Data Publication: outputs of the CLADDIER project

CM Jones [1], KA Bouton[2], JMN Hey [3], SE Latham [4] , BN Lawrence [4], BM Matthews[1], AJ Miles[1], S Pepler[4], K Portwin[1]

[1]e-Science Centre, Science an d Technology Facilities Council (STFC)
Rutherford Appleton Laboratory
HSIC, Didcot, OXON OX11 OQX, UK
Email: c.m.jones@rl.ac.uk

[2] National Centre for Atmospheric Science - Climate, University of Reading
Earley Gate, PO Box 243, Reading RG6 6BB, UK
Email: k.a.bouton@reading.ac.uk

[3] School of Electronics and Computer Science, University of Southampton
Highfield, Southampton, SO17 1BJ, UK
Email: jmnh@ecs.soton.ac.uk

[4] NCAS British Atmospheric Data Centre
Rutherford Appleton Laboratory, HSIC, Didcot, OXON OX11 0QW, UK
Email: s.e.latham@rl.ac.uk

ABSTRACT

Institutional publication repositories are becoming an established part of research communication, giving an opportunity to explore their relationship with the underlying data. The JISC funded Citation, Location and Deposition in Discipline & Institutional Repositories (CLADDIER) project in the UK has been investigating the issue of linking publications held in institutional repositories to the underlying data held in specialist repositories, such as Natural Environment Research Council (NERC) data centres, by developing the theme of citations, not only for publications but also for datasets. This paper will look at the key management issues identified in the project, for data repositories to be able to provide datasets that can be cited over a long term. It discusses the citation capture issues for publication repositories and discuss potential methods of disseminating these links between repositories.

Keywords: Data, publications, citation, institutional repositories

INTRODUCTION

The Joint Information Systems Committee (JISC) funded Citation, Location and Deposition in Discipline & Institutional Repositories (CLADDIER) project in the UK has been investigating the issue of linking publications held in institutional repositories to the underlying data held in specialist repositories, such as Natural Environment Research Council (NERC) data centres, by developing the theme of citations, not only for publications but also for datasets. The project partners are the University of Reading, the University of Southampton, the Science and Technology Facilities Council (STFC) (previously known as the Council for the Central Laboratory of the Research Councils - CCLRC) and the British Atmospheric Data Centre (BADC). The project is using the BADC as the test bed for data, the institutional repositories at STFC and University of Southampton for publications, and scientists at the University of Reading are testing the initial implementation.

The BADC acts as a primary archive for NERC funded atmospheric science datasets and additionally holds other subject related datasets so that they can be accessed more easily by researchers. These datasets hold a variety of different types of data, with the common theme that they are the results of experimental research programmes. In this domain, active researchers require access to both the written scientific record and the data to be able to explore further the topic under examination. At present the links between the data and the written record are fragmented. In particular, references to the data used will be within the text or acknowledgments in a non-standard format. This makes them hard to follow to find the original data source.

At the project's inception, CLADDIER created a use case to demonstrate the need to link the primary data to the publications produced and what benefits such links might bring. This is as follows:

> *A scientist, based at the University of Southampton, is researching the biology of seawater off the Cornwall coast. As part of her analysis she needs: publications and data on prior or similar work and data in the areas of ocean profiles, meteorology and remotely sensed ocean colour imagery. Once her work is complete she publishes a paper, citing the publications and datasets used and lodges her own publication and datasets in appropriate repositories.*
>
> *This new work is of interest to a scientist based at University of Reading and he will be able to find both the publication and the datasets used through the mutual citations held in the data and publication repositories.*

This use case is based on the fact that there will be, within this subject domain, a common understanding of the information needed to cite data (and hence a common method) so that the citation is usable by other domain experts. One of the main strands of the CLADDIER project has been to investigate this requirement and to suggest potential solutions.

DATA PUBLICATION AND CITATION

What does publishing data entail?

Data publication is not a new issue within the data centre community. There are other projects addressing this issue, such as the German STD-DOI project [1] which is working towards data publication and citation. CLADDIER aimed to take an atmospheric sciences point of view.

