AIMS UK event - notes

Revisiting Archival Principles from a digital preservation viewpoint,
Wellcome Collection Conference Centre, 10 June 2011

- the Powerpoint presentations from the day are available on the DPC website

Part One: The AIMS Project Approach and Model

A. Simon Wilson (Hull University Archives)

* After **William Kilbride's (DPC)** brief introduction, **SW** spoke about the structure of the day and gave a brief outline of the AIMS Project (An Inter-Institutional Model for Stewardship of born-digital archives). The event's structure was inspired by the structure of the Unconference which AIMS organised in Charlottesville in May. [This 2 day event was heavily shaped by the delegates themselves with considerable sharing of experience etc.]

* Within the AIMS Project, Hull University Archives acknowledge their beginners' status and decided to suggest Good Practice (rather than Best Practice). This also took into account that no single solution can be applied to multiple institutions.

* Involvement in AIMS gave us the courage to ask other depositors if they have any born-digital content – as a result a further 18GB of material has been deposited. Also - experience of working with one depositor keen to be involved in project but then reluctant to transfer digital material to the archives.

B. Judy Burg (Hull University Archives)

* Processes for born-digital archives in comparison to paper material – comparison of accessioning a floppy disk to accessioning a box without opening it.

* The four AIMS partners (Hull, Stanford, Yale and Virginia) will produce a White Paper – it will also discuss various functions for stewardship, including decisions to be made at each stage.

Part Two: Collection Development and accessioning

A. Chris Hilton (Wellcome Library)

* Starting point is to take existing principles and apply them to digital. Eventual aim is one interface for paper and digital.

* Proposed that rocket science is in fact simpler than the OAIS Reference Model, which boils down to: Get stuff > Put stuff somewhere > Keep it safe > Show stuff to people.

* Most digital material at the Wellcome Library is part of hybrid collections rather than purely digital – including hybrid collections that haven't yet been discovered – the problem of discs hidden in boxes of paper.

* Negotiating with existing depositors for their digital material. Problems arise with large organisations where there is an efficient records management system for paper, but digital material goes through different channels, i.e. 'that's IT's problem'. How will we find a way around this?

* Different timescale with digital material - having to 'talent spot' potential depositors to build a relationship with them - but how sustainable is gambling on the future success of your chosen academic/author/politician?

* The issue of trust and the idea that 'anyone can see it' – a different perception to paper material which is available to any visitor in the Reading Room. Our brand is Trust in handling *any* archives, and we will have to convince depositors that they can trust us with digital.

* Scrutinising archival processes for born-digital highlights the extent to which some processes for paper material have got 'baggy'.

* Discussion of appraisal and 'pre-cataloguing'. New digital accessions currently reside on a shared drive and are only ingested into the repository at the point of cataloguing. This will not be sustainable as backlog grows. The simple solution of describing at a series level brings the risk of restricted files becoming public (when we can't eyeball everything).

* What constitutes a 'reasonable producible unit'? Swiss National Archives say 30GB! [**Susan Thomas's** presentation later showed an example producible unit of 10 files].

* Consideration of ISAD(G) fields that may need adjustment: Physical Description (a description of the media the object originally resided on rather than content), Extent (needs to indicate what the user is up against - number of files rather than GB), Date (date created/last modified? i.e. the Photocopy problem).

B. Elinor Robinson (LSE Archives)

* As well as substantial amounts of digitised material LSE have 38 collections containing born-digital material; approx. 2% of the total. Accessioning digital material in CALM since 2009 though also had backlog of digital material discovered in existing collections. Main work has been on policy and technical infrastructure (Fedora).

* Archive institution places its trust in depositors – that they know what they are giving. LSE formulated a new detailed donor survey to be completed in situ by the Archivist rather than the depositor, helping solve current problem of uncompleted donor surveys.

* Institution commits to preserving what it gets – so has to know what it is getting. LSE assume new digital material is closed unless 100% confident that it can be accessed. They retrospectively pursue copyright and allow access only in the Searchroom while copyright is in force. Showing early commitment to preserve by earlier listing and appraisal reassures the depositor.

* Use DRAMBORA toolkit. Use CALM to catalogue.

