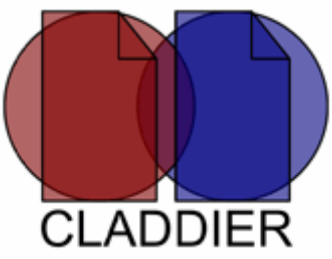

CITATION, LOCATION, AND DEPOSITION IN DISCIPLINE & INSTITUTIONAL REPOSITORIES

http://www.claddier.badc.ac.uk/

CLADDIER Project Report II

(Identifier Migration Issues for Repositories)

Preservation intent and collection identifiers

Sam Pepler, Kevin O'Neil

STFC, Rutherford Appleton Laboratory

Version: 1.2

Date: 4/02/2008

Document History

	Kevin O'Neil	Notes on identifiers
17/08/07	Sam Pepler	First Draft
31/12/07	Bryan Lawrence	Minor modifications
4/02/07	Sam Pepler	More on OAI-ORE and encapsulation methods, v1.2

Abstract

The report was motivated by the need to cite collection objects in scholarly communication, in particular datasets. The persistent identification of datasets is insufficient for scholarly communication citation. There needs to be an explicit statement of what the intended preservation of a dataset will imply; for example it may not necessarily imply bit stream preservation.

There is ambiguity in what type of object a datasets is; with different groups of users applying different connotations. More explicit language such as “data file collection” ensures that objects are well defined. Preservation, identification and object definition are intimately linked. Using what needs to be preserved by a particular user community is an excellent way to define the boundaries and properties of datasets.

Table of Contents

1 Introduction and Motivation

The CLADDIER project [CLADDIER] was funded by the JISC as part it's the Digital Repositories programme (March 2005). The aim of the project was to create a demonstration system that would link datasets held by the British Atmospheric Data Centre [BADC] to items held in institutional repositories at Science and technologies facilities council [STFC] and Southampton University [SOTON]. The approach of the project was to use the existing scholarly system of citation and adapt the repositories to store and communicate this linking information. If authors could reliably identify the data they used in their work using identifiers and bibliographic metadata then the data could be referenced like any other scholarly work.

This approach needs authors to change the way they refer to data in journal articles. In short, this is a culture shift for environmental scientists. The present practice is to reference papers that are associated with the data in some way, perhaps a paper describing the instrument that recorded the data. Another common practice is to refer to web pages describing the data at the data centre. These practices are inadequate because there is frequently insufficient bibliographic metadata to construct the citation and because there is no consistency in the identification of the dataset.

Recognising the need for an identifiable object to allow consistent citation was the driver for this work. As originally conceived this report was to "To identify and clarify issues in the automatic migration of datasets and their links to associated metadata". In particular the links between physical storage which changes when datasets were migrated to new systems or new organisational structures, and the metadata records which define the dataset. This has been broadened to examine other collection level items used in the other CLADDIER partner's repositories. This report looks at definitions of identifiers (section 2), common identifier practice (section 3) and which identifiers are currently used in within the CLADDIER partners repositories (section 4). The report then discusses the preservation of collection level objects (section 5).

2 Identifiers

An identifier allows the accurate and unambiguous identification of a resource. The resource can be anything; a person, house, colour, employee, journal paper or file. All identifiers require an issuing or assignment method, and a resolution method to find the identified object.

The assignment process has to ensure the uniqueness of the identifier. It requires governance processes sufficient to assure users of the identifier that it reliability refers to the same object. The resolution process must be usable by everyone employing the identifiers as part of their activities. This community of users must know how to assign identifiers, how to resolve them and how they should pass them to each other. An identification system is therefore characteristic of a community, which must support and manage it.

An example of an identifier is a bank account number and the associated community is the banking industry and the banking public. The assignment of bank account numbers is done by banks when an account is created. Direct debut instructions referencing account numbers are understood by users of the banking system and can be acted on.

