Cross-linking and referencing data and publications in CLADDIER

Brian Matthews[1], Katherine Bouton[4], Jessie Hey[3], Catherine Jones[1], Sue Latham[2], Bryan Lawrence[2], Alistair Miles[1], Sam Pepler[2], Katherine Portwin[1]

[1] *e-Science Centre, Science and Technology Facilities Council (STFC)*
Rutherford Appleton Laboratory
Chilton, Didcot, OXON OX11 OQX, UK

[2] *National Centre for Atmospheric Science (NCAS) British Atmospheric Data Centre,*
Rutherford Appleton Laboratory, Chilton, Didcot, OXON OX11 OQX, UK

[3] *School of Electronics and Computer Science, University of Southampton*
Highfield, Southampton, SO17 1BJ, UK

[4] *NCAS Centre for Global Atmospheric Modelling (CGAM), University of Reading*
Earley Gate, PO Box 243, Reading RG6 6BB

Abstract

Institutional 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 NERC data centres, by developing the theme of citations, not only for publications but also for datasets. In this paper, we discuss some of the aspects of this work, including policy for the publication of datasets, and architecture to support cross-citation within repository infrastructure.

1. Introduction

Institutional repositories are becoming an established part of the research communication landscape. Many institutions now maintain them to record, archive and disseminate their research output, typically documents produced for formal publication. Further, there are a number of established repositories which collect, store and distribute research data arising from experiments or observations. Whilst such data could be archived in institutional repositories, they are typically collected into subject based repositories in order to take advantage of discipline specific expertise and to maintain a close relationship with the relevant research community.

The emerging repository infrastructure gives the opportunity to explore the relationship between research publication and research data. From a publication it would be useful for readers to track back to primary data to inspect for themselves the quality of the data and the validity of the conclusions drawn, possibly running analyses of their own, and for authors to track forward to find further experimental studies influenced by the results which they reported. From a dataset, it would be useful to track back to discover the context the data was generated from the publications used to justify its collection, and to track forward to find the resulting publications reporting analysis of the data, not only the primary analysis of the data generator, but also secondary analyses by other users of the dataset.

Some journals now require the deposit of primary research data as supplementary material; this only partly solves the problem. The data deposited may be partial and only typically relates to one primary publication; the network of interlinked influence of data and publication is not maintained. We foresee the

potential of a network of mutual citations between data and its publication.

The JISC funded Citation, Location and Deposition in Discipline & Institutional Repositories (CLADDIER) project[1] has been investigating the issue of linking publications held in institutional repositories to the underlying data held in specialist repositories by developing the theme of citations, not only for publications but also for datasets.

The partners are the University of Reading, the University of Southampton, the Science and Technology Facilities Council, and a data repository: the British Atmospheric Data Centre (BADC). The latter, part of the National Centre for Atmospheric Science, NCAS)[2] is one of a range of discipline specific data centres commissioned by the Natural Environment Research Council (NERC).

The project is using the BADC as the test bed for data issues and the repositories at STFC and University of Southampton for publication issues. The BADC acts as a primary archive for relevant NERC datasets and in addition also holds other datasets in order 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 atmospheric-science related research programmes. In this domain, active researchers require access to both the written scientific record and data to be able to explore further the topic under examination and at present the links between the data and the written record are fragmented.

At first sight, this issue of cross-citation of data and publication seems straightforward. There are long established conventions for referencing publications from other publications – we can simply extend these to cover the case of citing datasets. Citation indexes and websites such as Citeseer track cross-citation. However, on closer consideration, a number of issues arise which need to be addressed.

- Data Citation
- Dataset Definition
- Data Publication.
- Citation of data in publication repositories
- Cross-searching data and publication archives.
- Addition of Cited-By links to Institutional Repositories.
- Notification of cross-references

[1] http://CLADDIER.badc.ac.uk/

[2] http://badc.nerc.ac.uk/

This paper presents a first look at some of the key citation issues, discusses citation capture issues for publication repositories, and presents potential methods of disseminating these links between repositories.

2. Data Publication

2.1 Data Citation

Atmospheric scientists historically have not cited data in the same way that they cite scholarly publications. To change this culture it is important to develop sensible citation practices that encapsulate the information required, and then to present scientists with examples of how data citations would appear in the bibliography of a paper. Here we construct an example citation and discuss the issues that result.

