Data publication

B.N. Lawrence, C.M. Jones, B.M. Matthews, S.J. Pepler.

NCAS British Atmospheric Data Centre, Rutherford Appleton Laboratory

This paper presents a discussion of the issues associated with formally publishing data in academia. We begin by motivating the reasons why formal publication of data is necessary, which range from simple fact that it is possible, to the changing disciplinary requirements. We then discuss the meaning of publication and peer review in the context of data, provide a detailed description of the activities one expects to see in the peer review of data, and present a simple taxonomy of data publication methodologies. Finally, we introduce the issues of dataset granularity, transience and semantics in the context of a discussion of how to cite datasets, before presenting recommended citation syntax.

1. Introduction

It has always been the case that the scientific method is predicated on the collection and analysis of data, yet the traditional metric of scientific merit is not the quality of the data itself, but of the discussion and analysis as presented in the peer-reviewed literature. The prevalence of this metric results in scientific communities which do not always analyse the merit or otherwise of the data itself, do not always rate highly the effort required to produce the data, and who cannot always reproduce their analysis because the basic foundation (of data) is not well enough described (or preserved) to support such reuse. Thus, for those communities, the basic thesis of science: that "the experiment" can be repeated is only true where an experiment can begin ab initio under controlled conditions, with no dependencies on external data.

The situation is compounded by the fact that the definition of data has become more blurred in recent years, following the advent of remote sensing and computerised combination of "measurements" and "models" of what is being measured. Thus dependencies on external data extend not only to observations that might have been collected about the physical properties of some specimen (of case study, flora, fauna, limb, organ, molecule or whatever) or specimens, but also observations which have been collected via remote sensing (in which case the details of the remote sensing algorithm are crucial to the interpretation) or which have been simulated (in which case the details of the simulation algorithms, initial conditions etc are crucial).

In general, all scientific communities do understand the importance of the observation provenance (whether simulated, sensed, or collected), it is just that 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. One of the reasons for this is as indicated above: the standard metrics of scientific merit do not recognise the importance of the work done to prepare data: even in cases of best practice (such as genomics, where details of gene-sequences must be deposited to accompany journal articles) the data itself is not subject to review and assessment. In general, it is the conclusions of the paper, and the analysis carried out in the paper which are assessed, the fitness of purpose of the data, any internal analysis, and it's availability or otherwise for re-interpretation are not assessed.

Of course, there are exceptions to this general state. There are papers where aspects of algorithms and experimental method are discussed, there are communities striving to record provenance, and 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. However, such papers almost never describe data in enough detail for the dataset to be recreated (in the case of simulations), or for a re-retrieval from raw data to performed (in the case of remotely sensed data), using the same methodologies. In part, this is because the "details" of such activities are normally deemed to be too voluminous and/or too technical for recording in a journal. (Often these omissions are explicitly seen as analogous to the "working" between mathematical equations to be left "as an exercise for the reader", even though all concerned realise there is a fundamental distinction between algorithms which may not be completely equivalent and mathematical manipulations which result in completely equivalent results). Nonetheless, the need for the primacy of the data itself alongside journal publications is becoming apparent both from social and funding policy perspectives (e.g. Arzberger, et.al., 2004; Klump, et.al., 2006) and from internal discipline requirements (e.g. Carr, et.al., 1997, Hancock, et.al.,2002, Brown, 2003).

Both the policy and scientific requirements are to publish data and make it available for re-use both within the original disciplines and to wider communities. In practice, despite examples of "table and figure data analysis" (e.g. Schriger et al, 2006), the requirement is for the publication of actual digital data on the Internet (as opposed to figures and tables presented within documents). There is an element of technology push to this requirement: like the establishment of journals - to some extent driven by the advent of the printing press and the advent of an efficient postal system (Schaffner, 1994) - the advent and requirements for internet data publication are catalysed by the fact that it is possible: both authors and readers are empowered to produce and consume digital data (Roosendaal and Geurts, 1997). However, like the development of the journal, it probably isn't the technology push that is the dominant factor, it is the changing nature of scholarly research itself (Van de Sompel, et.al., 2004) that matters. (Just as was the

case for the development of the journal; Schaffner op cit judges that the development of the experimental method - which required communicating small discrete units of information- was itself a key driver).

