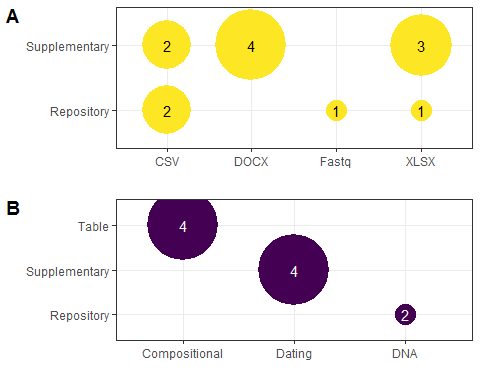


Figure 3 Summary of relationships between file types, data types, and data locations. A. File types and data locations, B. Data types and data locations.

## Data sets in repositories tracked by DataCite

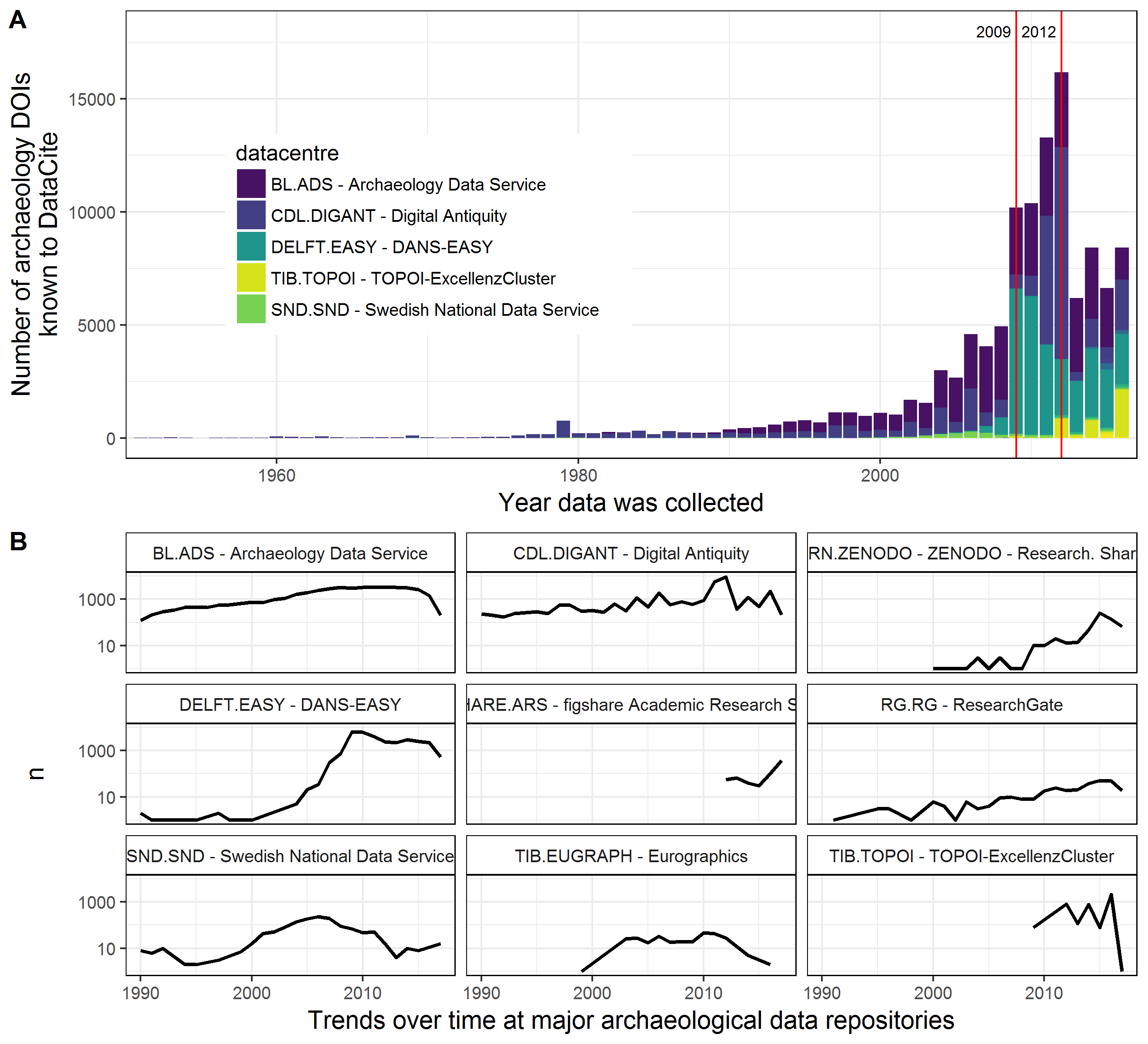


Figure 4 Change over time in archaeology datasets known to DataCite. A. Counts of datasets per year by the major data centres, red lines indicate major increases in the number of datasets in repositories. Note that the date on the horizontal axis indicates when the data was collected, not when it was deposited in the data centre, B. Temporal trends in the number of DOIs at individual data centres.

