Policy-making for Research Data in Repositories: A Guide

http://www.disc-uk.org/docs/guide.pdf

Supported by JISC

May 2009
Version 1.2

by
Ann Green
Stuart Macdonald
Robin Rice

DISC-UK DataShare project

Data Information Specialists Committee -UK

Table of Contents

INTRODUCTION

The *Policy-making for Research Data in Repositories: A Guide* is intended to be used as a decision-making and planning tool for institutions with digital repositories in existence or in development that are considering adding research data to their digital collections.

The guide is a public deliverable of the JISC-funded DISC-UK DataShare project (2007-2009)[1] which established institutional data repositories and related services at the partner institutions: the Universities of Edinburgh, Oxford and Southampton. It is a distilled result of the experience of the partners, together with Digital Life Cycle Research & Consulting. The guide is one way of sharing our experience with the wider community, as more institutions expand their digital repository services into the realm of research data to meet the demands of researchers who are themselves facing increasing requirements of funders to make their data available for continuing access.

The guide also can contribute indirectly to efforts to articulate the benefits of sound data management practices, as well as the goals of data sharing and long term access. In addition to setting up data repositories, institutions and libraries in particular can become more active by pursuing some of the following activities:

- Raising awareness of data issues within institutions and the benefits of actively managing research data
- Assisting in developing policies about data management and preservation
- Providing advice to researchers about data management early in the research life cycle; influencing the way researchers will be creating their data, the formats they will use and building a commitment to use the repository to publish/preserve their data
- Working with IT service colleagues to develop appropriate local data management capacity
- Training and introducing data management and curation concepts to research students
- Exploring methods of moving data from work-in-progress storage spaces to repositories in more seamless ways (Lewis, 2008).

This guide is largely based upon the online OpenDOAR Policy Tool (SHERPA, 2007), the OAIS Reference Model (CCSDS, 2002) and the TRAC checklist (OCLC, 2007). Although the focus was initially based on social science datasets, research institutions typically produce a vast heterogeneity of data types and so many other research outputs could be considered within the range of requirements listed in the report; other content includes images, texts, audio, video files as well as scholarly publications and 'grey literature'. The guide does not cover the value-added services that should be offered within a curatorial environment, details of selection and appraisal, nor does it cover

1 http://www.disc-uk.org/datashare.html

INTRODUCTION

advocacy, researcher requirements and data management considerations surrounding funders' mandates. Policies should be developed to address the complex issues related to access mechanisms and user support services as part of any service development process.

ACKNOWLEDGEMENTS

The production of this guide was a collaborative effort by the members of the DISC-UK DataShare project team. The authors would to like to thank Sally Rumsey, Luis Martinez Uribe, and Wendy White for their contributions. Our appreciation goes to editors Harry Gibbs and Jane Roberts, and graphic designer Jackie Clark.

HOW TO USE THIS GUIDE

Each section contains a set of data-related topics compiled from multiple sources that focus on research data quality, management, and preservation. Repository planners and developers can evaluate each set of requirements and choose the most applicable options for their policy and planning purposes.

Examples of how the requirements have been implemented in other instances, or related tools are presented in the shadow boxes.

Lists within the text are marked with 2 different symbols: examples are marked with diamonds and points for consideration are marked with arrows.

Authoritative references are supplied at the end of the guide for further study.

Repository policy examples:

Each of the three DISC-UK DataShare partners has a policy web page.

See Edinburgh DataShare's policies:
http://datalib.ed.ac.uk/DataShare/index.html#pol

See Oxford University Research Archive's policies:
http://www.ouls.ox.ac.uk/ora/ora_documents2/ora_policies

See e-Prints Soton policies:
http://eprints.soton.ac.uk/repositorypolicy.html

Content Coverage

1. CONTENT COVERAGE

1.a SCOPE: subjects and languages

⇨ What subject areas will be included or excluded?

⇨ Are there language considerations? Will translations be included or required? (will text within data files, metadata or other documentation in other languages be translated into English, for example?)

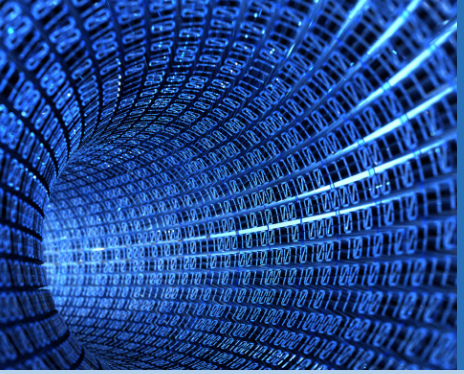
1.b. KINDS OF RESEARCH DATA

As stated in the Research Information Network's report, there are "hugely varied kinds of digital research data – from texts and numbers to audio and video streams." (RIN, 2008.)

> *The 2005 National Science Board (NSB) report entitled Long-Lived Digital Data Collections suggests that data can be differentiated … by their origins: whether they are observational, computational, or experimental. These latter distinctions are crucial when it comes to making choices for archiving and preservation. Observational data (e.g. observations of ocean temperature on a specific date) are essentially historical and cannot be reproduced, and so may be primary candidates for indefinite archiving. Computational data may require archiving complete information about the computer model and the execution (e.g. hardware, software, input data), but not of the data results themselves – which can in theory be reproduced. Experimental data may not be easily reproduced, given both cost considerations and the complexity of all the experimental variables* (Gold, 2007).

⇨ What kinds of research data will be included?

- Scientific experiments
- Models and simulations, consisting of 2 parts: the model with associated metadata and the computational data arising from the model
- Observations: from the astronomical to the zoological – of specific phenomena at a specific time or location, where the data will usually constitute a unique and irreplaceable record (this includes surveys, censuses, voting records, field recordings, etc.)



- Derived data: resulting from processing or combining 'raw' or other data (where care may be required to respect the rights of the owners of the raw data)
- Canonical or reference data: for example, gene sequences, chemical structures etc.
- Accompanying material (RIN, 2008)

Supplemental objects are normally included in the research data 'information package.' These accompanying materials may be coding instructions, interviewer guides, graphic flowcharts of data collection, questionnaires, information regarding the methods used and research techniques employed, codebooks, data collection instruments, summary statistics, database dictionaries, project summaries/ descriptions, and bibliographies of publications pertaining to the data.

1.c. STATUS OF THE RESEARCH DATA

Data production in science may be characterized in several phases and flavours. To begin with there is 'raw data' - data as it comes directly from the scientific instrument. There may be stages of validating and selecting the data. Then the data may be subjected to any number of standard or ad hoc processes that calibrate it – that translate it into a more general schema (e.g. sky maps, catalogues) (Gold, 2007).

⇨ Is the inclusion of the data into the repository determined by its status in the research process / lifecycle? Such as:

- 'raw' or preliminary data
- data that are ready for use by designated users
- data that are ready for full release, as specified in access policies
- summary/tabular data (could be associated with a publication)
- 'derived' data.

See: ARCHER
This is a suite of open-source, production-ready generic e-Research infrastructure components that have been independently tested and documented to provide better management of research data and assist researchers to
- collect, capture and retain large datasets from a range of different sources including scientific instruments
- deposit data files and datasets to e-Research storage repositories
- populate these e-Research data repositories with associated metadata
- permit dataset annotation and discussion in a collaborative environment
- support next-generation methods for research publication, dissemination and access.

For an overview of the ARCHER toolset see http://www.archer.edu.au/products.

Content Coverage

1.d. VERSIONS

> *Digital data can be copied, altered or deleted very easily... This makes it very important to be able to demonstrate the authenticity of data, and to prevent unauthorised access to data for ethical, legal and quality reasons. An important related concept is that of the master file, a formalised and checked final copy of the data (or other materials), or copy at a certain stage of development (as opposed to temporary working versions of data and other files)* (UKDA, 2008a).

Policy considerations for the deposit of multiple versions of a dataset:

- ⇨ The repository uses explicit version numbers which are reflected in dataset names.
- ⇨ The repository records version and status e.g. draft, interim, final, internal[2].
- ⇨ The repository stores multiple copies of a dataset in different formats.
- ⇨ The repository keeps the original copies of data and documentation as deposited.
- ⇨ The repository stores supplemental digital objects with the data file/s.
- ⇨ The repository records relationships between items[3] , such as 'supercedes' or is superceded by'.

VERSION CONTROL

> *It is of great importance to ensure that different copies of files, or materials held in different formats, or information that is cross-referenced between files are subject to version control, whereby checks and procedures are put in place to make sure that if the information in one file is altered, the corresponding information in other files is also altered* (UKDA, 2008a).