* Accessions: The Same But Different. Acceptance may be a qualified decision that comes later on in the relationship than with paper. LSE try to influence how potential depositors manage their born-digital records. Appraisal continues post-accession, post ingest.

* The changed depositor relationship is also affecting the LSE Archives, including heightened awareness of archives within the library and collaboration between the archives and IT department

Discussion:

Undeleting

* **Tim Gollins -** whilst digital material is 'just stuff', the significant difference is the vast volumes of digital material. How will we cope when we can't 'eyeball' all incoming material?

* **Chris H** agreed that sensitivity will have to be road-tested. Wellcome Library decided on a policy not to 'undelete' or routinely crack passwords – to ingest only what the depositor intends to deposit. **SW** - to build trust, we should tell depositors that we are capable of recovering deleted material. Further discussion on 'undelete': general agreement that with permission, undeleting would be ethical. **Jeremy John** agreed - but equally we definitely need to know from the depositor what we are receiving. Undelete is thin end of the wedge; **TG** - tracked changes in Word - could recover superseded material and risk authenticity.

* **ER** - LSE use undelete on a case-by-case basis. The 'locked-box' scenario: there is no point is having material you cannot access – would decline to accept material if they couldn't actually get at it. **TG**: Need a classification of the minimum possible action required in order to accept an object. **Richard Boulderstone** mentioned the three-tier classification system in use at the Bibliothèque Nationale de France ranging from 'don't understand, can't view' to 'can understand and view'.

Volume

* **JB -** we don't *have* to accept whole hard drives of material – in the paper world would try to get depositor to undertake some sorting. **Jacky Cox** pointed out that most institutions would not have resources /inclination to sort through years' worth of emails. **Catherine Hardman** raised the importance of *selling* the value of the institution and of born-digital archives in order to make it an attractive option – outreach must be a part of collection development.

Pre-ingest work:

* Many institutions' workflow: to receive, do basic manifest, assessment and quarantine, then put in 'temporary' storage until full cataloguing can take place. Do we need to develop a different workflow – to enable us to get stuff into a repository quickly?

* **RB** - BL needs more pre-ingest tools for its large volumes of content. **TG** agreed and pointed out the coming economic issues and the change in the 30-20 year rule – TNA will have to at least double their ingest capacity. Risk Management has to replace Risk Avoidance. **Chris H** – collecting repositories have the freedom to say no to potential deposits, but still have to sell their services. **ER** - the importance of permanent Digital Archivist posts. Suggested that pre-ingest work could anticipate costs – i.e. with particular formats.

Ownership and copyright:

* **JB** --In most cases the depositor will still have their own copies of the digital material. **Chris H** said Wellcome Library encourage depositors to consider digital deposits as being the same as handing over the only copy. No other solution apart from to trust in our depositors not to pass on additional copies? **EF** and **Chris H** said that both LSE and Wellcome prefer depositor to transfer ownership – we also create authenticity by our commitment to preserve material – the 'top copy'. **Chris H** - 'death of the original' is more a philosophical problem than a practical one, though may be market issues in the future.

* **Grant Young** asked about copyright in digital material. **ER** - LSE's deposit agreement includes two strands of copyright clearance, one for preservation copies and one for access. If the depositor does not agree, accepting the digital material is pointless. Several speakers pointed out that many of the same weaknesses in copyright law apply to paper archives too > risk management is the only option.

Part Three: Lunchtime Open Floor

A. Grant Young: brief discussion of Plato and the Evaluating Plato in Cambridge [EPIC] project

EPIC is exploring the feasibility of using Plato for Cambridge University's Digital Repository. The project worked on a set of word-processed research documents to test migration paths and identify vulnerable formats. In order to identify key objectives they asked their depositors and users for input on what the significant properties of the documents were.

* Findings will be shared in a paper to be published after the conclusion of the project in July 2011.

B. Neil Grindley: brief discussion of Curators Workbench

Curator's Workbench is a pre-ingest workflow tool developed at the University of North Carolina. Its development was prompted by need for more usable digital preservation tools. The software enables the description of items in MODS and METS as well as creating unique identifiers and checksums. The most novel feature of the tool is its crosswalking ability, which allows automatic matching of MODS records with certain files.

