Document Distribution Note
Release to Members: Immediate
Release to Public: 09/02/2012

APA Conference

London, 8th–9th Oct 2011

About the event

The Alliance for Permanent Access held its annual conference in London between 8=9th November. The conference was associated with a variety of parallel meetings including the APA's own AGM and board meeting, the ODE project, DigiCURV workshops, the APARSEN All hands meeting.

WK represented the DPC and various DPC members were present in their own right – Neil Grindley, (JISC), Laura Molloy (HATII), Martha Anderson (LoC), Kevin Ashley, Marieke Guy, and Martin Donnelly (DCC), Simon Lambert, Juan Biccaregui and Brian Matthews (STFC) and Tim Gollins (TNA), Nicky Whitsted (OU).

These notes are intended to provide an informal briefing for members of the DPC not able to attend the event. For an authoritative and comprehensive repost readers are encouraged to contact the organisers of the event or speakers directly.

Presentations and discussion

Keith Jeffrey – Introduction and Keynote

The European Commissioner Nellie Kroes sent her apologies and Keith who is outgoing chair of the Alliance offered to give the keynote instead. Keith reflected on the history of digital preservation from 1971 onwards. He emphasised the role of standards and the role of APA in pulling together research in DP research.

Javier Hernandez-Ros (Head of Unit, Digital Preservation Research, EC) – EU funded research on digital preservation

Relatively recently appointed though background in access to research data and technology policy. His talk had five themes – a short introduction to core policy issues, the past of research, current research, funding plans in the near future, and the longer term for EU research in this topic.

The i2010 programme in 2005 had a large chapter in digital preservation alongside digitisation and online accessibility. Recommendations from this work have been updated recently. A new call will be issued in early 2012 for improving access, management and preservation of scientific information. The EC made a significant contribution under FP6 to standards like TRAC, OAIS, Premis and others. Recent and ongoing work is familiar – Planets, CASPAR, Shaman, Presto and so forth - and some of this has led to the creation of competence centres and foundations such as OPF and Presto Centre. A wide range of agencies have been involved though in the early years it was research institutes and memory institutions, whereas in later years it has grown to include private and commercial partners.

ICT call in January will release another 30m Euros into DP research. (Information day on 15th November in Luxembourg) will fund more reliable and sercure preservation technologies and methods (via STREP); technologies and systems for intelligent management of preservation (via IP), interdisciplinary research networks (via a NoE) and schemes that promote the uptake of preservation research outcomes (via CSA).

Longer term will involve looking at new ways of doing things and there is a lot of interest in preservation-ready systems and resources. This is likely to change the nature of the agencies that are involved in these projects from the core memory institutions that were involved before.

The current framework programme will expire in 2014 and a new programme for 2014-2020 is currently in discussion. The debate is now about how much funding is available. However the proportion of resource for research and innovation funding is likely to increase provided a case can be made. The current proposal looks to increase the funding for science and innovation by 46%, and innovation and research are going to be united into a single framework. Three key themes for this investment: tackling societal challenges, creating industrial leadership and competitiveness and excellence in science. Digital preservation can be framed to fit all of these three headline developments. Digital preservation is interesting to the commission because of the nature of the jobs and competitiveness that it brings. Also key to this new framework is a new instrument called the ‘Connecting Europe Facility’ which are designed to strengthen cross border infrastructures in energy, transport and ICT in order to strengthen the internal market. This has existed for a long time as an instrument for the construction of physical infrastructure, but now is available for ICT and communications. 9.2 billion euros are currently allocated for broadband and key digital services for Europe in this period 2014-20.

Salvatore Mele (CERN) – ODE Project

ODE tells a number of tales that connect re-use, access and preservation, illustrating how these three elements work together and show the value and interconnection of each element. ODE has been listening to the way that connectedness and digital technologies have changed research and trying to reflect back real experiences to inform policy. CERN for example has been asking the question of ‘what is the universe made of’. It’s a large research group with thousands of researchers and hundreds of institutes. There are lots of incentives for research in this but very few incentives for preservation. However there are incentives for sharing data, partly because this improves the quality of research and partly because it allows more effective research when faced with a data deluge. A similar experience in life sciences where there are huge volumes of data. In climate science there are lots of data but different disciplines do not add up so that conclusions and ideas cannot be refined. Data sharing is not just a challenge within disciplines but is even more of a challenge across sciences. Citing data is hard but is a real incentive because it encourages re-use. It also changes the relationship between data and publishing – between innovation and knowledge. Journals and researchers and libraries need to change also. Evidence shows that scientists who publish their data are much more cited than those who do not. Change can come about in three ways – massive personal investment in each researcher one by one, generalised incentives and generalised requirements. There is also the possibility of sharing the science as well as the data.