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, proceed to a discussion of the procedures necessary to validate the quality of published data (as opposed to the procedures necessary to measure the uncertainty or error in the data itself), delineate some of the ways 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.

2. Data Publishing

The Oxford English Dictionary (OED Online) now (in the latest draft) defines the verb publish as "to make public or generally known", with, for our context the following explications: "To make generally accessible or available for acceptance or use (a work of art, information, etc.); to present to or before the public; spec. to make public (news, research findings, etc.) through the medium of print or the Internet." and "to undergo the process of publication". However, as far as the Internet is concerned, customary usage doesn't *require* a process before publication – the expectation (and reality) is that one can publish anything by making it available for download. Clearly, in the scientific context, we are very much interested in defining a process before publication, analogous to peer review, and we discuss this in detail below. For now, we simply state the publication process should enable us to make assertions about the trustworthiness and fitness for specific purposes of the data, and that the existence and nature of those assertions should be understood by the data consumers.

"Data" is a very ambiguous term and can mean more or less anything, and it is precisely this ambiguity that is of interest. Data publication poses the question - how do I publish a thing regardless of what it is? In the context of "internet publication", customary usage[2] expects that when a URL is de-referenced, the target can be directly rendered. That is, the immediate usage of the material is analogue: one way or

[1] We do not consider specifically 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 as 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!

another (whether via a printout or not), the material is aimed to be directly consumed by humans! By contrast, when one considers data, we should expect that the initial consumption of the material may be as digital binary objects. In many cases, volume and/or complexity mean the material is made available in binary (in some application specific format), and the target consumer is computer software, with human interpretation not expected without the intervention of (potentially) multiple layers of software. Once we admit to digital data, then there are issues to address about what it represents. We discuss this below in the section on definition and citation.

Finally, Internet publication also introduces issues associated with permanence. 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: (1) how do I find the material again (the "identifier" problem), (2) will the material identified have been moved, changed or removed, and (3) will I still have software capable of interpreting the object when I get back to it? In this paper we do not address the first two of these issues in detail, but we do discuss the third in the section on definition and citation.

In general then, for our purposes, we define to Publish (with a capital P) data, as to make (as permanently as possible) data available on the Internet that has been through a process which means it can appear along with easily digestible information as to its trustworthiness, reliability, format and content.

2.1 Data Publication Procedures

We have asserted that data Publication should consist of a procedure which results in the ability to make assertions about trustworthiness and fitness for purpose. In essence we are trying to define a process directly analogous to that which occurs in the existing scholarly communication process: Schaffner (op.cit) point out that the role of journals is related to the qualitative differences between formal and informal communications, what we would describe in Internet terms as the difference between publication (with a small p, i.e. "putting the data up on the web") and Publication (as defined earlier).

Van de Sompel (2004) identified five functions that process 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?

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 not like that, boundaries between data sets are often blurred and overlapping, the structure varies enormously, and the consumer needs considerable information in order to make sense of (or even "read") the data. We return to this theme later. The six functions can be grouped into two simpler categories:

- *Aids to reusability*: Things which 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" (a phrase coined in Willinsky, 2006).

Thus far, these functions are noncommittal about the extent, or even existence, of quality control, and this is a reflection of the range of methods used and the looseness of the term "scholarly communication". In practical applications in academia, the certification function is also known as "peer review": by passing the peer review process of Journal A, a paper published therein has reached a level of certification as to the quality (and possibly potential impact) of the material. Of course peer review itself is poorly defined: from editorial scrutiny to independent analysis, from the number of reviews to how they are used, every journal has a process that differs from another qualitatively and quantitatively. Nonetheless, the entire academic community understands the concept and benefits of peer review: as Armstrong (1997) puts it:

> "… most successful researchers find the current system of journal peer review to be effective. Journal peer review is commonly believed to reduce the number of errors in published work, to serve readers as a signal of quality, and to provide a fair way to allocate journal space."

