APA Conference 2014, 22nd-23rd October, Brussels

Session One

Welcome: Juan Bicarregui, STFC

- APARSEN project has focused on some of the key issues for the APA, in particular developing a Common Vision for digital preservation.
- The developing importance of open research data, the shift from theory to practice and the creation of the Research Data Alliance has also been a big influence.
- The launch of the 'Virtual' Centre of Excellence

EU-Funded Digital Preservation Research: Manuela Speiser, EC

- EU activities focus on policy coordination, research funding and developing legislation
- Recently published guidance information on digital preservation in relation to digitisation
- FP7 most recent research funding programme, 6 calls funding 23 projects with over 100 mil €
- Results from FP7 programme projects have included:
    - Formal models for characterisation
    - Tools for supporting preservation planning
    - Increased automation and scalability
    - Tools for web archiving
    - An emulation framework for obsolete media
- Have also developed a range of services for end-users:
    - OPF
    - PrestoCentre
    - APARSEN VCoE
- Projects ongoing until 2018 focusing on 3 areas:
    - Techniques and tools for recovering loss
    - Embedding reasoning and intelligence
    - Coordination actions (Includes 4C)
- Expected impacts from the programme:
    - Reduced information loss
    - Lower preservation costs
    - Strengthen EU service and technology providers
    - Restructuring the research landscape

EGI – A Community and Shared E-Infrastructure: Michel Drescher, European Grid Infrastructure

- Everyone is talking about 'big data' but it is an amorphous term and no one really knows what it means….
    - Covers: volume, variety, velocity….

- Estimates that 1/3 – ½ of data is 'dark data', where it is unknown how it was gathered and what it contains.
- Must have context and know relations to be able to turn data into information.
- EGI provides:
    - Uniform access to heterogeneous data and computer services – distributed storage services and grid and cloud platforms
    - Federation of services from public funded infrastructures, institutional infrastructures and commercial providers
    - Offered free at point of delivery or pay for use
- Has 340 data centres in 34 national grid infrastructures and more than 200 research centres
- Services available for long tail and big science
- Have a federated cloud infrastructure supported by organisations around the world
- Offering preservation as a service through the EGI cloud infrastructure, includes:
    - Shared services such as PID services, VM repository, Representation info repository
    - Customised services including provenance, data inspection, data packaging and brokerage
    - On-Demand services such as compliance checks and transformation services

Deploying APARSEN/APA CoE at a National Level: Silvio Salza, CINI

- Why specialise at a national level? To take into account national peculiarities e.g. national regulations, market peculiarity, language.
- In Italy there has been legislation for nearly 20 years, digital signatures have been legally valid since 1997
- Early regulations mostly focused on integrity and authenticity
- Since 1998 all public admin bodies have had to maintain an digital registry system for all mail, so immediately created a large market for DP services
    - Code of Digital Administration governs all practices
- Increasing momentum within Italy towards mandatory use of digital systems such as e-invoicing which drives need for DP
- Most are looking for 3rd party services rather than developing their own solutions
- Official certification process already in place based on ISO 16363
- Main targets of CoE in Italy and memory and scientific organisations who are less reliant on the commercial solutions
- Services to be offered will be around training, help with setting up and managing preservation processes, and audit and certification
- Currently working on setting up a qualified user group that will include national archives, libraries, central admin bodies, universities and quality minded private organisations
- Already have training at the level of formal qualifications and at an introductory level
    - Also producing online content adapted from in person courses and translations of APARSEN training content
- Services available to aid with planning and assessing quality of preservation services
- Will be an Italian version of the CoE portal

Session Two

ESA Perspective on Earth Science Data Preservation: Mirko Albani and Pier Giorgio Marchetti, ESA

- ESA has a policy for making all data it preserves free and open
- Now has more than 20 years of earth observation data available
- Data continuity is essential for comparison within Earth Science, so appropriate context information is essential
- All data captured is unique, capture not repeatable, fundamental for monitoring global change and effects policies and actions and has great potential for combination with other data sets.
- Also concerned with volume, value, variety and velocity of big data
- Main driver for Long Term Data Preservation (LDTP) at ESA:
    - Data and knowledge preservation
    - Data valorisation
    - Optimising investments
- Collaboration important, both in terms of projects and other space agencies
- Also working on internal processes, ensuring discoverability, access and data quality
- Have had success in processing data from the 70s, hardest part was finding the necessary contextual information
- ESA Research and Service Support service had the mission to support the EO community in exploiting the EO data, providing access and processing and scalable infrastructures
- Clear evidence of increased use of data sets for which ESA provides active support through an exponential increase in the number of publications.