ODE has learned that the next generation of infrastructure has to be built around incentives that enable data sharing rather than simply about data or computing. This will deliver benefits for the scientists but it will also contribute to changing science itself.

Juan Biccaregui (STFC) – Practice and Policy in Digital Preservation

STFC has a large range of responsibilities in respect of research data production, management and access mainly produced through large scale facilities which do research in particle physics, astronomy, materials, earth sciences and the like. At a strategic level they are involved in building research infrastructure and this includes research data infrastructure. The research lifecycle includes data creation and re-use all of which requires actual physical stuff, even if the researchers are not so interested in the building and specifications of the large scale infrastructure. Success is invisible: the better we are at preservation the less the researchers will notice. So for example we can pretty simply hide some of the metadata creation processes and we don't want researchers to have to enter metadata for every single measurement when there are such large volumes of data. Complex interconnected experimentation is going to require a range of supporting documentation from a range or partners. The amount of data being processed is doubling every 18months or so, and the 20petabytes currently available will last perhaps another 18months. This is a continuing problem. But the opportunities and the expectations are great. In reality STFC has built too many stand-alone architectures and is increasingly trying to establish a common set of services across the infrastructures, such as catalogues of users, preservation services and the like. This works well in theory but it's hard to do without a common policy framework.

Neil Grindley and Simon Hodson (JISC) – Current and future work in preservation and research

Neil and Simon introduced JISC and its work, identifying some existing and previous activities in preservation and research data management. Neil encouraged participants to look at the current JISC call on sustainability and described some new projects which are about to start with the broad theme of establishing the case for preservation. Simon introduced a number of projects on research data management. Success stories included those activities which had moved from being developmental projects to services. A new research data management programme is now being funded to develop institutional frameworks around data management which begin this month. The programme is going to be launched at the DCC conference. A large programme conference is being planned for 2013. JISC is also a partner for the Universities Modernisation Fund and is examining the role of clouds in data management. DCC has a role in this and there are a number of projects such as the Dataflow project at Oxford which is using SaaS techniques in the cloud.

Mark Dayer (Taunton and Somerset NHS Trust) – Digital preservation and the health service

IT services and projects in the NHS are notorious for producing lurid headlines of project failures. The NHS has enormous quantities of data and an enormous number diverse systems working locally and in unconnected ways. Core IT planning has not delivered benefits expected in the late 1990s and the problems are only increasing as systems proliferate. New types of medical scan produce new and exotic forms of data. There is a great need for improvement for ICT in the health services and there is need for innovation, but there are a lot of success stories. The failure costs twice over:

in terms of failed projects and in terms of opportunity costs for the profession who remain working with antiquated and inefficient systems. It wouldn't be so bad except that the health service needs to make £20 billion of savings. Better ICT is the way to save that money.

Martha Anderson (NDIIPP / LoC) Networks as evolving infrastructure for Digital Preservation

'When spiders unite they can bring down a lion'. Looking closely there are very few single webs: there are many small and interconnected ones. NDIIPP was created to help create networks between people to undertake preservation – communities working together as bilateral and multi-lateral alliances. Many common interests between organisations but these might not be obvious on first glance. It turns out that photographers and musicians who are really interested in metadata which are essential to their live businesses. Digital preservation has set off with concern about technology and how to build communities and something of a competition between the two. Infrastructure requires both to be meaningful. NDIIPP shows that boundaries between social and technical actions can be shifted in either direction. As multiple systems grow into networks and those networks into webs, so the early decisions will constrain or enable resilient integration and consolidation.

Michael Factor (IBM) – Riding the Wave on a Cloud?

