British Library Conference Centre on Tuesday 26th July 2005.

To open PDFs you will need Adobe Reader

The DCC/DPC joint Workshop on Cost Models for preserving digital assets was held at the British Library on 26th July, and was the first joint workshop between the two organisations. Around seventy delegates from the UK, Europe, and the US were treated to a rich and stimulating source of information and discussion on costs and business models with a number of key themes emerging.

Maggie Jones gave the welcome and introduction, on behalf of Lynne Brindley, and emphasised the need, not just to discover how much it costs to preserve X digital objects over time, but the implications of *inaction* and the strategic drivers which would motivate institutions to invest in digital preservation and curation. Laurie Hunter provided the keynote address (PDF 16KB)
and set the scene by placing digital preservation within a wider context of the business strategy of an organisation. The keynote stressed that there is a need to understand not just the costs but also the value of digital preservation and referred to the model scorecard as one tool which can be adapted for use in the digital preservation environment and which the eSPIDA project is investigating further.

James Currall referred to major obstacles to progress as including a very poor understanding of digital preservation issues among senior managers and creators and discussed some of the tools being developed by eSPIDA (PDF 182KB)
to help counteract those obstacles. Once again, the importance of the strategic direction of the organisation, was noted as being of critical importance. The eSPIDA approach to the model scorecard placed the information asset at the centre, with the other perspectives (customer, internal business process, innovation and development) tending to feed into the financial perspective. Currall noted that, while this was being applied within the University of Glasgow, the same principles can be applied anywhere.

Paul Ayris and James Watson gave a presentation describing the LIFE project (PDF 164KB) , which, like eSPIDA, has been funded under the JISC 4/04 programme. The LIFE project is a collaboration between UCL and the British Library. Paul Ayris described the context for the project, and drivers, which for UCL are the management of e-journals and the strategic issue of moving from print to e-journals. The BL needed additional information to help them manage multiple digital collections, acquired through voluntary and legal deposit, or created by them, and to maintain them

in perpetuity
. James Watson described the work to date in developing a generic lifecycle model which can be applied to all digital objects. The project also hoped to identify cost reductions and potential efficiencies. The major findings of this one-year project would be announced at a conference at the BL, in association with LIBER, on 12 December 2005.

The next sessions focussed on practical case studies. Anne Kenney described the work at Cornell (PDF 198KB)
on identifying the costs associated with taking on Paul Ginsparg's arXiv. A quote from a Victor Mature movie, "If we had some horses, we'd have a cavalry - if we had some men" seemed to appropriately sum up an attitude to digital preservation programmes, "we'd have a digital preservation programme, if we had some staff - if we had some content!". Kenney emphasised the importance of getting concrete cost figures since no senior management will be prepared to write a blank cheque. This reflects the recommendation Hunter made during his keynote address for digital preservation proponents to speak to senior management in concrete, economic terms. The presentation covered cost centers, which were principally staff costs, and also identified costs needed to support the work but which were often hidden. The arXiv.org archive is highly automated and is relatively cheap to maintain, with an estimated submission cost of between US$1-5. Expenses are minimised in this case by having a gatekeeper function at the beginning and having most cost of ingest borne by the depositor. Kenney also noted that the costs of the server had significantly reduced each year but cautioned that it was critical to ensure an ongoing annual budget, as it is not possible to skip a year in digital preservation.

The Cornell case study contrasted with the TNA case study (PDF 1.7MB) , presented by Adrian Brown. In this case, a publicly funded body with a mandate for preserving selected digital records so they must deal with a large number of formats. This illustrates the implications of organisational role and mission on potential costs. National libraries and archives will need to make different commitments to organisations who are more able to control the material they ingest. While TNA can influence creators, they cannot mandate that they will only accept certain formats. The TNA experience has shown some elements of costs for which there is a good understanding and others which there is little concrete knowledge of at this stage. Brown used the OAIS model to illustrate costs. Ingest costs represent the most substantial portion of costs and have been roughly calculated as £18.76 per file. As developments in automation progress and standards are agreed with creators, these costs may well fall over time. The time and human effort involved in creating metadata records for deposited materials was cited as a potentially high-cost element. Current research into automated metadata extraction could prove extremely beneficial in helping to minimise these costs. Data storage is relatively straightforward to prepare costs for but it is very difficult to predict transfer volumes over the next two years, and therefore difficult to plan longer term, so Preservation Planning is a major cost at the moment as it involves much R&D work. TNA also foresees opportunities to reduce costs through collaboration (not everyone needs to reinvent the wheel) and automation.

Erik Oltmans presented a model developed by the KB (PDF 350KB) , in collaboration with the Delft Institute of Technology, which compares costs over time of two key digital preservation strategies, emulation and migration. This is based on the assumption that migration must apply to every single object in a collection, while emulation does not. The emulation approach seems to work best with collections with very few formats - for example a large digital repository of pdf files. However, it can become much more costly when there are a vast range of formats to be emulated. Oltmans conceded that the model, may not be entirely realistic but provides a useful starting point. The KB experience indicates that volume is less of an issue regarding costs as the complexity of submissions.

The afternoon session began with David Giaretta discussing science data characteristics (PDF 323KB) and how these dictate the most appropriate and cost-effective strategy. For example, emulation is almost certainly not enough for science data, which is increasingly processed "on the fly" so the archive keeps the raw data and processes on the fly. Issues such as bandwidth are critical (how do you get data *into* the system and then how do you get it *out*?). Other issues are migrating a file (relatively straightforward) and migrating a collection (much more complex). The costs of keeping information *useable* were those which would be the most difficult.

