Enabling the citation of datasets generated through public health research

Jonathan Rans (DCC), Michael Day (DCC), Monica Duke (DCC) and Alex Ball (DCC)

Contents

1 Executive Summary 3

2 Introduction 4

3 Aims and Methods 5

4 Current Mechanisms in Data Preservation and Publication 6

5 Current Mechanisms in Citation 8

6 Recommendations 12

7 Recommendations for the PHRDF 15

8 Conclusion 18

9 Acknowledgements 18

10 References 18

11 Further information 21

Appendix One: Dryad 23

Appendix Two: Dataverse 25

Appendix Three: Archival Resource Keys and EZID 27

Appendix Four: DataCite, Digital Object Identifiers (DOIs) 29

Appendix Five: International Standard Name Identifiers 31

Appendix Six: Open Researcher and Contributor ID 32

Appendix Seven: ORCID DataCite Interoperability Network 33

Appendix Eight: Webtracks 34

Appendix Nine: Focus Group Notes 35

1 Executive Summary

There is a growing recognition of the value of research data as part of the scholarly ecosystem; infrastructure supporting the role of data in the academic discourse is starting to become established, of which mechanisms that enable the citation of digital data objects form a fundamental part.

The Wellcome Trust, on behalf of the Public Health Research Data Forum (PHRDF), commissioned the Digital Curation Centre to produce an examination of the current mechanisms enabling the citation of datasets generated by public health research. Information was gathered through desk research and interviews with key individuals working with data citation and was then presented to a panel of experts from the field of Public Health and Epidemiology.

These exercises informed a raft of recommendations for future work to be carried out by the PHRDF:

- The PHRDF should develop guidance on citation formatting and the target object for citation.
- The PHRDF should develop guidance on different persistent identifier schemes and investigate the possibility of establishing an umbrella organisation for assigning identifiers to public health datasets.
- The PHRDF should develop guidance for dataset owners to enable them to provide exportable citations in a standard format.
- The PHRDF should engage with researchers and publishers to encourage growth of citation infrastructure and adoption of best practice.
- The PHRDF should investigate initiatives developing methods of fine-grained attribution of credit for those working with public health datasets and potentially engage with them to help develop disciplinary standards.
- The PHRDF should investigate and engage with projects developing tools that enable versioning and fine-grained citation of datasets.

This report examines the opportunities for promoting dataset citation in the field of Public Health and Epidemiology. We provide a landscape review of the current initiatives addressing data publishing and data citation. Drawing on consultation with experts in the field of Epidemiology research we examine how these initiatives can be applied to the specific challenges associated with datasets generated in this area. We provide recommendations for future development, identifying priorities for further consultation.

2 Introduction

> "We need to create a culture within the research community that recognises the value of data sharing and recognises the contributions of those who do it well ... We also need to build the resources and tools to enable researchers to share and access data effectively, and which minimise the time and effort involved. ... we need to ensure that datasets of value are discoverable to potential users so that they can readily locate and access these resources."

- David Lynn, Head of Strategic Planning and Policy at the Wellcome Trust
  http://www.dcc.ac.uk/news/idcc11-preview-interview-david-lynn

Data is recognized as a valuable component of the modern scholarly ecosystem; infrastructure to embed data within the fabric of scholarly communication is emerging alongside more traditional journal publications. That infrastructure must contain elements to support the storage, access, discovery and re-use of data; citation mechanisms are a fundamental component of this. They must echo the role that traditional, journal citation has played in ensuring longevity of the scholarly record, acting as a bridge to permanent access and enabling reward systems.[1, 2]

Parallel to the development of the technical landscape, changes in grassroots practice and culture will be required to make data-driven citation an integral part of the ecosystem. There is a need to collaboratively develop discipline-specific guidance and standards, and to engage in advocacy with researchers, publishers and other stakeholders to ensure that best practice is embedded into workflows.[3]

The emerging infrastructure ranges in ambition from step-wise improvements in existing services to visionary international initiatives. Electronic journals which allow supplementary data to be submitted alongside journals papers, or to be linked to in a reference section, are an example of the former. Data journals, that move the focus onto data itself, rather than the outputs of its analysis, are a more recent (but growing) phenomenon. Whilst local or disciplinary-focused solutions are appropriate in some cases, the global and interdisciplinary nature of scholarly communications will make or break the success of other developments. The utility that can be leveraged from persistent identifier schemes, for example, is likely to be decided by the level to which they integrate with other academic systems. Repository and storage layers, which provide the base on which end-user services are built, can incorporate features designed to support data citation.[4, 5]

Any evaluation of citation initiatives in this rapidly-moving environment must consider not only the likely success of an isolated effort but also the potential for continued integration with the rest of the landscape.

Ultimately, the citation infrastructure must fulfill the following two broad functions to be considered a success:
- facilitate access to data (including discovery)
- enable technologies that track data, including those driving reward mechanisms

More granular requirements have been identified as components of these broad aims; these include: identification of datasets (and their subparts), mechanisms for locating and accessing the data, sufficient information to assign and compute credit, contextual information for the human user and supporting machine-readable information for consumption by automated agents.[6]

Technical solutions can only provide part of the answer; community guidance and agreement will be instrumental in driving change, and cultural barriers and drivers must be identified and addressed.

3 Aims and Methods

This report was commissioned by the Wellcome Trust on behalf of the Public Health Research Data Forum (PHRDF), an international organisation bringing together major funders of global health research. The PHRDF is dedicated to increasing the availability of health research data generated through their funding in ways that are equitable, ethical and efficient.

Our objectives in conducting this work were:

1. To provide the PHRDF with an up-to-date review of key initiatives enabling the citation of datasets generated through research.
2. To assess the opportunities and challenges for the application of emerging data citation mechanisms in the field of public health research.

To this end, we conducted a landscape review of general mechanisms used and emerging in the areas of data publishing and citation. We examined the aims and objectives of the initiatives; their coverage, both geographical and by discipline; and the services they provide. This information was gathered through desk research and interviews conducted with key individuals in the field and are presented in sections 4 and 5, and Appendices 1-8.

Findings from the landscape review were presented to an expert focus group of researchers and data managers from the field of Public Health and Epidemiology. The group's reactions to these findings informed a series of recommendations and priorities and identified potential challenges and incentives for implementing data citation mechanisms. A summary of the focus group discussion is presented in Appendix 9.

4 Current Mechanisms in Data Preservation and Publication

The traditional method for publishing research data, of attaching it to journal articles as a supplementary file, is growing increasingly unsuitable for supporting modern research outputs. File sizes and types are commonly restricted, data cannot be independently cited, and there is a lack of assurance relating to curation and preservation. There are a number of established and emerging mechanisms that address these issues; we examine some of the prominent ones here.[7]

4.1 Data Repositories

There are a huge variety of data repositories that can be used to archive, curate and expose datasets produced as an output of research, some associated with institutions and others with disciplines. All offer varying degrees of permanence and curation for deposited data.[8] Public health datasets, by their nature, may be incompatible with these systems, which are generally designed around providing open access to static datasets; however there are some interesting features worth noting.

Some public repositories, for example Dryad and Pangaea, have engaged vigorously with publishers to ensure that data deposited in the archives not only links to the article it supports but is also linked to from the article itself. This bi-directional relationship fosters visibility and enables tracking mechanisms, realising the full benefit of the citation.[9, 10]

The Dataverse system is a repository-building software tool that enables individual repositories to link into “dataverses” connected by a common discipline or organisation. Plugging in to a wider network develops a critical mass of disciplinary research, potentially improving visibility. This system offers persistent identifiers in the form of handles and has various data visualisation and manipulation tools included in the software. Of particular interest is their mechanism for validating cited datasets by including a check-sum style identifier, called a Unique Numeric Fingerprint (UNF) with each download. Researchers wishing to re-use a particular, cited dataset can check the cited UNF against the one provided with their own download to verify that the two are exactly the same.[11]

4.2 Enhanced Journals

GigaScience
A new approach to the publication of supplemental data is provided by the relatively new, open-access, open-data online journal GigaScience. Serving the biomedical and life sciences GigaScience places no restrictions on the size of supplemental files and provides any dataset in its database with a DOI assignment, allowing independent citation and aiding data visibility. Their aim is to provide a publishing option for difficult, “big-data” studies including cohort databases. At present all costs associated with publishing in GigaScience are defrayed by the Beijing Genomics Institute.[12]

Acta Crystallographica
Another approach to data publication is provided by the Acta Crystallographica Section E: Structure Reports Online, an open access megajournal publishing Crystallography data that ‘blurs the distinction between a journal and a database’.[13] Separating the data from the standard publishing process allows for rapid dissemination of results with an average publication time of less than one month.[14]

4.3 Data Journals

Data journals offer a forum for discussing the methods employed whilst planning and carrying out data collection, and subsequent non-trivial manipulations of the raw data. Analysis of results is not covered in the journal, which confines itself to high-level discussions of the dataset. This approach has had considerable impact in some fields; the Earth System Science Data journal being one example of a well-established publication with an open peer-review system.[15] The closest example to this kind of publication covering public health would be the International Journal of Epidemiology, which publishes descriptions of cohort databases.[16]

4.4 Alternate Data Publishing

Figshare
There is increasing recognition that the requirement to publish data that underlies published research leaves a great deal of data produced in the course of academic work in a grey area in terms of preservation and exposure. Figshare is an initiative that aims to address that issue by providing free, unlimited storage for this long-tail of data. Datasets can be uploaded without restriction; each receives a DOI, making it citable, and the platform provides visible metrics covering views and downloads of material.[17]

Nano-publication
A nano-publication takes the form of a richly annotated RDF-triplet of the form subject-predicate-object, for example "breast cancer>is a form of>cancer". It represents an unambiguous statement based on summary data and is citable in its own right. In order to automatically produce such a statement from a dataset or publication it is necessary for the discipline to have formal ontologies enabling precise definition of terms and relationships.[18]

Efforts are being made to publish a subset of the Genome Wide Association Study (GWAS) database by producing a collection of nano-publication statements. Application of controlled vocabularies such as the Medical Subject Heading (MeSH) and ontologies such as the Human Phenotype Ontology (HPO) to the data enables standardised statements to be produced.[19]

5 Current Mechanisms in Citation

Persistent identifiers offer a method for providing an unambiguous description of a given object by assigning it a unique numeric identifier to which descriptive metadata may be attached.
This robust identification can then be leveraged by systems that locate references to the object across the internet and others that search and harvest associated metadata.
In the context of data citation the objects we would wish to unambiguously identify are datasets and researchers.

