Digital Preservation in Data-Driven Science: On the Importance of Process Capture, Preservation and Validation

Andreas Rauber[1]

Department of Software Technology and interactive Systems
Vienna university of Technology
Favoritenstrasse 9-11, 1040 Vienna, Austria
rauber@ifs.tuwien.ac.at

Abstract. Current digital preservation is strongly biased towards data objects: digital files of document-style objects, or encapsulated and largely self-contained objects. To provide authenticity and provenance information, comprehensive metadata models are deployed to document information on an object's context. Yet, we claim that simply documenting an objects context may not be sufficient to ensure proper provenance and to fulfill the stated preservation goals. Specifically in e-Science and business settings, capturing, documenting and preserving entire processes may be necessary to meet the preservation goals. We thus present an approach for capturing, documenting and preserving processes, and means to assess their authenticity upon re-execution. We will discuss options as well as limitations and open challenges to achieve sound preservation, specifically within scientific processes.

Keywords: Digital Preservation, Processes, Context, eScience

1 Introduction

Digital preservation (DP) traditionally has a predominantly data-centric view on both its operations as well as the objects it is dealing with. Digital objects considered for preservation are usually (turned into, as far as possible) self-contained, static objects such as images resulting from scans, classical document-style objects, data sets, but also information packages containing software and other in principle dynamic objects as encapsulated files. Furthermore, objects are usually "removed" from an operational life-cycle into an archival life cycle, ingested into designated repositories for long-term maintenance, from which they are removed and re-inserted into a potentially new lief-cycle when needed in an operational manner again, only to be re-ingested as new objects after completion of their new life as now new archival objects. Consequently, also cost estimation and investments are largely based upon aggregated data item-level information, adding up processing costs, storage costs and others.

We claim that this traditional view is hitting severe limitations as we are observing a set of interesting changes in the preservation community: Most importantly, preservation is expanding beyond the traditional cultural heritage

community. While originating in the sciences, recognizing the need to maintain the investment made into data collected electronically (as evident by the early and strong commitment of institutions such as NASA, leading to the infamous OAIS model) the key drivers, expertise and know how and development has come from and taken place in the cultural heritage community. The key characteristic of this community is the dedication to preservation of information as a, or even the, primary mission, resulting in a holistic understanding of the scope of the problem and its long-term implication beyond individual technical or organizational issues. Yet, more recently we see a range of other communities facing the need for digital preservation: back to the origins of DP, science as a whole is becoming increasingly dependent on data as the core facilitator in virtually all scientific disciplines, leading to trends defined as data-driven science, e-Science, Big Data [9], the Fourth Paradigm [6], and others. But even beyond cultural heritage and science communities, both of which have been involved in DP for a long time, we find entirely new players / customers in need of DP solutions, many of them coming from a range of industrial backgrounds, and with a quite diverse set of motivations. These may range from specific legal / compliance requirements, via somewhat more ambiguous risk mitigation desires to serving dedicated business needs. Thus, while in principle being similar to the cultural heritage sector, there are some interesting challenges stretching beyond the ones encountered in more traditional settings.

This paper starts in Sec. 2 with a loose collection of observations on changes in the DP community, highlighting three areas of focus that we deem important, namely a shift towards risk management, viewing DP in the context of e-Governance frameworks, and the shift towards the preservation of processes rather than data objects. This will be followed by a more detailed look at two key aspects, namely process preservation and a framework for evaluating the quality of re-execution of preserved processes in Sections 3 and 4, respectively (largely adopted from [16]), before providing a brief summary in Section 5.

