Document Distribution Note
Release to Members: Immediate
Release to Public: 08/10/2014

PASIG 2014

Karlsruhe, 15-18 June 2014

About the event

From 15 to 18 September 2014 Pasig 2014 took place at the Center for Art and Media (ZKM) in Karlsruhe . You can find the programme at https://www.asis.org/FPasig/events/PASIG_2014_Brochure.pdf. The presentations are available online: http://web.stanford.edu/group/dlss/pasig/PASIG_September2014/.

Angela Dappert represented the DPC and the TIMBUS project Tuesday through Thursday. She presented a summary of the TIMBUS methodology and tools and participated on a panel on the preservation of complex digital objects, in particular databases and processes. DPC members (The University of Portsmouth, the Bodleian, Digital Repository of Ireland) and partners (OPF, nestor, National Library of Australia) were present in their own right. 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 report readers are encouraged to contact the organisers of the event (Arthur Pasquinelli and Tom Cramer) and the speakers directly.

Presentations

Monday

Monday focussed on Research Data Management, which was separate from the main conference. Legal matters, organizational challenges, best practices and infrastructure for data management were discussed.

Tuesday morning …

… was dedicated to Storage Technology 101 – an Oracle-sponsored session. This morning was very information dense. Art Pasquinelli started the day with an overview of long-term data retention trends.

Philippe Deverchere discussed disk trends, in particular new technologies, such as helium drives, shingled drives, heat assisted drives and bit patterned media. While providing more disk density, performance increase is limited. Phillipe discussed concerns for data protection and preservation: long-term retention, energy efficiency, proven and open formats (especially the need for avoiding complex storage software solutions, such as automatic storage deduplication), dual technology schemes as protection against hacking of all copies, storage cost, and reliability. He, and others later in the week, discussed the advantages of tiered storage combining flash, disk and tape to achieve an optimal balance against these concerns.

Christine Rogers covered tape trends. She discussed the vastly improving tape density over the years. Interestingly, recent published high-end improvements have only been achieved in the lab and this high-density demo tape can currently only be read using disk heads rather than tape heads, and therefore is not usable in practice.

Dan Deppen discussed improvements in the storage abstraction layer. This includes: analytics for monitoring how efficiently media are being used; policy-based media validation (for example based on last access, random sampling or newly entered cartridges) that can result in self-healing; mobile apps for monitoring your storage from anywhere; horizontal scalability by ingesting data through multiple stock servers rather than having to go through the bottle-neck of the metadata server. Additionally, import and export on remote sites is done by index access only rather than accessing content. He discussed cloud tape solutions and SWIFT integration, where 3 copies are randomly distributed amongst all the distributed storage nodes. He discussed the spectrum from highly performant and secure storage on premise with high initial but low total cost to public cloud solutions with less control, no guarantee of privacy but low initial cost, and all the solutions between them, including networks, remote sites and private cloud.

Donna Harland discussed Oracle's Optimized Storage Solutions Architect which provides sizing guidelines and creates customised proven complete storage configurations through preferred vendors, based on the requirements of individual organisations.

Philippe Deverchere discussed the properties of cloud use, including need for authentication, strong encryption that is not needed in a local solution, specific management, specific logging and reporting, bandwidth restrictions and possible use of compression, performance, security and cost. He introduced the different permutations of disk and tape use in the front-end or back-end of the cold storage: using tape for seeding during original ingest, for recovery, for migration from one cloud to another, or for tape use in the cloud.

Jason Goodman presented on the partnership of Cray and Oracle providing a joint solution for tiered storage for high-performance computing.

In the second half of the morning Thomas Ledoux from the French National Library, Krishna Chowdhury from the Qatar National Library and Mikko Tiainen from CSC Finland presented use cases of their storage architectures.

Tuesday afternoon …

… was dedicated to the customary Digital Preservation 101. This time it was presented by Tom Cramer, Stanford and Neil Jeffereys, Bodleian. They introduced introductory concepts of Digital Preservation. Neil went beyond basic introductory topics by discussing helpful design patterns for creating an architecture that suits current needs, but will not become a burden in the future. This includes: provisions for multiple ingest and dissemination paths; modularisation, layering and abstraction; use of simple standards, such as REST; txt and XML instead of binaries; open instead of proprietary; asynchronous workflows; parallelised workflows; scrupulous retention of provenance; provisions for versioning; adapting tools from digital forensics; taking immediate advantage of opportunities for acquiring metadata and content when they present themselves; using information provided in data management plans to plan for archiving provisions that will be needed downstream; and avoiding funding discontinuities.

Armin Straube from nestor gave an overview over their work on policy guidelines, curriculum development (nestor handbook and summer school) and certification. The nestor seal is based on DIN 31644 and is a simple self-assessment that is going to be integrated in a 3-tier trustworthiness certification approach, with nestor being the simplest, DSA above it, and ISO16363 as the most comprehensive.

David Minor from UCSD also covered trustworthiness and reported on their experience going through the equivalent of an ISO16363 certification. The certification is not officially possible until the certification governance, process and training are signed off. Chronopolis spent 18 months on it with 2 to 3 staff dedicated to it.