Matthew Addis and Ant Miller did a joint presentation on PrestoSpace (PrestoSpace Presentation One (PDF 1.8MB) and PrestoSpace Presentation Two(PDF 2.4MB)), an EU-funded project on audio-visual archives. The project began in February 2004 and will last for 40 months, and has 35 partner institutions. A key issue for a/v archives is that digital formats are rapidly becoming obsolete. Individual items on a shelf will cause huge logistical problems as they become obsolete. However once mass storage systems are developed, then it becomes imperative to have metadata in order to find and keep track of individual objects. The aim is to establish a framework for medium-large archives at this stage. Miller said that there is a need to "scare budget holders into action" but solid numbers are needed to back this up. Addis referred to the urgent need for planning as "whatever you put your stuff on will be obsolete at some stage." A workflow model was demonstrated, which enables decisions to be made on priorities for action. The next stage will be to test how well the model works against existing archives' plans. Some copies of the preliminary report were made available at the workshop for those interested in further information. The DCC and DPC will make the final version of this report available on their web sites when it is released later this year.

Andy Rothwell and Richard House provided the final presentation on costing EDRM programmes (PDF 605KB) . Rothwell echoed earlier discussion in indicating that the pre-ingest stage is crucial in driving down costs. It was also necessary to look at the implications of the Governments Modernising Government white paper, which has been a key catalyst in moving from paper to electronic records. When coupled with looking at the whole information space, it needed to be understood that only c. 2% of records ultimately end up at TNA, so organisations need to manage the other 98%. The value lies not so much in putting material in but in being able to access it, so search and retrieval capabilities are key. The costs of implementation are not trivial, and it can take anywhere for 18 months to 2 years to implement the change in management and to provide the necessary training to staff. These costs are often not considered and can be significant. Other issues to be considered are the volatility of the marketplace. A practical example used was when EDRM product A is no longer supported and needs to be migrated to EDRM Product B. Without tried and tested export facilities, this is not a trivial undertaking. Rothwell also noted that data migration costs are not currently being factored into EDRM programmes. House went on to make the point that the key issue is not replacing paper systems with electronic but rather the integration of paper and electronic records systems. In terms of costs, staff costs are substantial and classification system design is frequently underestimated.

The workshop concluded with a panel session of all speakers and was chaired by Chris Rusbridge (DCC Director). Questions raised during this session highlighted a range of issues that were explored during the workshop.

For instance, it will be essential to determine what level of fiscal responsibility content creators and end-users share for the long-term preservation of digital assets. End-users potentially stand to benefit most from the preservation of digital assets and, as such, should be made aware that they may have a role to play in bearing the costs of preservation. Related to this were questions regarding the costs of accessing and retrieving digital assets over time.

The issue of metadata and representation information was raised several times during the panel session. Many participants stressed that without quality contextual information being preserved with the digital asset, there is little to no value in preserving the object. For example, even if a statistical digital data set is preserved and accessible 100 years after its creation, unless key items are defined, such as table headings, the data will be unusable. Users could undertake archaeological processes to try and ascertain the meaning of table headings, but ultimately they would at best only be able to guess at their true meanings.

The limit to which digital repositories may dictate acceptable formats for deposit was also a topic discussed during this session. While it is widely acknowledged that most repositories will not have the capabilities to preserve every format, there was also concern about placing too many constraints on content creators and depositors. As noted during the TNA case study, some organisation will not have the luxury of selecting the formats they will accept due to the very nature of their organisations, though they may be able to influence creators. In other cases, user communities may influence the formats that are deposited within repositories. This was the case with arXiv who did not originally impose restrictions but found that most depositors used the LaTex standard. This illustrates that identifying preferred formats for deposit does not always come from the managerial level, but could indeed be user-driven. Ultimately, a compromise is needed between reducing constraints on creators and depositors but also with facilitating effective preservation activities over time. Where there are equally viable alternatives, it may be acceptable to suggest one choice of format over another.

Very few repositories will have the capacity to care for every format or will have staff with all the skills needed to carry out preservation activities. Many of the participants felt that sharing resources and skills across a wide range of repositories would be the most logical approach to ensuring long-term preservation. PrestoSpace has investigated the creation of a European market place in which repositories and service providers can benefit from a shared approach. Several participants thought that the DCC and DPC might be able to assist in facilitating such an approach in the UK.

Participants felt that determining the value of preservation itself rather than simply identifying the costs will be of paramount importance in securing funding for digital preservation activity. This reflects suggestions made by several of the speakers. For instance, Richard House argued that it will be crucial for organisations to identify potential benefits that are not only appreciated by senior management but also by their stakeholders as well. It was acknowledged by several participants that a given stakeholder community may change over time and, as such, identifying benefits could be quite a difficult task.

It is highly unlikely that repositories will be able to accept and care for everything that is offered to them. Accordingly, sound appraisal and selection processes must be established within organisations to determine exactly what they will and will not preserve. Again, an organisational mission statement can be very useful in selecting and appraising digital assets for preservation. Selection and appraisal policies may change over time as the organisation changes. As such, periodic review of these documents will be necessary. Indeed, such changes may result in holdings within the repository no longer fitting in with the overall organisational mission. Therefore, some type of de-accessioning or disposal policy must be taken into consideration.

Many of the questions highlighted that, as yet, we have very few concrete answers. As such, much more work must be done in determining useable cost models, in identifying practical benefits, and establishing the value of digital preservation. The DCC and the DPC are currently looking into making available the spreadsheets for the cost models presented at this event via our web sites. We will also endeavour to monitor the progress of current projects and to report major findings as they are released.