Opportunities for Data Exchange

SEVENTH FRAMEWORK PROGRAMME

REPORT ON BEST PRACTICES FOR CITABILITY OF DATA AND ON EVOLVING ROLES IN SCHOLARLY COMMUNICATION

21 June 2012

Rachael Kotarski [a], Susan Reilly [b,*], Sabine Schrimpf [c], Eefke Smit [d], Karen Walshe [a]

[a] The British Library, 96 Euston Road. London NW1 2DB. United Kingdom
[b] LIBER – Association of European Research Libraries, Koninklijke Bibliotheek, National Library Of The Netherlands. Po Box 90407. 2509 Lk The Hague. The Netherlands
[c] Deutsche Nationalbibliothek Informationstechnik, Adickesallee 1. D-60322 Frankfurt am Main. Germany
[d] The International Association of STM Publishers, Prama House, 267 Banbury Road. Oxford OX2 7HT. United Kingdom

* Corresponding author: Susan.Reilly@KB.nl

Please cite as: Kotarski R, Reilly S, Schrimpf S, Smit E, Walshe K (2012). Report on best practices for citability of data and on evolving roles in scholarly communication. Retrieved from [URL].

TABLE OF CONTENTS

EXECUTIVE SUMMARY

With the ever increasing availability of data, the best way to ensure its sharing and re-use is becoming a prominent issue. Finding data and understanding data are the first steps in such a process and good data citation is an important prerequisite to enable this. New roles are evolving to support researchers in this process with support in managing, archiving, discovering, interpreting and citing data.

This report sets out the current thinking on data citation best practice and presents the results of a survey of librarians asking how new support roles could and should be developed. The findings presented here build on the extensive desk research carried out for the report "Integration of Data and Publication" (Reilly, Schallier, Schrimpf, Smit, & Wilkinson, Sept 2011), which identified that data citation was an area of opportunity for both researchers and libraries. That report also recounted the findings of a workshop held at the LIBER 2011 Conference in Barcelona. The workshop, based on preliminary findings on the integration of data and publications, revealed that, although libraries saw the emerging research data landscape as an opportunity, there was a real need to define future directions and the scope of the role of libraries in data exchange. The issue of data citation was also identified as a fundamental issue to be addressed when exploring the way forward. This previous work is supported here with further information gathered through extensive desk research, structured interviews and an online survey of LIBER members to explore best practice in data citation and evolving support roles for libraries.

The following is a summary of the common findings for best practice in data citation and the role of the library in data exchange.

Data Citation

Data citation follows, to a large extent, the conventions for traditional publication citation and aims at acknowledging data creators and indicating availability of data. In citing data, there are however some unique considerations, due to the particular properties of datasets. Examples of this are:

- granularity: which elements inside the datasets are being referred to,
- versioning: in case of dynamic or regularly updated data, which version is cited.

Actions from those who have an active role in reuse have the potential to improve data citation:

- location of the data: apply persistent links such as Digital Object Identifiers and accession numbers to ensure sustainable access to the cited data,
- acknowledging creators: ensure that the credit is given to those who deserve it.

Another as yet unsolved issue concerns the location of the citation in a journal article. Depending on the way the data are integrated in the article, the citation location can vary. In the case of referencing data stored in repositories, outside the article, the common practice seems to converge around inclusion in the reference list, while certain journals will also include the database accession numbers inline or in footnotes. One interviewee advocated a separate data-section in the reference list to make it easier for users to find any data citations.

The findings of this report reflect a growing need and eagerness among those involved in scholarly communication to agree new conventions that are practical for all. In order to maximize the potential for uptake of new standards and practices, those developing data citation conventions should be cognisant of current practices which have already been evolving in data-intensive disciplines. Liaison roles could help to bridge gaps and foster understanding between different communities. These roles could also help promote awareness amongst researchers of the benefits of good data citation. There is consensus that the use of persistent identifiers such as DOIs and inclusion of these in reference lists, alongside better author instructions from publishers and metrics tools to track citation of data will also help to drive good practice going forward.

However, many ask for more clarity and unambiguous definitions of new concepts such as 'authorship of data', 'a data paper', and on the exact requirements for longevity and persistency of data.

The findings relating to data citation have implications for libraries too, as they could help promote awareness, develop liaison and embedded roles and focus in the future on better ways to help researchers manage, curate and preserve data. In order to provide training to researchers, libraries will need to develop the skills and roles to fill these gaps and cater for a wide range of disciplines and data types. The opportunity to develop new skills and roles was explored in more detail with the library community themselves through the online survey of LIBER librarians.

Some of the key learnings on best practice in data citation are:

- Citations with persistent identifiers should be listed in the references/bibliography to enable tracking of citation metrics.
- Publishers need to provide guidance for authors and referees on citation of data.
    - Researchers will be further encouraged to cite data where guidelines are provided on how to do it. To further ensure that citation of data is appropriate and included in the correct section of an article, publishers need to make peer reviewers aware of citation standards and formats they implement.
- There is confusion on what persistence and longevity of data is required for it to be citeable and cited.
    - Citation metrics for data could help inform this, by providing evidence on how long after publication data is cited and reused.
- There is a lack of clarity and agreement on what 'authorship' of a dataset means
    - Contributions beyond producing a dataset need acknowledgement via authorship, but the best way to do this still need to be decided.
- Researchers need to nurture awareness in their community of the benefits of data citation, and follow citation guidelines given by publishers and data centres.

- Many researchers do not appear to see the value and benefits of data citation. How different communities can work together to promote this activity and the status of datasets as primary research outputs and publishable works in their own right, is an issue that still needs to be addressed.
- This could be led from academic societies and institutions.

Librarians Survey

The LIBER survey aimed to explore the roles that libraries should fill in support of data citation and data management. Riley et al (2011) identified several opportunities for libraries, and in a workshop at the LIBER Annual Conference in Barcelona in 2011, it was established that research libraries are keen to engage in data management. The survey was designed to gather evidence on the current and expected roles of libraries in regard to data management in order to prescribe steps for the evolution of these roles.

This was done by assessing librarians' opinions on the following questions:

1. What is the perceived demand from researchers for support for data management from libraries?
2. In what areas does this demand exist?
3. What support is currently in place?
4. What skills are needed to meet the demand for support?

In total 110 responses were gathered, from a mailing to LIBER members that reaches approximately 800 people (response rate 13%). Additional responses were gathered from a dozen internationally recognized leading libraries (experts) in the field of data management support from the US and Australia. Their responses form a comparable benchmark.

The responses to the survey make it clear that librarians regard their involvement in support for research data as a new and important role. For the majority, the service level is still rather low, but librarians also appear keen to develop themselves in the area of data management, archiving and curation in addition to helping their researchers find data.

Over 80% of respondents report demand for support in the management of data, but this demand is not yet widely met. Only 19% support researchers in creating data management plans; 29% support researchers in making the data from their research available and almost 40% provide support in citing data. 50% indicate they have no plans to start offering such support.

The situation is a bit more positive for helping to retrieve data: the most important role indicated by 84% is to support findability of datasets. Again 65% state they do not offer these services yet, but would like to start. Around 65% have no strategies in place to ensure the retrievability of datasets via DOIs or other unique identifiers, and half of the respondents say the reason for this is lack of funding.

Yet, 70% of LIBER respondents (and 100% of librarians and data managers in the US and Australia) believe that datasets will become separately citable items and will become an integrated part of enriched scientific publications.

The skills section provides some interpretation for these results: Only 12% of librarians surveyed believe they already have the right skills to be prepared for these new data activities. Another 56% are investing in developing these skills. 82% believe they need more IT skills and 80% see the need for skills in data curation and archiving. 67% believe that subject specific research experience is needed. Interestingly enough, the expert libraries regard subject expertise as the most important skill by far. Almost unanimously, the libraries consider continuing professional development the best means to develop all such skills (93.4%).

Some of the key learnings on the role of libraries in supporting data exchange are:

- There is a demand for library support in the management and discovery of data. Although these are not yet widely met, libraries are starting to look at how they can offer support.
- The majority of libraries have no strategies in place to ensure the retrievability of datasets primarily due to lack of funding.
- The vast majority of respondents believe that datasets will become separately citable yet an integrated part of enriched scientific publications.
- Librarians appear keen to develop themselves in the area of data management, archiving and curation, although few of those surveyed believe they possess the right skills for these new data activities.
- IT and data curation skills are seen as the most important skills to develop, but experience may show that subject expertise is also very relevant. Almost unanimously, libraries believe continuing professional development is the best means of developing all such skills.

Common Findings

Much of what has been learned looking at data citation bears relevance for the definition of the role of the library in supporting data sharing and reuse. Many issues face a range of stakeholders in the data sharing, management and citation landscape and thus require further dialogue and discussion across these different perspectives, to develop potential solutions. In informing next steps in enabling data exchange it is therefore useful to draw together these common findings:

1. There are simple and practical steps that all parties can take to enable easier citation and tracking of data. These emerge from building consideration of data citation requirements into existing tools and services e.g. citation metrics and bibliographic management tools.
2. The emergence of the ‘data paper’ format is evidence for the increased appetite for data reuse and citation in some subject areas. But it is unclear how they will drive academic credit, data reuse and data citation in the long term, and their applicability to more diverse disciplines.

3. Current communication between data centres, publishers, libraries and scientific communities is poor and as a result standards, guidelines, support and training relating to data, if present, may not be relevant to community practices.

   A. Liaison roles would help to mediate interactions and bridge these gaps.

   B. Improved communication is also needed to navigate and develop solutions to the remaining challenges. This will include establishing what data can and should be cited, and how data citation can best provide the appropriate acknowledgement to all those involved in the data exchange process, from creation to reuse.

4. Many researchers do not appear to see the value and benefits of data citation. There is a gap, which could be filled by libraries, in advocacy for data sharing, the use of subject specific repositories, and best practice in data citation. These, if filled, would increase the number of researchers sharing and reusing data. The issue still to be addressed is how different communities can work together to promote this activity and the status of datasets as primary research outputs and publishable works in their own right.

5. Persistent identifiers should be used to uniquely identify and address datasets. These identifiers should be allocated by data publishers (data centres, repositories, libraries or publishers). Libraries have a role in promoting and supporting the use of persistent identifiers, through raising awareness on the use and reuse of identifiers within the library and research communities, and also through ensuring that they are findable within their search services. With their expertise in metadata, libraries should also be engaging in wider discussions surrounding the use of identifiers within metadata records and the agreement of standards for persistent identification.

6. Not all of the required skill sets currently exist in libraries and the profession may need to consider new ways of developing and attracting such skills and subject expertise in order to provide researcher support and training in citation, management, curation, and preservation, of data. Further dialogue must occur, both within the library community and in consultation with other stakeholders, about the type of skills that need to be developed and to explore how prepared libraries are for these new activities.

7. Data citation is key to the successful adoption of data sharing by researchers and libraries can help address some of the issues that need to be tackled if best practice in data citation is to be implemented. If libraries are to support researchers regarding opportunities for data exchange, there is a need to increase dialogue with researchers and for librarians to become more embedded in the research process.

8. Institutional repositories may need to support researchers whose data falls outside the remit of existing subject data repositories. Libraries and information support services that manage these repositories at academic institutions require expanding data-skills and also need strategies in place to support continued access to research data.

9. Investment is needed to increase the level of data management support in particular if European libraries are to meet emerging international best practice standards. This may also require institutional strategies and mandates from funders/institutions to be in place.

1. BEST PRACTICE IN DATA CITATION

1.1 Introduction

The report "Integration of Data and Publication" (Reilly et al., 2011) concluded that six criteria are the key to ensuring the long term success of linking data effectively to publications:

a. Availability
b. Findability
c. Interpretability
d. Re-usability
e. Citability
f. Curation/Preservation

That report also set out growing trends and potential opportunities for a range of stakeholders to promote data exchange through integrating data and publications. This chapter seeks to document existing best practices in data citation, mapping the gaps, issues and exploring the potential opportunities highlighted by Reilly et al. (2011) in order to identify ways to turn best practice into reality.

1.2 Methodology

The information on current data citation practices were captured utilizing a range of methodologies:

a. Desk research to map the current landscape, review literature and catalogue advice, standards and policies.
b. Eight structured one hour interviews with key stakeholders identified through desk research as opinion leaders in the area of data citation (researchers, data centres/repositories, publishers and library/information services) to assess a variety of perspectives on:

   - how best to cite available datasets
   - how different role types can work to make data sets more easily citable
   - how improving data citation practices would benefit various communities.

c. Fifty five interviews of key stakeholders in the broader data exchange landscape to assess general views on what good data citation should look like.

1.3 Why Cite?

The research process involves building on, reusing and critiquing the published body of evidence or knowledge in a given field. Citation allows an author to acknowledge and provide an entry point for readers into the background of the process that led up to their current work. There are a number of reasons given for why authors cite particular works. Hanney et al (2005) provide a useful summary of these various motivations, for instance drawing on dimensions such as refuting or supporting that work, simply noting it, reviewing a work or because it contains information that is being applied in some way. These motivations may equally apply to all works, regardless of their format or 'container', so the motives to cite data are likely to be a sub-set of the wider pattern of citation within a publication.

1.4 Is there a problem with citation of data?

The Report on Integration of Data and Publications (Reilly et al., 2011) highlighted that the integration of data and scholarly publication will help to address the various barriers currently facing those involved in creating an environment conducive to data sharing.

These barriers give rise to a landscape currently where the maximum return on investment in producing data is not being realised. The resulting push for data openness from funding agencies (Digital Curation Centre, n.d.; Hrynaszkiewicz, 2011; Pearce & Smith, 2011), requires researchers time, effort and resources to make data truly accessible (by providing adequate metadata, curating and preserving data). So the issue of data citation comes to the fore, as researchers ask 'What's in it for me?' Data authors want a system in place so they can gain the appropriate acknowledgement for the contribution that their data has made (Nature Biotechnology Editorial, 2009).

Data citation refers to the practice of providing a reference to data, in the same way researchers provide a bibliographic reference to research articles. As data have not been published in the traditional sense, there has been less formal acknowledgement of the role of data in the published literature. An increasing demand to share data has meant that mechanisms and policies to address this gap are required. Data have been shared in a number of domains, and particularly where publicly available databases have been established (e.g. DNA sequencing), practices for referring to data have been developed by academic and publishing communities. However a widespread culture for citing data has not kept pace with data availability.

The current lack of a data citation culture can be attributed to technical issues, such as infrastructure and standards, or cultural and social issues, such as authors being unaware of data citation requirements or unsure what data to cite, how to cite it, and when and where it can be cited.

Our interviewees largely felt that the primary issues facing citation of data are not technical, but social or cultural. We therefore explored these aspects in more detail to consider what can be done to overcome social or cultural barriers, and what has already been achieved in specific subject areas that could be applied more widely. There were technical barriers highlighted in the literature however, so these are also discussed.

1.5 What should citation achieve?

There are three main aims for citing knowledge used during research (based on Altman & King, 2007; Lane, 2008). While the 'container' or form of publication may range from a book, journal article, government report, thesis or research dataset, the aims behind citing them remain the same. These are:

a. To acknowledge and give credit to the producers of previous work

> *"Traditional citation enables career credit because it implies the object can be unambiguous, identified, has provenance, has been peer reviewed, is available in the exact form, is persistent" (Lane, 2008)*

The importance of traditional publication and citation records in building a researcher's reputation, and thus helping them to secure funding and promotion provide strong incentives to maintain the status quo. It will take time for data citation to feed into recognition systems in a similar way and will require not only changes in researcher practices, but also broader practices and policies (e.g. in the way research impact is assessed).

Metrics for data citation will need to be developed; interviewees noted that metrics that are appropriate and useful from a researcher perspective would be the biggest incentive for researchers to follow best practice. Persistent Identifiers provide one mechanism to enable metrics for data citation, but data citation should be built in to existing citation indexes.

> *"It's not technical, the data is there, it just needs to be switched on" (Interviewee: Publisher)*

Citation is also currently used as a proxy for assessing the impact of published materials; the more highly cited the publication the more likely it is to be considered important within a field[1]. This could equally apply to research data in the future, where citation of data could help to indicate the value of a dataset based on its reuse and how it contributes to future findings.

Data centres and research funders could also use citation metrics to inform their own decision making and strategies – they are currently unable to tell how well used their archived data is, as tracking usage is difficult without a standard citation format and without persistent identifiers. Even where they put in a large amount of effort into tracking usage, data centres may be underestimating the amount of research that uses their data by 30% (Lane, 2008; Sieber & Trumbo, 1995).

Acknowledgement and credit for data creation was mentioned by all interviewees, as the principal potential benefit of data citation. Interestingly this was seen as a key incentive to drive best practice, not only amongst researchers but across all role types.

b. To maintain the research record and how it has developed through previous work

New research is based on understanding what has come before. Citation of existing work is key to providing an overview of how a field has progressed, and

[1] Some studies have shown that citation metrics tally with research impact, see Cronin & Overfelt (1994), despite

the way the understanding of a subject has changed over time -datasets are an equally important aspect of this record.

c. **To allow readers to find the previous work for verification and reuse (discoverability)**

The intellectual work, insights and conclusions presented in a research publication, often include an interpretation of data and so allowing others to access the underlying data can allow for alternative interpretations and hypotheses to be derived. The data must therefore be open for re-analysis and verification. Providing access to data also allows for any mistakes or inconsistencies to be checked. For this to be possible researchers need to know where to find and access these works, and need to be provided with sufficient information about the dataset and how it was generated.

Data centres were the only group to mention provenance or verification of the data as a key aim; it is interesting that of the 12 researchers interviewed, none considered this.

As well as confirming these broad aims, interviewees made reference to additional potential advantages of data citation, such as: tracking data usage; providing an alternative route to publication; enriching publications; encouraging data sharing; increasing demand for data; increasing research efficiency; assuring long-term availability of data and increasing trust in research.

It is interesting to note that researchers mentioned few of these additional benefits themselves. Of the 12 interviewed, only 25% mentioned improved credit or impact of the data cited. This is worrying, because if researchers do not clearly see the added benefits of data citation for their careers, the community, or research itself, they may see the effort needed to enable citation (both in terms of providing adequate metadata for others to cite their own data, but also the effort to ensure they properly cite someone else's data that they have used) as being a chore, for which there are no rewards. This could be combatted if funders and other governing bodies were to incorporate data citation measures in funding decisions, and in recognition and reward systems.

1.5.1 Data citation lifecycle

Citation of data can help to drive a number of positive outcomes if carried out properly. Different aspects of the process of citation fulfil the three aims mentioned earlier (Allowing: 1. acknowledgement of authors, 2. maintaining the record 3. re-use and verification).

Citation of data can help to drive researchers to share data openly and can lead to further creation, reuse and publication of data. Researchers, who gain citation credit for making their data available, are able to demonstrate the value their work adds to the knowledgebase, which in turn can assist them in obtaining further funding, so they can continue to create, publish and reuse data. The Australian National Data Service (ANDS) provide an illustration of this data citation lifecycle, encompassing these drivers[2].