Philippe Devereche followed with a presentation on tiered storage.

Wednesday morning …

… was dedicated to research data, both small sciences and big data.

Neil Jefferies presented the Research Object approach of the Wf4Ever project. The TIMBUS and Wf4Ever projects have held several co-organised training events and training material can be found here: Slides http://timbusproject.net/resources/training/trainings-material#From%20Preserving%20Data%20to%20Preserving%20Research , Wf4Ever demo video http://youtu.be/TxW2wvreyoQ. Updated material will be available shortly. [Angela's addition: The differences between the TIMBUS and Wf4Ever approaches are that: Wf4Ever preserves open source workflows only, while TIMBUS preserves any type of processes. Wf4Ever is based on trust in the information provider, while TIMBUS tries to detect preservability risks. Wf4Ever focuses more on the front-end: they offer a dashboard for deposit of relevant materials, while TIMBUS focusses also on the back-end, automatically or semi-automatically extracting characteristics and dependencies of the research environment; rather than assuming the availability of information it helps to obtain relevant information. TIMBUS also supports an extensible approach that expands the preservation scope by permitting the addition of domain specific ontologies.]

Varsha Khodiyar spoke about Faculty of 1000, the methodological publication of open data. Nano publications, which are core scientific statements with associated context, offer a highly granular method for making research assertions. It is derived from the open annotation ontologies http://www.openannotation.org/spec/core and started with the Concept Web Alliance http://www.nbic.nl/about-nbic/affiliated-organisations/cwa .

David Wilcox from Duraspace presented the improvements introduced in Fedora 4 (see https://wiki.duraspace.org/display/FF/Fedora+4.0+Feature+Set) and gave a status report as to how far it is currently supported by the Fedora-based repositories, such as Islandora, Hydra, DuraCloud.

Wolfram Horstmann from the University of Göttingen spoke about the long-tail of research data, which comprises 48% of all projects that produce data of less than 1G or, 92.4 % producing data of less than 1 TB each. These projects have large heterogeneity of formats, standards, and access. Metadata is insufficient. All of this means that it is easy to find an individual data set, but it is hard to discover across data sets based on collection criteria. A valuable source of existing data sets is http://www.re3data.org/

David Minor from the San Diego Supercomputer Center/UC San Diego Library spoke about Leveraging High Performance Computing for Preservation and Curation. He introduced the architecture supporting their high-performance needs: high-speed networking within and without the organisation; interfaces grouped by performance; smart, self-healing storage; capacity. Common functions that are supported are fixity, replication, data tracking and registration. Challenges they face are issues of scale, short-term funding models, and having to work with the assumption that data can only be stored temporarily for analysis, rather than permanently. Rerunning analysis is not an option in these cases.

Jos van Wezel from the Karlsruhe Institute of Technology/Steinbuch Centre of Computing reported on their personal experience running the scientific data archive: costs, technologies, and challenges and brought interesting perspectives of someone who had to do so without having had much exposures to the concerns of digital preservation.

Thomas Schoenemeyer, Cray spoke on *Managing Data From High-Performance Lustre To Deep Tape Archives*. He presented LUSTRE, the Cray parallel distributed high performance file systems for computer clusters ranging in size from small workgroup clusters to large-scale, multi-site clusters. It works on administrative defined rules and allows purging, removal of directories, deferred removal, archiving and releasing based on built-in policies. A TAS connector provides a simple way to protect and move data across storage tiers and locations.

Erin Tripp from Discovery Garden spoke about Islandora's mining the open source ecosystem to achieve sustainable and reliable preservation solutions. A recent event report on Islandora for DPC members can be found here: http://www.dpconline.org/component/docman/doc_download/1224-islandora-camp-2014-05

Wednesday afternoon …

… started with lightning talks, amongst others, on the Versity Storage Manager, the EARK project, the TNA's cloud report, the BnF's SPAR system, Thorsten Langes' consultancy on digital preservation targeted to industry, EUDAT, and the NLNZ data centre.

This was followed by a presentation on the cost and impact studies for data centres done by Neil Beagrie in the economic, archaeological and atmospheric fields. As a case study, the ADS have, with a £1.2 m/ year investment and value, impact in the order of £13m on research, teaching and studying. Neil presented the range of methodologies used and the statements that can be made based on them.

The rest of the afternoon was dedicated to the preservation of data bases and processes. Janet Delve from Portsmouth University spoke on database archiving in the E-ARK Project; Neal Fitzgerald from Queensland State Archives spoke on preserving archival records in business systems; and the Swiss Federal Archives gave an introduction to SIARD. I was a speaker and formed part of the panel and, therefore, did not take any notes on this session. I presented on the methodologies and tools coming out of the TIMBUS project, which is finishing in December 2014. The slides can be found here: http://timbusproject.net/component/docman/doc_download/174-timbus-process-preservation-methodology-a-tools-

Thursday morning …

… was dedicated to preservation at scale.

Rainer Schmidt from the Austrian Institute of Technology represented the SCAPE project which has finished now. It has scaled digital preservation with respect to the size, number, complexity and heterogeneity of digital objects. ToMaR provides a solution to run preservation tools on a distributed Hadoop MapReduce cluster using existing command-line tools and Java applications. The tools do not need to be adapted to take advantage of the scalable environment.