Drivers for the use of the Cloud include the sheer volume of data if it is growing so quickly. Data is growing quicker than disk capacity and considerably faster than the capacity of management tools. Cloud computing is a model for enabling convenient, on demand access to a shared pool of configurable computing resources networks, storage, services, applications and the like) that can be rapidly provisioned and released with minimal management effort or service provider interaction. That means economies of scale in terms of storage as well as management; it offers speed and agility. But we also know the problems: data security and privacy; data being housed off shore; exit strategies of risks; then regulatory governance. But the cloud is a diverse thing and there are different models of clouds. If I have a private cloud for my own use and only accessible from the organisation then some of the issues go away but the pool of resources is relatively limited. At the far end the public cloud offers the greatest flexibility in terms of tools but may fail to offer the right regulatory framework. Somewhere in between is the concept of a 'community cloud' where uses are to some extent regulated but there is also a greater degree of flexibility.

This is one of the key issues for the ENSURE project which has three use cases including health care. The trick is to map OAIS AIPs to services in the cloud and to wrap it into a content aware data protection component.

Worth noting that the twitter feed showed quite a lot of disagreement about what the benefits might be. For example, one commentator said the definition of cloud sounded like bureau computing; another that some of the relationships between AIPs had already been dealt with via IRODS; a third that the issues of data storage were somewhat overplayed and that while they were correct they did not necessarily lead to a cloud infrastructure.

Monica Marinucci (Oracle) – Perspective on 'Riding the Wave'

The Riding the Wave report was commissioned for awareness raising and for vision and it has been very useful in this respect. It clearly articulates the need to develop a scientific e-infrastructure that supports seamless access use, re-use, access and trust in data. Data requirements go beyond capacity.

Tony Hey (Microsoft) – 4th paradigm and data sharing in sciences

Tony Hey has recently co-chaired a study for the NSF which was published around the same time as the Riding the Wave report and which was in many senses complimentary to it. (http://www.nsf.gov/od/oci/taskforces/TaskForceReport_Data.pdf) Changing the culture in universities is really slow and this needs some attention from academics and senior managers. Preservation and data science is a 'first class citizen' and needs to be funded for its own purpose rather than as a second thought. This includes capturing the provenance and understanding the workflows that created them. Microsoft has a lot of research interests in digital preservation given that their formats are in some senses the 'cause' of a lot of digital preservation action. A move to standards is good but is not enough because there are very many standards and they can be implemented in many different ways. So issues around engagement with the community are important and there is more here than might be realised. In addition to being engaged in PLANETS and SCAPE there's around 1000 in house computing science researchers in Microsoft and this group is making a considerable effort to engage in preservation research and development.

Christoph Best (Google) – Engineers in the Clouds: practical warehouse-scale computing

Most of the work in academic communities is about convincing academics to supply data, but the technology is not a problem – at least in a technical sense. We have moved from small data servers to large scale computing based on cheap consumer components that can be easily and readily replaced to create a much larger 'multi-computer'. Google file system means that thousands and thousands of computers can be plugged in and this changes the location of data management from the operating system to a specific file management system. It also allows scalability in terms of computing as well as storage. Data hosting and application hosting are therefore possible side-by-side. The trick is orchestrating the very large number of machines and the point is that they provide a new computing platform as well as a storage infrastructure. This has an impact on the scientific workflow. Thinking about the core questions if it will work – security has technical solutions (ie encryption) but cloud computing needs to be safe. That means tolerating errors as well as being reliable. If cloud providers are concerned with trust then we should expect them to develop mechanisms to signify trust – they will begin to look like banks (sort of).

Nigel Hickson (DCMS) – UK ICT policy development

European digital agenda follows up and has real dynamism. UK and other national plans can fit to this agenda. Key targets basic broadband for all by 2013, 30Mbit access to all citizens by 2020 and 50% with subscriptions for 100Mbis or more. Other key priorities – rights management infrastructure that is fit for the future, resolutions of orphan works, reform for data protection, broad spectrum for mobile broadband, universal broadband services … The internet is no longer like the skateboard – it is not going to go out of fashion. Government works best at a regulatory level and the mechanisms to do this are through agencies like the ITU – the International Telecoms Union. Regulation should be designed to be free which is a paradox. In the past telecoms providers have worked as a monopoly and have stifled development. That's not appropriate and probably no longer possible, so we need to prevent this from happening again.