[2] http://www.ands.org.au/guides/data_citation_poster.pdf

There is also a benefit for the citing author – there is evidence to suggest that a paper that links to or cites data is itself cited more often than papers that do not (Piwowar, Day, & Fridsma, 2007).

1.6 What should good data citation look like?

To achieve these benefits, citation guidelines and advice on citing data need to be provided to authors. There are already useful guides available on how to cite data, for example from the Digital Curation Centre (DCC, Ball & Duke, 2011) and the Australian National Data Service (ANDS)[3]. In general they should explain the desired mechanism for citation (e.g. use of persistent identifiers) and provide easy to follow guidance on:

- How to cite: using metadata;
- What to cite;
- When to cite: The location within a publication where a citation is to be given.

1.6.1 How to cite: using metadata

Publishers and journals advocate a range of different reference styles although the information included in a citation does not vary greatly. A summary of the different information suggested for inclusion (within a citation) in the literature and data centre and publisher's policies can be found in Appendix 1. All interviewees suggested that as far as possible the form of data citations should match that of traditional citations.

However, in contrast with journal articles, research data 'objects' can be referred to in many different ways, because the data themselves have distinct structures and formats (see discussion of versioning and granularity in 2.6.2). Data publishers vary in the way they ask for their data to be cited if they provide guidelines at all - one study found less than a third of repositories gave directions to researchers on how to cite their data (Weber, Street, Piwowar, Street, & Suite, 2011). Our interviewees pointed out that the problem is further compounded by a lack of discussion on the best way to unify differing approaches, to shape a coherent method for data citation.

Examples of ineffective data citation:

1. Taken from the journal Economic modelling (Green, 2009); in this example there is no way of locating this data or acknowledging authors.

   OECD, 2004 and CIS Statistics, 2003

2. From Hage (2011). The link to the data no longer works, although it may be possible to trace it, any authors who produced this work cannot be identified or acknowledged.

   ———. 2005. State system membership list. Version 2004.1. http://correlatesofwar.org/COW2%20Data/SystemMembership/system2004. csv (accessed January 8, 2008).

[3] http://www.ands.org.au/guides/data-citation-awareness.html

1.6.1.1 Metadata

Key to citation is the information that describes the data – metadata. There are a small number of recommended metadata schemas for research data, for example from DataCite (DataCite Metadata Working Group, 2011), the OECD (Green, 2009) and the Data documentation Initiative (DDI)[4]. A citation is essentially formed of a subset of metadata for the object.

Reference lists in scholarly publications usually include the publisher, the title of the work, the list of authors and the date of publication, and are key to discovery and access, as are markers to sections within that work, such as chapters and page numbers. These works potentially have multiple access routes, e.g. via a library hard copy, or institutional repository, a bibliographic database search or direct from the journal website.

For datasets, title and data distributor (e.g. the publisher or archive holding the data) are less helpful in finding and accessing a dataset, since there are few bibliographic resources that index or include this information. In other words, the dataset name is rarely sufficient to find it unless you know where to look – within a database or subject specific repository for example. It is therefore important for a citation to reference both the individual piece of data and location.

Typical web addresses are unreliable (Wren, 2008) for locating online resources, because they can move, change or disappear entirely. But persistent identifiers are fixed, with an infrastructure that allows for the location of the item to be updated. The result is that the identifier can provide persistent access to the data (Simons, 2012). DataCite provides such a service, and DOIs (used by DataCite) were by far the identifier most commonly mentioned by interviewees, closely followed by Handles (on which the DOI system is built). There was a keen preference for DOIs from interviewees because this is a system already used and understood by publishers for traditional publications and so the barrier to uptake would presumably be lower than for an entirely novel system. Interviewees proposed that a key requirement for good citation practice is that actionable links be created, which is possible for both DOIs and handles.

Citing a static dataset:

From DataCite recommendations:

> Creator (Publication Year): Title. Publisher. Identifier

Exemplified by:

> *Piguet, Bruno; Legain, Dominique; (2011): Tethered balloons*
> *CNRM Site 1; Météo-France, GAME.*
> *http://dx.doi.org/10.6096/BLLAST.TETHEREDBALLOONSCNRM*

Interviewees emphasised that while identifiers are important, they are only part of the solution, cultural change and perception shifts are equally important. This is highlighted by efforts in astronomy to give persistent identifiers to data that could provide a persistent, unique link to data referenced in an article. This ultimately failed due to lack

[4] See: http://www.ddialliance.org/

of researcher awareness and critical mass of publishers and data centres using the system (Accomazzi, 2011).

1.6.1.2 Acknowledgement

The authors or creators of the data are another important element of metadata that should be included in a citation. As well as providing provenance, it allows acknowledgement of those who produced the data. But there are still unresolved issues, even in conventions for listing authors:

- **Ordering**. If listed alphabetically, the author who gave the biggest contribution to a dataset may be lost somewhere in the middle, but the 'order of contribution' is a contentious point, so guidelines should be provided by data centres on how authorship is attributed.
- **Micro-attribution and attribution stacking**. The numbers of creators that need to be listed in the citation of a dataset may rapidly become impractical, especially when multiple datasets are combined to generate a new dataset. Equally, citing every single dataset separately may be unreasonable.

In some areas of practice, data providers have been requesting authorship on papers resulting from reuse of their data. While such acknowledgement would enable greater credit to be gained for creators of re-used data, Rohlfing & Poline (2012) highlight that this practice doesn't fit with the definition that journals have for 'authorship'.

One interviewee suggested a 'movie credits' approach to listing data authors. In this scenario, the 'director' or 'lead actors' would form the citation, while additional contributors would be listed in the metadata.

Many data centres also ask to be acknowledged in the citation. Data centres themselves need to justify their funding in order to continue managing and archiving research data. It is therefore important for them to be able to use citation of data they manage to demonstrate their value to the community. One organisation to look at this in depth is the Global Biodiversity Information Facility (GBIF). As a global infrastructure for aggregating biodiversity data, one of their key considerations in developing a data citation standard was:

> *"Several individuals, from the creator/collector of the data to the publishers and aggregators, play vital roles in the data life cycle, and each needs to be adequately recognized, attributed or credited"*

Their resulting recommendations not only include the Publisher, but also a 'contributor' to which all relevant persons or organisations can be included, with their role (Viswas Chavan, 2012).

1.6.1.3 Tools for citation

There are now a wide range of tools that enable easy management of references, that plug in to authoring tools, allowing authors to very quickly and simply insert citations in to their papers. Most tools have an option to add new references as a specific reference type, for instance as a 'Journal article', 'book' or 'website'. The metadata required for that reference then changes dependant on the type selected, and this is reflected in the formatted citation that is added to the bibliography during authoring.

As the use of such tools increases, ensuring that 'data' is one of the reference options would facilitate researchers in citing data more easily.

A survey of reference management tools (restricted to those that allow direct input of references during authoring), highlighted that few currently come with a data option predefined, although many allow customisable fields which could be used to format a data citation.

Software	**Data or dataset format included?**
Biblioscape	No*
Bibus	No, but users can format a 'dataset' option using custom formats (Martineau, 2005)
Bookends	No, but users can format a 'dataset' option using custom formats
Citavi	No (Meurer, Schultz, & Tejada, 2012)
Docear	No information found
EndNote	Yes ("x4 upgrade to x5, now slow," 2011)
JabRef	No, but users can format a 'dataset' option using custom formats (JabRef, 2012; Reed College, n.d.)
Mendeley	No*
Papers	Yes, as 'Database' or 'Table' (Papers for Windows, 2012)
Pybliographer	No information found
Qiqqa	No information found
Refbase	No†
Reference Manager	No, but users can format a 'dataset' option using custom formats (Thomson Reuters, 2008)
RefWorks	No (http://www.refworks.com/rwsingle/help/Reference_Types.htm)
Scholar's Aid	No, but users can format a 'dataset' option using custom formats (Shapland, n.d.)
Sente	Yes, as 'Data file' (Third Street Software, 2011)
WizFolio	No (Appleby & Leroux, 2011)
Zotero	No (Zelle, 2012), although it has been requested and flagged for development ("Issue #22: Data Set" 2011)

Table 1. Survey of bibliographic tools that support input of new citation records for research data. Only those tools that allow input of references during authoring in Microsoft Word or Open Office were surveyed, "Comparison of reference management software" (Wikipedia, n.d.).

* based on download and use of tool.

† based on use of demo database, http://demo.refbase.net/index.php

1.6.2 What to cite

As mentioned previously, data 'objects' have distinct structures and formats, and even within a single discipline can be very heterogeneous. Depending on the research in question, it can be: image data; numeric data; textual data or physical object; a small number of data points or millions of records generated from a single instrument, thousands of sensors or a single person making observations. Given this variability, it may be difficult for researchers to identify what data they need to cite.

The issue, summarised well by Laura Wynholds (2011), is the idea of data identity. The identity of cited data needs to be explicit in order to allow verifiability and reproducibility of the research and to allow accurate citation of the object. However, the definition and boundaries of a 'dataset' may change across disciplines and methodologies and from one researcher to another.

Two particular issues that influence the identity of a dataset are versioning and granularity.

1.6.2.1 Versioning

In the spirit of reproducibility, when data is cited within an article, exactly the same data should be available for future researchers to validate the work. However, it is not always possible to ensure that a dataset remains fixed in this way.

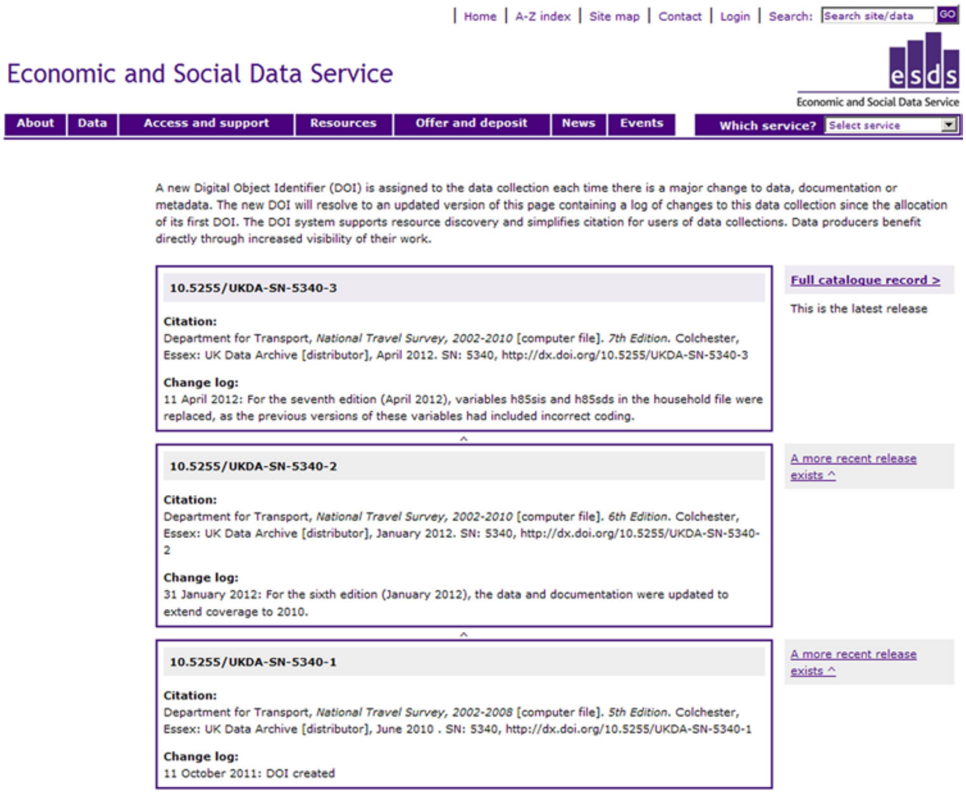
Digital datasets may change for a variety of reasons: the original data may be updated for new methodologies or technology refinement, or the data may be gathered on an incremental basis (such as observational time series) ad hoc or on a yearly, monthly, or even daily, basis.

Within social sciences for instance, longitudinal studies may collect data over a number of years. In such cases, social science data archives such as GESIS or the UK Data Archive (UKDA) store each addition or 'wave' of data as an individually citable object.

Example of citing different waves of data from GESIS

> *Förster, Peter (2012): Saxonian longitudinal study - wave 15, 2001. GESIS Data Archive, Cologne. ZA6233 Data file Version 1.0.0, doi:10.4232/1.11310*
>
> *Förster, Peter; Brähler, Elmar; Stöbel-Richter, Yve; Berth, Hendrik (2012):*

From the example above, it is not possible to tell from the citation of data from wave 15 in 2001 that data exists for 2010 (wave 24). If we imagine a situation where a dataset is cited, then subsequently updated, a researcher may follow the citation to access the data but wish to use the latest version available. To enable this, the UKDA provides links to both older and more recent versions on the access page for every dataset. Each version in this instance is still available as a unique, citable object, and so each can be cited separately, without reference to the other, but users trying to access the data will clearly see what else is available.


| Home | A-Z index | Site map | Contact | Login | Search: Search site/data GO
Economic and Social Data Service
esds
Economic and Social Data Service
About
Data
Access and support
Resources
Offer and deposit
News
Events
Which service? Select service

A new Digital Object Identifier (DOI) is assigned to the data collection each time there is a major change to data, documentation or metadata. The new DOI will resolve to an updated version of this page containing a log of changes to this data collection since the allocation of its first DOI. The DOI system supports resource discovery and simplifies citation for users of data collections. Data producers benefit directly through increased visibility of their work.

10.5255/UKDA-SN-5340-3

Citation:
Department for Transport, *National Travel Survey, 2002-2010* [computer file]. *7th Edition*. Colchester, Essex: UK Data Archive [distributor], April 2012. SN: 5340, http://dx.doi.org/10.5255/UKDA-SN-5340-3

Change log:
11 April 2012: For the seventh edition (April 2012), variables h85sis and h85sds in the household file were replaced, as the previous versions of these variables had included incorrect coding.

Full catalogue record >

This is the latest release

10.5255/UKDA-SN-5340-2

Citation:
Department for Transport, *National Travel Survey, 2002-2010* [computer file]. *6th Edition*. Colchester, Essex: UK Data Archive [distributor], January 2012. SN: 5340, http://dx.doi.org/10.5255/UKDA-SN-5340-2

Change log:
31 January 2012: For the sixth edition (January 2012), the data and documentation were updated to extend coverage to 2010.

A more recent release exists ^

10.5255/UKDA-SN-5340-1

Citation:
Department for Transport, *National Travel Survey, 2002-2008* [computer file]. *5th Edition*. Colchester, Essex: UK Data Archive [distributor], June 2010 . SN: 5340, http://dx.doi.org/10.5255/UKDA-SN-5340-1

Change log:
11 October 2011: DOI created

A more recent release exists ^

Figure 1: Screen capture showing UKDA access page for data, with linked versions of a dataset. http://dx.doi.org/10.5255/UKDA-SN-5340-3

However, it may not always be possible for a data archive to maintain each and every instance of a dynamic dataset. Take for example large scale environmental science datasets, such as the International Argo Project, where data is collected by over 3000 sensors throughout the world's oceans, generating over 100,000 data profiles per year (The Argo Science Team et al., 2003; Turton, 2003). Storing separate versions of data on this scale would require impractical volumes of storage.

There are currently two existing approaches to citing large-volume dynamic data. The first approach, as taken by the British Atmospheric Data Centre (BADC), is to only assign an identifier (making a dataset more easily citable) when a dataset is considered 'complete' and unchanging. In this case, there are no different versions of the data, and only the complete product. This may raise issues where data collection is on-going over long time periods, and makes the data less citable for researchers over the duration of data collection.

When data is updated regularly, the second option is to maintain the 'base' data and later changes e.g. data added through specific time periods are identified as separately citable objects (Altman & King, 2007). In this case, each data run would not be updated and the data would then be cited as a combination of the base data and later changes.

But when datasets change, what criteria should be applied to define a 'new' version? The National Snow and Ice Data Centre suggest that individual stewards should decide what constitutes a major or minor version, but generally changes that affect the dataset as a whole would constitute a new version (Federation of Earth Science Information Partners, n.d.). The UKDA have clear guidelines on what they consider to be a 'low impact' update that does not require a new version, and 'high impact' changes that will require the generation of a new version. High impact changes include miscoded data, addition of new variables or data series and file format changes (Corti & Bolton, 2012).

There are options available for ensuring that versioned data can be cited accurately, but factors including data structure and purpose can influence the appropriate method to use. One interviewee suggested that versioning is not an insurmountable issue in terms of data citation, but expectations need to be managed with regards to what data a citation relates to. As such, decisions governing practice in this area should be made by researchers and data centres which understand the issues and can establish community norms and methods for versioning and citation of versioned data.

1.6.2.2 Granularity

Data centres also need to agree with their user communities the level of granularity at which data should be cited. This applies particularly to datasets that have multiple levels of organisation e.g. structured databases. It is important to represent the granularity of data in citation, as reproducibility and findability of a dataset can be negatively affected by not doing so. .

An example of this is given by Buneman & Harmar (2006), who discuss the granularity of structured databases, particularly with reference to the IUPHAR (International Union of Basic and Clinical Pharmacology) Database[5]. They give three examples, where citations may refer to the database as a whole, to a discreet set of records within the database and to a specific data record. This same analogy can be drawn from other data sources, taking Pangaea as an example:

Examples of citing different layers of organisation in PANGAEA

> *Citing data from a specific cruise: Haardt, H; Maaßen, R (1983): Physical oceanography from the Drake Passage and Bransfield Strait during Meteor cruise M56. Institut für Angewandte Physik, Christian-Albrechts-Universität, Kiel, doi:10.1594/PANGAEA.737666*
>
> *Citing a single data profile from a cruise: Haardt, H; Maaßen, R (1983): Oceanographic and optical profile at station M56_127-235. doi:10.1594/PANGAEA.80634*

Buneman and Harmar suggest *"It should be possible to cite a database at varying degrees of coarseness. This does not mean that we need to cite a database at all levels of coarseness; rather that the citation system should allow more than one level if needed".*

Since citation at varying levels of granularity within the data requires explicit indication of the structure of a data centre and its datasets, the data centre would need to take responsibility for designing a citation format for each level of data available.

Altman & King (2007) refer to the granularity issue as 'deep citation' and suggest on a very simple level, that subsets of data can be referenced by citing the dataset as a whole and describing the subset within the main text. Similarly certain elements of journal

[5] http://www.iuphar-db.org/

articles are referenced in this way at present (for example, to cite a figure within an article, the figure number is given in the main text and the article is cited as a whole).

Blending these two approaches, the Digital Curation Centre (DCC, Ball & Duke, 2006) suggest that datasets should be cited *"at the finest-grained level available that meets your need. If that is not fine enough, provide details of the subset of data you are using at the point in the text where you make the citation".*

The Global Biodiversity Information Facility (GBIF) is looking at an alternative to current solutions to citation that account for both versioning and granularity. The GBIF provide access to data with a high level of complexity, both in its structure, but also in terms of who manages it.