2 Implications of Changes in Stakeholder Communities

The observed expansion in stakeholder communities has some interesting implications for the DP community: first of all, we are experiencing yet another clash of languages and cultures: after partially successfully consolidating the viewpoints of archival and library communities, merging the viewpoints of the museums community and even succeeding in getting computer science to listen and communicate on an increasingly shared level of mutual understanding, DP is currently being recognized, interpreted and contributed to by a whole range of new key players with completely diverse interpretation of the concepts widely accepted in the traditional DP community. Long-term may be as short as 7 years, preservation and loss is not necessarily measured on the level of the need for maintaining an object, but as best effort vs. risk trade-off, specifically not happening "at all costs" wherever possible, with deletion-as-early-as-permissible being a key factor,

Where DP is serving a business purpose, objects are rarely being perceived as frozen and deposited into an archive. Rather, they need to be maintained in an operational environment. Catch phrases such as business continuity capture a lot about the thinking behind this. More importantly, however, they also help to identify approaches and solutions to serve these (but also more traditional) DP needs that stem from different backgrounds: life cycle management of entire IT environments, redundancy, (IT) security, e-governance structures and methods have been developed, deployed, tested, customized and improved over long time spans in these communities. These solutions may prove valuable contributions to the DP community at large.

This integration of yet another heterogeneous set of communities poses severe challenges to a rather tightly-knit network of DP researchers and practitioners that has just started to evolve into a very young community of its own, with it's own jargon, events, and commonly understood basis of generic concepts. Yet, it also offers huge benefits. Apart from contributing new competences and tested solutions that can be adapted to serve more generic DP needs, it also broadens the basis amongst which to share the costs of research and development in DP. It allows us to integrate know-how from different groups and grows the market of where this know-how can be used.

In terms of new areas of activity, at least three major trends can be observed:

Preservation as a cost/benefit trade-off: The primary focus so far has been on preserving objects because they need to be preserved. Current thinking seems to be moving towards obsolescence as a risk, and DP as a risk mitigation strategy. While this concept is anything but new to the DP community, the key difference is on the juxtaposition of risk and benefit, and a much more pronounced and explicit willingness to sacrifice availability of objects when the investments necessary to maintain them would likely outweigh the business benefits or legal sanctions. Although the same principles are applicable in traditional settings as well, this focus shift forces a more explicit formulation of benefits/-value and a clearer specification of risk and costs, especially for more immediate/short term actions. This also will call for the application of existing frameworks for risk identification as well as cost/benefit estimation, with potentially strongly diverging valuations in the non-heritage domains.

DP as capabilities and maturity evaluations: Another important shift we may see coming is a shift from DP being something happening in a data archive or repository setting, i.e. a designated institution or department handling "old" objects, and being audited on its performance via any set of audit and certification routines to ensure proper preservation at the highest possible level. Rather, we may view DP as a set of capabilities that an institution has, as part of many other operational capabilities, and that are integrated with more routine processes. We are thus currently investigating opportunities of integrating DP in an eGovernance framework. Reference models such as TOGAF [17] and COBIT [8] provide a well-tested basis for designing and managing DP activities alongside other routine IT capabilities. It allows us to manage DP capabilities in the framework of IT governance, benefiting from the well-structured concepts

and processes in this domain, defining drivers, constrains and controls [2]. An example of how to model maturity levels for preservation capabilities is provided in [1].

Preserving processes rather than (only) data: The third, major area of activity currently, and probably the most relevant with respect to semantic technologies, relates to the shift from preserving static objects to the preservation of entire process chains. This shift is motivated by two core considerations: first of all, in many new settings, it is not static artifacts, but the need to be able to re-run processes in an authentic manner that are the key DP requirement. While the preservation of data as documentation may be sufficient to provide evidence about processes having been run they are no replacement for preserving the actual process.

But even when the focus is on the actual data, it may be advisable to preserve the process chain the data was subjected to. In an e-Science setting, while preserving the data is an essential first step for any sustainable research efforts, the data alone is often not sufficient for later analysis of how this data was obtained, pre-processed and transformed. Results of scientific experiments are often just the very last step of the whole process, and to be able to correctly interpret them by other parties or at a later point in time, also these processes need to be preserved.

