



KEEPING RESEARCH DATA

SAFE 2

REVIEW AND UPDATES OF KRDS1 ACTIVITY MODEL

Neil Beagrie, Brian Lavoie and Matthew Woollard

with contributions by the Universities of Cambridge, Oxford, and Southampton, the Archaeology Data Service University of York, and University of London Computer Centre.

Review Draft 31 July 2009

Prepared by:

Charles Beagrie Limited

www.beagrie.com

A study funded by

The logo for JISC, consisting of the letters 'JISC' in a bold, orange, sans-serif font.

Copyright HEFCE 2009. The authors have asserted their moral rights in this work

Comments on this draft can be sent to info@beagrie.com

REVIEW OF THE KRDS1 ACTIVITY MODEL

INTRODUCTION

The Keeping Research Data Safe2 (“KRDS2”) project aims to build on previous work on digital preservation costs for research data contained in the first Keeping Research Data Safe (“KRDS1”) report (Beagrie et al 2008).

All of our project partners undertook a detailed review of the activity model published in KRDS1 against their existing preservation activities and had an opportunity to suggest potential changes or areas of difficulty in the published model. The overall finding from this review was that the KRDS1 Activity Model was robust and broadly a good fit to their activities. Some changes were suggested, mainly to the wordings of definitions and edits to the existing text.

One specific area of concern for some was the use of OAIS terminology and its potential for acting as a barrier to understanding for some potential user groups. After discussion it was decided that the original justification for use of OAIS terminology where appropriate in KRDS still stood. OAIS terms are well-defined, published, and well-established in the preservation community. However, we believe it will be important for the wording of the activity table to be reviewed and adapted as needed locally by users for their intended audience and specific application.

In addition, three substantive changes or additions to activities were also identified by two or more reviewers and agreed as changes to the model for KRDS2:

- **The need to divide the “outreach and depositor support” sub-activity under Acquisition in the Archive phase in KRDS1.** Several national services reported that outreach providing data management advice was a significant activity for those charged with supplying advice and guidance to researchers preparing grant proposals in the pre-archive phase. A high percentage of these proposals would not be funded and would therefore not generate deposits. Other data producers than researchers could also be a significant target community for outreach. Similar concerns were raised by a university partner establishing a central support service for its researchers where outreach working with researchers from the moment they

create their datasets to ensure that appropriate preservation actions are taken early in the research life cycle; and audits to understand the research data management requirements of their research groups, will be crucial pre-archive phase activities. It was therefore agreed to introduce a new “Outreach” activity in the pre-archive phase and change the sub-activity under Acquisition to “depositor support” and amend definitions accordingly.

- **The need to divide the development of the archive’s Selection Policy and its application within the selection sub-activity of Acquisition.** Several reviewers pointed out the development of a selection policy is episodic as a cost and best separated out from the day-to-day application of policy. We have therefore inserted a new sub-activity for “develop policy and standards” under the administration activity and amended the selection sub-activity accordingly.
- **The need to cover staff training and development as a specific activity.** We have therefore inserted a new sub-activity for staff training and development under Common Services.

All changes to the KRDS1 activity model for KRDS2 are shown as tracked changes in section 4.2 with the new sub-activities shown with a blue background.

1.1. REVISIONS TO THE ACTIVITY MODEL

KRDS2 ACTIVITY MODEL	
Attribute	Scope Notes & <i>[source]</i>
Pre-Archive Phase	Primarily relates to research projects in universities creating research data for later transfer to a data archive. However activities can be adapted for first stages in piloting and development of a new data archive if required. <i>[Study Team]</i>
<u>Outreach</u>	<u>Guidance on best practice and archiving requirements and other support and training by the archive for researchers submitting funding proposals or creating research data. This may be targeted at potential depositors and/or broader communities and data</u>

	<u>producers.</u>
Initiation	Included to note any significant implications for preservation costs downstream. <i>[Study Team]</i>
Project design	Take into account implications of any data creation or acquisition activity including data formats; metadata; volume and number of files, etc. <i>[Study Team]</i>
Data management plan	Should include plans for future preservation and data sharing. <i>[Study Team]</i>
Funding application	Include <u>Full Economic Cost (FEC)</u> elements including activity relevant to preparation for preservation where applicable. <i>[Study Team]</i>
Project implementation	Allows for ramping up and staff investment in project starting-up activity. The project must define an 'implementation period' over which the implementation effort and cost are estimated. <i>[NASA CET]</i>
Creation	Included to note any significant implications for preservation costs or archive access/use downstream. <i>[Study Team]</i>
negotiate IPR/licensing/ethics	These need to be dealt with at the earliest stages <u>by the data creator</u> so that when data is <u>deposited/accepted</u> into an archive there are no residual issues around IPR, licensing, or ethics. These can be very difficult to resolve at a later stage. <u>Guidance on IPR, licensing and ethics may be available from the archive or funder to assist in this.</u> This is important because an archive, as custodian, will honour all applicable legal restrictions. An archive should understand the copyright concepts and applicable laws prior to accepting copyright materials into the archive. It can establish guidelines for ingestion of information and rules for dissemination