Some data has been generated by hundreds or even thousands of data creators and is in turn transformed, curated and preserved by multiple individuals, authors and data managers. Some data is dynamic, where observations of species may be added on a very regular basis to some of the contributing datasets, while others are static. Given the complexity of this data, GBIF are looking at a system where by authors would cite data utilising the search query they used to obtain the data through the GBIF interface (by species for example). However, this approach has intrinsic technical challenges in enabling a user to rerun a query to yield the same results at some point in the future.

1.6.2.3 Verification

Given the issues of versioning and granularity in data citation, it is important for researchers to be able to verify that data are identical to that cited in a publication. Altman & King (2007) suggest a verification method, by way of a Universal Numeric Fingerprint (UNF). The UNF algorithm generates a short character string unique to the data that summarises the content of the data and is format independent. The UNF will be different if any element of the data itself changes, but not if it is simply moved between software programmes or operating systems. This allows the data to be verified by cross-referencing the UNF provided within a citation with a UNF newly generated from the data. It can also be freely shared even when there are privacy or anonymity concerns over the data, as it is impossible to reverse engineer a dataset from the UNF.

Although Altman and King recommend the inclusion of a UNF within a data citation, we have found no evidence that this has been widely adopted. Although interviewees did highlight the need for citation to cite unchanging data, none highlighted any method of verification.

Example of citing a dataset, including a UNF for data verification

> *Richard Jessor; Shirley L. Jessor, 1991, "Socialization of Problem Behavior in Youth, 1969-1981", http://hdl.handle.net/1902.1/00782 UNF:3:bNvdfUO8c9YXemVwScJy/A== Murray Research Archive [Distributor] V2 [Version]*

1.6.2.4 Citation of data vs. data papers

Data can be 'published' natively, as a dataset, or may have a companion 'data paper'. A data paper is defined by Chavan and Penev (2011) as:

> *"a scholarly publication of a searchable metadata document describing a particular online accessible dataset, or a group of datasets, published in accordance to the standard academic practices.*
>
> *A data paper is a journal publication whose primary purpose is to describe data, rather than to report a research investigation. As such, it contains facts about data, not hypotheses and arguments in support of those hypotheses based on data, as found in a conventional research article"*

There are a number of journals publishing data papers already (e.g. Journal of International Robotics, Ecological Archives, Earth System Science Data, CMB data papers, BMC Data Notes) but there are few recommendations available on when a data paper should be cited over the dataset itself. Two interviewees, from publishing and data centre backgrounds proposed that citation of a data article should carry more prestige for its authors than citation of a dataset. They reasoned that the data article can provide more evidence of quality of a dataset, and richer metadata to enable re-use than is the case when simply making the data available through an archive. It may also provide an opportunity for peer review of the data. The citation of a data paper was described as a 'gold citation' as compared to a 'silver citation' for data in an archive.

As a data paper itself cites the dataset (making it discoverable for reuse), interviewees did not think that the archived dataset needs be cited as well. There isn't clear agreement on this, with other data archives requesting the contrary. For instance, The Dryad Repository requests that you "cite both the original article, as well as the Dryad data package". It should be noted that the 'original article' may not always be a data paper[6].

1.6.3 Where to cite: Location of citation within the publication

Ideally all relevant preceding research should be referenced and cited in a publication, such that, the specific part each played in the production of new knowledge is clear. In thinking about the appropriate location within the publication for data citation, it is therefore useful to consider and distinguish between different 'levels' of data (as outlined in the data publication pyramid from Reilly et al. (2011) and also whether authors are presenting and citing their own data relating directly to, and supporting, the research findings and conclusions in their research paper; or citing other people's, data.

The policies from a publisher perspective and the practical implementation of those policies may vary in these two cases.

[6] Taken from http://www.datadryad.org/using accessed 20/04/2012

1. **Data contained and explained within the article**

The National Information Standards Organization (NISO) provide the following advice(NISO Business Working Group, 2012):

> *"Any citation to Integral Content should cite the article as a whole. Citing the content separately is not good practice. Integral Content may be assigned a unique DOI to support linking from the article to the content. One approach might be to create a parent-child DOI structure by adding a suffix to the article DOI"*

1. **Data supplementary to articles**

Here the NISO guidance above also holds: citation of supplemental files is with citation of the full article, of which it is regarded to be a part. Some publishers have become more restrictive about the data that can be included within supplemental files to ensure its long term preservation (Maunsell, 2010) and reduce the burden of reviewing large supplemental files on peer reviewers.

2. **Referenced data in archives and repositories; data publications**

This category of data is held externally to the journal/publication and may relate to the authors own data or data produced by others. This category is perhaps where citation of data can particularly help with discovery, by linking data, which may be hard to find, with publications that provide context to the data (and the importance of bi-directional linking between the two was emphasised by interviewees). In many author instructions, authors are now encouraged to deposit their data in community-endorsed data repositories and add the accession numbers or DOI's from those repositories to their manuscripts. Instructions do vary however, some recommend inclusion in the reference list, others in-line in the article. When an article (in a data journal) is available that describes a dataset and its creation in more detail, this should be cited as well as giving a citation to the actual dataset (this was requested by interviewees and Penev et al., 2011).

3. **Data in desk drawers, desktop computers and data sticks**

Although data is available at this level, it is very difficult to adequately cite it. The main problem is the difficulty in linking, or providing access, to, unpublished data. Because of this, referencing data at this level in the pyramid may be more akin to citing a 'personal communication' as a source. Interviewees did highlight that greater credit for data sharing may encourage sharing behaviours (see also Sinnott, Macdonald, Lord, Ecklund, & Jones, 2005).

Many interviewees advocated citation of datasets as part of the full reference list, in order to ensure the citation is counted for purposes of tracking and credit. One interviewee suggested data citation in a separate (dataset) section of the reference list to enable specific and separate data citation indexing.

1.7 Different perspectives

Opportunities in data citation for different communities were presented in Reilly et al. (2011). Below, we expand on these, describing the opportunities in further detail, based on the findings reported here from the literature and interviews.

Perspective	Data citation opportunities
Opportunities for researchers	• Agree a convention for data citation • Follow metadata standards for datasets • Use of persistent identifiers such as DOIs
Opportunities for publishers	• Establish uniform data citation standards • Follow metadata standards for datasets • Use of persistent identifiers such as DOIs • Data Publications
Opportunities for data centres	• Engage in establishing uniform data citation standards • Support and promote persistent identifiers

Table 2. Opportunities for data citation from the perspectives analysed in Reilly et al. (2011).

1.8 Opportunities for researchers

1.8.1 Agree a convention for data citation:

In a handful of disciplines researchers have driven development of an agreed convention for data citation, due to the need to share data. For example, disciplines that rely on observational data, particularly when it takes place on a massive and shared scale (satellites, weather or marine data), or is very costly (large hadron colliders), have developed clear common practices regarding how to handle and refer to the available data.

In the biosciences, practice is advanced in relation to data sharing and research papers clearly refer to data available in the most popular data repositories (GenBank, WPDB, UniProt, ENA etc). The bioscience community was the force behind establishing the Bermuda Principles, which require gene sequences to be publicly released within 24 hours generation. These principles were driven by the large amounts of data generated by the multinational Human Genome sequencing effort and the need for a coordinated approach across all organisations involved. The pre-publication sharing of sequence data also made a new infrastructure necessary including, subject specific databases to house sequence information and unique identifiers (accession numbers) to allow versions of the same sequence to be tracked (Ostell, Wheelan, & Kans, 2004).

In other disciplines, where community supported data repositories do not exist, practice is less well developed. In cases where there is no common goal to help drive practice from the ground up, researchers generally conform to publishers policies when they exist (Sieber & Trumbo, 1995, also mentioned by interviewees), falling back on those of data centres as the next resort. If neither the data centre, nor publisher, gives instructions, citation does not occur.

To ensure publisher policies match with existing practice, community engagement is needed. Learned societies, as representatives of researchers as well as publishers, have an important role in bridging the gaps between niche community standards. They should act to grow citation standards from existing practice to cover their whole subject scope. Secondly, they should bring these through into their publishing standards and policies. Societies publish a significant proportion of research articles (over 30%, Ware & Mabe, 2009) and so can help to raise general awareness of data citation. Some also provide publishing guidelines that provide far more detail and background than can be found in author guidelines (Macrina, 2011).

1.8.2 Follow metadata standards for datasets:

That is, researchers should actively create metadata for their datasets, and submit these along with their data when depositing it in a repository. Basic metadata is the information included in the citation. As a minimum, it is important for the dataset to have a distinct title so it is easily distinguishable from titles of related papers and data. Metadata standards exist in many disciplines, however researchers may need support in following these; and this is a potential area librarians and data centres could help support.

1.8.3 Use of persistent identifiers such as DOIs:

Our interviews highlighted that researchers will use DOIs if it is easy to do so and if the benefits mentioned previously are clear. If citation and bibliographic management software included research dataset options, which would include a DOI or other persistent identifier, this could help to drive good practice by lowering the required effort on the part of the researcher.

1.9 Opportunities for Publishers

1.9.1 Establish uniform data citation standards

These standards may well be dependent on already existing citation practices across research communities, or guidelines within style guides. Interviewees urged publishers to consult with data centres within their subject areas (and vice versa), to work together to form those standards. It is equally important for publishers to make their citation requirements clear to authors and reviewers, and ensure they are followed (Mooney, 2011). Less than 10% of journals across ecology, evolutionary biology and environmental sciences give directions on citing data (Weber et al., 2011). Also, unless journals enforce their citation requirements through the peer review and editorial processes, the current situation will be much slower to improve.

1.9.2 Follow metadata standards for datasets

The importance of quality metadata in allowing re-use and in enabling citation is clear, and was underlined in interviews with publishers and data centres. Metadata standards would help, but for these to be used, they need to first be agreed and accepted by authors, data producers, data archives/libraries and most importantly by publishers. Interviewees would like to see publishers taking an active role in working with these different groups, to help unify existing recommendations and guidelines (DataCite, ICPSR, GESIS/da|ra etc.) and ensuring quality metadata through peer review. We have seen in the past that mandatory policies of journals like Science and Nature relating to referencing genomic data helped drive good practice in data management in that field.

1.9.3 Use of persistent identifiers for datasets

A summary of the existing requirements for citation are available in Appendix 1. In order to encourage researchers to use persistent identifiers for datasets, the recommendation or policy provided by a journal or data centre should actively mention the inclusion of a persistent identifier.

1.9.4 Data publications

Interviewees highlighted the opportunity for enhancing traditional publications through citation linking to underlying data. They also suggested that creating new types of data publications, such as data journals, would add an important element: i.e. the intellectual explanation of the particular value of the data. The potential for multiple formats and approaches arising means that there is no clear definition of what a data publication is.

1.10 Opportunities for data centres

1.10.1 Engage in establishing uniform data citation standards

Interviews demonstrated that data centres need to establish how they wish data to be cited, consider how this should be done at all levels (i.e. citing a whole database, a data collection, a data file) and should ensure clear guidelines and recommended citation formats are provided to users. Data centres could also work to engage with publishers to make them aware of their requirements.

Making data citation as easy as possible is crucial to encourage researchers to do it; data centres have an important role in this. Dataset landing pages – a webpage which provides the information relating to the dataset, could also include a recommended citation for the dataset in a copy and paste format, or a format that can be loaded into citation management software. Landing pages could also enable bidirectional linking with publications to provide users more context on the value and re-use of a particular dataset.

The Global Biodiversity Information Facility (GBIF) has been developing its data citation recommendations since 2008. Its white paper identified developments that were still needed for data citation to carry the same weight and credit as traditional citations, which could serve as a checklist for other data infrastructures:

1. Rethink 'copyright' for data
2. Involve learned societies
3. Clear understanding of how citation format must be flexible to accommodate differences in data types
4. Assurance of persistence
5. Technological incorporation
6. Requirements from publishers for underlying data to be published and cited

1.10.2 Support and promote persistent identifiers:

Again, as is the case for publishers, the requested citation format provided by data centres needs to mention persistent and resolvable identifiers. Data centres also need to decide the most appropriate citable unit for their data, as granularity issues can vary across disciplines. Within disciplines data centres could do more to help unify citation formats; Lane (2008) found that across 15 data aggregators in biodiversity, there were 11 different citation formats. This creates a confused picture for researchers, and across a single discipline it should be far easier to cooperate on a single standard.

1.11 Data citation: what we have learned and future directions

A number of issues regarding data citation were highlighted through desk research and community consultation. Some practical issues require active production of tools and services, or the uptake of these; while others point to ways in which various communities must consider working together to promote data citation best practices. Below is a list of some key learning points:

1. **Persistent identifiers should be used to uniquely identify and address datasets.**

   These identifiers should be allocated by data publishers (data centres, repositories, libraries or publishers) who should also make them easy to find and use, used by researchers when citing or referring to data and publishers should ensure via the editorial process that this requirement is met.

   Persistent identification solutions should be shared across peer organisations, so that learning and uptake are maximised.

2. **Citations with persistent identifiers should be listed in the references/bibliography to enable tracking of citation metrics.**

3. **Publishers and data centres need to provide guidance for authors and referees on citation of data.**

   Researchers will be further encouraged to cite data where guidelines are provided on how to do it. To further ensure that citation of data is appropriate and included in the correct section of an article, publishers need to make peer reviewers aware of citation standards and formats they implement.

4. **Citation metrics need to be built to monitor and assess citation of data.**

   To achieve this, publishers and other organisations already creating citation metrics for articles should consult with data centres and funders to develop data citation metrics specifications that meet their reporting needs.

5. **Vendors of bibliographic tools should recognise the need for data citation in their citation formats and search.**

   E.g. so that researchers can add references explicitly formatted as ‘datasets’ easily to a document from within their reference management software, and load bibliographic metadata provided by data centres/data searches as datasets.

6. **There is confusion on what persistence and longevity of data is required for it to be citeable and cited.**

   Citation metrics for data could help inform this, by providing evidence on how long after publication data is cited and reused.

7. **There is a lack of clarity and agreement on what 'authorship' of a dataset means.**

Contributions beyond producing a dataset need acknowledgement via authorship, but the best way to do this still need to be decided.

8. **The relatively recent emergence of the 'data paper' format means that authors (and publishers) are not clear on whether both a data paper and a dataset can or should be cited, and the weighting of each in terms of academic prestige.**

9. **Liaison roles would help to mediate interactions and bridge gaps between different communities in the data citation landscape**.

This could help improve communication between data centres, publishers and researchers, to ensure any new standards and guidelines developed account for existing community practices.

10. **Many researchers do not appear to see the value and benefits of data citation.**

How different communities can work together to promote this activity and the status of datasets as primary research outputs and publishable works in their own right, is an issue that still needs to be addressed.

11. **Researchers need to nurture awareness in their community of the benefits of data citation, and follow citation guidelines given by publishers and data centres.**

This can be led from academic societies and institutions.

2. DEFINING CURRENT BEST PRACTICE AND FUTURE ROLES

2.1 Introduction

The ODE report “Integration of Data and Publication” (Reilly et al 2011) concluded that these six criteria are key to ensuring the long term success of linking data effectively to publications:

1. Availability
2. Findability
3. Interpretability and Re-usability
4. Citability
5. Curation
6. Preservation.

It also showed that neither researchers nor publishers have the responsibility or resources to satisfy all of these criteria, whereas libraries and data centres have existing relationships with researchers and publishers and the related knowledge and skills to ensure that many of these criteria are met. A workshop at the LIBER Annual Conference in Barcelona in 2011, where initial results of the report were discussed with librarians, established that research libraries are keen to engage in data management. It moreover confirmed that there was a need to define the exact roles that libraries should fill and what the incentives and barriers are for researchers to work with libraries on data management if the potential of linking data to publications was to be truly realised.

The previous chapter, in exploring best practice for data citation also highlights opportunities for libraries in terms of providing support to researchers in making their data citable and in citing data.

The news roles and opportunities identified in the “Integration of Data and Publication” and in the chapter on best practice for data citation draw into question how well libraries are prepared to cope with these new roles and opportunities. They also have implications for the changing skill sets required of librarians.

2.2 The Survey

The survey was designed to validate some of the findings from the ODE report on Integration of Data and Publications. More importantly, it was designed to gather evidence on the current and expected roles of libraries with regard to data management, in order to prescribe steps for the evolution of these roles. The survey was structured roughly in accordance with the areas of opportunity identified in the linking data to publications report (Table 3).

Data Issue	Libraries and data centres opportunities (Chapter 4)
Availability	• Lower barriers to researchers to make their data available. • Integrate data sets into retrieval services.
Findability	• Support of persistent identifiers. • Engage in developing common meta description schemas and common citation practices. • Promote use of common standards and tools among researchers.
Interpretability	• Support crosslinks between publications and datasets. • Provide and help researchers understand meta descriptions of datasets. • Establish and maintain knowledge base about data and their context.
Re-usability	• Curate and preserve datasets. • Archive software needed for re-analysis of data. • Be transparent about conditions under which data sets can be re-used (expert knowledge needed, software needed).
Citability	• Engage in establishing uniform data citation standards. • Support and promote persistent identifiers.
Curation/ Preservation	• Transparency about curation of submitted data. • Promote good data management practice. • Collaborate with data creators • Instruct researchers on discipline specific best practices in data creation (preservation formats, documentation of experiment, ...).

Table 3. Libraries and data centres' opportunities in data exchange

2.3 Objectives and Scope

The main objective of the survey was to gain an understanding of how libraries can support researchers in data management and hence clarify the library role in data exchange.

This has been done through gathering answers from libraries related to the following questions:

1. What is the perceived demand from researchers for support for data management from libraries?
2. In what areas does this demand exist?
3. What support is currently in place?
4. What skills are needed to meet the demand for support?

Some questions were also reused from the Parse.Insight Survey in order to find out if there had been any changes in practices in, or views of, the field of data management that might affect the finding of this report.

Further questions were added to examine views on the future of linking data and publications. As they have been a traditional intermediary between researchers and publications it was also felt that it was important to establish whether libraries had a common vision of the future of data and publications and, in particular, how libraries believe data and publications could become better integrated in the future.

2.4 Target Group

The target group for this survey is the research library community. Having already established an overview of areas where researchers need support in the previous report, this survey sought to elucidate what demand were being made on libraries for support in data management. The libraries have an established relationship with their research communities, providing traditional library support services. Many of these traditional services have parallels or overlap with the increasingly important elements of data management, such as metadata and citation.

2.5 Validity of results

The survey was sent out via email to over 800 individual librarians from the 424 research libraries across over 40 European countries on the LIBER mailing list. The introduction to the survey emphasised that the survey was not just for libraries active in the area of data management, but was also for 'non-experts' interested in developing services in this area.

There were 110 respondents to the survey. This represents a response rate of over 13% of those contacted. Given such a response rate these results can be interpreted as broadly representative of the views of European research libraries as a whole.