5.1 Dataset Identification

There exist several schemes for assigning persistent identifiers to datasets, of these the ones with the most maturity and coverage are Handles, Archive Resource Keys (ARKS), Persistent URLS and Digital Object Identifiers (DOIs.) These address well known issues with the use of URLs as object location identifiers by providing an extra layer of administration that confers permanence to links citing a given object. It should be noted that persistence relies on the owner of the data object ensuring that associated metadata remain current and so the use of any of these schemes entails an on-going commitment of resource to administrative care.[20, 21, 22]

Typically a persistent identification scheme will consist of a naming authority; a resolution service that returns metadata about named objects, including its location; and associated protocols.

Key features of these four schemes are as follows:

Handle
Established in 1994, the Handle system was originally developed to provide unique identifiers to digital library objects. It represents a tried and tested naming and resolution service that supports the assignation of metadata to objects it identifies. With a wide degree of penetration, there can be a strong level of confidence that it will persist into the future. This persistence is supported by the system's independence from the Domain Name System (DNS).[23]

DOI
The Digital Object Identifier (DOI) was publically launched at the end of October 1998 and was originally used to identify publications. Since then, new naming authorities have been established that assign DOIs to other objects with the DataCite Organisation holding responsibility for assigning DOIs to datasets.

The DOI system is an administrative framework that defines common standards and practices. For the practicalities of naming and resolving identifiers, DOI utilises the Handle system but layers extra services on top of the base system. It provides a framework for the interaction of objects and services by assuring their compatibility. The DOI system is the only persistent identifier that has attained ISO standard accreditation. As it is based on the Handle system and has considerable penetration in its own right, there is little chance of the system disappearing in the foreseeable future.[24, 25]

ARK
The Archival Resource Key (ARK) was originally developed in 2001 as a method for identifying physical objects in an archive but has been adapted to make it suitable for persistently and uniquely

identifying digital objects. The focus of the ARK is on metadata and its delivery; the identifier allows access to object metadata, a statement on the persistence of the object and, of course, the object itself. In keeping with the idea that some objects have a limited life-span, ARKs can be destroyed. Technical barriers for the implementation of the ARK system are low and the system is DNS-linked. The ARK system enjoys wide-spread adoption in North America and has clusters of international clients. The level of uptake is such that the system is very likely to remain active into the future.[26]

PURL

The Persistent URL (PURL) system was initially implemented in 1996 and primarily focussed on providing a unique identifier that resolves to an internet location. The system allows the known history of object locations to be displayed on request. The software required to establish a PURL server and namespace is open source and freely accessible, meaning there is a low technical barrier to implementation and no reliance on a single resolving agency. However, federated responsibility for maintenance arguably makes the system more prone to link rot.[27, 20]

Of those four, arguably the scheme with the greatest amount of traction is the DOI. Originally used as a persistent identifier for electronically published content there now exist a number of authorities that assign DOIs to other digital objects. The DataCite organisation is responsible for assigning DOIs to research datasets hosted on the internet.

Individual researchers wishing to assign a DOI to their dataset do not go to DataCite directly but obtain one from the data centre they are hosting with. Data centres are required to register with DataCite before they can acquire DOIs for the datasets they hold. The centre must provide assurances that metadata will be made openly available and that data will be preserved according to fairly stringent DataCite stipulations. The rigorous quality control measures exercised over those using DOIs lend them a level of assurance reflected in their recognition as an ISO standard.[28]

Persistent identifier schemes tend to require an organisation wishing to assign IDs to register with a central agency. There is a cost associated with maintaining an ID service that may be passed on, through registration fees to the end user. For the Handle system there is a $50 registration fee followed by an annual $50 fee for every prefix that the organisation wishes to register. Each prefix can have an unlimited number of IDs associated with it. Registering to produce DataCite DOIs or ARKs varies in price depending on the organisation but is considerably higher than the Handle system registration.

5.1.1 ID Management Systems

A data centre can manage its identifiers through a third-party service such as the California Digital Library's EZID system. This service aims to simplify the process of assigning and managing persistent identifiers whilst also adding features such as the ability to reserve IDs before data has been published or even produced and allowing fine-grained control over metadata publishing. At present the EZID service enables the assignation of ARKs and DOIs, however, due to the restrictions on DataCite members only ARKs can be assigned to non-US users. The service is being used widely in the United States across numerous academic disciplines and also has users in parts of Europe.[29]

5.1.2 Presentation as part of a citation

Best practice is for persistent identifiers to be combined with the address of a resolver service, enabling electronic versions of the identifier to be robustly resolved to an internet location.

An example would be an identifier 10.1594/PANGAEA.762818 being coupled with the resolver address: http://dx.doi.org/ to give the persistent link: http://dx.doi.org/10.1594/PANGAEA.762818

5.2 Researcher Identification

There already exist well-established schemes providing unique identification of researchers such as ResearcherID or the Scopus Author Identifier; so far none has proved to have a broad enough scope to be suitable for global purposes such as attribution. There are two initiatives in development now that aspire to provide this reach; enabling next generation content management systems to leverage their framework. The schemes are currently separate but compatible, designed to be capable of future integration.[30, 31]

ISNI

The International Standard Name Identifier (ISNI) is a US initiative that aims to solve the disambiguation issues associated with identifying individuals in the media content industries. The scheme acts as a bridging service between proprietary party identification systems enabling collaborating industries, such as music and publishing, to identify individuals without the need to expose confidential metadata.[32]

ORCID

The Open Researcher and Contributor ID (ORCID) initiative offers an independent, discipline-agnostic database of academic researchers, unambiguously identified by a unique alphanumeric code. The system places control of the online profile in the hands of the individual researcher allowing them to populate and publish as much or as little information as they like. The software integrates with existing content management systems to facilitate profile population. Researchers creating a record do so for free; it is likely that organisations that wish to access the system will be subject to a registration fee.[33, 34]

5.3 Developing Citation-based Infrastructure

Establishing standards for the persistent identification of datasets is the first step towards constructing an ecosystem of technologies capable of leveraging this architecture to provide complex content management and integration systems.

ODIN

There are many projects addressing all manner of questions surrounding data citation, linkage and preservation. One with particular resonance to those working with cohort datasets is the ODIN project, an international collaboration between seven of the key organisations in the field of data publishing and citation.

ODIN officially started in October of 2012 and will examine the problems associated with complex citation of datasets from disciplines with known issues. As part of the work, ODIN will be engaging with researchers producing longitudinal datasets and developing approaches to aid the granular citation of dynamic datasets.[35]

Linkback Methodologies

There exist a variety of methods for website owners to receive notification when another site links to their own. A number of projects aim to develop similar tools for tracking links to datasets.

The trackback protocol was originally developed for blogging platforms to capture instances of re-blogging and linking. The CLADDIER and Storelink projects were JISC-funded projects that augmented this protocol for use as a citation notification system for repositories. This software enables the repositories to automatically collate citations for its digital objects.[36, 37]

The Webtracks project built on this work to produce a peer-to-peer protocol for linking collaboratively produced raw data via electronic notebooks; these aggregated links being a publishable, citable object.[38]

5.4 Measuring impact

Thomson Reuters

Thomson Reuters has announced the release, in late 2012, of the Data Citation Index on the Web of Knowledge platform. This index will include content from more than 80 established, curated repositories spanning multiple disciplines and represents a major step towards establishing datasets as an integral part of the scholarly record, capable of accruing credit.[39]

Altmetrics

Application of citation mechanisms enables not only the automatic discovery of journal citations but also the discovery of citations delivered through non-traditional publishing methods. Altmetrics tools enable researchers to track instances of dataset citation in blog articles, twitter feeds and other web-based publishing formats.[40]

6 Recommendations

6.1 What can be achieved?

We see the PHRDF's two, main aims in this exercise being the promotion of data sharing through increased visibility of high-value, cohort datasets and helping to construct a mechanism for tracking the impact of these datasets.

At present, dataset citation is not widely practiced in this discipline; this is to be expected, given that it is a relatively new concept and that cohort datasets have a number of fundamental properties that complicate sharing and citation.

Data producers rely, for attribution, on re-users incorporating an acknowledgement paragraph into their journal articles, a mechanism which can only be tracked manually at considerable effort. A move away from this system to one of formal citation would be welcome but most researchers would require guidance to make the transition.

The positive message is that there is a culture of data sharing in Epidemiology, albeit one that necessarily operates within restrictions. The opinion of the focus group was that clear benefits could be seen from implementing data citation mechanisms but that there was a considerable work to be done before those benefits could be fully realised.[41]

6.2 Specific issues surrounding the citation of Public Health datasets

As with any large academic field, Epidemiology is certain to have a great deal of heterogeneity in its associated datasets. One could not hope to identify all of the challenges that will be experienced by data producers in this area however, it is possible to suggest some of the common features that complicate efforts to cite and share datasets.

6.2.1 Location of the Datasets

Unlike other data centres that are designed to contain large numbers of separate datasets, institutions hosting public health data may only keep a small number of very large datasets. This may have an impact on the economy-of-scale approach that can be taken when assigning persistent identifiers, particularly relating to the use of DataCite DOIs where assurances are required from data owners relating to their standards and practices.

6.2.2 Size and complexity

Datasets produced during longitudinal studies can be extremely large, taking in huge numbers of subjects, each with considerable numbers of associated variables (e.g. height, weight, blood group etc.). The complexity of the data causes a number of issues both in terms of sharing with colleagues and in constructing a suitable citation.

For an academic to engage with work produced by a colleague is a non-trivial exercise, regardless of how close their research interests are. There is a huge amount of specific information surrounding any dataset that is vital to understand when developing a meaningful analysis. This is particularly true for public health datasets where context is all important and understanding why and how data were collected is key to knowing how to use them. Researchers will produce documentation describing datasets but even interpreting this without knowledge of the history of the project is very difficult, in many cases the knowledge is implicit and almost impossible to capture in writing.

The practical consequence is that data sharing is an exercise that, for many datasets, demands a commitment of resource *every time that it is undertaken*, both from the data producers and those re-using the data, in order to ensure that the science coming out of such collaboration is both valuable and robust.
Any move that increased the visibility of public health datasets would be likely to cause an attendant increase in the volume of requests for re-use. It is possible that dataset owners will be unable to attend to a major increase in data sharing applications without receiving specific allocation of funding council resources to do so.

