

Digital Archives for Archaeology and the Historic Environment – October 2011

1. Introduction

The DPC has been invited to facilitate a conversation between English Heritage, RCAHMS, RCAHMW and ADS to explore closer collaboration on the delivery of digital preservation services. Thoughtfully defined, closer collaboration could deliver advantages to all parties such as improvements in quality, more effective interoperability of processes, greater efficiencies in operation or economies in scale. Moreover, trusted preservation services benefit from peer scrutiny and transparency. So, even if closer collaboration proves to be impractical, these discussions may yet provide the basis for independent review and enhancement.

2. Preliminary Survey

In order to make the most efficient use of time, members have been asked to prepare a short overview of their digital preservation processes. This analysis is based on the functional areas of the Reference Model for an Open Archival Information System (OAIS), an ISO standard which assembles the building blocks of a long term preservation facility. The standard assumes the existence of six functional areas: Ingest, Archival Storage, Data Management, Administration, Planning and Access. A seventh area – Common Services – provides the local platform on which an OAIS is constructed. Each functional area in OAIS is composed of a group of components which work together to provide preservation services. Although the function of each component and the relationships between components are described, the standard offers only limited guidance on how any component is configured. From an organisational perspective, OAIS outlines a preservation architecture and common vocabulary, but it makes no assumption on whether the functional components are offered by a single agency or are shared between multiple partners, or outsourced. Therefore, even a preliminary mapping against OAIS enables agencies to plan how they might contract services, outsource them or collaborate as appropriate.

This document compiles all three responses received prior to the meeting from ADS, English Heritage and RCAHMS. It is intended as an informal guide to help the meeting rather than a comprehensive statement. Participants were asked to assess their own strengths and areas for improvement against each functional component of OAIS, to identify whether they think there is scope for collaboration in this area, to outline the sort of collaboration they think is possible and what they would hope to gain from it. Rather than presenting the surveys in order, answers for each functional area have been compiled together. It is hoped that this will help the meeting identify more quickly those areas where substantive discussion can be progressed, and eliminate those areas which are out of scope or unlikely to produce results.

3. Plan for the meeting

The meeting will be chaired by William Kilbride of the DPC and to encourage free and frank exchanges and participants are asked that the discussions be kept confidential.

The meeting will start with a question and answer session where each partner will be asked to present an element of the survey, then invite question and answer to clarify issues and opportunities. This element of the meeting will identify themes for the later part where participants will be asked to develop then present an action plan for collaboration on specific and clearly scoped areas, identifying also the benefits that are intended to accrue and the resources available to achieve them.

Proposals at the meeting will be written up by the DPC and shared among participants who will then be asked to make a formal response to the recommendations made, within a timescale agreed among partners. This gives members time to reflect on the process, discuss with their own stakeholders, and decide whether or not they are willing to progress any subsequent work.

4. More details

For more details on the OAIS standard, a good place to start is the DPC Technology Watch Report by Brian Lavoie of OCLC online at: http://www.dpconline.org/component/docman/doc_download/91-introduction-to-oais

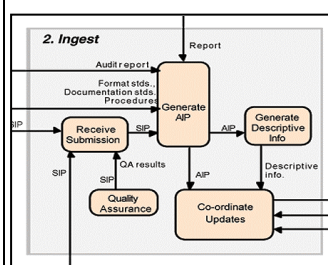
Survey of Digital Archive Functions

1. Introductions

Archaeology Data Service	English Heritage	RCAHMS
http://archaeologydataservice.ac.uk	http://www.english-heritage.org.uk/	www.rcahms.gov.uk
Stuart Jeffrey	Mike Evans	Emily Nimmo and Kirsty Lingstadt
Stuart.jeffrey@york.ac.uk	Mike.evans@english-heritage.org.uk	Emily.nimmo@rcahms.gov.uk ; kirsty.lingstadt@rcahms.gov.uk
Wednesday 12 th October	7/10/11	12.10.2011

2. Ingest

The 'Ingest' functional area is the interface between the OAIS and its producers and is primarily concerned with the taming of content and preparation for transfer into the archive. It includes five components: 'Receive Submission; Quality Assurance; Generate Archival Information Package; Generate Descriptive Information; and Co-ordinate Updates.