2.1 Persistent identifiers

Identifiers can be short lived things, as anyone in a ticketed queue will observe. Persistent identifiers can be defined as identifiers that set out to identify resources forever. The URN scheme states that “it is intended that the lifetime of a URN be permanent. That is, the URN will be globally unique forever, and may well be used as a reference to a resource well beyond the lifetime of the resource it identifies or of any naming authority involved in the assignment of its name.“ [URN] This persistent foresees the identifier outlasting the issuing community and even the resource it refers to. It still requires that the resolution method exist and is maintained forever, and hence there must be some commitment to the resolution method for an identifier to be persistent.

2.2 Citable identifiers

The CLADDIER project was principally concerned with incorporating data into the scholarly communication cycle. A references section of a paper lists references to the sources used and possibly the products of the work. The intention in providing these references is that others reading the article could return to the cited works and verify or refute the papers claims. For this reason references are more acceptable if they point to articles in respected journals. Journals are seen as persistent and publicly available. Centuries of experience have shown that articles are preserved as valuable objects by university libraries and legal deposit libraries. Anyone at anytime can look up the reference with confidence that they can find the same information as the author of the citing paper found. There maybe the cost of buying the journal, or an embargo period, or a translation problem, or a trip to a legal deposit library, but ultimately the **intent** is that the information is available. With this in mind a citable identifier is a persistent identifier where there is confidence that the resolution mechanism **and** the cited object itself are persistent. Clearly there are no watertight guarantees that the objects are persisted, only that the community was confident that the cited object would be preserved and that the means to find it would also exist.

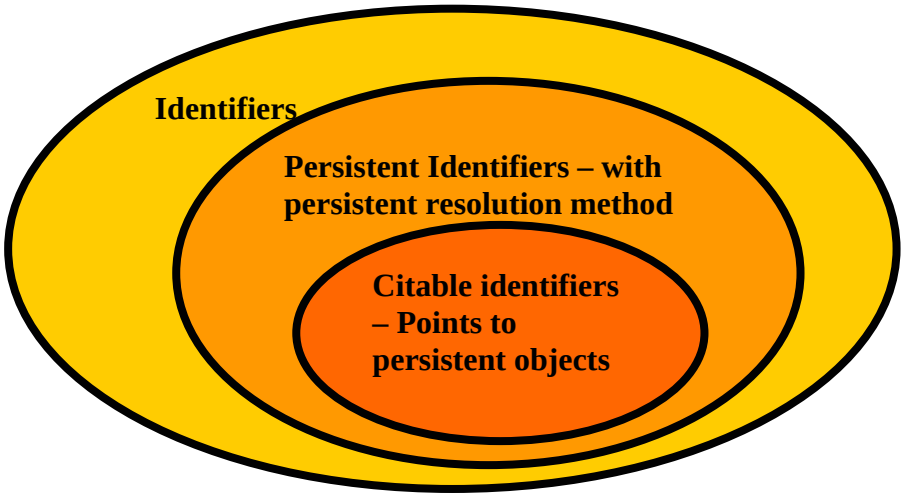

Figure1: Venn diagram showing the relationship between identifiers.

3 Currently cited identifiers

A number of identifiers systems are already in use for citation in scholarly communication. The list below is not exhaustive, but covers the ones that have a general applicability and could be extended to environmental datasets like those held by the BADC. Each method has its own level of trust in the resolution service, but also there is an associated level of trust in the community using the scheme will preserve the identified object.

3.1 Bibliographic metadata

The text string used in scholarly communication to express a reference can be considered an identifier. The combination of author, title, volume, page numbers and journal title represent a unique identification. The syntax for the expression is not necessarily consistent across all journals, nevertheless attempts have been made to standardise in the written form, for example ISO690 [ISO690] and medical reference format [Patrias], and via interfaces like OpenURL [Z39.88]. Clearly, this is the oldest system and the research community and publishers continue to support and promulgate it.

