Document Distribution Note
Release to Members: Immediate
Release to Public: 07/05/2013

APA Conference

Frascati, Italy, 6th–7th November 2012

About the event

The Alliance for Permanent Access (APA) is an international alliance set up in 2008 to support collaboration between different agencies in Europe that are trying to build capacity in the preservation of data – with special attention to research and scientific data. It provides a basis for project collaboration and runs an annual conference. DPC is a member of the APA and it is run by an international Executive Board which includes WK.

WK represented the DPC and several other DPC members were present in their own right – Julian Richads (ADS), Neil Grindley (JISC), Martin Donnelly (DCC), Juan Biccaregui (RCUK), Sabine Schrimpf (nestor) and Neil Beagrie (personal member).

These notes are intended to provide an informal briefing for members of the DPC not able to attend in person. They only represent the session that WK was able to attend. For an authoritative and comprehensive report, readers are encouraged to contact the organisers or speakers directly.

Presentations and discussion

Carlos Morais Pires, EC – Scientific Data Policies and Infrastructure

The EC is a policy maker and funder and it undertakes a lot of research of its own. The Final ODE report – a project funded by EC - makes a really strong statement about how funders can target their limited resources. It emphasises that maximum impact from research is only possible with investment in data sharing and a co-ordinated approach to policy. This is the thinking which will inform the Horizon 202 programme which will come after FP7. Three keys programmatic themes are emerging: Open Scientific Content, Open Culture and Open Infrastructures, as stated in thinking like the ALLEA declaration, while the recommendation on Scientific Information from July 2012 was explicit about long term preservation. These are references which will be carried forward into the next 7 years of funding. Recommendation 4 is explicit that the EU member states should 'reinforce the preservation of scientific information' and that they should 'further develop e-infrastructures underpinning the system for disseminating scientific information'. Need to link national and local infrastructures to a European and global system for preservation and access, and ensuring that this is scalable for large scale scientific data sets. There's a basket of issues in here such as training, provenance, policy harmonisation and what not. The geographic span and the disciplinary specialism mean that there is a real danger in trying to homogenize everything. A flexible approach is needed in which authority is distributed.

Researchers, funders and policies are increasingly global in scale and they recognise that this global scale requires a global response. So NSF, EC, Australian Govt, Canadian Govt and others have been discussing how to work together. They have decided that a global Research Data Alliance is the way to proceed. There are 4 threats to global research infrastructure – unmanaged, disconnected,

invisible, single use. We can move to managed, connected, findable, re-usable. The funders have agreed a set of principles for their own negotiations: openness, balance, consensus, harmonization, voluntary and non-profit. RDA has a lot of energy behind it and will be formally announced in March 2013.

Antonella Calvia Gotz, European Investment Bank – The value of data

All universities and public research organisations, companies, consortia and partnerships can be beneficiaries of EIB funds for e-infrastructure, network and service investments, though agencies have to be credit worthy, must be legal entities with borrowing capacity, must be financially viable. Investment over 50million euros needs a direct appraisal to ensure they are credible and offer genuine value back to society. All tangible and intangible costs can be funded, so this includes physical components of things like data centres, as well as research and operational costs. Challenges in the funding include the timing and sufficiency of budget commitments, the uncertain value versus the costs (value is often over-estimated for e-infrastructure research and costs are under-estimated), and short-term planning horizons for long term service delivery. The result is that there are typical 'market gaps' for financing e-infrastructure. The field is typically risky, agencies (such as consortia) are typically weak and funding models are not very long term. That means investment can be hard to acquire for projects that are socially viable and potentially very impactful. This is the gap that EIB can fill, bridging the gap between research and development work and independent long-term sustainable business models. EIB prefers a long and early engagement and provides a lot of technical support to projects and initiatives. 2-3 years is not unusual between early discussions and signing a contract. Financing is spread through time, and we can use funding to keep high-value staff employed at a time of austerity. Europe needs to invest in innovation as this will be the basis for economic recovery: this is a material way to achieve that. The traditional projects were about bridges and hard infrastructure but the future will be about data and scientific infrastructure. This is a loan not a grant, and there is a need to ensure sustainability and plan this from the outset. Investments are monitored and are expected to be long term. Staff can work without worrying about financial problems, especially those countries where cuts are most severe. A new way to protect that investment.

