CAMBRIDGE BIG DATA

UNIVERSITY OF CAMBRIDGE

EPSRC
Engineering and Physical Sciences
Research Council

Our Digital Future - Multidisciplinary Perspectives on Long Term Data Preservation and Access

Day 1

Author: Dr Adam Farquhar
Affiliation: British Library

Title: Diamonds are forever. What about research data?

Abstract: In this talk, I'll highlight some of the goals and challenges around holding research data for the longer term. We'll consider what comes after there is robust long-term storage for the bits. There are fracture lines in the scholarly record where the evidence base for results reported in articles neatly fits. We'll look at ways to close these gaps. Even the best current practices leave enormous challenges for future data re-use. We'll look at recent technical developments that may help researchers without radically changing their workflows. Emerging infrastructure promises to help researchers get appropriate credit for the additional work that they do to make data re-usable now and in the future. We'll review some key international developments.

Author: Dr William Kilbride
Affiliation: Digital Preservation Coalition

Title: Past performance is no guide to the future value: remembering and forgetting in the digital age

Abstract: In the 1990's fashionable neo-liberal commentators confidently predicted the end of history, the result of complex social and economic forces, the end of the Cold War, and the triumph of the market. It made good headlines but it wasn't true. By the early 2000's a different group of experts were furrowing their brows, holding out the prospect of a digital dark age, the result of over-reliance on unstable technical paradigms. It made wonderful headlines, but it's not happened yet. More recently, these two narratives have fused in the guise of fashionable lefty economics. So we are now promised the end of capitalism resulting from the frictionless exchange: the tyrants of corporate greed be defeated by the twin forces of abundant, inexhaustible data, and open source everything. Perhaps. Perhaps not. Digital preservation - the process of ensuring that the value of digital resources can be explored and released beyond the boundaries of technological obsolescence, media failure or operator change - finds itself suddenly and unexpectedly relevant. Dull-but-worthy, our tools, techniques and approaches have been resolutely unremarkable. It's not that we have shunned the limelight: the limelight has shunned us. But now, pushed unexpectedly into the glare and glory of public policy debate, how will we be judged. Our time has come. What shall we say? Past performance will be no guide to future value.

Author: Professor Jane Winters
Affiliation: Institute for Historical Research

Title: Making sense of the archived web

Abstract: The web is an integral part of our daily lives, whether we are shopping online, booking cinema tickets, registering to vote or checking whether or not it is going to rain today. It is also of enormous importance to researchers in the humanities and social sciences: as the site of digitised historical material, as a primary source in its own right, and as a means of promoting and communicating research to the widest possible audience. It is hard to imagine how you would write the history of the late 20th and early 21st centuries without access to all of this data. Where once we had handwritten diaries, we now have blogs; letters are superseded by Facebook status updates; our newspapers have become major online resources; and carefully curated Flickr collections have taken the place of photo albums.

The archiving of this vast range of material is increasingly occupying national memory institutions such as the British Library and The National Archives in the UK. However, as things stand, we do not have the expertise, the tools or indeed the legal framework to allow us to exploit this invaluable resource effectively. This presentation will explore some of the challenges of working with the archived web, and focus in particular on the work of the Big UK Domain Data for the Arts and Humanities project.

Author: Dr Jamie Shiers
Affiliation: CERN

Title: Data Preservation for Re-Use: from tens of TB to tens of EB

Abstract: This talk discusses the challenges in storing, sharing and curating data from High Energy Physics experiments at CERN and elsewhere. Data volumes range from tens of TB to tens of EB and the target period for which the data should remain fully usable runs to a few decades. (For example, data from LEP, where the active data taking period ran from 1989 to 2000, is expected to be fully usable until around 2030 with the bits available for much longer).

The talk differentiates between what is needed to preserve the data (and the necessary “knowledge” so that it remains usable) from the infrastructure and services required to share the data with future users – sometimes for previously unknown purposes. It also discusses how we measure if we are achieving our goals, including the use of Active Data Management Plans, Certification and agreed metrics for knowledge capture and analysis reproducibility.

