#OurDigitalFuture – Multidisciplinary Perspectives on Long Term Data Preservation and Access

For any questions or to request re-use of the content of this report, please contact Clare Dyer-Smith (coordinator@bigdata.cam.ac.uk).

Purpose of the conference

As the worldwide volume of digital data undergoes exponential growth, Big Data technology allows unexpected value to be derived from existing and new datasets, and increasingly huge datasets to be recorded across all areas of academic research. As data volumes grow, and electronic storage deteriorates, the recoverability of this data is dependent upon curation of electronic archives and replacement of storage media, along with the ability to discover and access the data stored using technologies that may soon be obsolete. Decisions will need to be made about which data is kept, how it is stored, and how it can be accessed, in order that the scientific and human record from the current digital age is appropriately preserved for the future.

The focus of the conference was largely around scientific and research data, but also looked at the challenges in national archives and memory institutions. The issue of personal data archives was not looked at in great detail at this conference but is a concern in areas of big data research particularly those involving new forms of data, such as social media, online interactions and personal digital archives (for example email and photo) which have unique sensitivities, for example their vulnerability to changing commercial policies and discontinuation of services.

The specific objectives were

- To assemble a broad and diverse community of interest
- To identify key shared challenges and share knowledge and expertise in digital preservation
- To better define the required areas of research, including technology research
- To assess and define additional areas of training, education and skills development in long tern data preservation for science and research
- To inform the case for sustained investment in preservation and in education around preservation models and their associated cost

The range of disciplines covered in the talks included high energy physics, astronomy, infrastructure modelling, bioinformatics, libraries, archives, history, policy, medical research and law.

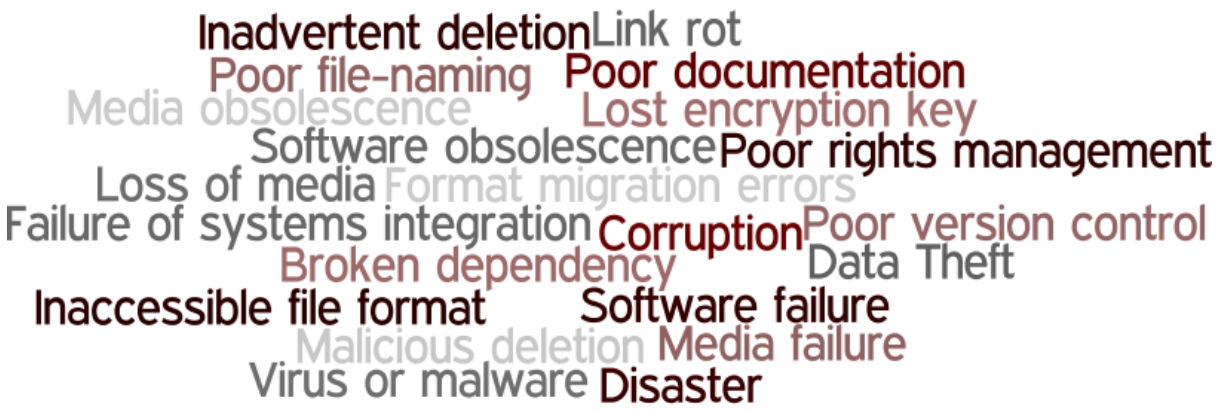
Videos of the keynote talks, and slides for the majority of the talks on both days of the conference are available at the conference website; this report is a summary of the key issues to be taken away and links to additional resources.

Why preserve, curate and manage data?

- Data are often expensive to create and their value may increase over time
- Data is valuable because there is so much of it and it can be combined in new ways
- Data re-use is emerging as a crucial component of scientific endeavour.
- Investing in data and software is an investment in infrastructure for the future
- Data and big data analysis are expected to increasingly drive decision-making – provenance and trust in those decisions presents technical and behavioural challenges

Specific challenges for big, digital and new forms of data

- There are several types of threat to digital data (see below for examples)
- New forms of data present new challenges for example ephemerality in and online data and the commercial nature of much personal online data resulting in uncertainty over future access and rights
- There are new challenges in big data for archives – going through 100TB collections of documents is not feasible in the same way as with paper, so new methods are required
- In several fields (particularly astronomy, high energy physics), simulated data volumes are as large as or larger than those from experiments. There is a lack of standards around simulated data and the requirements for preservation
- Data creation is itself often expensive: preservation and active management must be factored in as a lifetime cost
- There are potential technological and economic threats to long term models beyond ~20-30 year horizon (limited number of suppliers in the marketplace; breakdown of Moore's Law)



It's not just data! Software, workflows, documentation and dependencies

- Data without description, documentation and software are rarely useful
- There isn't a clear distinction between metadata and data – and both are vital
- There is a need to preserve: software, software environments (document and emulate or migrate), data and software dependencies, units, calibration information

- Scientific software in particular has wide variation in quality and consistency – scientists are software 'artisans', lack formal training, often write code in isolation and frequently write software as a means to an end and not an end in itself
- Provenance and workflow preservation are also vital

