The International Journal of Digital Curation
Issue 2, Volume 6 | 2011

Citation and Peer Review of Data: Moving Towards Formal Data Publication

Bryan Lawrence,
NCAS British Atmospheric Data Centre,
STFC Rutherford Appleton Laboratory

Catherine Jones & Brian Matthews,
e-Science Centre,
STFC Rutherford Appleton Laboratory

Sam Pepler & Sarah Callaghan,
NCAS British Atmospheric Data Centre,
STFC Rutherford Appleton Laboratory

Abstract

This paper discusses many of the issues associated with formally publishing data in academia, focusing primarily on the structures that need to be put in place for peer review and formal citation of datasets. Data publication is becoming increasingly important to the scientific community, as it will provide a mechanism for those who create data to receive academic credit for their work and will allow the conclusions arising from an analysis to be more readily verifiable, thus promoting transparency in the scientific process. Peer review of data will also provide a mechanism for ensuring the quality of datasets, and we provide suggestions on the types of activities one expects to see in the peer review of data. A simple taxonomy of data publication methodologies is presented and evaluated, and the paper concludes with a discussion of dataset granularity, transience and semantics, along with a recommended human-readable citation syntax.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. ISSN: 1746-8256 The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre.

Introduction

Why Publish Data?

The foundation of the scientific method is the collection and analysis of data, but traditionally the quality of scientific work is judged on the conclusions drawn from the data as presented in the peer-reviewed literature, rather than the quality of the data itself. The fitness for purpose of the data, any internal analysis, and the data's availability or otherwise for re-interpretation are not assessed. This results in communities who do not always rate highly the effort required to produce data, and who cannot always reproduce their analysis because the data is not well enough described or preserved. For these communities, the basic thesis of the reproducibility of the experiment is only true where an experiment can begin under controlled conditions, with no dependencies on external data.

The situation is complicated by the fact that the definition of data has become more blurred in recent years, following the advent of remote sensing and the combination of measurements and data resulting from computer models of what is being measured. Dependencies on external data extend not only to observations that might have been collected about the physical properties of some specimen(s), but also to observations collected via remote sensing (which require details of the remote sensing algorithm), or which have been simulated (which require details of the simulation algorithms, initial conditions etc).

In general, scientific communities do understand the importance of the observation provenance, even though not all communities put in place methodologies to ensure that the provenance is recorded and published in enough detail for an analysis or experiment to be repeated. Some exceptions do exist; for example there are papers where aspects of algorithms and experimental method are discussed, and communities strive to record provenance. There is a general culture that where a dataset is produced for external consumption that there will be a "paper of record" describing the dataset; though these papers almost never describe data in enough detail for the simulated dataset to be recreated or for a re-retrieval from raw data to performed using the same methods. Often this is because the details of such activities are normally deemed to be too voluminous, too derivative and/or too technical for publishing in a journal. Nonetheless, it is acknowledged that data, alongside journal publications, are first class research outputs, both from social and funding policy perspectives (Arzberger et al., 2004; Klump et al., 2006; Costello, 2009) and from internal discipline requirements (Carr et al., 1997; Hancock et al., 2002; Brown, 2003).

Best practice is to publish data and make it available for re-use both within the original disciplines and the wider community. This is usually done by publishing actual digital data on the Internet, rather than as figures and tables presented within documents (Schriger et al., 2006). There is an element of technological push to this behaviour as both authors and readers are empowered to produce and consume digital data (Roosendaal & Geurts, 1997). However, the changing nature of scholarly research itself is the primary driver, rather than the technological possibilities provided by the Internet (Van de Sompel, et al., 2004).

Scope of this Paper

In this paper we present some of the issues to be addressed in making data publication on the Internet[1] a full, peer-to-paper publication, with similar standards of respect for output and quality. We begin by defining what we mean by publication and proceed to a discussion of the procedures necessary to validate the scientific quality of published data through a process of peer review. We describe some of the ways that data publication can be organised, and then conclude by presenting a notation which could be used to identify citations to published data. Although we have motivated the discussion of data publication from the perspective of the wider scientific community, when we get to details, we concentrate on the issues for data publication in the environmental sciences.

The discussion in this paper is presented in the context of work carried out by the Research Information Network (2008) for their report "To share or not to share: Publication and quality assurance of research data outputs". In that report the authors present the results of a series of exhaustive interviews carried out with over one hundred interviewees. Their conclusions agree with many of the motivations discussed in this paper, and provide extra support for the proposal that data publication can and should be happening.

Data Publishing

What Does Data Publishing Mean?

We are familiar with the definitions and processes of publishing as they apply to physical media, such as books and journal articles; so much so that publication has become synonymous with processes of ensuring the quality and longevity of the published item. However, as far as the Internet is concerned, one can publish anything by simply making it available for download, and it is commonly known that resources published in this way are less reliable. From a scientific data point of view, we are very interested in defining a process of data publication that will carry with it the ability to make assertions about the trustworthiness and fitness for specific purposes of the data which are understood by the data consumers.

Data publication poses the question: "How do I publish a thing regardless of what it is?" In the context of "internet publication" of things such as news articles or videos, customary usage[2] expects that when a URL is de-referenced, the target can be, or is, directly rendered and therefore can be directly consumed by humans. By contrast, when a URL for a dataset is de-referenced the expectation is that the data will be supplied in a computer readable format with human interpretation made possible through other software. It is for this reason we propose the introduction of a data publication to bridge the gap between the human and the computer. The definition of data and its representation give rise to issues, which are discussed further in the section on definition and citation.

[1] We do not specifically consider the publication of data onto physical media. We consider that to be a subset of the wider data publication problem, sharing some of the same wider issues as regards peer review and Internet publication, but for which more traditional publication methodologies are amenable.
[2] Customary usage; i.e. web browsing. We explicitly exclude Web Services from customary usage.

In this paper we define to Publish (with a capital P) data, as: "To make data as permanently available as possible on the Internet." This Published data has been through a process which means it can appear along with easily digestible information as to its trustworthiness, reliability, format and content.

Data Permanence and Publication

The expectation of publishing on paper is a level of permanence not generally achieved by publishing on the Internet. Permanence issues for electronic material revolve around three problems:

- How do I find the material again (the "identifier" problem)?
- Will the material identified have been moved, changed or removed?
- Will I still have software capable of interpreting the object when I get back to it?

The first question is particularly relevant given the fragility of most URLs. A solution that is gaining momentum in scientific circles is the use of Digital Object Identifiers (DOIs) to provide a permanent identifier and locator for datasets. Work in this area is currently ongoing and brings together interested parties on a global scale. The DOIs provided for data sets are registered through DataCite[3], which is an international organisation with the goal of establishing a not-for-profit agency that enables organisations to register research datasets and assign persistent identifiers to them, so that research datasets can be handled as independent, citable, unique scientific objects.

The second of these questions is primarily an archiving problem, and as such falls out of the scope of this paper. However, an essential part of citation and publication of data is providing the assurance that the dataset that was cited yesterday may be located and retrieved in the same condition, or as near as possible, in perpetuity. Although the details of archive provision are beyond this paper, we take it as a Publication requirement that data to be Published are held by an organization which can provide appropriate long term curation. As well as addressing the long term persistence directly, such curation would be expected to address the third question by migrating the data interfaces to support the software and software interfaces in current use at any given time. For some datasets this might well require that the archive hosting the data have discipline specific experts, for others, this might well be achievable without such expertise, but either way, an active process of curation is a necessary component of the Publication activity.

Data Publication Procedures

Data Publication should consist of a procedure which allows the community to make assertions about the trustworthiness and fitness for purpose of the data. To do this we are defining a process directly analogous to that which occurs in the existing scholarly communication process. Schaffner (1994) points out that the role of journals is to provide the qualitative differences between formal (journal publication) and informal ("putting the data up on the web") communications.

[3] DataCite: http://www.datacite.org/. Retrieved January 22, 2010.

In traditional publishing the defined unit is a block of text (journal paper, chapter etc). This is well defined, does not overlap with other items, has well understood characteristics and can be referred to without ambiguity. Data are more complicated, with boundaries between data sets which are often blurred and overlapping, and with varying structure. Often the data consumer needs considerable information in order to make sense of (or even "read") the data.

Van de Sompel et al. (2004) identified five functions that the process of Publication should perform:

- **Registration**: which allows claims of precedence for a scholarly finding.
- **Certification**: which establishes the validity of a registered scholarly claim.
- **Awareness**: which allows actors in the scholarly system to remain aware of new claims and findings.
- **Archiving**: which preserves the scholarly record over time.
- **Rewarding**: which rewards actors for their performance in the communication system based on metrics derived from that system.

In the context of data, there is an important extra function that is required:

- **Definition**: what is it that is being published?

The six functions can be grouped into two simpler categories:

- **Aids to Reusability**: things that make publications permanently available and the knowledge within useable in other contexts (Archive, Awareness, Definition); and
- **Recognition Enablers**: things that make it possible to measure and recognize the value of work (Registration, Rewarding, Certification).

