Enabling use of DOIs for data citation in longitudinal studies

A report on the DataCite UK Workshop for CLOSER

Contents

Executive summary 1

Recommendations 2

Workshop Report 4

1. Introduction 4
2. Benefits of persistent identifiers for longitudinal study data 5
3. Implementation of DOIs and Persistent Identifiers in CLOSER studies to realise the benefits 6
4. Next steps 11

Executive summary

The UK's longitudinal studies (including cohort and household panel studies) have been making their data available to researchers beyond their immediate staff for decades. Over this time, they have developed various models of data sharing and citation that have increased the value of the data collected.

DataCite[1] was created in 2009 to encourage data sharing by allowing better practice in data citation using **Digital Object Identifiers (DOIs)**. The British Library is a founding member of DataCite and works with UK organisations to ensure that their data, software and other research items can be uniquely identified DOIs. **The British Library DataCite service is DataCite UK**.

Use of DOIs for data citation has since become recommended practice, but as yet very few longitudinal studies in the UK are providing DOIs for their users to cite study datasets.

DataCite UK wanted to understand why this is the case, and with the support of the CLOSER project held a workshop at the British Library in May 2018 to discuss and explore the issues. Ultimately, we want to ensure that UK longitudinal research can be appropriately cited, increasing its reach and impact and our ability to understand and measure that reach and impact.

Attendees of the workshop included: representatives of UK longitudinal and cohort studies; representatives of the UK Medical Research Council, Economic and Social Research Council and Wellcome, as the funders of longitudinal studies; DataCite UK; and the UK Data Service. The aims of the workshop were to highlight the benefits of DOIs for the citation and management of longitudinal

[1] DataCite is an international non-profit organisation that provides persistent identifiers (DOIs) for research data: https://www.datacite.org

datasets, examine the impact and implications that DOI implementation may have on studies, and look at the work that might be needed to realise the benefits of having DOIs for data citation for longitudinal studies.

Community benefits

All workshop participants agreed that one of the strongest benefits of having DOIs for longitudinal study data are the improved metrics and ease of finding out where study data are used. The burden of reporting on data use is also diffuse, as any person or organisation is able to make use of tools such as Google Scholar or Altmetrics to look at a study's data citations. The effort required to do so is much less than that required to search for less standardised and esoteric forms of data acknowledgement and citation. Other community benefits discussed included the ability of funders to understand which funding decisions had the greatest impact on social and biosocial research and policy (as measured by citation of the data demonstrating its use and re-use); better reproducibility and trust in research or decisions based on specific data; and better context for data in the research landscape.

Implications of implementation

The workshop demonstrated that implementation of DOIs would be easier for some studies than for others. Some studies archive their data with the UK Data Service and so already have DOIs for the data and can start to realise the benefits. However, studies where data are held locally in a relational database, and extracted on a case-by-case basis, will have more difficulty in implementing DOIs. At a local level, they require additional support in planning for implementation, to ensure versioning and granularity practices can be represented. Approaches to these may affect their data workflows and infrastructure.

Benefit realisation

Beyond local implementation, the benefits of DOIs can only be realised with a fundamental level of buy in from the research community, funders and publishers. Researchers need to cite data with DOIs as requested by studies to enable citation and impact metrics. Funders and publishers can help to ensure this through (financial) support and appropriate mandation. Tools such as CrossCite can join-up citation practice across scholarly communication. These approaches can be reinforced if, as a community, longitudinal studies enable best-practice citation with DOIs.

A number of actions were discussed and DataCite UK has refined and extended these based on needs articulated at the workshop. There are recommendations for a number of stakeholders, and the British Library, through DataCite UK, will continue to work with CLOSER, the UK Data Service, funders and publishers on the implementation of these recommendations and realisation of the benefits of citing study data with DOIs.

Recommendations

1. In support of best practice citation of data, and to enable better metrics on data use, longitudinal studies should implement DOIs for their data. The first step will be to consult with their host institutions to make use of existing DataCite UK accounts, and assign at least high-level DOIs for

acknowledgement and citation of the study's data in a broad sense, before working to implement DOIs for data that enable reproducibility**.**

2. DataCite UK, CLOSER and the UK Data Service should communicate the benefits of DOI use above current practice for data citation and metrics for longitudinal data, and support the studies to encourage adoption and implementation.