The Global View: Professor John Wood, RDA

- Many different projects producing huge amounts of data: things like the LHC and Square Kilometre Array
- RDA is about building bridges to overcome hurdles for easy data access, data sharing, and interoperability by facilitating collaboration. Reduce current fragmentation.
- A bottom-up organisation formed from a number of working and special interest groups. Discussing at plenary sessions and online community.
- Focus on practical outcomes over producing endless reports.

APARSEN/APA Centre of Excellence: Simon Lambert, STFC and David Giaretta, APA

- Looked a variety of different models for CoEs, both physical and virtual. Especially those within the DP domain (i.e. IMPACT, VCC-3D, PrestoCentre, OPF)
- 3 aims seem to be key:
    - Developing the excellence
    - Structuring the excellence
    - Spreading the excellence
- CoE looking to bring together tools, services, training and consultancy to answer the key questions relating to DP

- Used the 'business model canvas' approach to help understand the interaction between infrastructure, offerings, customers and financial viability
- Key questions that were investigated:
    - What/who is inside/outside the CoE?
    - What is the demand/market?
    - What should the blueprint from the CoE?

Session Three

Opportunities for CoE Offerings: Ruben Riestra, INMARK

- General discussion of the opportunities that should be investigated particularly by people looking to bid for H2020 funding.
    - Potential for funding in the region of 115000 members of staff
    - Lots of untapped markets for digital preservation work in cultural heritage sector and well as more broadly e.g. nuclear power

SCIDIP-ES Services: Fulvio Marelli and David Giaretta

- An overview of the services that are being developed out of the SCIDIP project that have been tested in the Earth Sciences domain
- Services offered include:
    - Orchestration service (brokerage?) – notification of changes in context
    - Gap Identification service – identification of the implications of change
    - Preservation strategy toolkit – allow some estimation of the costs and implications of preservation actions
    - Representation Information Registry service – allow users to easily fill gaps in repinfo
    - Authenticity toolkit
    - Storage and packaging service
    - Data and Process virtualisation services
- Not all fully developed, some still need more work. All available as open source software.

Introduction to APARSEN Standards Database: Hans-Ulrich Heidbrink, InContec

- VCoE can be a unique neutral platform for sharing knowledge on DP, this is particularly relevant to input to standards and may provide a useful channel for contributions to development of standards
- Have been collecting a variety of information about standards: about the standards themselves as well as the scope of standards, relations between them and gaps as well as projects that have had standardisation activities.
- This information has all been assembled in a database: http://fenugreek.fernuni-hagen.de:8080/StandardsWeb/home/standardsRegister.xhtml

Session Four

Digital Preservation at the Exa-Scale - Challenges of the Next 10 Years: Jamie Shiers, CERN

- Is it realistic for service providers to say they can provide a long-lived service?
- Must understand the costs of exa-scale digital preservation, a challenge to a number of projects
- Significant economies of scale in shared data repositories
- Need huge improvements in efficiency and speed to be able to handle the amounts of data that are project, more sites in a network will not be enough
- Very large volume tape storage systems are very intolerant to things like dust. Require 'clean rooms' to ensure reliability
- Current data management solutions will not scale to the volume required. Will require much more automation.