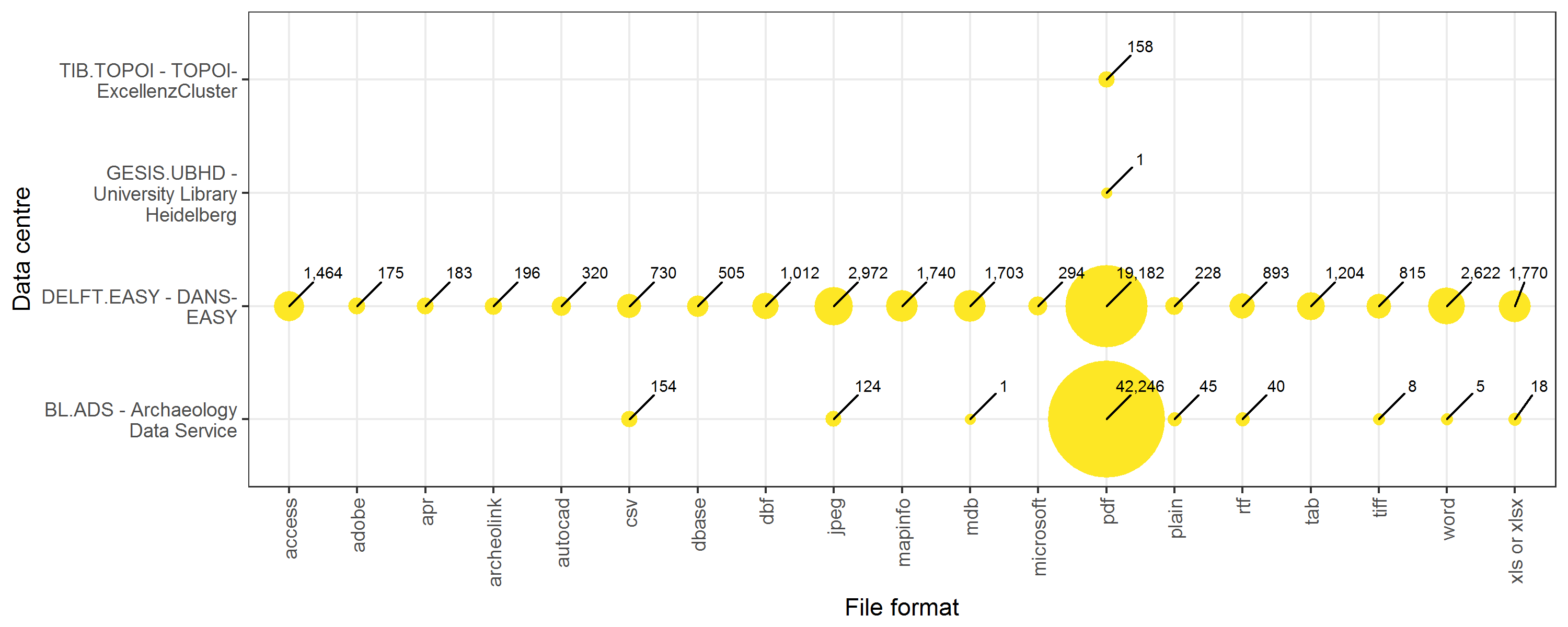


Figure 5 Relationships between data centres and file formats. The ten most frequently found file formats are shown here, in the data centres with these files from the top ten data centres by number of DOIs. Open file formats for structured data, such as CSV, are not common.

Our third investigation of archaeologists' data sharing behaviours is an analysis of all records (n = 118,311, as of August 2017) of archaeological data files in online repositories tracked by DataCite (an organisation that provides persistent identifiers, or DOIs for research data). The majority of records come from the Archaeology Data Service (ADS) in the UK, followed by the Dutch Archiving and Network Services (DANS-EASY). Although the earliest data dates to 1640, file deposition started and steadily increased during 1990-2008, as personal computers became more affordable and data easier to digitize. There is a substantial increase in 2009 when the DANS-EASY repository saw in influx of records from Dutch CRM firms. This relates to the introduction of Dutch archaeology legislation, the *Wet op de archeologische monumentenzorg* (Wamz), in 2007 that formally obligated archaeologists in the Netherlands to archive their data via DANS within two years of completion of their projects (Keers et al. 2011). Another spike occurs in 2012 when tDAR had a large increase in deposits, likely due to them waiving their deposit fees for that year (Ellison 2012). After 2009 the volume of deposits declined in many repositories, with the exception of generic repositories such as Figshare, Zenodo and ResearchGate, which continued to grow.

Of the 75,632 records with copyright information, the majority are released under the ADS conditions of use (n = 43,661, 58%), which is broadly similar to the CC-BY-NC (Moore and Richards 2015). The next most frequently used condition is the EU Open Access license (n = 22,198, 29%), similar to the CC-BY license. Creative Commons licenses such as CC-BY, CC-0 and others are used by less than 3% of records.

Many of the records are grey literature such as unpublished reports, produced by commercial consultants or research projects. Of the 66,121 (56%) records that include information about the file types in the deposit, 86% (n = 56,935) contain only PDFs. Structured data files such as spreadsheets, databases or shapefiles are found either alone or together with other file types (typically PDFs) in 7% (n = 4,481) records. The highest proportion of records with structured data files are found in the DANS-EASY repository (n = 4,304, 14%), followed by the ADS with 0.40% (n = 174). Note that file format information is not present for all records in this sample, and some of the smaller repositories do not supply this information to DataCite at all, so it is possible that the true proportion of structured files in these records is higher than observed here.

From this pilot study of repository data we see that policy mandates in the UK and the Netherlands have been effective in populating repositories with archaeological data. With the 2012 spike in records deposited in tDAR we also see that archaeologists are sensitive to the price of sharing data. Nearly all records are made available under conditions that permit others to use the data while requiring that the original author of the data receive credit in reuse. In most cases this is because the repository hosting the data requires the author to choose one of these permissive licenses as a condition of making the deposit. However, although many items are *able* to be reused because of the licenses, the ease of reuse of these items is low. Only a small fraction of the items deposited are machine-readable structured data files. The majority of items are PDFs, which require extensive manual handling to extract tabular data for convenient reuse and combination with other datasets.