3. To enable longitudinal studies to apply DOIs appropriately, DataCite UK should work with CLOSER and the UK Data Service to develop and provide specific guidance to studies on the application of DOIs, with particular reference to versioning and subsets of their data**.**

4. To promote further use of DOIs, CLOSER should continue to display DOIs within Discovery and where appropriate within other tools and services that it develops.

5. DataCite UK, CLOSER, and the UK Data Service should work together to examine further opportunities to realise the benefits of DOIs for the studies. A first step could be an initial test of DOI-based reporting to examine the reach and impact of the studies and their data.

6. DataCite UK should work with CLOSER, the studies, UK Data Service, key publishers, UKRI, Wellcome Trust and other relevant stakeholders to support practice change for data citation and realise the benefits of DOIs for longitudinal studies and their users.

Section 4 provides further as to what these recommendations would require, and who would be involved.

Workshop Report

1. Introduction

Longitudinal studies in the UK have been making their data available to researchers beyond their immediate staff for decades. The studies have developed various models of data sharing and citation over this time that have enabled broader analysis and knowledge to be derived from the data. In general, data sharing across research has become a more formal activity, with funders, publishers and institutions alike reinforcing the expectation that data from publicly funded research will not only be shared, but also cited, acknowledging the creators and contributors to that data.

DataCite was created in 2009 to support data sharing with better practice in data citation using **Digital Object Identifiers (DOIs)**. DOIs enable consistent data citation and researchers are thereby able to share their data with less fear of being scooped or their contribution not being recognised. In 2014, best practice for data citation was formalised in to the Joint Declaration of Data Citation Principles[2], eight principles for the purpose, function and attributes of good data citation. Use of DOIs supports a number of these principles but as yet, few longitudinal studies in the UK identify their data with DOIs.

DataCite UK[3] wanted to understand why this is the case, and so proposed to hold a workshop exploring the issues with CLOSER. CLOSER brings together eight longitudinal studies to maximise the use, value and impact of longitudinal research. It is funded by the Medical Research Council (MRC) and Economic and Social Research Council (ESRC).

On May 4, 2018, DataCite UK hosted a workshop at the British Library for CLOSER stakeholders, providing an introduction on the use and benefits of persistent identifiers such as DOIs. The workshop had the following aims:

- Highlight the benefits of using DOIs for enabling citation of and links to longitudinal study data;
- Discuss the impact of the use of DOIs on the studies and their users;
- Investigate the implications of assigning persistent identifiers to study data and the steps needed to realise benefits.

The workshop provided an introduction to DOIs, followed by case studies of use. This session led into a discussion to uncover benefits that had not previously been identified or articulated by the existing DataCite UK community. Following examples of DOI implementation, further discussions examined what barriers or obstacles might be impeding DOI uptake and implementation by longitudinal studies. The workshop ended with an exploration of how various stakeholders, including funders, publishers, studies, individual researchers and DataCite UK itself could improve DOI uptake. In particular, it asked: what are the 'next steps' for studies to start the process of DOI implementation?

[2] Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014 https://doi.org/10.25490/a97f-egyk

[3] Run by the British Library as a DataCite member providing DOI services to organisations in the UK

This report summarises the information presented and the discussions heard at the workshop, and provides a set of recommendations that address some of the issues preventing the application of DOIs to longitudinal data by studies in the UK.

2. Benefits of persistent identifiers for longitudinal study data

DOIs are a persistent identifier subject to an international standard[4]. DOIs enable data citation by providing a unique link to a digital object such as a journal article or dataset, a link that will always take the user to the same object, even if its online location changes. This is of particular benefit to long-lived cohort studies that have to manage references to data that remain relevant beyond the timescale of any one website or following changes to a website's structure.