We begin with data produced by the NERC Mesosphere Stratosphere Troposphere radar facility (MSRF,[1]). 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 regularly as more data becomes available. The existing primary discovery metadata record for this data is single catalogue entry 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 [2] for databases and retrieval systems.

The initial attempted citation follows the form:

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.

Taking these elements in turn we have;

- *Author*: A corporate author is most appropriate, i.e. NERC and the facility name. However if would also be good to credit the principle investigators.

- *Title*: The data is referred to by 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. All of these could be counted as different editions, and there is the complication that the data is regularly updated. 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 are at liberty to create 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. The updating issue is addressed elsewhere.
- *Place of Publication*: This holds little value in an internet world littered with virtual centres. We are leaving this out even though the field is required to follow the guidelines.
- *Publisher*: The organisation responsible for maintaining the primary copy of the data is the BADC, or perhaps NERC. Is this what is meant by the publisher? Publishers also perform some quality control functions, for example peer review. For arguments sake we use the BADC here.
- *Date of Publication*: For this data 1990 onwards. The use of onwards indicates that the dataset is not static, and being updated.
- [*Date of Update/Revision; Date of Citation*]: Again, as this data is updated every hour, the citation date is most useful, but even this is not completely unambiguous, because a citation consumer will not be able to know exactly what the last date was that the user of the data actually had available.
- *Extent*: this is the size of the data and is optional. We do not use this.
- *Series*: Optional and not appropriate in this example.
- *Availability*: A URL from which the data is available.
- *Language*: Optional. Although the data can be considered language neutral, all the metadata, both within the data files and in accompanying documents is in English.
- *Notes*: Optional notes

The citation for the MST dataset now looks 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 four issues raised by this exercise.

- The edition of the data is being defined by the author of the referencing paper rather than clearly defined by the “publisher”.
- 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).
- There is considerable ambiguity as to what data was actually in the dataset at the time of “consumption” (the citation date may not indicate the last date of data in the dataset at that time, and the data centre may not keep that information either).
- Without a priori knowledge, the consumer of the citation has no idea what “type” of data is being cited (this is usually obvious for a traditional publication, e.g. book, journal article etc).

2.2 Dataset definition

In our example we have noted that a broad subset, “version 2, Cartesian products”, can be arbitrarily chosen by the author of the referencing paper as the edition. It should be clear from the data set documentation what “version 2, Cartesian products” refers to, however, as files are added daily to the MSTRF data, this is still not an explicit description of the files that make up this edition or what they contain.

The Climate Science Mark-up Language (CSML, [3]) is the proposed solution the problem of defining a data set. CSML is a standards-based data model and Geography Mark-up Language (GML, ISO 19136) application schema for atmospheric and oceanographic data. It describes the data in terms of feature types and encapsulates the location of the data within the storage system (usually files). It is this latter property that we propose to use to clearly mark the boundaries of datasets for the purpose of citation.

Having established a mechanism to explicitly define datasets, the question remains as to where these boundaries should lie. To answer this question a number of active scientists were asked what they would like to reference as a dataset. Unsurprisingly this varied according to the type of data that they used. Common to all was a desire to define data in terms of the instrument, model or programme that collected the data. There is a need for datasets to be broad scale to avoid frequent references to data that is very similar. These are large-scale aggregations of data files. The reason why such broad data sets are preferred is to avoid multiple references to data that is very clearly related.

2.3 Peer review

The relation between the data and the data centre in the example above is much the same as the relation between a technical report and an institutional library. The data has been lodged at the data centre because NERC wishes to retain its data assets in the same way as an institutional library is charged with curation of technical reports. In both cases any quality control procedures are internal to the organisation, although the procedures use different criteria.

For data to be recognised as an equal of journal articles then independent peer review needs to be performed. Then there would be two classes of publication which we need to deal with: the first is un-refereed material, which by analogy with the grey-literature, one would not expect to see in citation lists, and the second is formally refereed ("published").

There are two mechanisms by which dataset peer review could be achieved: a panel of scientists or experts convened by the data centre, or by using a traditional publishing organisation.