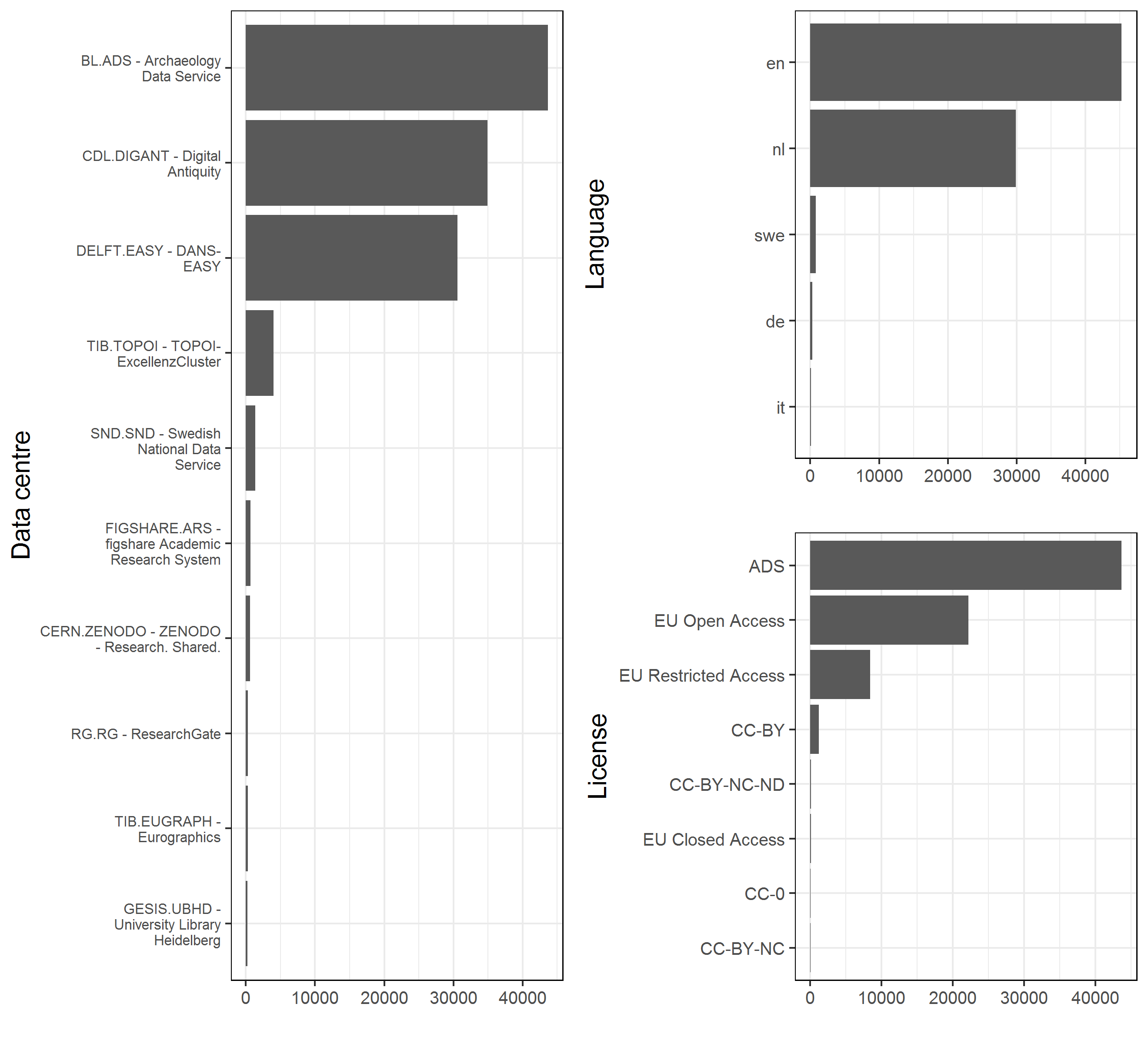


Figure 6 Distributions of data centers, languages (en = English, nl = Dutch, swe = Swedish, de = German, it = Italian) and licenses (see text for details) among archaeology datasets known to DataCite. Horizontal axis is the number of DOIs.

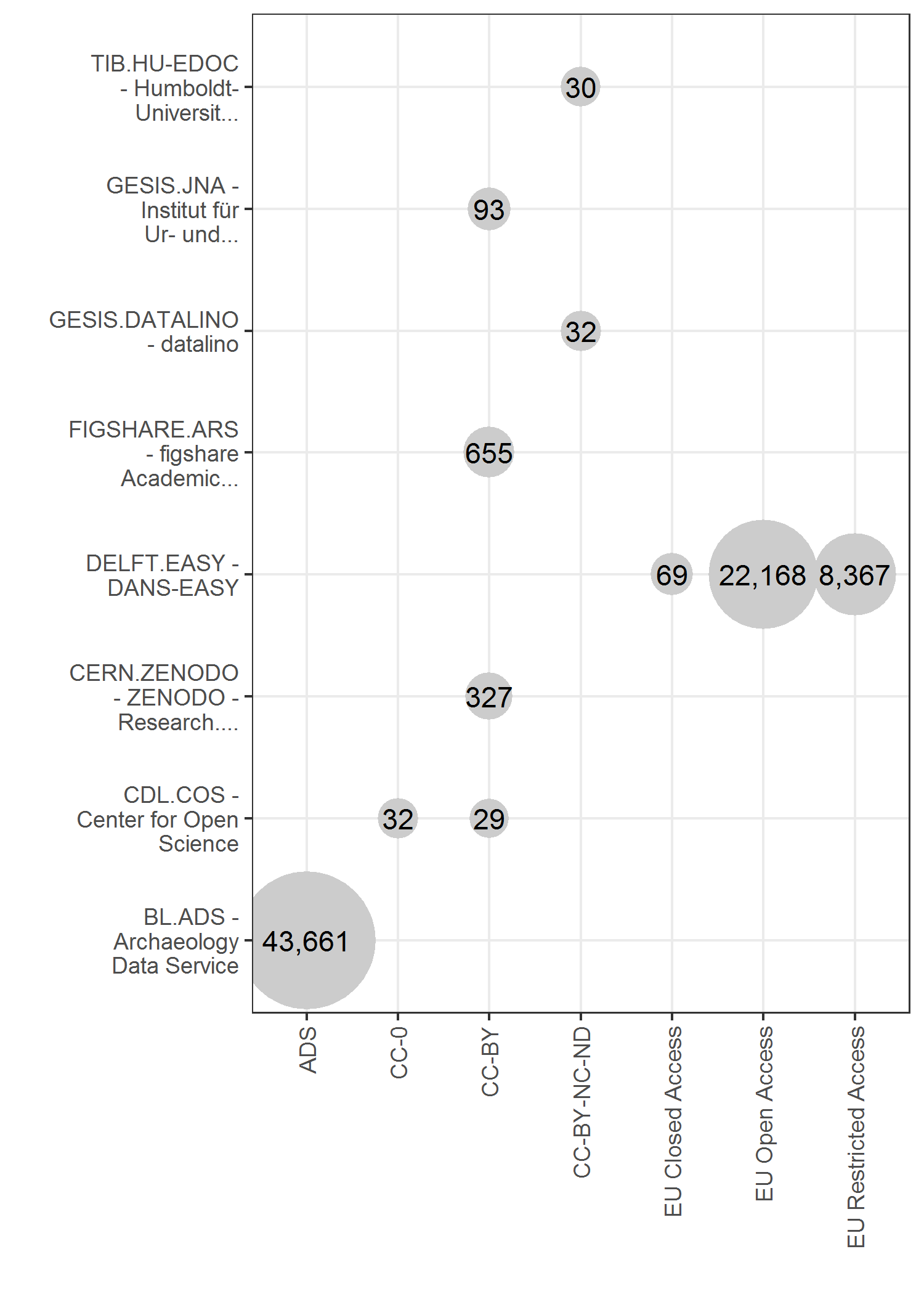


Figure 7 Summary of licenses used by data centers with archaeological data

# How to share archaeological data?

In the pilot studies reported above we see three common approaches to sharing data: sending files privately by email, following a personal request; including data as supplementary files submitted for publication with a journal article; and sharing data by depositing files in a trustworthy, public, DOI-issuing repository. The third method, via a public repository, is by far the most widely recommended method both specifically by archaeologists (Kintigh 2006; Kansa et al. 2011; Kansa and Kansa 2013), and in other sciences (cf. Cranston et al. 2014; Penev et al. 2017; Mounce 2014; Thessen & Patterson 2011; Whitlock 2011). List of trustworthy repositories are maintained by the editors of the journal *Scientific Data* (n.d) and <http://www.re3data.org>, a global registry of research data repositories (Pampel et al. 2013). There are many guidelines and conventions from other disciplines that are relevant to sharing archaeological data (ICPSR 2012; Strasser et al. 2012; Wilkinson et al. 2016). Table 3 summarises some of these, and recommendations specifically written for archaeologists can be found in UK Archaeology Data Service (n.d.), ARCHES project. (n.d.). Many of these guidelines may create new technical and practical challenges to archaeologists who are not accustomed to sharing data, and many not want to dedicate resources to sharing data. However, we should not let the perfect be the enemy of the good. In our own work we try to follow these guidelines with no dedicated resources for data sharing, and with an incomplete knowledge of some of the technical details of repository operations. While we recognise that our data sharing practises fall short of all the guidelines, we still see value in sharing data imperfectly, rather than not at all.