One concern regarding the results of the survey was the gap between perceived and actual demand for data management services. It was felt that libraries already providing data management support services would have a better insight into the demand from researchers for these services as they are better embedded in the data management/research workflow. Libraries not providing these services were more than likely basing their answers on perception. As the survey responses show, there is currently a relatively low rate of data management support service provision across European libraries. To address this gap between perception and experience, several internationally recognized leading libraries (experts) in the field of data management support provision from the US and Australia were identified and asked to respond to the survey. A comparison of these responses with the European responses allows for the identification of the gaps between experience and perception. However the responses from the US and Australia are from a select few and cannot be necessarily interpreted as representative of the state of play internationally.

2.6 Survey results

2.6.1 Demand for data management support

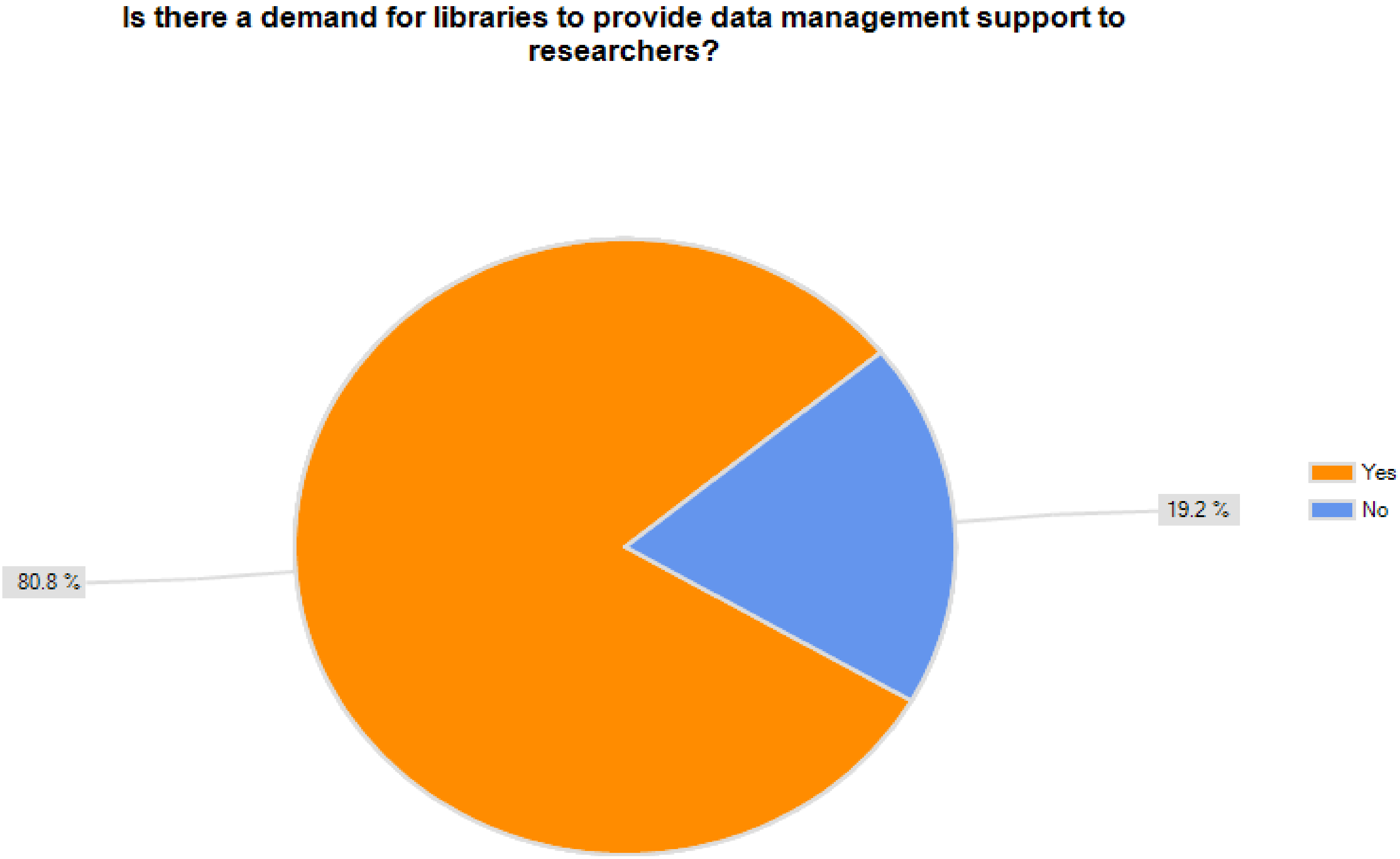


Figure 2. Demand

The overwhelming majority of survey participants (81%) believe that there is a demand for libraries to provide data management support to researchers.

2.6.2 Demand vs. supply

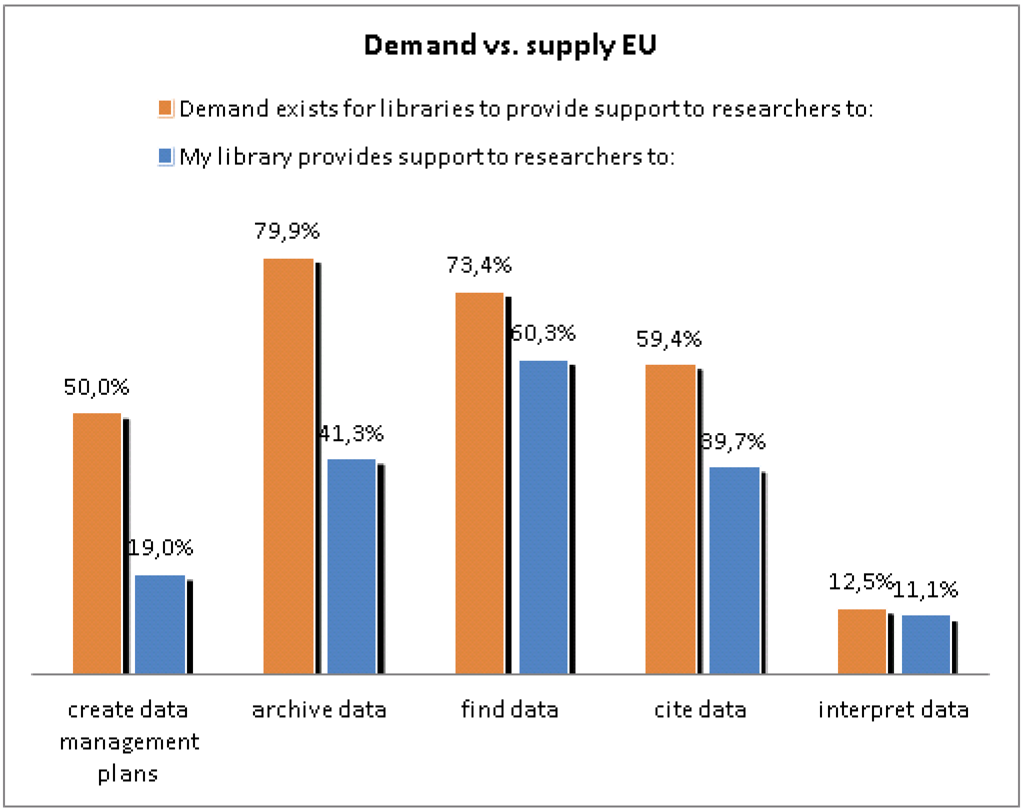


Figure 3. Supply

The strongest demand is seen for support in data archiving (80%), followed by support in finding data (73%), citing data (60%) and writing data management plans (50%). Demand for support in interpreting data ranks lowest with 12.5%. Linking data sets was specifically mention by both the European and the Expert libraries.

It is noticeable that supply of such services significantly lags behind these figures. Libraries are particularly weak on the level of support for the creation of data management plans. This type of support may not be provided due to a lack of skills, such as technical and subject expertise, and also to the possibility that these libraries are not properly embedded in the research process.

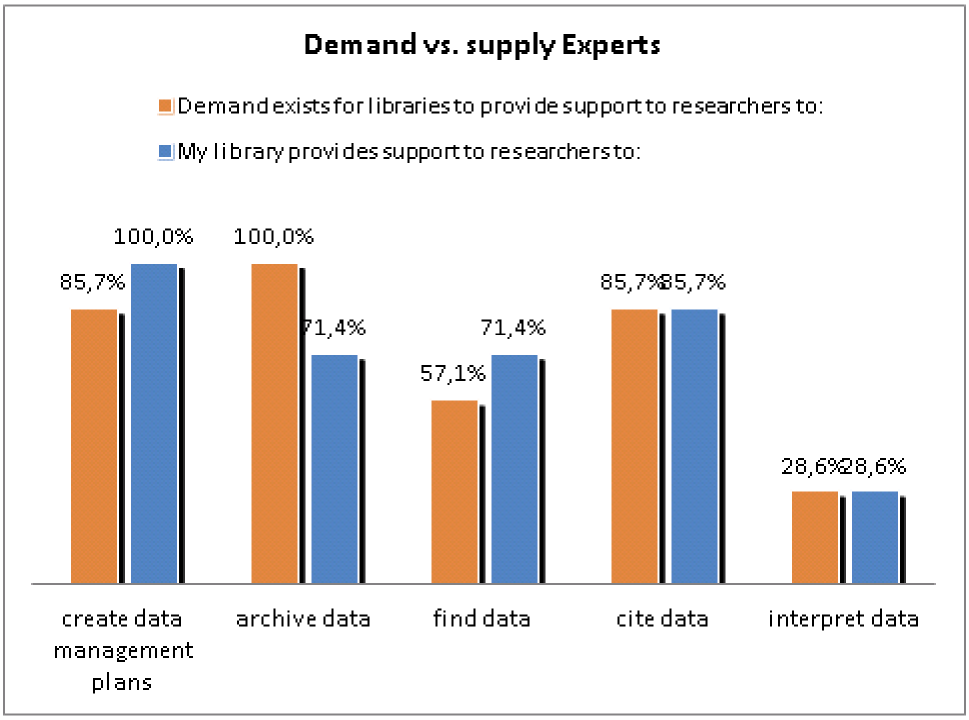


Figure 4. Experts Supply

Because the Expert librarians represent libraries that are already active in data management support provision, their supply rates are significantly higher in comparison to the European libraries: 71% provide support to find data, 86% to cite, and71% to archive. The biggest discrepancy between the European and the Expert libraries is the provision of support in writing data management plans, which is provided by 100% of the Expert libraries, but only by 19% of the European libraries. In fact, it seems that there is an oversupply of this type of support amongst the Expert libraries. It is worth noting that the Expert libraries come from regions where data management plans are mandated by research funders.

2.6.3 Linking data to publications

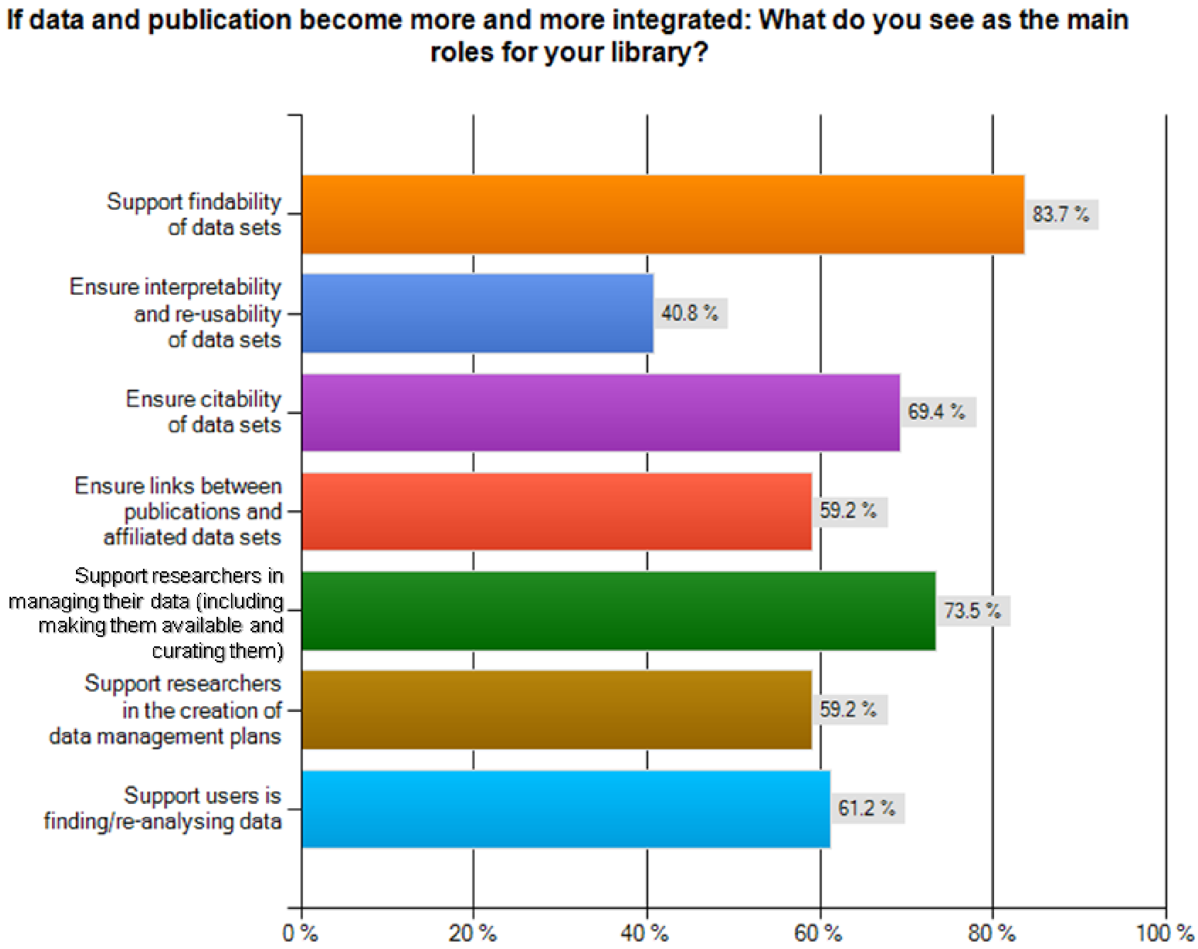


Figure 5. Integration roles

Libraries are aware that they have a role to play in relation to the integration of data and publications. 83% of the surveyed libraries see their main role as supporting the findability of data sets. This is an area which draws on traditional library skills such as cataloguing and metadata. 73.5% see support for researchers in managing their data as a major role for libraries, which points to a need to develop skills in data curation. Only 41%, of libraries prioritise supporting interpretability and re-usability of data sets. The responses to this question may need to be explored further as re-usability also relates to licencing, which is identified as an area that libraries could help address in chapter 2.2 of this report.

2.6.4 Who should be responsible for maintaining and selecting research data in archives?

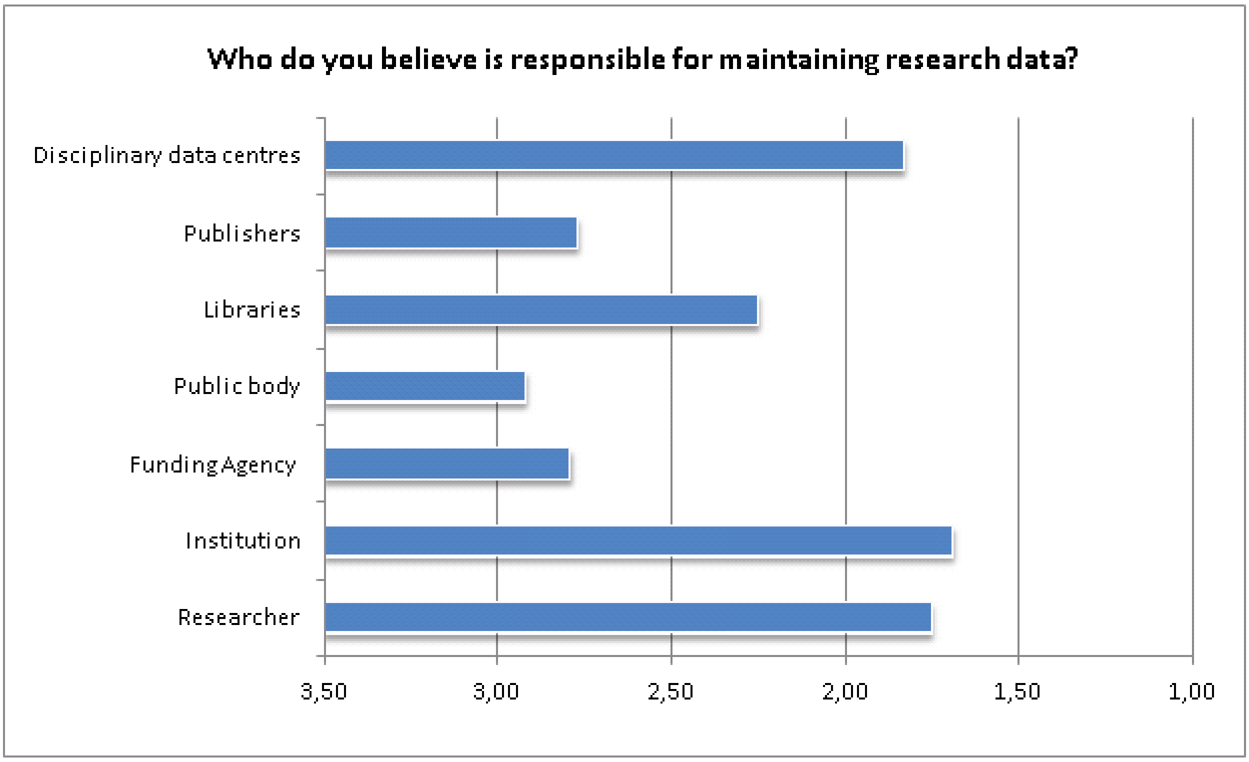


Figure 6. Responsibility for archiving

The institution (in which research data is created) is considered most responsible for maintaining research data, directly followed by the researchers themselves and disciplinary data centres. In this question, the response rates are very similar between the European and the Expert libraries. Libraries rank themselves in fourth place in terms of responsibility. It is worth noting when interpreting these results that it is common for libraries either to have responsibility for, or a close relationship with, the institutional repository. Publishers, Funding Agencies and Public Bodies are seen to have the least responsibility for maintaining research data.