* The creators are keen to receive feedback and anyone who would like to help evaluate Curator's Workbench through the user group should contact **NG**.

Part Four: Arrangement and Description:

A. Susan Thomas (Bodleian Library)

* All cataloguing is done in EAD using Oxygen XML editor. Collections are prepared and then passed to dedicated cataloguers and all medium-to-large collections are catalogued on a project basis.

* Digital material is quickly ingested into the BEAM Preservation Store after receipt. They then identify the content , appraise it and decide on what is 'in scope' – to what extent passwords will be broken, browser histories exported etc.

* FTK is used to create a filelist and metadata with files exported from disc images and migrated to current formats. Aim to gain basic intellectual control of the material. The cataloguers will only view ingested material so media photos supply the original context.

* For small collections, the cataloguer is supplied with media photos, an inventory, a file list from FTK and associated metadata. For larger collections, the cataloguer is also granted access to FTK file viewer.

* They also use a tool called Collection Builder which auto-generates metadata and allows input of supplementary human-generated metadata, in order to create DIPs for Drupal to handle. Description is more by use of keywords than long detailed descriptions.

* The objectives of A&D of digital material are the same as paper i.e. removal of non-archival material, packaging the material well etc.

Issues with current approach:

* Cataloguing projects necessitate a lead-time before the cataloguer is in post and raises issues if extra material is discovered mid-project.

* Cataloguers are users too – need to access all metadata and info

* This approach is based on fairly standard word-processed documents no model yet for more complex multimedia files.

* Tools need development to be more user-friendly. Scale issues already - only have a one-desk licence for FTK.

B. Jeremy John (British Library)

* Personal Digital Manuscripts Project (3 years). eMss lab combines curatorial examination with digital forensics*.*

* Creating curatorial digital objects (photography of creator's landscape, panorama's of creator's office and interviews – which can be linked to objects), digitised personal objects (images of media) and Virtual Archival Computing (emulating).

* Capture material and then arrange it; often a complicated mix of hard drives and floppy disks. Use two tools to check checksums are identical and compare metadata from various sources to maintain metadata quality.

* Unique identifiers that make clear which version it is - i.e. an original resource, digital object, replicate, facsimile, access copy.

* Using visualisations alongside some manual cataloguing for large amounts of material. **JJ** proposes running analysis likely to be most used, but leaving less popular analysis to be performed by users.

* Need to provide researchers with a way of citing their experience/access event – the environment in which digital material was viewed

* Use emulator to present content in Reading Room – put facsimiles into folders and replicate disk image

Discussion:

* **Susan Corrigall** and **Owain Roberts** asked about photographing the media and whether the original media is disposed of once ingested. **ST** explained that photos are for cataloguers' reference and are stored as a piece of metadata alongside the file.

* **TG** brought up the issue of dealing with material stored on the cloud or webmail services and whether there would be authenticity issues. **ST** said that they had successfully exported Wendy Cope's webmail account and **William Kilbride** suggested that virtualising a desktop would enable one to drop in various services from the cloud.

* **JB** considered the archivist's role to portray the structure of the organisation/depositor to the user – with digital personal papers the user can explore the existing structure themselves . **Jacky Cox** pointed out that the archivist's role in describing record-keeping methods is challenged since lots of people do not have any for digital material.

* **RB** suggested use of web 2.0 as an opportunity for the public to help produce metadata through crowdsourcing (though a minimum level of description would be necessary to facilitate this). **TG** described the need to get the community excited about the content – to consider it purely as 'cheap metadata' would be short-sighted. **JB** - example of the Self-Archiving Legacy Toolkit (SALT) at Stanford which allows this to happen purely between the depositor and the repository.

* **ST** - Need to provide researchers with a way of citing individual digital objects (even if not described individually).

* **OR** discussed the current mentality that we have to follow up every single IPR and copyright issue before we 'liberate' the material and allow access – is there a way to manage risk and produce material earlier without contravening legislation? **RB** suggested that we may have to accept that access will have to wait until we have solved these issues. **Catherine H** – ADS's depositors' keen to get stuff online but are unaware of behind-the-scenes work.