Please briefly describe the typical elements of your ingest processes?

We follow the OAIS reference model so cover all the elements above using our internal Collections Management System. Our ingest manual is available here:
<http://archaeologydataservice.ac.uk/advice/preservation/>

We have an ingest system targeted primarily at our own internal research teams:

- We get a submission spreadsheet from the photographic or research teams, to a specified format
- QA of information on spreadsheet carried out, but no QA of embedded metadata
- Descriptive information then developed in our catalogue (AMIE), within 6 weeks

- Negotiate submission
- Receive submission
- Virus check
- Transfer Data and Metadata to Temporary Storage Area
- Audit/Appraise Data and Metadata
- Complete Accession Record
- Transfer Data and Metadata to Digital Archive
- Generate and Send Acknowledgement Letter
- Create Catalogue Records with Digital Instances
- File Paperwork
- Store Original Media in Negative Room

Which parts of your ingest processes do you think are particularly strong?

The ADS puts a lot of effort into negotiation with depositors to ensure a well formed SIP – this is time consuming, but essential.

- Submission spreadsheet is robust and works reasonably well with depositors
- Good flow through to descriptive information, ensuring no backlog

- Negotiation with depositors
- Deposit agreements and licenses

Which parts of your ingest processes would you like to improve given the resources?

Streamlining of this process via a more automated ingest, particularly with regard to controlled vocabularies and data formats.

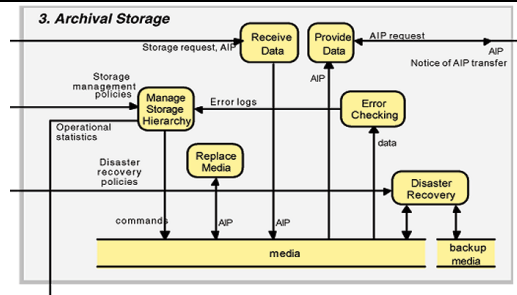
- Need to develop processes to cope with external deposits
- Increase amount of automatically generated metadata and reduce need for handcrafting
- Improve communication with depositors to establish proper

- Quarantine of data and metadata before virus check
- Creation and (automatic) checking of fixity values
- Transfer of metadata in a machine-readable format

	audit trail	<ul style="list-style-type: none"> Automation of file transfer, file structure, naming and migration to preservation format Automation of cataloguing
Are there any elements of the ingest process which you think would be improved through collaboration?		
Collaboration with LA and Museum Services have led to some streamlining of this process. However collaboration with other repositories in the form of the SIP would ultimately make it easier to aggregate resource discovery metadata (i.e. all deposits, wherever held, are formed in mutually intelligible and searchable forms).	<ul style="list-style-type: none"> Possibly development of solutions for automating metadata generation (both at submission point and as part of generating descriptive info) Standardisation of ingest requirements/formats could help encourage deposit by external depositors 	<ul style="list-style-type: none"> Sharing examples of best practice in relation to all areas of ingests process. Developing automation of processes Standardization of accepted file formats and required metadata
What would you hope to gain from such collaboration?		
Archive portability (a sustainability benefit) and potentially increased visibility of archives through external metadata aggregators.	<ul style="list-style-type: none"> Access to technical expertise Improvement of deposit across sector 	<ul style="list-style-type: none"> Benefit from experience of others, avoid issues they may have experienced when undertaking similar work. Economies to be found sharing resources to address issues relevant to all. Standardization of ingest requirements will facilitate more efficient negotiation and transfer from depositors and hopefully an increase in compliance.
Are there any elements of the ingest process which can only be carried out locally?		
Negotiation can currently only be done per repository as each has its own metadata requirements and archival storage procedures.	<ul style="list-style-type: none"> QA in any but the most technical sense benefits from a relationship with record creators and an understanding of their work. Generating descriptive info is for us integrated with our cataloguing of non-digital material, which has to be local 	<ul style="list-style-type: none"> Negotiation Appraisal Accessioning Cataloguing Storage

3. Archival Storage

The Archival Storage area is oriented around the management of robust storage, placing data on media, ensuring the integrity of data stored and recovering data from media as required. It includes six functional components: Receive Data, Provide Data, Error Checking, Disaster Recovery, Replace Media, Manage Storage.