Data sharing is a relatively new behaviour for most archaeologists, so it is important to clarify what does not count as data sharing. Making data available on a web site that is not a trustworthy, persistent-identifier-issuing repository is not data sharing. For example, data files hosted on a personal faculty webpage or the web server of a research project is problematic because these files could disappear at any time, the data can change in untracked ways, and published links to the files are broken when these websites are reorganised, making the files inaccessible. Sharing data via supplementary files attached to a journal article is problematic because they are inaccessible to non-subscribers for paywalled journals, and because publishers often alter the data file formats and file names during the production process. For example, machine-readable format such as CSV are sometimes presented as PDF or Microsoft Word files in supplementary files. The modification of supplementary file names during the article production process can break scripts that attempt to read files by referring to file names as they were prior to the publisher handling the files. Finally, authors who write in their publications that 'the data are available by request' are not sharing data. Not only are the data not actually available, but there is no way to enforce that the authors will actually honour any requests for their data. A simple strategy for avoiding these pseudo-sharing behaviours is to deposit data files in a trustworthy data repository, and use persistent links to the files in reports and manuscripts.

# How to cite archaeological data?

The Joint Declaration of Data Citation Principles is the current standard in many disciplines for data citation within any research object (Data Citation Synthesis Group 2014). The result of these principles is a citation formation as follows (the exact order of the items may vary slightly across different referencing styles):

Creator (PublicationYear). Title. Version. Publisher. Identifier

The 'PublicationYear', 'Title' and 'Publisher' are familiar from traditional citation formats, except that the publisher may not be a traditional book or journal publisher, but a dedicated data repository organisation. 'Creator' is equivalent to the author or primary collector and arranger of the data (e.g. Open Context or tDAR). The 'Version' and 'Identifier' fields are unique to data citation, and can take a variety of forms. The identifier is a special type of persistent internet address. The identifier must have two important qualities: the address contains some metadata (such as the identity of the publisher, but not necessarily in a human-readable format), and they remained fixed to a specific file or bundle of files over the lifetime of those objects, even if their location on the internet changes. There are several varieties of these in use, with DOIs (digital object identifiers) currently the most common form. When a data publisher changes the internet address of a data set, they are responsible for informing the DOI-issuing authority so that the DOI resolves to the new location of the files.

A DOI consists of three parts:

* the proxy, usually <https://doi.org> but sometimes also <https://hdl.handle.net/> or just doi:
* the prefix, usually a string of numbers that is unique to the organisation publishing the DOI. For example, tDAR prefix is 10.6067
* the suffix, a random-looking string of any combination of characters and numbers that uniquely identified the file or bundle of files, this is also known as a local identifier (local ID) or an 'accession', an identifier which, by itself, is only guaranteed to be locally unique within the database or source (McMurry et al. 2017)

A DOI alone is not sufficient for citing a dataset because, like many things on the internet, the files can change over time. For example we might upload data files to the repository when we submit a paper for peer review, then get a DOI and cite that in our manuscript so the peer reviewers can see them (some services provide a private URL that can be shared with reviewers but is not publicly accessible or discoverable by a search engine). Then when we resubmit our manuscript after peer review, we might update the repository to include additional data files relating to additional analyses requested by the reviewers. We still have the same DOI in our manuscript for this data set, but the files have changed because of the revisions we did in response to the peer reviewers' comments. Most data repositories will keep each version of the files that you upload, so in this scenario there will be at least two versions of the files in our data repository. Even if we do not follow this exact sequence of updating the files, we may still change them in the future, after our paper has been published. Because of this potential to change the files stored at the DOI, it is necessary to identify specific file versions in our citation of the data. A simple method to record the version is to include the date that the data files were accessed, as when citing any other resource found on the internet. Some repositories also provide DOIs that have a version number suffix like '.v2', '.2', '/2'. However this is not formally a functionality of the DOI system, so including the date accessed is a more robust and general method of indicating the version. Technological solutions to issues of transience, permanence and granularity in data citation are still in development.

When citing data in the text of a manuscript or report, the ideal practice is to cite the dataset and the traditional publication that first described it, like so (Penev et al. 2017):

"This paper uses data from the [name] data repository at <http://doi.org/xxx/xxx> (Author YYYY), first described in Author YYYY"

For example:

"This paper uses data from the Open Science Framework data repository at <http://doi.org/10.17605/OSF.IO/32A87> (Marwick 2017), first described in Marwick et al. 2017."

Citing both the dataset and the publication helps to resolve ambiguity about the source of the data, and gives context to the data. Furthermore, the authorship of the data files may differ from those of the author's paper describing the data, and it is important that both authors (or sets of authors) receive credit for their work through the normal mechanisms of citation recognition.

In the paper in which these data are first described, the citation statement should be included in the body of the paper, in the Material and Methods section. In addition, the formal data reference should be included in the paper's reference list, using the recommended journal's reference format. For example, here is a statement that would appear in the paper in which the data are first presented:

“The data underpinning the analysis reported in this paper were deposited in the Open Science Framework Repository at <http://dx.doi.org/10.17605/OSF.IO/32A87> (Marwick 2017).

With the reference list of that paper including Marwick (2017) as:

Marwick, B. (2017). Dataset and R code for "Movement of lithics by trampling: An experiment in the Madjedbebe sediments", <http://doi.org/10.17605/OSF.IO/32A87>, Open Science Framework, Accessed 7 Sept 2017.

# What are the ethics of data sharing?

A full treatment of the ethical issues of data sharing is beyond the scope of this paper, but we can highlight some points specifically relevant to the question of when *not* to share archaeological data. Much of the discussion on ethics and data sharing generally relates to appropriate attribution (assigning credit for the data) in data reuse (Duke et al. 2013). The default position for archaeological data, like data from many other sciences, is that they are part of the global human patrimony, and should be accessible to other scholars and the public (Zimmerman et al. 2003; Vitelli & Colwell-Chanthaphonh 2006). However, archaeologists, like other researchers working with local and indigenous communities, governments, and international collaborators, generate research data that is intellectual property with many stakeholders. With diverse stakeholders, archaeological data can become entangled in economics, nationalism, cultural politics, and identity (Nicholas and Wylie 2009). Often, one of these stakeholders is an indigenous community, and archaeologists may find differences between Western and indigenous conceptions of what data are, how they may be generated, used, shared, and 'owned' (Harding et al. 2012).

One of these differences is that Western researchers often have individual autonomy in sharing data, and once data are shared, it is free for all to use. In contrast, within indigenous communities, data are part of the group's identity and property, akin to tangible heritage (Nicholas 2014), with certain members entrusted to keep that data on behalf of the group, and sometimes it may be inappropriate for these people to share the data with other members of the group, or beyond the group (Tsosie 2007). Differences between Western and indigenous concepts or data are not only cultural - substantial inequalities exist in access to resources for converting archaeological data into economic and political benefit (e.g. graduate degrees, employment, publications, grants, promotions). On one hand, making archaeological data openly available benefits indigenous communities, because it removes some of the financial and technical barriers to accessing that data (e.g. subscription fees to journals). On the other hand, this may be little more than a useless courtesy (Hymes 1972), because simply making data openly available does not give indigenous people access to the contexts where the data can be used to generate economic and political benefits. Thus, open data has different values for Western and indigenous communities, and if archaeologists want to make a compelling case to their indigenous partners for data sharing, they should also make a commitment to reducing some of the inequalities in making use of the data, for example, by providing some training in data analysis, or adapting the data into a context that is valued by the indigenous community.