Thus, specifically in an e-Science setting, preserving processes together with the data helps us to meet two goals at the same time: on the one hand, the processes are essential aspects of representation information, allowing to trace the various (pre-)processing steps applied to the data, any bias that might have been introduced, or errors stemming from faulty processing. On the other hand, it also allows us to re-run these processes on new data, learning how models and views evolved, and to discover discrepancies from earlier analyses.

We thus need to go beyond the classical concerns of Digital Preservation research, and consider more than the preservation of data. The following section takes a closer look at some of the activities centering around process preservation, covering both process capture as well as evaluation of re-executed processes – and the resulting requirements at the level of process capture. We will choose examples from the e-Science/Data curation domain as an exemplary setting. The following sections are largely adopted from a position paper summarizing our considerations to data quality aspects in data curation for a recent workshop by the National Science Foundation [16].

3 Capturing Processes

Curation of business or E-Science processes requires capturing the whole context of the process, including enabling technologies, different system components on both hardware and software levels, dependencies on other computing systems and services operated by external providers, the data consumed and generated, and more high-level information such as the goals of the process, different stakeholders and parties. The context of information needed for preserving processes

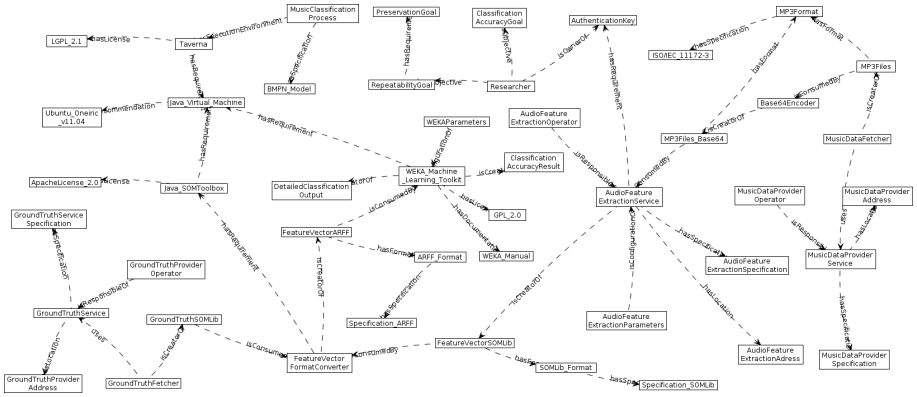

Fig. 1: Context Model of musical genre classification process [12], top-level view

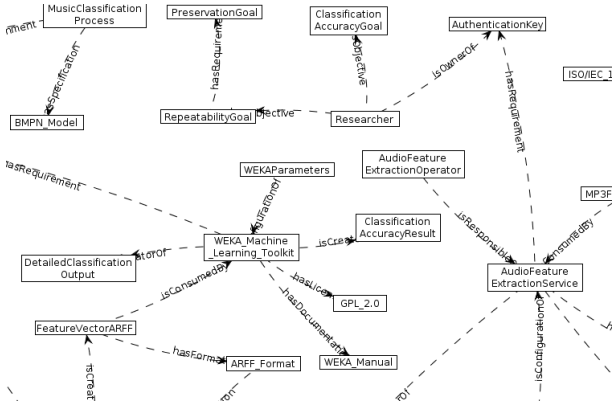

Fig. 2: Excerpt of Context Model of musical genre classification process

is considerably more complex than that of data objects, as it not only requires dealing with the structural properties of information, but also with the dynamic behavior of processes. Successful curation of an eScience process requires capturing sufficient detail of the process, as well as its context, to be able to re-run and verify the original behavior at a later stage, under changed and evolved conditions. We thus need to preserve the set of activities, processes and tools, which all together ensure continued access to the services and software which are necessary to reproduce the context within which information can be accessed, properly rendered and validated.

