Preserving Data to Preserving Research: Curation of Process and Context

Angela Dappert
Digital Preservation Coalition Senior Project Officer
DPC c/o The British Library,
Floor 5, Room 14, 96 Euston Road,
London NW1 2DB +44 (0) 20 7412 7028
angela@dpconline.org

Rudolf Mayer, Stefan Pröll, Andreas Rauber
Secure Business Austria,
Vienna, Austria
mayer@ifs.tuwien.ac.at
sproell@sba-research.org
rauber@ifs.tuwien.ac.at

Raul Palma
Poznan Supercomputing and Networking Center,
Poland
rpalma@man.poznan.pl

Kevin Page
University of Oxford e-Research Centre,
United Kingdom
kevin.page@oerc.ox.ac.uk

Daniel Garijo
Universidad Politecnica de Madrid,
Spain
dgarijo@delicias.dia.fi.upm.es

ABSTRACT

Awareness of the need to provide digital preservation solutions is spreading from the core memory institutions to other domains, including government, industry, SME and consumers. In many of these settings we are, however, faced with preserving more than just data. In the domain of eScience, for example, investigations are increasingly collaborative. Most scientific and engineering domains benefit from building on the outputs of other research by sharing information to reason over and data to incorporate in the modeling task at hand.

This raises the need for preserving and sharing entire eScience workflows and processes for later reuse. We need to define which information is to be collected, create means to preserve it and approaches to enable and validate the re-execution of a preserved process. This includes and goes beyond preserving the data used in the experiments, as the process underlying its creation and use is essential.

The TIMBUS project and Wf4Ever project team up for this half-day tutorial to provide an introduction to the problem domain and discuss solutions for the curation of eScience processes.

General Terms

Infrastructure, preservation strategies and workflows

Keywords

e-Science, data preservation, workflows, semantics, Research Objects, Context Models

1. TUTORIAL STRUCTURE

The tutorial will cover the following topics:

Introduction to Process and Context Preservation: The introduction will motivate the need for process and context preservation, illustrate how this task is difficult in an evolving domain, and introduce a common use case, based around the work of a researcher in Music Information Retrieval [1], which is used in the rest of the tutorial to illustrate approaches and tools for the rest of the tutorial to illustrate approaches and tools.

Data Citation: Data forms the basis of the results of many research publications, and thus needs to be referenced with the same accuracy as bibliographic data. Only if data can be identified with high precision can it be reused, validated, verified and reproduced. Citing a specific data set is not trivial, however: it exists in a vast plurality of specifications and instances, can potentially be huge in size, and its location might change. We will provide an overview over existing approaches to overcoming these challenges. We will also present the issue of creating data citations of data held in databases, especially of dynamic data sets where data is added or updated on a regular basis.

Re-usability and traceability of workflows and processes: The processes for creating and interpreting data are complex objects. Curating and preserving them requires special effort, as they are dynamic, and highly dependent on software, configuration, hardware, and other aspects. We will discuss these issues in detail, and provide an introduction to two complementary approaches.

The first approach is based on the concept of Research Objects, which adopts a workflow-centric approach and thereby aims at facilitating the reuse and reproducibility. It allows us to package the data, along with the scientific context information of how these resources were used or produced, as one Research Object, and thus to share and cite it. This enables publishers to grant access to the actual data and methods that contribute to the findings reported in scholarly articles.

A second approach focuses on describing and preserving a process and the context it is embedded in. The artifacts that may need to be captured range from data, software and accompanying documentation, to legal and human resource aspects. Some of this information can be automatically extracted from an existing process, and tools for this will be presented. Ways to archive the process and to perform preservation actions on the process environment, such as recreating a controlled execution environment or migration of software components, are presented. Finally, the challenge of evaluating the re-execution of a preserved process is discussed, addressing means of establishing its authenticity.

2. INTENDED AUDIENCE

The tutorial is targeted at researchers, publishers and curators in eScience disciplines who want to learn about methods of ensuring the long-term availability of experiments forming the basis of scientific research.

3. EXPECTED LEARNING OUTCOMES

The tutorial participants will become familiar with:

- Motivations and challenges of process preservation;
- Motivations, stakeholders and challenges of making data citable;
- How data is cited today, best practices, guidelines and metadata standards;
- Available technologies for identifiers: Archival Resource Key (ARK), Digital Object Identifiers (DOI), Extensible Resource Identifier (XRI), HANDLE, Life Science ID (LSID), Object Identifiers (OID), Persistent Uniform Resource Locators (PURL), URI/URN/URL, Universally Unique Identifier (UUID);
- Approaches and Initiatives for citing data: CODATA, Data Cite, OpenAire, challenges and opportunities: granularity, scalability, complexity and evolving data sets current research questions;
- Ontologies needed to capture research objects: Core Ontology of the RO family of vocabularies, workflow centric ROs, provenance traces, life cycle of research objects;
- Wf4Ever Toolkit / technological infrastructure for the preservation and efficient retrieval and reuse of scientific workflows: software architecture, functionalities, software interfaces to functionalities, reference implementation as services and clients:
    - Collect, manage and preserve aggregations of scientific workflows and related objects and annotations
    - Workflow sharing through a social website
    - Execution of workflows;
    - Testing completeness, execution, repeatability and other desired quality features;
    - Testing the ability of a Research Object to achieve its original purpose after changes to its resources;
    - Recommendations of relevant users, Research Objects and their aggregated resources;
    - Converting workflows into Research Objects;
    - Search for workflows by input parameters or frequency of use;
    - Collaborative environment;
    - Access and use of research objects and aggregated resources;
    - Synchronization with remote repositories;
    - Visualization of research object evolution;
- TIMBUS context model and tools to semi-automatically capture the relevant context of a business process for preservation:
    - The scope of context regarding business process preservation - technology, application and business context, aligned with enterprise architecture;
    - The context meta-model, with domain independent and domain specific aspects;
    - Demonstration of a context model instance of example processes (in the eScience domain);
    - Tools to automatically capture some parts of the context (software dependencies, data formats, licenses, etc);
    - Outlook on reasoning and preservation planning, based on the context model.