Policy considerations for the deposit of multiple versions of a dataset:

- ⇨ The earlier version may be withdrawn from public view. There will be links between earlier and later versions, with the most recent version clearly identified.
- ⇨ The repository ensures that different copies of files or materials held in different formats are subject to version control.

2 The first version might come from the data collection process, the second version might emerge from data cleaning or reformatting, the third might result from composite variable construction, and so forth.

3 The repository maintains the relationship between the parent object and any subsequent child objects (i.e. tracking of unique identifiers for child and parent objects; possible inheritance of parent object identifier to child objects).

⇨ The item's persistent identifier will always link to the latest version.

See Versions Toolkit (Versions of Eprints - User Requirements Study and Investigation of the Need for Standards) at: http://www.lse.ac.uk/library/versions/about.html
Note: currently focuses on academic papers, not other content types, but principles may be applicable.

1.e DATA FILE FORMATS

Preferred formats are formats designated by a data repository for which it guarantees that they can be converted into data formats that will remain readable and usable. Usually, these are the de facto standards employed by that particular community. (DANS, 2008 p. 8)

⇨ Consider what formats will be accepted for deposit, and which are preferred.

In some cases the repository will accept any format, but it won't necessarily guarantee continued access and preservation of obsolete or obscure formats. It is worth remembering that what is generically called a 'dataset' can be made up of data and application specific content in the data file or in a separate program file. 'Raw' data can usually be stored without problems, but the accompanying application can present problems for use and may have licensing issues. The management of file formats over time is covered in section 5.c, File Preservation.

See examples of preferred formats at:
Economic and Social Data Service website:
http://www.data-archive.ac.uk/sharing/acceptable.asp
ICPSR Guide, see section entitled 'Final Data Preparation: File Formats' in:
http://www.icpsr.umich.edu/ICPSR/access/dataprep.pdf

EXAMPLES OF FILE FORMATS

♦ Statistical (quantitative) data:

Application formats: SAS, SPSS, Stata, XML, Excel, etc.
⇨ Will it be required that ASCII files be accompanied by data definition statements for a specific software application?
⇨ Must Excel files be converted to tab- or comma-delimited text?

Content Coverage

- Non-statistical data:
  - Qualitative (textual) data: RTF, HTML, ATLAS.ti, NUD*IST, NVivo, XML etc.
  - Supplemental materials: PDF, Word, image files etc.
  - Digital audio data: for example, WAV files of human speech signals, Audio Interchange File Format (.aiff), MP3 etc.
  - Plasmids encoded in genbank format
  - Java webstart files for software to render data
  - 3D image data: TIFF stacks, RAW, DICOM, and similar format
  - Image and Digital video data: MPEG-2, JPEG 2000.

Policy considerations for acceptable data file formats from producers and/or depositors:

- ⇨ Will the repository accept only file formats which are well-developed, have been widely adopted and are *de facto* standards in the marketplace?
- ⇨ Will the repository *guarantee* that specific file formats will be converted into data formats that remain readable and usable? (See section 5.c, File Preservation.)
- ⇨ Will the repository minimise the number of file formats to be managed as far as is feasible/desirable?
- ⇨ Will the repository accept only 'open' non-proprietary, well-documented file formats wherever possible?
- ⇨ Will compression formats be accepted? (e.g.7zip, gzip, winzip)

Normalisation: File format conversion by the repository when data are submitted. For consideration:

- ⇨ Will the repository convert digital objects of a particular type (e.g. statistical software system files) to a single chosen file format that is thought to embody the best characteristics of functionality, longevity and preservability?
- ⇨ Will the repository convert proprietary formats to non-proprietary formats?
- ⇨ Will the repository create plain text versions of datasets (encoded in either ASCII or Unicode character sets)?
- ⇨ Will the repository retain the original bitstream (file) with the item, in addition to its converted formats?

Dissemination formats:

- ⇨ Will the repository accept formats for the purposes of transfer, storage and distribution to users, which do not meet the conditions of long term access?
- ⇨ Will the repository move data into more current software versions,

Examples of file normalisation:
1. At the University of Edinburgh, Microsoft Access data is converted to SQL with a normalized database schema using SQL Scripts, exporting to text files (Unicode). The schema and the Unicode tables are then zipped and ingested into the repository along with the original version of the database.

2. In a blog posting. Modelling and storing a phonetics database inside a store. Ben O'Steen from University of Oxford describes the steps taken to put a database (a collection of files, grouped by file folder into a hierarchy) into a Fedora Commons-based repository. The first step was to 'describe a physical view of the file system in terms of objects and then to represent this description in terms of RDF and Fedora objects'. Subsequent steps were to 'add simple Dublin Core metadata for both the objects and the individual files and to indicate how the data is interrelated.' Ben also added as an object in the databank the basic description of the custom data format. See:
http://oxfordrepo.blogspot.com/2008/10/modelling-and-storing-phonetics.html

even though they are proprietary? e.g. some repositories store multiple versions of data in SPSS, Stata and XML formats.

File format tools have been developed that are used for file type identification and file format validation.

The JHOVE project (see boxed text) draws a distinction between these two actions as follows. 'Format identification is the process of determining the format to which a digital object conforms. Format validation is the process of determining the level of compliance of a digital object to the specification for its purported format.'

1.f VOLUME AND SIZE LIMITATIONS

Consider any restrictions on the number of files per study or overall size of the study in advance of deposit.

- ⇨ Will there be restrictions by the number of bytes, or number of separate files, or other conditions?
- ⇨ Will the repository make use of compression software to bundle multiple files (e.g. zipfiles)?

In certain circumstances repository systems may be unable to accept or deliver large data files. There are a number of alternative options each dependent upon the technical infrastructure of the host institution.

- ♦ Storage Area Networks (SANs) interconnect seamlessly with different kinds of data storage devices to provide fast access to

Content Coverage

Format identification:

'PRONOM is an on-line information system about data file formats and their supporting software products. PRONOM holds information about software products, and the file formats which each product can read and write. PRONOM is a resource for … impartial and definitive information about the file formats, software products and other technical components required to support long-term access to electronic records and other digital objects of cultural, historical or business value.'
See: http://www.nationalarchives.gov.uk/pronom

Global Digital Format Registry (GDFR) – 'Peer-to-peer network of independent, but cooperating registries of format communicating over a common protocol… will provide sustainable distributed services to store, discover, and deliver representation information about digital formats.'
See: http://www.gdfr.info

Format validation: often used by repositories when adding digital objects

JHOVE (JSTOR/Harvard Object Validation Environment) – a tool used to recognize and validate a limited number of 'popular' file formats. It is both a file type identification tool and a file format validation tool. JHOVE reads through an entire file and determines the degree of compliance to a format specification.
See: http://hul.harvard.edu/jhove

'DROID (Digital Record Object Identification) provides automated file identification information using the PRONOM registry. It 'is designed to… be able to identify the precise format of all stored digital objects, and to link that identification to a central registry of technical information about that format and its dependencies. DROID is a platform-independent Java application, and includes a documented, public API, for ease of integration with other systems.'
See: http://droid.sourceforge.net/wiki/index.php/Introduction

on-line data. The activities that SANs support include disk mirroring, backup and restore, archival and retrieval of archived data, data migration from one storage device to another, and the sharing of data among different servers in a network.

- The Storage Resource Broker (SRB) is a data grid application developed by San Diego Supercomputer Centre aimed at federating collections of distributed data and presenting them to the user as a unified collection. Its features include management, collaboration, controlled sharing, publication, replication, transfer, and preservation of distributed data. The SRB system is middleware in the sense that it is built on top of other major software packages including repository systems such as Fedora (see: http://www.itee.uq.edu.au/~eresearch/projects/dart/out-comes/FedoraDB.php) and DSpace (http://wiki.dspace.org/index.php/DspaceSrbIntegration).
- iRODS™, the Integrated Rule-Oriented Data System, is an open-source data grid software system developed by the Data

Content Coverage

Storage Area Networks and Repositories

At Edinburgh there are two versions of SAN available. One is for research purposes and is under the direction of the Edinburgh Parallel Computing Centre (EPCC). The other is a general SAN run by the Edinburgh Compute Data Facility (ECDF) available to anyone from the university who wishes to back up large research data stores (see: http://www.ecdf.ed.ac.uk). An option for researchers wishing to share large-scale datasets is to create a descriptive metadata record in the Edinburgh DataShare repository pointing to a remote storage location, including the SAN, and providing contact details for requesting access.