Review by a panel convened by the data centre is the method used by the NASA Planetary Data System [4] to review datasets. This is very much like the traditional publisher organising peer review of journal articles. The purpose of the review is to determine that:

- The data are complete (e.g., no missing calibration files)
- The data are suitable for archiving (i.e., of sufficient quality and with enough documentation to be useful and intelligible in the distant future)
- The PDS standards have been followed

While such an approach could be used at the BADC, using this for all data would conflict between the requirement to be a primary archive and facilitator, as the effort involved would result in significant time delays before data was available.

An alternative is for a traditional publishing organisation to review the data using existing peer review mechanics for paper publication. This second possibility has the advantage of associating the kudos of the publisher's paper journals with the data publications. Many data centres and data providers could submit data to the "data journal" allowing a broader range of data to be included. For the BADC a suitable organisation might be the Royal Meteorological Society (RMS) which publishes a number of respected journals in Atmospheric Sciences and meteorology. An "RMS Data Publications" series would need to be constructed to index the peer reviewed material (the data itself should still lie with the data provider, in this case the BADC)

If this method was implemented then the example citation above 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.

There are still issues associated with the nature of the dataset being cited, what the citation actually points to (metadata, or data?), and the ability to cite within large datasets (an ability analogous to providing page or chapter numbers in large documents). CLADDIER has been addressing these too, and the results of that analysis will be presented elsewhere.

3. Cross-searching Repositories

An aim of CLADDIER is to explore new and automated ways of linking and discovering data and associated publications. A common mechanism to search across these different object types is needed to enable scientists to discover both data and publications from different sources at a common point

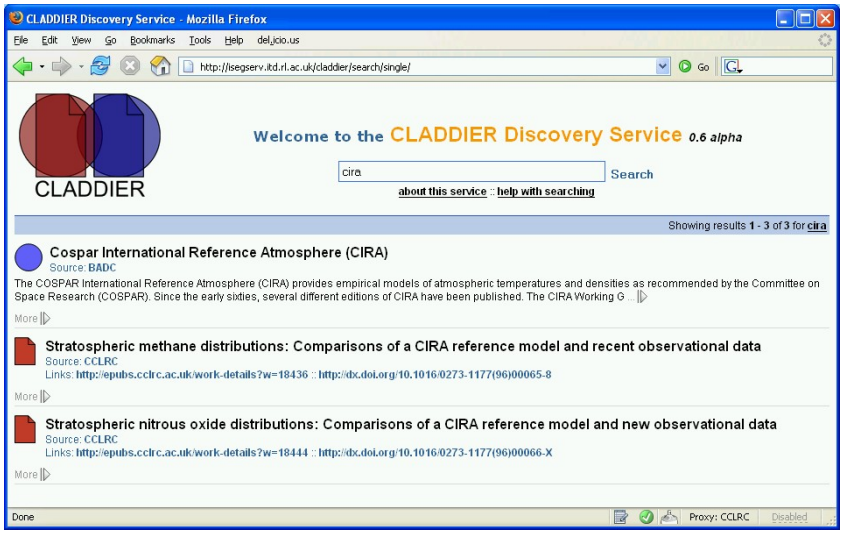

Figure 1: Example of Discovery Searching across Repositories

The CLADDIER Discovery Service[3] is a first step towards that goal. It federates the University of Southampton[4] and STFC ePubs (formerly the CCLRC ePubs IR)[5] Research Publication Repositories as exemplars of Institutional Repositories in the UK. The British Atmospheric Data Centre (BADC), likewise, is an exemplar of a discipline based data archive. The CLADDIER Discovery Service harvests metadata records from these repositories via the OAI-PMH protocol, indexes them, and makes them available via this search interface.

An example of the use of the discovery service is given in Figure 1. The user has entered a key term *CIRA*, and the results are displayed. In this case, the CIRA dataset held by the BADC is discovered, together with two publications on this data. Summary metadata and appropriate links to objects are displayed. The Metadata can be expanded, and the links followed to allow the discovery of items of common interest. The digital object type (publication or data) is represented by the icon displayed with each reference, as well as the repository holding the object.