Please briefly describe the typical operation of your Archival Storage?

All elements mentioned above are again covered by the ADS including data integrity (e.g. MD5), deep storage on and off site and a well formed disaster recovery plan.

Files are stored on a RAIDed disk array, managed by an off-the-shelf digital asset management system Portfolio v9.5. Files are also backed up on 2 sets of hard drives created at time of ingest.

There is no systematic system for error checking.

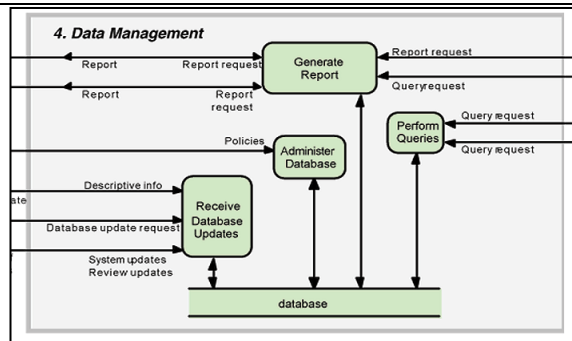
Disaster recovery is via our outsourced IT providers – but is based on the hard drives and a tape of the Portfolio indexes, not a tape of the whole system

- All digital archive material will be retained permanently on multiple drives within RCAHMS' network storage. This storage area is read-only to RCAHMS staff except those who are directly involved in accessioning and cataloguing the digital archive
- As part of pre accessioning of externally generated material this data is copied to a temporary archive location for evaluation. A temporary storage location for evaluation of internally generated RCAHMS material is provided on network storage so that it is accessible to creators and the digital archivist.
- A major upgrade of RCAHMS storage systems timetabled for 2012/13 that will enable the creation of 8Tb volumes
- The current storage solution (EVA SAN) is out of warranty in 2012 and will either need to be replaced or storage moved to the cloud
- A daily incremental backup is performed alongside a weekly and monthly full backup. Back-ups are written to disc array and tapes which are stored on site in a fire proof safe. Off-site storage is currently under negotiation.
- A rolling program of storage media re-refresh is performed over a 3-5 year timetable dependent on project pressures, budget and warranty expiration.

Which parts of your archival storage operations do you think are particularly strong?		
Deep storage is provided both by UoY systems and by the UKDA. The ADS are very committed to ensuring migration of the AIP to non-proprietary formats for long term preservation.	It has worked!?	<ul style="list-style-type: none"> • Access control • Back-up • Disaster recovery
Which parts would you like to improve given the resources?		
ADS are in the process of moving to a FEDORA based repository which has a number of technical advantages.	<ul style="list-style-type: none"> • The backup mechanism is not robust or fast. The system has survived a major failure of the raid array, but it took over a month to fully restore. I would like to move to full a mirrored system • System needs to be fully scalable to allow for ongoing growth • We need a data checking regime 	<ul style="list-style-type: none"> • We are investigating utilizing cloud storage for backup and off-site storage of tapes. • The planned upgrade will facilitate a separation of archival and dissemination copies of digital objects which will aid performance.
Are there any elements of the archival storage which you think would be improved through collaboration?		
Consensus on migration paths and preservation formats would improve alignment between repositories.	<ul style="list-style-type: none"> • Possibly sharing of backups to improve survivability • Possibly storage of large specialist file types and/or rarely used data 	Mirroring of data
What would you hope to gain from such collaboration?		
Repositories could cooperate on disaster recovery procedures.	<ul style="list-style-type: none"> • Improved survivability • Use of shared expertise to preserve particularly technical or complex data • More cost effective storage, near or off line 	Lowering the cost of externally supplied storage and minimizing the risks of any data loss through multiple copies.
Are there any elements of archival storage which can only be carried out locally?		
Local back-up, storage management and disaster recovery.	<ul style="list-style-type: none"> • QA in any but the most technical sense benefits from a relationship with record creators and an understanding of their work. • Generating descriptive info is for us integrated with our cataloguing of non-digital material, which has to be local 	At least one networked copy of data