In short, there is more to the ethics of data sharing than avoiding risk of damage. Redacting the locations of archaeological sites to prevent looting is an obvious case of preparing data for sharing to reduce the risk of damage to the sites (and in some jurisdictions, mandated by law). However, because of cultural, economic and political differences between Western archaeologists and the communities they work with, an archaeologist cannot consider their own risk assessment to be comprehensive and exhaustive (Nicholas and Wylie 2012). It is crucial for archaeologists to discuss data sharing plans during negotiations with representatives authorized by the indigenous peoples whose cultural heritage is the subject of investigation. This negotiation process is important because only indigenous peoples themselves can identify potential adverse outcomes to data sharing, and they can do this only if they understand the proposed research and anticipated results. The concrete outcome of this process is often a Memorandum of Understanding between the indigenous community and the archaeologist's employer that documents the outcome of negotiations about data sharing. If the archaeologist's indigenous partners have reasons not to share data in public, then the archaeologist should respect that position, and note in publication that the data are restricted at the request of the indigenous collaborators.

# Summary and Conclusion

The status of research data is slowly changing in many fields from a gift-culture where data are privately-held commodities traded for co-authorship, invitations, and other status tokens (Hagstrom 1982; Wallis et al. 2013), to part of the public commons, accessible to anyone with an internet connection. This change has been fastest in data-rich 'big science' fields, such as astronomy, physics, and oceanography, that use expensive instruments that generate data shared with large teams of collaborators to produce new discoveries (Reichman and Uhlir 2003). However, in 'smaller' data-poor sciences such as archaeology, where collaborative groups are small, data sharing is still mostly ad hoc and disaggregated. Our pilot studies, while small and limited in many ways, support this statement with an overall sharing rate of 20% in responses to private requests for data, and 53% of sampled journal articles with openly available data, in a wide variety of formats and locations. In examining archaeological data deposited in repositories, we found that most of the contents are PDF files of reports, and only a very small fraction of the data files are structured data such as spreadsheets. Archaeologists still have a long way to go in making their data easier to reuse by sharing data in plain text structured data formats such as CSV. We found that legislation requiring data sharing resulted in a substantial increase in the rate of data deposition, and that archaeologists are sensitive to fees charged by a repository.

Among the many challenges of increasing the rate of data sharing in archaeology and other 'small' sciences is that data sharing is unfunded, unrewarded and is only rarely required. The challenge of funding can be addressed by professional societies, such as the Society of American Archaeology, communicating to funding bodies, such as the National Science Foundation, about the importance of allocating funds specifically for data sharing and requesting a chance in policy. The challenge of requiring data sharing can also be addressed by funding bodies, but may be most effectively tackled by journal editors and editorial boards, for example by adopting the Transparency and Openness Protocols of the Centre for Open Science (Nosek et al. 2015). Changes in journal-specific submission guidelines can result in desirable changes in authors’ research practices (Giofrè et al 2017). For example, if archaeology journals adopt incentives such as badging (Kidwell et al. 2016), and require as a condition of publication that submitted articles include in the text a properly formatted citation to the data analysed in the article (or an explanation why the data are not available), authors would quickly adapt to this new requirement to ensure their article will be published. As peer reviewers, we can also communicate to editors that we require the manuscripts that we review to include citations of datasets analysed in the manuscript.

The challenge of data sharing as an unrewarded behaviour was the primary motivation for this article. By specifying a standard for the scholarly citation of archaeological data we provide a method for giving credit to the authors or compilers of the data. This rewards those authors with visibility and recognition for their effort with citations, a currency that is valued broadly in the discipline. The data citation standard outlined above is easy for humans and search engines to understand, and is highly functional, with a full internet address to the data contained in the citation. The standard can accommodate the addition of additional information required by different citation styles and publishers' requirements. The data citation standard is not dependent on any specific choices of software, archive, data producer, publisher, or author and is easy to adapt into existing writing and research workflows.

# Acknowledgements

This paper emerged out of discussions at the 2017 SAA meeting, specifically the forum ‘Beyond Data Management: A Conversation about “Digital Data Realities”’, organised by Sarah Kansa and Eric Kansa, and the SAA publications committee meeting. Thanks to Canan Cakirlar for insights into the history of archaeological data sharing in the Netherlands. BM was supported by an Australian Research Council Future Fellowship (FT140100101).

# References

Alsheikh-Ali AA, Qureshi W, Al-Mallah MH, Ioannidis JPA (2011) Public Availability of Published Research Data in High-Impact Journals. PLOS ONE 6(9): e24357. <https://doi.org/10.1371/journal.pone.0024357>

American Psychological Association. (2001). Ethical standards for the reporting and publishing of scientiﬁc information. In Publication manual of the American Psychological Association (5th ed., pp. 387–396). Washington, DC

ARCHES project. (n.d.). The Standard and Guide to Best Practice in Archaeological Archiving in Europe. Retrieved September 7, 2017, from <http://archaeologydataservice.ac.uk/arches/Wiki.jsp?page=ARCHES%20Introduction>

Bierer, B.E., Crosas, M., Pierce, H.H., 2017. Data Authorship as an Incentive to Data Sharing, New England Journal of Medicine. 376(17):1684-1687

Borgman, C. L. (2015). Big data, little data, no data: Scholarship in the networked world. MIT press.

Borgman, C. L. (2012). The conundrum of sharing research data. Journal of the American Society for Information Science and Technology, 63(6), 1059–1078. <doi:10.1002/asi.22634>

Chambers, C. (2017). The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice. Princeton, NJ: Princeton University Press.

Clarke, Mary. 2015. The Digital Dilemma: Preservation and the Digital. Archaeological Record. Advances in Archaeological Practice. 3(4): 313-330.

Daston, Lorraine, Galison Peter. (1992) The Image of Objectivity. Representations. 40:81–128.

Dorch, B. F., Drachen, T. M., & Ellegaard, O. (2015). The data sharing advantage in astrophysics. Proceedings of the International Astronomical Union, 11(A29A), 172-175.