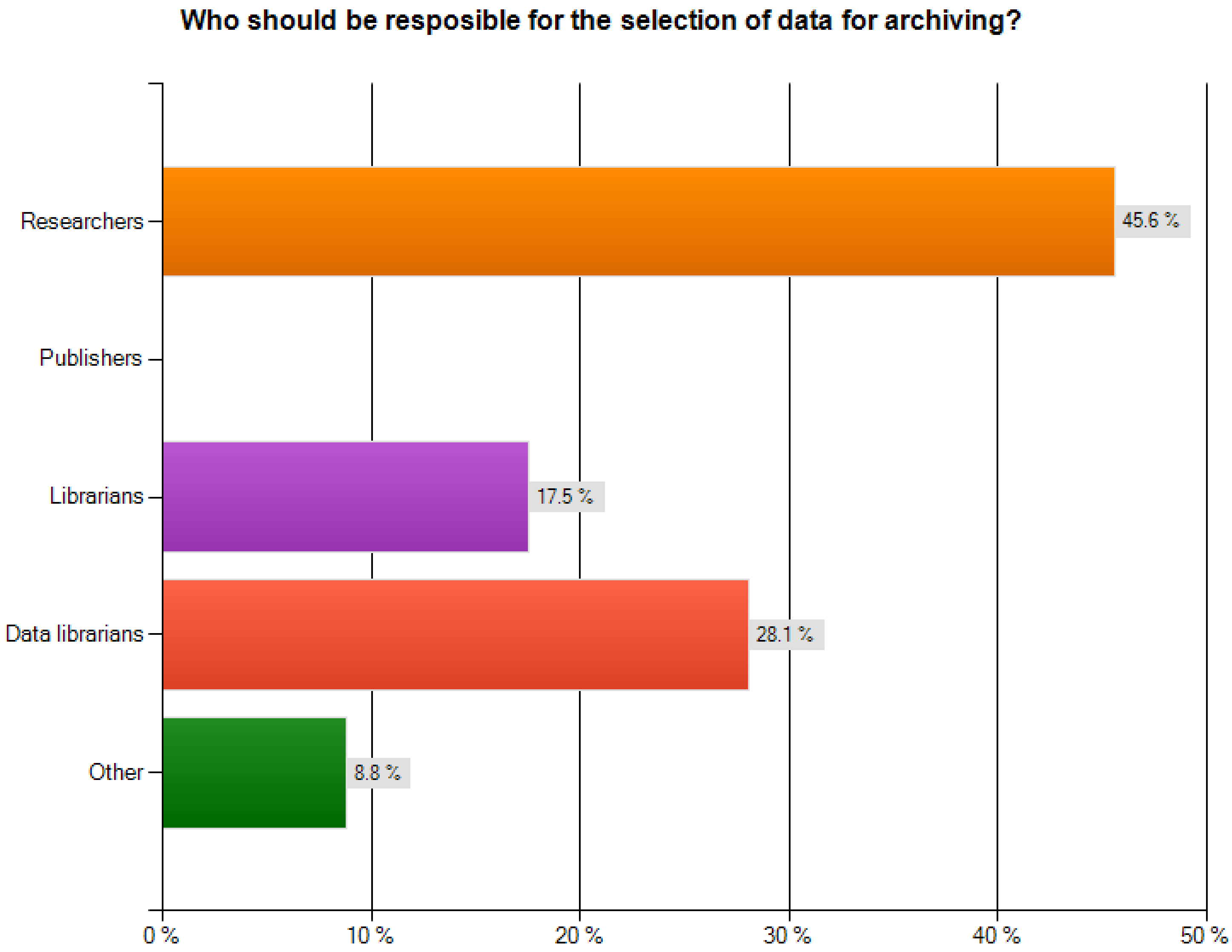


Figure 7. Responsibility for selections

Responsibility for the selection of data for archiving is overwhelmingly seen as lying with researchers, followed by data librarians, librarians and others. Interestingly, the Expert libraries differ with this opinion in that they all agree that only the researchers should be responsible for the selection, and no one else. If researchers are to be solely responsible for this, then perhaps libraries should begin to consider how they can support researchers to make these decisions? In their free text answers several of the Expert librarians pointed to this as the way forward:

"Ideally a combination of responsibilities. Researchers have the most domain expertise but may not necessarily think about future users outside their discipline so the view of an information professional (not necessarily a librarian) would also be useful in assessing long-term value."

2.6.5 Availability

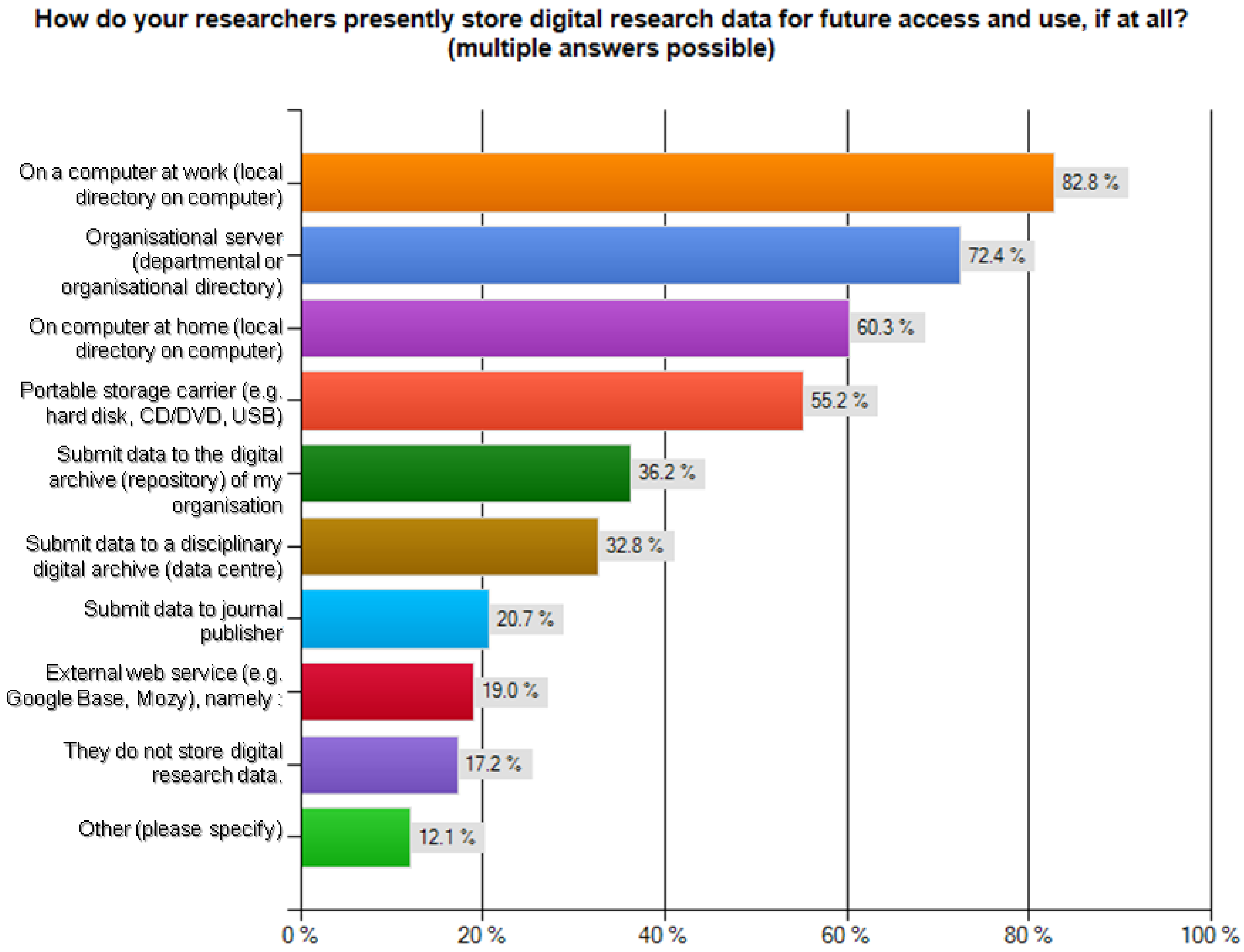


Figure 8. Storage practices

When it comes to ensuring availability of research data, there is still much to be achieved. To establish a baseline, we repeated a question in the survey from the PARSE.Insight survey 2009. Whereas a wide range of researchers of all disciplines, data managers and publishers, answered the PARSE.Insight survey this current survey was answered mainly by librarians, who report their view on the behaviour of their researchers. We regard the answers as largely reliable, because research librarians know their user base well.

Responses show that the storage habits of researchers have improved somewhat with regard to availability of data in comparison to 2009. Disciplinary data centre and digital archive of institution rank significantly higher (36% as compared to 6% and 33% as compared to 14%), and also storage on the organizational server, which secures a minimum amount of availability for the data, jumped from 59% in 2009 to 72%. But a strong preference to store data privately still exists amongst researchers. Should libraries be doing more to advocate for storage options that facilitate data sharing?

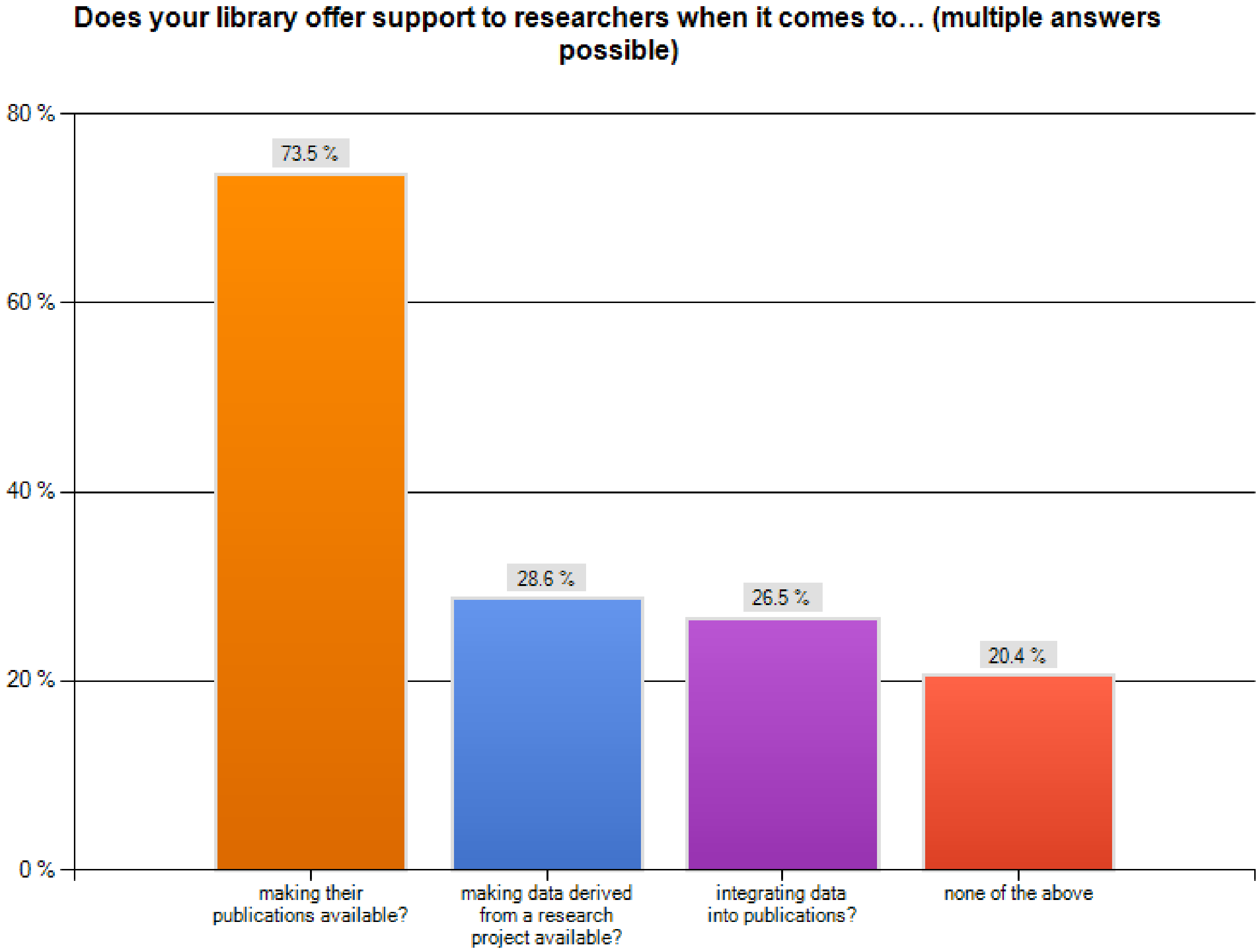


Figure 9. Support for making data available

Libraries could support their researchers in making their data more widely available, but, here too, our survey establishes a gap. While 74% of the surveyed libraries offer support to their researchers when it comes to making their publications available, only 29% are prepared to offer support when it comes to making data from a research project available. Only 27% provide support to integrate data with publications, and 20% offer no such services.

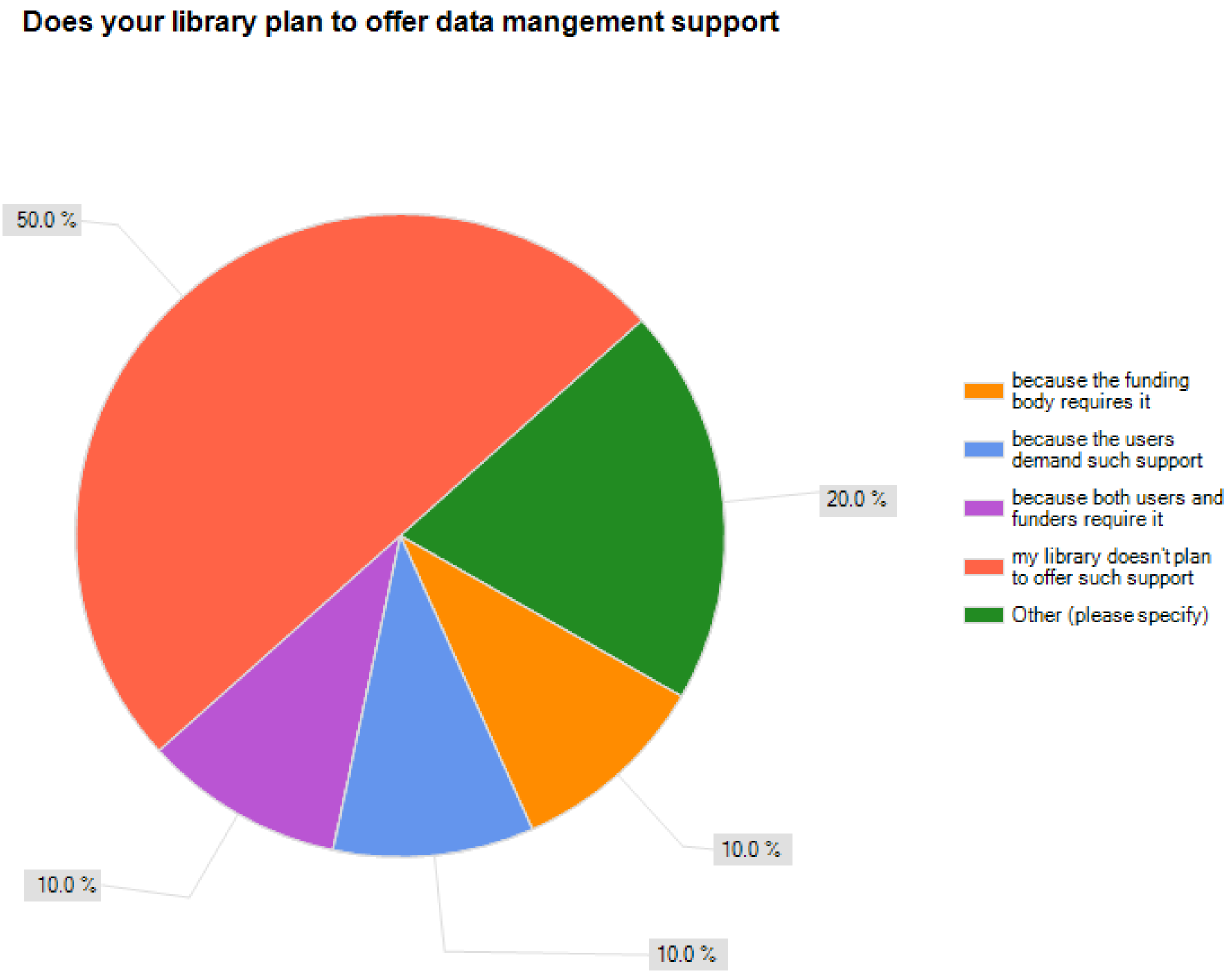


Figure 10. Future plans for data management support

Moreover, only 50% of the surveyed libraries plan to offer such support, for which they provided a variety of reasons. The other half of the libraries state that they do not plan to offer any support. When compared to the high agreement rates in question 2.3 concerning the roles that libraries have to play in this area (73.5% agreed that libraries have a role to play to support researchers in managing their data, including making them available), these answers reveal a worrying gap. Is this something that needs to be mandated before efforts will be made to bridge this gap?

2.6.6 Findability

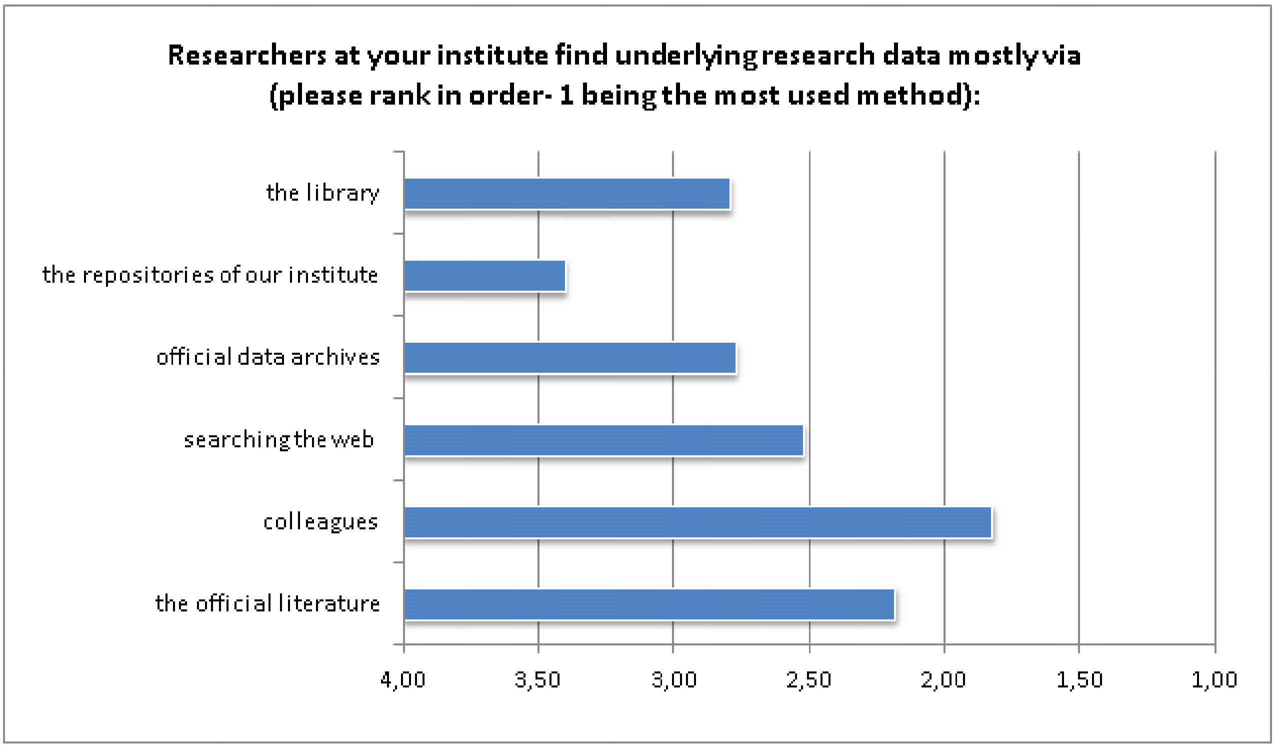


Figure 11. Finding data

When asked how their researchers find research data, librarians assume that they mostly turn to their colleagues. The official literature is ranked next, followed by "searching the web". Official data archives, the library, and the repositories of the institution rank (in this order) last. The high position of official literature re-emphasises the need to establish and maintain close links between publications and the underlying data. The relatively low ranking of libraries proves that libraries have not yet established a prominent role in supporting researchers in finding research data (see also next question). That official data archives/subject repositories are not better used may point to a gap in advocacy for the use of such archives.