4. Data Management

The Data Management functions are primarily concerned with ensuring that descriptive information about the contents of the archive is maintained and made available for internal administrative and reporting. In this context 'data' is distinguished from the information packages which the archive receives, preserves or distributes. This functional area is therefore oriented around a collection management database. It includes four components: Receive Database Updates, Administer Database, Perform Queries, and Generate Reports.



Please briefly describe the typical elements of data management in your preservation facilities?

The ADS has a well-developed CMS that allows us to perform all the stated data management functions, including the tracking of migration events.

Our cataloguing system AMIE is used as a collections management database in the sense of holding descriptions of archive packages, allowing queries and reports to be run

- Descriptive metadata identifying and describing the collection of archived material for RCAHMS both physical and digital is stored in the catalogue tables of an Oracle database.
- Metadata recorded is ISAD-G compliant and it is supplemented with additional 'Digital Instance' information relating to individual files
- The database is primarily structured around geographical location and also collection.

Which parts of your data management processes do you think are particularly strong?

The ADS CMS is fairly integrated in that it tracks collections, people (and organizations) and objects from the point at which a negotiation is opened with a depositor right through the preservation lifecycle. The CMS is also the (live) source of the majority of data used in the ADS delivery system, i.e. the website is generated dynamically from CMS content.

- We use robust data standards, which are common to other EH systems and help support cross searching of descriptive info
- Our query tools are powerful (but not very user friendly)
- Our cataloguing systems integrate descriptive records for digital and non-digital material

Robust, well-structured database system adhering to established standards.

Which parts of your data management processes would you like to improve given the resources?

Closer integration between CMS and delivery – particularly with regard to the creation of web services (this is in train). There is also the potential for conversion to LD formats for appropriate datasets – but perhaps this is best pushed back to the ingest/negotiation function

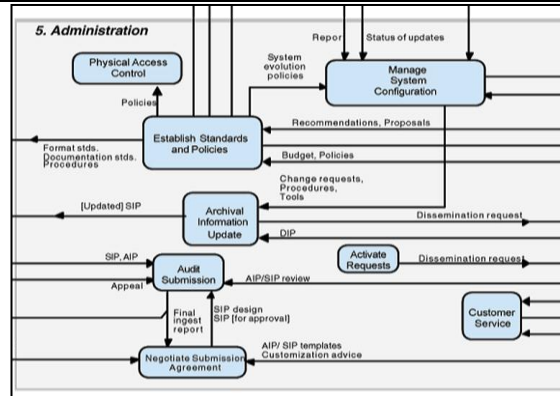
AMIE and the DAMS Portfolio are not properly integrated, leading to them falling out of step, the double handling of information, etc etc. They need to work much more closely together.

Increased commonality between descriptive datasets, data annotation and improvement.

Are there any elements of the data management process which you think would be improved through collaboration?		
No answer	No answer	<p>Aligning practices to facilitate interoperability where this could enhance user interaction with collections.</p> <p>Utilizing the OASIS submission process to capture user generated metadata. This could expedite ingest and cataloguing of digital materials and eliminate possible duplication of effort for our depositors.</p>
What would you hope to gain from such collaboration?		
This would be contingent on a more automated ingest system and consensus on AIP form and other data standards, I think.		<p>Enhanced user and depositor experience</p> <p>Efficiencies in time and/or money.</p>
Are there any elements of your data management processes which can only be carried out locally?		
Even with full collaboration on data standards, automated ingest and functions such as disaster recovery – part of the definition of a repository would be its ability to carry out some degree of data management locally.	Specialist nature of subject matter and integration with non-digital holdings argue for local processes	Some scope for data upgrading to be carried out externally through an online interface (see SURE project), other elements must be managed within RCAHMS.