Intensive Cyber Environments group at University of North Carolina at Chapel Hill (developers of the SRB). The iRODS system is based on applying SRB technologies in support of data grids, digital libraries, persistent archives, and real-time data systems. The set of assertions these communities make about their digital collections are characterized in iRODS Rules which are interpreted by a Rule Engine to decide how the system is to respond to various requests and conditions (Moore et al, 2007).

Storage Resource Broker and Repositories

The PLEDGE project (Smith et al. 2006), a collaboration between the MIT and University of California, San Diego Libraries, and the San Diego Supercomputer Center successfully demonstrated an SRB-managed and grid-enabled storage architecture for DSpace. This was achieved by standardising archival policies for their digital objects including research data through the use of rules engines to produce scalable, interoperable digital archives and preservation environments.

Distributed Data Collections

The Purdue e-Data repository (see Witt, M. 2008) has been built using a Fedora Web Services framework to provide functionality for remote datasets in addition to datasets being stored locally. In the case of very large datasets for which it is not practical or possible to ingest into a central repository, middleware such as OAISRB has been developed locally to provide an Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) interface to the SRB to enable the harvesting of metadata from datasets residing on a storage grid so that they can be represented alongside local data collections.

Metadata

2. METADATA

A number of decisions need to be made by the digital repository regarding metadata of various types. 'Administrative, descriptive, technical, structural and preservation metadata, using appropriate standards, are used to ensure adequate description and control over the long term' (DCC, 2008).

2.a ACCESS TO METADATA

Considerations:

- ⇨ Anyone may access the metadata free of charge.
- ⇨ Access to some or all of the metadata is controlled.

2.b REUSE OF METADATA

Considerations:

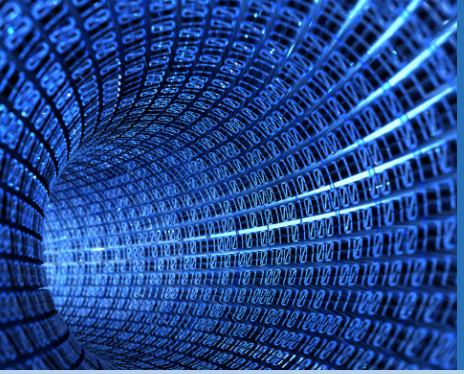
- ⇨ May the metadata be reused in another medium without prior permission provided there is a link to the original metadata and/or the repository is mentioned?
- ⇨ Will it be permissible to reuse the metadata for commercial purposes? Is formal permission required?
- ⇨ Will the repository system allow metadata harvesting of dataset descriptions by other institutions following the OAI-PMH guidelines, or other harvesting protocols?
- ⇨ What level of metadata is re-usable? Dataset descriptions? Full descriptive metadata (e.g. DDI XML record)?
- ⇨ Are data providers required to allow reuse of metadata?

2.c METADATA TYPES AND SOURCES

It is important that researchers deposit additional files known as *documentation* that describe their dataset in more detail, and especially the processes used to create it. Examples of documentation include: a codebook file for statistical data; code for a software program; a format specification; a technical report explaining the research protocol or methodology. Without such documentation, a dataset may not be fit for re-use.

In addition, most repositories attach metadata fields to each deposited item, which conform to some standard or schema. Dublin Core (DC) elements (properties) include descriptive information such as data creator(s), date produced, abstract and subject. DC metadata can be configured within the repository software to conform to an XML-based standard exchange protocol called OAI-PMH, which allows the content to be 'harvested' by web-services and other repositories.[4]

4 Open Archives Initiative Protocol for Metadata Harvesting,



Metadata

Metadata can take several forms, some of which will be visible to the user of a digital library system, while others operate behind the scenes. The Digital Library Foundation (DLF), a coalition of 15 major research libraries in the USA, defines three types of metadata which can apply to objects in a digital library-

- ***descriptive metadata:*** *information describing the intellectual content of the object, such as MARC cataloguing records, finding aids or similar schemes*
- ***administrative metadata:*** *information necessary to allow a repository to manage the object: this can include information on how it was scanned, its storage format etc (often called technical metadata), copyright and licensing information, and information necessary for the long-term preservation of the digital objects (preservation metadata)*
- ***structural metadata:*** *information that ties each object to others to make up logical units (e.g. information that relates individual images of pages from a book to the others that make up the book itself).*

In general, only descriptive metadata is visible to the users of a system, who search and browse it to find and assess the value of items in the collection. Administrative metadata is usually only used by those who maintain the collection, and structural metadata is generally used by the interface which compiles individual digital objects into more meaningful units (such as journal volumes) for the user (University of Oxford, 2005).

The repository must make choices about what kinds of metadata will be required within the repository and from where each type will be produced.

DESCRIPTIVE METADATA

- Bibliographic description/s, e.g. Dublin Core, MODS, MARC21, MARCXML, ONIX.
- A structured catalogue record, or study description, is created for each dataset. Domain-specific descriptive metadata: DDI, SDMX, FGDC, EAD, TEI etc. (For social science data, particularly the DDI, see ICPSR, 2005, chapter 3 *Best Practice in Creating Technical Documentation*).
- Full information relating to the content, structure, context and source of the data; information about the methods, instruments, and techniques used in the creation or collection of the data
- References to publications pertaining to the data.
- Information on how the data have been processed prior to submission to the repository.

Metadata

ADMINISTRATIVE METADATA

- Preservation metadata maintained over the lifecycle of the data, documenting actions taken at submission, curation and dissemination: PREMIS[5]
- 'Event history' information is stored and linked to the digital objects?
- Rights management metadata
- Technical metadata (storage format etc.)
- Representation Information: how data are internally coded, necessary for rendering data in an understandable form. (See section 1.e on File Formats)

A Dublin Core-based dataset profile is in use at the the University of Edinburgh's DataShare repository (http://www.disc-uk.org/docs/Edinburgh_DataShare_DC-schema1.pdf) and the University of Southampton's ePrints repository (http://www.disc-uk.org/docs/ePrints_Soton_Metadata.pdf).

STRUCTURAL METADATA

Structural metadata indicates how different components of a set of associated data relate to one another.

The most straightforward example is Relational Database metadata, described in Wikipedia as follows:

> *Each relational database system has its own mechanisms for storing metadata. Examples of relational database metadata include: Tables of all tables in a database, their names, sizes and number of rows in each table. Tables of columns in each database, what tables they are used in, and the type of data stored in each column. In database terminology, this set of metadata is referred to as the catalogue.*

Other examples of structural metadata:

- FOXML[6] is used in the Fedora repository software, where compound objects are treated as a single file.
- OAI-ORE[7] defines compound objects distributed on the Internet through the creation of resource maps which use unique URLs for each component.

---

5 PREMIS, http://www.loc.gov/standards/premis/

6 Introduction to Fedora Object XML (FOXML) http://www.fedora.info/download/2.0/userdocs/digitalobjects/introFOXML.html

7 Open Archives Initiative Object Reuse and Exchange (OAI-ORE). http://www.openarchives.org/ore/

- METS[8] is used as a 'wrapper' for compound digital objects, allowing them to be identified as such and acted upon, e.g. importing and exporting in repositories.
- RDF[9] provides a simple way to make statements about Web resources, often expressed in RDF/XML, in the form of "triples," i.e. subject-predicate-object expressions that relate objects to one another. (For example, A is version of B.)

About OAI-ORE:
'Open Archives Initiative Object Reuse and Exchange (OAI-ORE) defines standards for the description and exchange of aggregations of Web resources. These aggregations, sometimes called compound digital objects, may combine distributed resources with multiple media types including text, images, data and video. The goal of these standards is to expose the rich content in these aggregations to applications that support authoring, deposit, exchange, visualization, reuse, and preservation. Although a motivating use case for the work is the changing nature of scholarship and scholarly communication, and the need for cyberinfrastructure to support that scholarship, the intent of the effort is to develop standards that generalize across all web-based information including the increasingly popular social networks of web 2.0.' See http://www.openarchives.org/ore/toc.

2.d METADATA SCHEMAS

Repositories may need to put in place additional metadata schemas to support the ingest, management, and use of data in their collections.

Some repositories implement additional or extended metadata schemas for domain specific datasets. For example, creating a new community/ / collection (e.g. for Astronomy or Space Physics) – the SPDML (Space Physics Data Markup Language) schema could be 'plugged-in' to DSpace. This would mean that researchers wishing to deposit Space Physics data would be presented with an ingestion interface based on the metadata schema for their particular data type, thus capturing the richness of that particular domain dataset (which otherwise could be lost by the default DC-based schema).

