From Preserving Data to Preserving Research: Curation of Process and Context

Rudolf Mayer[1], Stefan Proell[1], Andreas Rauber[1,2], Raul Palma[3], and Daniel Garijo[4]

[1] Secure Buisness Austria,
Vienna, Austria
[2] Vienna University of Technology,
Austria
[3] Poznan Supercomputing and Networking Center,
Poland
[4] Universidad Politecnica de Madrid,
Spain

Abstract. In the domain of eScience, investigations are increasingly collaborative. Most scientific and engineering domains benefit from building on top of the outputs of other research: By sharing information to reason over and data to incorporate in the modelling task at hand.
This raises the need to provide means for preserving and sharing entire eScience workflows and processes for later reuse. It is required 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.
This tutorial thus provides an introduction to the problem domain and discusses solutions for the curation of eScience processes.

Awareness for the need to provide digital preservation solutions is spreading from the core memory institutions to other domains, including government, industry, SME and consumers. Likewise, in the domain of eScience, the documentation and preservation of research processes, to allow later understanding and re-execution of e.g. experiments that may be the basis of scientific discoveries, is understood as an important part of the research, and thus gaining more interest. However, comprehensive solutions are still rarely applied. This tutorial is therefore aimed at providing a holistic view on the challenges and solutions of preservation of eScience processes.

As the very core of eScience, *data* forms the basis of the results of many research publications. It thus needs to be referenced with the same accuracy as bibliographic data. Only if data can be identified with high precision, it can be reused, validated, falsified, verified and reproduced. Citing a specific data set is

however not always a trivial task. Research and business data exist in a vast plurality of specifications and instances. Additionally, data sets can be potentially huge in size, and their location might change as they are transferred between different institutions. This tutorial thus starts with an overview of current data citation practices.

In some scenarios, the data base itself is *dynamic*, e.g. new data gets added on a regular basis, or existing elements were changed or deleted. Such settings pose new challenges to citation, as the cited source should be unchanged over time. We thus provide an introduction into the topic of dynamic data citation and present potential solutions for this area.

On top of the research data, the re-usability and traceability of *workflows and processes* performed thereon is vital for preservation. The processes 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. These aspects form the *contextual information* of the processes, and are thus the primary concern of preservation. This tutorial presents these challenges in detail, and provides an introduction to two complementary approaches to alleviate them.

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 packaging the data and the methods as one Research Object to share and cite it, and thus enable 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 artefacts 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.

Throughout the tutorial, the challenges and problem domain are demonstrated on practical examples taken from eScience research workflows. On these examples, the tools and methods presented earlier are applied, and the solutions are discussed for the adequacy.