The Future of the Past: May 2011

1. Introduction

On 4-5th May 2011 the EC hosted an invitational event to discuss and consult on future plans for digital preservation in Europe. It sought to look beyond the current Framework Programme and establish priorities for research in digital preservation for the next (FP8). William Kilbride attended on behalf of the DPC though there were various members present in their own right – BL, JISC, DCC, Portsmouth University - there were 64 people present.

The action plan for FP8 is progressing rapidly and the first complete draft will be complete within a couple of weeks. Suggestions for priority additions can still be made but suggestions need to be submitted within the next two to four weeks if they are to make the first draft. DPC members are therefore encouraged to review this document and contact WK early if they wish to make proposals or amendments. See section five.

These informal notes are intended to give DPC members an informal briefing. They are not intended as an official record: an official report will be published by Clive Billeness and published on the EU website. DPC thanks Neil Grindley of JISC for his contribution to the document.

2. Welcome and scene setting

Javier Hernández-Ros of the European Commission, (DG INFSO) introduced the day by laying out the medium term plans for the European Commission's investment in research and information technology. He noted that FP7 was already half way through and that there has been investment in DP research throughout this programme. Another call is imminent in January 2012 but the EU is already looking beyond this. The next framework programme will not be called FP8 – instead it will merge a range of research activities and it will develop a strategic framework for research and innovation. Javier is the newly appointed head of the Unit "Cultural Heritage and Technology Enhanced Learning".

3. EU-Funded research on digital preservation – past and present

Ross King opened the event by presenting an overview of previous and current research in digital preservation. It was based in part on a white paper about to be published by Andreas Rauber and colleagues.

The basics of DP remain the same. For example, the digital universe is expanding and the rate of growth is extraordinary. The trends in growth of data production - including born-digital and digitised data - are greater than the trends in data storage capacity. What's worth preserving (everything?)? How do we preserve it and ensure the quality of our preservation? How do we access the data? We know the challenges – hardware dependency, formats, contextual information and so on … and there are a series of approaches to answer this question including migration, emulation and redundancy.

There are 18 large scale EU projects on preservation, some of them long completed and some of them only now starting. Early projects like Erpanet (2001) were about awareness raising, moving then to frameworks of tools and services to manage data sets, projects like PLANETS or SHAMAN. The relatively simple data types have expanded and made more sophisticated data types have come on board with new projects. The newer projects have also typically addressed scale. The majority of projects have been concerned with 'how to' preserve, rather than the access, appraisal and scalability issues. Bit preservation is absent from the research, and therefore seems to be a fixed problem – but experts will tell you even this topic requires a lot more thought than has been given to it. Similarly only one project has really engaged with emulation. Distributed storage has been researched in the US and North America rather than in the EU. Linked data, security and trust, quality assurance and virtualisation have been dabbled with rather than fully explored. FP7 has already spent three times more money than FP6. The partners have been mainly scientific and research institutes, memory institutions and commercial suppliers. But industry has been missing from the list. This has been a perennial problem for the DP community in part because the economic incentives for DP are misaligned with the short time scales of the industrial community.

So what's next? There have been a number of research agenda and roadmaps published (DPE, Parse.Insight), and financial problems also need to be considered (LIFE and BRTF). Other topics like 'preservation as a service'

seems like a natural extension of storage as a service. Preservation-ready file storage and point-to-point integrity checks are important. Databases have hardly been worked on – is export to XML really the only way forward? Software preservation looks interesting as a feature of virtualisation.

A new wiki – http://sokrates.ifs.tuwien.ac.at/wiki/index.php/Main_Page - emerged from a small invitational event in 2010.