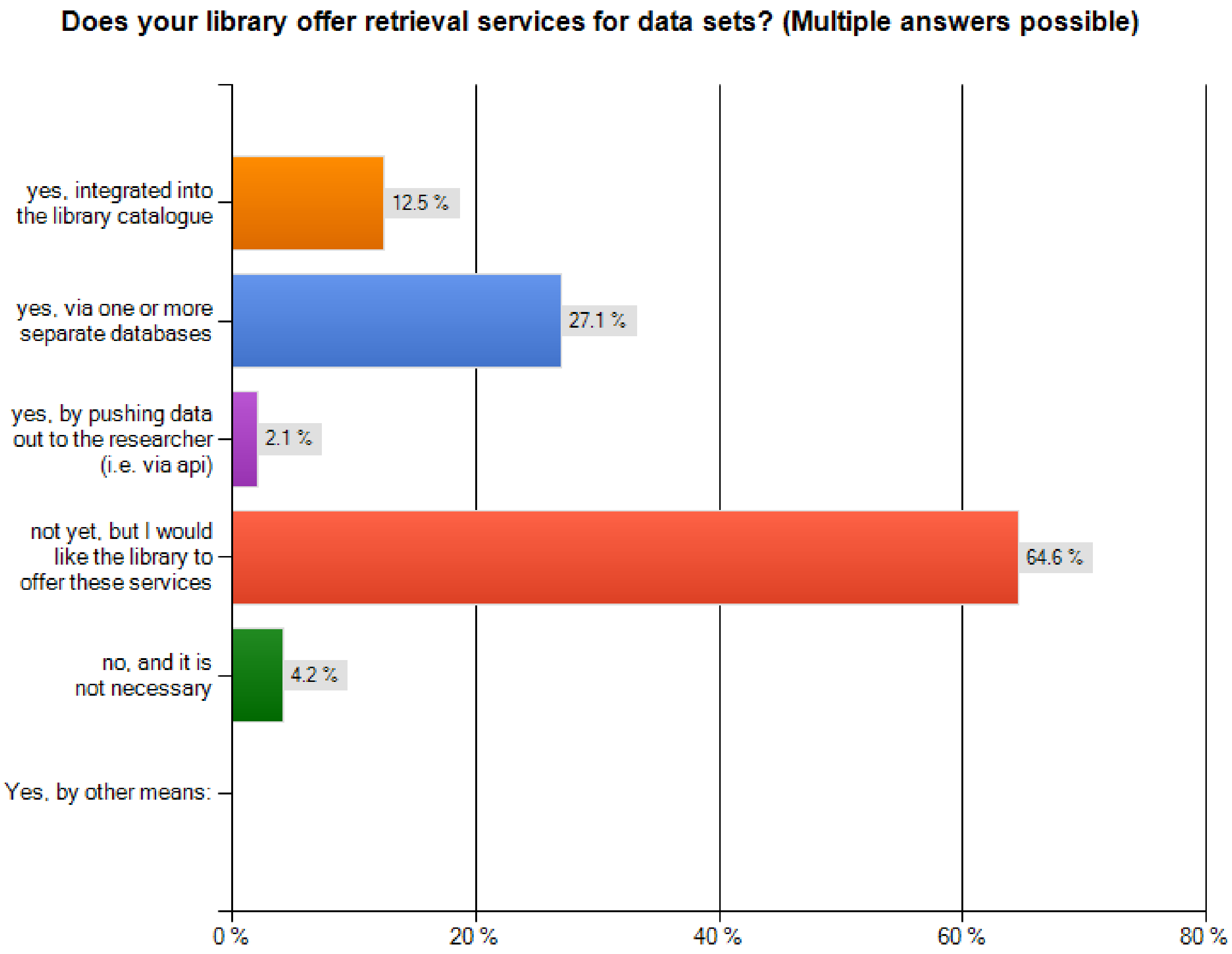


Figure 12. Retrieval services

Only 12.5% of the surveyed libraries have integrated retrieval services for data sets into their library catalogue. 27% offer data retrieval services via separate databases. The considerable majority of 64.5% does not, but would like to, offer such services. Only 4% believe that it is not necessary to provide a data retrieval service. One opinion was expressed that libraries do not have the 'know-how' to develop such integrated search services.

2.6.7 Interpretability / Re-usability

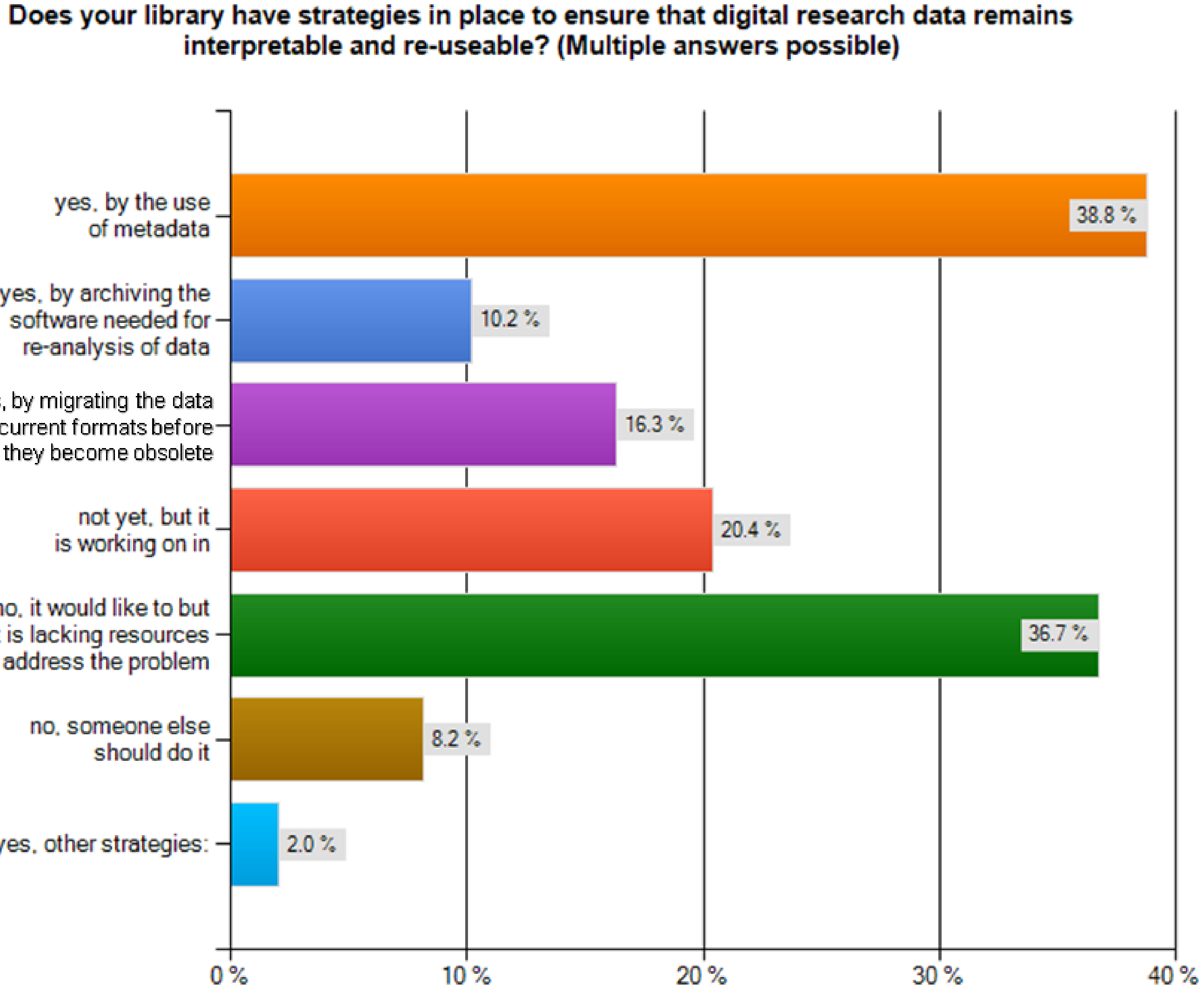


Figure 13. Interpretatibity & reusability

The surveyed libraries support interpretability and re-usability of research data primarily through the use of metadata (39%). The next largest group indicates that they lack resources to address the problem (37%). 20% report that they are working on developing strategies. Only 10% archive the software needed for re-analysis of data, and only 16% have established migration strategies for research data.

This question addressed the technical interpretability and re-usability of research data. Another way to support interpretability and re-usability (intellectually, semantically) is to maintain close links between research data and the related article, where the raw data are explained and interpreted in full length. This is explored in the next question.

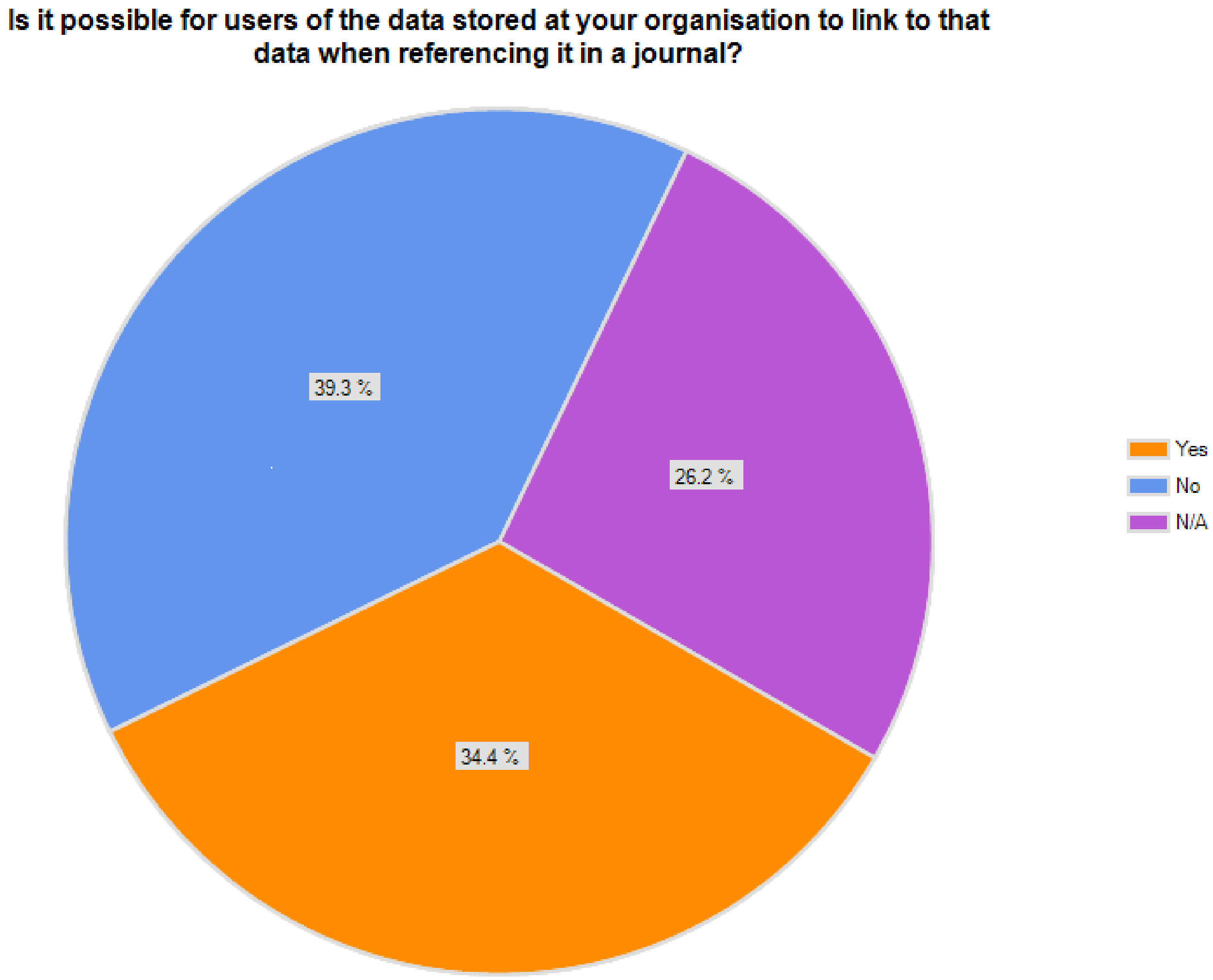


Figure 14. Linking to data from journals

Roughly a third of the surveyed libraries state that research data stored at their institutions can be linked to journal articles. Almost 40% state that this is not possible and 26% said that this question would not be applicable. The figure of 26% should be regarded in conjunction with data availability (section 3.5 of this document) as data that is not made available in disciplinary or institutional archives cannot be linked to when referenced in an article.

Here, the gap between the European libraries and the selected Expert libraries is pronounced, as 87.5% of the US/AUS libraries state that their users can link to data at their organizations.

2.6.8 Citability

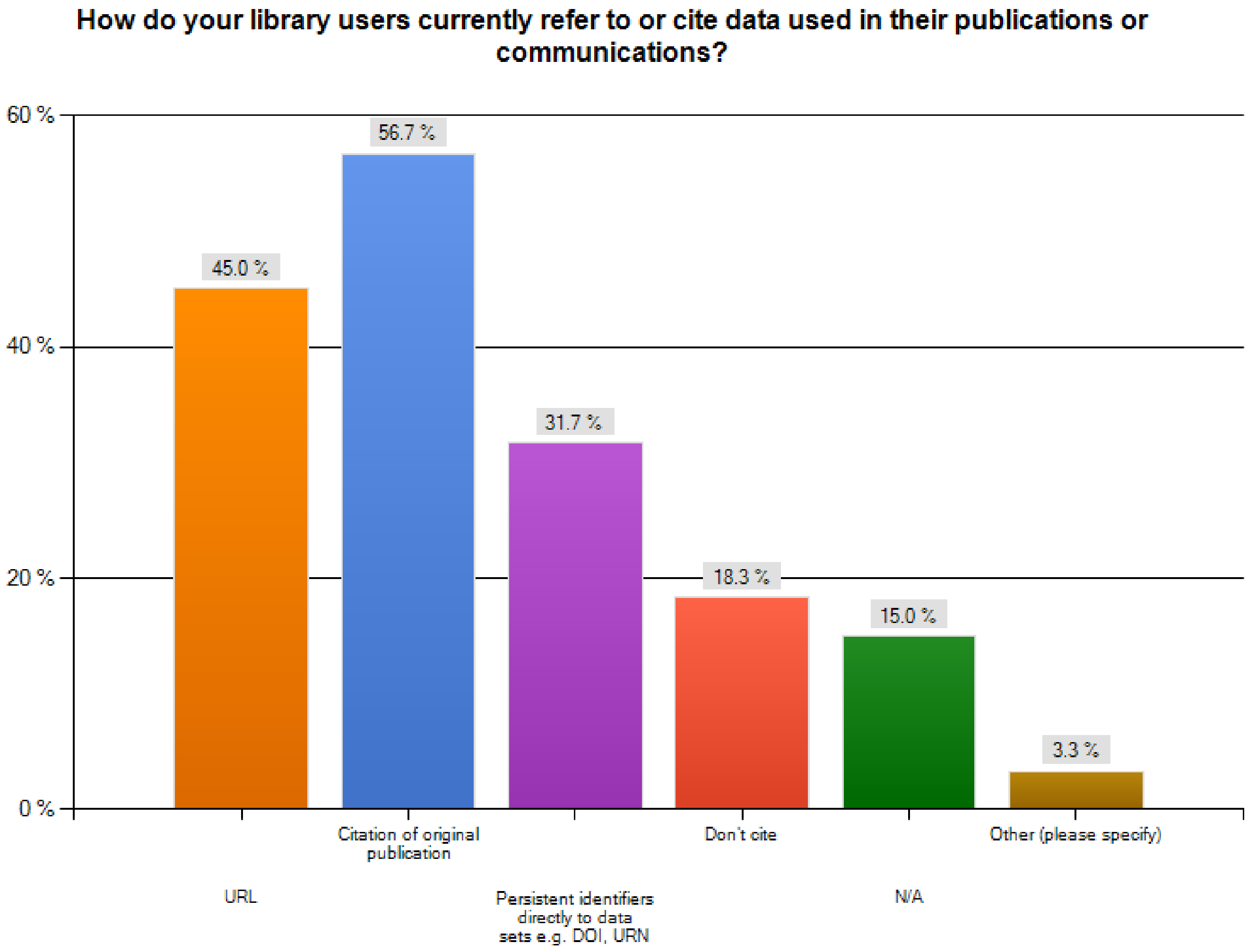


Figure 15. Citation practices

The response to this question very much relates to the previous chapter on best practice in data citation. It points to a very clear gap between what should be best practice and what libraries see as the current practice amongst researchers.

Currently the most widely used method of citation is citation of the original publication (50%). The use of URLs (45%) is also popular. Only 32% of researcher actually link to the original data using a persistent identifier. This is a worrying indication of current practice and point to a need for support in educating researcher on the use of persistent identifiers to enable citability and proper citation of data.

The response from the Expert libraries shows that there is almost an inverse relationship between the citation of the original publication (37.5%) and the use of persistent identifiers (75%).