Author: Niklas Blomberg, Ph.D
Affiliation: Director, ELIXIR

Title: ELIXIR: Enabling European-wide sharing of data in the life sciences

Abstract: Open access to bioinformatics resources provides a valuable path to discovery. ELIXIR is identifying core data resources that are essential to the larger international community and is developing a robust framework to secure their long-term sustainability and accessibility. Some of these datasets are highly specialised and would previously only have been available to researchers within the country/state in which they were generated.

ELIXIR, a research infrastructure founded by 17 European countries and EMBL-EBI, has been formed to orchestrate the collection, quality control and archiving of large amounts of biological data produced by life science experiments. By coordinating local, national and international resources – hosted at over 120 institutes - the ELIXIR infrastructure will meet the data-related needs of Europe's 500,000 life-scientists.

The challenges in storing, integrating and analysing the data from modern biological experiments needs a coordinated effort that involves both national and international resources. ELIXIR is currently constructing a distributed e-infrastructure of bioinformatics services – a data nodes network - built around established European centres of excellence. This talk will discuss our approaches to handling, analysing and archiving large and also highly diverse data-sets. Furthermore the talk will discuss experiences in data integration and the need for establishing data-management plans within projects that address the issues of meta-data annotation and long term archiving.

Author: Carole Goble
Affiliation: The University of Manchester, UK; The Software Sustainability Institute UK

Title: Method Preservation: workflows and models matter

Abstract: Data preservation is, of course, important. But so is the preservation of method. In practice the exchange, reuse, reproduction and preservation of data-centric experiments requires the bundling and exchanging the experimental methods, computational codes, algorithms, workflows and so on along with the narrative and the data. "FAIR Research Objects [1]" are composite and evolutionary, just as research is not "finished": codes fork, data is updated, algorithms are revised, workflows break, service updates are released.

In the EU Wf4ever project we set out a Research Object Framework for preserving computational workflows, in particular those using remote and independently stewarded datasets and third party services [2]. This framework is the basis of the community effort on the Common Workflow Language and has been developed for models in the FAIRDOM System Biology Commons [3] and STELAR Asthma eLab [4]. The RO term has gathered momentum as the NIH BD2K program is building a Research Object Commons.

ROs are metadata objects for explicitly describing aggregations or packages of content: boxes of components, and assembling instructions, with a shipping manifest for what is in the box and where it is from. We specify the ontologies needed to construct manifests (aggregation and annotation) and to guide their content (checklists, provenance, versioning, dependencies). The RO container is implemented using off-the-self platforms, like Zip, BagIt, and Docker. The RO content is not all physically within but likely logically held outside – the containers have "holes" in them. We need ways of knowing where their content is and it has changed and identifiers to glue the whole thing together.

In this talk we will discuss workflow/computational method reproducibility and how the Research Object metadata framework helps preserve computational artifacts alongside their data. I will also raise the importance of the sustainability of software in the preservation landscape [5].

[1] Bechhofer et al (2013) Why Linked Data is Not Enough for Scientists, Future Generation Computer Systems, doi:10.1016/j.future.2011.08.004 http://www.researchobject.org

[2] Belhajjame et al (2015) Using a suite of ontologies for preserving workflow-centric research objects, Web Semantics: Science, Services and Agents on the World Wide Web, doi:10.1016/j.websem.2015.01.003

[3] Wolstencroft et al (2015) SEEK: a systems biology data and model management platform BMC Systems Biology doi: 10.1186/s12918-015-0174-y http://www.fair-dom.org

[4] Custovic et al (2015) The Study Team for Early Life Asthma Research (STELAR) consortium 'Asthma e-lab': team science bringing data, methods and investigators together, Thorax doi:10.1136/thoraxjnl-2015-206781 (https://www.asthmaelab.org)

[5] The Software Sustainability Institute UK, http://www.software.ac.uk

Author: Dr Ripduman Sohan
Affiliation: Computer Laboratory, University of Cambridge

Title: Guaranteed Computing

Abstract: Big data could mean big problems if results are outdated, wrong or inaccurate. In this talk I will outline the problems of, and principles required to support guaranteed or reliable computing in future computation systems operating on big data with emphasis on reproducibility and verification. I will also outline some of the work being carried out at the Cambridge University Computer Laboratory focused on providing support for guaranteed computing.