Comments from the floor and discussion after the presentation:

- One of the problems is that it's hard to trace the past projects, especially the older ones and there's a great deal of fragmentation between sectors. For example library or audio-visual communities are not brought together well even though they are working on similar problems.
- We need to think about standards and interoperability more. The demonstrators and proofs of concept will not impress industry alone. It's not surprising that they've not yet been adopted more widely. We have a major set of assets here and we can use this now – the timing is right to look now at the return of investment.
- Selection of digital information will be required. Should we work really intensively on relatively simple data sets that we can solve or should we work slowly across a much broader front. Focus on the main forms and means of communication will provide a satisfactory basis for 'most' DP needs. Web archiving exemplifies this.
- The costs and responsibilities for preservation cannot be entirely borne by content creators. There needs to be a balanced relationship between industry and memory institutions which aligns the tools, means and incentives for preservation.

4. Discussion sessions

The large group split into small groups and was presented with three questions which could inform the agenda for the next day. The following notes summarise the discussions in WK's groups.

Q1: How do you see the digital preservation landscape today?

- Community growing quickly: a dynamic community which has made dramatic progress in 10 years.
- Too much focus on problems, not enough on opportunities and benefits that accrue from long term access.
- The landscape is fragmented and has too little impact from previous research.
- Too much jargon – a barrier to participation
- Too much theory and not enough practice
- Too little fundamental research
- Too few people to fill the gaps between technologists and collections managers
- Although DP training is high quality there is too little understanding of fundamental digital technology by students and practitioners which means that the specialised DP training is often misaligned.
- Co-operation within EC has been very positive but co-operation with US and north America has been surprisingly weak.
- Need to invest in people as well as technology – that also helps with dissemination and return on investment in technology

Q2: How do you see the future of digital technology? What would you like DP to be like in 2020?

- Less niche, more mainstream
- Less jargon, more clarity
- DP as a regular background knowledge of staff – like using the web or email.
- Need to have had industrial scale examples of migration and emulation
- More research in virtualisation and software preservation
- More research and experience of preservation of complex objects
- More effort to enable confident deletion and de-selection
- More effort and engagement in the green ICT agenda
- More integration with forensics and adaptation of forensic tools
- Greater clarity from services like Amazon, YouYube, Facebook, Flickr etc about what will happen to the data that they serve up to the community
- Need to crack the 'demand-side' of the DP problem – by developing preservation ready systems

- Better diagnostic tools to anticipate DP problems before they arise

Q3: provide 2 research topics that the EC should invest in over the next decade:

- Automation
    - Automation of preservation processes
    - Automation of quality assurance
    - Appraisal, de-duplication and de-selection
    - Systems that scale up as well as down, especially for small institutions
    - Characterisation, metadata extraction and forensics
    - Transparent (ie invisible!) preservation services
    - Deployment of preservation-ready systems
- Models that are applicable and realisable
    - Models that helps us to preserve complex digital objects
    - Metadata modelling
    - Practical and achievable workflows
    - Hybrid models of preservation: migration and emulation/ virtualisation; preservation on demand
    - Tool evaluation and analysis of service dependencies
    - Modularity of tools and components to ensure complimentarity of existing and new tools

A round up at the end of this session provided 32 core themes for discussion:

- preservation-aware systems, on demand access and procedures
- new approaches for bit preservation and redundant storage
- emphasis for framework for dp across the EU - speak to the community
- more basic tools for use by anyone
- integrated access across time, systems and communities - understanding reuse better
- trigger for preservation actions, evidence for doing dp, value, risk, costs, benefits, impact
- marketing dp solutions, standardisation
- systems that preserve knowledge
- systems for quality assessment, certification, transparency
- services to curators, consumers, producers
- preservation of complex objects- the amount and complexity of data is increasing
- personal archiving - It will touch every citizen in the EU. It's about making DP a commodity. Promote it as a social practice
- anticipatory preservation planning - modelling and forecasting preservation issues before they arrive. responding to format obsolescence - doing some forecasting in advance
- integrating preservation into the lifecycle of digital objects. Still scope to do this, particularly the tracking of provenance in digital environments
- move away from problematizing DP and treating it as an engineering problem. Treat it as a process and think about what is good about DP
- semantic and dynamic objects. Can't pin down these ideas like butterflies, we need different ways of describing content. Do it from search and discovery pov and see what that means for DP
- shift the way we think about DP. It's a solution not a problem.
- managing objects in a lifecycle way: want to make preservation easy and not to worry about it. Objective is to make it as simple as possible
- not just preserving content but also preserving systems and processes
- intelligent techniques - improved techniques for appraisal, description. Need automated processes to end up with better metadata. analysis of the structure of objects.
- really reliable and credible access to the whole gamut of digital objects. With persistent identifiers and to work with computer scientists and academics to embed this at source
- develop an integrated emulation and preservation system that is widely tested and to thoroughly evaluate the user experience. Investigate hybrid preservation/emulation/virtualisation approaches
- virtualisaiton of access and reuse. Make access and reuse easier.
- autonomic self-preserving systems. Robust learning systems that advise users on the best course of action for preserving collections