(Armstrong has statistics that back up this assertion).

What is needed then is community acceptance of peer review procedures to enable data Publication. However, a priori not all communities may accept peer review of data, regardless of method. Brown

(2003) summarized the arguments against 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 situation should be entirely analogous to the situation with preprints – which enable sharing, and improve the impact - Hitchcock, 2007 - of formal *validated* published articles at the same time).

Exactly what data review procedures should include is the key point that we expand upon here. Data publication procedures have traditionally concentrated on the preservation and long-term access issue, with an emphasis on associative metadata. The institutions attempting this procedure often have policies that require such metadata, but in practice have methodologies that can err anywhere on a spectrum from obtaining near meaningless free text entries in defined categories to obtaining far too limited a subset of information from strongly controlled vocabularies. It is rare for such metadata to receive 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 likely data they might be asked to publish. The importance of such 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."

Leaving aside the definition of "most", the assertion that data is never looked at again is not true for environmental data: re-use is expected and crucial. Thus, from our perspective, the situation should be the same for data production as it is for paper readership. Current readers (data consumers) are desired, but future, or latent, readers may also be important. Schaffner quotes Johannes Kepler on this topic:

> "I am writing a book either for my contemporaries or for posterity. It is all the same to me. It may wait for a hundred years for a reader, since God has also waited six thousand years for a witness"

Of course this future expectation is directly related to the "Archive" function, and so the peer review process is not only asserting statements about quality, it should also address the ability of the information to be curated in an archive for the long term. This is an extra requirement for data publication: the default situation for document publishing is that the ability for libraries to persist (multiple) discoverable copies requires little a priori action by the publishers (beyond using quality paper and ink)! By contrast, in data publication, the concepts and requirements are significantly more onerous. In particular, we need to distinguish between preserving the bits and bytes (archival) and preserving and migrating the information between formats and user-services (curation). Where data formats are not common and/or the services required by the user communities are complex, the job of ensuring that both 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.

In terms of the metadata itself, what material is needed? In an analysis of the expectations of data users, Wang and Strong (1996) found that the beyond intrinsic data quality, but of equal importance, the quality of the information which allowed users to identify the applicability of the data to their task was important, 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 contextural 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 (e.g. see Simmhan et al., 2005, for a provenance taxonomy). Here, we define provenance metadata more widely to include not only automatically generated material but as much as is feasible of 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 resusable.

Lawrence et.al. (2008) have introduced a taxonomy of metadata, which defines

- A-Archive metadata as being 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 as being the material needed to put the archive contents in their scientific context, so this will include provenance metadata etc,

- C-Character metadata includes all the assertions about the importance of the material, including subsequent citations and annotations, and
- D-Discovery metadata, which is a subset of information useable to find the data in search engines and other online discovery services and catalogues.

The data refereeing procedure must ensure that all such metadata is as complete as possible, but it must also address other qualities expected of Publication class material. What of the quality of the data in terms of its 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 unlike journal articles, much 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 – for measurements of the earth atmosphere one has different realistic maximum and minimum expectations than one has for solar atmospheric measurements. One can also make additional requirements of data: for example, explicit assessments of measurement (or simulation) uncertainty can be required.

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 it's own right. This is more likely where there is some chance of future legal challenge to conclusions that might be drawn from the data.

The importance and impact of data is also difficult to address, given that often individually unimportant data measurements can gain value from being aggregated, and 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" (e.g. Wilson, 2001, and our own experiences with the requirements of our funding body: the Natural Environment Research Council). However, like Kepler, many data producers (and hence data journal value assessors) need to have a view of the far distant future. Clearly observations of time-varying real world phenomena are the most important class of data, but even simulations and data which results from (potentially repeatable) analysis may be worthy of publication; even if only so that repeatability can be

verified! Again, in practice, there needs to be a value judgment applied within the peer review process, and the value will depend on the target readership of the "data Journal".