To address these challenges, we have devised a context model to systematically capture aspects of a process that are essential for its preservation and verification upon later re-execution [12]. The model consists of approximately 240 elements, structured in around 25 major groups. It corresponds to some degree to the representation information network [10], modeling the relationships

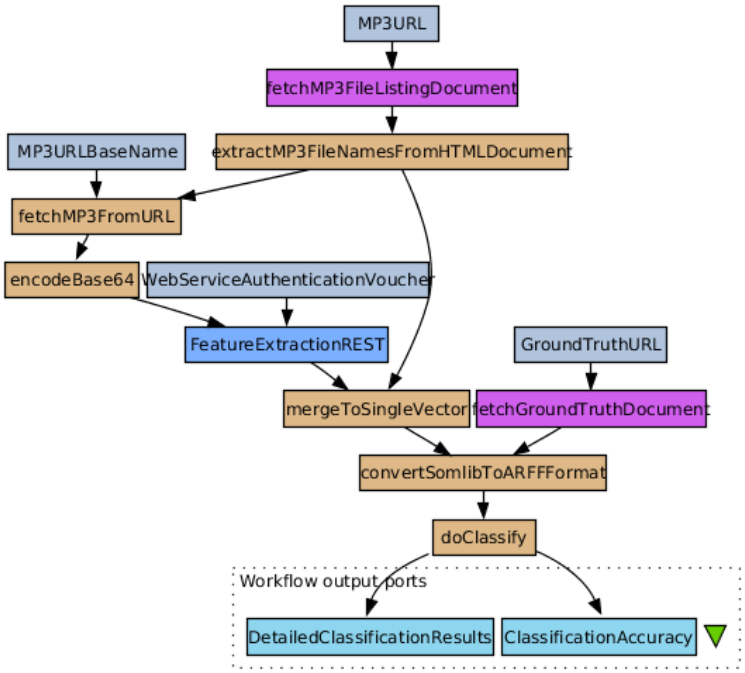

Fig. 3: Musical genre classification, including fetching of data, modelled in the Taverna workflow engine [11]

between an information object and its related objects, be it documentation of the object, constituent parts and other information required to interpret the object. This is extended to understand the entire context within which a process, potentially including human actors, is executed, forming a graph of all constituent elements and, recursively, their representation information. Specific emphasis is given to the identification of distributed components of a process: identifying, for example, external web services is essential to ensure that a process can be preserved, as specific measures need to be devised to ensure their availability in the future. Approaches include an integration of a web service into the process, legal agreements on the preservation of web services (at specified versions) by the service provider, deposit regulations using ESCROW regulations [7], and others. The model is implemented in the form of an ontology, which on the one hand allows for the hierarchical categorization of aspects, and on the other hand shall enable reasoning, e.g. over the possibility of certain preservation actions for a specific process instance. While the model is very extensive, it should be noted that a number of aspects can be filled automatically – especially if institutions have well-defined and documented processes. Also, not all sections of the model are equally important for each type of process. Therefore, not every aspect has to be described at the finest level of granularity.

Figure 2 provides an overview on the concrete instances and their relations identified as relevant aspects of the process context for a music classification process. The process basically represent a typical machine learning experiment in music information retrieval (MIR), where features are extracted from audio data and used as the basis for training a classifier system, sorting pieces of music into different musical genres. For a detailed description of this process and on how to make it fit for preservation, refer to [11]. An excerpt is provided in Fig. **??**, showing, for example, the links to the preservation goal specifications, some of the modules involved (the AudioFeatureExtractor and the WEKA Machine Learning toolkit) and their respective contexts (licenses (GPL_2.0), file format and their link to specifications, manuals, etc.

To move towards more sustainable E-Science processes, we recommend implementing them in workflow execution environments. For example, we currently use the Taverna workflow engine [13]. Taverna is a system designed specifically to execute scientific workflows. It allows scientists to combine services and infrastructure for modeling their workflows. Services can for example be remote web-services, invoked via WSDL or REST, or local services, in the form of predefined scripts, or user-defined scripts.