Sharon Webb from the Royal Irish Academy spoke on Preservation as a Service. The Digital Repository of Ireland is an interactive trusted digital repository for contemporary and historical, social and cultural data held by Irish institutions. It links and preserves the data held by Irish institutions, providing a central internet access point and multimedia tools. Sharon spoke about their funding, status, leaflets, fact sheets and membership model. She explained how the services offered cover a range of community needs that varies for depositors, end user, institutions that wish to leverage the infrastructure and open data initiatives.

Mike Quinn from Preservica presented the company's digital preservation offerings. A recent recorded webinar for DPC members can be found here: http://www.dpconline.org/members/conference-reports/1208-dpc-webinar-technology-bytes-preservica-7-may-2014

David Minor from UCSD spoke on the Digital Preservation Network (DPN). Academic Preservation Trust (APTrust), Chronopolis, Hathi Trust, Stanford Digital Repository (SDR) and the University of Texas Digital Repository (UTDR) use a federated approach to preservation. Their pre-existing digital repositories for long-term preservation and access are used to replicate multiple dark copies of these collections in diverse nodes that are only accessible for preservation actions. Rather than creating one central repository they function as fall-back solutions for each other. David described the architecture. Key concepts are First Nodes as point of entry where contracts are negotiated and service levels agreed, and Replicating Nodes. Every node has a staging area and uses AMQP messaging. The transfer mechanism still needs to be scaled up. BagIt is used for content packaging. The project is currently developing the exact workflow for each scenario in which DPN applies: Ingest and replication, restoration of content, cessation of a first node and successioning.

Adi Alter of Ex Libris presented the company's Rosetta digital preservation offerings. A recent recorded webinar for DPC members can be found here: http://www.dpconline.org/members/conference-reports/1222-dpc-webinar-technology-bytes-ex-libris-21-may-2014

Philippe Devechere, presented the *Four-Tier Computing: Linking On-Premise and Cloud Infrastructures* that had already been discussed in the Oracle special session earlier in the week.

Alex Wade from Microsoft Research spoke about Azure4Research, which is an initiative by Microsoft to make their Azure cloud platform available for free to researchers who propose interesting cloud research projects. The website has hands-on labs and online training materials available for anyone. Alex presented examples of interesting work that had been performed on previous grants. The

SCAPE project, for example, has a conversion matrix that migrates deposited files into all alternative, appropriate file formats. It then runs analyses on all output formats for potential indicators that suggest that migration loss has happened. It shows in colour the regions of documents or images that differ before and after migration.

Matthew Addis from Arkivum presented the company's digital preservation offerings. A recent recorded webinar for DPC members can be found here: http://www.dpconline.org/members/conference-reports/1165-technology-bytes-arkivum-26-march-2014

Michael Selway from Cray discussed the Cray approach of using an open archive for big data and super-computing. He argued that it is necessary to establish a tiered data management architecture for total data management in which the core of the solution is a well-established software with a plan for technology refresh.

Thursday afternoon …

… was dedicated to audio-visual media preservation. Neither of the scheduled speakers had been able to attend the conference. The very competent replacement speakers presented ad-hoc.

Christine Rogers from Oracle gave an overview over the storage architecture needs in AV industry that need to support production, post-production, distribution and archiving (48%). Increasing digitisation, file-based work-flows, changing video formats (with 4 TB/hour standard on 4k films at the moment) and changing business opportunities all impact the shift in storage architectures. Solutions need to address the questions on total cost of ownership and storage efficiency; meet potentially infinite retention requirements; manage data growth and complexity; survive tech refreshes and data migrations; ensure data protection, integrity and security; and maintain access to data as applications become obsolete.

Erin Tripp from Discovery Garden / Islandora presented a case study of the MIRC-DVR project at the University of South Carolina that preserves the Fox Movietone News Collection. Erin presented the content model architecture that is tiered for preservation, media production and streaming web access, which is controlled by XACML policies. The backend is synced to a Filemaker database that supports metadata versioning and complex workflow tracking, from inspection of the film original to digitization and transcoding. Metadata includes PBcore, MODS, and DC. The pilot now has been launched. Challenges remain the large storage size; performing fixity checking on such large files; defining policy-based management; off-site back-up for disaster recovery; durable linking and embedding; end-user contributed metadata; and providing time-based metadata.

Finally, Tom Cramer from Stanford presented on the Avalon project which forms part of the Hydra project and is an open source system for managing and providing access to large collections of digital audio and video. It enables curation, distribution and access provision. Key requirements include: support robust, standards-based metadata for description and annotation of time-based media; support authentication, authorization and copyright; accommodate special requirements for asset preservation and long-term archiving; integrate with preservation repository services; leverage other open-source higher education projects such as Hydra. An interesting metadata challenge is that often lots of snippets from different sources are combined in one AV object presenting a difficult copyright situation.

This session was followed by a further series of lightning talks.

About this document

Version 1	Written	07 October 2014	AD
Version 2	Distributed	08 October 2014	DPC members