We expect, that like the traditional journal world, data publishers will appear providing publications that are recognized to have a range of subject matter, quality and impact.

Returning to the issue of metadata, and the difficulty in obtaining it, along with the laborious nature of collecting it one of the reasons asserted by some scientists for not doing the work is that the information required will appear anyway 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, papers are generally geared towards persuasion by constructing an argument by narrative: the data is new, the method is important, the result is interesting, and it is all based on facts which are not always explicitly identified and fully qualified within the paper. (There is a whole body of research developing methods for identifying the structure of the narrative arguments which appear within papers, see for example, de Waard, 2007). 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 2007-08-01 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 2007").

There is one final issue before we leave data publication: 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 law being applicable in the UK and the US (Rusbridge, 2007). In what follows, we neglect issues of ownership given the uncertainties involved. We also note that even when ownership issues are clear, once we introduce the concept of publication, other legal frameworks become relevant, for example, in the UK, there is the possibility that the data might become subject to the Legal Deposit Libraries Act 2003, which may have ramifications for the issues of persistence and curation discussed above.

2.2 A practical guide for peer review of data.

Thus far we have motivated the importance of peer review of data, and covered some of the things it must cover. 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.

2.2.1 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.

2.2.2 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[3], there are a number of characteristics of the PDS that have generic interest and are worth summarising here:

[3]While quite 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…

- 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. (op cit) 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.

2.2.3 Summary: 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, based on the discussion presented earlier. 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).

Data Quality

1. 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?
2. Are data values internally consistent?
3. Does the data represent reality with sufficient accuracy to use? Is the data of tolerable precision?

4. 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.)
5. Are the data values reported physically possible and plausible?
6. Is the data validated against an independent dataset? (Has it been calibrated?)

Metadata Quality

7. Is there sufficient quality metadata describing the format and physical content? (See for example, the requirements of the PDS.)
8. Is there sufficient quality metadata describing provenance and context? Has the data changed in someway since it was measured? Is the processing chain visible and well documented? (See for example, questions from section 2.2.1.)
9. 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?
10. Is there suitable quality discovery metadata? At a bare minimum, can Dublin Core be constructed?
11. Does the metadata use appropriate controlled vocabularies?
12. 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?
13. Is all the available metadata conforming to standards?

General

14. Is there an existing user community? Is that community happy that the data is usable?
15. What is the track record of the data source? Is it/her/him/they reliable?
16. Are the intellectual property rights for the data established?
17. Is the data available at the correct network address?

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

18. 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.

3. Data publication models

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:

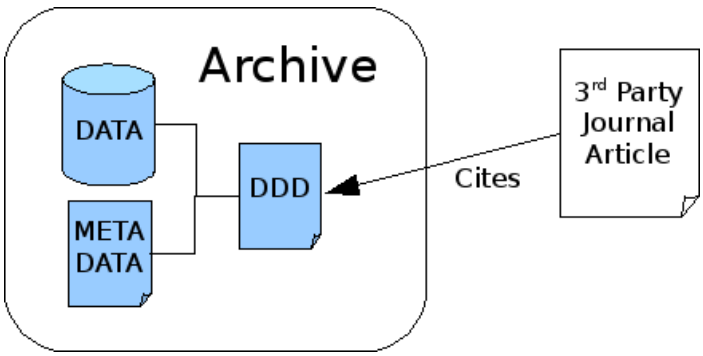
1. Author: data creator, normally required to meet the initial data format specifications of the curator.
2. Resolver: maintains a document that includes links to the data itself and any metadata, and which is the primary citable entity.
3. Identifier Manager: Controls how identifiers are distributed between data entities, archives, resolvers, and documents.
4. Review Controller: Controls the peer review process (if any).