From the point of view of citation, the complexity of the base dataset makes it difficult to unequivocally link to the subset of data that informed the work in which the citation appears. This is an issue of granularity that has been addressed by some projects; the ARK ID's suffix pass-through system enables datasets to be cited at different levels of granularity. The Dataverse system also enables the extraction of subsets of a main dataset and uses the UNF system as a validation should that subset need to be extracted again by another party. However, all of these methods are predicated on being able to return to a static, unchanging dataset.

6.2.3 Subject to regular updates

Typically databases used in longitudinal studies are not static items but dynamic resources, constantly being added to, refined, reassessed and updated.
This feature of the datasets presents issues for researchers attempting to cite them as it is difficult ensure that there is sufficient information included to locate the correct version.

One approach to this issue, adopted by the Dryad repository, amongst others, is to allow updates to datasets but to require that a whole new copy of the set is deposited, including revisions, which is then assigned its own DOI and linked to the original dataset to aid discovery. In the case of vast cohort databases it is impractical to consider keeping a copy every time that a change is made; indeed it is likely to be impractical to keep a copy even when a major revision has been carried out.

Despite this, there is still merit in retaining a record of the version of the database that was accessed, whether by formal versioning information or the less precise method of recording the date of access. Database owners should know what changes have been made to a database across versions and ought to be able to explain why a given analysis, run at two different times on the same database, returned different results. The question is whether it is necessary to be able to recreate a particular set of results or whether it is sufficient to be able to explain why the results are different. In either case it is necessary to be able to identify which version of a dataset produced a particular set of results.

This is an emerging area in which community standards have not yet been developed; the Data Documentation Initiative (DDI3) group are working on a standard approach to versioning datasets generated and maintained in longitudinal studies.[42]

6.2.4 Subject to Consent

There is an ethical imperative for data custodians wishing to re-use data gathered as part of a public health study to ensure that the original participants have consented to having their personal details used for this purpose.

In reality this can be a difficult issue to resolve with consent taking many different forms, verbal, written and implied by proxy, amongst others. This is further complicated by the timeframes that many cohort studies work across; consent given by a cohort member decades ago could not explicitly cover the myriad uses to which that data could be put today.

In cases where it is impossible to return to the subjects and obtain updated consent that specifically covers data re-use the approach adopted is to assume a level of implied consent on behalf of the subject. In practice, as consent was originally given for academic use the assumption is it can be extended to cover academic re-use. There can be a degree of discomfort, therefore, in making this data available without restriction as that will expose it to agencies, such as pharmaceutical companies, that the cohort might not wish to support. These objections could be quite separate from concerns about private information being put into the public sphere.

However, should a robust citation architecture be developed, improved tracking of outputs would enable dataset owners to provide cohort members with a much more detailed picture of the downstream uses of their data. This would help to demonstrate the positive impact that their contribution continues to make.

6.2.5 Subject to Concerns of Privacy

Naturally a major concern when dealing with sensitive personal information is protecting the privacy of individuals who have contributed data to a study. Opinions in the focus group varied widely on the scope for publishing datasets containing sensitive personal information.

The truth is that, if privacy were the only concern, most datasets probably could be pared of identifiable data and aggregated to the point where they would be safe to share with the public at large. The question is how much utility remains and does the effort of preparation justify the end result.[43]

Of potential interest in this area are initiatives that allow researchers to query restricted databases and return aggregated, non-sensitive results. The Dataverse project is currently exploring this area with a view to developing querying tools.

6.2.6 Ownership

Sometimes there are complicated issues of ownership. Local governments in countries where longitudinal studies are conducted can feel that they have ownership of the data that has been collected within their borders, recognising it as a valuable asset.

7 Recommendations for the PHRDF

There is a clear need to work with researcher and data managers to develop a tailored approach to data citation; this could be achieved through a programme of consultation exercises and technical workshops.

Once clear, practical guidance has been developed the PHRDF can use its influence to promote good citation practice ensuring that responsibilities are well defined, both for data producers, who need to make their data properly citable and for data re-users who have a responsibility to ensure that the datasets they have accessed are cited in accordance with disciplinary standards.

The content of the guidance and the issues that it will need to address will largely be decided by what the PHRDF and the community feel that the citation ought to achieve. Here we present the three main uses for a citation and the attendant issues that will need to be addressed to realise those benefits.

7.1 To provide a robust link between the dataset and the work that it supports, creating a visible, trackable element that can be used to assess impact.

7.1.1 Citation Target

The first issue to address is the target for the citation, best practice dictates that it ought not to point directly at the dataset itself but should instead point to a landing page for the dataset that either includes an onward link or guidance on procedures for accessing restricted data. Also included on the landing page could be a fairly broad description of the dataset and the research questions it aims to address along with other, non-confidential, high-level metadata. Guidance on suitable contents for the landing page should be developed in collaboration with data producers.

7.1.2 Format

At present the use of data citations is extremely variable, although half of journals recommend style manuals that include appropriate guidance, many articles lack adequate data citations.[44, 45] Good practice tends to be driven by discipline, with those that have a stake in open data forging ahead of the crowd.

To address this, the PHRDF could develop guidance on a suitable, standard format for the citation, and promote its use through policy development. If the citation simply has to provide a link between articles and datasets something fairly basic could be used, as the utility of the citation increases further elements would need to be incorporated.[46]

A variation on the DataCite standard format is given as an example:

Project (Access date): *Title*, Publisher, Identifier/Location

7.1.3 Persistent Identifiers

Guidance will be needed for data producers on the various methods for persistently identifying their datasets. Cost will certainly be part of the consideration but weight should be given to the level of integration displayed by the different schemes and the ease with which they can be implemented.

The PHRDF should investigate the need for a central facility enabling institutions hosting a single, live public health dataset to acquire persistent identifiers. If such a facility aspired to have an international reach it may be incompatible with some identifier schemes or require a federated structure. Such a scheme would not necessarily have to limit itself to the use of a single identifier type but might offer a range of options with advice to help researchers choose the most appropriate for their needs.

It should be noted again that persistence is dependent on the continued involvement of data owners in administering associated metadata records.

7.1.4 Developing exportable citations

One practice that is common amongst established repositories that could be of use to large dataset owners is that of providing ready-made, exportable citations in various formats. Dryad offers citations on dataset landing pages that can be cut and pasted or exported to citation management software in RIS or BibTex formats.[9] Figshare offers a cut and paste citation, which includes an automatically generated timestamp, and also exports references to Reference Manager, Endnote and Mendeley.[17] This effort achieves two aims, ensuring that the people best placed to populate the citation with up-to-date information are in charge and lowering the barriers for data re-users to include proper citations in their journal articles.

7.1.5 Engaging with researchers

There will be a need for best practice to be promoted to researchers through a programme of advocacy and policy development. This must promote awareness of the roles and responsibilities in relation to data citation and emphasise the obligation that they have to their funders and the data producers to discharge those duties.

7.1.6 Engaging with publishers

The PHRDF should engage with publishers of public health research to encourage them to allow data citations to be included in journals and in a way that enables machine discovery. Best practice in this regard is for citations to be included in the reference section rather than the text as its accessibility enables discovery by automated systems.

7.2 To enable the attribution of credit to those engaged in the production of public health datasets.

The general response of participants in our focus group was, while it would be a good thing to assign credit to people traditionally overlooked by standard attribution, there were serious complications involved. This is partially due to the sheer number of people who contribute to public health datasets but also due to the difficulty of retrospectively assigning credit to people who may have left the project decades before.

7.2.1 Decide who should be included on the citation

The PHRDF would need to work with the public health community to identify which individuals associated with the production of a public health dataset should be included on the citation. One suggestion that arose from the focus group was that only senior staff, both academic and technical, be included on the citation but that all other staff involved were made discoverable perhaps by being linked from the dataset landing page.

7.2.2 Investigate models of microattribution

There are on-going studies, such as the ODIN project, that are developing models for attributing credit to large numbers of researchers attached to a single project. The PHRDF should investigate these projects and potentially contribute to controlled vocabularies describing the unique roles and contributions of staff working on large public health datasets.[47]

As a corollary to this, methods of assigning persistent identification to researchers (e.g. ORCID) should be investigated as a way of unambiguously placing researchers into large attribution schemes.

7.3 To enable the reproduction/validation of research associated with the dataset.

In order to reproduce or validate work supported by a cited dataset it is essential that the citation is able to identify the exact set of data used by the researcher. For a large, dynamic dataset this presents two problems. The first is versioning, identifying the exact point at which a dataset has been accessed. The second is one of granularity, successfully identifying the specific subset of information that was accessed. Individually, these problems have been addressed in various ways but they interact with each other to complicate the situation.

7.3.1 Granularity and versioning

To address these issues the community needs to decide what level of reproducibility is practical or desirable, is it necessary to be able to run the same analysis on exactly the same data and return the same result or is it enough to be able to explain why the result is different? Identifying the needs of the public health community in this area will enable the PHRDF to contribute to existing efforts to tackle these problems.

The PHRDF should identify and engage with initiatives in the wider academic community developing standard versioning tools for complex datasets; for example, the work being done in this direction by the Data Documentation Initiative and the ODIN project.[35, 42]

Additionally, the PHRDF should investigate methods for providing granular citation, at present some datasets that require multi-level citation achieve this by breaking the data down into subsets and then assigning each subset its own identifier. Given the size and mutability of cohort datasets this may not be a practical approach, although the focus group indicated that in some cases subsets of data generated during collaborative work are saved separately from the main dataset and could provide a suitable hook for an identifier. As a corollary to this, it should be noted that tracking impact from one, large dataset would entail integrating impact metrics from multiple identifiers.

8 Conclusion

The field of dataset citation is fairly new yet already there are emerging mechanisms that appear robust enough to provide a base for future development. The PHRDF needs to collaborate with public health researchers to develop discipline specific standards and engage with the wider community to embed these standards in the citation landscape. Guidance and advocacy will be required to foster a culture of dataset citation.