* **ER** - The forensic team extract content from depositor's storage medium then make it accessible to cataloguers for their work. Only add descriptive metadata at series level – but give each digital object (file) an individual reference/'shelf mark'.

Part Five: Discovery and Access:

A. Catherine Hardman (Archaeology Data Service)

* ADS is a purely digital archive which is embedded in a profession already comfortable with sharing its research and donating to archives. ADS makes [almost] all of its holdings available online for free download, and hopes that by selling itself to its users they will eventually become depositors themselves.

* The OASIS project makes unpublished fieldwork reports (grey literature) available online - these would otherwise be inaccessible. OASIS created an online upload form which allows contractors to upload reports directly, cutting out the steps of sending reports from contractor to local authority archive to national authority (plus backlogs at each step).

Challenges and issues:

* Building community buy-in – and encourage local authority to adapt and validate

* With higher volumes of reports has inevitably come the problem of poor quality reports and/or poor quality metadata – but the material is peer-reviewed by the world, which encourages improvement!

* Technical issues for some files/images – work on a 'best effort' caveat

* Old backlogs from before OASIS mean that metadata has to be hand-created, a time-consuming process

Opportunities:

* 80-90% of current research is now uploaded to OASIS (they receive approx. 500 reports per month) and the depositors themselves create almost all of the metadata at the point of upload through the OASIS form.

* Have managed to attract significant depositors including Time Team and volunteer groups.

* Will be attaching DOIs and making material citable (previously only at collection level). Also considering use of Natural Language Processing.

B. Tim Gollins

* **TG** started by demonstrating TNA's new search facility (currently in beta) -- currently for paper, the work has applications for digital archives - will be easier to integrate paper and digital in one interface.

* Entries have been tagged with up to 6 taxonomies allowing quick and effective filtering (e.g. by date range) and some results have also tagged with geo-locations. This search function stems from consideration of users' needs, based on Amazon and Ebay.

* Most users just want to access the 'stuff', not necessarily by navigating the intellectual hierarchy. Still described in EAD but stepping away from EAD technology. Treating records as information assets rather than physical objects. Collections hierarchies are still there but they have been relegated, aimed at academic researchers.

* SIPs could contain metadata which directs the item to the correct series.

* TG considered vast volumes of data that we are facing. TNA currently holds only 800GB of born-digital material (plus 100TB other digital data), but expects approx 1 PB by 2016. By 2014 will be receiving 35,000 images per day, every day.

* The volume of material needing to be accessioned will require automation, just a question of how to cope with this. TG suggests we need to reconsider the trend to granularity.

Discussion:

* **Chris H** – need to manage user expectations. **RB** - Move away from archivist creating catalogue which describes stuff. Could just create list of what is there and let researcher search for themselves. Refocus cataloguing effort towards quality of metadata.

* **TG** described the process of creating taxonomy categories by using Boolean searches for particular keywords – when a record returns x number of hits it is allocated to that category. The human has to be taken out of the loop.

* Some concern about the future of archival principles. **Chris H** – we might remove some detail from catalogues but vital to retain the hierarchical structure. Easier because digital is more easily browsable than paper?

* But Institutions which collect public records are in a better position to predict the volumes they will receive than collecting institutions.

* Metaphor: peering up the cartoon hosepipe to await the deluge of data!

Drawing it all together:

Judy Burg summarised some of the themes raised during the day.

* It is all just stuff, but we're more concerned about the challenges in dealing with digital stuff

* Ethics and informed consent

* Authenticity

* Scalability (i.e. by horizontal expansion)

* Roles and relationships

* Redefining the relationships between users, archivists and depositors

* Our role is mediator between user and depositor?

* We don't yet have an examples to relate to – i.e. Salman Rushdie is only visible depositor and no mature research community yet

Discussion:

* **RB** – Digital archives require similar processes to paper archives but using different tools

* **TG** - Intangibility of digital material. Need to replicate how the eye 'sees' stuff. Otherwise you are just 'feeling' the documents. Forensic tools 'open the box' and let you see the documents

* We need a user-friendly and affordable tool designed for archivists (like FIDO - Forensic Investigation of Digital Objects at Kings College London), which would be a natural progression from all the work done on format characterisation.

Thanks to all our speakers and delegates for taking part and making the day a success.