5. 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)
6. Metadata Editor: Carries out editorial functions (including assembly and definition) for the dataset metadata.
7. Metadata Creator: the author of documents that describe the data.
8. Reviewer: assesses fitness of the data against publishers predefined and/or community accepted review criteria
9. Archiver: responsible for the persistent storage of the datasets.
10. 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.

3.1. Stand alone data publication



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 provide 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 more rare: examples include the Planetary Data System (http://pds.jpl.nasa.gov/), and the putative Earth System Atlas (http://earthsystematlas.sr.unh.edu/).

Role	***Archive***	***Journal***	***Author***	***3rd Parties***
Author			B	
Resolver	B	←		
Identifier Manager	B	←		
Review Controller	B,P	←		
Gatekeeper	B	←		
Metadata Editor	B	←		
Metadata Creator	Some	←	Some	
Reviewer	B			P
Archiver	B			
Curator	B			

Table 1: the distribution of roles in stand alone data publication. Basic functionality for publication is denoted with a B, extra activities to promote Publication are denoted with P (note that the primary metadata creation should be carried out by the author where Publication is intended). In this case, the archive subsumes many of the traditional "publisher" roles (implied by the arrows in the journal column).

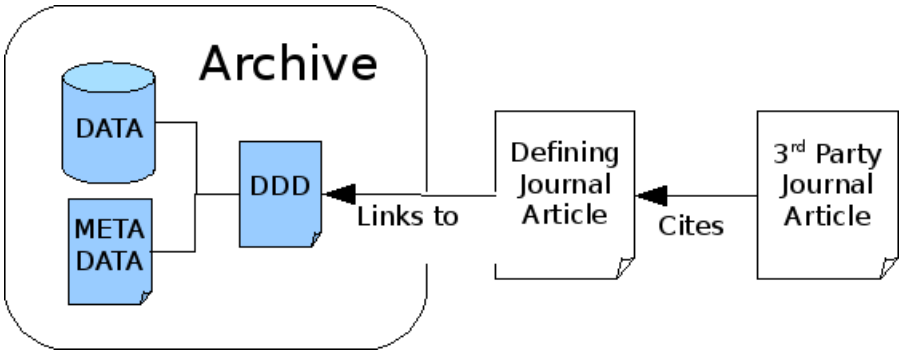
In this case, a third party citing the data will use an identifier to the DDD provided by the archive, and use that directly. In most cases, academic journal will not allow such identifiers to appear in the formal reference list.

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, but 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, one of the requirements of peer review is that it should be completely independent of the data submitters[4].

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

3.2. 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. The paper generally describes the project and aspects of the algorithms and data, but is generally 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 etc of the data and its documentation. Nearly all journals accept papers of this sort.

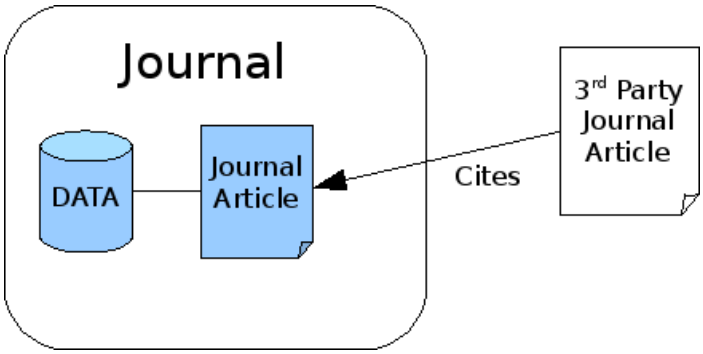
In most cases, the author of the dataset and the author of the proxy paper is the same. The paper often refers to quality procedures involved with the data production, but this is often shoehorned into a paper

[4] 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).

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 centres 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.