From an author point of view these functions respectively enable the "right to know" and the "right to be known" (Willinsky, 2006).

Thus far, these functions do not make any reference to quality control. In practical applications in academia, this quality certification function is carried out by peer review. If a paper has passed the peer review process of Journal A and has been published, it can be assumed to have reached a level of certification as to the quality and possibly impact of the material. Of course, peer review itself is poorly defined, as different journals have different methods for dealing with editorial scrutiny, independent analysis, number of reviews and how they are used, etc. Nonetheless, peer review is the accepted process in the scientific community for evaluating the quality of scientific work.

Data Publication Procedures for Data

Community acceptance of peer review procedures to enable data Publication is now needed. However, not all communities may accept peer review of data, regardless of method. Brown (2003) summarized the arguments against peer review from a small group of molecular biologists that were arguing that existing data sharing methodologies were more than adequate:

> "… without public sharing, access to the data would not have been possible, and… peer review would only serve to complicate and slow down the scientific process."

Neither argument is persuasive: the entire purpose of peer review is to enable reliable sharing of information, and there is no reason why the introduction of peer review should preclude rapid sharing of data that has not been reviewed.

The key point is exactly what should data review procedures include? Data publication procedures have traditionally concentrated on the preservation and long term access issue, with an emphasis on associative metadata. The institutions attempting data publication often have policies that require such metadata, but in practice have yet to find the balance between near meaningless free text entries in defined categories and far too limited a subset of information from strongly controlled vocabularies. This metadata rarely receives independent scrutiny, and quality control issues of the data content itself are generally out of scope. Few institutions aiming to publish data are likely to have the in-house expertise to carry out such a procedure for all the data they might be asked to publish.

The importance of metadata is undoubted, as Gray et al. (2002) put it:

> "Data is incomprehensible and hence useless unless there is a detailed and clear description of how and when it was gathered, and how the derived data was produced."

However:

> "It's fine to say that scientists should record and preserve all this information, but it is far too laborious and expensive to document everything … And besides, who cares? Most data is never looked at again anyway."

The assertion that data is never looked at again is not true for environmental data; re-use is expected and crucial. From that perspective, the situation should be the same for data production as it is for paper readership. Current readers (data consumers) are desired, but future readers will also be important.

Of course, this future expectation does depend on the ability of the information to be curated in an archive for the long term. Unlike document publication, this is an extra requirement for data publication. In particular, we need to distinguish between preserving the bits and bytes (data archival) and preserving and migrating the information between formats and user services (data curation). Where data formats are not common and/or the services required by the user communities are complex, the job of ensuring that the data are fit for use as computing systems and interfaces change, and that the associated metadata are complete and use the appropriate vocabularies, will require expert teams of curators who have both computing and discipline knowledge. The peer review process needs to facilitate curation by minimising the proliferation of new vocabularies and formats, and maximising the metadata.

Required Metadata

In an analysis of the expectations of data users, Wang and Strong (1996) found that intrinsic data quality was of equal importance with the quality of the information that allowed users to identify the applicability of the data to their task, as was the information which aided direct interpretation, understanding and the accessibility. They characterized these last three as: contextual data quality, representational data quality, and accessibility data quality. For the purposes of metadata, it is the context and representation metadata that is important.

Much contextual metadata includes provenance metadata. In computer science, provenance metadata is usually understood to be automatically generated material that tracks changes as information products pass through some workflow (see Simmhan et al., 2005, for a provenance taxonomy). Here, we define provenance metadata more widely to include not only automatically generated material, but also human generated annotations and correlative information. Clearly, the process of collecting this can be exceedingly onerous, yet is important if the data is to be reusable.

Lawrence et al. (2009) introduced a taxonomy of metadata, which defines:

- **A**: Archive metadata – the material needed to manipulate the archive contents (also called Representation Information) and understand what the actual physical measurements might be (in terms of, for example, controlled vocabularies);
- **B**: Browse metadata – the material needed to put the archive contents in their scientific context, so this will include provenance metadata etc;
- **C**: Character metadata – including all the assertions about the importance of the material, including subsequent citations and annotations; and
- **D**: Discovery metadata – a subset of information useable to find the data in search engines and other online discovery services and catalogues.

All these classes of metadata are crucial to the Publication process, but the current methods of constructing metadata usually involve groups other than the data originators providing much of this metadata, since the originating scientist (or instrument) does not provide the data. It will be seen that there is an issue as to authorship and authenticity if some (or all of) this material originates from the publisher as well as the author.

One of the reasons asserted by some scientists for not providing metadata is that the information required will appear in "the paper of record" which describes the dataset and collection methods or algorithms used to produce it. We would assert that while this might be true, it generally is not, as papers are normally geared towards persuasion by constructing an argument by narrative. By contrast, the underlying data, and the metadata upon which the arguments are constructed, need to be explicitly identified for data publication. The distinction can be demonstrated with a short (very contrived) example: one might write in a paper "I watched steam rise from my coffee and from this I deduced it was hot," whereas, the underlying data has the following fact "Steam rose from my coffee (coffee looked at 2009-08-01 09.38)". While, the observation time might not have been germane to the argument being built in the paper, it might be to subsequent users of the data ("The coffee won't be hot by December 2009").

Data Refereeing Procedure

Peer review is vital to the processes of establishing scientific consensus, and carries with it concepts of authority and validity. A dataset which has been through peer review can be considered to have been through a process of scientific quality assurance. This check of scientific quality must be done by domain experts, and hence is out of scope for the data archive staff, though they may be able to make judgments about the dataset's technical quality (suitability of format, completeness of metadata etc). Historically, peer review has always been carried through processes developed by academic journals, and it makes sense to piggy-back on these already existing processes to peer review data – in other words, producing a data journal.

The data peer review procedure must ensure that all metadata is as complete as possible, but it must also address other qualities expected of Publication class material, such as the data's internal self-consistency, the merit of the algorithms used, the data importance, and its potential impact.

Internal self-consistency is relatively easy to evaluate: in the case of data, many checks can be automated, but it still requires a human to make summary judgments on the results. For example, a temperature dataset with units of Kelvin can be easily rejected if negative numbers appear, but if the units are Celsius, a different discrimination might be needed. In either case, a human might need to decide what the bounds of realistic numbers might be. One can also make additional requirements of data: for example, by requiring explicit assessments of measurement (or simulation) uncertainty.

In many cases, data will consist of observations of phenomena made with state-of-the art instrumentation, but even when the instruments are not the best or the latest, observations of real world phenomena still have value. The situation is more difficult where the data is produced by analysis or simulation using algorithms which are not regarded by the community as the best or most complete. In these cases, value judgments will need to be made as to whether future use of the data (without reproduction) is likely, and whether or not the data has some merit as evidence in its own right. This is more likely where there is some chance of a future legal challenge to conclusions that might be drawn from the data.

Similar judgments will need to be made regarding the importance and impact of data. Often individually unimportant data measurements can gain value from being aggregated, and within that, in many cases, there is a continuum of measurements which needs to be rather arbitrarily divided into datasets (or in our case "Publishable entities"). Traditional metrics of the value of databases are predicated on "usage equals value" (Wilson, 2001). However, many data producers will need to have a view of the far distant future. We expect that, just like the traditional journal world, data publishers will appear providing publications that are recognized to have a range of subject matter, quality and impact.

Data Ownership and Copyright

The list of functions associated with publication is silent with respect to ownership. With documents, copyright law subsumes the ownership issue. The authors may or may not assign exclusive rights to the publisher, but the copyright status of the publications should be clear. Unfortunately, in the case of data and databases, even the

appropriate area of law that might be applicable to published data is not clear (Waelde & McGinley, 2005), with the possibility of different laws being applicable in the UK and the US (Rusbridge, 2007).

Science Commons[4] and the Panton Principles[5] are endeavoring to address copyright issues and data licensing in open data. Earth System Science Data (ESSD) is an open access journal which allows authors to keep copyright of their data. Similarly, data repositories such as the British Atmospheric Data Centre[6] use a variety of data licenses, while the data stored in Pangaea[7] is provided under a creative commons license. The issues surrounding licensing and data ownership are complex, and so are considered to be out-of-scope for this paper.

Examples of Peer Review of Data.

In this section we briefly summarise the publicly available information about peer review in two existing publication scenarios, before presenting our own guide to the issues that need to be addressed in the environmental sciences.

X-Ray Absorption Fine Structure (XAFS) Spectroscopy.

Over a period of years, and a number of workshops, the XAFS community developed a reviewer's checklist to help referees assess papers presenting XAFS results. In this case, the publication methodology is essentially that of publication by proxy, with no backup raw data archive. Nonetheless, because the experimental method was so crucial, a checklist to help assess the data collection was developed. The checklist appears in Koningsberger (1993), and covers:

- The experimental procedure (with detailed questions about the experimental setup);
- Data reduction (with detailed questions about the methodology and explicit requirements for raw data);
- Data analysis (again, detailed questions about methodology, data analysis packages used, with requirements for raw spectra and explicit values of analysis parameters).