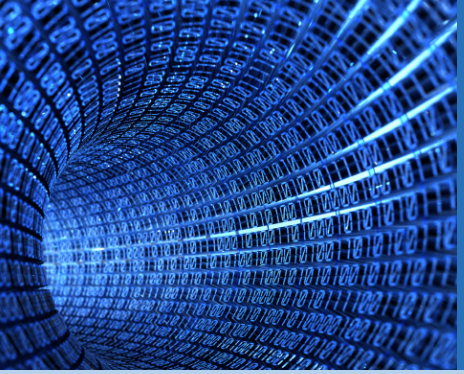
The same could be applied to Learning and Teaching Objects (LOM – Learning Object Metadata), biological species observational data (Darwin Core), SPASE Data Model (standard metadata for space science data description) etc.

8 Metadata Encoding and Transmission Standard (METS). http://www.loc.gov/standards/mets/

9 Resource Description Framework (RDF). http://www.w3.org/RDF/

Metadata

DDI, the Data Documentation Initiative, is an XML-based metadata schema that can describe not just the dataset as a whole but also descriptive material drawn from the life-cycle of a data resource.This can include information, for example, about the source of funding and methodology used in collecting data to an entire set of survey questions and resultant variables. It was initially created as a metadata schema for codebooks but has developed broader application for time series data, complex hierarchical data files, and tabular data. The DISC-UK DataShare project explored the potential of enhancing institutional data repositories through the use of DDI metadata in a briefing paper (Martinez, 2008).



Submission of Data (Ingest)

3. SUBMISSION OF DATA (INGEST)

3.a ELIGIBLE DEPOSITORS

⇨ Will eligibility be restricted by status?

Deposits may be made by, for example:

- Accredited members, academic staff, registered students, employees of the institution, department, subject community or delegated agents
- Data producers or their representatives ('self deposit')
- Only repository staff.

⇨ Will eligibility be restricted by content?

For example, eligible depositors:

- may only deposit their own work
- must enter descriptive metadata for all their data
- are limited to depositing complete datasets as defined by the repository
- may only deposit data of a certain type or subject.

⇨ Will the repository provide a confirmation of receipt to the depositor including a request to resubmit a digital object in the case of errors resulting from the submission?

3.b MODERATION BY REPOSITORY

Considerations:

⇨ Are submissions checked to ensure that data integrity has been fully maintained during the transfer process? If so, spot checks, or all submissions?
⇨ The repository checks metadata records for accuracy.
⇨ The repository adds Digital Object Identifiers (DOIs) or another persistent identifier, such as the Handle system.
⇨ Does the repository's administration review items for the following:
- eligibility of authors/depositors?
- relevance to the scope of the repository?
- valid formats?
- exclusion of spam?

3.c DATA QUALITY REQUIREMENTS

3.c.1 RESPONSIBILITY

The repository should have clear and concise depositor agreements written in plain language that are presented to depositors with each acquisition. In most cases, data producers are responsible for the quality of the digital research data. The repository is responsible for the quality of storage and availability of the data. For example, the following statement could be part of a submissions policy:

> *The validity and authenticity of the content of submissions (all materials submitted by the depositor, including full data and metadata) is the sole responsibility of the depositor, and is not checked by the repository* (SHERPA, 2007).

In these cases, the repository accepts no responsibility for mistakes, omissions, or legal infringements within the deposited object. There may be situations in which the depositor does not guarantee that the dataset is accurate and the depositor indemnifies the repository's institution against any legal action arising from the content of the dataset. One way to mitigate against legal risks is to have a 'take-down' policy for removal of objectionable items. (See section 6, Withdrawal).

In some cases, licenses are presented to depositors to cover the range of requirements for reuse of the data (see section 4, Access and Reuse of Data).

3.c.2 QUALITY ASSESSMENT

If the repository evaluates data quality in order to make decisions about whether to accept content or not, the repository can choose to determine the quality by assessing the following:

Their intrinsic scientific quality through assessment by experts and colleagues in the field. Considerations:

- ⇨ Are the research data based on work performed by the data producer (researcher or institution that makes the research available) and does the data producer have a record of academic merit?
- ⇨ Was data collection or digitization carried out in accordance with prevailing criteria in the research discipline?
- ⇨ Are the research data useful for certain types of research and suitable for reuse? (DANS, 2008 p. 7)

⇨ The required contextual information (metadata) has been provided by the data producer. Descriptive, structural and administrative metadata must be provided, or created by the repository, in accordance with the applicable guidelines of the repository.

Submission of Data (Ingest)

- ⇨ Data files are inspected to ensure that variables and values are accurate according to the documentation supplied and are sufficiently labelled for secondary use.
- ⇨ Checks are made to verify that metadata in a data file matches metadata in descriptive documentation (e.g. variable names in a dataset match variable names in a codebook).

3.d CONFIDENTIALITY AND DISCLOSURE

Data Seal of Approval from DANS (Data Archiving and Networked Services, The Netherlands)

'DANS was given the task ... to develop a seal of Approval in order to ensure that archived data can also be found, recognized and used in future. Such a seal of approval can be applied for and awarded to research which meets a number of recognizable criteria in the area of quality, durablity and accessibility of the data. The seal of approval can also be requested by and awarded to data repositories that want to store research data perminently and make them accessible. The seal of approval contains guidelines for applying and checking quality aspects of the creation, storage and (re)use of digital research data in the social sciences and humanities. These guidelines serve as a basis for granting a "data seal of approval.' (DANS, 2008).

Repositories will often require that data depositors ensure that data meet requirements of confidentiality and non-disclosure for data collected from human subjects. In some cases, the repository may alter sensitive data to create anonymised data that can be distributed to its user community.
In the case of one social science data archive, data collections acquired by the repository "undergo stringent confidentiality reviews to determine whether the data contain any information that could be used - on its own or in combination with other publicly available information - to identify respondents. Only after the completion of those reviews are data made available from the repository. Should such information be discovered, the repository alters the sensitive data after consultation with the principal investigator to create public use files that limit the risk of disclosure" (ICPSR, 2007c).

3.e EMBARGO STATUS

Some repository infrastructure systems have the technical capacity to embargo or sequester access to data until the content has been approved for release to the public. Agreements about the embargo – its length and what triggers its ending – need to be made between the repository and its contributors.

Submission of Data (Ingest)

In the UK, the Data Protection Act defines personal, confidential and sensitive data and sets out parameters under which data processors (researchers) can use them. Furthermore, Research Ethics Committees may approve or disapprove data sharing plans based on the researchers' methods and what sort of consent has been sought.

A thorough overview of 'consent, confidentiality and ethics in data sharing' for researchers has been written by the UK Data Archive as part of their suite of web pages dedicated to managing and sharing data (UK Data Archive, 2008d).

> *Institutional repositories typically do not permit content to be removed once submitted. However, a variety of legitimate circumstances might require an institution to limit access to particular content to a specific set of users. These circumstances might include copyright restrictions, policies established by a particular research community (limiting access to departmental working papers to members of that department, for example), embargoes that an institution's Sponsored Programs Office might require to keep the institution in compliance with the terms of sponsor contracts, and even monetary access fees for certain data. Implementing these policy-based restrictions requires robust access and rights management mechanisms to allow or restrict access to content -and, conceivably, to parts of digital objects - by a variety of criteria, including user type, institutional affiliation, user community, and others,* (Johnson, 2002).

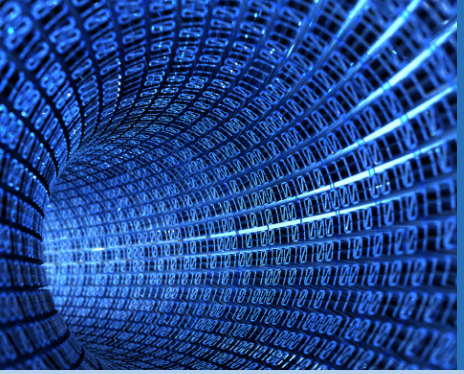
Considerations:

- ⇨ Will the embargo status and length of embargo be determined by repository staff or data producers (or their representatives)?
- ⇨ Will the repository choose to hold materials where the metadata is publicly accessible but the data are embargoed or restricted in some way?
- ⇨ Can the repository software be configured to automatically release the data on the end date of the embargo or will manual procedures be needed?

3.f RIGHTS AND OWNERSHIP

The repository service may wish to enter a license agreement with the depositor upon transfer of the data item through a written or click-through Depositor Agreement.

- ⇨ How extensive are the *rights given to the digital repository* over the deposited material?
- ⇨ Are there limitations to what the digital repository is allowed to do with the material?