	and duplication of the information when necessary. [Study Team OAIS-RM]
generate research data	Conceive and plan the creation of both raw and derived data created throughout the duration of the project, including capture method and storage options. <i>[Study Team]</i>
generate descriptive metadata	Generating This function extracts the Descriptive Information for research data. This will form part of the Archival Information Package deposited with the Archive at a later stage from the Archival Information Packages (AIPs) and collects Descriptive Information from other sources to provide to Coordinate Updates, and ultimately Data Management. This includes metadata to support searching and retrieving AIPs (e.g., who, what, when, where, why), and could also include special browse products (thumbnails, images) to be used by Finding Aids. [Study Team OAIS-RM]
generate user documentation	The producer of the data needs to take into account whether users outside of the project may access the data and document accordingly. <i>[Study Team]</i>
generate customised software	This includes custom interfaces and applications if required. Such software will require specification, testing and implementing and include detailed documentation. Standardising on a set of supported software will be more cost effective and should be encouraged. <i>[Study team]</i>
Data management	Services and functions for populating, maintaining, and accessing a wide variety of data by the project. <i>[OAIS RM]</i>
create submission package for archive	Format/contents and the logical constructs used by the Producer and how they are represented on each media delivery or in a telecommunication session. Submission Information Package

	(SIP): An Information Package that is delivered by the Producer to the <u>archive</u> OAIS for use in the construction of one or more Archival Information Packages. <i>[OAIS RM]</i>
Archive Phase	
Acquisition	In LIFE model but not in OAIS reference model, apart from negotiate submission agreement. <i>[Study Team]</i>
Selection	The <u>application</u> development of the <u>archive's</u> Selection Policy and its application. <i>[Study Team LIFE]</i>
negotiate submission agreement	The specification of submission requirements for producers/depositors together with communication and negotiation of <u>submission agreements</u> with producers/depositors. <i>[Study Team LIFE]</i>
outreach and depositor support	Support and training for researchers submitting funding proposals that include creating research data, and <u>support and encouragement for researchers and others</u> with data to deposit. <i>[Study Team]. N.B. Poorly captured in most other models – probably equivalent to technical co-ordination in NASA GET</i>
Disposal	Poorly captured in most other models and added by the study team – destroy is also in draft DGC curation lifecycle model (Higgins 2007). <i>[Study team]</i>
transfer to another archive	Transfer material to an archive, repository, data centre or other custodian. Adhere to documented guidance, policies or legal requirements. <i>[Study Team].</i>
destroy	Destroy material which has not been selected for long-term curation and preservation. Documented policies, guidance or legal requirements may require that this be done securely. <i>[Study Team].</i>
Ingest	The ingest functional area includes receiving, reading, quality checking, cataloging, of incoming data (including metadata,

	documentation, etc.) to the point of insertion into the archive. Ingest can be manual or electronic with manual steps involved in quality checking, etc. <i>[NASA CET] & [OAIS]</i>
receive submission	This provides the appropriate storage capability or devices to receive a submission of data. Submissions may be digital delivered via electronic transfer (e.g., FTP), loaded from media submitted to the archive, or simply mounted (e.g., CD-ROM) on the archive file system for access. Non-digital submissions would likely be delivered by conventional shipping procedures. The Receive Submission function may represent a legal transfer of custody for the Content Information and may require that special access controls be placed on the contents. This function provides a confirmation of receipt to the Producer, which may include a request to resubmit in the case of errors resulting from the submission. <i>[OAIS RM]</i>
quality assurance	The Quality Assurance function validates (QA results) the successful transfer of the data submission to the staging area. For digital submissions, these mechanisms might include Cyclic Redundancy Checks (CRCs) or checksums associated with each data file, or the use of system log files to record and identify any file transfer or media read/write errors <i>[OAIS RM]</i> . In addition to these basic integrity checks, it may also include many more discipline-specific tests on the quality of data and metadata.
generate Information Package for Archive	This deals with the transformation of the submitted data (or information package) into a format suitable for the archive. Archival Information Packages within the system will conform to the archive's data formatting and documentation standards. This may involve file format conversions, <u>redaction, disclosure checking</u> , data representation conversions or <u>other</u> reorgan ^{is} zation of the content