Duke, C. S., & Porter, J. H. (2013). The Ethics of Data Sharing and Reuse in Biology. BioScience, 63(6), 483–489. <https://doi.org/10.1525/bio.2013.63.6.10>

Editors, (2013). Guide for Authors. Journal of Archaeological Science. Retrieved September 7, 2017, from <https://web.archive.org/web/20130125222940/https://www.elsevier.com/journals/journal-of-archaeological-science/0305-4403/guide-for-authors>

Ellison, L. A. (2012). Just 20 days left!! Upload your archaeological information FREE to tDAR!! | The Digital Archaeological Record. Retrieved September 7, 2017, from <https://www.tdar.org/news/2012/12/just-20-days-left-upload-your-archaeological-information-free-to-tdar/>

Eglen, S. J., Marwick, B., Halchenko, Y. O., Hanke, M., Sufi, S., Gleeson, P., Poline, J.-B. (2017). Toward standard practices for sharing computer code and programs in neuroscience. Nature Neuroscience, 20(6), 770–773. <https://doi.org/10.1038/nn.4550>

Fienberg, S. E., Martin, M. E., & Straf, M. L. (1985). Sharing research data. National Academy Press, Washington DC.

Giofrè, D., Cumming, G., Fresc, L., Boedker, I., & Tressoldi, P. (2017). The influence of journal submission guidelines on authors’ reporting of statistics and use of open research practices. PLoS ONE, 12(4). <https://doi.org/10.1371/journal.pone.0175583>

Gleditsch, N. P., Metelits, C., & Strand, H. (2003). Posting your data: will you be scooped or will you be famous. International Studies Perspectives, 4(1), 89-97.

Hagstrom, W O (1982). Gift Giving as an Organising Principle in Science In: Barnes, B and Edge, D eds. Science in Context: Readings in the Sociology of Science. Milton Keynes, United Kingdom: The Open University Press, pp. 21–34.

Hart, E. M., Barmby, P., LeBauer, D., Michonneau, F., Mount, S., Mulrooney, P., Hollister, J. W. (2016). Ten Simple Rules for Digital Data Storage. PLOS Computational Biology, 12(10), e1005097. <https://doi.org/10.1371/journal.pcbi.1005097>

Henneken, E. A., & Accomazzi, A. (2011). Linking to data-effect on citation rates in astronomy. arXiv preprint, arXiv:1111.3618.

Huggett, J. (2017, May 23). Citing Data Reuse. Retrieved September 7, 2017, from <https://introspectivedigitalarchaeology.wordpress.com/2017/05/23/citing-data-reuse/>

Hymes, D. (1972). 1974. Reinventing anthropology. New York: Pantheon

Inter-university Consortium for Political and Social Research (ICPSR). (2012). Guide to Social Science Data Preparation and Archiving: Best Practice Throughout the Data Life Cycle (5th ed.). Ann Arbor, MI

Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L. S., & Errington, T. M. (2016). Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency. PLoS Biol, 14(5), e1002456.

Keers, G., van der Reijden, H., and van Rossum, H. 2011. Planning Archaeology: A Synthesis of the Thematic Sub-reports. Amsterdam: Rigo.

Kratz JE, Strasser C (2015) Researcher Perspectives on Publication and Peer Review of Data. PLOS ONE 10(2): e0117619. doi: 10.1371/journal.pone.0117619

Kratz J and Strasser C. Data publication consensus and controversies [version 1; referees: 1 approved with reservations]. F1000Research 2014, 3:94 (doi: 10.12688/f1000research.3979.1)

Kyzas PA, Loizou KT, and Ioannidis JP. (2005) Selective reporting biases in cancer prognostic factor studies. Journal of the National Cancer Institute 97, 1043-1055.

Leonelli, S. (2016). The philosophy of data, In Floridi, L. (Ed.) The Routledge Handbook of Philosophy of Information. Routledge, NY.

Marwick, B. (2017). Computational reproducibility in archaeological research: Basic principles and a case study of their implementation. Journal of Archaeological Method and Theory, 24(2), 424-450.

McCullough BD, McGeary KA, Harrison TD. (2006) Lessons from the JMCB Archive. Journal of Money, Credit, and Banking 38(4), 1093-1107. <doi:10.1353/mcb.2006.0061>

McCullough BD and Vinod HD. (2005) Comment: Econometrics and Software. Journal of Economic Perspectives 17(1), 223-224.

McMurry, J. A., Juty, N., Blomberg, N., Burdett, T., Conlin, T., Conte, N., … Parkinson, H. (2017). Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data. PLOS Biology, 15(6), e2001414. <https://doi.org/10.1371/journal.pbio.2001414>

Merton, R.K. (1942) "The Normative Structure of Science". In: Merton, R. K., The Sociology of Science: Theoretical and Empirical Investigations, Chicago, IL: University Of Chicago Press.

Moore, R., & Richards, J. (2015). Here Today, Gone Tomorrow: Open Access, Open Data and Digital Preservation. In Andrew T. Wilson, Ben Edwards (eds) Open Source Archaeology: Ethics and Practice. De Gruyter, Netherlands

Nicholas, G. (2014). Indigenous cultural heritage in the age of technological reproducibility: Towards a postcolonial ethic of the public domain. In Coombe, R. J., Wershler, D., & Zeilinger, M. (Eds.) Dynamic fair dealing: creating Canadian culture online. University of Toronto Press. pp. 213-24.

Nichols, D.M., Twidale, M.B., (2017). Metrics for openness, Journal of the Association for Information Science and Technology 68, 1048-1060.

Nicholas, G. P., & Wylie, A. (2012). Do Not Do Unto Others…’: Cultural misrecognition and the harms of appropriation in an open-source world. Appropriating the Past: Philosophical Perspectives on Archaeological Practice, 195-221.

Nicholas, G. P., & Wylie, A. (2009). Archaeological finds: legacies of appropriation, modes of response. The Ethics of Cultural Appropriation, 11-54.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Yarkoni, T. (2015). Promoting an open research culture: Author guidelines for journals could help to promote transparency, openness, and reproducibility. Science, 348, 1422–5.

Nuijten, M. B., Borghuis, J., Veldkamp, C. L. S., Alvarez, L. D., van Assen, M. A. L. M., & Wicherts, J. M. (2017, July 13). Journal Data Sharing Policies and Statistical Reporting Inconsistencies in Psychology. Retrieved September 7, 2017, from psyarxiv.com/sgbta

National Science Board. (2005). Long-lived digital data collections. Retrieved September 7, 2017, from <http://www.nsf.gov/pubs/2005/nsb0540/>

National Science Foundation. (2010). Data Management & Sharing Frequently Asked Questions (FAQs). Retrieved September 7, 2017, from <https://www.nsf.gov/bfa/dias/policy/dmpfaqs.jsp#1>

National Science Foundation. (2016). Dissemination and Sharing of Research Results. Retrieved September 7, 2017, from <https://www.nsf.gov/bfa/dias/policy/dmp.jsp>

Pampel, H., Vierkant, P., Scholze, F., Bertelmann, R., Kindling, M., Klump, J. & Dierolf, U. (2013). Making research data repositories visible: The re3data.org registry. PloS one, 8(11), e78080.