Submission of Data (Ingest)

Considerations:

- ⇨ Are file format changes limited by authenticity requirements?
- ⇨ Is the repository free to change the original submitted material as it sees fit during the preservation processing? (Beedham et al, 2005 p. 105)
- ⇨ Can the repository translate, copy or re-arrange datasets to ensure their future preservation and accessibility, and keep copies of datasets for security and back-up, or can depositors notify the repository that specific restrictions apply?
- ⇨ Can the repository migrate datasets to another repository (e.g. subject-based, institutional) on condition that all metadata are migrated with the dataset, and that no charge will be levied by the destination repository?
- ⇨ Can the repository incorporate metadata or documentation into public access catalogues for the datasets it holds?
- ⇨ Will the repository be under any obligation to reproduce, transmit, broadcast or display a dataset in the same format or software as that in which it was originally created?
- ⇨ While every care will be taken to preserve the dataset, will the repository be liable for loss or damage to the dataset or any other data while it is stored in the repository or a repository to which the dataset is subsequently migrated?
- ⇨ Do depositors retain the right to deposit the item elsewhere in its present or future version(s)?

Depositors' agreements regarding *copyright* may cover the following:

- ♦ The content of deposited dataset does not breach any law.
- ♦ It is original and does not infringe the copyright of any other person (e.g. it is not derived from a licensed or commercial product).
- ♦ If it contains material that is copyright of a third-party, the depositor has secured permission from the rights-holder or their representative to include such material in the dataset (including a commercial or academic partner in a research project).
- ♦ Any third-party materials for which the depositor has not secured the necessary permissions have been deleted from the dataset before deposit.
- ♦ If the dataset has been sponsored or subsidised by any institution or organisation other than the depositor's employer, s/he has fulfilled all obligations to that institution or organisation regarding publication.

The institution's legal office may wish to sign off on the final Depositor Agreement.

Submission of Data (Ingest)

Does data have copyright?

Facts are not copyrightable in any jurisdiction, but there are varying levels of protection for data in different countries. In the European Union, the Database Directive has led to greater protection for compiled databases than in the USA which is governed solely by copyright law. The Digital Curation Centre has produced a briefing paper that clarifies IPR in Databases within UK law (McGeever, 2007).

Sample copyright statements for inclusion in a Depositor Agreement:

Any copyright violations are entirely the responsibility of the authors/depositors. If the repository receives proof of copyright violation, the relevant item will be removed immediately.

The repository shall not be under any obligation to take legal action on a depositor's behalf or on behalf of any other rights holder in the event of breach of intellectual property rights or any other right in the material deposited.

Depositors retain all moral rights to the work including the right to be acknowledged.

An example of a depositor agreement (for the Edinburgh DataShare repository) is here: http://datalib.ed.ac.uk/DataShare/Depositor-Agreement.pdf.

Access and Reuse of Data

Open Data:

'Open Data is a philosophy and practice requiring that certain data are freely available to everyone, without restrictions from copyright, patents or other mechanisms of control. It has a similar ethos to a number of other "Open" movements and communities such as open source and open access' (Wikipedia - http://en.wikipedia.org/wiki/Open_data).

Recent technological advances in web services and infrastructures have democratised the processes allowing data to be used, stored, visualised, analysed in collaborative ways.

Examples of collaborative utilites which use 'community-driven' Web 2.0 technologies to visualise or 'mash' numeric data include Many Eyes, Swivel and Infochimps. There are also a whole range of spatial visualisations or mashups using mapping tools, earth viewers or open geo-browsers such as Google Earth, OpenStreetMap, Open Layers, which capitalise on and utilise the preponderance of location-based information. These utilities allow researchers to upload and analyse their own data in 'open' and kinetic environments (Macdonald, 2008 a, b).

By opening up their code to repository developers (e.g. through APIs) or by the development of new plug-ins or tools, numeric and spatial data visualization could be enhanced within the repository environment. This would have the potential of engaging potential depositors, to enhance output, and to provide analysis and visualisations as part of 'value-added' functionality.

See also: non-profit organisations such as the Open Data Foundation (http://www.opendatafoundation.org/), and the Open Knowledge Foundation (http://www.okfn.org/)

4. ACCESS AND REUSE OF DATA

4.a ACCESS TO DATA OBJECTS

Digital research data should be easy to find, and access should be provided in an environment which maximises ease of use; provides credit for and protects the rights of those who have gathered or created data; and protects the rights of those who have legitimate interests in how data are made accessible and used (RIN, p.10).

OPEN ACCESS

The open access publishing movement was started by the Budapest Open Access Initiative and its signatories in February 2002. 'Open Access means access to material via the Internet in such a way that the material is free for all users to read and use.' (Wikipedia, http://en.wikipedia.org/wiki/Open_access)

Considerations:

- Will access to the content in the repository be open to the public? Note restrictions on reuse may apply even though *access* is allowed. (see section 4.b, Use and Reuse of data Objects.)

CONTROLLED ACCESS

If access to some or all items is controlled, the repository might be required to limit access based upon:

- User type/status (general public, research organization, membership, administrative staff)
- Location – access restricted to specific IP location or physical location
- The number of concurrent users of an object at a given time.

RESTRICTED ACCESS

The repository may be required to restrict access to data for a number of reasons e.g. datasets might contain confidential information that could lead to the identification of respondents or datasets may be used to develop a patent or commercial product. How is this implemented? (Dulong de Rosnay, 2008).

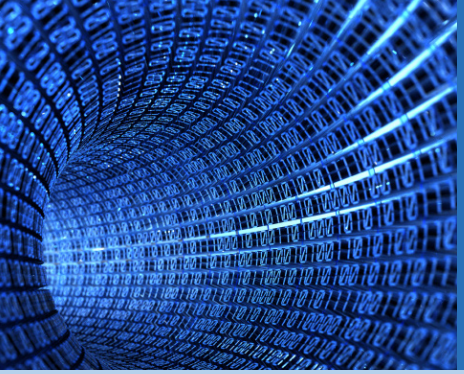
How will restricted access conditions be implemented?

- The repository stores data on a secure non-networked server and has clear policies regarding the physical safekeeping and use of the restricted datasets when in the researcher's possession.
- The repository offers a Data Enclave for restricted data with confidentiality issues for which there is heightened sensitivity to disclosure, as indicated either by the depositor or the repository. The only form of access to such data is through on-site analysis in the repository's secure data enclave with very controlled conditions (ICPSR, 2007a).

REGISTRATION

Considerations:

- Will registration be compulsory before downloading or accessing data?
- Will registration be compulsory for depositors only?
- Will a local registration system be implemented or will it interoperate with other systems (e.g. UK Access Federation or campus single sign-on)?



Access and Reuse of Data

The following digital repository systems are used by social science data archives and may be implemented locally, though they are not open source and may involve payment. They offer a range of data management and online data analysis features.

The Dataverse Network Project at Harvard University: 'The extensive digital library services of each dataverse include data archiving, preservation formatting, cataloguing, data citation, searching, conversion, subsetting, online statistical analysis, and dissemination. Each dataverse presents a hierarchical organization of datasets, which might include only studies produced by the dataverse creator (such as for an author or research project), those associated with published work (such as replication datasets for journal articles), or datasets collected for a particular community (such as for a journal's replication archive, or a college class or subfield.)' (http://thedata.org)

NESSTAR at Norwegian Social Science Data Services: 'Nesstar is a software system for data publishing and online analysis. The software consists of tools which enables data providers to disseminate their data on the Web. Nesstar handles survey data and multidimensional tables as well as text resources. Users can search, browse and analyse the data online.'
(http://www.nesstar.com)

Survey Documentation and Analysis (SDA) from the University of California at Berkeley:
'SDA is a set of programs for the documentation and web-based analysis of survey data. There are also procedures for creating customized subsets of datasets.'
(http://sda.berkeley.edu)

- ⇨ Will access be managed at the institutional/departmental level, user registration level, or at the dataset level?
- ⇨ Are all datasets individually tagged with differing rights, permissions, and/or conditions?
- ⇨ Are users required to confirm their acceptance of the terms and conditions of access?

ACCESS METHODS

Considerations for data delivery:

- ⇨ The repository provides a link to download entire data files.
- ⇨ Data can be accessed through a batch feature.
- ⇨ Data can be accessed through a query-based system.
- ⇨ Extracts of data may be chosen for download, and descriptive statistics may be created by the user.
- ⇨ Analytical routines to use with specific software applications will be provided online (e.g. set up files or system files).
- ⇨ Visualization and mapping applications will be provided online.
- ⇨ Other web services may access the data in the system.