	<p>information.</p> <p>[modified from OAIS RM]</p>
generate administrative metadata	<p>Metadata about the preservation process:</p> <ul style="list-style-type: none"> • pointers to earlier versions of the collection item • change history <i>[OAIS RM]</i>
generate/upgrade descriptive metadata and user documentation	<p>Includes the development (or upgrading of received) data and product documentation (including user guides, catalogue interfaces, etc.) to meet adopted documentation standards, including catalogue information (metadata), user guides, etc., through consultation with data providers. <i>[NASA CET]</i></p>
co-ordinate updates	<p>Provides a mechanism for updating the contents of the archive. It receives <i>change requests, procedures</i> and <i>tools</i> from Manage System Configuration. <i>[OAIS RM]</i></p>
reference linking	<p>The <u>semantic</u> linking of primary data to textual interpretations of that data. Pioneering projects such as JISC-funded eBank, have demonstrated that this is a very powerful and valuable feature. It is now being explored by a number of other JISC-funded repository projects such as SPECTRa¹, GLADDIER² and a joint follow-on project, StoreLink³. There is also some evidence that such virtual links may facilitate real connections between physical services i.e. between data centres and institutional repositories in libraries <i>[Study Team]</i>.</p>
Archive Storage	<p>Services and functions used for the storage and retrieval of Archival Information Packages (AIPs). <i>[OAIS RM]</i></p>

<p>receive data from ingest</p>	<p>The Receive Data function receives a <i>storage request</i> and an <i>AIP</i> from Ingest and moves the <i>AIP</i> to permanent storage within the archive. This function will select the media type, prepare the devices or volumes, and perform the physical transfer to the Archival Storage volumes. [OAIS RM]</p>
<p>manage storage hierarchy</p>	<p>The Manage Storage Hierarchy function positions, via <i>commands</i>, the contents of the <i>AIPs</i> on the appropriate media based on <i>storage management policies</i>, operational statistics, or directions from Ingest via the storage request. It will also conform to any special levels of service required for the <i>AIP</i>, or any special security measures that are required, and ensures the appropriate level of protection for the <i>AIP</i>. [OAIS RM]</p>
<p>replace media</p>	<p>This provides the capability to reproduce the Archival Information Packages (<i>AIPs</i>) over time. [OAIS RM]</p>
<p>disaster recovery</p>	<p><u>Disaster recovery is the process, policies and procedures related to preparing for recovery or continuation of technology infrastructure critical to an organisation after a natural or human-induced disaster. Disaster recovery planning should include planning for resumption of applications, data, hardware, communications (such as networking) and other IT infrastructure. It is a subset of a larger process known as business continuity planning that includes planning for non-IT related aspects such as key personnel, facilities, and crisis communication. It should provide a plan for and testing of mechanisms</u> for duplicating the digital contents of the archive collection and storing the duplicate in a physically separate facility <u>and recovery from them</u>. This function is normally accomplished by copying the archive contents to some form of removable storage media (e.g., digital linear tape, compact disc),</p>

	but may also be performed via hardware transport or network data transfers. The details of <i>disaster recovery policies</i> are specified by Administration. <i>[Study Team and OAIS RM]</i>
Error checking	Provides statistically acceptable assurance that no components of the <i>AIP</i> are corrupted during any internal Archival Storage data transfer. It requires that all hardware and software within the archive provide <i>notification of potential errors</i> and that these errors are routed to standard <i>error logs</i> that are checked by the Archival Storage staff. <i>[OAIS RM]</i>
provide copies to access	The archive design will reference the preservation strategy and policy, considering off-site copies and any discipline requirement for multiple versions or editions. The number of versions and copies affects storage and management costs. <i>[Study Team]</i>
Preservation Planning	The services and functions for monitoring, providing recommendations, and taking action, to ensure that the information stored in the archive remains accessible over the long term, even if the original computing environment becomes obsolete. <i>[Study Team modified from OAIS RM]</i>
monitor designated user community	The Monitor Designated Community function interacts with archive Consumers and Producers to track changes in their <i>service requirements</i> and available <i>product technologies</i> . Such requirements might include data formats, media choices, and preferences for software packages, new computing platforms, and mechanisms for communicating with the archive. <i>[OAIS RM]</i>
monitor technology	The Monitor Technology function is responsible for tracking emerging digital technologies, information standards and computing platforms (i.e., hardware and software) to identify technologies which could cause obsolescence in the archive's