This simple list is presented here because it summarises quite succinctly the provenance metadata which needs to appear alongside data.

The NASA Planetary Data System (PDS).

The planetary data system provides high quality peer reviewed datasets, targeted to the very specific requirement of supporting NASA's planetary science. While this is very discipline specific[8], there are a number of characteristics of the PDS that have generic interest and are worth summarising here:

[4] Science Commons: http://creativecommons.org/science.
[5] Panton Principles: http://pantonprinciples.org/.
[6] British Atmospheric Data Centre: http://badc.nerc.ac.uk.
[7] Pangaea: http://pangaea.de.
[8] While discipline specific, the PDS does support a range of data types, as well as sub-disciplines, with mission data, as well as astronomical observations and laboratory measurements covering aspects of planetary science from planetary geoscience to atmospheres, rings and small bodies etc (PDS Standards Reference, 2009).

- There is a peer review process that requires that: a) the data are complete (e.g. there are no missing calibration files); b) the data are of sufficient quality and with enough documentation to be useful and intelligible in the distant future, and c) the PDS standards are followed.
- The PDS standards cover data format, content and documentation. Because the PDS supports a very wide variety of input data types, the format requirements are not onerous (it is allowable to construct complex new binary representations) but the concomitant documentation requirements are therefore much more specific. Following Lawrence et al. (2009) we would describe these as strong requirements on the A-Archive metadata, although there are also requirements for catalog files (D-Discover metadata).
- There is a very extensive data proposal process which defines what is needed to carry out data ingestion into the PDS. It is not simply a case of simply providing conformant data.
- Because the data holdings are relatively arbitrary binary and there are not machine understandable description documents. The PDS does not provide services layered over the data beyond discovery, file browsing, and download.
- There is a recommended citation format, and the citations for datasets are explicitly provided as part of the metadata.

A Generic Data Review Checklist

In this section we present a stratified summary list of activities that we believe are important parts of the data review process. Not all of these activities result in a pass or fail: there is considerable scope for subjective reviewer expertise, but some of them are rather mechanical and amenable to automated checking (although it should be noted even the objective tests are against the subjective criteria of the publication process).

In many cases, criteria can be made completely objective, in that we can write the criteria before the data is collected. But it is not possible to anticipate all possible future uses for a dataset, so some criteria may only be defined after the data is collected. Inevitably these criteria are "reactive" to some events, and to some extent subjective, even if it is then possible to automate their application.

This also raises the question of who makes the judgment about the bounds of the criteria, especially when it comes to outlier and/or unexpected results. In the first case it should be the dataset author, but though defining accuracy limits for a set of particular dataset uses is possible, it is not possible to define accuracy limits for all possible uses of the dataset. It is reasonable to expect that an anticipated list of dataset uses and limits should be provided.

The data review will vary according to discipline and type of data, given a full spectrum of possibilities from unrepeatable irreplaceable measurements to repeatable improvable simulations, though we have attempted to make the checklist as generic as possible.

Data Quality:

- Is the format acceptable? If so, is there an automatic format checker available, and if there is, does the dataset/file pass the automated checks?
- Are data values internally consistent? Do they fall within an appropriate range for the phenomenon being measured/observed/simulated? (For example, does a temperature dataset with values in degrees Kelvin have negative values?)
- Does the data represent reality with sufficient accuracy to use? Is the data of tolerable precision? (In the case of simulations, can the simulation be trivially repeated, in which case publication of the data is probably unwarranted.)
- Does the extent and coverage of the data match expectations? Does the coverage (spatial and/or temporal) add significant value to what is already available? (If not, is there added precision or some other reason for its publication? See also the discussion below on granularity.)
- Are the data values reported physically possible and plausible? (This requires significant domain knowledge, or a clear definition of what the data values range should be.)
- Is the data validated against an independent dataset? (Has it been calibrated?)

Metadata Quality:

- Is there sufficient quality metadata describing the format and physical content? (See for example, the requirements of the PDS, 2009.)
- Is there sufficient quality metadata describing provenance and context? Has the data changed in some way since it was measured? Is the processing chain visible and well documented? Have all the human interactions with the data prior to ingest/publication been recorded?
- Is there existing metadata (or are there references) already making assertions about the quality and usefulness of the data? If so, are these included in the metadata?
- Is there suitable quality discovery metadata? At a bare minimum, can Dublin Core be constructed?
- Does the metadata use appropriate, controlled vocabularies?
- Can all internal references (both electronic, e.g. URL/DOI, and traditional, e.g. to ISO690) be resolved to real entities? Are the external electronic references stored in a trusted repository? If not, can they be cached with the metadata?
- Is all the available metadata conforming to standards?

General:

- Is there an existing user community? Is that community happy that the data is usable? (This can be tracked after publication through citation, or before publication through the use of user surveys or comments in a process of open peer review.)

- What is the track record of the data source? Are they reliable?
- Are the intellectual property rights for the data established?
- Is the data available at the correct network address?

In some cases there will be electronic services, such as visualizations, associated with the data, in which case the reviewer will need to address the service/data compatibility and function:

- Do the advertised services work with the data? Is it likely that these services can be maintained with time?

This list is not exhaustive, but does display the range of possible checks. Obviously many of the checks above are metadata checks rather than data checks. This is indicative of the fact that quality data is not possible without quality metadata. It will be seen that the metadata checks essentially follow the metadata taxonomy from A-Archive to D-Discovery discussed above. In practice then, given complete and accurate data, the syntactical correctness and semantic completeness of the metadata is the key requirement of the review.

It is not possible to evaluate the reproducibility of results produced from a published data set by peer review of that dataset, as it is only the dataset, not its associated results, that are being published. However, once a dataset is Published it allows anyone peer reviewing the analyses resulting from that data to check the reproducibility of the results.

Data Publication Models

Data publication provides a bridge between human and computer in that it bridges the gap between data, which is consumed by a computer directly, and information, which is produced from data, but consumed by humans. Given that we are advocating the Publication of data, what methodologies to achieve this are possible? An analysis of existing publication activities yields the following basic classes:

1. Standalone Data Publication
2. Data Publication by Proxy
3. Appendix Data
4. Journal Driven Data Archival
5. Overlay Publication

These classes are discriminated in the main by how the roles involved in publication are distributed between the various actors. In this section, we identify the key roles and actors in the data publication process before using these roles and actors to discriminate between the classes defined above. The section concludes with a discussion of these models in terms of their overall strengths and weaknesses and where the responsibilities for data review lie.

Key Roles

- **Author**: Data creator, normally required to meet the initial data format specifications of the curator.

- **Resolver**: Maintains a document that includes links to the data itself and any metadata, and which is the primary citable entity.
- **Identifier Manager**: Controls how identifiers are distributed between data entities, archives, resolvers and documents.
- **Review Controller**: Controls the peer review process (if any).
- **Gatekeeper**: Controls access to the data and/or the metadata for validated users. Here "validated" users might be those who have "paid" subscriptions, or simply those for whom access has been granted via some other criteria, such as provision of an email address).
- **Metadata Editor**: Carries out editorial functions, including assembly and definition for the dataset metadata.
- **Metadata Creator**: The author of documents that describe the data.
- **Reviewer**: Assesses fitness of the data against publishers predefined and/or community accepted review criteria.
- **Archiver**: Responsible for the persistent storage of the datasets.
- **Curator**: Responsible for ensuring that the interfaces, format and metadata, are refreshed as necessary with time. Defines the acceptable data formats at ingestion.

It will be seen in the analysis of the individual classes that is useful to consider how these roles are distributed amongst the following traditional actors, who, in some cases, are themselves the same entity:

1 **The Journal**: Responsible for a process and "item of record."
2 **The Archive**: Responsible for data.
3 **The Author**: Creator of original material.

Third parties, outside Journal, Author or Archive, may act as consumers of the data, citers, or reviewers.

Stand Alone Data Publication

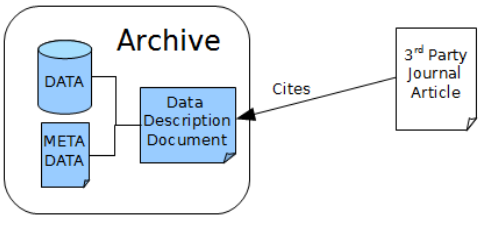

Figure 1. Stand alone data publishing.

The data is a publication in its own right, with no requirement for a co-existing standard journal article describing the data. The data archive provides systems which provide a data description document (DDD) as the citable item, and the data is obtainable either directly and electronically via links from that record, or via an application process which is accessible from that record. The requirements and definition of the data description document are varied, ranging from a simple web page with links to controlled format documents with standardized fields. While many data

archives (such as the British Atmospheric Data Centre) provide standalone data publication and carry out their own internal review as to whether to accept data, the extra procedural steps to regard such archives as Publishers (as defined above) are rarer: examples include the Planetary Data System[9], and the putative Earth System Atlas[10].