Collaboration across disciplines and stakeholder groups

- Many projects are multi-stakeholder, including private and public sector, with varying levels of engagement across project lifecycle
- As for many cross-disciplinary grand challenges, stakeholders must work together across disciplines, and across the public and private sector
- Archives and memory institutions have long experience of making decisions around data preservation and selection. Is there a distinction between scientific and research data, and digital heritage?
- Definitions of 'forever' and 'what to keep' are disciplinary/context dependent – one size does not fit all.
- Interoperability between disciplines is becoming increasingly relevant (data linkage, multi-modality – examples from medicine, infrastructure)
- Interoperability requires as aligned a policy or set of standards as possible – important to define community standards

Behaviour change and skills

- High standards (keep everything) are not achievable and should not dictate practice
- Determining what can and should be preserved is challenging and context dependent. Some disciplines' standards for determining data preservation needs are more straightforward or well developed than others'. Can we get better at predicting future uses and throwing the 'right' things away?
- Getting recognition for good practice by data and software citation can drive behaviour change in science and research. For example data citation gives credit where it is due, but allows others to access and build on the research of those generating the data. Embargo periods are applied in some disciplines (e.g. astronomy, bioinformatics) to balance data publication against priority of authorship for data creators
- Data producers know their data very well and with support can annotate it properly. But they don't often understand the technical requirements of storage systems and so require specialised support staff

Projects and resources

Data citation and repositories

- DataCite is a worldwide initiative providing persistent identifiers for datasets in the form of DOIs. DOIs are familiar to researchers and place datasets on the same level as articles, allowing finding, access, attribution of credit and establishment of provenance
- Digital repositories (e.g. Dryad) make the data underlying publications available to others including reviewers of the article during the peer review process
- ORCID identifiers to allow researchers to link all their work together; data as well as publications
- THOR is a H2020 project aiming to link persistent identifiers together
- Zenodo and Github are repositories for research outputs (broadly defined) and software respectively. A partnership between them enables code to be made citable
- HEPData is a repository of data from publications in High Energy Physics
- E-ARK provides open Access tools, services, metadata specifications, including data mining tools for business intelligence.

Workflow and model preservation

- FAIRDOM helps researchers to be in control of collecting, managing, storing, and publishing your data, models, and operating procedures
- myExperiment is a collaborative environment where scientists can safely publish their workflows and *in silico* experiments, share them with groups and find those of others
- BioVel offers "workflows" to process large amounts of data, particularly aimed at scientists in biodiversity conservation. It also provides tools for designing and running workflows.
- Workflow and History programming (GigaScience)
- ResearchObject aims to map the landscape of initiatives and activity in the development of **Research Objects**, an emerging approach to the publication, and exchange of scholarly information on the Web
- Workflow4Ever is creating an architecture and tooling for the access, manipulation, sharing, reuse and evolution of Research Objects in a range of disciplines
- **Common Workflow Language**, CWL (http://www.commonwl.org/) is designed to express workflows for data-intensive science, such as Bioinformatics, Medical Imaging, Chemistry, Physics, and Astronomy. CWL builds on technologies such as JSON-LD and Avro for data modeling and Docker for portable runtime environments.
- The Astropy Project is a community effort to develop a single core package for Astronomy in Python and foster interoperability between Python astronomy packages
- The CERN Open Data Portal releases (eventually large) subsets (copies) of the data with documentation, software and environment to run it
- PERICLES aims to address the challenge of ensuring that digital content remains accessible in an environment that is subject to continual change. This can encompass not only technological change, but also changes in semantics, academic or professional practice, or

society itself, which can affect the attitudes and interests of the various stakeholders that interact with the content

Data Provenance

- W3C working group primer on PROV Data model for provenance
- FRESCO – a Fabric for Reproducible COmputing tracing a complete history of each piece of data and the transformations it has undergone. This project (the work of the DTG in the Cambridge computer Lab) has developed many tools including:
- IPAPI – Improved Provenance API, a library for provenance support allowing the behaviour of complex systems to be traced.
- OPUS – Observed Provenance in User Space. This is an interposition system running in the background behind other programs, recording all I/O and major state events of a program. Implementation for research data allows the exact computations, data, software and processing of data to create particular outputs (e.g. figures for a journal paper) to be captured.
- HadoopProv: Provenance for MapReduce, operates across Hadoop workflow with low (<10%) computational overhead.
- Resourceful offers fine-grained resource accounting for the cloud with fine grained attribution to trace compute and overall usage levels.
- CamFort offers automatic evolution (refactoring, modernisation, porting to new run environments) and verification of code for computational science; addressing challenges in legacy code and systems, outdated languages and complexity in numerical codes.

Status reporting – examples from CERN

- Official Database of CERN Experiments: "The Grey Book" – Experiments, Institutes and Scientists
- DPHEP Portal: Access to Data Preservation Status of HEP institutes worldwide (and, where applicable, other portals)

Policy initiatives

- PasteurOA
- RECODE project
- Research Data Alliance – interest group on Active Data Management plans

Other activities, organisations and conferences

- Software Sustainability Institute
- Digital Preservation Coalition
- UNESCO Persist
- Open Science, Open Scientists: 2017 workshop at CERN focussing on practical experience of **data sharing**, **re-use**, **reproducibility** of results, **linking** publications to data (and other objects with DOIs) etc
- Threats to Openness in the Digital Age