3.2 DOI – Digital Object Identifier

Digital Object Identifiers [DOI] are based on the Handle system. The DOI system consists of a numbering syntax, a resolution service (based on the handle system), a data model, and policies and procedures for the implementation of DOI through a federation of Registration Agencies. DOI assignment is standard practice in the publishing community and is thus clearly citable. It's most common use is to label journal articles. DOI is a general scheme that extends to any digital object. It has already been used with data objects, see Paskin 2005 [Peskin]

3.3 ISBN and other publication numbers

International Standard Book numbers identify books. An ISBN is assigned to each new edition and variation of a book by the books publisher. There is a global governance process for ISBN numbers [ISBN] that has been in place since the 1970s. They are accepted globally as the standard for identification of books, however the scheme is tied to the book concept and is not extendable to datasets. However, International Standard Series Numbers (ISSN), used to identify journal series, show that extending the ISBN principle to a different object type is possible if a clear definition for the object can be found.

3.4 URL

Universal Resource Locators (URL) are probably the most globally recognisable identification system. They are resolvable be any web browser by any anyone. They are frequently used as references, but these are understood to point to transitory objects and are generally augmented with a *cited on* date.

3.5 PURL – Persistent URL

A Persistent Uniform Resource Locator [PURL] is functionally a URL. However, instead of pointing directly to the location of an Internet resource, a it points to an intermediate resolution service. The PURL resolution service associates the PURL with the actual URL and returns that URL to the client. This allows the identifier to be persisted through changes in organisation and

websites. There is no guarantee in the persistence of the resource pointed to, but the act of assigning the PURL indicates that the object pointed to has some level of intended persistence.

3.6 LSID

Life Science Identifiers **[**LSID] are persistent, location-independent, resource identifiers for uniquely naming biologically significant resources including species names, concepts, occurrences, genes or proteins, or data objects that encode information about them. LSID has explicit versioning enabling determination of successor objects. It is a URN scheme maintained by the life science community.

3.7 Multiple identifiers

As stated above identifiers are distinctive to their user community and their need to distinguish objects. A corollary to this is that if an object needs referencing by multiple communities then this may require multiple identifiers. In the banking example an International Bank Account Number (IBAN) enables banks to participate the international banking, but the national bank account number is still valid, resolvable and maintained.

The use of multiple identifiers in any system also allows for consistency checks to be performed. For example, say an article cites a paper using a DOI, but also includes, in the conventional manner, title, journal name, volume and page numbers. If resolution of the DOI yields a paper with a different title then there has been a mistake in writing the citation.

There is a tendency to invent a new identification scheme for every new type of object characterised. This is not sustainable in the long-term and most popular identifier schemes are sufficiently flexible to allow reuse. The need for multiple identifiers should not be seen as an invitation to invent unneeded schemes, but a reminder that objects can be categorised in a host different ways.

4 Identified objects within the CLADDIER partners repositories

The two main type of item under consideration by the project are journal articles in institutional repositories and datasets in the BADC.

The Southampton ePrints archive [ePrints] stores material generated by the university staff and associated projects. The items identified are small collections of files, predominately journal articles, research reports and presentations. The preserved object is the collection of files with a common research theme. The collection can be added to in the initial stages and then the item is fixed and becomes a collection of preserved bit streams. The repository uses URLs to identify these objects in its system, but also holds alternative locations including an “official URL”. Typically an eprint is description of a journal article with an associated file containing the “full text” of the article.

Using the Functional Requirements for Bibliographic Records [FRBR] terminology, the STFC ePubs repository identifies the work as preserved concept. This allows the associated files attached to the work to be added to indefinitely. The key preserved object being the conceptual work item. The repository uses URL and OAI identifiers for these objects. The file object attached to each work, *manifestations,* are referenceable via a URL. As with the Southampton

repository an ePubs work is typically a description of a journal article with an associated file containing the “full text” of the article.