In the federated discovery service, the notion of data publication is key; a particular dataset needs to have a formal metadata format as discussed above, which can be expressed as a Dublin Core profile, transmitted using the OAI-PMH protocol[6], which can then be merged with publication records.

This is a relatively naïve approach and detail is lost, which means that searches are unfocussed. A refinement of this search engine would be to combine Dublin Core metadata profiles giving more detailed information, so comparisons and links could be inferred more accurately. The recently developed ePrints profile would be appropriate for publication archives [6]; however, no such appropriate profile exists for scientific data: this would be a suitable topic for future work.

4. Cross-citation of resources

A major issue in CLADDIER is how to track cross citation of objects. The citation situation which CLADDIER addresses is illustrated in Figure 2. In this figure, P1 and P2 are papers, and D1 a dataset, and all un-labelled arrows represent citation links.

Traditional publishing uses one directional citation which is entered by the author. Typically such citations will only reference publications. Datasets are typically not cited in those citations, although it is frequently the case that the "inspiration" paper behind the data set is. Data archives wish to track who has been using data resources and thus want to keep track of forward links ("cited-by" links) – they may be informed of a citation from a communication, or from a usage report for example. Once a data archive has recorded a paper as arising from a particular dataset, then the citation from the paper to the data set can be

3 http://isegserv.itd.rl.ac.uk/CLADDIER/search/single/

4 http://eprints.soton.ac.uk/

5 http://epubs.cclrc.ac.uk

6 http://www.openarchives.org/OAI/openarchivesprotocol.html

added, using the data citation form discussed above; this is not necessarily added by the author, but rather by the repository managers.

Thus in CLADDIER, we wish to move from the traditional situation a) in Figure 2 to situation b) where cited-by links are added.

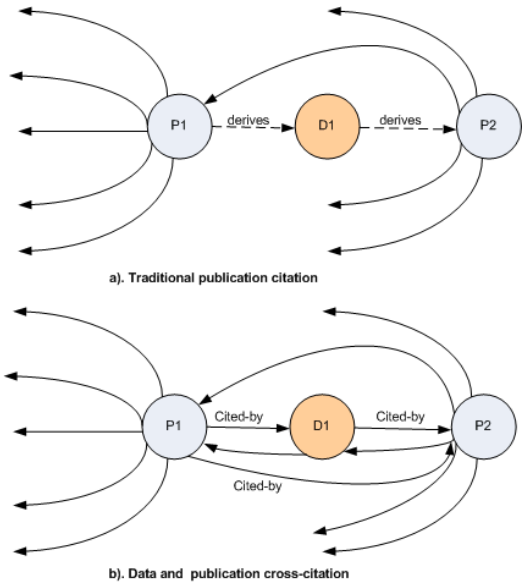

Figure 2: Cross-Citation Linking

However, data and publication repositories typically do not maintain such cross-citation information. Typically they either do not maintain citation information at all (such as ePubs), or they maintain at best a list of references culled from the free-text entry of the paper. A typically example of this is the ePrints archive at the ECS, University of Southampton[7] where citations are extracted from documents. However, links to the cited documents are not recorded, although the user is given the facility to search through the archive, using the Paracite system to locate articles from raw references[8].

In CLADDIER, a more rigorous yet simple mechanism was sought, similar to the citation linking in a citation aggregation service such as Citeseer[9]. A simple data model for citations has been developed, as illustrated in the entity relation diagram in Figure 3. The citation is kept as a simple text block (`referenceText`); it was decided that to break the citation into its component parts was too complicated for the general user (an important consideration for us), and not required for the demonstration, while still maintaining compatibility with Dublin Core guidelines [6]. Each citation has a set of links identified and distinguished (`Uri`); there are typically more than one link associated with each citation, as each citation may give for example a Digital Object Identifier (DOI) and a direct URL to an archive or a journal website. Each citation also maintains an ordering (`seq`), as it was observed that for some citation formats the order that they appear may be significant. The same model can be used to represent both a backward "cites" link and a forward "cited-by" link; thus in the model, a `linkDirection` attribute is maintained to record the direction of the link.