Author: Victoria Tsoukala, PhD
Affiliation: National Documentation Centre/National Hellenic Research Foundation

Title: Policy Responses to Research Data Preservation Challenges

Abstract: As a result of the data deluge researchers, their institutions and funders are faced with the issue of preserving research data in the long term, in view of enabling their maximum reuse. At the same time, it is clear that long-term preservation has not been at the center of focus with respect to research data, and policies addressing the issues of long-term preservation are far from mainstream. The presentation discusses current challenges related to long term data preservation and good practices and policies adopted by national and international stakeholders in ensuring data preservation. An overview of current policies and good practices is provided to suggest potential directions, especially in the planning and policy realm. In doing so, the presentation also discusses related work from the RECODE project (www.recodeproject.eu)

Author: Dr David Giaretta
Affiliation: Giaretta Associates

Title: Data preservation Policies: from creation to exploitation

Abstract: Data preservation presents special challenges, many different from those presented by documents. Additional information must be collected during its creation in order to ensure that data is preservable, and that its preservation may be justified though its exploitation. This presentation will describe policies which support these requirements.

Author: Dr Fiona Reddington, Head of Population
Affiliation: Prevention and Behavioural Research Funding at Cancer Research UK

Title: CRUK and Big Data – past, present and future

Abstract: The avalanche of data and information emerging from cancer research in recent years means that the application of computational tools in cancer research and cancer care has become a vital and rapidly developing field. Bioinformatics has primarily developed to address this for high throughput or data rich systems such as genomics and proteomics. However, large datasets are also produced in cell biology, physiology, pathology, imaging, therapeutics, clinical trials and epidemiology. We are currently not in a position to make maximal use of the existing or future data sets generated by cancer research.

There is a unique opportunity to integrate all these different types of data and apply them systematically to improve understanding of cancer and its management and support the commercial exploitation of discoveries emerging from research. This talk will highlight some of the current, challenges and the potential for partnership

Author: Tim Gollins
Affiliation: National Archives/National Records Scotland

Title: The parsimonious anthropologist OR How string figures and a friar from the 14th Century might inform the preservation of digital research data

Abstract: In this paper I will try to take a step back from the detail of current Digital Preservation models and Digital Preservation systems and present some thoughts around Digital Archives that have been emerging in the last few years. This will enable me to pose a number of questions and challenges to what might be regarded as conventional thinking in this domain. I will use the ephemeral nature of String Figures to provide a tangible thought experiment to consider what should be preserved. Then I will then use the idea of the Principle of Parsimony to challenge the trend to increasing complexity and cost in preservation systems. This will implicitly re-balance the value-cost equation that is always at the heart of memory institutions. I believe these perspectives will be able to inform some responses to the challenges faced by data repositories in the context of Long Term Data Preservation and Access.

Day 2 - Workshop 1: What should we keep? Lessons from history for the shift to digital

Author: Dr Anthea Seles, Digital Transfer and Records Manager
Affiliation: The National Archives

Title: Understanding our digital realities to plan for our digital future

Abstract: The presentation will provide an overview of early digital preservation practices in order to contextualise how we best move forward in the brave new digital world. It will touch on what are the feasible expectations we can have when trying to preserve bits and bytes for future access, the impact of crisis of trust around digital and finally considerations as we move forward in the era of big data.

Author: Dr David Willcox and Lucie Jordan
Affiliation: The National Archives

Title: New challenges in finding and transferring digital government records for public access: Appraisal, selection and sensitivity review

Abstract: Born-digital records pose fundamentally different challenges for UK government compared to paper records. Methods for appraising, selecting and sensitivity reviewing paper records for transfer to The National Archives are long understood. However, the recently piloted born-digital records transfer process shows that high volumes coupled with poor structure is a challenge even for starting to make sense of collections. Add the fact that born-digital records will also need to be sensitivity reviewed before they can be released to the public and the challenges appear overwhelming.

The presenters will outline these major challenges and discuss the exciting opportunities, provided by technology, currently being explored by The National Archives. These opportunities could see an even more accurate process than the one established in the paper world.

