Datasets, Open Data and Digital Preservation (full day)

David Tarrant, University of Southampton, UK, davetaz@ecs.soton.ac.uk
Carl Wilson, Open Planets Foundation, carl@openplanetsfoundation.org
Rebecca McGuinness, Open Planets Foundation, rebecca@openplanetsfoundation.org

At the core of digital preservation is the aim of archiving the public record to be available for future generations. This record consists of thousands of different types of records, files and artifacts and digital preservation approaches both help preserve and expose this record to a wider audience. We cannot question the role that that web has played opening up an unimaginable amount of resources to people of all creeds, genders and race.

Over the years many significant projects have been established as preservation hubs. Interestingly the vast majority of these are government (or public) funded, e.g. National Archives and Libraries, Educational Institutions and other Government Organisations (e.g. Office of National Statistics etc). Each of these organisations is collecting and preserving a vast amount of content, much of which the general public is yet to gain ubiquitous access to.

Governments around the world are now demanding that the public be able to access data and resources freely, driving up transparency, improving trust and also saving money at the same time. The UKs Open Data movement[1] and the recent executive order[2] mean that organisations are now mandated to open up not only their content, but also the raw data.

This movement is not limited to government organisations however. In the academic community, funders and publishers are now demanding publication of supporting data and datasets as well as just the final publication[3,4]. The benefits of being so open and transparent are already clear. In the government sector there have already been many examples of where data can lead to huge savings[5] and boost the economy at the same time[6]. In the academic sector, opening up data allows results of experiment to become reproducible, allowing us to avoid potentially dangerous misinterpretation of data[7].

This new influx of data as well as mandate to make current data open will provide a substantial challenge to many organisations. However, with digital preservation practices well established, this presents an incredible opportunity to push digital preservation practices out to a much wider audience. With the open data movements focussed on the web as a platform, there is a key opportunity to crash the two communities together for the good of all involved. In this tutorial we are going to do just that!

1 - Web creator backs UK Open Data Institute (http://www.bbc.co.uk/news/technology-18161759)
2 - US Executive Order (http://cdn.govexec.com/media/gbc/docs/pdfs_edit/050913jm1.pdf)
3 - Journal Open Data Policies (http://oad.simmons.edu/oadwiki/Journal_open-data_policies)
4 - Turning Government Data into Gold (http://europa.eu/rapid/press-release_IP-11-1524_en.htm)
5 - Millions in prescriptions savings identified (http://tinyurl.com/qyfv83c)
6 - Review of PSI Re-Use Studies (http://epsiplatform.eu/content/review-psi-re-use-studies-published)
7 - Paper disputing basic science of climate change is fundamentally flawed (http://tinyurl.com/3jjdg85)

Tutorial Overview

This full day tutorial focuses on dataset publication and the lessons learned in the digital preservation community as well as the directions that the web standards bodies are following. By bringing together top speakers from the digital preservation and open data communities we introduce techniques for dataset publication, management and long term digital preservation.

Suitable for all attendees we look at both the technical and non technical aspects surrounding dataset publication and gives hands on experiences of a number of best practice techniques and tools that can help assist the processes of long term data management.

At the end of the tutorial attendees should have a greater understanding of what open data mandates and executive orders mean and know how to approach the problems that will be encountered along the way. Further, this tutorial will address the opportunities for managing these new requirements with existing digital preservation methodologies, practices and platforms.

Learning Outcomes

By the end of the course you will:

- Have an overview of open data publishing and analyse the application potential
- Understand the changing requirements for dataset publication, executive orders and mandates
- Be able to apply digital preservation tools and techniques to dataset management and archiving
- Evaluate a number of best practice approaches for dataset classification to help with key decisions in the preservation and publishing process.

Session Overview

Session 1: Data publication: Apparently i'm the expert!?!

This first session introduces the drivers behind the open data movement and aims to allow attendees to discuss developments around europe to get a grasp on the current state of the art and benefits to the public. By introducing a number of key projects and organisations we look at some of the best practices for both identifying and publishing properly licensed and protected datasets.

Session 2: Machine readable: Isn't a PDF and a ZIP file enough?

In the second session we look at the current best practices and guideline standards that can help you get started on the path towards perfect dataset preservation. We look at why the 5-Stars of linked data is not suitable to guide the process of its own and how the new Open Data Certificate, which integrates a great many digital preservation principles, can help.

Session 3: So I have some data, now what?

This final session introduces attendees to the notion of using a version control system (in this case GitHub) to manage the process of curating and managing a dataset for long term preservation. We look at how the features provided by such platforms directly map to the needs and requirements laid out in the previous sessions. To conclude this session we shall look at the next steps and show an example of how this approach can be used as part of an automated testing process for digital preservation tools, where test data is published as Open Data and shared with the Digital Preservation community.

Speakers

Dr David Tarrant

David Tarrant is a Lecturer at the University of Southampton and Postgraduate Training Associate at the Open Data Institute in the UK. He is a leader of Open Data advocacy, responsible for main training courses and outreach events. He is also a long standing member of the digital preservation community in which he has worked on many data publication projects, most recently including the Results Evaluation Framework (REF) platform introduced at last year's iPres conference.

Carl Wilson

Carl currently works for the Open Planets Foundation promoting software development and testing best practises to members and providing technical input to external projects such as SCAPE. Prior to that he worked for The British Library's Digital Preservation Team on internal and external projects, including a brief spell as Technical Co-ordinator for the SCAPE project. He spent three years as a technical lead on the Planets Project, developing the Interoperability Framework and Service Interface definitions. He also helped organise and run the Planets Service Developer's Workshops. Carl particularly enjoys Hackathon style events, and is a regular attendee at the OPF's events, as well as those organised as part of the SPRUCE and AQuA projects.