5. Administration

The Administration functions ensure that the OAIS remains aligned with the goals of the agencies which sponsor it. It is a relatively complex area and interfaces with technology and resources as well with the administrative relationships an archive is required to develop with consumers and producers. It includes eight components: Physical Access Control, Establish Standards and Policies, Manage System Configuration, Archival Information Update, Audit Submission, Negotiate Submission Agreement, Activate Requests and Customer Service.



Please briefly describe the typical elements of Administration in your preservation facilities?

As above, the ADS administers ingest and data management via its CMS and associated procedures.

- Our standards and policies are incomplete and the process for managing them currently adhoc and not properly integrated with wider EH management structures.
- Reasonable system for controlling access to files by allowing users different levels of access using passwords
- We have good mechanisms for negotiating submission agreements through a dedicated flowlines post – but primarily for internal deposits as previously noted.

The administrative function of our digital archive is performed by the Digital Archivist reporting to the Operational Manager for collections and Head of collections, alongside a steering committee for the development of a Trusted Digital Repository. The Digital Archive Policy sets out objectives and responsibilities. The Digital Archivist maintains open lines of communication with external depositors, undertaking user education and facilitating negotiation of submission agreements alongside implementing and maintaining archive policies and standards.

Which parts of your Administration do you think are particularly strong?

It is a real strength of the ADS that it has well developed administration policies and process documentation. It is also a strength that this is all made freely available to depositors/potential auditors (e.g. DSA). The ADS are very involved in data standards development nationally and internationally - not necessarily a core repository role, but essential where this is not being done by other bodies. We make great efforts to adopt and follow existing standards where they exist. We are also involved more generally in archiving policy (e.g. AAF, EAC) Being University based is a real advantage for the ADS, both in terms of international contacts, but also in terms of access to

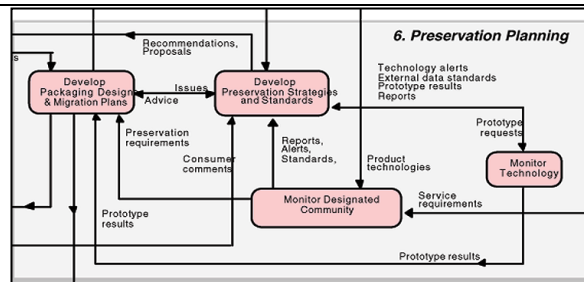
Negotiation of submission agreements and policies around this

Oversight, inclusion in RCAHMS' strategic plan and digital archive policy.

national and international research funding to develop procedures and infrastructure.		
Which parts of your Administration would you like to improve given the resources?		
Customer service – in the sense that the repository should not rely on its own access points, but engage fully with the wider range of organizations, projects and initiatives which provide access to distributed data.	We need to put more structured and systematic management of the system in place – to support both adoption of standards and policies and a joined up approach to systems configuration	<ul style="list-style-type: none"> • Greater granularity in policies and standards. • Clearer definition of management structure and budget.
Are there any elements of the Administration which you think would be improved through collaboration?		
A real need for agreement on licensing agreements arises when thinking about sharing data (even metadata). Since the inception of the ADS the licensing environment has moved on and a sector wide consensus on the use of, say, CC licensing, at least for metadata would be very useful and avoid confusion for users.	A common approach to core standards and policies	<ul style="list-style-type: none"> • Sharing examples of best practice, examples of policy/planning documents. • Greater consistency across policies and standards of participating organizations.
What would you hope to gain from such collaboration?		
Clarity for users (and also for depositors and repositories!)	A consensus would help with making the case locally for resources to support this.	Benefit from experience of others, avoid issues they may have experienced when undertaking similar work. Help to eliminate the duplication of effort and also ensure the most robust policies possible through critical peer review and wider range of expertise in-putting in the process.
Are there any elements of Administration which can only be carried out locally?		
As with data management there is a core subset of administrative functions that need to be carried out locally – this is especially true without the sector wide consensus on data standards, formats and licensing mentioned above.	Establishment of standards and policies has to align with local business priorities to get management buy –in	At this time all of the administration function can only be carried out locally, although there is scope for collaboration as outlined above.