DOIs are now an established, trusted way of making datasets findable, accessible and re-usable. They support best practice for data citation (the Joint Declaration on Data Citation Principles) and the FAIR data principles[5]. Even where data are restricted or are not currently available, DOIs provide a link to an authoritative source of information about that data.

DOIs allow tracking of the use and re-use of the data, providing a verifiable means of data citation and referencing. As they are globally unique, they can be used in searches to find out who is using or citing data that any given DOI is assigned to. Along with the tools being developed by DataCite and other organisations, this means that their use can dramatically cut the effort and cost involved in measuring the citations of longitudinal data.

Use of DOIs for data further enables tracking of secondary uses, including use of data in novel ways in other disciplines. The use of DOIs for citation and tracking can therefore enable collaborative and interdisciplinary research, as DOIs enable data discovery outside of the remit of a particular study. It is possible to search based on study names for data reuse, but this is time consuming and inefficient. As an example, the Inter-university Consortium for Political and Social Research (ICPSR) estimates they annually allocate 2.5 FTE to uncovering citations to their data, purely because of poor citation practice[6].

Where DOIs are assigned to waves or subsets of longitudinal data, this kind of tracking can also help studies to understand which particular subsets of their data are cited. Such insight is a valuable way of feeding back on outcomes to participants, and may also support planning of future data collection. Quantified usage information based on DOIs will provide studies and their funders with improved insights into their users, where they are publishing and how study data are being combined with other datasets.

---

[4] International Organisation for Standardization (2017) *Information and documentation -- Digital object identifier system*. ISO 26324. Available from https://bsol.bsigroup.com/Bibliographic/BibliographicInfoData/000000000030177056

[5] Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, 160018. https://doi.org/10.1038/sdata.2016.18

[6] Moss, E; Lyle, J. (2018). Opaque data citation: Actual citation practice and its implication for tracking data use. http://hdl.handle.net/2027.42/142393

Of particular importance for longitudinal data that is continually evolving is accurate citation of data subsets at specific points in time. DOIs support the versioning and time stamping of data, and the identification of subsets within larger datasets. This enables research reproducibility, and so trust in the outcomes of research. In addition, it allows for research credit to both the study and to individual creators of the data who have worked on the study through time.

Longitudinal studies put significant effort into collecting, processing, documenting and managing their data to make it available for researchers beyond the studies themselves. The metadata supporting DOIs, particularly around the creators and contributors to data, means that when data are cited, individuals as well as the studies themselves can be acknowledged for their contributions. Furthermore, DOIs enable discovery and citation (and therefore increased research impact) as studies will reach the notice of potential users who may not have previous knowledge of the studies or their research. This may result in increased opportunities for funding and collaboration.

For funders the improved tracking of data use has follow-on benefits. It will help them to identify case studies for impact as well as providing a decision-making tool with fairness, transparency and accountability. In conjunction with ORCID iDs[7] and article DOIs, data DOIs can provide an additional dimension to tracking, to cover not just what is used and how much, but also by whom, for what and in what discipline. Beyond tracking, the links between DOIs will enable funders to visualise how the studies interact and collaborate, either at an official level but also where researchers may use data from multiple studies in their analyses.

Users of longitudinal data will also benefit. They will be able to trust that data they cite is robust and will not change, which also enhances the trust in their work. DOIs will also help them see where their outputs sit in the research landscape with others that have used the same data. This may lead to increased opportunities for collaboration and will certainly increase the visibility of their own work.

In summary, DOIs provide a robust international standard for data citation, discovery and tracking.

We recommend that DataCite UK, CLOSER and the UK Data Service continue to articulate the benefits of DOI use to the CLOSER studies, their users, their funders and key academic publishers, and support their adoption and implementation.

3. Implementation of DOIs and Persistent Identifiers in CLOSER studies to realise the benefits

In order to realise the benefits of assigning DOIs to data, some further developments in both research culture and infrastructure are needed. The studies associated with CLOSER fall into two broad groups in terms of their approach to data access. The first group provide datasets to the UK Data Service. These can then be accessed subject to one of three different licence agreements. The majority of datasets are subject to a licence agreement that allows the whole dataset to be downloaded by the end user.