The British Atmospheric Data Centre [BADC] is primarily a repository for data sets. A data set is a collection of data files with a common theme. These are identified using URLs of pages describing the data set, with links to services from which the data files can be retrieved. The number of files in the data sets varies from tens to tens of millions. Frequently the files are added to a dataset as the collection expands. An example is the UK MST radar dataset [MST] where data are ingested into the data set every 15 minutes. Supporting documentation, related metadata also have URL identifiers. Large data sets are not necessarily preserved at the bit level. If a new reprocessed version of the data is available then resource constraints may dictate that only the latest version is stored. Because of this, references to data files via a URL are not encouraged.

5 Preservation of collection level objects

All the repositories use compound collection level objects as their main catalogued items. For the BADC a dataset is a collection of files, ePub explicitly applies the work concept from the FRBR model, and in the Southampton repository each eprint is a collection of files with associated metadata. Outwardly these look like very similar models; collection level items that contain a set of files. Nevertheless the differences are exposed by asking questions how these objects are defined and specifically what properties of the object are intended to be persevered. The Southampton and STFC repositories have a well defined structure and hence a well defined conceptual model of the items in the repositories. The BADC has a less clear definition of the bounds on a dataset.

The diagram below illustrates the number of ways in which the concept of a dataset is used and interpreted. The orange box is the collection of data files that make up the dataset, the collection dataset. This is what the BADC currently define as the dataset. The blue box labelled Meta-dataset indicates the metadata that points to the data files. This is the most common way to reference the dataset, as a reference webpage describing the dataset. A feature collection, represented by the light blue box, is a representation of semantically meaningful features of the data. For the BADC an example of a feature is the temperature measured at weather stations over the last hundred years. Features may span files and only refer to parts of files. Work is underway to create feature descriptions for the BADC data using the Climate Science Modelling Language [CSML]. The enclosing green box labelled “packaged dataset” refers a dataset concept of a self contained information package along the lines of the OAIS reference model [OAIS]. These differing conceptual models of a dataset are used interchangeable with little thought. If we are to progress the identification of datasets, and from there citation, then it is essential that we more clearly define what a dataset is.

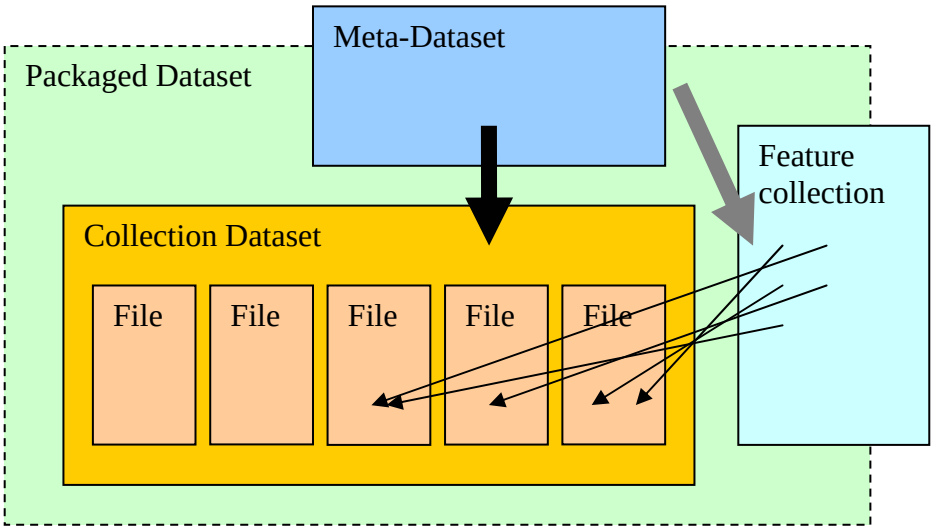

Figure 2: Differing meanings of "datasets"

5.1 Preservation intent

As defined above a citable identifier is one where there is confidence that the object going to be preserved. If the objective of an archive is to preserve identifiable objects, then the definition of those objects is fundamental. Preservation and object definition are also linked in a more subtle way. If an object is changed sufficiently (i.e. not preserved) it is regarded as a different object then it needs a new identity. That is, the criteria governing the creation of object are intimately linked to object properties that are intended to be preserved.