The term "publishing" has a standard meaning: ***1.*** *to prepare, produce and distribute (printed material, computer software, etc) for sale to the public.* [2] For academic publishing it has many implicit meanings, including the concepts of longevity of access and some form of quality control. Both the formal definition of publishing and the additional concepts are relevant to data publication. Whilst publication for written works, such as journal articles, is a well-understood and long term practice, it is still under development for data. The term "data" does not describe a homogenous thing in the same way that the term "journal article" does. Data can come in many shapes and sizes, produced by humans or automatically. Much of the data used in the Atmospheric Science discipline is collected over years and is constantly being added to and revised, drawing some of the issues around data publication into sharp focus. In CLADDIER we have considered three components: (1) citation requirements and styles, (2) making objects permanently available and (3) quality control processes.

The act of citation implies that the cited item will be retrievable by the reader using the information provided. In the textual domain there are many standard formats which provide this functionality. There is also the implication, especially for printed material, that the referred-to information will be retrievable in the form that the author provided it. This is becoming less true as the trend for citing less well-preserved material such as web sites increases. The act of making information available via a website can be considered to be synonymous with publication, but in this case the additional publication concepts discussed above are rarely considered. However it is important to remember how internal judgements on the relevance and trustworthiness of references impacts on the behaviour of the reader and so it is important for data citations to be as long-lived and reliable as those to printed publications.

User requirements for citing data

One of our project aims was to propose a format for citing the data, as at present atmospheric scientists do not formally cite data in the same way as they cite scholarly publications. As part of the project we investigated user requirements for citing BADC data with practising scientists so that we could develop a citation scheme that they would be willing and able to adopt. Key issues identified 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 which produced the data, and an unambiguous reference to the source of the data.

These 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 can be used for service metrics and paper location. The main concerns were that it should be traceable to the data provider and to be recognised as intellectually equivalent to academic papers.

These different stakeholders were brought together at the CLADDIER workshop, held on 15th May 2007 to discuss these issues further [3]

Citation style

Having identified some of the key needs for citing data, we now construct an example citation using the NERC Mesosphere Stratosphere Troposphere radar facility (MSTRF,[4]) data which is held by the BADC and discuss the resulting issues. This data consists of a time series of wind speed profiles for altitudes between 2 and 80 km. As well as the primary wind data there are numerous other parameters measured by the radar and other instruments co-located with the radar. The data from the radar is processed and then stored at the BADC as a series of data files, and the dataset is updated hourly as more data becomes available. The existing primary discovery metadata record for this data is single entry within the BADC catalogue that provides a description of the MSTRF. Additional metadata are located in the file headers, and in documentation available with the data.

To create a citation for this data we start by following the National Library of Medicine Recommended Formats for Bibliographic Citation [5] for databases and retrieval systems:

> *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.*

Considering these elements in turn:

> *Author*: The author of an incremental data set is hard to identify. Ideally both the principle investigators and the corporate body who provided the means to get the data should be recognised.
>
> *Title*: This should identify the data resource, i.e. the facility name.
>
> *[Content Designator Medium Designator]*: For simplicity “Internet”.

Edition: The data has several versions of processing and levels of product. Additionally there are data from ancillary instruments and all of this is being regularly updated. In this complex picture it is difficult to decide what constitutes an edition in the way envisaged by NLM. For this example we chose "version 2", a processing version number, and "Cartesian", a product level description. As there is no guidance, at present, on what an author should class as the edition they can use what they believe is useful, for instance, "Mesosphere, Spectra Widths" is an equally valid description of the data edition. This potential ambiguity needs to be addressed.

Place of Publication: This has no value for an internet resource so we have chosen to omit it.

Publisher: The organisation responsible for maintaining the primary copy of the data is the BADC, or perhaps NERC. For arguments sake we use the BADC here.

Date of Publication: For this data it is 1990 onwards. The use of onwards indicates that the dataset is being updated. This date is not linked to edition in the same way as it might be in other material types.