9 Acknowledgements

Thank-you to Christine Borgman (UCLA), Todd Vision (Dryad), Todd Carpenter (NISO), Adam Farquhar (BL), Joan Starr (CDL), Mike Taylor (Elsevier), John Kaye (BL), Brian Matthews (STFC) , Geraldine Clement-Stoneham (MRC), Philip Curran (MRC Unit for Lifelong Health and Ageing), Judith Glynn (LSHTM) Debbie Hart (KCL Department of Twins Research), Catherine Jones (STFC), Christine McMahon (UCL Institute of Child Health), Victoria Vazquez (KCL Department of Twins Research), Sarah Jones (DCC), Angus Whyte (DCC) and Kevin Ashley (DCC) for helpful contributions.

10 References

[1] Ball, A. & Duke, M. (2012). 'How to Cite Datasets and Link to Publications'. *DCC How-to Guides*. Edinburgh: Digital Curation Centre. Available online: http://www.dcc.ac.uk/resources/how-guides

[2] Managing *and Sharing Data*. (2011). UK Data Archive [Report]. Retrieved on 17 October 2012 from: http://www.data-archive.ac.uk/create-manage

[3] *Open to all? Case studies of openness in research.* (2010). Research Information Network and National Endowment for Science, Technology and the Arts. Retrieved 13 September 2012, from: http://www.rin.ac.uk/our-work/data-management-and-curation/open-science-case-studies

[4] Hrynaszkiewicz, I. (2012) *Citing and linking data to publications: more journals, more examples… more impact?* [Blog Post]. Retrieved 17 October 2012 from: http://blogs.biomedcentral.com/bmcblog/2012/01/19/citing-and-linking-data-to-publications-more-journals-more-examples-more-impact/

[5] Edmunds, S., Pollard, T., Hole, B, Basford, A. (2012). *Adventures in data citation: sorghum genome data exemplifies the new gold standard.* BMC Research Notes 2012, 5:233. Doi: 10.1186/1756-0500-5-223 http://dx.doi.org/10.1186/1756-0500-5-223

[6] 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*. [Report]. Opportunities for Data Exchange, Alliance for Permanent Access. Accessed on 17 October 2012 from: http://www.alliancepermanentaccess.org/index.php/community/current-projects/ode/outputs/

[7] Piwowar, H. (2010) *Supplementary Materials is a Stopgap for Data Archiving*. [Blog Post]. Retrieved 11 September 2012 from the Research Remix blog:

http://researchremix.wordpress.com/2010/08/13/supplementary-materials-is-a-stopgap-for-data-archiving/

[8] List of Repositories and Data Centres. http://datacite.org/repolist

[9] Dryad Website. http://datadryad.org/

[10] Pangaea Website. http://www.pangaea.de/

[11] Dataverse Website. http://thedata.org/

[12] GigaScience Website. http://www.gigasciencejournal.com/

[13] Patterson, M. (2011). *Open-access megajournals – find out more in Estonia*. PLOS Blogs [Blog Post]. Retrieved on October 17 2012 from: http://blogs.plos.org/plos/2011/06/open-access-megajournals-%E2%80%93-find-out-more-in-estonia/

[14] Acta Crystallographica Section E Website. http://journals.iucr.org/e/

[15] Earth System Science Data Website. http://www.earth-system-science-data.net/

[16] International Journal of Epidemiology Website. http://www.oxfordjournals.org/our_journals/ije/about.html

[17] Figshare Website. http://figshare.com/

[18] Mons, B., Velterop, J. (2009). Nano-Publication in the e-science era. [Conference Paper] Netherlands Bioinformatics Centre. Retrieved 17 October 2012 from: http://www.nbic.nl/uploads/media/Nano-Publication_BarendMons-JanVelterop.pdf

[19] Beck, T., Thorisson, G., Brookes, A. (2011). Applying ontologies and exploring nanopublishing in a genome-wide association study database. In *Proceedings of the 4th International Workshop on Semantic Web Applications and Tools for the Life Sciences* (SWAT4LS '11). ACM, New York, NY, USA, 1-2. DOI:10.1145/2166896.2166897 http://doi.acm.org/10.1145/2166896.2166897

[20] – Hilse, H.-W., Kothe, J. (2006). *Implementing Persistent Identifiers: Overview of concepts, guidelines and recommendations*. London / Amsterdam: Consortium of European Libraries and European Commission on Preservation and Access, 2006. http://nbn-resolving.de/urn:nbn:de:gbv:7-isbn-90-6984-508-3-8 [Persistent Identifier]

[21] – Hakala, J. (2010). *Persistent identifiers – an overview*. [Report]. Technology Watch. Retrieved on 17 October 2012 from: http://metadaten-twr.org/2010/10/13/persistent-identifiers-an-overview/

[22] Tonkin, E. *Persistent Identifiers: Considering the Options*. (July 2008) Ariadne Issue 56. Retrieved on 22 August 2012 from: http://www.ariadne.ac.uk/issue56/tonkin/

[23] Handle Website. http://www.handle.net/

[24] DataCite Website. http://www.datacite.org/

[25] Starr, J. (2011). isCitedBy: A Metadata Scheme for DataCite. D-Lib Magazine, Jan/Feb2011, Vol. 17, Number 1/2. doi: 10.1045/january2011-starr http://dx.doi.org/10.1045/january2011-starr

[26] ARK Website. https://confluence.ucop.edu/display/Curation/ARK

[27] PURL Website. http://purl.oclc.org/docs/index.html

[28] *Minting DOIs for research data in the UK*. (2012). Kaptur [Blog Post]. Accessed on 18 October 2012 from http://kaptur.wordpress.com/2012/05/28/minting-dois-for-research-data-in-the-uk/

[29] EZID Website. http://www.cdlib.org/services/uc3/ezid/

[30] – ResearcherID Website. www.researcherid.com/

[31] Scopus Website. http://www.scopus.com/home.url

[32] ISNI Website. http://www.isni.org/

[33] Wilson, B. and Fenner, M. (2012). *Open Researcher and Contributor ID (ORCID): Solving the Name Ambiguity Problem* [online] Educause review online. Retrieved 10 September 2012, from http://www.educause.edu/ero/article/open-researcher-contributor-id-orcid-solving-name-ambiguity-problem

[34] Butler, D. (2012) *Scientists: your number is up* [online] Nature news. Retrieved 10 September 2012 from http://www.nature.com/news/scientists-your-number-is-up-1.10740

[35] ORCID and DataCite Interoperability Network Website. http://odin-wp.gthorisson.name/

[36] CLADDIER Project Page.
http://www.jisc.ac.uk/whatwedo/programmes/digitalrepositories2005/claddier

[37] Storelink Project Page.
http://www.jisc.ac.uk/whatwedo/programmes/digitalrepositories2007/storelink.aspx

[38] Webtracks Project Page.
http://www.jisc.ac.uk/whatwedo/programmes/mrd/clip/webtracks.aspx

[39] Data Citation Index Website: http://wokinfo.com//products_tools/multidisciplinary/dci/ Accessed 25 September 2012.

[40] Altmetrics Website. http://altmetrics.org/manifesto/

[41] Pisani, E., Abouzahr, C. (2010). *Sharing health data: good intentions are not enough*. Bulletin of the World Health Organization 2010;88:462-466. doi: 10.2471/BLT.09.074393 http://dx.doi.org/10.2471/BLT.09.074393

[42] Castillo, F., Clark, B., Hoyle, L., Kashyap, N., Perpich, D., Wackerow, J., Wenzig, K. (2011). *Metadata for the Longitudinal Data Lifecycle*. DDI Working Paper Series, ISSN 2153-8247. http://dx.doi.org/10.3886/DDILongitudinal03.

[43] Hrynaszkiewicz, I., Norton, M., Vickers, A., Altman, D., (2010). *Preparing raw clinical data for publication: guidance for journal editors, authors and peer reviewers.* BMJ 2010;340:c181 doi:10.1136/bmj.c181 http://dx.doi.org/10.1136/bmj.c181

[44] Mooney, H, Newton, MP. (2012). The Anatomy of a Data Citation: Discovery, Reuse, and Credit. *Journal of Librarianship and Scholarly Communication* 1(1):eP1035. http://dx.doi.org/10.7710/2162-3309.1035

[45] Green, T. (2009). We Need Publishing Standards for Datasets and Data Tables. *OECD Publishing White Papers*, OECD Publishing. doi: 10.1787/787355886123 http://dx.doi.org/10.1787/787355886123

[46] 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 http://dx.doi.org/10.1045/march2007-altman

[47] Giardine et al. (2011). *Systematic documentation and analysis of genetic variation in hemoglobinopathies using the microattribution approach*. Nature Genetics, Vol. 43, pp.295-301. doi:10.1038/ng.785. http://dx.doi.org/10.1038/ng.785

[48]Walport, M., Brest, P. (2011). *Sharing research data to improve public health.* The Lancet, Vol. 377, Issue 9765, pp537-539, http://dx.doi.org/10.1016/S0140-6736(10)62234-9

11 Further information

The question of what constitutes 'Data' is addressed in these papers:

Borgman, Christine L., *The Conundrum of Sharing Research Data* (June 21, 2011). Journal of the American Society for Information Science and Technology, pp. 1-40, 2011. Available at SSRN: http://ssrn.com/abstract=1869155 or http://dx.doi.org/10.2139/ssrn.1869155
And
Borgman, C., Wallis, J., Mayernik, M. *Who's Got the Data? Interdependencies in Science and Technology Collaborations*. (2012). Computer Supported Cooperative Work [Springer Netherlands] http://dx.doi.org/10.1007/s10606-012-9169-z

The MRC Data Support Service provides a gateway to population and patient study data: http://www.mrc.ac.uk/Ourresearch/Ethicsresearchguidance/datasharing/DSS/index.htm

The Force 11 initiative is working to integrate data into the future of scholarly communication: http://force11.org/

Fundref, a project facilitated by Crossref, provides a formal taxonomy of funders, enabling them to plug into mechanisms for tracking impact:
http://www.crossref.org/fundref/index.html

Work on the provenance of web objects is being done by the W3C Provenance Group:
http://www.w3.org/2011/prov/wiki/Main_Page

The Malaria Open Data project (MAP) being run in Oxford is a good example of open access public health data.
http://www.map.ox.ac.uk/

The Biosharing project is working in the area of standardisation, providing a focal point for different initiatives.
http://www.biosharing.org

The idea of a living paper that evolves after publication and has an open peer review system represents a change in approach. The faculty of 1000 is making significant contributions in this area.
http://f1000research.com/2012/07/16/citations-and-indexing-a-novel-approach/

The DataOne portal provides access to environmental data and represents a step forward for technology supporting open data.
http://www.dataone.org/

Mike Weiner's DataShare project at the University of California is an example of biomedical data sharing having a positive impact.
http://datashare.ucsf.edu/xtf/search

The CODATA/ICSTI/BRDI group is doing a great deal of work to develop citation standards and practices.
http://sites.nationalacademies.org/PGA/brdi/PGA_064019

Appendix One: Dryad

The Dryad data repository is a well-established initiative with an international reach that aims to preserve the underlying data supporting peer-reviewed articles in scientific journals. To this end, they maintain strong links with publishers and offer an enhanced service to partner journals, integrating manuscript and data submission. In turn, the journals encourage scientists publishing with them to deposit with Dryad and ensure that links to supporting data appear in the publication. Dryad began as a repository for Evolutionary Biology research and now offers broad coverage of biology and medicine; however, any scientific publication is legitimate and open for inclusion.