For differing forms of dataset objects above the intended preservation criteria manifest themselves in differing ways. For the file collection dataset the following could be examples of preservation intent statements.

Collection dataset preservation

All the data files in this dataset will not change their bitwise content. This dataset is complete and *no more data files will be added, changed or removed. A full listing of the files and their checksums is available.*

All the data files in this dataset will not change their bitwise content. This dataset is incomplete and files will be added periodically, however no data file will be removed, or changed.

This dataset is incomplete and files will be added and updated periodically. The files may be reorganised and renamed

All the data files in this dataset will not change their numerical content. This dataset is complete and no more data files will be added, changed or removed. This data may be migrated to a different format that preserves file level metadata and the numerical content to at least 6 decimal places.

Examples of meta-dataset preservation intent are more likely to include a scope or theme that is preserved.

Meta-Dataset preservation

This dataset contains data from the XYZ project. Any documentation and metadata for this dataset will change as new information becomes available. All version of the documentation will be preserved indefinably.

This dataset contains data from the ABC instrument. Any documentation and metadata for this dataset will change as new information becomes available. The data files associated with this dataset will be preserved until at least 2020, however this dataset record will be preserved indefinitely.

As digital objects become more complex the preservation intention becomes more esoteric and complex. To resume the bank account example used above. A bank account is identified by a bank number it identifies a pot of money kept by a bank and controlled by an identifiable person. What is preserved? The identifiable person can't change and the money in the account must balance as debits and credits are made. Preservation in this instance is clearly not dependent on making sure the information is static, but on ensuring it changes so that it is conceptually preserved.

To avoid confusion preservation intent should be explicitly stated. This allows clear definition of what the object is.

5.2 Use driven preservation intent

Who decides what is to be preserved? The link between preservation intent and object definition gives the answer. As the user community decides the properties that determine object creation, so they decide on what must be preserved. If an object is regarded as the same by its user community then it has been preserved. The difficulty with this is as the user community expands to include different groups the preservation intention may need to change.

The development of the Meta-dataset is driven by the need to discover the dataset in as wide a context as possible. As the CLADDIER project has demonstrated it is possible for datasets to be discoverable along side journal articles. Clearly the user community for these objects are researchers, but also librarians. The users of the feature collection information (CSML) are the researchers and also developers of interoperable software that understand the feature types used. The file collection dataset is used by researchers and by data centre staff concerned with bit level preservation.

An example of how use has driven the BADC policy on preservation is that of numerical weather prediction data. The data are stored as double precision floating point numbers. For nearly all applications single precision is more than adequate, however some users may need the data to restart the weather model in the exactly same state to ensure repeatability, and hence the decision is to store the data at double precision in the original proprietary format.

Extending the use of datasets enhances preservation by increasing the number of preserved properties this in turn tightens the definition of the dataset and this allows clearer and more explicit citation.

5.3 Methods for defining the compound objects

A compound object is one in which there are identifiable constituent parts. This implies that all objects are compound. A book contains chapters, a chapter contains paragraphs, a paragraph contains sentences and sentences contain words. Even a simple bit-stream contains bits. These examples of information objects are well known and cause little problem for a person to interpret because, chapters and bits carry meanings that are familiar to the most people. Nevertheless, as digital objects become more exotic and complex implicit knowledge of what the objects are and where there boundaries lie become less focused and need to be defined with structured semantics.