Peng RD, Dominici F, and Zeger SL. (2006) Reproducible Epidemiological Research. American Journal of Epidemology 163, 783-789 doi: 10.1093/aje/kwj093

Penev L, Mietchen D, Chavan V, Hagedorn G, Smith V, Shotton D, Ó Tuama É, Senderov V, Georgiev T, Stoev P, Groom Q, Remsen D, Edmunds S (2017) Strategies and guidelines for scholarly publishing of biodiversity data. Research Ideas and Outcomes 3: e12431. <https://doi.org/10.3897/rio.3.e12431>

Piwowar HA, Day RS, Fridsma DB (2007) Sharing Detailed Research Data Is Associated with Increased Citation Rate. PLOS ONE 2(3): e308. <https://doi.org/10.1371/journal.pone.0000308>

Piwowar, H. A., & Vision, T. J. (2013). Data reuse and the open data citation advantage. PeerJ, 1, e175. <https://doi.org/10.7717/peerj.175>

PLoS Medicine Editorial and Publishing Policies. 13. Sharing of Materials, Methods and Data. PLoS Medicine Editorial and Publishing Policies. Retrieved September 7, 2017, from <http://journals.plos.org/plosmedicine/policies.php#sharing>.

Reichman, J. H., & Uhlir, P. F. (2003). A contractually reconstructed research commons for scientific data in a highly protectionist intellectual property environment. Law and Contemporary problems, 66(1/2), 315-462.

Reidpath DD, and Allotey PA. Data Sharing in Medical Research: An Empirical Investigation. Bioethics 15(2), 125- 134 (2002). doi: 10.1111/1467-8519.00220

Rowhani-Farid, A., Allen, M., & Barnett, A. G. (2017). What incentives increase data sharing in health and medical research? A systematic review. Research Integrity and Peer Review, 2(1), 4. <https://doi.org/10.1186/s41073-017-0028-9>

Savage CJ, Vickers AJ (2009) Empirical Study of Data Sharing by Authors Publishing in PLoS Journals. PLoS ONE 4(9): e7078. <doi:10.1371/journal.pone.0007078>

Sawyer, S. (2008). Data wealth, data poverty, science and cyberinfrastructure 1. Prometheus, 26(4), 355-371.

Scientific Data Editors (n.d) Recommended Data Repositories. Retrieved September 13, 2017, from <https://www.nature.com/sdata/policies/repositories>

Sears, J.R.L (2011) Data sharing effect on article citation rate in paleoceanography. Presented at Fall Meeting of the American Geophysical Union. Retrieved September 7, 2017, from <https://doi.org/10.6084/M9.FIGSHARE.1222998.V1>

Spencer, Hilary (2010). Thoughts on the sharing of data and research materials and the role of journal policies. Retrieved September 7, 2017, from <https://web.stanford.edu/~vcs/Nov21/hilary_spencer_rdcscsJan2010.pdf>

Strasser, C., Cook, R., Michener, W., & Budden, A. (2012). Primer on Data Management: What you always wanted to know. EScholarship. Retrieved September 7, 2017, from <https://doi.org/10.5060/D2251G48>

Stodden, V. (2009a). The legal framework for reproducible scientific research: Licensing and copyright. Computing in Science & Engineering, 11(1), 35-40.

Stodden, V. (2009b) Enabling Reproducible Research: Open Licensing for Scientific Innovation. International Journal of Communications Law and Policy. Retrieved September 7, 2017, from <https://ssrn.com/abstract=1362040>

Stodden, V. 2010, The Scientific Method in Practice: Reproducibility in the Computational Sciences, Columbia University Academic Commons, Retrieved September 7, 2017, from <http://hdl.handle.net/10022/AC:P:11417>.

Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., … Dorsett, K. (2015). Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide. PLOS ONE, 10(8), e0134826. <https://doi.org/10.1371/journal.pone.0134826>

Tsosie, R. (2007). Cultural challenges to biotechnology: Native American genetic resources and the concept of cultural harm. The Journal of Law, Medicine & Ethics: A Journal of the American Society of Law, Medicine & Ethics, 35(3), 396–411. <https://doi.org/10.1111/j.1748-720X.2007.00163.x>

Uhlir, P.F., & Cohen, D. (2011). Internal document. Board on Research Data and Information, Policy and Global Affairs Division, National Academy of Sciences. 18 March 2011.

UK Archaeology Data Service. (n.d.). Guides to Good Practice: Main. Retrieved September 7, 2017, from <http://guides.archaeologydataservice.ac.uk/>

Vanpaemel, W., Vermorgen, M., Deriemaecker, L., & Storms, G. (2015). Are We Wasting a Good Crisis? The Availability of Psychological Research Data after the Storm. Collabra, 1(1), Art. 3. DOI: <http://doi.org/10.1525/collabra.13>

Vines, Timothy H, et al. 2014 The Availability of Research Data Declines Rapidly with Article Age. Current Biology 24(1):94-97.

Vitelli, K. D., & Colwell-Chanthaphonh, J. S. (Eds.). (2006). Archaeological ethics. Rowman Altamira.

Wallis, J. C., Rolando, E., & Borgman, C. L. (2013). If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology. PLoS ONE, 8(7), e67332. <doi:10.1371/journal.pone.0067332>

White E, Baldridge E, Brym Z, Locey K, McGlinn D, Supp S. (2013) Nine simple ways to make it easier to (re)use your data. Ideas in Ecology and Evolution. 2013;6: 1–10. doi: 10.4033/iee.2013.6b.6.f.

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. American Psychologist, 61(7), 726.

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Bourne, P. E. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, 160018. <https://doi.org/10.1038/sdata.2016.18>

Womack RP (2015) Research Data in Core Journals in Biology, Chemistry, Mathematics, and Physics. PLOS ONE 10(12): e0143460. doi: 10.1371/journal.pone.0143460

Zimmerman, L. J., Vitelli, K. D., & Hollowell-Zimmer, J. (Eds.). (2003). Ethical issues in archaeology. Rowman Altamira.