[7] A globally unique and persistent identifier for disambiguating researchers, https://orcid.org

The second group manage access to secondary data users themselves. They provide users with bespoke subsets of data on request, rather than whole datasets. The process of accessing data for these studies typically involves users providing the studies with a summary of their research questions to assess if data from the study are appropriate for use in their research. The study then works with the user to determine the data variables and timeframes relevant to the work to ensure that provision of access is appropriate. The studies then extract and provide this data to the user under strict access and use terms.

In discussing implementation of DOIs, CLOSER workshop participants raised the broad issue around a lack of guidance on how to create DOIs and apply them to data, particularly with regard to creating DOIs at varying levels of granularity. The DataCite UK team gave an overview of this for workshop participants, which includes advice on how best to use metadata to represent the relationships between data subsets, as well as when new DOIs might be recommended for updated data.

We recommend that DataCite UK and the UK Data Service provide improved guidance on implementation of DOIs specific to longitudinal studies.

Some of the content of such guidance is summarised below and this document could therefore serve as a starting point for guidance development in collaboration with CLOSER, the studies and the UK Data Service. Such guidance would need to take account of the governance and processes that the longitudinal studies already have in place for data discovery, access and use.

3.1 Getting started

The first step in using DOIs is to understand how to work with DataCite UK.

DataCite UK works with organisations to ensure that responsibility for managing and maintain DOIs is embedded within a stable part of the organisation. As of June 2018, 90 UK organisations are DataCite UK users, including over 70 higher education institutions. Faculties, departments and projects should work with their host institutions to access DataCite UK services. Of the eight CLOSER studies, all are hosted by an institution that is a DataCite UK client, and so will find it easier to access DOIs. Staff at host institutions are also another potential source of guidance on the use and application of DOIs.

DataCite UK can help to support the relationship between both the study and institution in implementation of DOIs for longitudinal studies. Studies can contact DataCite UK if they are unsure of their local institutional contact

3.2 Meeting requirements

There are a number of mandatory requirements for applying DOIs to data. These are:

- Provision of an open landing page that the DOI takes users to;
- Provision of mandatory metadata[8] for each DOI;
- Ability to provide and maintain the data for the long term;

[8] DataCite Metadata Working Group. (2017). DataCite Metadata Schema Documentation for the Publication and Citation of Research Data. Version 4.1. DataCite e.V. https://doi.org/10.5438/0014

- Maintenance of the DOI in perpetuity.

Host organisations contractually commit to these in their arrangement with DataCite UK and so must ensure any studies they work with can also meet them. There were no issues raised by workshop participants representing the studies around the provision of metadata, or maintenance of DOIs. The ability to provide landing pages and maintain data for the long term were only seen as an issue in conjunction with discussions around assigning DOIs to data subsets and versioning.

3.3 Versioning and subsets

Assigning DOIs to new versions of data and to data subsets enable reproducibility and trust in research, and is also required under the Joint Declaration on Data Citation Principles. The Research Data Alliance (RDA) working group on data citation has reported on approaches to citing dynamic, changing data[9]. The UK Data Service has a rule-based definition of 'significant change' to a dataset that enables a new version of the data to be created and a new DOI to be assigned when required. Significant changes include addition of variables, corrections of miscoded data, changes in data format, changes in access conditions and significant changes in documentation.

However, being able to assign DOIs at a granular level according to the RDA recommendations is a particular difficulty for some of the studies. Assigning DOIs to new versions and subsets of data require them to be able to provide metadata and a landing page for each DOI as well as ensuring that the data can be retrieved exactly as cited. The ability to assign and manage DOIs at this level may require a rethink of a study's data management infrastructure and processes.

DataCite UK working with the UK Data Service should provide additional guidance on data versioning and subsets, and work with studies to understand what changes, if any, might be necessary to implement citation.

This guidance should adhere to existing RDA recommendations as far as possible.

3.4 High-level DOIs

One first step towards data citation is to ensure each study has an over-arching DOI that enables all the study's data to be cited as an intellectual entity. Although this does not support research reproducibility *per se*, this 'umbrella' or 'mother' DOI still enables some benefits, such as acknowledgement and tracking, and does not require the study to change any existing workflows or data publishing practices.