The guiding principle of the repository is to encourage data deposit by providing a low-bar route to submission. This was achieved in a number of ways, primarily by removing any restrictions on the type of data that could be deposited but also by providing a simple data collection method and a relatively low cost per package. The hope was that this low-friction approach would encourage the preservation of commonly lost, "long-tail" data.

The repository layers other services on top of its main offerings of data storage and access. When uploaded, data is scanned for viruses and old formats; for certain data types, simple curation is supported with automatic migration tools available. Data is secured and an embargo can be placed on its publication by the uploader, provided that the length of the embargo period is supported by the policy of the associated publisher. Under certain circumstances, when an embargo beyond the standard is required, it can be extended with clearance from the relevant journal.

Dryad has developed a standard set of minimum metadata that enable a suitable level of data discovery and are a requirement of deposit. There will always be niche subjects that use their own standards and, as much as is possible, they should be accommodated but there must be a framework and some kind of standardisation is necessary.

Considerations for Epidemiologists

There is recognition that the field of Epidemiology is one where there is considerable potential for engagement and support. An important step to realising that potential was taken by the DryadUK project which engaged with the online journal, BMJ Open to provide data deposit support to the medical community.

Bearing in mind that circumstances change, there are some current limitations to the repository that may be of interest to Epidemiology researchers, the main one being that of versioning. Dryad does not allow changes to be made to deposited datasets and requires that any changes should be made to a new copy of the dataset that is submitted in its own right, given its own DOI and then linked to the parent dataset. Should there be a complicated relationship between a parent and derived dataset or two versions of the same set it is expected that a readme file submitted by the author would provide clarification.

Directly linking datasets to physical objects, for example samples in a biobank, is too difficult although there are ways to work around that, for example, by including biobank IDs as part of a dataset. In cases where this would be useful, the assumption is that the information required to find the object would be contained within the journal article associated with the dataset.

One policy worth noting is that all data deposited with Dryad is subject to the CC0 creative commons waiver, which allows unrestricted use of the data. This increases the utility of the dataset by lowering the bar to reuse but, for researchers using data with privacy restrictions, applying such a licence could be inappropriate.

Developments for the future

The mechanisms to support the Dryad revenue model have been developed over the past year and will soon be deployed. Fees will be collected from journals, in the main, but could also be taken from authors or institutions supporting data deposit. As a benchmark, the figure associated with data deposit from an author publishing in a non-affiliated journal will be around £30-40; this cost covers processing and the preservation of deposited data in perpetuity.

The future success of the repository depends on a continuing dialogue between the repository and scientific publishers; there will be an increase in the number and spread of associated journals; if journal policy allows deposit, Dryad will accept the data. To facilitate this process it would be a positive development to see a greater involvement of funders with journals and researchers in order to develop standards of what data is suitable for submission.

Based on an interview with Todd Vision.

Appendix Two: Dataverse

The Dataverse Network project provides open source software enabling the creation of virtual archives for preserving, publishing, sharing and analysing research data. Archives, or 'dataverses' contain collections of individual data packets and associated metadata records and can be linked with other dataverses to form networks that share a common discipline or host facility, improving the visibility of stored material. As with other repository systems, the Dataverse Network is designed to support the archiving of largely static datasets although it does allow discovery of related versions of the same dataset.

Harvard University's Institute for Quantitative Social Science (IQSS) will host dataverses created by any researchers, anywhere in the world as a way of encouraging the sharing of academic data. Alternatively, groups or institutions can use the open source software to create and host their own dataverse, still with the option of connecting it to dataverses hosted elsewhere. Content held within a dataverse instance can be made visible to standard search engines.

The Dataverse Network project began officially in 2006 but has roots in a number of projects developed in Harvard since 1987. Broadly speaking the aim of the project is to make the research area more accessible to others by supporting the sharing of data, either completely openly, if possible, or within a limited sphere, if not. Additionally the project aimed to improve citation of data, facilitate better archival formats, improve dataset visibility and promote the link between journals and datasets.

Project reach

At present there are more than 400 dataverses holding the outputs of over 50,000 individual studies; these studies can contain multiple datasets as part of their data package. Other than Harvard there are 20 institutions that are running instances of the dataverse software and are connected with the main hub. There are new dataverses being added all the time expanding the geographical and disciplinary reach of the network, recently an astronomy dataverse was launched to bring together data from this field. In addition there are examples of institutions that have used the open source software to build their own collection of repositories independently of the main system; several universities in the Netherlands are using dataverse to link their datasets within a closed system.

Features of interest

The dataverse software adds value to the archived dataset by assigning a unique persistent identifier, using the handle system; this identifier system was chosen because of its pricing model but it is in the project roadmap to support the DataCite system. The software also provides statistical analysis and data visualisation tools; and by facilitates the migration of obsolete formats. Although all data formats can be ingested into the dataverse system only certain types receive migration support. For researchers downloading from the system there is the option to select a subset of the main dataset, particularly useful when the main holding is extremely large.

One particular feature of interest unique to the dataverse system is the provision of each dataset with a validation number referred to as a Unique Numeric Fingerprint (UNF). This can only be applied to quantitative datasets, such as plain tables, and provides a way of validating the contents to ensure that values have not been altered between uses. This enables researchers reusing a dataset linked to from a journal article to verify that the cited dataset is the same as the one that they have downloaded by comparison of associated UNFs.

There are several steps to creating a UNF but it boils down to converting the table into text format, which removes the differences between file formats and ensures that two sets of exactly the same data will return the same UNF value. When a dataverse user downloads a subset of a dataset the download includes a file containing the citation for that data along with a freshly calculated UNF.

The Dataverse Network offers individual control over data licencing conditions; individual Dataverse curators are responsible for making licencing decisions and the software supports datasets with restricted access rights and varying terms and conditions; there is no requirement to make the data open or to apply a particular licence.

Future Work

There is a project being conducted by the Harvard Data Privacy Lab in conjunction with the Dataverse Network that aims to develop tools for handling sensitive data[1]; this project is funded by the NSF for the next four years. The aim is to enable users to access and query restricted databases, returning meaningful information without compromising confidential data. This project will also look at differential privacy, testing to see how methodologies such as data masking affect the re-use of datasets, and whether it is possible to protect data in such ways without compromising its utility for others.

There is on-going, integration work with PKP's Open Journal System that prompts researchers to deposit associated data to the Dataverse Network when they submit a paper; this project is funded by the Sloan Foundation.[2] Broadening the base of associated publishers through on-going engagement is a key priority. The project will also be expanding its reach across disciplines with medical and neuroscience data being a key area for engagement. The Gates foundation is driving the use of the Dataverse Network by encouraging projects receiving their funding to deposit data with the system.

On the technical front there is work planned to make the deposit of data more frictionless using APIs, this will lower the bar to adoption for lone PIs and small groups. There is a constant expansion of the number of file formats that receive migration and visualisation support, and the Network aims to provide future support for large datasets and streamed data.

Based on an interview with Mercè Crosas.

[1] *New tools will make sharing data safer in cyberspace*. (2012). Harvard [Press release]. Accessed 08 Nov 2012 from https://www.seas.harvard.edu/news-events/press-releases/new-tools-will-make-sharing-research-data-safer-in-cyberspace

[2] *IQSS and PKP to develop data sharing system for journals.* (2012). Harvard [Blog post]. Accessed 08 Nov 2012 from http://www.iq.harvard.edu/announcements/iqss-and-pkp-develop-data-sharing-system-journals

Appendix Three: Archival Resource Keys and EZID

Archival Resource Keys

Archival Resource Keys (ARKs) were developed at the California Digital Library (CDL) in 2001 in response to a need for long-term unique identifiers in the archival space. Initially these were developed to describe physical objects but, latterly, have been applied to all types of digital object. The scheme provides an administrative framework for naming and resolving identifiers; naming is through a federated system of Name Assigning Authorities (NAA).

Project Reach

As of 2012 there were 138 registered NAAs covering a broad range of disciplines[1]; the majority are based in North America but there is also quite significant uptake in Europe, particularly in French institutions.

Items of Interest

The ARK system was designed with a focus on object metadata; in addition to returning a location for objects assigned an ARK it is also possible to query, by appending either ? or ??, available descriptive data relating to the object or a formal statement of the object's permanence, respectively.

In this way the ARK scheme differs from the DataCite DOI model; DOIs have a notion of perpetuity; they suggest that the object that they attach to has permanence. For a digital object this entails long-term storage of an unchanging object in a stable and heavy-duty repository. With ARKs it is possible to assign that permanence, if required, but it also recognises that some objects will be preserved for a limited period. It is possible to use the identifier to convey the length of time an object will be held for; when that period has expired and the object has been deleted it is possible to destroy its associated ARK and provide a smooth user experience, for example, through the use of a 'tombstone' page.

Future Developments

The ARK project has been working on a way to cope with the granular citation of datasets and have developed a method called suffix pass-through. The method employs a two-stage resolution process in which the top level of a dataset is registered with the ARK authority, the first part of the key resolves to that location. At this point the remainder of the key, the suffix, is passed to local resolution software integrated with the repository. This points to a subset of the data and enables many subsets of a single dataset to have unique identifiers whilst only requiring a single ARK to be registered. This update is currently at the QA stage of development.

[1] List of Name Assigning Authorities. Accessed 17 October 2012 from: http://www.cdlib.org/services/uc3/naan_table.html

EZID

The development of the EZID project began when the California Digital Library identified a need for software capable of handle large numbers of persistent Identifiers. This allows the easy creation and assignation of identifiers to datasets, and facilitates the storage and updating of metadata including associated URL locations. The interface allows for automated operation of changes enabling the large-scale handling of identifiers.