Access and Reuse of Data

Creative Commons 'provides free tools that let authors, scientists, artists, and educators easily mark their creative work with the freedoms they want it to carry so others can share, remix, use commercially, or any combination thereof.' (http://creativecommons.org)

Science Commons is a US-based project that explores legal and cultural ramifications of data sharing. They have concluded that using Creative Commons licenses on data is not appropriate, but others, such as the Open Knowledge Foundation in the UK, have contested this. A Digital Curation Centre briefing paper covers Science Commons' conclusions on this topic (McGeever, 2009).

Open Data Commons provides forms of data licenses for data that are consistent with the open data movement, for example, the Public Domain Dedication and Licence (PDDL) and Open Database License (DDbl). (http://www.opendatacommons.org)

4.b USE AND REUSE OF DATA OBJECTS

The repository may have a policy informing users of possible limitations. Prior to downloading data, will the user be required to agree to the terms of an online Terms of Use statement?

Considerations:

- The data are in the public domain and reuse is not limited.
- The data are covered by contractual restrictions for attribution, limitation to non-commercial usages, prohibition to modify data, or other constraints on their redistribution or modification.
- All reuse of data is prohibited.
- Restrictions are applied for data users on the right to reformat and redistribute.
- Restrictions can be lifted on a case-by-case basis (e.g. by request).
- Stipulations are made for the data to be used in an ethical manner or responsible manner.

Repository services may wish to enable depositors to attach a Creative Commons (CC) license to their work. The following CC licenses are commonly used in internet applications and apply to copyrighted works.

- *Attribution:* The repository allows data users to copy, distribute, display, and reuse your copyrighted work, and derivative works based upon it - but only if they give credit in the required manner.
- *Non-commercial:* The repository allows others to copy, distribute, display, and perform your work, and derivative works based upon it - but for non-commercial purposes only.
- *No Derivative Works:* The repository allows others to copy, distribute, display, and perform only verbatim copies of your work, not derivative works based upon it.

Access and Reuse of Data

See: CLADDIER briefing on data publication –
At present written publications cite data sources within text and acknowledgements, making the data difficult to discover automatically. The CLADDIER project in the UK investigated and identified a number of user requirements for citing data including: need for an unambiguous reference to a well defined permanent entity; the reference/citation needs to be understandable for humans; an unambiguous data reference should include the activity or tool which produced the data (where practicable); the source of the data (i.e. the repository) may be as important as the producer and needs to be unambiguous.

They also found that data producers have certain requirements for citation, namely it should be traceable to the data provider/producer, usable for usage metrics, recognised as intellectually equivalent to academic papers, and able to be used to search for papers citing data. (http://claddier.badc.ac.uk/trac/raw-attachment/wiki/wp10-mtg/CLADDIER-datapub-briefing-20070514.pdf)

- *Share Alike:* The repository allows others to distribute derivative works only under a licence identical to the licence that governs the original work.

Citation

- Will users of the data be required or requested to cite the dataset/s?
- What is required? For example will authors, title and full bibliographic details be given in addition to any of the following:
  - a hyperlink or URL for the original metadata page
  - the original copyright statement
  - the original rights permission statement
- Is mention of the repository mandatory in institutional policy?

Copies

What restrictions, if any, will be placed on making copies of the data and accompanying materials?

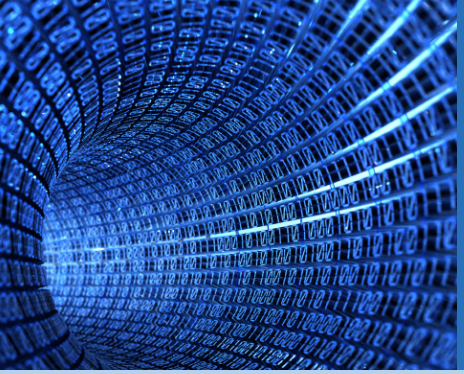
- Copies can be reproduced, displayed or given to 3rd parties in any form or medium
- Copies can only be made for personal research or study, educational or not-for-profit purposes
- Copies can be made for commercial purposes without prior permission or charge
- Full items must not be sold commercially without formal permission of the copyright holders.

Harvesting of metadata and textual content:

- ⇨ Will it be permissible for data or metadata items to be harvested by robots for full-text indexing or citation analysis?
- ⇨ Will the repository system allow searching of its information objects by search engines such as Google and Yahoo, according to current protocols and security policies?

4.c TRACKING USERS AND USE STATISTICS

- ⇨ Will all access mechanisms be sufficiently granular to allow the identification of individual users in order to maintain logs of actions performed by users?
- ⇨ Will all actions relating to access to the material be recorded?
- ⇨ What repository use statistics will be made available and to whom? For example:
  - ♦ Repository staff
  - ♦ Depositors
  - ♦ Data producers
  - ♦ Repository users
  - ♦ Research funders / publishers
  - ♦ Institutional / organisational purse holders / senior management?



Preservation of Data

5. PRESERVATION OF DATA

Repositories may have different purposes, from data sharing to long-term archiving, but some policies and planning around preservation issues must be taken from the outset, to ensure continuing access as long as is needed.

5.a RETENTION PERIOD

Define a dataset retention period, for example:

- Items will be retained indefinitely.
- Items will be retained for at least xxx years from the date of deposition.
- Items will be retained for the lifetime of the repository.
- Retention periods may be set for individual items, as required.

5.b FUNCTIONAL PRESERVATION

It may not be possible to guarantee the readability of some file formats due to software obsolescence, but the repository may choose to promise to maintain the *usability and understandability* of the specific file formats over time.

⇨ If the repository promises usability and understandability over time, what specific file formats will be included in this guarantee?

Database Curation:
'Database archiving focuses on archiving data that are maintained under the control of a database management system and structured under a database schema, e.g. a relational database. When archiving scientific and reference data, database archiving is frequently regarded as maintaining a collection of database snapshots over time…. This form of database archiving involves making off-line copies of the data and managing these copies efficiently.' For further information, see the DCC briefing paper (Müller, 2009).

5.c FILE PRESERVATION

The earlier section on data file formats (1.e) covers which file formats will be accepted for deposit. This section deals with how the repository will manage datasets over time.

Preservation of Data

'Over time, items stored in DSpace will be preserved as is, using a combination of time-honoured techniques for data management and best practices for digital preservation. As for specific formats, however, the proprietary nature of many file types makes it impossible to make guarantees' (MIT Libraries, 2002).

Considerations:

- ⇨ Will varying levels of support be offered for various file formats?
- ⇨ If not all formats are supported, will some formats be preserved only at the bit level (no migrations or transformation are planned)?
- ⇨ Will the repository use encryption or compression for archival files?
- ⇨ Will the repository regularly back up its files according to current best practice?
- ⇨ Will data files be migrated to new file formats where necessary to preserve access to their intellectual content?

Institutional preservation strategies for research data:

DataStaR, a Data Staging Repository hosted by Albert R. Mann Library, at Cornell University promotes the publishing or archiving data and high-quality metadata to discipline-specific data centers and/or Cornell's own institutional repository. It also supports collaboration and data sharing among researchers by providing data curation services early in the research lifecycle, and then promotes the transmission of data to repositories better suited for long-term curation and preservation.
See: http://datastar.mannlib.cornell.edu/

Monash University introduces the concepts of Domains, Data Stores and Curation Boundaries to address data preservation and curation throughout the research lifecycle. Data travels from a private research domain to a shared research domain to a public domain through a series of curation boundaries where the data objects themselves can be appraised, described, controlled and migrated to corresponding data stores (Treloar et al, 2007).

5.d FIXITY AND AUTHENTICITY

The OAIS standard (CCSDS, 2002) defines fixity as information which can be used to validate the authenticity of information extracted from a digital object. Fixity checks such as checksums, message digests, and digital signatures are used to verify that a digital object has not been changed between two points in time or events. Information created by these fixity checks provides evidence for the integrity and authenticity of the digital objects.

Preservation of Data

When should fixity information be created and verified?
Fixity information can be created or verified at numerous stages in the preservation workflow. The management policy for the archive, its context and the level of confidence required will dictate when and what type of fixity information is created and how often it is verified. Appropriate points in the lifecycle for generating and verifying fixity information include:

- At point of creation
- At point of accession
- At point of ingest
- At point of transformation
- As part of normal maintenance routines
- At point of dissemination (Paradigm Project, 2007).

To ensure authenticity of the digital objects and the metadata:

- Will the repository establish protocols and audit trails to show who has accessed each dataset, and who has enhanced or annotated it? (RIN, 2008 p.14)
- Will existing relationships between datasets and explicit links be maintained?