6. Preservation Planning

The Preservation Planning functions are primarily concerned with ensuring that the actions and standards followed by the OAIS remain current and effective through time, and that any actions are assessed and validated where possible. It includes four components: Monitor Technology, Monitor Designated Community, Develop Preservation Strategies and Standards, and Develop Packaging Design and Migration Plans. OAIS makes only rudimentary recommendations for what ought to be included in a preservation plan.



Please briefly describe the typical elements of preservation planning in your preservation facilities?

Apart from OAIS recommendations the ADS has worked hard to develop (and publish) data preservation plans covering the elements above. See: <http://archaeologydataservice.ac.uk/advice/preservation>

- There is adhoc identification of formats or media at risk – often sparked by a particular business need. Eg we are transferring Images of England project digital files from CD to disk array to improve accessibility as well as preservation.
- We have good links with internal depositors and regularly discuss preservation and dissemination requirements eg for GIS and websites.

We are still in the planning and development stages of our TDR and have not yet agreed a preservation planning strategy.

Which parts of preservation planning do you think are particularly strong?

Monitoring of technologies, again it is a strength of the ADS that our technical team are active in the repository community more broadly both in technical discussions and in management practice (e.g. DPC, RDMF), (although there are not always the resources to respond immediately to changes in technology)

Links with internal depositors

N/A

Which parts of preservation planning would you like to improve given the resources?

Migration plans are more easily created than enacted given other pressures on resources.

Systematic review of formats/media ie monitor technology

Implementation of a preservation plan for each object type.

Are there any elements of preservation planning which you think would be improved through collaboration?

Consensus on migration paths and standards. Also many (sector specific - i.e. Cultural Heritage) repositories share the same of similar designated communities and given that monitoring these is a specialist function in itself this could be shared between repositories.

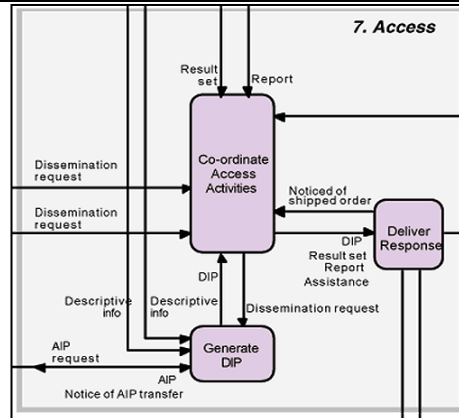
Monitoring technology and development of preservation strategies

Sharing of existing preservation plans and migration strategies and systems. Collaborative development where policies/systems are not in place for common formats and object types.

What would you hope to gain from such collaboration?		
Shared responsibility for migration paths and standards as well as community monitoring could reduce the pressure on resources.	Sharing information would give access to expertise. Seeking consensus on preservation strategies would help develop understanding of options and implications – even if we didn't all come to same conclusion	Help to eliminate the duplication of effort and also ensure the most robust policies possible through critical peer review and wider range of expertise in-putting in the process.
Are there any elements of preservation planning which can only be carried out locally?		
Those elements of preservation planning that are contingent on local infrastructure will always need to be done locally.		RCAHMS deals with a variety of materials which may be out of scope for other organizations making collaboration in those areas inappropriate.

7. Access

The Access functions manage the relationship between the OAIS and its consumers. In doing so it interfaces with the administration functions and archival storage. The scale and nature of access is not defined so the consumer could be broker who provides access to the public, a customer, or the public via the Internet. There are only three components: Generate Dissemination Information Package, Co-ordinate Access Activities and Deliver Response.



Please briefly describe the access functions within your preservation facilities?

ADS delivers all its data on-line via its website. The website is complex and utilizes some sophisticated technologies (e.g. NLP and faceted classification). The DIP is generated from the resource discovery metadata developed at ingest and held in the CMS.