Role	Data Archive	Author	3rd Parties
Author		*Yes*	
Resolver	*Yes*		
Identifier Manager	*Yes*		
Review Controller	*Yes*		
Gatekeeper	*Yes*		
Metadata Editor	*Yes*		
Metadata Creator	*Some*	*Some*	
Reviewer	*Some*		*Some*
Archiver	*Yes*		
Curator	*Yes*		

Table 1. The distribution of roles in stand alone data publication. In this case, the archive subsumes many of the traditional "publisher" roles. Formal publication can be done by external actors, such as journals, but this publication model is often more informal. Note that the primary metadata creation should be carried out by the author where Publication is intended.

In this case, a third party citing the data will use an identifier to the data description document provided by the archive, and use that directly. In most cases, academic journal will not allow such identifiers to appear in the formal reference list, though exceptions are being made for those data description documents which are identified by DOIs.

The advantages of this system of publication are that the material describing the data forms part of the internal metadata of the data, and it should not be possible for a dataset description to exist in the absence of the dataset, or for it to become incorrect with time. A proper data archive will carry out curation functions over time including format migration etc, which could mean that independently managed data descriptions become invalid. The main disadvantage is that the methods of citation (from journals and between datasets) are not standardized, and despite explicit peer review and internal community regard, it is rare for wider communities to regard these publications as worth of academic recognition (as defined above). There is also an additional issue: by and large, such archives are embedded in academic or research institutions, which both submit their own data and organize their own peer review. In the UK at least, this is frowned upon, as one of the requirements of peer review is that it should be completely independent of the data submitters[11].

There is a variant on stand alone data publication, which is stand alone database publication. In some cases, rather than a dataset being embedded within an archive, the database itself is the publishable entity.

[9] Planetary Data System: http://pds.jpl.nasa.gov/.
[10] Earth System Atlas: http://earthsystematlas.sr.unh.edu/.
[11] For example: the British Atmospheric Data Centre (BADC) is funded by the Natural Environment Research Council (NERC) who require all awardees who produce atmospheric science data to "deposit" their data with the BADC. If the BADC were to organize their own peer review, there would be a bias towards acceptance (increased funding, meeting NERC goals etc).

Publication by Proxy

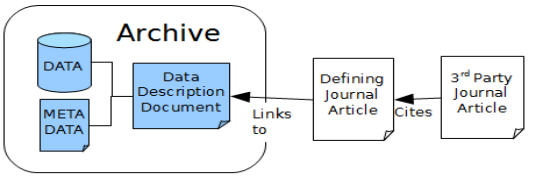

Figure 2. Publication by proxy.

In this case, the data is published independently of a conventional article written with the aim of both describing the data and providing a hook to the data location and/or access methodology. Such a paper generally describes the project and aspects of the algorithms and data, but is far from a complete description of the data that would enable a user to manipulate the data without reference to much other material, not all of which may be in the public domain. The refereeing procedures for the paper do not generally cover constraints on the quality control of the data and its documentation. Nearly all journals accept papers of this sort.

Role	Data Archive	Journal	Author	3rd Parties
Author			*Yes*	
Resolver		*Yes*		
Identifier Manager		*Yes*		
Review Controller		*Yes*		
Gatekeeper	*Yes*			
Metadata Editor	*Yes*			
Metadata Creator		*Some*	*Some*	
Reviewer				*Proxy paper only*
Archiver	*Yes*			
Curator	*Yes*			

Table 2. The distribution of roles in publication by proxy. In this case, the proxy paper is treated as any other academic paper with the same publication roles fulfilled by the journal. The data archive performs a more limited set of the publication roles quite separately for the data.

In most cases, the author of the dataset and the author of the proxy paper is the same. The paper often refers to quality procedures involved with the data production, but this is often shoehorned into a paper designed to describe scientific findings. However, where such procedures are described, they are within the purview of the journal paper reviewer, but in practice the data itself is subject to the same publication procedures outlined in the standalone data production case.

The advantages of this model are that it fits naturally with existing publication paradigms. The disadvantage is that the long term preservation of the data is separated from the paper (what worth is the paper without the data?), and is constrained by the policies and funding of a data centre host institution that may have little or no incentive to retain the data (particularly during periods of low usage). The data holdings themselves will adhere to the data centre's syntactic and information requirements, which may or may not be those required by the journal article user communities. A subsequent scientific activity citing the data would normally cite the journal paper, not the dataset itself.

Another disadvantage is that there would need to be safeguards put in place to ensure that the separately held data is not changed. The archive would need to have established and well applied policies on quality assurance, curation etc, in order to ensure that the dataset remains frozen.

Appendix Data

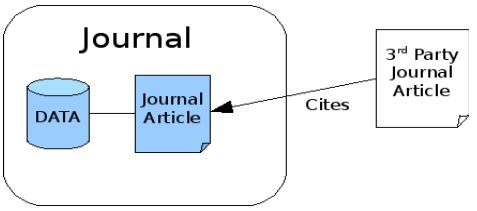

Figure 3. Appendix data.

In this case, data appears as supplementary material to a paper, and are submitted along with it. This is the model used by Nature[12] as well as a range of other electronic journals. In general, there are both size and format constraints on the supplementary material, and it is expected that the material will be reviewed along with the paper. There are not normally ancillary metadata: data needs to be fully described in the paper.

Role	**Journal**	**Author**	**3rd Parties**
Author		*Yes*	
Resolver	*Yes*		
Identifier Manager	*Yes*		
Review Controller	*Yes*		
Gatekeeper	*Yes*		
Metadata Editor	*Limited*		
Metadata Creator		*Limited (the paper is the metadata)*	
Reviewer			*Review of article including appendix data*
Archiver	*Yes*		
Curator	*Limited*		

Table 3. The distribution of roles in Appendix Data. In this case, the journal handles everything (but in general does not deal with metadata). The archival function is subsumed by the journal, but not all curation functions may be addressed.

The advantages of this method are that the paradigm is a natural extension of existing publishing options. Along with size and format limitations, the disadvantages include the expectation that the data is limited to only that germane to support the arguments presented and there is no evidence that long term curation issues are understood by traditional scientific publishers (although this is somewhat mitigated against by format restrictions). Citation is accomplished by citing the parent paper, but the data will not be independently discoverable. It is not obvious that review procedures are targeted at the data quality itself, or anything about the data, per se.

[12] For biological data, Nature operates the journal driven data archive model as well. These two models co-exist peacefully.

Journal Driven Data Archival

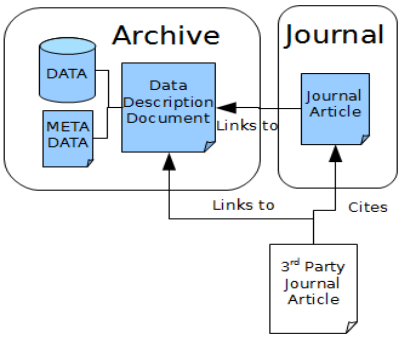

Figure 4. Journal driven archival.

In this type of data publication the need to publish data with papers along with constraints on journal space, has resulted in the creation of an ecosystem of databases which both serve a data sharing function and an archive of record function. The bioinformatics community abound with examples, one of which is the PloS Genetics Journal[13] who require that: "All appropriate datasets, images, and information should be deposited in public resources. Please provide the relevant accession numbers…"

PloS Genetics then recommend a set of public databases. Similarly, in the geophysics community, the American Geophysical Union has a data policy[14] which states:

> "… data cited in AGU publications must be permanently archived in a data center or centers that meet the following conditions:
>
> a) are open to scientists throughout the world.
> b) are committed to archiving data sets indefinitely.
> c) provide services at reasonable costs.
>
> … To assist scientists in accessing the data sets, authors are encouraged to include a brief data section in their papers. This section should contain the key information needed to obtain the data set being cited."

In the former case, it can be argued that the archives have been created to support the requirement to deposit and reference data, in the latter, the archives specifically listed were pre-existing national and international data centres. However, whatever the heritage, there is now a growing symbiotic relationship between a class of journals and a class of data centres. While that relationship is probably most mature in the biosciences, it is maturing rapidly in most disciplines. A sign of the maturity is the requirement by the journal that data integral to the argument be deposited before the publication can be accepted (by this metric, the geophysics community is still immature). In general, electronic journals that require this activity also allow "Appendix Data" if there is no suitable repository and volume and format issues can be resolved.

[13] The Public Library of Science (PloS) Genetics Journal: http://genetics.plosjournals.org.
[14] American Geophysical Union Data Policy: http://www.agu.org/pubs/authors/policies/data_policy.shtml.