5.3.1 Implicit technological boundaries

Implicit boundaries are typical in most objects, a file, a relational database or a physical book are all bound together by the technology. These are all discernable as discreet entities with boundaries defined by the technologies used to bind them together. For example, database records are part of the database because they are contained within the database system; the chapters are contained in the book because they are bound together in a physical volume. There is an assumed coincidence of the technological boundary and the conceptual boundary in most cases. Implied technological boundaries are an imperfect solution as technical limitation rather than semantics may determine the boundary position. This is the case for a multi-volume book, or a database with an extra table added to store unrelated administrative information. Even with these limitations the prevalence of these technologies allows universal assumptions to be made, for instance a row in a database table contain field values that are related to one another.

Digital objects binding technologies provide ways to group bit-streams into more complex objects. Examples of this are:

- A file system uses directories to bind files and sub-directories together
- A database system uses a table to bin a group of records together in a row
- A database system uses a row to bind a field values together.
- A zip file uses a set file format to bind a set of files into a single bit-stream

5.3.2 Explicit Semantic boundaries

The alternative to using implicit technological boundaries is to explicitly state what the object is and thus define its scope. To illustrate this consider a book, as an aggregation of pages bound in its covers, it has an inherent extent. It can also be thought of as an aggregation of chapters marked out in the table of contents. The table of contents is a way of encoding the semantics of book elements, chapters and sections, into a summary that define the boundary of the book. This usually coincides with the implicit technological boundary, which are the book's covers. It is instructive to imagine adding an extra chapter to the end of a book without adding it to the table of contents. Is the new chapter in the book or not? A person examining the book is likely to conclude that the chapter was left out of the table of contents accidentally. If however a chapter is removed from the book that exists in the table of contents the assumption is likely that a chapter has been accidentally miss out. This thought experiment shows how we use and trust both implicit technologic object boundaries and semantic boundaries.

5.3.3 OAI-ORE

The Open Archives Initiatives – Object Reuse and Exchange (OAI-ORE) has provided specifications to allow the aggregation of URI identified resources using a resource map. The resource map is an RDF network that specifies explicitly which URI belong to an aggregation. Because it is an RDF based solution it offers an extensible method to provide type information and include semantics from existing schema. OAI-ORE is thus a method of binding semantics and meaning to bit-stream objects identified by URI. The draft alpha document describing the ORE data model [Lagoze 2007][Lagoze 2008] outlines the suggested relationships and elements.

Appling the ORE model to BADC datasets would be straight forward using the collection of files definition of a dataset. The files within the BADC are aggregated in a hierarchal directory structure and each directory could be given a ORE resource map to explicitly describe the content. This would allow the clear inclusion of all the data files and the exclusion of such items as readme files. These resource maps could also be used from versioning of the dataset content for rapidly changing datasets.

The ORE model also provides an excellent method of presenting the meta-dataset style dataset. The dataset's metadata could be made to explicitly include the data collection dataset rather than the ambiguous state of affairs that the current un-typed link displays. The packed information object dataset can also make use of the ORE method to package the meta-dataset and additional representation information.

However, the feature collection dataset using CSML does not map to ORE model. CSML points to sub-bit-stream elements this is out side of the current ORE scheme. It may be possible to introduce a more complex ORE scheme that deals with parts of bit-streams, but this would go against its philosophy of using standard URI identification.

Overall the ORE model would allow the explicit description of datasets and connect to existing data models.

6 Conclusions

The persistent identification of objects is insufficient for scholarly communication citation. There needs to be an explicit statement about the object preservation, which is trusted by the community of paper authors. For traditional journal article publication the most commonly adopted method is the DOI system.

Preservation intent for collection level objects with complex semantics is very different from that of bit level preservation.

The dataset objects within the BADC have confused identity and this needs to be improved on perhaps by exploring object packaging technologies such as OAI-ORE and CSML. This will lead to datasets having explicit, well-defined preservation intent, which will avoid potential confusion of mismatched intentions.

Extending the use and user communities for preserved objects enhances preservation practice by more clearly defining the objects themselves. Including new user communities is essential for good preservation.