- defining a self-preserving object able to check its own integrity
- simplicity - demonstrating a transparent non-custodial system. And defining where this might be applicable. Should be low-cost and highly scalable and should be integrated into the bas business processes. How far can you go with being very simple.
- reformulate dp problems as key computer science questions. A design principle for digital objects.
- once we have this data in the archives, what do novel methods look like to analyse these data once they are stored.
- looking at automation that breaks down into variety of tools: selection, appraisal, de-duplicating, id, etc . Could scale up and scale down for large and small institutions. Forensics for md and transparent and invisible preservation
- a set of portable models that are applicable and useful: for complex objects and md modelling: for workflows: for service dependencies. And return to some fundamental questions of dp such as is migration the way to go with the new technologies that we have or might emerge.

5. Day two: refining the innovation and research agenda for DP

The day started with a discussion of the many topics which had been raised as potential research topics in previous discussions. It was recognised that some of the proposals from the first day were duplicated or overlapping and moreover that some of the ideas were not in fact sustainable or appropriate to the EU. So the original proposals were narrowed down to 14 topics for research and subjected to scrutiny by break-out groups. A detailed report of the day will follow but the 14 high level themes were as follows:

- **New approaches to bit preservation** – no one wanted to join this discussion and it was dropped.
- **Extraction of preservation information** - Some need for coordination on the syntax level. hard for multimedia. the goal should be measured in scalability, interoperability. success would be the amount relevant extracted info and lower cost of tools and metadata. How make our life better. Create trust in digital objects. Might be an optimisation of EU funding by aligning with other initiatives going on.
- **Integrated access to digital resources** – time, system and community - Scenario of someone in 2300. Will they really be able to know what information was produced in 2011? People would have wider access to all sorts of data. There would be more competition for accessing data. If you open up the data and tell people what you are doing, you increase transparency.
- **Reformulate digital preservation as a computer science question in order that preservation-ready systems can be created -** network of sustainability in preservation. Want computer scientists and economists to work together. Take a long view of the digital info management problem. Roadmap for sustainable computing. no-one currently responsible for bringing this into our degrees. Want it to be a core module in courses. embedded into society, feeding into the knowledge economy. Engineering methods should be common in DP rather than social science methods.
- **Integrated emulation systems** - it's a challenge because we want to provide an easy user experience and will probably be difficult. Should be clicking on an object that then fires up the appropriate emulator. But there is a migration component in emulation as well. if an emulator is static then it will fail. Whilst the tech underpinning would change the ways of accessing it would need to reflect practice. Integration of virtualisation withe emulation and they should be more compatible. Need an emulation action right when the software is produced. Much more feasible and economic to do the emulation early. Vital to raise general awareness of emulation as a strategy across all communities and at different levels. Computer museums may be able to maintain documentation about technical environments. coordination action or network of excellence to raise awareness. Then projects to tackle issues identified by the first step.
- **Knowledge preservation** - Lots of talk about representation information: got to understand the disciplinary area. things change over time and we have to do our best to link all this stuff together. There is sufficient disagreement about what knowledge is needed so a supporting action is required followed by an IP. (some people in the know are suggested specific funding instruments), STREPS and SME's subsequently. It would be easier to tackle the mass of data out there. Would allow us to link to more things but it would stop us drowning by helping is to link to more specific and relevant bits of data. It would allow users of information to be much more discerning.
- **Quality assessment** – influential across every other topic. Especially important just now given mass digitisation but not only about digitisation.