### Colophon

This document contains 15,785 words and was generated on using the following computational environment and dependencies:

#> setting value   
#> version R version 3.4.1 (2017-06-30)  
#> system x86\_64, mingw32   
#> ui RTerm   
#> language (EN)   
#> collate English\_Australia.1252   
#> tz America/Los\_Angeles   
#> date 2017-09-28   
#>   
#> package \* version date   
#> assertthat 0.2.0 2017-04-11  
#> backports 1.1.0 2017-05-22  
#> base \* 3.4.1 2017-06-30  
#> bindr 0.1 2016-11-13  
#> bindrcpp \* 0.2 2017-06-17  
#> bookdown 0.5 2017-08-20  
#> broom 0.4.2 2017-02-13  
#> cellranger 1.1.0.9000 2017-05-31  
#> changepoint \* 2.2.2 2016-10-04  
#> colorspace 1.3-2 2016-12-14  
#> compiler 3.4.1 2017-06-30  
#> cowplot \* 0.8.0 2017-07-30  
#> data.table 1.10.4 2017-02-01  
#> datacitation \* 0.0.0.9000 2017-09-13  
#> datasets \* 3.4.1 2017-06-30  
#> devtools 1.13.3 2017-08-02  
#> digest 0.6.12 2017-01-27  
#> dplyr \* 0.7.3 2017-09-07  
#> evaluate 0.10.1 2017-06-24  
#> forcats 0.2.0 2017-01-23  
#> foreign 0.8-69 2017-06-21  
#> ggplot2 \* 2.2.1.9000 2017-09-07  
#> ggrepel \* 0.6.5 2016-11-24  
#> glue 1.1.1 2017-06-21  
#> graphics \* 3.4.1 2017-06-30  
#> grDevices \* 3.4.1 2017-06-30  
#> grid \* 3.4.1 2017-06-30  
#> gridExtra \* 2.3 2017-09-09  
#> gtable \* 0.2.0 2016-02-26  
#> haven 1.1.0 2017-07-09  
#> highr 0.6 2016-05-09  
#> hms 0.3 2016-11-22  
#> htmltools 0.3.6 2017-04-28  
#> httr 1.3.1 2017-08-20  
#> jsonlite 1.5 2017-06-01  
#> knitr 1.17 2017-08-10  
#> koRpus 0.10-2 2017-04-04  
#> labeling 0.3 2014-08-23  
#> lattice 0.20-35 2017-03-25  
#> lazyeval 0.2.0 2016-06-12  
#> lubridate 1.6.0 2016-09-13  
#> magrittr 1.5 2014-11-22  
#> memoise 1.1.0 2017-04-21  
#> methods \* 3.4.1 2017-06-30  
#> mnormt 1.5-5 2016-10-15  
#> modelr 0.1.1 2017-07-24  
#> munsell 0.4.3 2016-02-13  
#> nlme 3.1-131 2017-02-06  
#> papaja 0.1.0.9492 2017-09-17  
#> parallel 3.4.1 2017-06-30  
#> pkgconfig 2.0.1 2017-03-21  
#> plyr 1.8.4 2016-06-08  
#> png 0.1-7 2013-12-03  
#> psych 1.7.5 2017-05-03  
#> purrr \* 0.2.3 2017-08-02  
#> R6 2.2.2 2017-06-17  
#> Rcpp 0.12.12 2017-07-15  
#> readr \* 1.1.1 2017-05-16  
#> readxl 1.0.0 2017-04-18  
#> reshape2 1.4.2 2016-10-22  
#> rlang 0.1.2 2017-08-09  
#> rmarkdown 1.6 2017-06-15  
#> rprojroot 1.2 2017-01-16  
#> rvest 0.3.2 2016-06-17  
#> scales 0.5.0.9000 2017-09-07  
#> stats \* 3.4.1 2017-06-30  
#> stringi 1.1.5 2017-04-07  
#> stringr \* 1.2.0 2017-02-18  
#> tibble \* 1.3.4 2017-08-22  
#> tidyr \* 0.7.1 2017-09-01  
#> tidyselect 0.2.0 2017-08-30  
#> tidyverse \* 1.1.1 2017-01-27  
#> tools 3.4.1 2017-06-30  
#> utils \* 3.4.1 2017-06-30  
#> viridis \* 0.4.0 2017-03-27  
#> viridisLite \* 0.2.0 2017-03-24  
#> withr 2.0.0 2017-09-05  
#> wordcountaddin 0.2.0 2017-05-19  
#> xml2 1.1.1 2017-01-24  
#> yaml 2.1.14 2016-11-12  
#> zoo \* 1.8-0 2017-04-12  
#> source   
#> CRAN (R 3.4.0)   
#> CRAN (R 3.4.0)   
#> local   
#> CRAN (R 3.4.0)   
#> CRAN (R 3.4.0)   
#> CRAN (R 3.4.1)   
#> CRAN (R 3.4.0)   
#> Github (rsheets/cellranger@024d5ba)  
#> CRAN (R 3.4.1)   
#> CRAN (R 3.4.0)   
#> local   
#> CRAN (R 3.4.1)   
#> CRAN (R 3.4.0)   
#> local   
#> local   
#> CRAN (R 3.4.1)   
#> CRAN (R 3.4.0)   
#> local   
#> CRAN (R 3.4.0)   
#> CRAN (R 3.4.0)   
#> CRAN (R 3.4.0)   
#> Github (tidyverse/ggplot2@c592e32)   
#> CRAN (R 3.4.1)   
#> CRAN (R 3.4.0)   
#> local   
#> local   
#> local   
#> CRAN (R 3.4.1)   
#> CRAN (R 3.4.0)   
#> CRAN (R 3.4.1)   
#> CRAN (R 3.4.0)   
#> CRAN (R 3.4.0)   
#> CRAN (R 3.4.0)   
#> CRAN (R 3.4.1)   
#> CRAN (R 3.4.0)   
#> CRAN (R 3.4.1)   
#> CRAN (R 3.4.0)   
#> CRAN (R 3.4.0)   
#> CRAN (R 3.4.1)   
#> CRAN (R 3.4.0)   
#> CRAN (R 3.4.0)   
#> CRAN (R 3.4.0)   
#> CRAN (R 3.4.0)   
#> local   
#> CRAN (R 3.4.0)   
#> CRAN (R 3.4.1)   
#> CRAN (R 3.4.0)   
#> CRAN (R 3.4.1)   
#> Github (crsh/papaja@ede6845)   
#> local   
#> CRAN (R 3.4.0)   
#> CRAN (R 3.4.0)   
#> CRAN (R 3.4.0)   
#> CRAN (R 3.4.0)   
#> CRAN (R 3.4.1)   
#> CRAN (R 3.4.0)   
#> CRAN (R 3.4.1)   
#> CRAN (R 3.4.0)   
#> CRAN (R 3.4.0)   
#> CRAN (R 3.4.0)   
#> CRAN (R 3.4.1)   
#> CRAN (R 3.4.0)   
#> CRAN (R 3.4.0)   
#> CRAN (R 3.4.0)   
#> Github (hadley/scales@d767915)   
#> local   
#> CRAN (R 3.4.0)   
#> CRAN (R 3.4.0)   
#> CRAN (R 3.4.1)   
#> CRAN (R 3.4.1)   
#> CRAN (R 3.4.1)   
#> CRAN (R 3.4.1)   
#> local   
#> local   
#> CRAN (R 3.4.0)   
#> CRAN (R 3.4.0)   
#> Github (jimhester/withr@eff4818)   
#> local   
#> CRAN (R 3.4.1)   
#> CRAN (R 3.4.0)   
#> CRAN (R 3.4.0)

The current Git commit details are:

#> Local: master D:/My Documents/My Papers/Advances in Archaeological Practice/Data sharing/datacitation/  
#> Remote: master @ origin (https://github.com/benmarwick/datacitation)  
#> Head: [7d3c426] 2017-09-16: second submit