Role	Data Archive	Journal	Author	3rd Parties
Author			*Yes*	
Resolver	*For data*	*For article*		
Identifier Manager	*For data*	*For article*		
Review Controller		*Yes*		
Gatekeeper	*Data is usually open in this model*	*Yes*		
Metadata Editor	*Yes*	*Yes*		
Metadata Creator		*Some*	*Some*	
Reviewer				*Review of article and data*
Archiver	*Yes*			
Curator	*Yes*			

Table 4. The distribution of roles in journal driven archival. The journal controls the publication workflow allowing it to take on the role of gate keeper.

In most cases, paper authors will be data authors, and as such are responsible for submitting both the data to an archive, and the papers to journals. These are two separate and independent submissions, linked only by the condition that if data is referred to in the paper it should be submitted first to the data archive in order that an appropriate resource reference pointing to a data description document (an accession number in the case of the biosciences) is generated. The data review process is again the same as the stand alone data publication situation, with the journal reviewer not responsible for looking at the underlying data.

The advantages and disadvantages of this methodology are similar to those in "Publication by Proxy". Additional advantages are the requirement by the journal that the germane data is deposited, that there is a defined methodology of referencing, and the references to the datasets can appear in the reference list, thus enabling the development of citation metrics for the data itself. The disadvantage remains that the archive retention policies are not under the same governance and persistence policies of the referencing journal, and that the data itself is not explicitly reviewed.

Overlay Data Publication

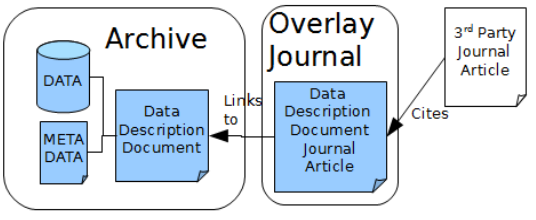

Figure 5. Overlay data publication.

In this case, the journal does not control the primary material, but controls some material that makes assertions about the primary material, and controls the all-important refereeing procedures. The concept of an overlay publications was apparently (Enger, 2005) first introduced by Ginsparg (1996), who stated:

> "… we can imagine a relatively complete raw archive unfettered by any unnecessary delays in availability. Any type of information could be overlayed on this raw archive and maintained by any third parties."

Here we are defining overlay data publication as data publication via an overlay journal explicitly targeted at data publication. In this case, the key content would be a data description document, which adds all the content which might be missing in the primary archive: for example, making additional assertions about the importance of the data set, and possibly provenance (although that may appear with dataset metadata). A particular requirement of the data description document is that it should not include anything that might age, as the data is managed by the underlying archive (for example, the format might be migrated, so format descriptions should remain with the archive).

The review process carried out by the overlay journal would be expected to make demands of the quality of the data held in the archive, and of the metadata within the archive as well. Reviewers might well make comments that would result in changes to both, which we would anticipate requiring new versions of either to appear in the archive: it would be those that were published rather than the originals.

Authors would be expected to submit data to the archive, and data description documents to the data journal as well (it could be argued that a third party might do this). The author would be expected to respond to reviewers for changes in data, metadata, and data distribution document content. Metadata created by the curator might also need to be modified in response to review.

Like the situation with journal driven data publication, the overlay journal could point to data held in multiple different repositories. The key distinction between the two models would be the expectation that the overlay journal would be dedicated to data publication, with procedures (and relationships with data archives) targeted towards delivering respected data review.

Role	Data Archive	Journal	Author	3rd Parties
Author			*Yes*	
Resolver	*For data*	*For data article*		
Identifier Manager	*For data*	*For data article*		
Review Controller		*Yes*		
Gatekeeper	*Yes*	*Data article should be open in the model*		
Metadata Editor	*Yes*	*Yes*		
Metadata Creator	*Some*		*Some*	
Reviewer				*Review of data article and data*
Archiver	*Yes*			
Curator	*Yes*			

Table 5. The distribution of roles in Overlay Data Publication.

The advantage of this methodology is that it combines the rapid information dissemination aspects of publishing without a process with the ability to subsequently assess (and potentially modify) the material through peer review. It also allows a data overlay journal to describe/index data holdings in multiple repositories. The disadvantages are shared with the Journal Driven Archiving model: the overlay journal does not necessarily control the persistence and reliability of the underlying archive.

There is a special case of the overlay journal: where the journal controls the underlying archive and the overlays are constrained to point only to it. This would mitigate against the major disadvantage, but would make the situation more analogous to stand alone data publication. However, it would still be possible to allow a third party to control the quality control procedures without controlling the delivery, as is done with research societies contracting out their journal publishing. This does raise the issue that there will need to be clear lines of accountability if third parties are allowed to control the quality control without controlling the delivery of the data. The mechanisms for an overlay journal could easily be managed through the use of a blog, but this would not convey the authority and validity that Publication requires. Overlay journals would therefore need to be run by a well respected organization, which is already trusted by the scientific community, and could extend that trust to the data archive controlling the data delivery.

Further work regarding the business cases required to use overlay journals as a technology for data publication was carried out as part of the JISC-funded Overlay Journal Infrastructure for Meteorological Sciences (OJIMS) project. User surveys carried out by the project team determined that there is a significant desire in the meteorological user community for a data journal, which would allow scientists to receive academic recognition (in the form of citations) for their work in ensuring the quality of datasets. The sponsors and funding bodies for the experimental campaigns that produce these data would also benefit, as it would encourage scientists to submit their data to accredited data repositories, where they would be archived and curated. Further information on the OJIMS project can be found in Callaghan et al. (2009a and 2009b).

Currently Existing Data Journals

Copernicus Publications currently operates the journal Earth System Science Data (ESSD) [15], which publishes data and has the following stated aim:

> "Earth System Science Data (ESSD) is an international, interdisciplinary journal for the publication of articles on original research data(sets), furthering the reuse of high (reference) quality data of benefit to Earth System Sciences."

ESSD has a two-stage publication process where, in the first stage, papers that pass a rapid access peer review are immediately published on the Earth System Science Data Discussions (ESSDD) website. They are then subject to Interactive Public Discussion, during which the referees' comments (anonymous or attributed), additional short comments by other members of the scientific community (attributed) and the authors' replies are also published in ESSDD. In the second stage, the peer review process is completed and, if accepted, the final revised papers are published in

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

ESSD. ESSDD and ESSD are both ISSN-registered, permanently archived and fully citable, ensuring that a lasting record of the scientific discussion surrounding a paper is maintained and that publication precedence for authors is confirmed.

Responsibilities for Review in Data Publication

It can be seen that with the exception of the Appendix Data model, all four other publication methodologies are built around the existence of a functional data archive. Of those four, with the exception of the stand alone data publication, actors outside the Archive carry out most of the extra roles that result in the sobriquet "Publication". It is the distribution of those roles that distinguish the classes, and for the purposes of this paper, the key ones are those associated with review.

We do not discuss the Appendix Data model further, as we believe it is functionally limited given that there is no explicit data curation, and very limited scope for the direct integration of the published data into academic workflow. As such, it doesn't meet the driving requirement of the changing nature of research. Of the four remaining, we have argued that stand alone data publication and the overlay journal model both potentially support direct and complete review of the data and metadata, while publication by proxy and journal driven only support indirect review via whatever is included within journal article. Thus there are two models that support data publication per se: The main distinctions between the overlay journal and the stand alone data publication are the explicit decoupling of responsibilities between the data archive and the overlay journal, and the ability for the overlay data journal to support multiple primary data archives. Whilst the latter provides a significant advantage over stand alone data publication, for the purpose of this paper, our main interest is how the explicit decoupling could be arranged.

We have seen that the review procedures consist of both objective and subjective analysis of both the data and the accompanying metadata. We have also seen that the creation of content itself (data, and metadata) might be decoupled. A key question then is how these sub-roles could be optimally distributed to get the best results for the peer review process both in terms of primary (data) author experience and output product. Clearly it is desirable to minimise the requirement for data resubmission, so as much as possible the objective data checking criteria should be handled during the original archival ingestion process (where it might be possible to verify – and reject if necessary – individual data files rather than entire datasets). Similarly, all syntactic checks for format and vocabulary conformance should be carried out as much as possible within the archival ingestion process (whether automatically or by archival staff), leaving the semantic and completeness checks for the review process to be carried out in the overlay review process. This requires the author to take on the entire A and B metadata creation.

This would be not only optimal in terms of the author experience, but also in terms of the requirements on archive systems and staff: as much as possible of the "discipline expertise" must be offloaded to the external review, minimising requirements on broad discipline expertise within data centre staff. Of course, these are not minimised to the extent of their not being needed: an important role of the data centre remains the curation function, which does require discipline knowledge and good expert relationships with user communities.