3.3. 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[5] 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	***Archive***	***Journal***	***Author***	***3rd Parties***
Author			P	
Resolver		P		

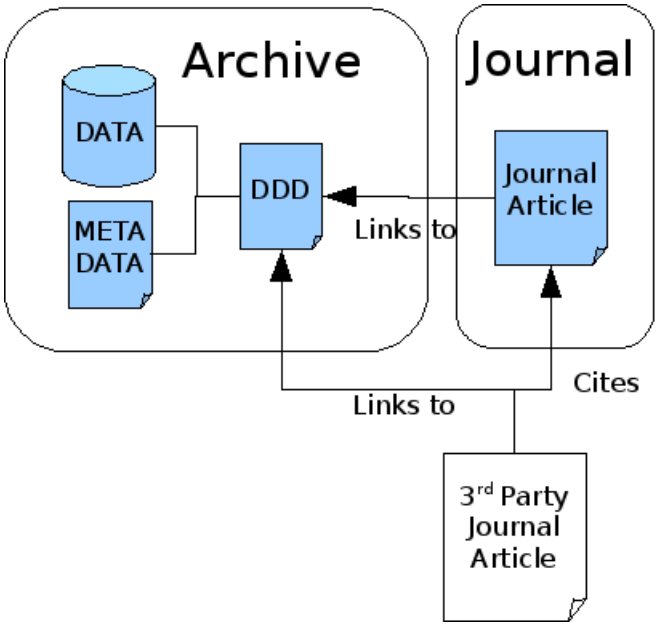
[5] For biological data, Nature operates the journal driven data archive model as well. These two models co-exist peacefully.

Identifier Manager		P		
Review Controller		P		
Gatekeeper		P		
Metadata Editor				
Metadata Creator				
Reviewer				Some
Archiver	→	P		
Curator	→	Some		

Table 2: 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 consumed by the journal (indicated by the harrows), 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.

3.4. Journal driven data 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[6] Genetics Journal (http://genetics.plosjournals.org) 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 [7] 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."

[6] The Public Library of Science (PloS), http://www.plos.org

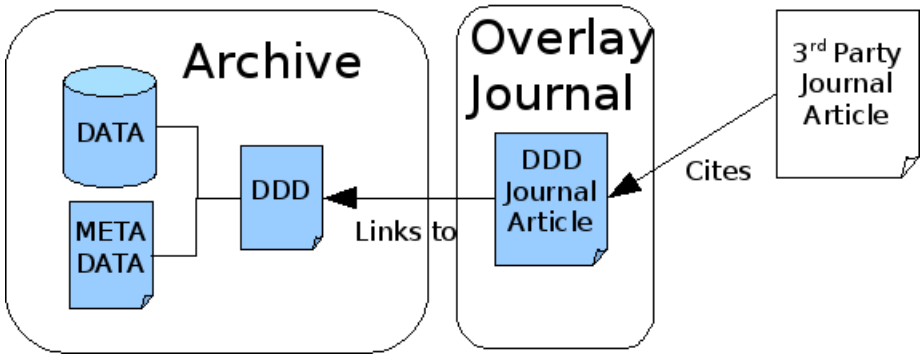
[7] Americal Geophysical Union Data Policy: http://www.agu.org/pubs/data_policy.html

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 preexisting 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.

In most cases, paper authors will be data authors, and as such are responsible for submitting both 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 (accession number in the case of the biosciences) is generated. The data review process is again the same as the standalone 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.

3.5. 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	***Archive***	***Journal***	***Author***	***3rd Parties***
Author			B,P	
Resolver	B	P		
Identifier Manager	B	P		
Review Controller	B	P		

Gatekeeper	B	P		
Metadata Editor		P		
Metadata Creator	Some		Some	
Reviewer	B			P
Archiver	B			
Curator	B			

Table 3: The distribution of roles in Overlay Data Publication. The basic functionality to present the data on the Internet is denoted with B, the additional functionality for Publication is denoted with P.

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 standalone 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).