- **Preserving complex digital objects** - current dp practices and tools can't cope with simple objects. EC should stimulate industry interest and play a pathfinder role. Clear need of a development of models and data warehousing, we need to get as close to the creation of objects as possible. The dp tools should be as easy to use as anti-virus tools. We need to promote engagement with end-users. They can be a prompt to industry to develop new tools and services. And what are the ey features of digital preservation. Face shift needed to think about complex objects.
- **Automation** - The scale of dp means it cannot be done by human actions. QA will become interested if automated. knowledge management and most other topics are dependent. (Except self-preserving objects where it is built in). Automatic means hi-throughput and we need to be able to scale up for 5-10 years time. There are techniques like rule-based systems - can start that tomorrow. Some things could be dealt with by artifical intelligence. Success indicator is that large amounts of data can be handled economically. Should be affordable by all types and sizes of organisation. And useability. automation should hide complexity. should reduce costs and should be broad takeup. The role of curators will change. They will move to the next layer of abstraction to make policies. There will be provenance assurance in matters of trust. Overall increase in archive quality and decrease in loss.
- **Ease of use – implicit and invisible preservation services – for personal archiving and mass market -** people are notoriously bad at backing up. We're talking about billions of people using no specific metadata with no expertise. No-one has their data in one environment. So the preservation needs to be implicit. Needs to be lo-cost. So the research question is what are the limits on taking effective action. what is the least you can for lowest cost that can feasibly and sustainably be carried out. Ease of use is bad description. Moving dp from the explicit realm of experts to the implicit realm of the wider public. How can we use infrastructure that is being built for other purposes. From industry perspective, need HCI experts to get it embedded. If this is noticed then it has failed - because it needs to be seamless. The research question is how to reduce this down to nothing.
- **Integration of DP in digital asset management** - why a challenge for research and innovation...still a lot of unresolved issues with complex objects. there are many things that need investigation, a lack of strategies and methodologies to take action. In engineering you trace problems with a change management system. how to calculate the financial consequences is still needed. The value of the digital object needs more research. To support policies. Must be composed as a set of individual component parts.It will improve business processes and improve the financial justification of those processes. You need to make this change visible. Needs early adopters and success stories.
- **Standards development** – eg metadata, persistent id's and certified access. And the need for a standards body. EU should exercise its role here. Why carry on research? The costs. There are various different types of stakeholders with interests in standards. Also the need to break down DP infrastructure to portable and replaceable components requires well designed interoperability which is likely to benefit from standards.
- **Markets and costs/benefits** - Develop simple tools for users to find the information. There are issues around paying for preservation. Need a economic model. The projects and programme need to be attractive for industrial partners. How would this change the world.
- **Self-preserving objects** - we need mechanisms to manage objects. preservation usually done at this level. This is the panacea of preservation. if we can attach it to objects we will reduce the costs of dp. it's the holy grail for researchers. We need to get software companies on board though. a consortium model in Europe is a powerful way to go. The timing couldn't be better because we now have tools that we didn't have during the previous framework programmes and there is a broad consensus that this is a good way to go. It is an engineering problem. And we can borrow from spyware and malware communities. stakeholders are everyone who uses digital objects. Will change the info lifecycle. Will change how we produce software and how we develop. we will encourage people to think harder about the objects as they design and create. Preservation self-awareness and impregnation programme! Going to be successful if it is opaque and we don't see it working. It may create wealth, if there are rights and micro-objects embedded in objects. This is a fantastic research question.

6. About this document

Version 1	Document initiated	04/05/2011	WK
Version 2	Document completed and distributed to colleagues present	05/05/2011	WK, NG, PW, CB, JD, DG
Version 3	Contribution from NG	06/05/2011	NG
Version 4	Distributed	06/05/2011	WK