[*Date of Update/Revision; Date of Citation*]: As this data is continually updated, the citation date is most useful, but this is not completely unambiguous because a citation user will not be able to know exactly what the dates of the dataset used were, as the citation date may not be the same as the use date.

Availability: A URL from which the data is available.

The following fields are optional and we have not used them in this example: *Extent Series, Language* and *Notes*

The citation for the MSTRF dataset is as follows.

> *Natural Environment Research Council, Mesosphere-Stratosphere-Troposphere Radar Facility [Thomas, L.; Vaughan, G.] . Mesosphere-Stratosphere-Troposphere Radar Facility at Aberystwyth, [Internet]. Version 2, Cartesian products. British Atmospheric Data Centre (BADC), 1990- [cited 2006 Apr 25]. Available from http://badc.nerc.ac.uk/data/mst.*

There are several issues raised by this exercise: how unambiguous is the reference when it is not clear how to establish what was in the dataset at the time of use due to the use of citation date; the edition is defined by the author and the reference does not enable the user to identify what "type of data" it is (unlike textual citations). Finally the role of the data centre as publisher is not fully understood by the data centre, the authors, or even those using (and citing the data). The next sections consider how to resolve some of these issues.

Making data permanently available

This has three aspects to it: defining what is to be kept, ensuring that it is described effectively and identifying who is responsible for managing it. Of the three, the latter two are more resolvable by using standards such as Climate Science Modelling Language (CSML) [6] and Geography Mark-up Language [7] to describe and encapsulate the data within the storage system and the trusted repository audit guidelines to ensure that the data centre is well-managed [8]. Defining what is to be kept is inextricably linked with data publication and decisions on granularity. It may be decided to keep an entire dataset as an entity for simplicity; however this poses problems for datasets which are still growing and additionally it is unlikely that a scientist will use the whole of the dataset so that citing the entire dataset will not aid another scientist to easily re-evaluate the analysis using the same data. On the other hand it is not technically and administratively feasible to keep each sub-set of data generated by a scientist as an unique version. Some mid-point between these approaches needs to be identified. In addition to the issues surrounding granularity there are the issues of versions of the dataset generated due to reprocessing or repackaging and format translations for preservation processes.

Considering our citation example, the granularity is set by “version 2, Cartesian products” and the access date. However this is not an explicit description of the data used, or which data files make up the data.

Peer review

An important part of the publication process is quality control. For publications in academic journals this is done through a peer-review procedure for the intellectual content and copy-editing for style and correct use of language for that journal title. Although data centres apply a form of quality control when acquiring the datasets, this is likely to be simply at the usability level to ensure that the data can both be reused by others and preserved effectively. This is an important role, but if data publication is to be comparable to written work publication then there needs to be an independent peer-review process.

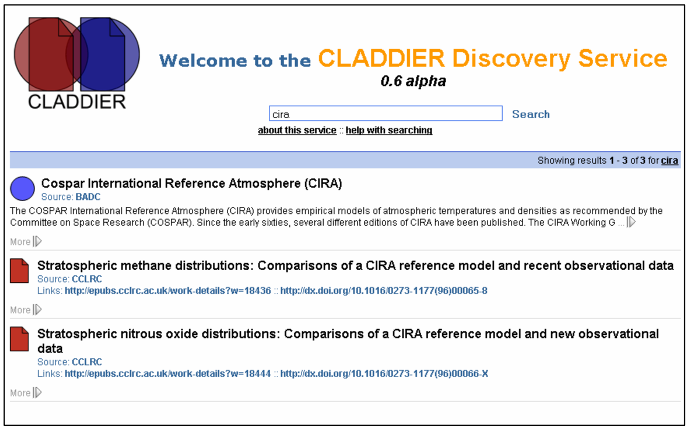
There are two potential approaches to achieving peer-review: linking it to ingest in a data centre or using a traditional publishing organization. The former is already in use by the NASA Planetary Data System (PDS) [9] where a panel is convened comprising of Data Centre experts, domain experts and a data producer representative who consider whether the data are complete, suitable for archiving and that the PDS standards have been followed. This approach could be adopted by the BADC but would introduce delays into the ingest process. An alternative is to consider the latter, and for the BADC subject domain an obvious traditional publisher is the Royal Metrological Society. The data centres and providers could submit the data to a “data journal” which would carry out the quality control whilst allowing the data to reside in an appropriate location. If this were to be implemented then our citation would become:

> *Natural Environment Research Council, Mesosphere-Stratosphere-Troposphere Radar Facility [Thomas, L.; Vaughan, G.] . Mesosphere-Stratosphere-Troposphere Radar Facility at Aberystwyth, [Internet]. Version 2, Cartesian products. RMS Data Publications, 1990- [cited 2006 Apr 25]. Available from http://badc.nerc.ac.uk/data/mst. [doi:10233/23498234]*

Note the change of publisher and the addition of the publisher’s identifier for this item.

DISCOVERY TOOLS

One of the results of the CLADDIER project is a simple resource discovery tool which uses OAI-PMH [10] to harvest metadata records to enable federated searching across both publication and data repositories to aid the location of information and cross-linked material. This prototype allows scientists to search for data held in the BADC, University of Southampton and STFC repositories. The screen shot in Figure 1 shows the results of a search for “CIRA”. It has found a data set held at the BADC and two publications held within the STFC (CCLRC) institutional repository, as indicated by the icons identifying the type of material located. From the metadata records discovered the user can navigate to items of interest.



Welcome to the CLADDIER Discovery Service

0.6 alpha

cira Search

about this service :: help with searching

Showing results 1 - 3 of 3 for cira

Cospar International Reference Atmosphere (CIRA)
Source: BADC
The COSPAR International Reference Atmosphere (CIRA) provides empirical models of atmospheric temperatures and densities as recommended by the Committee on Space Research (COSPAR). Since the early sixties, several different editions of CIRA have been published. The CIRA Working G ...
More

Stratospheric methane distributions: Comparisons of a CIRA reference model and recent observational data
Source: CCLRC
Links: http://epubs.cclrc.ac.uk/work-details?w=18436 :: http://dx.doi.org/10.1016/0273-1177(96)00065-8
More

Stratospheric nitrous oxide distributions: Comparisons of a CIRA reference model and new observational data
Source: CCLRC
Links: http://epubs.cclrc.ac.uk/work-details?w=18444 :: http://dx.doi.org/10.1016/0273-1177(96)00066-X
More

Figure 1: Screen shot of the discovery service

For this approach to be effective all the repositories need to be able to express their metadata in Dublin core and to be able to provide an OAI-PMH set. For dataset providers the concept of a published set which can be described effectively through metadata is key.

This approach whilst providing cross-searching functionality does reduce the precision of the search due to the quality of the metadata offered through OAI-PMH. One way to improve this would be to use Dublin Core metadata profiles, such as the ePrints [11] for scholarly works; however as yet there is no profile developed for scientific data which could be applied to the BADC metadata records.

CAPTURING CITATIONS

To be able to navigate from data to textual work and vice-versa the citations must be within the metadata records, in the relevant repositories, in a form that can be used for this purpose. At present only textual works have citations and they are handled in a naïve fashion within repositories: the University of Southampton Research Repository sometimes includes a list of citations and STFC ePubs did not hold citations. One of the considerations is the tension between what the author is prepared to input and how automatically identifiable the separate parts of the reference are.

For the first iteration we took the pragmatic approach that some information was better than none and for each individual citation there are the reference text, sequence and link direction indicators and the identifiers for the item. There is the possibility that there would be more than one identifier for an item as it might be available from more than one location. This data model is generic enough to be implemented in any of the project repositories and has been done so for the ePubs repository. We are now extending this to give an option for a more expressive structure for references.