	<p>computing environment and prevent access to some of the archives current holdings. <i>[OAIS RM]</i></p>
<p>develop preservation strategies and standards</p>	<p>The Develop Preservation Strategies and Standards function is responsible for developing and recommending strategies and standards to enable the archive to better anticipate future changes in the Designated Community service requirements or technology trends that would require migration of some current archive holdings or new submissions. <i>[OAIS RM]</i></p>
<p>develop packaging designs and migration plans</p>	<p>The Develop Packaging Designs and Migration Plans function develops new <u>Information Package</u> designs and detailed migration plans and prototypes. This activity also provides advice on the application of these <u>Information Package</u> designs and Migration plans to specific archive holdings and submissions. <i>[OAIS RM]</i></p>
<p>develop and monitor SLAs for outsourced preservation</p>	<p>Where a decision is made to outsource some or all archive functions a contractual relationship will be established and to ensure service requirements are understood and met a Service Level Agreement needs to be put in place and monitored. Not in other models. <i>[Study Team]</i></p>
<p>preservation action</p>	<p>Preservation Action covers the process of performing actions on digital objects in order to ensure their continued accessibility. It includes evaluation and quality assurance of actions, and the acquisition or implementation of software to facilitate the preservation actions <i>[LIFE]</i>. Preservation has a feedback loop back into/through Ingest functions in activity model. <i>[Study Team]</i></p>
<p>generate preservation metadata</p>	<p><u>the information an archive uses to support the digital preservation process. Specifically, the metadata supporting the functions of maintaining viability, renderability, understandability, authenticity, and identity in a preservation context. Preservation metadata thus spans a number of the categories typically used to differentiate types of metadata: administrative (including rights and</u></p>

	<p><u>permissions), technical, and structural. The documentation of digital provenance (the history of an object) and to the documentation of relationships, especially relationships among different objects within the archive.[adapted from PREMIS]</u></p>
<p>First Mover Innovation</p>	<p>Where preservation functions and file formats are evolving a high-degree of R&D-expenditure might be required in implementation phases and in developing the first tools, standards and best practices. This cost is highly variable for individual institutions and significantly dependent on how much is done solely by the institution or by a wider community. Communities or vendors can make significant up-front investments in first solutions and standards which affect downstream preservation costs. Most data archives participate in these activities to some degree although leadership and significant effort may be restricted to a few large institutions. Not in other models – added as has significant implications for cost modelling or potential for use/re-use. <i>[Study Team]</i></p>
<p>develop community data standards and best practice</p>	<p>Whilst preservation functions are evolving professional involvement in developing community standards and best practises is a cost effective approach to the delivery of efficient solutions. <i>[Study Team]</i></p>
<p>Share development of preservation systems and tools</p>	<p>Combining effort with others in the community can deliver significant developments for relatively small cost to individual institutions, and may even attract external funding. <i>[Study Team]</i></p>
<p>engage with vendors</p>	<p>This might include beta-testing, participation in user groups, and development of commercial partnerships. <i>[Study Team]</i></p>
<p>Data Management</p>	<p>The services and functions for populating, maintaining, and</p>

	accessing both descriptive information which identifies and documents archive holdings and administrative data used to manage the archive. <i>[OAIIS RM]</i>
administer database	Responsible for maintaining the integrity of the Data Management database, which contains both Descriptive Information and system information. Descriptive Information identifies and describes the archive holdings, and system information is used to support archive operations. <i>[OAIIS RM]</i>
perform queries	Receives a <i>query request</i> from Access and executes the query to generate a <i>result set</i> that is transmitted to the requester. <i>[OAIIS RM]</i>
generate report	Receives a <i>report request</i> from Ingest, Access or Administration and executes any queries or other processes necessary to generate the <i>report</i> that it supplies to the requester. Typical reports might include summaries of archive holdings by category, or usage statistics for accesses to archive holdings. <i>[OAIIS RM]</i>
receive database updates	Adds, modifies or deletes information in the Data Management persistent storage. The main sources of updates are Ingest, which provides <i>Descriptive Information</i> for the new AIPs, and Administration, which provides <i>system updates</i> and <i>review updates</i> . <i>[OAIIS RM]</i>
Access	Services and functions which make the archival information holdings and related services visible to Consumers. <i>[OAIIS RM]</i>
search and ordering	This includes providing access to catalogue information and a search and order capability to users, and receiving user requests for data. “Order” implies a request /permission step, regardless of how implemented (e.g. manual or automated), where a request for a set of data or product instances, perhaps the results of (or a