Citing Published Data

Even in a community where the dependency on data archival is required (such as the bioinformatics Genbank community), there is no standard way to refer to the data in the archive. In a study of genomic and proteomic database usage, Brown (2003) found that the citations into the databases were reported in a variety of ways. While individual journals became more explicit in their instructions (how to cite gene sequences in databases) to authors over time, at that point no convergence of syntax was reported.

While this might be acceptable within one discipline, where a kind of "received wisdom" can be developed so that the methodologies and interpretation of the citations on a journal-by-journal basis become passed on via collaboration networks, this is not conducive to use by the wider academic community. It also hides a number of difficulties with data citation that become apparent when the data being cited conforms to more complex data models. Buneman and Silvello (Buneman, 2006; Buneman & Silvello, 2010) looked into this problem from the perspective of curated scientific databases, where the aim is to automatically generate citations which are machine readable, but at the same time are understandable to a human. Their findings do not all agree with our suggestions, but there is much common ground.

Even where journals are trying to establish codes of practice that might be aimed at wider applicability, there is a sort of "citation hysteresis" in the notation: aspects of citation information which are appropriate for traditional paper publication are still being required. For example, the American Geophysical Union has guidance for references to data:

> "The format for the reference will be specified in AGU's guide for contributors. The following elements must be included in the reference: author(s), title of data set, access number or code, data center, location including city, state, and country, and date."

There seems to be no benefit in requiring the physical location of the data centre in an Internet based data reference.

In order to establish an appropriate convention for data citation, we have canvassed active scientists about what information would be necessary in a reference, held a workshop to address the results and come up with a recommended citation format.

Key issues identified in the interviews were the need for a human understandable, unambiguous reference to a well-defined permanent entity. To make the reference unambiguous, the following pieces of information would be required:

- Author,
- Publication year (or equivalents),
- Activity or tool that produced the data,
- An unambiguous reference to the source of the data.

The practicing scientists also had some concerns about the process of publishing and citing data. In particular, they felt the granularity of the dataset needed to be addressed. For example where there is a facility providing data from a set of instruments, does the facility or a particular instrument comprise the dataset level? There were concerns about publishing incremental data, the versioning of data and the need for the granularity to have meaning for users of the data rather than for the convenience of the data producers. Data producers have requirements about citation of their data so that it could be used for service metrics and paper location; however, their main concerns were that it should be traceable to the data provider and to be recognised as intellectually equivalent to academic papers.

These concerns echo and extend those of similar work reported by Klump et al. (2006), who listed persistence and quality as the two issues most important for data publication. In their work, the issue of persistence was dealt with, in part, by constructing Digital Object Identifiers for datasets registered with the Technical University of Berlin. They did not address data quality, which has been the main thrust of this paper.

In the remainder of this section we expand on the issues of granularity and transience that are specific to the citation of data, discuss what the target of citation should actually be, and what it means. We then discuss existing best practice in citation before introducing our recommended syntax.

Issues of Transience, Permanence and Granularity

The issue of transience does not exist in traditional data based publication, in which case the date of publication has real and immutable meaning. Likewise, permanence is a matter of arranging the safe custody of a paper – preferably multiple copies thereof. In the case of Internet publication, an identifier may refer to a resource which has changed (or even disappeared). In the short history of Internet citation, this has been dealt with by appending the common syntax of, for example, "accessed on 31/12/2009" to a URL. However, this syntax does not support the requirement of data publication to have an unambiguous and resolvable reference to a dataset as it was when cited. Nor does it address persistence. As outlined above, we are not addressing persistence here in any detail, but we assume that the persistence of the relevant bits and bytes can be arranged. The issue then is that they should not be changed, and that the citation should always point at the relevant ones!

In the most part, issues of transience should be dealt with during the process of Publication. To be peer reviewed, a dataset should not be transient: neither being updated by appending data (as might happen with time series of climatological data) or by replacement (as might happen when erroneous measurements are replaced). Both cases should be dealt with the issuing of new editions of data (and re-review). It has been suggested that automatically generated data, with automatic provenance, might introduce new problems here, but we would argue that the introduction of automation changes nothing, since such automation *precedes* the decision to publish (and thus set a *specific version* of the data in stone). However, the requirement of new editions of data as more data is collected/produced or the data is better analysed, means there is an obvious issue as to granularity: How many new records should be collected before submitting a new dataset for publication? While this question is really a question for the review process (it's a subjective decision for the publisher to provide criteria and

the reviewer to judge), it still leaves questions for the citation mechanism: How does one cite into an aggregated dataset? How does one denote "new" editions of data?

What Should a Citation Refer To?

As discussed above, a citation will usually (but might not always) resolve to a human readable document that we have called a "data description document". To that extent, the notion of a citation is immediately transparent. However, in the same way as most existing citation notations immediately give guidance as to whether the target of the citation is a book, thesis, journal article, CD, DVD or web site, there is much to be gained from the citation giving guidance about what is being cited in the case of data.

Again, as discussed above, the concept of data citation admits a wide range of citation targets. Examples might include digital spectra output from instruments, images output from cameras, binary datasets produced from simulations and gene sequences as tabulated codes. However, while it is obviously possible to include text in a citation, such as that in the previous sentence, it's not obvious that such broad textural descriptions are enough. The more specific the information in the citation, the more easily the reader can evaluate the necessity of accessing the target information. The reason why existing citations work so well in providing this information is that the number of types of target (book, DVD etc) is small, and the nouns (book, DVD etc, whether explicitly named in the citation or implicit via the syntax of the citation) are well known to all readers. The situation is not the same for data. A data citation needs to both indicate the class of item being referenced, and potentially include the equivalent of page numbers to identify portions of the citation target.

In terms of a notation to describe what is cited, there are already pre-existing international standards which provide context for describing things in the real world: ISO19101 (2002) introduces the notion of a "feature" as an abstraction of a real world phenomenon, and ISO19110 (2005) introduces the concept of dictionaries and registers of features. While both have been introduced in the geospatial domain, the concepts are far more widely applicable: that we can name features of the real world, define their attributes of interest, and register their descriptions. From the point of view of citation, that means that if we can use as part of the citation a defined feature name from a defined registry to identify the target, a human reader will either be instantly able to recognise the feature name, or take advantage of the feature registry to resolve the nature of the target. If the citation points to a data description document, that data description document should also point, amongst other things to the same feature descriptions!

In the most cases, citeable datasets will consist of feature collections (such as collections of gene sequences and aggregations of remote soundings from radars), but in many cases citing authors will want to indicate specific targets within the datasets (e.g. specific genes or soundings). With the concept of a feature available, we not only have the notion of defined feature collections being a feature in their own right, but we then have the data analogy of a page, with specific features being potentially identifiable within collections (with or without separate authorship). We present examples of this below.

The concept of a feature description should not be confused with the format (syntax and/or encoding) of the citation target. The feature description provides information as to the semantic nature of the target so that, for example, the notion of a profile number referring to a portion of a profile collection makes sense without any knowledge of how the profiles are formatted. The format and syntactical descriptions would be expected to appear as part of the metadata.

Regrettably, while the notion of features is well established, and there is an established methodology – the Geographic Markup Language (ISO19136, 2005) – constructing machine readable descriptions of (geographic) features, there is little best practice in terms of feature type registries. Nonetheless, the notion of pointers to features and feature-type registries is enough to allow us to proceed with citations based on definitions of these being made available as part of the metadata, with the anticipation that eventually community-governed permanent registries will become common.

Existing Citation Formats

Examples of existing best practice include:

- The PDS citation format[16], which can be summarised as:

  Author(s), "Title", Journal (always NASA Planetary Data System), Dataset ID, (Optional) Volume ID, Year (of publication).

  For example:

  Christensen, P.R., N. Gorelick, G. Mehall, and K. Bender, "Mars Global Surveyor Thermal Emission Spectrometer Standard Data Record", NASA Planetary Data System, MGS-M-TES-3-TSDR-V1.0, vols. MGST_0001 – MGST_0061, 1999.

- The German project "Publication and Citation of Scientific Primary Data" (Klump et al., op.cit., Brase & Schindler, 2006), which uses DOIs to construct references expected to be for the form:

  Author(s) (Year): Title (doi:opaque_assigned_identifier).

  For example:

  Hal, G (2005): IPCC-DDC_CSIRO_SRES_A2: 140 YEARS MONTHLY MEANS Commonwealth Scientific and Industrial Research Organisation Australia (doi:10.1594/WDCC/CSIRO_SRES_A2).

[16] Policy for Citations of PDS Data: http://pds-geosciences.wustl.edu/citations.html. Retrieved December 11, 2009.

- DataCite[17] have developed a metadata scheme for providing the metadata that should be attached to a dataset which has had a DOI assigned to it. It consists of five mandatory and twelve optional properties, which may be consumed by computer or assembled to create a human-readable citation string. DataCite remains discipline-agnostic concerning matters pertaining to academic style sheet requirements, as its members come from a wide range of scientific disciplines. It therefore recommends rather than requires a particular citation format using the mandatory properties of the metadata scheme:

  Creator (PublicationYear): Title. Publisher. Identifier.