Withdrawal of Data and Succession Plans

6. WITHDRAWAL OF DATA AND SUCCESSION PLANS

Review of conditions for withdrawal of datasets:

- ⇨ Will items be removed from the repository?
- ⇨ Under what conditions will the repository choose to remove items? Reasons for withdrawal by repository might include:

  - ♦ copyright violation
  - ♦ legal requirements and proven violations
  - ♦ national security
  - ♦ falsified research
  - ♦ confidentiality concerns etc.

- ⇨ Will items be removed at the request of the depositor?

  If items are withdrawn, indicate the terms of the withdrawn items, for example:

  - ♦ Withdrawn items are deleted entirely from the database.
  - ♦ Withdrawn items are not deleted, but are removed from public view.
  - ♦ Identifiers/URLs for withdrawn items are retained and for how long? For example, indefinitely, transiently, or not at all.
  - ♦ URLs will continue to point to 'tombstone' citations, to avoid broken links and to retain item histories together with a link to replacement versions, where available, and with reasons for withdrawal.

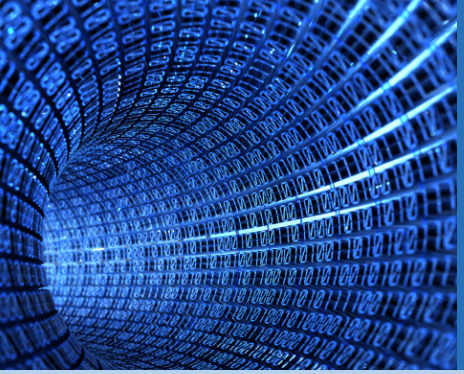
Metadata for withdrawn items:

- ♦ Should a dataset be removed by either the repository or the depositor, the repository reserves the right to retain its metadata record in the repository as trace of the dataset.
- ♦ The metadata of withdrawn items will / will not be searchable.

CLOSURE and SUCCESSION

Considerations:

- ⇨ In the event of the repository being closed down, will the data be transferred to another appropriate archive?
- ⇨ Will items be returned to their originators?
- ⇨ Are there other institutional or domain-specific repositories that can offer succession arrangements?



Next Steps

7. NEXT STEPS

Across the globe, institutions are driven by an increasing awareness of research data outputs as highly valued resources within an integrated digital knowledge base. Mandates to share, manage, and preserve this digital heritage present challenges to researchers, their home institutions and national archives, as well as their own disciplines to create trusted, shared, understandable, and usable solutions. The promises of globally open data fuel the demand to build the infrastructure, context, and access methods that further enable a dynamic digital future.

This guide proffers a decision-making framework for institutions at a time when rapid innovation for the support of the curation lifecycle of research data is under way (in the UK, this is being promoted by JISC, RIN, and the DCC[10]). As evidenced by the resources we cite from longstanding social science data archives (for example, ICPSR, DANS, and the UKDA[11]), expertise in data archiving has been evolving since the 1960s. This guide seeks to bring the best practices and standards employed by these organisations into the context of the accelerating development of institutional repositories.

The DISC-UK DataShare project's interactions between data support specialists and their institutional repository colleagues informed the production of the guide. We propose that this collaborative effort across and among institutions can be a model for future development as we move from discussion to full implementation of policies and practices related to bringing data into repository environments. Among the project's findings are the importance of policy for data management, as well as shared tools and strategies that support solutions to the many challenges of research data stewardship and persistent access.

The guide can move forward conversations and policy development for the range of questions that emerge as institutions plan digital services for data management and sharing. Please consult the references for further authoritative sources of information and advice.

10 Joint Information Systems Committee, Research Information Network, Digital Curation Centre.

11 Inter-University Consortium for Political and Social Research at the University of Michigan, Data Archiving and Networked Services in the Netherlands, UK Data Archive at the University of Essex.

References

ARTS AND HUMANITIES DATA SERVICE (2004a) *AHDS Guides to Good Practice*. London: AHDS. Available from: http://ahds.ac.uk/creating/guides/index.htm

ARTS AND HUMANITIES DATA SERVICE (2004b) *How to Deposit with the AHDS*. London: AHDS. Available from: http://ahds.ac.uk/depositing/how-to-deposit.htm

AUSTRALIAN NATIONAL UNIVERSITY (2008) *ANU Data Management Manual: Managing Digital Research Data at the Australian National University, v. 1.03*. Canberra : ANU. Available from: http://ilp.anu.edu.au/dm/ANU_DM_Manual_v1.01.pdf

BEEDHAM, H., MISSEN, J., PALMER, M. and RUUSALEPP, R. (2005) *Assessment of UKDA and TNA Compliance with OAIS and METS Standards*. Essex: UK Data Archive. Available from: http://www.jisc.ac.uk/media/documents/programmes/preservation/oaismets.pdf

CARR, L., WHITE, W., MILES, S. and MORTIMER, B. (2008) Institutional Repository Checklist for Serving Institutional Management, Version 0.2. OR2008, Third International Conference on Open Repositories, Southampton 1-4 April 2008. Southampton: University of Southampton. Available from: http://pubs.or08.ecs.soton.ac.uk/138/1/IRChecklist.pdf

CONSULTATIVE COMMITTEE FOR SPACE DATA SYSTEMS (2002). *Reference Model for an Open Archival Information System (OAIS).* Washington DC: CCSDS. Available from: http://public.ccsds.org/publications/archive/650x0b1.pdf

CREATIVE COMMONS (2008) *License your work*. Massachusetts: Creative Commons. Available from: http://creativecommons.org/about/license/

DATA ARCHIVING AND NETWORKED SERVICES (2008) *Data Seal of Approval*. The Hague: DANS. Available from: http://www.datasealofapproval.org/

DIGITAL CURATION CENTRE (2006) Glossary of terms. Edinburgh: DCC. Available from: http://www.dcc.ac.uk/resource/glossary/

DIGITAL CURATION CENTRE (2008) *The DCC Curation Life Cycle Model*. Edinburgh: DCC. Available from: http://www.dcc.ac.uk/docs/publications/DCCLifecycle.pdf

References

DIGITAL PRESERVATION COALITION (2008) P*reservation Management of Digital Materials: The Handbook* (section 5.5). York: DPC. Available from: http://www.dpconline.org/graphics/medfor/recommendations.html

DULONG DE ROSNAY, M. (2008) *Check Your Data Freedom: A Taxonomy to Assess Life Science Database Openness*. London: Nature Precedings. Available from: http://dx.doi.org/10.1038/npre.2008.2083.1

ECONOMIC AND SOCIAL DATA SERVICE (2008A) *Depositing Data*. ESDS: University of Essex and University of Manchester. Available from: http://www.esds.ac.uk/aandp/create/depintro.asp

ECONOMIC AND SOCIAL DATA SERVICES (2008b) *End User License*. ESDS: University of Essex and University of Manchester. Available from: http://www.esds.ac.uk/aandp/access/licence.asp

ePRINTS SOTON (2006) *University of Southampton Research Repository - Repository Policies*. Southampton: University of Southampton. Available from: http://eprints.soton.ac.uk/repositorypolicy.html

GOLD, A. (2007) Cyberinfrastructure, Data, and Libraries, Part 1: A Cyberinfrastructure Primer for Librarians. *D-Lib Magazine* [online], 13 (9/10). Available from: http://www.dlib.org/dlib/september07/gold/09gold-pt1.html

INTER-UNIVERSITY CONSORTIUM FOR POLITICAL AND SOCIAL RESEARCH (2005) Guide to Social Science Data Preparation and Archiving: Best Practice Throughout the Data Life Cycle, 3rd ed. Ann Arbor: ICPSR. Available from: http://www.icpsr.umich.edu/ICPSR/access/dataprep.pdf

INTER-UNIVERSITY CONSORTIUM FOR POLITICAL AND SOCIAL RESEARCH (2007a) Guidelines for Depositing Data. Ann Arbor: ICPSR. Available from: http://www.icpsr.umich.edu/ICPSR/access/deposit/guidelines.html

INTER-UNIVERSITY CONSORTIUM FOR POLITICAL AND SOCIAL RESEARCH (2007b) Levels of Access to Data: Public and Restricted. Ann Arbor: ICPSR. Available from: http://www.icpsr.umich.edu/ICPSR/access/restricted/index.html

INTER-UNIVERSITY CONSORTIUM FOR POLITICAL AND SOCIAL RESEARCH (2007c) *Confidentiality Review*. Ann Arbor: ICPSR. Available from: http://www.icpsr.umich.edu/access/deposit/conf-review.html

JOHNSON, R. K. (2002) Institutional Repositories: Partnering with Faculty to Enhance Scholarly Communication. *D-Lib Magazine*. 8 (11). Available from: http://www.dlib.org/dlib/november02/johnson/11johnson.html