Author: Dr David Erdos
Affiliation: University of Cambridge

Title: The Challenges for Long-Term Data Preservation and Use under the General Data Protection Regulation

Abstract: The recently agreed EU General Data Protection Regulation will put in place some of the strictest rules in the world on the collection, storage and dissemination and use of personal information.

At the same time, it also includes special provisions for research and archiving in the public interest, options for wider national restrictions and provisions to protect freedom of expression. This presentation will take a first look at this new framework and what it may mean for the future of long-term preservation and use of personal information.

Author: Rachel MacGregor
Affiliation: Lancaster University

Title: Blurring the lines

Abstract: The infrastructure and technical support required for digital preservation has much continuity with the processes for managing physical archives. In terms of access and discoverability there are significant differences in approaches required for the digital and the physical. However the impact of the digital on discovery methods is probably more significant than actual format and it is here that a break from the past is most urgently required. I will look at these challenges and how a diversity of approaches from different disciplines, blurring the traditional lines between archivists, librarians, data managers etc could bring great benefits.

Author: Tim Evans
Affiliation: Department of Archaeology, University of York

Title: Twenty years of digital curation at the Archaeology Data Service: challenges for archive and access

Abstract: Founded in 1996, the Archaeology Data Service (ADS) is a discipline-specific digital archive. After twenty years, the organisation curates over 320,000 digital objects comprising over 2 million files and amounting to 10Tb of data. As the amount of data held continues to increase, so do the challenges of ensuring preservation, management and access. The paper presents practical case-studies from the ADS, focussing on our procedures for ensuring data integrity, and the result of a recent migration of vector images. The paper also examines accessibility and re-use, as well as the practical challenges of presenting large datasets.

Author: Alex Taylor
Affiliation: Division of Social Anthropology, University of Cambridge

Title: Bunkering Data: Sowing the Seeds of the Digital Future?

Abstract: Drawing from my ethnographic research on the emergence of subterranean data centres, I will discuss how the increasingly commonplace practice of preserving data in ‘future-proof’ underground bunkers connects to cultural anxieties surrounding the fragility and ephemerality of the digital and how digitisation/data preservation projects shape (and are shaped by) imaginaries of a digital future that is haunted by the spectre of disaster – from routine Disaster Recovery plans to the science-fictional long-term dormantisation of data in the event of a civilisational crisis.

Author: Jenny Bunn
Affiliation: University College London

Title: Preserving collections and samples in the era of Big Data

Abstract: In speaking of a ‘common assumption that only a small portion of that whole can be preserved’ it is also necessary to uncover our assumptions about the relationship between wholes and parts of wholes. Archivists have long sought to preserve wholes that are ‘greater than the sum of their parts’, but they have not always been clear about what that means. Melding archival theory with concepts from Cybernetics, I will outline how the whole we strive to keep is not synonymous with keeping everything, but with managing and using data in a certain way.

Day 2 - Workshop 2: Current and Future perspectives on technology for data preservation and sharing

Author: Ana Trisovic
Affiliation: University of Cambridge

Title: Data Preservation at CERN

Abstract: The LHCb experiment at CERN currently has two active projects on data preservation. The first one is a data dependency database, a system implemented as a graph database which tracks all information related to the LHCb data and its links to the software used to process them. In addition, it contains information about the compatibility between various software versions and the data. The second project is called CAP (CERN Analysis Preservation), which is a web-based portal that allows the logging of detailed information about an analysis from its inception to the publication. I would like to present these projects, their use cases and how they will help us rerun legacy software and reproduce old analyses in the future.

Author: Marc O'Brien
Affiliation: Cancer Research UK Cambridge Institute

Title: We will keep it forever (from a tin perspective)

Abstract: The CRUK Cambridge Institute generates genome, multi-modality imaging and histopathology data, and is rapidly approaching 2PB of stored data. Here we discuss a scientist driven, ‘keep forever’ approach to data that supports publications.