  DataCite also recommend the following form when information about Version and ResourceType is required:

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

  For example:

  Irino, T; Tada, R (2009): Chemical and mineral compositions of sediments from ODP Site 127-797. Geological Institute, University of Tokyo.doi:10.1594/PANGAEA.726855. http://dx.doi.org/10.1594/PANGAEA.726855.

  Geofon operator (2009): GEFON event gfz2009kciu (NW Balkan Region). GeoForschungsZentrum Potsdam (GFZ). doi:10.1594/GFG.GEOFON.gfz2009kciu.http://dx.doi.org/10.1594/GFZ.GEOFON.gfz2009kciu.

  Denhard, Michael (2009): dphase_mpeps: MicroPEPS LAF-Ensemble run by DWD for the MAP D-PHASE project. World Data Center for Climate. doi: 10.1594/WDCC/dphase mpeps. http://dx.doi.org/10.1594/WDCC/dphase_mpeps.

Probably the most complete analysis of citation methodologies for databases on the Internet is that of Patrias (2007) for the U.S. National Library of Medicine, who has multiple pages of recommendations for citing databases and retrieval systems online. Patrias addresses three different scenarios: citing entire databases and/or retrieval system, citing parts of such systems, and citing contributions. In doing so, the issues of granularity and transience we have outlined above are partially addressed, but not the issue of semantics. Without an understanding of what is being referenced, and without the concept of features, there is a rather clumsy methodology for providing length:

[17] DataCite Metadata Scheme for the Publication and Citation of Research Data Version 2.1: http://dx.doi.org/10.5438/0003.

> "Provide the length of the part to a database when possible. Calculate the extent of the part using the best means possible, i.e., number of paragraphs, screens, bytes, or pages if printed. Since screen size and print fonts vary, precede the estimated number of screens and pages with the word about and place extent information in square brackets, such as [about 3 screens]."

In this definition we can see the problems of trying to apply print media concepts to data.

Buneman and Silvello (Buneman, 2006; Buneman & Silvello, 2010) also propose a rule-based citation system for structured and evolving datasets/databases to allow computers to automatically generate citations.

Recommended Citation Syntax

Despite the problems and limitations of Patrias' (2007) exposition, we believe it is the best starting point for constructing a generic data citation syntax. That syntax can be summarised as:

> Author(s). Title [Content Designator Medium Designator]. Edition. Place of Publication: Publisher. Date of Publication [Date of Update/Revision; Date of Citation]. Extent. (Series). Availability. (Language). Notes.

We now consider these elements in turn, addressing the issues we have identified, before presenting our modified version with examples.

Author: Note that the author of an incremental dataset may be hard to identify. Both the principle investigators and any corporate body providing the means to get the data might be recognised. If this is the case, individuals should be named in parentheses after corporate names.

Title: This should identify the data resource, which may or may not include a facility name.

[Content Designator Medium Designator]: This is an opportunity to introduce the feature type. Because the feature type should be a registry member, or at the very least, an entry in a controlled vocabulary, the URN or URL of that member should also be included. So, we would replace this with *FeatureName, FeatureURN.* Note that we believe that including "Internet" here is redundant, as the appearance of a URL or DOI later in the citation carries the information that the material is on the Internet.

Edition: Data may have several versions of processing and multiple levels of product (e.g. measurements below the orbit tracks for satellite data may be one level of product, and a gridded global product may be another). In practice, the review process will provide a nomenclature for the edition that is appropriate for the data type. We would advise that this nomenclature should be chosen from a controlled vocabulary.

Place of Publication: This has no value for an internet resource so we recommend its omission.

Publisher: The organisation responsible for hosting the data.

Date of Publication: This, like a traditional journal publication, should be the date at which the peer review process has completed *and* the data has appeared. Note that the data may well appear before the peer review process has completed!)

[*Date of Update/Revision; Date of Citation*]: As this data has been through a review process, and is expected to be permanently available, this section can be omitted.

Extent Series. We would use this to put in a universal resource name (URN) which might differ from the URL at which the data is downloadable, but which is intended to be persistent. Where it is desirable to point into a larger dataset or collection to a specific feature member or members, we would add notation as follows: either *fid* or the letter *f* followed by the feature id, feature id list, or range.

Availability: A URL from which either the DDD or the data is available. This would be omitted if the URN provided was a Digital Object Identifier (DOI). Note that in both cases, the link might point to a different distributor website than the implicit publisher website.

The following fields would remain optional: *Language* and *Notes*.

If we followed the same order of material, our citation would then be:

> Author, Title [featurename, featureID]. Publisher. Year. DOI or (urn:URN, fid:x [Available at URL]).

However, the workshop participants also recommended moving the date away from the URNs etc, to make it easier to scan, so we have:

> Author (Date). Title [Featurename, featureID]. Publisher. DOI or (urn:URN, fid:x [Available at URL]).

To summarise the differences from Patrias (op.cit.), we see that:

1. We are dealing with published data. We can remove the citation date.
2. We have introduced a URN and an optional feature identifier.
3. We are always using URLs or DOIs that indicate we've got Internet media. We lose the [Internet] designator.
4. We have introduced a feature descriptor after the title to define what it that is being reviewed.

Five examples follow. The first two are contrived versions of the examples of existing practice presented earlier, and then three following are hypothetical, since the datasets involved have not been through any "peer review" procedure.

> Christensen, P. R., N. Gorelick, G. Mehall, and K. Bender (1999). Mars Global Surveyor Thermal Emission Spectrometer Standard Data Record [volumes, http://pds.jpl.nasa.gov/documents/sr/] NASA Planetary Data System. urn: MGS-M-TES-3-TSDR-V1.0 fid: MGST_0001 - MGST_0061. [Available from http://starbrite.jpl.nasa.gov/pds/viewDataset.jsp?dsid=MGS-M-TES-3-TSDR-V1.0]

As well as the date reorder, note the addition of (i): a (fictitious) feature type (trying to define the volume concept, but it should really try and define the nature of the spectrometer records themselves); and (ii): a real url that can be resolved.

Commonwealth Scientific and Industrial Research Organisation Australia [Hal, G.] (2005): IPCC-DDC_CSIRO_SRES_A2: 140 YEARS MONTHLY MEANS. [GridSeries, http://ndg.nerc.ac.uk/csml2/GridSeries]. World Data Centre for Climatology. doi:10.1594/WDCC/CSIRO_SRES_A2.

Note that (i): now it looks (correctly) like the Commonwealth Scientific Industrial Research Organisation is a corporate author, rather than the publisher, which is the World Data Centre (who organised the review), and (ii) the type of the data is now clear and that the feature type definition doesn't have to be owned by the publisher.

> Iwi, A. and B.N. Lawrence. A 500 year control run of HadCM3. [GridSeries, http://ndg.nerc.ac.uk/csml2/GridSeries] British Atmospheric Data Centre, 2004. urn: badc.nerc.ac.uk__coapec500yr. [Available from http://badc.nerc.ac.uk/data/coapec500yr].

This example differs from the previous one in that there are no corporate authors, and the syntax is URN, [Available at] rather than the DOI version.

> Iwi, A. and B.N. Lawrence. A 500 year control run of HadCM3. [GridSeries, http://ndg.nerc.ac.uk/csml2/GridSeries] British Atmospheric Data Centre, 2004. urn: badc.nerc.ac.uk__coapec500yr. fid:jaekfxy [Available from http://badc.nerc.ac.uk/data/coapec500yr].

This example adds the option of identifying a specific grid within the gridseries via the feature id. The same notation could be used to identify a specific spectrum within a collection of spectra or gene sequence within a collection.

> Natural Environment Research Council, Mesosphere-Stratosphere-Troposphere Radar Facility [Thomas, L.; Vaughan, G.] (2001) Mesosphere-Stratosphere-Troposphere Radar Facility at Aberystwyth: The 1990 Decade. [ProfileSeries, http://ndg.nerc.ac.uk/csml2/ProfileSeries]. Version 2, Cartesian Products. British Atmospheric Data Centre (BADC). [urn:badc.nerc.ac.uk__mst1990s]. Available from http://badc.nerc.ac.uk/data/mst/1990s.

In this case, we see a complex corporate authorship and the use of a facility name in the title of the data. The hypothetical review process has imposed the decadal granularity in what is an ongoing collection of data. Clearly different granularities would be possible (campaigns, months, years etc). The appropriate granularity will be an "editorial" decision for the publishers. There is also an edition number, and a product designator (not, regrettably, from a controlled vocabulary).

There are obviously many more cases that could be examined, but these suffice to show the intent.

Conclusions

In this paper we have begun by motivating the necessity for peer review of data, described some of the aspects of such a review, introduced some possible methodologies for data publication and discussed how peer reviewed data might be cited.