The scheme also allows a choice of identifiers to be offered from a single access-point. Currently this extends to ARKs, which can be assigned to datasets from any area, and DataCite DOIs, which can only be assigned to North American datasets, as per DataCite convention. These will soon be joined by Universally Unique Identifiers, which are being included due to their popularity in certain parts of Europe. There is an expectation that support will be extended to other identifier schemes in future.

The majority of the scheme's direct users are libraries or repositories that form a point of contact for researchers, the end users of the identifiers. The EZID scheme supports this federated model of engagement by providing open-access training materials, such as powerpoint presentations, postcards and webinars.

Project Reach

The EZID system became available for use in 2011 and there are now a number of major institutions in North America using the software to manage their persistent identifiers. This includes the Dryad digital repository; a partial list of users is available on the EZID website.[1]

Items of interest

The EZID software is licenced in an open-source, red-hat model allowing other institutions to run their own versions; for example, Perdue University run their own EZID clients. Technical support for instances of the EZID system run outside of the CDL is provided by the project's own technical team; users have access to two permanent members of staff.

EZID will reserve identifiers by holding them without registering them with the agency. This is useful if you would like to cite a supporting dataset in a paper before that dataset is public; there is no need to add metadata until the object is published.

Future Developments

EZID is looking to develop fine-grained metadata control as part of future developments. Researchers will be able to set a 'do not index' flag for metadata; this could be used for datasets under embargo or for those whose owners would like their website to be the destination for indexed searches.

Based on an interview with Joan Starr

[1] EZID partial user list. Accessed 18 October 2012 from: http://n2t.net/ezid/home/community

Appendix Four: DataCite, Digital Object Identifiers (DOIs)

DataCite is a well-established organisation that aims to promote the stability and visibility of research data on the internet whilst also increasing its acceptance as a legitimate contribution to the scholarly record.

DataCite is a member of the International DOI Foundation, a federated organisation committed to the provision of a social and technical infrastructure for delivering persistent, digital identification of both physical and digital objects. The foundation has several members supporting use of DOIs in particular areas; DataCite has responsibility for implementing identifiers for datasets. DOIs already have considerable traction in the academic sphere where they have been used for some time to permanently and uniquely identify journal articles.

Technically, Digital Object Identifiers (DOIs) make use of the handle system for providing and resolving unique identifiers but considerably enrich it by layering extra services on top of the existing infrastructure. The quality of this enrichment was recognised earlier this year when DOIs became the first persistent identification scheme to be awarded ISO standard status. There are cost implications associated with the maintenance of this improved service and registering to mint DOIs for datasets is more expensive than the handle system, specific costs vary depending on the DataCite member that an organisation registers with (for UK datasets, the British Library is the responsible DataCite member.)

Through a process of engagement with datacentres, publishers and researchers the organisation has helped to develop standards and strategies for permanently connecting the outputs of scholarly research with the underlying data in a discoverable, linked fashion.

Recent Developments

Over the past year, the DataCite organisation has increased its membership footprint and scope, the total number of datasets with associated DOIs has increased substantially, as has the number of registered datasets with the recommended level of metadata.

DataCite operates a federated model of membership as datacentres have an international consumption but are funded either nationally or by subject. Typically, there is one DataCite member per country although in some areas, such as Germany, where members are not discipline agnostic, several exist covering different domains. This membership is expanding both geographically and by discipline; two notable additions in the previous year are the Conference of Italian Rectors (CRUI), providing coverage of the Arts and Humanities, and the Canada Institute for Science and Technology (CISTI), Canada's first DataCite member. This expansion will continue in the future with attention being paid to filling domain gaps in countries that have an existing DataCite member.

There have been important developments in the technical infrastructure and shared services over the past year. Working with CrossRef, DataCite has developed a new content resolver that broadens the way in which links and metadata can be processed, paving the way for increased access by automated consumers. The resolver enables new applications such as the automated citation formatter[1] which constructs citations automatically using metadata associated with a given DOI plugged in to particular journals' style guides.
The philosophy of this initiative is to achieve faster and greater interoperability by leveraging what already exists in the web rather than inventing specialised protocols.

[1] CrossCite automated citation formatter website. http://crosscite.org/citeproc/

Future Directions

DataCite has just started the ODIN project in collaboration with ORCID, the British Library and the STFC looking, amongst other things, at how you deal with multiple contributors on a single project and how you deal with multiple papers that link to a single dataset. The project will be working with datasets generated in high energy physics and longitudinal cohort studies, both seen as typifying those issues. The project will include data from CERN being made public, as it is data with a high, public profile that should generate considerable interest in the project. [See also: Appendix Seven: ODIN project.]

DataCite will also be engaging in more community-building activities, particularly in an effort to support the work done by datacentres and data stewards, CISNI has made some great statements in this area. In June DataCite released a joint statement with the STM publishers calling for bi-directional linkage between journal articles and the data that supports them. In future, there will be more joint statements made in an effort to drive community practice; the key goals at the moment are to get datasets cited and to ensure that they are cited correctly.

Based on an interview with Adam Farquhar.

Appendix Five: International Standard Name Identifiers

The International Standard Name Identifier (ISNI) scheme is an ISO standard methodology for assigning persistent, unambiguous identifiers to individuals and entities associated with various areas of the media. At present the system does not include academic publishers but it has been organised to be interoperable with ORCID, leaving the way clear for future linking of the two systems.

The ISNI system addresses disambiguation issues by assigning a single, unique, persistent identifier to an entity and enabling them to link their various public identities through it. This identifier can be applied to individuals, groups of individuals, companies or organisations and could be any way by which an entity is identified in the public sphere; this includes things such as band names, author pseudonyms, record labels etc. The system supports complex ways of relating names together, for example institutional identifiers can use nesting and linking metadata to connect companies through departments to people. At present, this level of metadata organisation is considerably more sophisticated that that used by journals.

One major advantage that this system affords is in the payment of royalties, which can be a complex, time-consuming process. The approach taken by ISNI is that, if you can assign an identifier and keep on using it, your systems will become progressively simpler to manage.

ISNI operates as a bridge identifier which spans systems, providing opportunities for interoperability between traditionally discrete databases. At present the system has traction across a range of media, there are currently eight industries taking part, including television, music, film and publishing. Several hundred thousand records have already been ingested into the ISNI system from the registries of participating agencies. The central registry keeps the bare minimum amount of metadata required for disambiguation purposes (at present that amounts to nine fields) with its network of registration authorities holding their own, more detailed metadata locally.

The ISNI system is using automatic ingest of large quantities of data to pre-populate its database; by accessing the OCLC's Virtual International Authority File, an amalgamation of data from 27 authorities, ISNI can seed the system with 8 million names. At present the project is working on ironing out the bugs in identity matching algorithms to enable this level of large-scale, automated ingest.

Future Developments

The next 24 months are going to be concerned with implementation and rollout of the system; it is expected that new industries will be amalgamated into the system during this period. Once the identifiers are in place, the infrastructure will allow development of new content management systems. This is not seen as a short term goal, it will take in the region of two years to get the content management systems in place and then several years after that for second generation systems to be designed and implemented.

Based on an interview with Todd Carpenter

Appendix Six: Open Researcher and Contributor ID

The Open Researcher and Contributor ID (ORCID) system has been developed to address the issue of ambiguity over the identification of academic authors. For researchers with common names it can be almost impossible to identify their complete scholarly record, indeed it has been estimated that disambiguation algorithms armed with reasonable amounts of metadata are still unable to resolve identity correctly 5-10% of the time.[1] The ORCID project aims to provide a database of scholarly records that can be utilised by individual researchers to avoid duplication of effort by automatically populating CVs and online forms whilst at the same time constructing a robust network of links between researchers and their scholarly outputs that can be leveraged by metrics systems and institutions that wish to keep track of their scholarly output.

There have been previous attempts to address this issue but none have gained the multi-disciplinary, global support required to be recognised as a standard. ORCID will release their underlying code under the MIT Open Source Licence Framework making it available for alteration and re-use. Public data in the repository will be regularly deposited with partners. The hope is that this open, international approach coupled with the guarantee of permanence delivered by its backers will establish the project as a global standard. There are over 280 participants in the ORCID initiative, mainly academic institutions but also including publishers, scholarly societies and various other organisations.

The project went live towards the end of 2012 and enables researchers to build an on-line profile of their scholarly output, linked to a unique identifier. The system places control of the profile in the hands of the researcher allowing them to publish as much or as little of their information as they like.

Population of the profile lies with the researcher; the scheme supports bi-directional linkage with various author profile and manuscript submission systems enabling the user to pull down a list of publications that are likely to be associated with them. The researcher then manually selects the publications to add to the profile; problems with disambiguation mean that pre-populating profiles automatically cannot be completed accurately enough to satisfy ORCID. Additionally, this avoids creating vast quantities of redundant profiles for researchers who have no use for an ORCID profile, undergraduates, postgraduates and co-writers who have one publication to their name and no intention of producing more. The ORCID profile links and synchronises with existing systems such as Scopus, ResearcherID and RePec. Many organisations have announced applications that will integrate with the system and will be participating in the ORCID launch partners program.

Initially, the system only focuses on articles in scholarly journals however, ORCID's aspiration is to contribute to the collection of claims about all relevant scholarly contributions ranging from grants, patents and books to research datasets, indeed it is almost certain that data DOIs will be integrated into the system next year. The project aims to offer as broad a range of scholarly outputs as possible, reasoning that if an academic asserts that an object, such as a blog post, has scholarly value, then it does; the community, particularly funders, will determine which outputs hold value.

In terms of costs, the intention is for individual researchers to be able to acquire an ORCID profile and associated unique ID at no charge whereas organisations, such as universities, will pay a tiered subscription charge.

Based on an interview with Mike Taylor.

[1] Wilson, B. and Fenner, M. (2012). *Open Researcher and Contributor ID (ORCID): Solving the Name Ambiguity Problem* [online] Educause review online. Retrieved 10 September 2012, from http://www.educause.edu/ero/article/open-researcher-contributor-id-orcid-solving-name-ambiguity-problem

Appendix Seven: ORCID DataCite Interoperability Network

The ORCID DataCite Interoperability Network (ODIN) project is a two-year collaboration, financed by the EU that seeks to address some of the major open questions in the area of data publishing and referencing by developing standards and protocols for citation, attribution and tracking. There are seven partner organisations involved, the British Library, CERN, DataCite, ORCID, Dryad, arXiv and the Australian National Data Service.