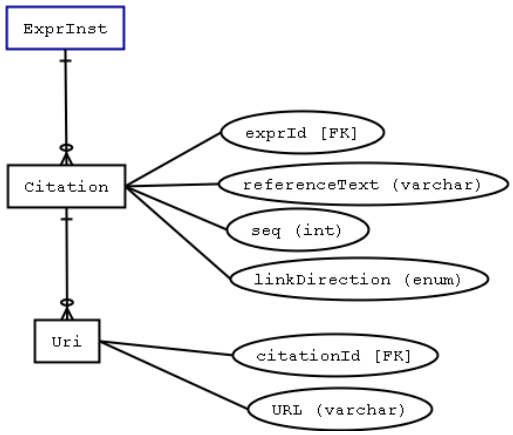

Figure 3: Data Model for Citation

The ePubs system uses the FRBR data model for representing information about publications [7]. FRBR distinguishes between a *Work*, the intellectual creation encapsulated by the entry in the repository; an *Expression*, a realisation of the work in say a particular paper, article, or presentation; a *Manifestation*, the physical embodiment of an expression as say a PDF or Word format; and an *Item* a particular copy of the manifestation files. The citation model thus needs to fit with FRBR. A work may go through a variety of expressions during its development – for example, an idea may go through an abstract, poster presentation, conference paper and presentation, and journal paper. At each stage there may be more or fewer citations, and this set may not remain the same during these stages. Thus it was appropriate to associate the `Citation` entity with the `ExprInst` entity, the particular expression of the work in question. This can happen in both directions, although in practice when a cited-by link is uncovered, it is not always clear which expression of the work is referred to.

This data model is generic enough to cover citations to any digital object type, and thus fit

[7] http://eprints.ecs.soton.ac.uk/
[8] http://paracite.eprints.org/about.html
[9] http://citeseer.ist.psu.edu/

into any of the repositories in CLADDIER. The first implementation has been in the STFC ePubs system. We give an example in Figure 4

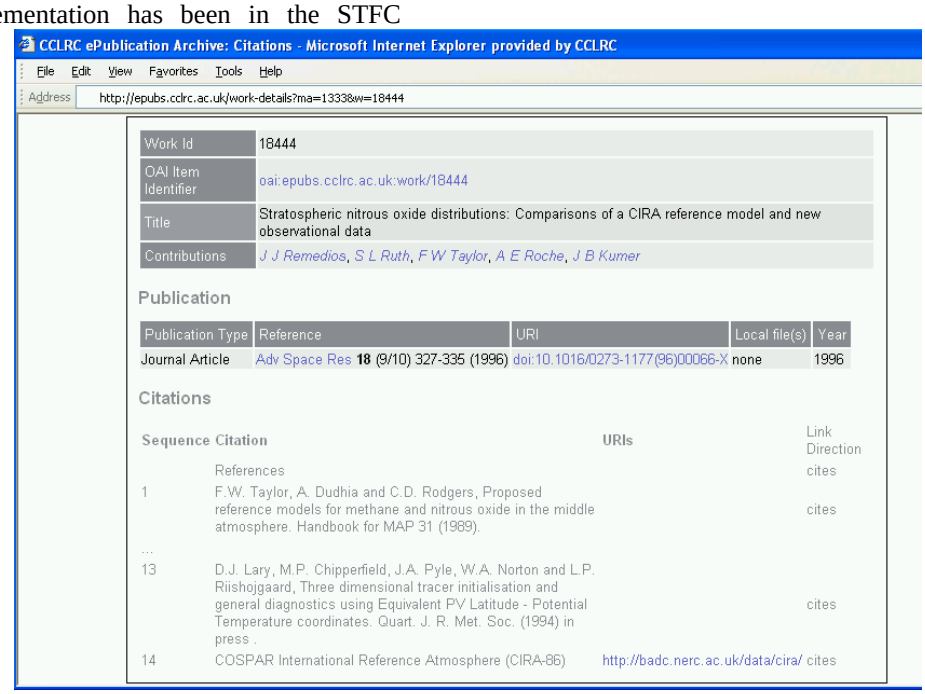

Figure 4: Example of Data Publication Linking in ePubs

(edited for brevity). The link to one of the citations uncovered in the CLADDIER discovery service for the CIRA dataset has been followed to the ePubs archive. One inspecting the entry, an option to view citations is given. This opens up a new window, which is given in the Figure. This lists a number of citations as given in the paper, together with a citation to the data used to derive the results in this paper.