Kostas Glinos (Head of E-Infrastructures, EC) – Reflections on open knowledge infrastructures

The vision for the e-infrastructure is to 'empower research communities through ubiquitous trusted and easy access to services for data computation communication and collaborative work'. It cuts across disciplines and countries. Infrastructure develops slowly and collects individual systems and services to become greater than the whole. It's not just technical but also social and regulatory and experience suggests it is generally much better if it can be developed from the ground up rather than with a top down view. Key elements here are trust – the researcher or the user of an infrastructure will only use it if they can trust it and there are all sorts of issues about convincing people that it is 'safe' to use. Open infrastructure is likely to be more trusted because they are transparent and science that uses such an infrastructure is, in turn, more likely to be trusted by researchers and the public. So data management plans for example will help preservation and transparency. Innovation requires ambitious uses for the infrastructure (though not simple duplication) and because of the scale it is likely to support international collaboration.

David Giaretta (APA) – APARSEN and PARSE.Insight

David introduced and gave an update on APARSEN. This has been covered before in DPC notes. Ash Hunter then introduced a specific area of work for APARSEN on testing environments for digital preservation, Mariella Guercio introduced work on authenticity and a fourth speaker introduced a study on the use of persistent identifiers on behalf of Maurizio Lunghi of FRD who was unable to attend.

Mirko Albani (European Space Agency) – SCIDEP ES project

The SCIDEP ES project was introduced, a project which seeks to put together the relevant research infrastructure for digital preservation for the Earth Science community which have been tested, validated and used. Although the service is intended for the earth sciences domain, the components are going to be both generic and standards based. Moreover this domain is itself quite widely defined. This means that the outcomes of the project will be something of a demonstrator for practitioners and researchers in many other domains. An early target will be to trial this infrastructure with all of the scientific data held by ESA, not just the earth observation data. The project has only just started. The OAIS information model which is conceptually quite simple, is the key to making it work. It requires that AIPs are properly assembled with sufficient representation information. In some senses this is a development of the CASPAR project because it turns the ideas and recommendations of CASPAR into an operational service.

Pirjo-Leena Forrstrom (CSC) – EUDAT project

EUDAT is a new European initiative that will deliver a Collaborative Data Infrastructure (CDI) with the capacity and capability for meeting future researchers' needs and enabling cross-disciplinary science in a sustainable way. Its design will reflect a comprehensive picture of the data service requirements of the research communities in Europe, and beyond. This will become increasingly important over the next decade as we face the challenges of massive expansion in the volume of data being generated and preserved (the so called 'data tsunami') and in the complexity of that data and the systems required to provide access to it. EUDAT has a large number of partners

Rainer Schmidt (AIT) – SCAPE project

SCAPE is taking up some of the themes of the PLANETS project on things like the integration of preservation tools, the development of integrated workflows, quality assurance, automated planning and technology watch. It also has a strong component of dissemination training and sustainability. The tools will be tried out with high volume and heterogeneous data sets like web archives, scanned newspapers, scientific data sets and broadcast. Characterisation of large scales of files, large scale migrations and large scale quality assurance are issues present specific and new problems for digital preservation but they are closer to the real experience of many institutions and need to work in order that we can work with the volumes of data becoming available. For

Eefke Smit (STM) ODE Project Integrating Data and Publications

Why research data is so important: Papers in Nature on DNA/The Human Genome- In 1952 a one page article, in 2000 a long article with fold-out pages, in 2010 an e-journal with clickable links. Journals PDFs now come with links to interactive data that you can manipulate. Number of datasets is outpacing the number of journal articles that are being published. Authors often sell data to publishers but some have had to stop accepting it as they were unsure of how to use it. Bringing in rules that data must be directly linked to the article. For ODE project used a Data Publication Pyramid first proposed by Jim Gray. Found that around 75% of data may never be openly available for reuse, also many disciplines do not have nominated data archive centres. In an ideal world the majority of data would either be included in publications or deposited in archives and therefore available. Publishers can help improve data availability by instituting stricter editorial policies, recommend archives, guidelines for citation etc. Possible structures for articles in the future: articles become layered; data and multimedia becomes independently citiable; underlying data will become part of articles through interactive PDFS; articles will become interactive with semantic tagging; there will be links between data archives and publications.

Max Wilkinson, British Library: The Currency of Data – At researcher perspective of integrating data and publication

Opportunities available:

Availability – a 1st class research resources, need to get people to loosen control and also identify roles and responsibilities for those managing it.

Findability – need persistent identifiers.

Interpretability – recognise the importance in metadata and work towards best practice in the development of standards.

Reuseability – Need for preservation.

Citability – Need to agree a convention for data citation, use PIs.