7 References

[BADC] The British Atmospheric Data Centre (BADC), organisation website [cited 2008-01-28] available from http://badc.nerc.ac.uk/

[CLADDIER] Citation, Location, And Deposition In Discipline & Institutional Repositories (CLADDIER) Project website, [cited 2008-01-28] available from http://claddier.badc.ac.uk/

[CSML] Andrew Woolf, Dominic Lowe, Climate Science Modelling Language, Version 2, User's Manual, 2005. [Cited on 2008-01-28] Available from http://ndg.nerc.ac.uk/csml/

[DOI] The Digital Object Identifier System, organisation website [cited 2008-01-28] available from http://www.doi.org/

[ePrints] Southampton University, Southampton ePrints archive, repository website [cited 2008-01-28] available from http://eprints.soton.ac.uk/

[ePubs] Science and Technology Facilities Council, STFC ePublication Archive, repository website [cited 2008-01-28] available from http://epubs.cclrc.ac.uk/

[FRBR] K . G. Saur, Functional Requirements for Bibliographic Records Final Report IFLA Study Group on the Functional Requirements for Bibliographic Records UBCIM Publications – New Series Vol 19, München 1998 available from http://www.ifla.org/VII/s13/frbr/frbr.pdf

[ISBN] International Standard Book Number System for Books, Software, Mixed Media etc. in Publishing, Distribution and Library Practices [cited 2008-01-28] available from http://www.isbn-international.org/

[ISO690] Excerpts from International Standard ISO 690-2, Information and documentation -- Bibliographic references - Part 2: Electronic documents or parts thereof, 2004, available http://www.lac-bac.gc.ca/iso/tc46sc9/standard/690-2e.htm

[Lagoze 2007] Carl Lagoze and Herbert Van de Sompel, Compound Information Objects: The OAI-ORE Perspective, 2007 available from http://www.openarchives.org/ore/documents/CompoundObjects-200705.html

[Lagoze 2008] Carl Lagoze and Herbert Van de Sompel, ORE User Guide – Data Model Overview, 2008, available from http://www.openarchives.org/ore/0.1/datamodel-overview

[LSID] Life Science Identifiers resolution project, project website [cited 2008-01-28] available from http://lsids.sourceforge.net

[MST] Natural Environment Research Council, Aberystwyth Radar Facility, [Hooper, D.]. The NERC Mesosphere-Stratosphere-Troposphere (MST) Radar Facility at Aberystwyth, [Internet]. British Atmospheric Data Centre, 1989-, *[Cited on 2008-01-28]*. Available from http://badc.nerc.ac.uk/data/mst/

[OAIS] Consultative Committee for Space Data Systems, Reference Model for an Open Archival Information System (OAIS) 2002. Also adopted as: ISO 14721:2003. Available from http://public.ccsds.org/publications/archive/650x0b1.pdf

[Paskin] Digital Object Identifiers for scientific data, Norman Paskin, Data Science Journal**,** Vol. 4 (2005) pp.12-20, doi:10.2481/dsj.4.12

[Patrias] Patrias, K. Library of Medicine Recommended Formats for Bibliographic Citation, Supplement: Internet Formats, 2001, National Library of Medicine, Reference Section, Bethesda, USA.

[PURL] Online Computer Library Center (OCLC), Persistent Uniform Resource Locator, organisation website [cited 2008-01-28] available from http://www.purl.org/

[SOTON] Southampton University, organisational website [cited 2008-01-28] available from http://www.soton.ac.uk/

[STFC] The Science and Technologies Facilities Council (STFC), organisation website [cited 2008-01-28] available from http://www.scitech.ac.uk/

[URN] K. Sollins, L. Masinter, Functional Requirements for Uniform Resource Names, RFC 1737. 1994 http://www.ietf.org/rfc/rfc1737.txt

[Z39.88] The OpenURL Framework for Context-Sensitive Services. ANSI/NISO Z39.88-2004 ISSN: 1041-5653 http://www.niso.org/standards/resources/Z39_88_2004.pdf