Access to digital archives is either via an inquiry to our Research Service staff or our websites (Viewfinder, EH Archives). Websites provide access to versions of a limited range of material. Some records can only be obtained via staff

Canmore is the web front end of the RCAHMS database. It is at the heart of the RCAHMS archive, providing searchable, map-based information on over 280,000 buildings and archaeological sites throughout Scotland, as well as a catalogue of the collection items held. Over 130,000 digital images are available to browse and purchase online.

It brings together the results of the survey and collections material into one place and combines location information, site details and images on more than 300,000 archaeological, architectural, maritime and industrial sites throughout Scotland.

Canmore offers users the ability to:

- Conduct searches and advanced searches to find information on specific sites
- Identify where types of site are located
- Search for digital images on sites or locations throughout Scotland
- Collect direct references to specific sites from a wide range of textbooks and journals
- Know which of these we hold in our Search Room
- Users can now share their own information and images with the national collection by registering with Canmore

		Access to RCAHMS' Collections material through Canmore will be radically improved by planned developments within the SWISH program. This will include access to hierarchical catalogue records, PDFs files and dissemination surrogates for CAD drawings.
Which parts of access do you think are particularly strong?		
We have developed a search interface and supporting systems intended to make discovery and access easier – however this could still be enhanced (always the case). Our development of web services, bibliographic, Monument Inventory and archive based will be a key strength in the future – potentially delivery of LoD datasets also.	We provide effective retrieval of material. Both ERS staff and websites provide good searching facilities, with fast delivery of well documented files	<ul style="list-style-type: none"> • Variety of search facilities • Delivery of digital images. • User engagement through adding their own images to a Flickr collection
Which parts of access would you like to improve given the resources?		
Online discovery and access can always be improved.	<ul style="list-style-type: none"> • The relationship between our catalogue (AMIE), the DAMS (Portfolio) and the Catalogue needs to be much more integrated to allow for quick automatic updating of content • Development of automated on-line delivery mechanisms for a full range of files , including sale and licensing options where appropriate • Development of “scan-on-demand” system for creating digital surrogates for analogue material 	<ul style="list-style-type: none"> • Further innovative ways for our depositors and consumers to interact with our collections. • Automated generation of dissemination surrogates for more object types, access to original data.
Are there any elements of access which you think would be improved through collaboration?		
Yes, in a dream scenario, where data standards and shared repository processes are well developed then cross-searching of repositories and deep and stable linking of distributed datasets could benefit all repositories – essentially created a seamless pool of mutually intelligible data sets wherever they are held.	Only where a partner can provide access to specialist expertise (eg for a certain type of archive) or to a specialist audience	<p>Cross organization/ collection search facilities might bring benefits for consumers.</p> <p>Automated generation of dissemination surrogates.</p>
What would you hope to gain from such collaboration?		
The ADS is neutral on branding of datasets and careful to appropriately attribute data they present that is ultimately drawn		Enhancement of the user experience

from other sources. So even where collaboration is desirable in terms of shared infrastructure, the 'public face', and access modes for data are not our key concern – except where there the deposition of data is funded directly by an organization with a very specific remit and audience.(e.g. via a research council)		
Are there any elements of access which can only be carried out locally?		
Certain types of access to ADS data (i.e.) via the ArchSearch interface cannot be replicated elsewhere so would not be abandoned even though broadcast of data availability via web-service will become more widespread.		e-commerce

8. Any other comments

Are there any other areas of strength you would like to mention?	None
Are there any other areas for improvement you would like to mention	None
Are there any other areas for collaboration you would like to propose	None
Do you think this short survey has been useful? What should we do with the results?	None
Is there anything else you would like to add?	None

9. About this document

Version 1	Document initiated by WK and distributed to members as draft	30/09/2011	WK
Version 2	Updated with comments and distributed	04/10/2011	WK
Version 3	Return individual survey responses	13/10/2011	SJ, EW, ME
Version 4	Compile responses and distribute to members	14/10/2011	WK