This mother DOI can be linked to the DOIs assigned to subsets of the data at varying levels of granularity. The relationships between the mother DOI and any associated DOIs does not need to be made explicit through the DOI name itself, but should be made explicit through the metadata submitted to DataCite on the creation of a DOI, particularly the use of 'relatedIdentifier' fields. This

[9] Rauber, A., Asmi, A., Uytvanck, D. V., & Proell, S. (2016). Data Citation of Evolving Data: Recommendations of the RDA Working Group on Data Citation (WGDC) (Version 1). https://doi.org/10.15497/rda00016

aspect of the metadata can also be used to indicate relationships between dataset DOIs and DOIs linked to publications such as journal articles, conference proceedings or other research outputs.

Several studies present at the workshop agreed that their first step would be to assign an over-arching DOI for the study's data as a single entity.

3.5 Implementation

For some studies there are numerous points at which DOIs could be implemented, both in terms of workflow, and in terms of who could assign the DOIs. For example, four of the CLOSER studies already send data to the UK Data Service, and so their data already obtain DOIs for citation. Some data held at the UK Data Service are the only source of data for secondary users and so these studies may not need to implement DOIs themselves.

Where access to the data is provided outside of the UKDS and is highly mediated, this will not be an option. In such a scenario, new study data are usually added into a relational database, changing the whole dataset, and so applying DOIs to discrete versions is not possible. However, the guidance on dynamic data citation suggests an approach of providing DOIs to extracted subsets. Extracted subsets could be stored and made available separately from the relational data. Alternatively, the queries used to extract the data could be assigned DOIs. Providing the queries are able to persistently extract identical data this would support reproducibility.

Such changes may require technical resource and developments that funders could support, and DataCiteUK and UK Data Service could help on some aspects of implementation.

3.6 Using DOIs as a foundation for metrics

Once applied, DOIs can be used to collect metrics on the use of longitudinal data. Both traditional and alt-metrics can be collected using DOIs as a globally unique way of identifying citations or mentions of the data. Funders and the studies themselves can use that information to look at the impact of the data.

Where the guidance for data versioning and subsets with DOIs is followed as per section 3.3 and studies created both an over-arching DOI and DOIs for query- or time-based subsets of the data, there is a risk that studies will end up with multiple DOIs, even numbering in the hundreds, to track for use metrics. This may result in double counting of citations if a user decides to cite the over-arching DOI as well as those for multiple other granular DOIs for individual datasets. This highlights a need for being able to identify duplicate citations when tracking usage of a study's data, although the true implications and scale of this problem will not be known until longitudinal studies start to assign DOIs in this way and users cite the data according to the Joint Declaration on Data Citation Principles.

DataCite UK through DataCite itself can represent the needs of and concerns of longitudinal studies on this matter. This also requires studies and their funders to engage with DataCite UK on their wider metrics needs.

3.6 Changing practice

Use of DOIs for data citation is now the mainstream in many areas of the life sciences e.g. with Springer Nature[10] journals. In contrast, the use of DOIs for dataset citation in longitudinal study based research is still emerging. A number of steps are required to incentivise change for practice in this regard. This situation is somewhat analogous to the establishment of accession numbers in bioinformatics, when it only became the norm as key publishers and funders insisted on their use. This form of citation now provides a more stable, standard practice across the discipline.

Workshop participants highlighted that host institutions, funders, journals, professional bodies and the studies themselves all have a responsibility for encouraging researchers to cite data responsibly and consistently.

Studies can provide standard or configurable citation recommendations for instance with the use of the CrossCite tool[11] and journals can include it as part of normal reference checking. The UK Data Service supports the users of social data with guidance in data citation, and this guidance can revised and extended to cover the citation of longitudinal study data. Funders and publishers can review their guidance to grantees and authors to join-up the practice of data citation from grant award through to tracking research outputs and longer-term impact.

DataCite UK can work with CLOSER, the Studies, UK Data Service, UKRI, Wellcome Trust and key publishers to support practice change for data citation across the longitudinal study research domain.