3.6 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 standalone 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 standalone 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 standalone 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. The latter provides a significant advantage over standalone data publication, however, 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. 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 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 automatic or carried out by archival staff), leaving the semantic and completeness checks for the review process to be carried out in the overlay review process. This is 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.

4.0 Citing Published Data

Even in a community where the dependency on data archival is required (the bioinformatics genebank 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 (a to 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 known "by the grapevine", 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.

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.

No one we have consulted can see any reason for requiring the physical location of the data centre in an Internet based data reference!

As part of the project funding this work, we have both canvassed active scientists about what information would be necessary in a reference, and 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, and an unambiguous reference to the source of the data. The practising 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, what comprises the dataset level: the facility or a particular instrument? 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. (op cit), 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, and then discuss existing best practice in citation before introducing our recommended syntax.

4.1 Issues of Transience 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. In the case of Internet publication, an identifier may refer to a resource which has changed. In the short history of internet citation, this has been dealt with by appending the common syntax of, for example, "accessed on 31/12/2007" 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.

In the most part, issues of transience should be dealt with in 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 introduced 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?

4.2 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 website, 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, gene sequences as tabulated codes etc. 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 standard which provide context for describing things in the real world: ISO19101 (2002) introduces the notion of a "feature" 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 (etc collections of gene sequences, aggregations of remote soundings from radars etc), 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).

4.3 Existing Citation Formats

Examples of existing best practice include:

1. The PDS citation format[8], 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.

2. The German project "Publication and Citation of Scientific Primary Data" (Klump et.al., op.cit., Brase and 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).

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

[8] Policy for Citations of PDS Data, http://pds-geosciences.wustl.edu/citations.html, accessed 31st December 2007.

semantics: what is being referenced, and without the concept of features, there is a rather clumsy methodology for providing length:

> "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!

4.4 Recommended Citation Syntax

Despite the problems and limitations of Patrias (op cit) 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 considering these elements in turn, we 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, 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 carrying out the review process, which may not be the same as the organisation hosting the data itself.

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 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 an 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

(i) We are dealing with published data. We can remove the citation date.

(ii) We have introduced a URN and an optional feature indentifier.

(iii) We are always using URLs or DOIs that indicate we've got Internet media. We lose the [Internet] designator.

(iv) 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.

1. 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.

2. 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 CSIRO is a corporate author, rather than the publisher, which is the World Data Centre (who organised the review), (ii) the type of the data is now clear and that the feature type definition doesn't have to be owned by the publisher!

3. 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.

4. 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 etc.

5. 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.

5.0 Summary

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 standalone 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.

The authors acknowledge the support of both the Joint Information Systems Committee and the Natural Environment Research Council in carrying out this work.

References

Armstrong, J.S. (1997). Peer Review for Journals: Evidence on Quality Control, Fairness, and Innovation. Science and Engineering Ethics, 3, 63-84.

Arzberger, P., P. Schroeder, A. Beaulieu, G. Bowker, K. Casey, L. Laaksonen, D. Moorman, P. Uhlir, and P. Wouters (2004). Promoting Access to Public Research Data for Scientific, Economic, and Social Development. Data Science Journal, 3, 135

Brown, C. (2003). The Changing Face of Scientific Discourse : Analysis of Genomic and Proteomic Database Usage and Acceptance. J. Am.Soc.Inf.Sci.Tech.,54, 926-938.

Brase, J. and U. Schindler (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.

Carr, T.R., R.C. Buchanan, D. AdkinsHeljeson, T.D. Mettilee, and J. Sorenson. (1997) The future of scientific communication in the earth sciences : The impact of the Internet. Computers and Geosciences, 23, 503.

De Waard (2007). A Pragmatic Structure for Research Articles, in Buckingham Shum, S., Lind, M. and Weigand, H. (2007), (Eds). Proceedings ICPW'07: 2nd International Conference on the Pragmatic Web, 22-23 Oct. 2007, Tilburg: NL. ISBN 1-59593-859-1 & 978-1-59593-859-6. Archived in: ACM Digital Library & Open University ePrint: 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 Tromso. http://www.enger.priv.no/files/2005/master.pdf. Accessed December, 21, 2007.