Curation – Develop sustainable and realistic data management plans. Need to collaborate with public data archives.

Preservation – Emphasis on the sustainable element of the data management plan.

Is data a currency? Perhaps, but if so then we need to treat it as a valuable asset. Need to educate researchers about the role making their data available can play in advancing their career.

Sabine Schrimpf, DNB – Custodians of Data
Libraries and Data Centres provide different services at different points during the research lifecycle, their areas of expertise are often different. Data centres tend to have more subject specific knowledge, libraries tend to have more general experience and knowledge. But with data and publications becoming more integrated they must begin to find more common ground and work together. Need to consider preservation and curation of many different forms of data and publications from integrated publications through to datasets published separately. A number of different initiatives are trying to deal with the issues:

- DataCite: persistent identification and citation
- Dryad and Dataverse: Both looking at integration of data publication and management support.
- Pangaea: A data publication and archiving service.

Although these initiatives exist there continue to be gaps and dilemmas: availability of data (lack of awareness, lack of incentives and fragmentation of services), findability (metadata, PIs, retrieval services), Interpretability & Re-usability (Documenation standardisation, curation and preservation). Libraries and data centres should turns these gaps into opportunities by acting co-operatively to provide richer information services.

Hans Pfeiffenberger, Tales of Drivers and Barriers to Data Sharing
Carried out a number of interviews to capture opinions for people in many different roles relating to data creation, publication and preservation about the opportunities for data sharing. Conclusions that came out of the work included:

- The necessity of an efficient infrastructure to enable data sharing.
- Money must be made available to cover data management tasks carried out by researchers.
- There needs to be an international data infrastructure and international support for this work.
- Support should include a 'help desk' and training.
- Premature release of data should not be enforced.
- Data misinterpretation is no reason for not sharing data.

Produced a list of barriers and drivers to data sharing, information of these is available in the work package report. Conclusions that have come out of this will now form the basis for quantitative research on the issues identified.

David Giaretta, APA: Audit and Certification

What is wanted in terms of audit and certification? Something comfortable, low cost, low trouble, comfirmation doing a good job. By funders? Is money being well spent? Also an ISO standard.

Challenges of developing an audit and certification approach:

- Does anyone have enough experience?
- What kind of judgement can realistically be made?

Still using OAIS as the basis (new version out next month). TRAC provided an initial draft, then developed by a CCSDS working group. The metrics that will be published as ISO 16363 is available from http://wiki.digitalrepositoryauditandcertification.org. Should be possible for the metrics to be used by a repository for a self-audit. Are the metrics sufficient? Should be regarded as a guide for auditors, the areas to focus on, other standards etc should also be considered.

There is a hierarchy of ISO standards concerned with good auditing. Additional document, ISO 16919, 'Requirements for Bodies providing Audit and Certification', defines the primary TDR Authorisation Body (PTAB) and process for accrediting auditors and creating (national) accrediting bodies – to allow more in-depth knowledge of local legislation.

The final steps in testing the draft standards were test audits to improve common understanding of the metrics within the PTAB (3 in EU, 3 in US). Standards then finalised, PTAB established as a formal body, will accredit new auditors (accredited training courses, will conduct audits with existing auditors), when enough auditors in a country will set up national audit bodies. Working on the assumption that there will be a demand from public and private/commercial organisations/service providers. Process designed to scale up. Hope that in the next 2 to 3 years there will be a full system in place for audit.

Have agreed a European framework for audit and certification:

- Basic – Data Seal of Approval
- Extended – Supervised self-audit
- Formal – Full audit

Need more audit cycles before finalising the handbooks for auditors and repositories.

What would certification look like? Not a simple statement of approval, instead a process of improvement. Information on the PTAB and the audit process: http://www.iso16363.org/

John Wood: Final Reflections

Going back to a report 'Preparing Europe for a New Renaissance', which called for a more holistic approach. Taking a researchers point of view, how do we train researchers in these issues? Need to give them the knowledge required to interface with all of the resources available to them without diluting their own specialism. Example of a pan-European resource and technology infrastructure: CLARIN.

Issues relating to take-up and legal and regulatory issues. Given volume and complexity of data may need a new data scientist role as an extension of the work carried out by archivists and librarians.

About this document

Version 1	Written on day	8-9/11/2011	WK
Version 2	Distributed	25/11/2010	DPC members, POCOS project