The ODIN project will leverage persistent identification infrastructure provided by the ORCID and DataCite projects to produce mechanisms for linking researcher and dataset IDs across multiple services whilst also defining the nature of these relationships and developing standard terms of reference. A key aim of the project is to enable the complex tracking of datasets through multiple instances of re-use potentially demonstrating impact in areas hitherto hard to qualify, such as influence on government policy.

ODIN will also be working with pilot groups in two research areas associated with complex outputs to produce proofs of concept that address key barriers to the uptake of data citation. High energy physics groups at CERN provide an excellent test-bed for models of complex attribution, commonly several hundred researchers will contribute to the production of a single paper. Additionally, the British Library will focus on social sciences through the British Birth Cohort Studies to look at how citations and relationships can be built up over a longer time frame and to track the impact of data collected between 1946 and 2010. The method for engaging with the research community has not yet been finalised, it is likely that the project members will aim to use existing data to build a model and then take it to researchers to test.

The Birth Cohort Studies proof of concept will address some of the key questions of interest to public health research including citing datasets that are confidential or under embargo and investigating fine-grained control of data and metadata exposure, allowing selected subsets to be published. At present only data centres and data creators have been approached to take part in the study, the next stage of engagement could take in institutional repositories.

On the development side DataCite and ORCID will ensure that the processes will mesh well between themselves and other identifier systems and the non-European partners provide a strategic input and advice on international scalability.

Based on an interview with John Kaye

Appendix Eight: Webtracks

Webtracks was a JISC-funded project that ran from August 2010 to November 2011 and developed new protocols for linking data objects; this work lead on from two precursor projects, CLADDIER and Storelink, also funded by JISC.

The initial project, CLADDIER, examined various models of data publishing and citation, investigating different ways in which data could be associated with a journal paper. The project worked with a modified version of the TrackBack protocol, used to connect blog posts, and implemented it as a citation notification service for repositories. As a practical proof of concept this protocol was implemented in the STFC's ePubs repository and the BADC repository. This work was further developed by the Storelink project which implemented the protocol in the EPrints repository architectures, working with eCrystals at Southampton University.

Webtracks looked at ways of keeping track of large quantities of raw data produced as part of collaborative projects, building on the work of the two previous projects to develop methods of propagating linked data rather than just citations.

Pilot studies centred on large scale scientific facilities, such as the Rutherford Appleton Lab's ISIS neutron source, and worked with the Smart Research Framework's Labtrove software to develop a peer-to-peer protocol for linking raw research data. Electronic notebooks are used as focal points for the aggregation of these links and constitute a publishable, citable resource enabling data to be published almost as soon as they are captured.

At present the protocol is designed for internal use on specific projects however, if there was sufficient interest, it could receive an open source release allowing other groups to continue development.

Based on an interview with Brian Matthews

Appendix Nine: Focus Group Notes (inc. Basia Zaba Interview)

Summary of Data Citation Focus Group Meeting, Wellcome Trust, London, 18 September 2012

Chair: Jonathan Rans (DCC); Present: David Carr (Wellcome Trust), Geraldine Clement-Stoneham (MRC), Philip Curran (MRC Unit for Lifelong Health and Ageing), Michael Day (DCC), Judith Glynn (London School of Hygiene and Tropical Medicine), Debbie Hart (KCL Department of Twins Research), Catherine Jones (STFC), Sarah Jones (DCC), Christine McMahon (UCL Institute of Child Health), Victoria Vazquez (KCL Department of Twins Research)

Introductions

David Carr (Wellcome Trust) provided a brief introduction to the Public Health Research Data Forum, a consortium of health funding bodies that have signed up to a joint statement on the sharing of data in the public health domain.[1] The Data Forum is working to implement the statement by exploring three main areas: 1) capacity and skills; 2) culture and incentives (including data citation); 3) infrastructure and tools (including standards, metadata, the discoverability of data, etc.). In developing an implementation work plan, there will be a need to build incentives for data sharing - e.g. providing ways of tracking the downstream use of assets. The DCC have been commissioned to look at the existing landscape, so that funding bodies can have an evidence based focus on knowledge (thus this focus group).

Sarah Jones (DCC) then provided brief introduction to the Digital Curation Centre before all other participants were given the opportunity to introduce themselves and their involvement in data sharing.

- *Philip Curran* (MRC Unit for Lifelong Health and Ageing) - looks after the LSHD 1946 cohort. This has lots of external collaborators (300+); data sharing is a large part of their work. The quality of the data varies over time. Often, lots of work is needed to bring much older data up to modern standards. He has also contributed to the MRC Data Support Service (DSS), providing a dataset that can be accessed via their gateway. The MRC unit develops its own tools for data delivery of multiple variables.
- *Geraldine Clement-Stoneham* (MRC) - open access publishing lead at MRC. They are collaborating with other research councils to harmonise data policies. MRC requirements state that researchers are expected to cite the underlying data in publications so she was interested in learning about methods to do this.
- *Debbie Hart* and *Victoria Vazquez* (KCL Department of Twins Research) - work on the Twins UK bio-resource project. The project has been tracking 12,000 twins, and is now entering its 21st year. They have lots of data, including biological data. They support open access via a governance committee; any access difficulties are addressed through this committee for independent governance.
- *Judith Glynn* (London School of Hygiene and Tropical Medicine) - working on longitudinal studies covering leprosy, TB, HIV and other diseases in Malawi. It has been Wellcome Trust funded for the last 15 years. The records are complex, e.g. on-going and totally linked. Like the Twins UK study, they also link to biological, genetic data.
- *Christine McMahon* (UCL Institute of Child Health) - PhD student, interested in metadata citation
- *Catherine Jones* (STFC) - has worked on projects about linking data e.g. CLADDIER. This investigated granularity, i.e. how you could pick the bit of data you want to cite out of a

[1] Walport, M., Brest, P. (2011). *Sharing research data to improve public health.* The Lancet, Vol. 377, Issue 9765, pp537-539, http://dx.doi.org/10.1016/S0140-6736(10)62234-9

- big atmospheric data study. She was also project manager for the 2nd phase of MRC DSS. They produced a gateway to data from 5 cohort studies, one of which was deposited by Dr Curran.
- *Basia Zaba* (London School of Hygiene and Tropical Medicine) – works on the ALPHA network project that links 10 longitudinal studies gathering demographic and HIV data. Subsets of the data are pooled and analysed. [views gathered by interview after the main focus group session and incorporated into notes]

Session 1: Data publication methods

Are the various public health studies publishing/sharing data?

Yes, and more than is normally recognised. It was felt that reports promoting data sharing present things much more negatively than is the case in real life. Data sharing happens across the board.

All the studies have some kind of governance structure in place. Sharing is typically mediated through them. Data sharing takes place by direct negotiation with collaborators, not by open publication or hosting data online.

Typically only subsets of data are made available as it would not be realistic to make the whole dataset open.

Twins UK give out data to anyone who requests it. You don't have to meet bona-fide researcher checks. They also provide a lot of support to data requesters. It is a similar model to that used by the University of Bristol's ALSPAC (Avon Longitudinal Study of Parents and Children) study.

The ALPHA network was given as an example of data being shared across multiple studies via data sharing agreements. It's not open, but is a way to share data with a level of consistency.

The increased publication of questionnaires was suggested, as there would be no major ethical issues involved.

The Alpha Network links 10 studies containing demographic and HIV data. Analysis is performed on data pooled from the participating studies. Currently this is not public but the project would like to put it in a public repository. The expectation is that most access will be from other academics.

At present the technical barriers to sharing are as big as any other.

How is access decided / granted?

Access is typically managed by the study itself. Standalone publishing is the most prevalent model.

MRC-funded studies tend not to deposit in a central archive, partially due to real privacy risks. The cohort is the most valuable asset studies have, so they need to be extremely sensitive to their attitudes. Participants are typically happy for their data to be shared with other researchers in epidemiology but less so with pharmaceutical companies.

It was noted that there were big differences of data sharing culture between the ESRC and MRC, so jointly-funded projects would typically deposit metadata rather than the actual data in the UK Data Archive – see section on *sharing metadata rather than data* (2.8).

Access is typically granted based on the requests that come in from researchers. These would need to be framed in terms of a research question or hypothesis they hope to answer. Twins UK have recently had a couple of requests that have asked for so much data that they could potentially replicate the entire phenotype. These seem to be just phishing-type exercises, trawling everything. They are managing this by asking for very hypothesis driven requests.

Some studies restrict access to 'bona-fide' researchers e.g. those with an e-mail address with an ac.uk domain name or registered charities, etc. Others may check CVs, etc.

The MRC unit retain control of all data when working with collaborators on a study. Requests for data sharing sent to the grant's PI would need to be sent on to the MRC unit for approval. The MRC unit typically gets an SPSS/Stata file of derived variables back from researchers who have used data, which is put back in the pool. New derived data items become part of the datasets that were shared with documentation about how these were created. The MRC unit have a large derived data library. One statistician is responsible for keeping this organised.

One would expect to have some control over who uses the data; access is through a data management committee. The approach could mirror that of the DSS.

The ALPHA project doesn't keep copies of subsets of data after analysis has been run but they do keep the files which describe the subsets and the analysis.

Data access restrictions are a way to control the quality of science

The restrictions on access are not really controls on sharing but a way to control the quality of science that is output. Data providers want to make sure that people fully understand the data they've requested and that any products of data sharing pass the peer-review process. There is a potential reputational risk if this is not done.

It was felt that data providers have a requirement to facilitate access but those requesting data also have a responsibility to learn about it so they can handle, manipulate and analyse it appropriately.

In the end, data sharing is about risk management. In epidemiology, there is debate over the validity of data anonymisation, so data providers are applying the only other controls that they have, i.e.:

1. to judge the validity of those that request data, e.g. a researcher or journalist;
2. to limit the amount of data people can have access to.

For some, anonymisation does work, public use datasets are never going to be at the same level of detail as original datasets but that's OK. The DSS only provide dates to the nearest month, for example. How closely do you describe the geography? You need to amalgamate to a certain extent.

The MRC Unit and Twins UK request to see draft papers before publication, to ensure that any use of the data (e.g. in tables) is appropriate. Also, one researcher had asked Twins-UK for various similar questionnaires to be run but had not published them, so to protect the cohort, the governance committee pushed for publication of results before any further studies were approved.