Session Five

APARSEN Training: Kirnn Kaur, BL, Sharon McMeekin, DPC et al

- At beginning of project APARSEN did a comprehensive gap analysis of current training provision versus need
- Produced a list of topics not currently covered and 8 simple recommendations for training providers:
    1. Mix theory with experience-led training
        i. Case studies welcomed
    2. Include practical exercise & worked examples
    3. Need a mix of courses covering broad issues and more specialised content
    4. Targeted training for specific audiences
        i. By sector or role
    5. A continuing professional development framework is required
    6. Need to be more responsive to needs of the community
    7. Responsive to new development, remains current and authoritative
    8. Embed DP training in broader information or risk management training courses
- CoE training material in multiple streams
    - Training events
    - Online Training – Online Training Portal (OTP)
    - Training material gathered previous projects
    - Training offered by partners
- OTP – contains 3 elements:
    - Register of training and qualification opportunities
    - Training material – mostly video or audio presentations
    - Full online training material – still under development
- OTP currently available for free to all users
- Training courses are structured as follows:
    - Pre-course test

    - Learning modules – AV content and reading
    - Post-course test
    - Certification of achievement
- Welcome other contributions to the OTP, in particular new entries to the registry of courses and from projects looking for a home for their training material
- Training content from APARSEN Members

CINI

- Developing training at formal qualification, face-to-face and online course leves
- Mostly focused on e-gov and cultural heritage domains
- Responding to a need for training in Italian

LIBER

- Active at other workshops and conferences
- Hold own digital curation workshops
- Range of training at LIBER conference
- Partner in FOSTER project – developing online training portal, gathering material from training events
- Hoping to support training courses to allow them to repeat and be more sustainable
- Encouraging people to get involved through calls for events, training

nestor

- Training and education one of many elements of the nestor network
- Have a yearly practitioner day on specific topics
- Also have a yearly nestor school that has presentations from experts as well as discussions amongst students, get formal qualification points from training

SBA

- Provide a variety of different training material/courses either on their own or at larger events
- Cover a number of specific topics:
    - Preservation Planning
    - Content Analysis
    - Software Escrow
    - Data Citation
    - Data Security

Session Six

CoE Consultancy Services: David Giaretta, APA and Ruben Riestra, INMARK

- APARSEN partners will offer a variety of different consultancy through the CoE, covering most aspects of DP
- Ultimately DP is a managerial issue, looking specifically at three parts of APARSEN Common Vision model: value, business cases and business models
- Need to be able to understand the intangible digital objects can deliver value through clear benefits

- **Business Plan** shows how you will deploy the **Business Models** which details how you will achieve **Value**

Computing Ecosystems and Digital Legacy: Natasa Milic-Frayling, Microsoft Research

- Dependent on software for access to digital objects and software is in the hands of the ICT industries which is a fundamental issue in digital preservation
- Companies are constantly focused on innovation so even contemporary can be unstable, needing continual patching
- Makes preservation of the digital ecosystem very difficult, need a new model for this type of work within the ICT industry
- User experience is becoming increasingly important, format migration will not be enough to preserve the apps of today
- Working on solutions in the cloud, including migration and virtualisation – demonstration of solitaire and Encarta on Windows 98
- Need to make demand clear for software companies, once they are aware of the demand they will work to fill it
- Virtualisation cannot be a complete solution, must be combined with migration and management of the ICT ecosystem

DANS Research Data Services and the APARSEN CoE: Rene van Horik, DANS

- DANS part of the Dutch National Academy, mandate to do outreach relating to research data management as well as offering a variety of data services
- Offer training on research data management which encompasses material on digital preservation, including online training material: http://datasupport.researchdata.nl/en/
- DANS services incl a repository (EASY), scholarly info gateway (NARCIS), research data site and PI resolver
- Sharp rise in both number of data sets deposited and instances of data reuse in the DANS repository
- Have implemented a federate data infrastructure in a similar framework to that described in the Riding the Wave report with levels relating to a technical infrastructure, back office and front office

Session Seven

Pericles Project: Simon Waddington, KCL

- http://www.pericles-project.eu/
- Doing case studies within a variety of disciplines including space data and digital art.
- Must have flexibility in the types of licences offered as different creators (particularly in digital art) are happy with different levels of interaction with the objects/software they create
- Preservation often seen as an ‘end of life’ activity, Pericles trying to move away from this point of view, thinking of content changing continuously over time.

- ‘Preservation by design’ – looking at capturing information about the environment and context in which the object was created. Starting with what they are calling ‘Significant Environment Information’
- First tool – PERICLES Extraction Tool (PET) – Automatic collection of SEI information. Aims to be generic, modular and domain agnostic. Collection by observation, unstructured workflows.