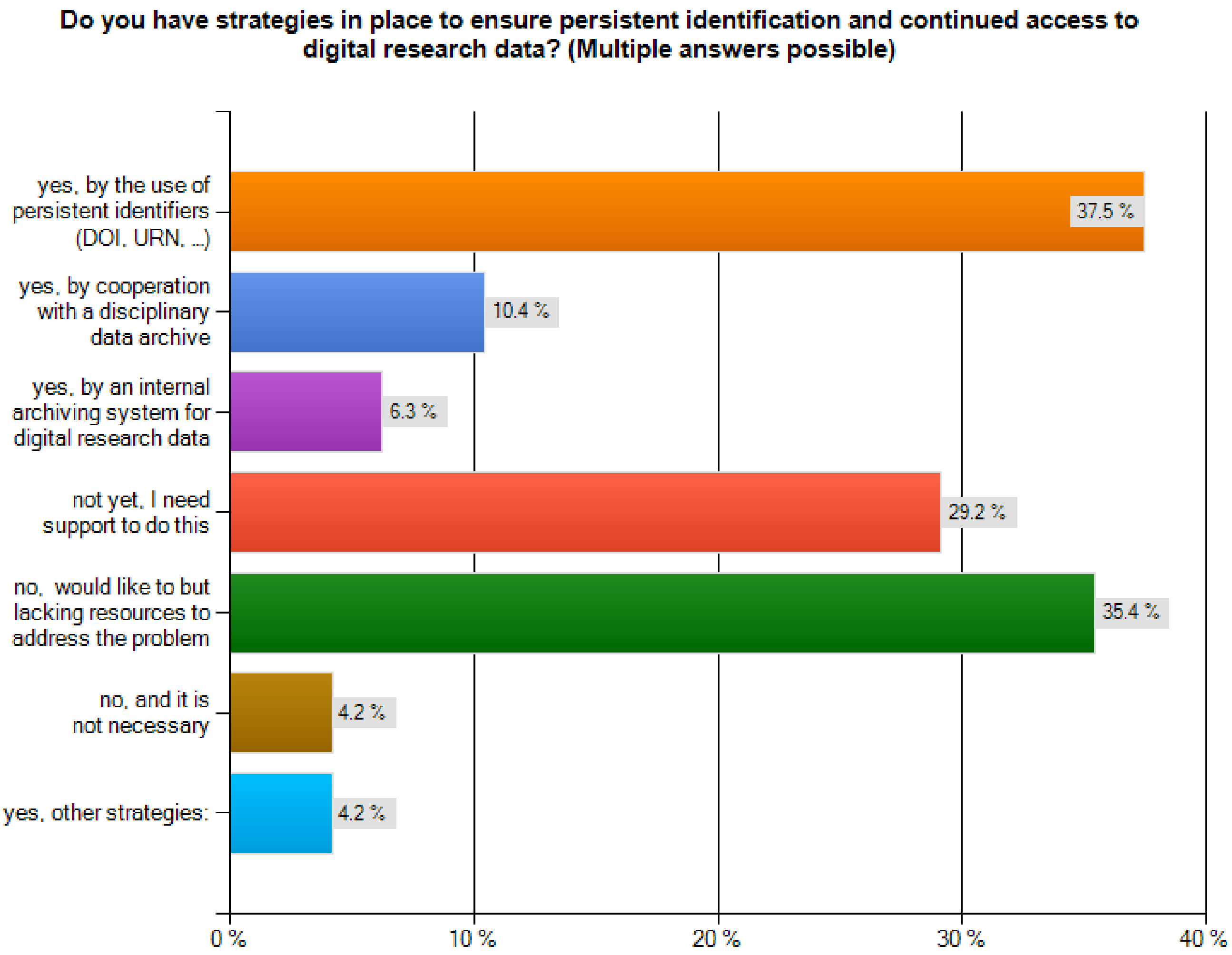


Figure 16. Persistent identification strategies

It seems that libraries are aware of the need to support persistent identification. A very large proportion (35%) state that they would like to put strategies in place but they are lacking the resources to address this problem. A need for support (29%) was also identified as a barrier to putting strategies in place. 37.5% of libraries are using persistent identifiers, whilst only 10.4% actually cooperate with disciplinary data archives to do this.

In contrast 75% of the Expert libraries use persistent identifiers and 25% actually cooperate with disciplinary data archives. This suggests that there is still scope for libraries to increase cooperation with disciplinary archives, particularly if they feel they are lacking the support to implement such strategies on their own.

2.6.9 Preservation

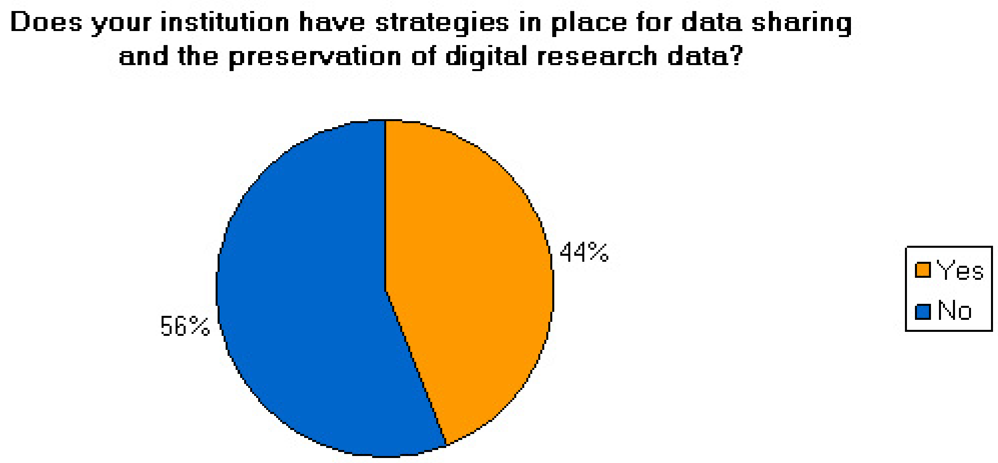


Figure 17. Preservation strategies

Unsurprisingly, 44% of libraries have strategies in place for data sharing and preservation in comparison to 88% of the Expert libraries. Again this is a large gap. If libraries see the need to provide support for data management, and clearly they do, then this gap needs to be addressed. The reasons why libraries do not currently have these strategies in place need to be explored.

2.6.10 The future of data publishing

To elicit some opinions on the future of data publishing, some potential scenarios were included for the survey participants to agree or disagree with. The options are listed here in the order of their popularity.

"About the way publications and data are integrated":

Scenario	**Agree**	**Expert Comparison**
The best place for underlying data is in official data repositories and archives	84%	75%
Publications should always contain links to the underlying research data	74%	75%
Data archives should have a system in place for persistent identifiers that properly support citation of datasets	74%	87.5%
Research journals should have much stricter editorial policies on data availability	64%	25%
Underlying data should be part of the peer review process	54%	37.5%
Underlying data should be cited separately in the reference list	46%	37.5%
Publishers and editors should only accept in supplements the summary datasets that are of direct relevance to the article	26%	12.5%
There are not sufficient trustworthy data archives available for authors to deposit their data	24%	62.5%
The best way to make underlying data available is via supplementary files to journals	16%	0%
Supplementary files to journal articles only make sense if they are interactive	4%	25%
Other (please specify)	no answers	37.5%

Table 4. Responses to scenarios on the integration of publications and data

From the replies, it is clear that the role of official, trustworthy repositories is important. In fact, they are far more important than supplements to journals, which do not rank highly in the responses. It is interesting to note that data are expected to play a more prominent role in the enrichment of articles, and as separately citable entities. The suggestion that publications should always contain links to the underlying data, and that citation of datasets should be supported by persistent identifiers, rank among the three most popular statements. The most notable difference between the European libraries and the Expert libraries is that in almost all cases the Expert libraries have a stronger opinion. The Expert libraries seem to strongly believe that there are not sufficient trustworthy data archives available for authors to deposit their data.

"Future scenarios for better integration of data and publications":

Scenario	Agree	Expert Comparison
Data sets will become separately citable items, supported by their own citation framework.	71%	100%
Underlying data will increasingly become an integrated part of enriched articles via special viewers, pointers and interactive pdf's.	71%	100%
Web developments will present new opportunities to better integrate underlying data with publications and to improve their discoverability.	67%	75%
The present linear research article will gradually be replaced by more modular presentations of research with a customisable, hierarchical or layered presentation in which each reader can choose the preferred level of detail (also for data sets).	46%	62.5%
Research data will evolve into a publication of its own kind, independent of traditional publications.	37.5%	62.5%
There is a need for dedicated Data Journals that describe data sets and their methodologies.	37.5%	25%
Research journals will have to play a guiding role in the proper management of underlying research data, e.g. data management plans and deposits at suitable data repositories/archives	33%	12.5%
Other (please specify)	6%	12.5%

Table 5. Responses to scenarios on improving integration of data and publications

Again, with these future scenarios, the opinion from the Expert libraries is similar but much more pronounced. An important trend for both researchers and libraries is that data sets will become separately citable items, supported by their own citation framework. This highlights the impetus to implement and support best practice in data citation.

2.6.11 Skills

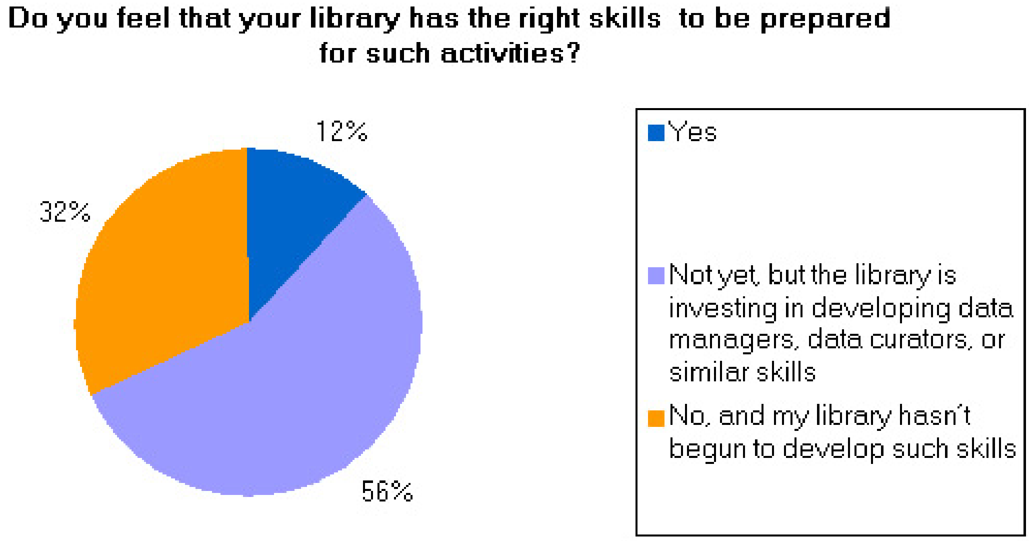


Figure 18. Current librarian skills

Only 12% of libraries feel they have the right skills to address the opportunities that data sharing presents. 56% are investing in developing these skills. The 32% who have not begun to develop these skills shows that there will still be a gap between demand and actual support provision in the future. Comments from the Expert libraries such as "We have a dedicated full-time data management role, and are increasingly embedding data management advisory functions within our teams of subject librarians" provide a sharp contrast in terms of the state of play regarding the skills that are in place.

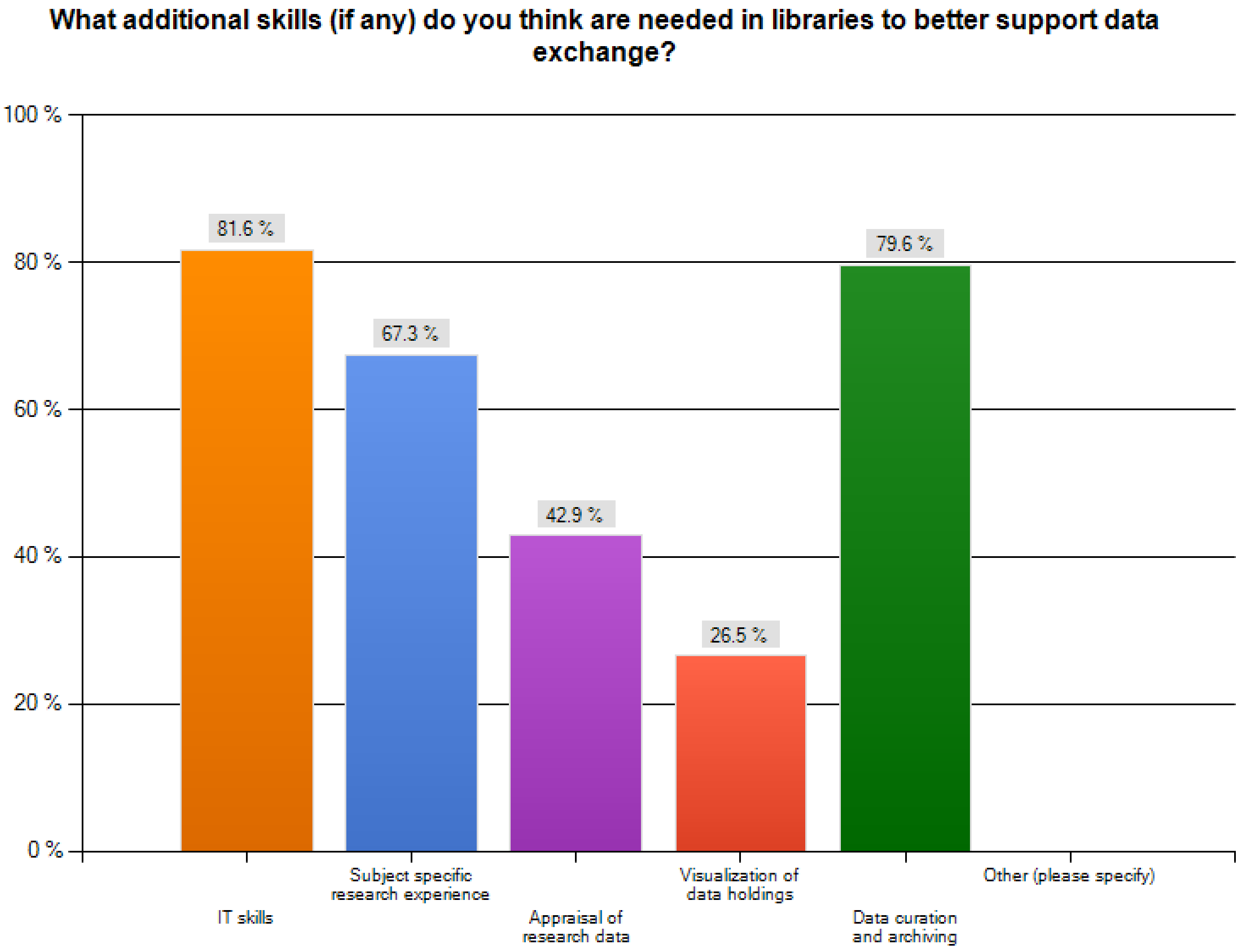


Figure 19. Additional skills for librarians

82% of libraries consider IT skills to be the most needed skills in terms of supporting data exchange. Skills in data curation and archiving are also considered important (80%). The difference between the perceived importance of subject specific expertise, for European libraries (67%) and the Expert libraries (88%) is pronounced. The Expert libraries see this skill as the most important by far. Subject expertise has been found to be more valued by researchers and more likely to engender trust in advice from libraries[7].

[7] http://www.jisc.ac.uk/media/documents/programmes/RIM/RIMReport_FINAL.pdf

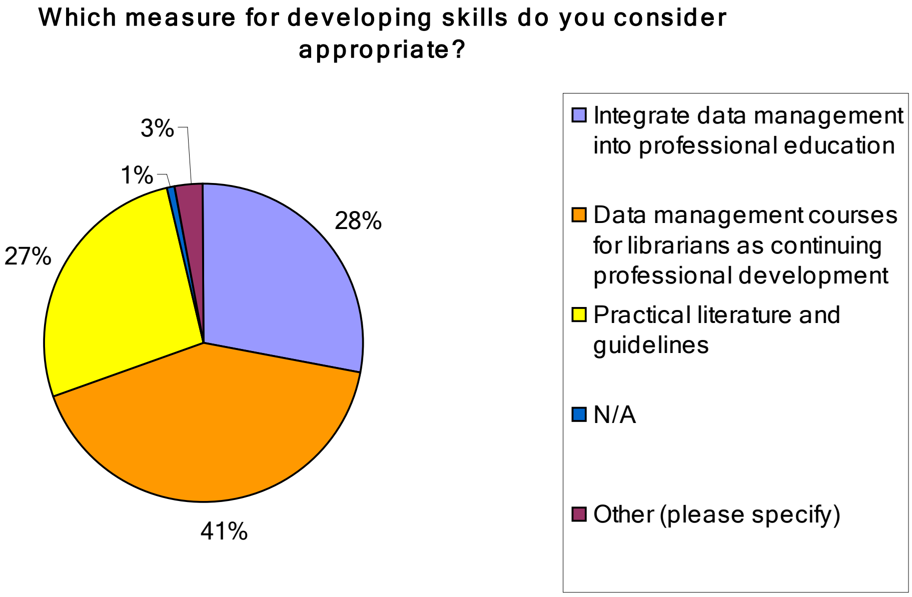


Figure 20. Development of skills

The best means of developing all such skills, according to the libraries, is through the provision of continuing professional development. This is possible in all skills areas except for subject specific research expertise, which implies the need to recruit librarians with experience of discipline specific research.

The integration of data management into professional training courses was also a favoured option (63.3%), followed by practical literature and guidelines (61.2%). These preferences were mirrored in the Expert responses. Other suggestions for developing skills included the engagement of libraries in related professional organisations such as the Data Management Association (DAMA) and the introduction of coaching/project secondment opportunities in order to build relevant organisational knowledge.

2.7 Library Roles: what we have learned and future directions

There are a number of issues that have been raised in this survey. Some results also point to future directions that libraries must consider when contemplating their next steps in terms of harnessing opportunities in data exchange.

Below is a list of some key learnings and future considerations drawn from the survey report:

1. **Libraries recognize that there is a demand to provide data management support for researchers**, and that they are well positioned to fulfil at least parts of this demand.
2. Currently, the **libraries that provide data management support** are in the **minority**.
3. Libraries clearly need to **work together to develop the new skill sets necessary to address the demand for support in data management**. Further dialogue must

occur, both within the community and in consultation with other stakeholders, about the type of skills that need to be developed.

4. Not all of the required skill sets currently exist in libraries and the **profession may need to consider new ways of developing and attracting such skills e.g. subject expertise**.
5. **There is a gap, which could be filled by libraries, in advocacy for data sharing**, the use of subject specific repositories, and best practice in data citation in order to increase the number of researchers sharing and reusing data.
6. There is a **need to increase dialogue with researchers** regarding opportunities for data exchange and to become more embedded in the research process.
7. **Investment must be made in increasing the level of support for data management plans** in particular if European libraries are to meet emerging international best practice standards.
8. It may be that an increase in support for writing data management plans will only occur if there are **mandates from funders/institutions** for this activity.
9. Researchers will need **training and guidance on how to make their data citable** and on how to cite data in order to ensure that they can fully benefit from a future where data may become a publication on its own right. Libraries are well positioned to provide this type of support and should move to do so.
10. There are still a great many institutions without **strategies in place for the preservation of research data**. Why this is the case should be considered and addressed.
11. Coordinated action might be considered to **support libraries to put strategies in place to support persistent identification** and continued access to research data.

Establishing the **library role in data curation should be prioritised**. Further work needs to be done to map out exactly what the role of libraries should be in data curation and explore how prepared libraries are for this activity.

3. COMMON ISSUES

Here we aim to summarise some of the key themes and issues relating to data citation and emerging roles for libraries in supporting data exchange, captured through desk research, community consultation and the online survey of librarians.

Many issues face a range of stakeholders in the data sharing, management and citation landscape and thus require further dialogue and discussion across these different perspectives, to develop potential solutions. A 'one-size-fits-all' solution will not be possible in most areas; some specific issues will need to be addressed at a community-based level, as solutions will be discipline specific.

There should be clear global communication of where enablers have been found, to allow shared learning and streamlining of approaches where possible.