4. BIOGRAPHY OF THE PRESENTERS

Angela Dappert is Head of Research and Practice at the Digital Preservation Coalition. She also serves on the PREMIS Editorial Committee. In both capacities she is involved with the issues of modelling and defining metadata for computer environments. She has worked at the British Library on data carrier stabilization, digital asset registration, digital preservation planning and characterization, eJournal ingest, and digital metadata standards. Before this she worked for Schlumberger, the University of California, Stanford University and Siemens. Angela holds a Ph.D. in Digital Preservation from the University of Portsmouth, an M.Sc. in Medical Informatics from the University of Heidelberg and an M.Sc. in Computer Sciences from the University of Texas at Austin.

Daniel Garijo is a PhD student in the Ontology Engineering Group at the Universidad Politecnica de Madrid. His research activities focus on e-Science and the Semantic Web, specifically on how to increase the understandability of scientific workflows using provenance and metadata. He has been a member of the W3C Provenance Working Group, and previously participated in the Wf4Ever project.

Rudolf Mayer is a researcher at Secure Business Austria, as well as the Department of Software Technology and Interactive Systems at the Vienna University of Technology. His research interests cover digital preservation, specifically the preservation of processes, information retrieval (specifically on text documents and music), data analysis and machine learning. He has many years of lecturing experience in these subjects. He has been involved in the DELOS and PLANETS projects, and currently works on digital preservation aspects in the FP7 projects APARSEN and TIMBUS.

Kevin Page is a researcher in the Oxford e-Research Centre, University of Oxford, UK. His work on web architecture and the semantic annotation and distribution of data has, through participation in several UK, EU and international projects, been applied across a wide variety of domains including sensor networks, music information retrieval, clinical healthcare, and remote collaboration for space exploration. His current research focuses on the application of semantic web architecture to information management systems for scientific workflows, musicology, and social machines, and the common approaches that underly these seemingly disparate subjects. He has previously organized and presented tutorials at the Extended Semantic Web Conference, the International Society for Music Information Retrieval conference and the Oxford Digital Humanities Summer School.

Raul Palma is a researcher at Poznan Supercomputing and Networking Center (PSNC). His research interests cover digital preservation, particularly of scientific methods, provenance and evolution of digital artifacts, ontology engineering and distributed technologies. He has participated in several EU projects, including the Network of Excellence Knowledge Web, NeOn, e-Lico and WF4Ever. He has many years of lecturing experience in related topics, both at the university and private institutions. He has authored or co-authored several vocabularies and ontologies, such as the Research Object evolution Ontology, Ontology Metadata Vocabulary (OMV) and different extensions for describing ontologies and related resources, models for collaborative ontology construction and digital multimedia repositories

Stefan Pröll is a researcher at SBA Research. His primary research focus lies on digital preservation, especially on security aspects of digital archives, including authenticity and provenance of digital objects. Further areas of interest are databases and data citation. Currently he is working on FP7 projects APARSEN and TIMBUS focusing on security and provenance related topics. Before he joined SBA in April 2011, he was working in international organizations in the area of Web development, Linux server and database administration.

Andreas Rauber is Associate Professor at the Department of Software Technology and Interactive Systems at the Vienna University of Technology. He is involved in several research projects in the field of Digital Libraries, focusing on the organization and exploration of large information spaces, as well as Web archiving and digital preservation. His research interests cover the broad scope of digital libraries, including specifically text and music information retrieval and organization, information visualization, as well as data analysis and neural computation. He has been involved in numerous initiatives in the area of digital preservation (DELOS, DPE, Planets, SCAPE, TIMBUS, APARSEN). He has been lecturing extensively on this subject at different universities, as part of the DELOS and nestor summer schools on digital preservation, as well as during a range of training events on digital preservation.

5. REFERENCES AND CITATIONS

[1] Kevin Page, Raúl Palma, Piotr Holubowicz, Graham Klyne, Stian Soiland-Reyes, Daniel Garijo, Khalid Belhajjame, and Rudolf Mayer. "Research Objects for Audio Processing: Capturing Semantics for Reproducibility." In Audio Engineering Society Conference: 53rd International Conference: Semantic Audio. Audio Engineering Society, 2014.