Julian Richards, ADS – Sustainable digital archiving and value-added

Archaeologists have an unusual relationship to data, create lots of data and increasingly things are born digital. The ADS's position has always been that there is no point in preservation without (some kind of) access and the ADS has been particularly focussed on ensuring that access is enhanced. The number of users grow year on year and there has been a real escalation in usage, and there has been a really strong support for ADS when researchers are asked about the importance of the services that they provide. Initially funded by grant from JISC and then AHRC. ADS has engaged with English Heritage and commercial archaeological users. Data free at the point of use, so the costs are met by those funding the fieldwork. A one-off deposit charge is widely understood in archaeological research. Generally the costs for preservation are 3% of the total project costs, based on number of files, complexity and size. Costs of preservation flatline at a certain point, around 20 years, so that's the timescale over which costs are calculated. A recent analysis has examined the impact of the ADS, investigating that impact quantitatively and

qualitatively. Detailed analysis of value suggests something in the region of 2.5 and 8 fold return of investment in the data.

Peter Doorn (DANS) – 'Cost comes before profit'

'Cost comes before profit' is a common axiom of dutch traders. Paul Wheatley has reviewed cost models and noted that they proliferate. Julian has shown a straightforward model which works in practice and delivers clear value. In reality there is a lot of doubt and debate on the topic of costs and there is a range of ways of approaching it: and the value of data is not very well understood. What is 'added-value'? How can you measure the impact of scientific progress? What are the limits of growth? The more data you create leads to increased costs of managing data. Open access policies imply that you cannot really charge users so the costs of data have to be met by creators and/or funders of data. The scenarios in which DANS works are not simply economic but political. They do not charge for re-use but they need to earn back additional storage and handling costs, so they do charge for back up services for example. Storage is calculated over 5 years and for then it is free – similar to ADS's proposition.

Monica Marinucci (Oracle) – Data across time and organizations

Preservation enables three things - access, meeting corporate or legal requirements or to fulfil a mission. Preservation drivers for industry include saving money, reducing risk and increasing productivity. These drivers exist in government, justice, banking manufacturing and health care.

Mirko Albani (ESA) – Value and scientific data

ESA has a large number of earth observation missions ongoing and numerous completed projects. So the data is growing and heterogeneous and highly expensive to gather. Also the data can become more valuable through time with time series comparison – arctic ice for example needs to be measured across multiple satellites over several decades. Different instruments means that data needs to be interoperable through time. So to exploit the data to its greatest extent there is a need for careful preservation. The data is planned long before the mission starts and therefore data management planning starts very early. But the long term use after the mission is completed can be harder to understand ahead of time as the nature of earth science research challenges and policy development changes. All data is unique and un-repeatable and because it is global in scale it is a 'humankind asset'.

Eefke Smit (STM Publishers' Association) – The value of data to publishers

Publishers understand value very clearly, both in terms of the wider social mission but with a particular view on corporate assets and maintaining value for shareholders. So publishers have been involved in the Alliance from the very start and recognise that preservation needs a multi-pronged approach.

Neil Grindley (JISC) – Value from data now and into the future

JISC values information because it is accessible, flexible, smart. Moreover it enables innovation. JISC has been going through a series of changes lately. The new legal entity which JISC will become will

continue to invest in development and innovation, with about £2M allocated for investment in 2013. JISC is a funder but it's also a service provider, especially as a home to infrastructure and policy development. Thinking about digital preservation, this helps innovation in a variety of ways. Research methods are changing, and new business models to exploit data. JISC can help people achieve that.