	selected subset of the results of) a search, is processed and accepted or denied. <i>[NASA CET]</i>
generate information package for dissemination to user	This function accepts a dissemination request, retrieves the Archival Information Package from Archival Storage, and moves a copy of the data to a staging area for further processing. The types of operations, which may be carried out, include statistical functions, sub-sampling in temporal or spatial dimensions, conversions between different data types or output formats, and other specialized processing. <i>[OAIS RM]</i> . <u>See also generate Information Package for Archive in Ingest – as some archives may generate archive and dissemination version simultaneously,</u>
deliver response	The Deliver Response function handles both on-line and off-line deliveries of responses (Delivery Information Packages, result sets, reports and assistance) to Consumers. <i>[OAIS RM]</i>
user support	The user support functional area includes support provided in direct contact with users by user support staff, including <u>training for users, user demonstrations,</u> responding to queries, taking of orders, staffing a help desk (i.e., staff awaiting user contacts who can assist in ordering, track and status pending requests, resolve problems, etc.), etc. User support staff includes science expertise to assist users in selecting and using data and products. <i>[adapted from NASA CET]</i> .
new product generation	Initial generation and reprocessing with quality checking of new data products produced from data or products previously ingested, or generated <i>[NASA CET]</i> . Note that this has as a feedback loop back into/through Ingest functions.
Support Services	

Administration	Services and functions needed to control the operation of the other functional entities on a day-to-day basis. <i>[OAIS RM]</i>
general management	Management includes management and administration at the data service provider level (“front office”) and direct management of functional areas. Management also includes staff with overall responsibility for internal and external science activities, information technology planning, and data stewardship. <i>[NASA CET]</i>
customer accounts	To facilitate billing and payment receipts from “customers”. Also useful for reporting usage and restricting access as appropriate to closed collections with specific license conditions. <i>[Study Team]</i>
Administrative support	Administrative support and control provided by office managers, personal assistants and secretaries. <i>[Study Team]</i>
<u>Develop policies and standards</u>	<u>This function is responsible for establishing and maintaining the archive's standards and policies. These include initial format standards, documentation standards, model deposit agreements, the archive's selection policy and the procedures to be followed during the Ingest process. They will normally involve a large initial effort to develop and then regular review and small updates over time and rarer major re-drafting.[adapted from OAIS RM]</u>
Common Services	These are the other shared supporting services supplied by the institution or located within the archive. <i>[Study Team]</i>
operating system services	Provide the core services needed to operate and administer the application platform, and provide an interface between application software and the platform. <i>[OAIS RM]</i>
network services	These provide the capabilities and mechanisms to support distributed applications requiring data access and applications

	interoperability in heterogeneous, networked environments. <i>[OAIS RM]</i>
network security services	Network security services include access, authentication, confidentiality, integrity, and non-repudiation controls and management of communications between senders and receivers of information in a network <i>[OAIS RM]</i>
software licences and hardware maintenance	Ensure that correct software licenses are in place and that they are renewed in a timely way. Also, determine the most appropriate level of hardware maintenance for the configuration and put in place call procedures and reporting with the supplier. Renew in a timely way. <i>[Study Team]</i>
physical security	With reference to facility and infrastructure. The service will have a Disaster Recovery Plan to deal with all eventualities and to mitigate risk. <i>[Study Team]</i>
utilities	Supply of uninterrupted power supply, air conditioning, water etc. <i>[Study Team]</i>
supplies inventory and logistics	Management of supply chain, movement of goods, and recording of purchases and deliveries. <i>[Study Team]</i>
<u>Staff training and development</u>	<u>Support for training or developing Archive staff to carry out particular roles. <i>[Study Team]</i></u>
Estates	Estates management and attendant costs includes leasing of premises, space management and maintenance. Treated as a cost element in TRAC separate from other common services and charged at variable rates according to function e.g. laboratory/non-laboratory <i>[Study Team]</i> .

REFERENCES

Beagrie, N., Chruszcz, J. and Lavoie, B., 2008, *Keeping Research Data Safe: a cost model and guidance for UK Universities*, (Joint Information Systems Committee 2008).

<http://www.jisc.ac.uk/publications/publications/keepingresearchdatasafe.aspx>