Author: T. Masood, G. Yilmaz; D.C. McFarlane; A.K. Parlikad
Affiliation: Distributed Information and Automation Laboratory, University of Cambridge

Title: Long-term Data Preservation and Access – Infrastructure Assets Perspective

Abstract: Long-term data preservation and access is challenging across many disciplines including infrastructure assets that serve societies for long-term. However, asset information is lost due to technological and organisational challenges and disruptive events over long time. Even though organisations take some actions but largely mechanisms for information future-proofing are missing. A number of semi-structured interviews and workshops were conducted with leading organisations dealing with infrastructure assets to understand the challenges involved. An information future-proofing assessment approach is presented here, which has been applied in case studies of bridges and structures, a department building, and underground transport infrastructures.

Author: Professor Richard McMahon
Affiliation: Institute of Astronomy, University of Cambridge

Title: Data preservation issues in astronomy

Abstract: I will review the progress, challenges and lessons learnt in ensuring long term preservation of astronomical data ranging from photographic astronomical data prior to the 1980's, through the digital age. I will cover the approaches used by astronomers working at different wavelengths from radio to the X-rays and both ground based and space based observations. Key issues are metadata, data compression and software than can read the data. As well as preservation I will describe progress in data discovery. Whilst I will primarily focus on astronomical data many of the lessons learnt (some painful and recurring) are relevant to other domains.

Author: Ripduman Sohan
Affiliation: Computer Laboratory, University of Cambridge

Title: OPUS - Keeping Track Of Your Research Data

Abstract: EPSRC's open data requirements now mean it's necessary to track the data and metadata involved in the creation of every publication. Manual collection of this information is cumbersome, tedious and error prone while adding automatic collection is likely to be a domain-specific, time consuming and expensive exercise.

At the Computer Laboratory we have developed a system called OPUS that assists with the task of data and metadata collection for Linux applications. OPUS seamlessly and transparently monitors the programs run on machines and records a variety of information such as the files that were accessed and the user running the program. This is achieved with negligible performance overhead.

We have adapted OPUS to satisfy the EPSRC open data requirements.

During the talk we will introduce OPUS, show you how it can help you to meet EPSRC's requirements and outline our plans for the future.

Author: Dr Simon Waddington
Affiliation: King's College London

Title: PERICLES – Management of change to enable long term reuse

Abstract: We will describe recent work of the EU FP7 PERICLES project on digital preservation, which is developing models, methods and tools to deal with the management of change to enable long term reuse of digital content. Motivated by examples in space science experiments and time-based media art, we focus on the preservation of complex digital objects, which include not only final datasets, but also intermediate data, software, documentation, policies and user communities. The talk will outline our approaches to dealing with technological, policy and semantic change, based on capture and modelling of the environment.

Author: Sven Schlarb
Co-authors: Janet Delve, Rainer Schmidt, Richard Healey
Affiliation: Austrian Institute of Technology

Title: The Use of Big Data Techniques for Digital Archiving

Abstract: The background for this paper is work in progress in E-ARK: an EC FP7 pilot B project[1] having as its main objective the creation of an open source, digital archiving system with attendant standards and tools to be deployed in seven pilot instances. Hence practical application is at the heart of the project, which is led by archivists, researchers, SMEs, digital preservation / archiving membership organizations and government home offices, who together seek to fill the current digital archiving lacuna. E-ARK is a wide-ranging project: we are taking and integrating existing best practices into a digital archiving system, so that it is suitable not only for national archives and government agencies, but also for regional, local, business and research archives of all shapes and sizes. A legal study taking account of varying national legal directives delineates how the archiving system can be deployed against a pan-European backdrop.

[1] E-ARK is funded by the European Commission's FP7 PSP CIP Pilot B Programme under Grant Agreement no. 620998.

Author: Professor Frank Krauss
Affiliation: University of Durham

Title: HEPData - Long Term Data Preservation in High Energy Physics

Abstract: In this talk the HEPData project will be introduced, which is the central long-term data preservation facility for data from collider-based experiments in particle physics. The talk will introduce some of the challenges in providing such a service and emphasize the changes in the different roles of users and the actual service providers since HEPDatas inception about 30 years ago. Some strategies and ideas for extended data content and services will also be discussed.