Susan Reilly (LIBER) – Research passport

Only 22% of cultural heritage agencies have long-term preservation plans in place, and this is a problem because the objective of access can only be delivered if preservation is carried out. There has been something like £10BN investment in digitisation so that's a lot of money to waste. We need to spend a lot more time thinking about preservation of cultural heritage resources. Three things are needed: trust in content and infrastructure; infrastructure for access, reuse and deposit; and sustainability of roles, mandates and partnerships. Issues around legislation and mandates need to be cleared out, and we really need to understand shared infrastructure better.

Project Fair – short papers

WK chaired an extended session of short papers which were a mix of informal presentations and tasters of larger research projects and initiatives which would be presented later in the conference. Papers included the following:

- Mirko Albani (ESA) - SCIDEP-ES project
- Neil Beagrie – Keeping Research Data Safe
- Luigi Carotenuto (Telespazio) – Ulysse and Circe projects
- Martin Donnelly (DCC) – Data Management Planning
- Michael Factor (IBM) – Ensure and Vision projects
- David Giaretta (APA) - APARSEN
- Neil Grindley (JISC)- 4C project
- Ross King (AIT) – SCAPE Project
- Ari Lukkarinen (CSC) – EUDat Project
- Maurizio Lunghi (FRD) – DigiCurV Project
- Rudi Mayer (SBA) – TIMBUS Project
- Cezary Mazurek (Poznan) - Wf4Ever Project
- Salvatore Mele (CERN) - ODE
- Susan Reilly LIBER – LIBER Research passports
- Thomas Riise (L3S)– ARCOMEM
- Barbara Sierman (KB) – The Atlas of Digital disasters
- Jamie Shiers - DPHEP
- Eefke Smit (STM) – Codata and Data Citation
- Miroslav Serl – CNZ 'what will be left behind'
- Beiamino Di Martino – MOSAIC

Document Distribution Note
Release to Members: Immediate
Release to Public: 07/05/2013

DigitalPreservationCoalition

Hans Pfeiffenberger (Alfred Wegener Institut) – researchers and funders: challenges and changes

There are a lot of barriers to data sharing in research, and the reasons are complex. ODE has tried to simplify this and provide clear and concise material to help researchers understand the opportunities and advantages of data sharing, and how practice may need to change. Data sharing is not just about preservation but there is clearly confusion in some quarters about this. For example, in different disciplines digital archives go to different types of institution – library, repository or library. Learned societies and editorial boards tend to be influential in formulating policy on data sharing, providing a solid and contextually specific sense of the ethics of data sharing: this matters because researchers can influence these groups. If there are capable archives then researchers should be clearly identify them through guidance from learned societies and editorial boards; and if there are not then researchers should complain to funders.

Susan Reilly (LIBER) – Evolving Roles in Scholarly Communications

The cultural change in libraries and scholarly communication has been really great in the last decade or more. ODE examined this by exploring a number of themes – for example how data and publications are linked together, which means that libraries need to have preservation strategies in place for data as well as publications and dependable links between them. We need to define new roles for librarians and a better understanding of the skills that will be needed. This analysis of skills was based on a major survey of skills in 7 areas including preservation and it allowed libraries to articulate their own vision for the future. Subject specific expertise is rated over technical skills and an emphasis is placed on re-training existing staff rather than hiring new people. There is a strong sense of 'doing' rather than 'waiting'. Skills were a real concern, and a broader dialogue was needed with publishers and researchers to identify the gaps.

Eefke Smit (STM) – Changing Publishing

Journals are now working much more effectively in the presentation and linking of data. A few years ago data was too often 'dumped' in supplementary files which were poor to access and hard to preserve. In the last 2 years we have much better examples of data publications which not only have data in repositories well linked and managed, but also with their own viewers and browsers for data. A 'Data publication Pyramid' describes the types of data publication with a large amount of unpublished or raw data at the bottom, and a small amount of data publications at the top. Typically the amounts of data are quite small in size – the files are seldom more than a megabyte or so. It's not the size it's the complexity that matters, as well as the growing number of publications which include data. There are still too few 'data publications'.

About this document

Version 1	Written at conference	6-7/11/2012	WK
Version 2	Distributed	7/11/2012	DPC members