4.1 Citation Notification

In order to complete the desired cross-citation graph as in Figure 2, we need to populate the repositories with cross-citations. In particular, we need to inform repositories that their entries have been cited so that they can add "cited-by" entries. We propose that this citation would be undertaken by a notification service; this is the subject of our current work within CLADDIER.

A number of notification schemes were considered, and a Peer-to-Peer approach was proposed, as sketched in Figure 5. This approach adopts a well-known systems for notification of article referencing used in the Blogging community known as Trackback. Trackback is a "framework for peer-to-peer communication". Essentially, TrackBack involves sending a "ping" request over HTTP, saying "resource A has a link to (cites) resource B". TrackBack was first released as an open specification in August 2002[10], and is supported by blogging software such as MoveableType.

TrackBack uses a two stage protocol. In simple terms, once a repository B discovers a citation URI in a resource B1 to a resource A1 in repository A, it accesses the resource via a conventional http call. Within the returned page Repository B looks for a "TrackBack URL". If it finds one, it calls it with a http Post, delivering a URI to the resource B1. Repository A can then augment the metadata of A1 with the reference to B1, thus completing the cross-reference. This simple protocol is thus well-suited for the task of adding cross-references. We have extended this protocol in two ways; Firstly, the protocol is vulnerable to misuse, so we added the notion of a "whitelist" so that only trackbacks from knows locations are accepted. Secondly, we have extended the data carried in the protocol to allow arbitrary metadata to be passed.

The TrackBack mechanism proves to be robust and simple to implement . With the extension of the protocol to carry arcbitrary metadata, there is potential as a P2P mechanism for inter-repository communication.

10 http://www.sixapart.com/pronet/docs/trackback_spec

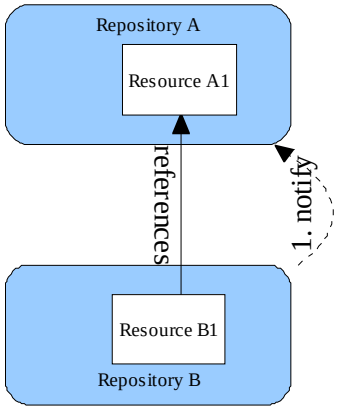

Figure 5: Notification Service

5. Conclusions

Linking traditional publications and data publications involves a range of issues, from the semantic nature of the link (what material should be capture in a citation and how should it be formatted?), through to identifying and deploying technical methods of accomplishing bidirectional navigable links between catalogues which follow these links/citations.

In this paper we have presented some of the work of CLADDIER in this arena, concentrating on three classes of activity: the semantics of citation (and the formal textural syntax), a prototype combined data/document discovery service, and our plans for implementing bidirectional links between the individual partner catalogues.

It will be seen that CLADDIER has asked as many questions as it has answered. There is much best practice to be identified in citation semantics, the use of common (lowest-denominator) metadata schema such as Dublin Core leads to unfocussed searches (so what should we do instead?), and there are issues yet to be resolved with navigable bidirectional links. Nonetheless, and even though the tools developed in the project are relatively simple prototype exemplars. taken together and combined with the theoretical work on data identification and publishing, CLADDIER provides a powerful infrastructure for representing and utilising the network of cross-citations for users and repository managers.

The project is currently undergoing user consultation and testing. The project participants intend to take the work of the project into service into the data repositories. Further, the work is of more general use in other areas of science, and the participants intend to explore how the experience learnt in CLADDIER maybe applied elsewhere, such as in support of the STFC facilities.

Acknowledgements

We would like to thank the support of the JISC Repositories and Preservation Capital Programme for their support in this project.

References

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

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

[3]. 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)

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

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

[6]. Guidelines for Encoding Bibliographic Citation Information in Dublin Core Metadata (2005). http://dublincore.org/documents/dc-citation-guidelines/

[7]. Functional Requirements for Bibliographic Records, IFLA Report (1998). http://www.ifla.org/VII/s13/frbr/frbr.pdf