1. There are simple and practical steps that all parties can take to enable easier citation and tracking of data. These emerge from building consideration of data citation requirements into existing tools and services e.g. citation metrics and bibliographic management tools.
2. The emergence of the 'data paper' format is evidence for the increased appetite for data reuse and citation in some subject areas. But it is unclear how they will drive academic credit, data reuse and data citation in the long term, and their applicability to more diverse disciplines.
3. Current communication between data centres, publishers, libraries and scientific communities is poor and as a result standards, guidelines, support and training relating to data, if present, may not be relevant to community practices.
   - i) Liaison roles would help to mediate interactions and bridge these gaps.
   - ii) Improved communication is also needed to navigate and develop solutions to the remaining challenges. This will include establishing what data can and should be cited, and how data citation can best provide the appropriate acknowledgement to all those involved in the data exchange process, from creation to reuse.
4. Many researchers do not appear to see the value and benefits of data citation. There is a gap, which could be filled by libraries, in advocacy for data sharing, the use of subject specific repositories, and best practice in data citation. These, if filled, would increase the number of researchers sharing and reusing data. The issue still to be addressed is how different communities can work together to promote this activity and the status of datasets as primary research outputs and publishable works in their own right.
5. Persistent identifiers should be used to uniquely identify and address datasets. These identifiers should be allocated by data publishers (data centres, repositories, libraries or publishers).Libraries have a role in promoting and supporting the use of persistent identifiers, through raising awareness on the use and reuse of identifiers within the library and research communities, and also through ensuring that they are findable within their search services. With their expertise in metadata, libraries should also be engaging in wider discussions

surrounding the use of identifiers within metadata records and the agreement of standards for persistent identification.

6. Not all of the required skill sets currently exist in libraries and the profession may need to consider new ways of developing and attracting such skills and subject expertise in order to provide researcher support and training in citation, management, curation, and preservation, of data. Further dialogue must occur, both within the library community and in consultation with other stakeholders, about the type of skills that need to be developed and to explore how prepared libraries are for these new activities.
7. Data citation is key to the successful adoption of data sharing by researchers and libraries can help address some of the issues that need to be tackled if best practice in data citation is to be implemented. If libraries are to support researchers regarding opportunities for data exchange, there is a need to increase dialogue with researchers and for librarians to become more embedded in the research process.
8. Institutional repositories may need to support researchers whose data falls outside the remit of existing subject data repositories. Libraries and information support services that manage these repositories at academic institutions require expanding data-skills and also need strategies in place to support continued access to research data.
9. Investment is needed to increase the level of data management support in particular if European libraries are to meet emerging international best practice standards. This may also require institutional strategies and mandates from funders/institutions to be in place.

4. REFERENCES

Accomazzi, A. (2011). Linking Literature and Data: Status Report and Future Efforts. *Future Professional Communication in Astronomy II*, (2010). doi:10.1007/978-1-4419-8369-5_15

Altman, M., & King, G. (2007). A Proposed Standard for the Scholarly Citation of Quantitative Data. *D-Lib Magazine*, 13(3/4). doi:10.1045/march2007-altman

Appleby, J. W., & Leroux, C. (2011). Manual Entry - WizFolio - LibGuides at Scholars Portal. *Scholars Portal*. Retrieved June 1, 2012, from http://guides.scholarsportal.info/content.php?pid=221965&sid=1842663

Ball, A., & Duke, M. (2006). *How to Cite Datasets and Link to Publications*. DCC How To Guides (Vol. 11). Edinburgh: Springer Netherlands. Retrieved from http://www.dcc.ac.uk/resources/how-guides

Ball, B. A., & Duke, M. (2011). *Data Citation and Linking*. DCC Briefing Papers. Edinburgh. Retrieved from http://www.dcc.ac.uk/resources/briefing-papers/

Buneman, P., & Harmar, T. (2006). How to cite curated databases and how to make them citable. Scientific and Statistical Database Management, 2006. *18th International Conference on Scientific and Statistical Database Management* (pp. 195-203). Vienna. doi:10.1109/SSDBM.2006.28

Chavan, Viswas. (2012). *Recommended practices for citation of data published through the GBIF network* (p. 12). Copenhagen. Retrieved from http://links.gbif.org/gbif_best_practice_data_citation_en_v1

Chavan, Vishwas, & Penev, L. (2011). The data paper: a mechanism to incentivize data publishing in biodiversity science. *BMC Bioinformatics*, 12(Suppl 15), S2. BioMed Central Ltd. doi:10.1186/1471-2105-12-S15-S2

Comparison of reference management software. (n.d.). *Wikipedia*. Retrieved May 31, 2012, from http://en.wikipedia.org/wiki/Comparison_of_reference_management_software#Word_processor_integration

Corti, L., & Bolton, S. (2012). Presentation: Data Identifiers: How to ensure your data is cited properly. Retrieved April 19, 2012, from http://www.jisc.ac.uk/media/documents/events/2012/JISC_WEBINAR_DOIs_FINAL.pdf

Cronin, B., & Overfelt, K. (1994). Citation-based auditing of academic performance. *Journal of the American Society for Information Science*, 45(2), 61-72. doi:10.1002/(SICI)1097-4571(199403)45:2<61::AID-ASI1>3.0.CO;2-F

DataCite Metadata Working Group. (2011). *DataCite Metadata Schema for the Publication and Citation of Research Data*. Version 2.1. doi:10.5438/0003

Digital Curation Centre. (n.d.). Overview of funders' data policies. Retrieved May 15, 2012, from http://www.dcc.ac.uk/resources/policy-and-legal/overview-funders-data-policies

Federation of Earth Science Information Partners. (n.d.). Interagency Data Stewardship/Citations/provider guidelines. Note on Versioning and Locators. *ESIP Federation Wiki*. Retrieved May 31, 2012, from http://wiki.esipfed.org/index.php/Interagency_Data_Stewardship/Citations/provider_guidelines#Note_on_Versioning_and_Locators

Green, T. (2009). We Need Publishing Standards for Datasets and Data Tables. *OECD Publishing White Paper*. doi:10.1787/603233448430

Hage, F. M. (2011). Choice or Circumstance? Adjusting Measures of Foreign Policy Similarity for Chance Agreement. *Political Analysis*, 19(3), 287-305. doi:10.1093/pan/mpr023

Hanney, S., Frame, I., Grant, J., & Buxton, M. (2005). Using categorisations of citations when assessing the outcomes from health research. *Scientometrics*, 65(3), 357-379. doi:10.1007/s11192-005-0279-y

Hrynaszkiewicz, I. (2011). The need and drive for open data in biomedical publishing. *Serials*, 24(March), 31-37. doi:10.1629/2431

Issue #22: Data Set. (2011).*GitHub*. Retrieved June 1, 2012, from https://github.com/ajlyon/zotero-bits/issues/22

JabRef. (2012). Customizing entry types. *JabRef reference manager*. Retrieved May 31, 2012, from http://jabref.sourceforge.net/help/CustomEntriesHelp.php

Jones, C. M., Bouton, K. A., Hey, J. M. N., Latham, S. E., Lawrence, B. N., Matthews, B. M., Miles, A., et al. (2007). *Data Publication: outputs of the CLADDIER project What does publishing data entail?* Computer (pp. 1-7). Retrieved from http://epubs.cclrc.ac.uk/bitstream/1838/PV2007_Jones_CLADDIER-final.pdf

Lane, M. A. (2008). *Data Citation in the Electronic Environment. Philosophy.*

Macrina, F. L. (2011). Teaching authorship and publication practices in the biomedical and life sciences. *Science and engineering ethics*, 17(2), 341-54. doi:10.1007/s11948-011-9275-1

Martineau, P. (2005). Interacting with OpenOffice.org: Formatting the bibliography. *Bibus documentation*. Retrieved May 31, 2012, from http://bibus-biblio.sourceforge.net/bibus_doc/html/en/usingOOo.html#mozTocId650609

Maunsell, J. (2010). Announcement regarding supplemental material. *The Journal of Neuroscience*, 30(32), 10599-10600. Retrieved from http://www.neuro.cjb.net/content/30/32/10599.short

Meurer, P., Schultz, J., & Tejada, A. (2012). Citavi Manual. Swiss Academic Software GmbH. Retrieved from http://www.citavi.com/service/en/docs/Citavi_3-Manual_EN.pdf

Mooney, H. (2011). Citing data sources in the social sciences: do authors do it? *Learned Publishing*, 24(2), 98-108. doi:10.1087/20110204

NISO Business Working Group. (2012). *Recommended Practices for Online Supplemental Journal*. Retrieved from http://www.niso.org/apps/group_public/document.php?document_id=7964&wg_abbrev=suppbusiness

Nature Biotechnology Editorial. (2009). Credit where credit is overdue. *Nature Biotechnology*, 27(7), 2009.

Newton, M. P., Mooney, H., & Witt, M. (2010). A Description of Data Citation Instructions in Style Guides. *Libraries Research Publications*, Paper 121. Retrieved from http://docs.lib.purdue.edu/lib_research/121

Ostell, J. M., Wheelan, S. J., & Kans, J. A. (2004). The NCBI Data Model. In A. D. Baxevanis & B. F. F. Ouellette (Eds.), *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins* (pp. 19-44).

Papers for Windows. (2012). Glossary of Icons in Papers. *Papers for Windows Support*. Retrieved June 1, 2012, from http://pfw.mekentosj.com/kb/getting-started/glossary-of-icons-in-papers

Pearce, N., & Smith, A. H. (2011). *Data sharing: not as simple as it seems*. Environmental health : a global access science source, 10(1), 107. doi:10.1186/1476-069X-10-107

Penev, L., Mietchen, D., Chavan, V., Hagedorn, G., Remsen, D., Smith, V., & Shotton, D. (2011). Pensoft Data Publishing Policies and Guidelines for Biodiversity Data, 1-34.

Piwowar, H. a, Day, R. S., & Fridsma, D. B. (2007). Sharing detailed research data is associated with increased citation rate. *PloS ONE*, 2(3), e308. doi:10.1371/journal.pone.0000308

Reed College. (n.d.). Making a .bib file with JabRef. *Help Deck Computing & Information Services*. Retrieved May 31, 2012, from http://web.reed.edu/cis/help/latex/Jabref.html

Reilly, S., Schallier, W., Schrimpf, S., Smit, E., & Wilkinson, M. (2011). *Report on Integration of Data and Publications* (p. 87). Retrieved from http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2011/11/ODE-ReportOnIntegrationOfDataAndPublications-1_1.pdf

Rohlfing, T., & Poline, J. (2012). Why shared data should not be acknowledged on the author byline. *NeuroImage*, 59(4), 4189-4195. Elsevier Inc. doi:10.1016/j.neuroimage.2011.09.080

Shapland, M. (n.d.). Evaluation of Scholar's Aid. Retrieved June 1, 2012, from http://eis.bris.ac.uk/~ccmjs/scholar.htm

Sieber, J. E., & Trumbo, B. E. (1995). (Not) giving credit where credit is due: Citation of data sets. *Science and Engineering Ethics*, 1(1), 11-20. doi:10.1007/BF02628694

Simons, N. (2012). Implementing DOIs for Research Data. *D-Lib Magazine*, 18(5/6). doi:10.1045/may2012-simons

Sinnott, R., Macdonald, A., Lord, P., Ecklund, D., & Jones, A. (2005). Large-scale data sharing in the life sciences: Data standards, incentives, barriers and funding models (The Joint Data Standards Study). *The Biotechnology and Biological Sciences Research Council; The Department of Trade and Industry; The Joint Information Systems Committee for Support for Research; The Medical Research Council; The Natural Environment Research Council and The Wellcome Trust*, (August 2005). Retrieved from http://www.dcs.gla.ac.uk/publications/paperdetails.cfm?id=8109

The Argo Science Team, Roemmich, D., Boebel, O., Desaubies, Y., Freeland, H., Kim, K., King, B., et al. (2003). Argo: The Global Array of Profiling Floats. In C. J. Koblinsky & N. R. Smith (Eds.), *Observing the Oceans in the 21st Century* (pp. 248 - 258). Melbourne: GODAE Project Office and Bureau of Meteorology. Retrieved from http://www.argo.ucsd.edu/batch32d.pdf

Third Street Software, I. (2011). Sente Academic Reference Manager for Mac OS X. Retrieved June 1, 2012, from http://www.thirdstreetsoftware.com/site/SenteForMac.html

Thomson Reuters. (2008). Customizing Reference Types. *Reference Manager 12* (pp. 370-381). Thompson Reuters. Retrieved from http://refman.com/support/docs/ReferenceManager12.pdf

Turton, J. (2003). Argo: An Array of Free-Drifting Profiling Floats. *Sea Technology*, 44(10), 33-36.

Ware, M., & Mabe, M. (2009). *The stm report*. Retrieved from http://www.stm-assoc.org/industry-statistics/the-stm-report/

Weber, N. M., Street, E. D., Piwowar, H. A., Street, W. M., & Suite, A. (2011). Evaluating Data Citation and Sharing Policies in the Environmental Sciences. ASIS&T '10 *Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem* (pp. 1-2).

Wren, J. D. (2008). URL decay in MEDLINE--a 4-year follow-up study. *Bioinformatics (Oxford, England)*, 24(11), 1381-5. doi:10.1093/bioinformatics/btn127

Wynholds, L. (2011). Linking to Scientific Data : Identity Problems of Unruly and Poorly Bounded Digital Objects. *The International Journal of Digital Curation*, 6(1), 214-225.

x4 upgrade to x5, now slow. (2011)."EndNote General" forum. Retrieved May 31, 2012, from http://community.thomsonreuters.com/t5/EndNote-General/x4-upgrade-to-x5-now-slow/m-p/22351/highlight/true#M5706

Zelle, R. M. (2012). Adding Items to your Zotero Library. *Zotero*. Retrieved June 1, 2012, from http://www.zotero.org/support/getting_stuff_into_your_library

5. APPENDICES

A summary of recommended information to be included in data citations, from the literature and from data centres. We have not reviewed data citation instructions in style guides, as a useful review is already available (Newton, Mooney, & Witt, 2010).

Element	Reference	Data centres/aggregators
Title	(Sieber & Trumbo, 1995); (Altman & King, 2007); (Green, 2009); (DataCite Metadata Working Group, 2011); (Jones et al., 2007)□	ADS; AHDS; BADC; DataVerse; Dryad; ESDS; GBIF; GESIS; ICPSR; IPY; NIST; NSIDC; ORNL-DAAC; PANGAEA
Author/Depositor	(Sieber & Trumbo, 1995); (Altman & King, 2007); (Green, 2009); (DataCite Metadata Working Group, 2011); (Jones et al., 2007)	ADS; AHDS; BADC; DataVerse; Dryad; ESDS; GBIF (plus multiple other 'contributors'); GESIS; ICPSR; IPY; NSIDC; NIST; ORNL-DAAC; PANGAEA
Distributor (data centre/publisher)	(Sieber & Trumbo, 1995); (Green, 2009); (DataCite Metadata Working Group, 2011); (Jones et al., 2007)	ADS; AHDS; BADC; DataVerse; Dryad; ESDS; GBIF; GESIS; ICPSR; IPY; NIST; NSIDC; ORNL-DAAC
Date of publication	(Sieber & Trumbo, 1995); (Altman & King, 2007); (Green, 2009); (DataCite Metadata Working Group, 2011); (Jones et al., 2007)	ADS; AHDS; BADC; DataVerse; Dryad; ESDS; GBIF; GESIS; ICPSR; IPY; NIST; NSIDC; ORNL-DAAC; PANGAEA
Unique identifier	(Altman & King, 2007); (Green, 2009); (DataCite Metadata Working Group, 2011)	GESIS; ICPSR; PANGAEA; Dryad; BADC; ADS; ESDS; ORNL-DAAC; GBIF; DataVerse
Actionable link	(Altman & King, 2007); (Green, 2009); (DataCite Metadata Working Group, 2011); (Jones et al., 2007)	ADS; BADC; DataVerse; Dryad; ESDS; GBIF; ORNL-DAAC
Date of survey/collection	(Sieber & Trumbo, 1995)	GESIS; IPY; NSIDC
Location of distributor	(Sieber & Trumbo, 1995)	ADS; AHDS; ICPSR; ESDS; IPY; NIST; ORNL-DAAC
Author institution		BADC; PANGAEA
Research funder	(Sieber & Trumbo, 1995)	
Indication if data is machine readable	(Sieber & Trumbo, 1995)	
Information on code book	(Sieber & Trumbo, 1995)	
Universal numeric fingerprint	(Altman & King, 2007)	DataVerse
Date accessed or cited	(Green, 2009); (Jones et al., 2007)	ICPSR; IPY
Study identifier		AHDS; ESDS;GESIS
Object type or media		ADS; AHDS; ESDS; GBIF; GESIS; ICPSR; IPY; NSIDC; ORNL-DAAC
Version number/identifier		ESDS; GBIF; GESIS; ICPSR; NIST; NSIDC
Number of records		GBIF

Table notes:

GESIS: Minimal standard recommended for the citation of research data (vers. 1.3 as of 07.03.2012) retrieved from http://www.gesis.org/en/services/data-analysis/data-archive-service/citation-of-research-data/ Accessed: 05/04/2012

PANGAEA: http://wiki.pangaea.de/wiki/Citation accessed: 05/12/2012

Dryad: http://datadryad.org/using#howCite accessed: 05/04/2012

BADC (British atmospheric Data Centre): http://data.datacite.org/10.5285/E8F43A51-0198-4323-A926-FE69225D57DD accessed 05/04/2012

ICPSR (Inter-University Consortium for Political and Social Research): http://www.icpsr.umich.edu/icpsrweb/ICPSR/support/faqs/2008/10/why-and-how-should-i-cite-data accessed: 05/04/2012

ADS (Archaeological Data Service): Taken from the specific example: http://dx.doi.org/10.5284/1000002 accessed: 05/04/2012

UKDA/ESDS (UK Data Archive/Economic and Social Data Service): http://esds.ac.uk/orderingData/citing.asp accessed 05/04/2012

ORNL-DAAC (Oak ridge Nation Laboratory): http://daac.ornl.gov/citation_policy.html accessed 05/04/2012

IPY (International Polar Year): http://ipydis.org/data/citations.html accessed: 05/04/2012

GBIF (Global Biodiversity Information Facility): GBIF (2011). Recommended practices for the citation of data published through the GBIF Network. Version 1.0 (Authored by Vishwas Chavan), Copenhagen: Global biodiversity Information Facility. Accessible at http://links.gbif.org/gbif_best_practice_data_citation_en_v1

NSIDC (National snow and Ice Data Center): http://nsidc.org/data/gla03.html accessed 20/04/2012

AHDS (Arts and Humanities Data Service): http://www.ahds.ac.uk/history/collections/citation.htm accessed 20/04/2012

DataVerse: http://thedata.org/citation/standard accessed 20/04/2012

NIST (National Institute of Standards and Technology): NIST ask that their data is cited as if it were a book: http://www.nist.gov/srd/frequent.cfm