Challenges to sharing – context / documentation

It would be very difficult to simply publish all data because it would be impossible to use without an extensive understanding of its nature and context, e.g. information on how the data was collected, how project aims may have changed over time.

Context is important, but can be very hard to capture, particularly given the complexity of large field epidemiology studies, as these are not one thing. Studies evolve over time, with data being collected in different studies for different purposes. It is, therefore, difficult to publish data in a form that would not potentially be misleading.

You don't always know what you know to use the data – there's so much implicit knowledge. It is very difficult to document that, even for insiders. Having documentation that's understandable to outsiders who are coming in without knowing the background and context would be something completely new (projects are not typically interested in developing a training area for secondary use). You would certainly need additional time and funding to provide this. Joint grant funding might be one way to do this.

The documentation has to trace the data from collection through to final analysis, although simply providing that information isn't enough; input is required to aid interpretation.

Challenges to sharing – resourcing

50 phenotypes of data were cleaned by a team of 10 in the Twins UK project. It was a real slog to get this done in 2 years and it's only 50 phenotypes out of thousands. They've had so many enquiries since that the statisticians are getting lots of extra work they're not paid to do. When people know there's new data available there is an increased demand. It is time consuming to deal with enquiries, so there is a need for added support and training.

The issue with publishing anything is that there are more requests and no additional time or money to deal with these. The MRC view is that such requests should be submitted as grant applications, assuming that there is a viable research question. The MRC makes a deliberate choice not to fund data management and sharing for its own sake. Nobody has the resource to invest in cleaning and managing all data without knowing whether it would be used or not. Fundamentally, there has to be a research question that will be answered.

The MRC unit gets involved with potential collaborators at the proposal stage so they can help advise on what the right set of variables are and to make sure they're not doing something that has been covered by earlier studies.

Challenges to sharing – consent

Consent ranges from verbal to written consent depending on the nature of the study and participants. It also varies as studies evolve, throwing up further confusion. Interestingly, consent for sharing in the LSHTM Malawi studies is granted by the Malawian government rather than participants as they have a sense of ownership of that data. LSHTM are currently discussing developing a biomedical resource that would be owned and managed by Malawi, i.e. the curators / gatekeepers would physically be based there. Consent is expected to get harder to obtain, as people become much more aware of the value of personal health data.

The Twins UK study seeks new consents every time they ask new questions. They looked at global consent but didn't think it would work for them. They are going to do a questionnaire to see what participants understand by consent, withdrawal, etc. as they anticipate most people just tick the box to confirm.

Whenever there is a large education, wealth or power gap between researchers and subjects there is a difficulty with consent, people are overawed.

The Demographic Surveillance System doesn't ask for consent as the data comes through from a different project; information is described by proxy so consent is implied.

Challenges to sharing – embargoes

What's permissible in terms of embargo periods is still very hazy – it's just termed as what's 'reasonable' for the discipline. When a request comes in on something active the Twins-UK team tend to put them in touch with the lead investigator. Embargoes can be problematic as the time to clean and analyse can vary so much for different types of data. The MRC Unit try to clarify this at the proposal stage.

Sharing metadata rather than data

People aren't always aware of the studies that are taking place. The MRC supports the discovery of information about epidemiological studies through the MRC Data Support Service (DSS) and of project information through their website and RCUK's Gateway to Research (GtR). There is a perceived need to share descriptive metadata to raise awareness and to help enable new research. Individual studies can define how much metadata they would happy to release. The level of constraints depends on who you're making information available to, studies typically being more open to the research community than to the general public.

Data sharing would help research funding bodies like the MRC demonstrate that there are real outcomes from the projects that they fund, and helps justify their getting additional funding to for subsequent programmes. There is a growing expectation that all publicly-funded research outputs (including data) should be made available online, but this is not realistic as the time and effort needed to do this would in many cases outweigh the benefits.

There was a feeling that the focus should be on publishing metadata where possible rather than publishing datasets. Data would be accessible on request, as at present, but there would be no need for all datasets to be published publicly for anyone to access anytime.

This discussion fed into a short debate about why we share data. Is it to encourage new kinds of research using the existing studies? If it's just to satisfy the taxpayer in terms of knowing what has been created, then it would be sufficient to make just basic descriptive metadata available.

Data citation

Current state of play

None of those attending were currently using persistent identifiers for data. However, there was a desire to do so in future to demonstrate use for reporting back to funding bodies. The pressure is the impact agenda and use of identifiers is not particularly driven by the science.

Attendees do not provide guidelines on how to cite the specific subset of the data provided. References can instead be made to papers that describe the study.

At present the most studies request some kind of acknowledgement in published papers. Acknowledgements have two purposes: 1) explaining where the data came from; 2) are a courtesy to the creator.

Acknowledgements can be published in a standardised form but differences in paper formatting means that it would be difficult to harvest this automatically from all publications. Most data holders will therefore need to keep their own records of usage. Twins UK is setting up an audit system at present but this still requires somebody to undertake the role.

The studies typically ask collaborators to provide papers that are based on data sharing. This could be enforced by asking for the publication information (if this has not already been provided) before access can be granted to the data a second time - or in the case of STFC's ISIS, before granting repeated beam time. Most people do send the details of publications but some do forget. If they don't do this, would they remember to include a DOI? There's a growing view that publishers shouldn't accept the publication if it doesn't link to the underlying data.

A question was asked about whether participants produce data papers? Overview papers describe the cohorts and are updated periodically, but they don't have data publications as such. These overview papers are published mainly in the *International Journal of Epidemiology*.

Pros of using DOIs

Advantages could be seen in the potential of linking a DOI with their version control system, helping to avoid a lot of checking that is currently undertaken manually. At present they need to backtrack to figure out which exact version was used, which can be quite laborious.

Assigning DOIs to fixed subsets is the practical thing to do. These DOIs ideally take you to a landing page with as much relevant information as deemed useful by the individual / area. Metadata for landing page could potentially come from grant proposals. The only question would be getting permission to publish these abstracts. MRC already do it for their grant proposals, so permission could potentially be included in agreements.

Providing a suggested citation style may make it easier for metadata about data to be copied or exported, potentially lowering the barriers for researchers to reuse.

STFC has just signed up to be an issuer of DOIs. As the facility that runs ISIS, they can keep a copy of all the data produced from the beam so they will mint their own DOIs via iCAT, their data catalogue. There is an embargo for 3 years to the originating PI.

Cons of using DOIs

There is an open question about what level of granularity to cite. Each experiment has multiple runs, each of which has lots of datasets. It is unclear where DOIs fit within this process?

Concerns were also raised about ensuring DOIs resolve persistently. As a data owner you're asserting that you are able to get back to a specific version. Some felt this was impossible as data are constantly evolving, meaning that you would need to regularly issue new versions and mint DOIs for them. Some felt you wouldn't want to keep an incorrect version of the data - e.g. one with an incorrect variable - just because someone has used it and may want to get back to that version. However, others would keep a version of the subset archived separately and would not update it to reflect changes in the main dataset. The dataset (a partial snapshot of the full database) would be created with its DOI and left 'as is' as a reference version. The Twins UK study do similar - every data request has its own place in the database and the dataset that's sent out is stored uniquely there.

Versioning is important; even if the dataset that was used can't be recovered there is a need to know why the results of the same analysis have varied.

The DDI3 initiative has thought about versioning tools; if the owners of datasets had some standardised tools that could handle versioning that would be very helpful.

The role of journals in providing guidelines

There was no emerging consistency on how journals allowed citations to be made, i.e. whether DOIs for associated datasets were included, etc. There was a need to push journals to reach a consistent view on where to put data citations (or acknowledgments) in a paper to ensure they can be harvested. It could be included in the text as a part of the methods section or included in the references, which may be easier to share given that some publishers are more willing to make these available. It was suggested that it might be better to keep separate infrastructures for citing publications and data, the latter capable of identifying particularly well-used and productive datasets.

What should be cited / acknowledged?

The overwhelming view was to reference the study or project not individuals. It would be far too messy to decipher who had exactly what role in the data collection, analysis, curation, etc. and you could not do it retrospectively. Also, where do you draw the line? Too many people have a role, so data can only really be cited at a higher (project or institute) level. There was no clear view on how microattribution would work in this context.

But you also need incentives for sharing, so perhaps those who make it available for others or maintain it should be the ones who are cited? Data curators find it hard to move on in an academic environment as they have no publications / rewards.

Some journals have begun to use author lists that break down the individual roles of each contributor. Such a long list, however, could potentially devalue those who've made the real effort. Maybe there was a need for a separate structure to acknowledge curation effort? It was unclear whether this could be done sufficiently via citations.

There should be recognition for the people who don't usually receive it. You could have attribution for senior researchers, project managers and data managers; others should be discoverable.

Summary and conclusions

From the data publication session, the main issues identified were:

- The complexity of public health data, e.g. the need for collaborators to gain a deep understanding of the data being used, the loss of implicit knowledge
- Privacy - noting complexities around consent and ownership
- Resources - sharing data takes time and effort; the costs of publishing all data would far exceed any benefits gained thereby
- A focus on good quality science - collaborators can make the case to work with specific datasets (and data holders) by applying for research grants. These would be science-led (and peer reviewed) and retain some level of control over data use
- Risk management - avoiding reputational damage

There was a desire for this exercise to emphasise the degree of sharing that takes place rather than focusing on the restrictions. Sharing happens across the board, even if datasets are not published or made openly available. The restrictions are primarily not to control sharing but to preserve the quality of the science that is output.

On data citation, public health requirements are more about keeping track of the derivations that have come out and the impact of this than about individual attribution.

Developing standards and promoting good practice can be done but requires money. Hold technical meetings to address the key issues involve documentalists, researchers and data managers.

It was asked what would be needed to incentive data sharing. The replies suggested:

- Money - as more highly-skilled statisticians and data managers would be needed
- Long-term commitments to help maintain longitudinal studies
- When new research is funded money could be retained until people’s data is available

There is a real cultural difference between epidemiological research and other research domains. For example, unlike ESRC, MRC primarily funds scientific research (rather than data management) and is quite ruthless in terms of looking at outputs and impact. In this environment, researchers are not rewarded for being altruistic and working a dataset up for sharing more widely. The data sharing that flows from this situation will always tend to be ad hoc and mediated by the data owners.