References

LEWIS, M. (2008) University Libraries in the UK Data Curation Landscape, Keynote. *4th International Digital Curation Conference*, Edinburgh 1-3 December 2008. Edinburgh: DCC. Available from: http://www.dcc.ac.uk/events/dcc-2008/programme/presentations/0840_Martin_Lewis.ppt

LYON, L. (2007) *Dealing with Data: roles, rights, responsibilities and relationships*, Consultancy Report. Bath: UKOLN. Available from: http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/reports/dealing_with_data_report-final.pdf

MACDONALD, S. (2008a) Data Visualisation Tools: Part 1 - Numeric Data in a Web 2.0 Environment. DISC-UK, January 2008. Available from: http://www.disc-uk.org/docs/Numeric_data_mashup.pdf

MACDONALD, S. (2008b) Data Visualisation Tools: Part 2 - Spatial Data in a Web 2.0 Environment and Beyond. DISC-UK, September 2008. Available from: http://www.disc-uk.org/docs/spatial_data_mashup_V2.pdf

MARTINEZ, L. (2008) The Data Documentation Initiative (DDI) and Institutional Repositories. DISC-UK, February 2008. Available from: http://www.disc-uk.org/docs/DDI_and_IRs.pdf

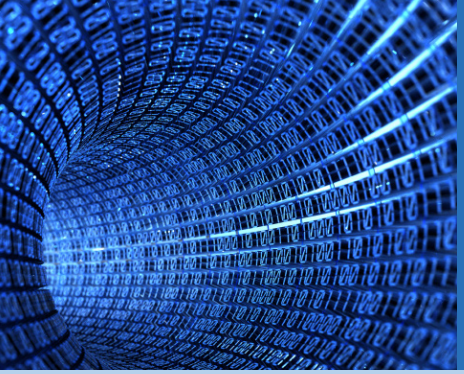
MARTINEZ-URIBE, L. (2008) *Findings of the Scoping Study and Research Data Management Workshop: Scoping Digital Repository Services for Research Data Management*. Oxford: University of Oxford, Office of the Director of IT. Available from: http://www.ict.ox.ac.uk/odit/projects/digitalrepository/findings.xml

MASSACHUSETTS INSTITUTE OF TECHNOLOGY LIBRARIES (2009) *Data Management and Publishing*. Cambridge, MA: MIT Libraries. Available from: http://libraries.mit.edu/guides/subjects/data-management/

MASSACHUSETTS INSTITUTE OF TECHNOLOGY LIBRARIES (2002) *DSpace Format Support*. Cambridge, MA: MIT Libraries. Available from: http://libraries.mit.edu/dspace-mit/build/policies/format.html

MCGEEVER, M (2007) IPR in Databases. Edinburgh: DCC, Oct. 2007. Available from: http://www.dcc.ac.uk/resource/legal-watch/ipr-in-databases/

MCGEEVER, M (2009) Science Commons. Edinburgh: DCC, Mar. 2009. Available from: http://www.dcc.ac.uk/resource/legal-watch/science-commons/



References

MOORE, R. W., RAJASEKER, A., and MARCIANO, R. (2007) Implementing Trusted Digital Repositories. *DigCCurr2007 International Symposium in Digital Curation*, Chapel Hill, North Carolina, April 2007. Available from: https://www.irods.org/pubs/DICE_DigcCur-Trusted-Rep-07.pdf

MÜLLER, H (2009) *Database Archiving*. Briefing Paper. Edinburgh: DCC. Available from: http://www.dcc.ac.uk/resource/briefing-papers/database-archiving/

ONLINE COMPUTER LIBRARY CENTER (2007) Trustworthy Repositories Audit & Certification: Criteria and Checklist. Version 1.0. Dublin, OH: OCLC. Available from: http://www.crl.edu/PDF/trac.pdf

OPEN KNOWLEDGE FOUNDATION (2009) *Comprehensive Knowledge Archive Network*. Cambridge: Open Knowledge Foundation. Available from: http://ckan.net/

OXFORD UNIVERSITY RESEARCH ARCHIVE (2008) *ORA Policies*. Oxford: University of Oxford. Available from: http://www.ouls.ox.ac.uk/ora/ora_documents2/ora_policies

PARADIGM PROJECT (2007) *Metadata for Authenticity: Hash Functions and Digital Signatures*. Universities of Oxford and Manchester. Available from: http://www.paradigm.ac.uk/workbook/metadata/authenticity.html

PURDUE UNIVERSITY LIBRARIES (2007) D2C2, *Distributed Data Curation Center.* Purdue University Libraries, West Lafayette, IN, USA. Available from: http://d2c2.lib.purdue.edu/

REPOSITORIES SUPPORT PROJECT (2008) *Repository Policy Framework*, Briefing Paper. Bristol: JISC. Available from: http://www.rsp.ac.uk/pubs/briefingpapers-docs/repoadmin-policyv2.pdf

RESEARCH INFORMATION NETWORK (2008) *Stewardship of digital research data - principles and guidelines*. London: RIN. Available from: http://www.rin.ac.uk/data-principles

RURAL ECONOMY AND LAND USE PROGRAMME DATA SUPPORT SERVICE (2006) Guidance on Data Management. Essex: University of Essex. Available from: http://www.data-archive.ac.uk/relu/reluaug2006.pdf

RURAL ECONOMY AND LAND USE PROGRAMME DATA SUPPORT SERVICE (2009) *Data Management Plan*. Essex: University of Essex. Available from: http://www.data-archive.ac.uk/relu/plan.asp

References

SCIENCE COMMONS. *Submit your database policy*. San Francisco: Science Commons. Available from: http://shirleyfung.com/mbdb/submit.php

SHERPA. (2007) *OpenDOAR Policies Tool*. Nottingham: University of Nottingham. Available from: http://www.opendoar.org/tools/en/policies.php

SMITH, M. and MOORE, R. (2006) Digital Archive Policies and Trusted Digital Repositories - *2nd International Digital Curation Conference Digital Data Curation in Practice*, Glasgow, 21–22 November 2006. Glasgow: DCC. Available from: http://pledge.mit.edu/images/6/6f/Smith-Moore-DCC-Nov-2006.pdf

TRELOAR, A., GROENEWEGEN, D., and HARBOE-REE, C. (2007) The Data Curation Continuum: Managing Data Objects in Institutional Repositories. *D-Lib Magazine*, 13 (9/10). Available from: http://www.dlib.org/dlib/september07/treloar/09treloar.html

UK DATA ARCHIVE (2008a) Data Formats and Software. Essex: University of Essex. Available from: http://www.data-archive.ac.uk/sharing/formats.asp

UK DATA ARCHIVE (2008b) *Manage and Share Data - Authenticity and Versioning of Data*. Essex: University of Essex. Available from: http://www.data-archive.ac.uk/sharing/version.asp

UK DATA ARCHIVE (2008c) *Manage and Share Data - Data Documentation and Metadata*. Essex: University of Essex. Available from: http://www.data-archive.ac.uk/sharing/metadata.asp

UK DATA ARCHIVE (2008d) *Manage and Share Data - Consent, Confidentiality and Ethics*. Essex: University of Essex. Available from: http://www.data-archive.ac.uk/sharing/confidential.asp

US NATIONAL LIBRARY OF MEDICINE (2007) *Digital Repository Policies and Functional Requirements Specification. Prepared by the NLM Digital Repository Working Group, Version 1*. Bethesda, MD: USNLM. Available from: http://www.nlm.nih.gov/digitalrepository/NLM-Digital-Repository-Requirements-rev032007.pdf

US NATIONAL SCIENCE BOARD (2005) *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century.* Arlington, VA: National Science Foundation. Available from: http://www.nsf.gov/pubs/2005/nsb0540/

UNIVERSITY OF OXFORD (2005). *Metadata in the Oxford University Digital Library*. Oxford: University of Oxford. Available from: http://www.odl.ox.ac.uk/metadata.htm

WITT, M. (2008) *Institutional Repositories and Research Data Curation in a Distributed Environment*. Baltimore: Library Trends. Available from: http://muse.jhu.edu/journals/library_trends/v057/57.2.witt.html

EDINA and University Data Library
Information Services Division
University of Edinburgh
160 Causewayside
Edinburgh, EH9 1PR
Scotland, United Kingdom

EDINA home page: http://edina.ac.uk
Email: edina@ed.ac.uk
Phone: +44 (0)131 650 3302

Data Library home page: http://datalib.ed.ac.uk
Email: datalib@ed.ac.uk
Phone: +44 (0)131 651 1431 / 651 1744