THE TRACK-BACK MECHANISM

Having designed and implemented both forward and backward citation capture, the next step was to design a mechanism whereby the repository where a work with a citation was deposited could notify the repository where the cited item was located that this item had been cited by the deposited work. Several notification schemes were considered and a Peer-to-Peer approach was adopted; this uses a well-known protocol used in the blogging community know as "trackback". Trackback is a framework for allowing communication from one resource to inform another that there is a citation and was released as an open specification in August 2002 [12]. Trackback has a two stage protocol using http requests, known as a "trackback ping": once a repository B discovers a citation URI in a resource B1 within the repository B to a resource A1 in repository A, it accesses A1 via a http call. Repository B looks for a "trackback URL" embedded within the page returned by A1; if there is one there repository B calls A's trackback URL with a http Post delivering B1's URI. Repository A can then augment A1's metadata with a "cited by B1" thus completing the cross-reference. We have extended this basic protocol in two ways: one to add a notion of a "whitelist" so that trackbacks are only accepted from known and trusted repositories, and secondly to increase the data carried in the transactions so that richer metadata can be passed.

CONCLUSION

CLADDIER has looked at some complex issues surrounding publishing data, discovering linked data and textual publications and how to make the navigable bi-directional links more effective. It can be observed that it has raised more questions about these issues during the activities undertaken. In particular this paper has highlighted unresolved issues with: the semantics of the data citation are still under discussion to ensure that the resulting reference is both full enough to be of use without being too difficult to understand, the unrefined searches that Dublin core brings, and issues with trackback and trusted repositories.

However, despite these unresolved issues, the CLADDIER project has produced prototype tools which demonstrate the potential for cross-linking and these are currently undergoing user testing by research scientists the University of Reading. These tools and the underlying theoretical work on data publication will be taken forward within the partners to enhance textual/data linkages in other areas of science in support of the STFC facilities.

REFERENCES

[1] – Publication and Citation of Scientific Primary data project http://www.std-doi.de/

[2] - As defined: "publish." *Chambers 21st Century Dictionary*. 2001. Xreferplus. 05 January 2007 http://www.xreferplus.com/entry/1222933

[3] – Lawrence, Bryan, Pepler, Sam, Jones, Catherine, Matthews, Brian, McGarva, Guy and Coles, SimonHey, Jessie (ed.) (2007) Linking data and publications in the environmental sciences: CLADDIER project workshop, Chilworth, Southampton, UK 15th May 2007. Southampton, UK, CLADDIER project http://eprints.soton.ac.uk/46207/

[4] - Hooper, D: The Natural Environment Research Council (NERC) Mesosphere-Stratosphere-Troposphere (MST) Radar at Aberystwyth http://mst.rl.ac.uk/ (2006)

[5] - National Library of Medicine: Recommended Formats for Bibliographic Citation, Supplement: Internet Formats. http://www.nlm.nih.gov/pubs/formats/internet.pdf (2001)

[6] - Woolf, A.,et.al.: Climate Science Modelling Language: standards-based markup for metocean data. Proceedings of 85th meeting of American Meteorological Society http://ams.confex.com/ams/Annual2005/techprogram/paper_86955.htm (2005)

[7] – Geographic Information – Geographic Mark-up Language ISO 19136:2007

[8] - Research Libraries Group: An Audit Checklist for the Certification of Trusted Digital Repositories (2005) available at http://www.rlg.org/en/pdfs/rlgnara-repositorieschecklist.pdf

[9] - National Aeronautics and Space Administration, Planetary Data System: Peer Reviews. http://pds.jpl.nasa.gov/data_services/peer_reviews.html (2006)

[10] – OAI-PMH protocol available from http://www.openarchives.org/OAI/openarchivesprotocol.html

[11] - Eprints Application Profile, (2006), http://www.ukoln.ac.uk/repositories/digirep/index/Eprints_Application_Profile

[12] – Track back technical specification available from http://www.sixapart.com/pronet/docs/trackback_spec

ACKNOWLEDGEMENTS

We would like to thank the JISC Repositories and Preservation Capital Programme and the participants of the Linking data and publications in the environmental sciences: CLADDIER workshop for their support in this project.