Ginsparg, P. (1996). "Winners and Losers in the Global Research Village" http://xxx.lanl.gov/blurb/pg96unesco.html. Accessed: 2007-12-21.

Gray, J., A.S. Szalay, A.R. Thakar, C. Stoughton, J.v.d.Berg (2002). Online Scientific Data Curation, Publication and Archiving. Microsoft Research Publications, MSR-TR-2002-74.

Hancock, W.S., S.L. Wu, R.R. Stanley, E.A. Gombocz (2002): Publishing large proteome datasets: scientific policy meets emerging technologies. Trends in Biotechnology, 20, S39-S44.

Hitchock, Steve, 2007. The effect of open acces and downloads ("hits") on citation impact: A bibliography of studies. Open Citation Project. http://opcit.eprints.org/oaicitation-biblio.html (accessed December 21, 2007).

Hurd, J., C.M. Brown, J.Bartlett, P.Krietz, G. Paris (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, 452

ISO19101 (2002) Geographic information — Reference Model

ISO19110 (2005) Geographic information — Methodology for feature cataloguing

ISO19136 (2005). Geographic information – Geography Markup Language

Klump, J., R. Bertelmann, J. Brase, M. Diepenbroek, H. Grobe, H. Hock, M. Lautenschalger, U. Schindler, I. Sens and J. Wachter. (2006). Data Publication in the Open Access Initiative. Data Science Journal, 5, 79.

Koningsberger, D.C., Report on activities of Committee on Standards and criteria in XAFS Spectroscopy., Jpn. J. Appl. Phys., 32, 877-88.

Lawrence et.al. (2008). Information modelling and DataGrids, Phil Trans (to be submitted April 2008). Full reference to be supplied with page proofs.

OED Online (2003) , Oxford University Press, Draft entry cited 19th December, 2007. http://dictionary.oed.com/cgi/entry/50191830

Patrias, K. Citing medicine: the NLM style guide for authors, editors, and publishers [Internet]. 2nd ed. Wendling, Daniel L., technical editor. Bethesda (MD): National Library of Medicine (US); 2007 [Accessed on 31/12/2007]. Available from: http://www.nlm.nih.gov/citingmedicine

Roosendaal , H.E. and P.A.Th.M. Geurts (1997). Forces and functions in scientific communication : an analysis of their interplay, in, CRISP 97, Cooperative Research Information Systems in Physics, M. Karttunen, K. Holmlnund, and E.R. Hilf (ed). Available at http://www.physik.uni-oldenburg.de/conferences/crisp97/.

Rusbridge, C. 2007. Open Data Licensing: is your data safe? [Internet] [Accessed on 31/01/2008]. Available 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, 239-247.

Schriger, D.L., R.Sinh, S. Schroter, L. Py and D.G. Altman. 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, 750-756.

Simmhan, Y.L., and B. Plale and D. Gannnon. A Survey of Data Provenance in e-Science. SIGMOD Record, 34. Online at http://www.sigmod.org/sigmod/record/issues/0509/p31-special-sw-section-5.pdf , accessed December 21, 2007.

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

Van De Sompel, et.al., 2006. An Interoperable Fabric for Scholarly Value Chains., D-Lib, 12, doi:10.1045/october2006-vandesompel

Waelde, C. and M. McGinley, 2005: Public Domain; Public Interest; Public Funding: Focussing on the 'three Ps' in Scientific Research. SCRIPT-ed, 2, 71-97.

Wang, R.Y and D.M. Strong (1996). Beyond accuracy: What data quality means to data consumers. J. Man. Inf.Sys., 12., 5

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

Wilson (2001). Informatics; new media and paths of data flow. Taxon, 50, 381-387.