In our discussion of peer review, we have presented criteria which mainly address the completeness and accuracy of the metadata, but the raw quality of the data, along with its relevance, context and provenance is obviously important. We reiterate that in practice we imagine that different publishers will introduce different review strategies, that will result in a spectrum of different data publications in terms of subject matter, completeness of review, and both implicit and explicit data qualities, much as exists in the traditional academic journal world.

The introduction to publication methodologies introduced five basic classes which differed in the main around how and where the peer review would occur. We argue that only stand alone data publication and overlay data journal publication (as we have defined them) offer comprehensive review of the metadata and data itself and thus offer "true" data publication.

We have introduced a citation syntax that would clearly identify what is being cited, as well as provide clear differentiation between the publisher and the distributor. A key component of the citation syntax is the presence of a feature type description, as well as the ability to cite features within feature collections.

Acknowledgements

The authors acknowledge the support of both the Joint Information Systems Committee and the Natural Environment Research Council in carrying out this work. The authors would also like to thank the reviewers for their insightful comments and criticism.

References

Armstrong, J.S. (1997). Peer review for journals: Evidence on quality control, fairness, and innovation. Science and Engineering Ethics, 3(1), 63-84. doi:10.1007/s11948-997-0017-3

Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G., Casey, K., Laaksonen, L., Moorman, D., Uhlir, P., & Wouters, P. (2004). Promoting access to public research data for scientific, economic, and social development. Data Science Journal, 3, 135-152. doi:10.2481/dsj.3.135

Brase, J. & Schindler, U. (2006). The publication of scientific data by the world data centers and the national library of science and technology in Germany. Data Science Journal, 5, 205-208. doi:10.2481/dsj.5.205

Brown, C. (2003). The changing face of scientific discourse: Analysis of genomic and proteomic database usage and acceptance. Journal of the American Society for Information Science and Technology, 54, 926–938. doi:10.1002/asi.10289

Buneman, P. (2006). How to cite curated databases and how to make them citable. In Proceedings of the 18th International Conference on Scientific and Statistical Database Management (SSDBM). Vienna, Austria: IEEE Computer Society.

Buneman, P. & Silvello, G. (2010). A rule-based citation system for structured and evolving datasets. IEEE Data Engineering Bulletin, 33(3), 33-41. Retrieved July 12, 2011, from http://sites.computer.org/debull/A10sept/buneman.pdf.

Callaghan, S.A., Hewer, F., Pepler, S., Hardaker, P., Gadian, A. (2009 a). Overlay journals and data publishing in the meteorological sciences. Ariadne, 60. Retrieved July 7, 2011, from http://www.ariadne.ac.uk/issue60/callaghan-et-al.

Callaghan, S.A., Pepler, S., Hewer, F., Hardaker, P., Gadian, A. (2009 b). How to publish data using overlay journals: The OJIMS project. Ariadne, 61. Retrieved December 11, 2009, from http://www.ariadne.ac.uk/issue61/callaghan-et-al/.

Carr, T.R., Buchanan, R.C., Adkins Heljeson, D., Mettilee, T.D. & Sorenson, J. (1997). The future of scientific communication in the earth sciences: The impact of the Internet. Computers and Geosciences, 23(5), 503-512. doi:10.1016/S0098-3004(97)00032-0

Costello, M.J. (2009). Motivating online publication of data. BioScience, 59, 418-427. doi:10.1525/bio.2009.59.5.9

De Waard (2007). A pragmatic structure for research articles. In S. Buckingham Shum, M. Lind & H. Weigand (Eds.), Proceedings ICPW'07: 2nd International Conference on the Pragmatic Web. Tilburg: NL. Retrieved July 7, 2011, from http://oro.open.ac.uk/9275.

Enger (2005). The concept of 'overlay' in relation to the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). Masters thesis, University of Tromsø, Norway. Retrieved July 26, 2011, from http://hdl.handle.net/10760/6960.

Ginsparg, P. (1996). Winners and losers in the global research village. Retrieved June 27, 2011, from http://xxx.lanl.gov/blurb/pg96unesco.html.

Gray, J., Szalay, A.S., Thakar, A.R., Stoughton, C., van de Berg, J. (2002). Online scientific data curation, publication and archiving (Technical Report MSR-TR-2002-74). Redmond, WA: Microsoft Research. Retrieved July 26, 2011 from http://research.microsoft.com/pubs/64568/tr-2002-74.pdf

Hancock, W.S., Wu, S.L., Stanley, R.R. & Gombocz, E.A. (2002). Publishing large proteome datasets: scientific policy meets emerging technologies. Trends in Biotechnology, 20(12), S39-S44. doi:10.1016/S1471-1931(02)00205-7

Hurd, J., Brown, C.M., Bartlett, J., Kreitz, P., & Paris, G. (2002). The role of "unpublished" research in the scholarly communication of scientists: Digital preprints and bioinformation databases. Proceedings of the American Society for Information Science and Technology, 39(1), 452-453. doi:10.1002/meet.1450390153

ISO19101 (2002). Geographic information — Reference model.

ISO19110 (2005). Geographic information — Methodology for feature cataloguing.

ISO19136 (2005). Geographic information – Geography Markup Language.

Klump, J., Bertelmann, R., Brase, J., Diepenbroek, M., Grobe, H., Hock, H., Lautenschalger, M., Schindler, U., Sens, I. & Wachter, J. (2006). Data publication in the Open Access Initiative. Data Science Journal, 5, 79-83. doi:10.2481/dsj.5.79

Koningsberger, D.C. (1993). Report on activities of Committee on Standards and Criteria in XAFS Spectroscopy. Japanese Journal of Applied Physics, 32(Suppl. 32-2), 877-878. Retrieved July 26, 2011, from http://jjap.jsap.jp/link?JJAPS/32S2/877.

Lawrence, B.N., Lowry, R., Miller, P., Snaith, H. & Woolf, A. (2009). Information in environmental data grids. Philosophical Transactions of the Royal Society A, 367, 1008-1014. doi:10.1098/rsta.2008.0237

Patrias, K. (2007). Citing medicine: The NLM style guide for authors, editors, and publishers. Bethesda, MD: National Library of Medicine. Retrieved June 27, 2011, from http://www.nlm.nih.gov/citingmedicine.

PDS Standards Reference, Version 3.8 (2009). Retrieved 12th July 201, from http://pds.nasa.gov/tools/standards-reference.shtml.

Research Information Network (2008). To share or not to share: Publication and quality assurance of research data outputs. Retrieved June 27, 2011, from http://www.rin.ac.uk/our-work/data-management-and-curation/share-or-not-share-research-data-outputs.

Roosendaal, H.E. & Geurts, P.A. (1997). Forces and functions in scientific communication: An analysis of their interplay. In M. Karttunen, K. Holmlnund, & E.R. Hilf (Eds.) *CRISP 97, Cooperative Research Information Systems in Physics.* Retrieved July 12, 2011, from http://www.physik.uni-oldenburg.de/conferences/CRISP97/.

Rusbridge, C. (2007). Open data licensing: Is your data safe? [Web log message]. Retrieved June 27, 2011, from http://digitalcuration.blogspot.com/2007/07/open-data-licensing-is-your-data-safe.html.

Schaffner, A. (1994). The future of scientific journals: Lessons from the past. Information Technology and Libraries, 13(4), 239-247.

Schriger, D.L., Sinh, R., Schroter, S., Py, L. & Altman, D.G. (2006). From submission to publication: A retrospective review of the tables and figures in a chort of randomized controlled trials submitted to the British Medical Journal. Annals of Emergency Medicine, 48(6), 750-756. doi:10.1016/j.annemergmed.2006.06.017

Simmhan, Y.L., Plale, B. & Gannnon, D. (2005). A survey of data provenance in e-science. SIGMOD Record, 34(3), 31-36. Retrieved June 27, 2011, from http://www.sigmod.org/publications/sigmod-record/0509/p31-special-sw-section-5.pdf.

Van De Sompel, H., Payette, S., Erickson, J., Lagoze, C., Warner, S. (2004). Rethinking scholarly communication. D-Lib Magazine, 10(9). doi:10.1045/september2004-vandesompel

Van De Sompel, H., et al. (2006). An interoperable fabric for scholarly value chains. D-Lib Magazine, 12(10). doi:10.1045/october2006-vandesompel

Waelde, C. & McGinley, M. (2005). Public domain; public interest; public funding: Focussing on the 'three Ps' in Scientific Research. SCRIPTed, 2(1), 71-97. doi:10.2966/scrip.020105.71

Wang, R.Y. & Strong, D.M. (1996). Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems, 12(4), 5-33.

Willinsky, J. (2006). The access principle: The case for open access to research and scholarship. MIT Press: Cambridge, MA.

Wilson, H.D. (2001). Informatics: New media and paths of data flow. Taxon, 50(2, Golden Jubilee Pt. 4), 381-387. Retrieved July 26, 2011, from http://www.jstor.org/stable/1223887.