3.7 Sharing best practice

As more studies implement DOIs, there may be a diversification of how they are applied. To maximise on any investment in DOI implementation, the CLOSER Studies should continue to collaborate and share what they learn and what they have done. If they keep each other updated on the implementation of DOIs into their research workflows these will act as case studies that reach beyond CLOSER to other longitudinal studies.

DataCite UK can work with CLOSER, the studies and UK Data Service to help develop and share this knowledge base of best practice for data citation in longitudinal studies' research.

[10] Springer Nature data availability statement: https://www.springernature.com/gp/authors/research-data-policy/data-availability-statements
[11] https://citation.crosscite.org

4. Next steps

Six recommendations have been proposed as a result of the workshop, involving a number of different stakeholders, and are set out on page 2. Further details of the next steps required to achieve these recommendations, and who would be responsible, are set out in the following table.

Responsible party	**Recommendation**	**Specific next step(s)**
Longitudinal Studies	Apply DOIs to their data. Includes over-arching DOIs for the study's data as a whole, as well as more granular DOIs. Provide users with recommended or configurable data citations that include DOIs, and encourage use of them for data citation. Update user guidance and use CrossCite tool to reflect best practices for data citation.	Contact their host institution for information on how to make use of the institution's DataCite UK Agreement (DataCite UK can provide details of institutional contacts). Apply at least an over-arching DOI for the study's data. Feed in requirements around DOI-based metrics for longitudinal data reuse to other responsible parties.
DataCite UK	Work with UK Data Service to develop guidance on the implementation and use of DOIs specifically for longitudinal studies. This guidance will include specific details on assigning DOIs for data versions and subsets. Work with studies to understand what changes they might need to make to assign DOI according to best practice. Represent the needs of UK longitudinal studies on metrics driven by DOIs, and especially for data versions and subsets. Communicate the benefits of DOI use above current practice. Work with UK Data Service and CLOSER to examine further opportunities to realise the benefits of DOIs for the studies and their data.	Discuss how CLOSER and UK Data Service can feed in to and support this guidance.

Responsible party	Recommendation	Specific next step(s)
UK Data Service	Work with DataCite UK to develop guidance on the implementation and use of DOIs specifically for longitudinal studies. This guidance will include specific details on assigning DOIs for data versions and subsets. Guide and monitor emerging practice on assigning DOIs to longitudinal datasets, and apply to own holdings of longitudinal study data. Communicate the benefits of DOI use above current practice. Work with DataCite UK and CLOSER to examine further opportunities to realise the benefits of DOIs for the studies and their data.	Revisit and revise existing data citation guidance and include or develop new guidance specifically for citing study data. Reassign DOIs to match community practice on longitudinal data.
CLOSER	Ensure DOIs are visible within appropriate tools and services that it develops. Communicate the benefits of DOI use above current practice. Work with UK Data Service and Datacite UK to examine further opportunities to realise the benefits of DOIs for the studies and their data.	Discuss with DataCite UK for guidance.
Study Funders	Provide resource for implementation and development of study infrastructure to accommodate DOIs. Begin to use DOIs to examine the reach and impact of studies and their data. Data policies to advise use of persistent identifiers. Actively encourage and expect best practice citation of longitudinal data, using DOIs.	Discuss with studies to understand what support might be necessary and appropriate. Feed in requirements around DOI-based metrics for longitudinal data reuse.

Responsible party	Recommendation	Specific next step(s)
Host Institutions	Encourage and provide guidance on best practice citation of longitudinal data, using DOIs.	Liaise with DataCite UK for guidance on use of DOIs as an integrated infrastructure for research data citation, discovery and tracking
Publishers	Encourage best practice citation of longitudinal data, using DOIs and reference checking to ensure best practice is followed.	Open dialogue with journal editorial boards relevant to CLOSER studies on DOI uptake and adoption.
Professional bodies	Encourage best practice citation of longitudinal data, using DOIs.	Liaise with DataCite UK for guidance on use of DOIs as an integrated infrastructure for research data citation, discovery and tracking