Implementing such a research workflow in a system like Taverna yields a complete and documented model of the experiment process – each process step is defined, as is the sequence (or parallelism) of the steps. Further, Taverna requires the researcher to explicitly specify the data that is input and output both of the whole process, as well as of each individual step. Thus, also parameter settings for specific software, such as the parameters for a machine learning tool or feature extraction, become explicit, either in the form of process input data, or in the script code.

Figure 3 shows an example of the music classification experiment workflow modeled in the Taverna workflow engine. We notice input parameters to the process such as the URL of the MP3 contents and the ground truth, and also an authentication voucher which is needed to authorize the use of the feature extraction service. The latter is a bit of information that is likely to be forgotten frequently in descriptions of this process, as it is rather a technical requirement than an integral part of the scientific process transformations. However, it is essential for allowing re-execution of the process, and may help to identify potential licensing issues when wanting to preserve the process over longer periods of time, requiring specific digital preservation measures.

During an execution of the workflow, Taverna records so-called *provenance data*, i.e. information about the creation of the objects, on the data transformation happening during the experiment. Taverna uses its proprietary *Janus* format, an extension on the Open-Provenance Model[14] that allows capturing more details. Such data is recorded for the input and output of each process step. It thus allows to trace the complete data flow from the beginning of the process until the end, enabling verification of the results obtained. This is essential for being able to verify system performance upon re-execution, specifically when

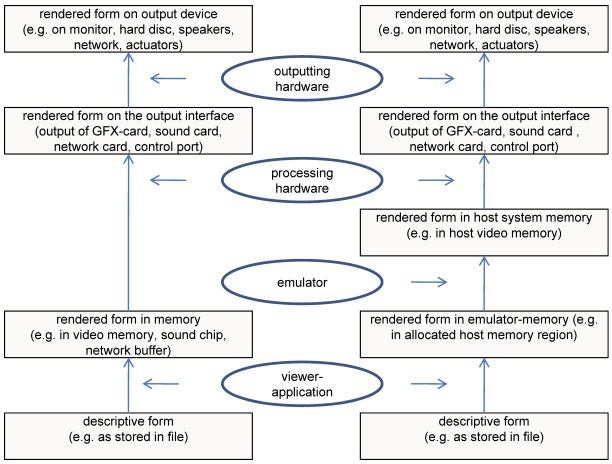

Fig. 4: Different forms of a digital object in a system's memory. On the left the layers in an original system are shown, on the right the layers in the system hosting the emulator are shown. [3]

any component of the process (such as underlying hardware, operating systems, software versions, etc.) has changed.

4 Evaluating Process Re-Execution

A critical aspect of re-using digital information in new settings is its trustworthyness, especially its authenticity and faithful rendering (with rendering being any form of representation or execution and effect of a digital object, be it rendering on a screen, an acoustic output device, or state changes on ports, discs etc.). Establishing identity or faithfulness is more challenging than commonly assumed: current evaluation approaches frequently operate on the structural level, i.e. by analyzing the preservation of significant properties on the file format level in case of migration of objects. Yet, any digital object (file, process) is only perceived and can only be evaluated properly in a well-specified rendering environment within which faithfulness of performance need to be established. In emulation settings, this evaluation approach is more prominently present, yet few emulators support the requirements specific to preservation settings. We thus argue that, actually, migration, emulation and virtually all other approaches to logical/structural data preservation need to be evaluated in the same way, as they are virtually no different from each other as all need to be evaluated in a given rendering/performance environment. [5].

We also devise a framework for evaluating whether two versions of a digital object are equivalent [3]. Important steps in this framework include (1) a description of the original environment, (2) the identification of external events influencing the object's behavior, (3) the decision on what level to compare the

two objects, (4) recreating the environment, (5) applying standardized input to both environments, and finally (6) extracting and (7) comparing the significant properties on suitable levels of an object's rendering. Even though the framework focuses mostly on emulation of environments, the principles are also applicable specifically for entire processes, and will work virtually unchanged also for migration approaches, when complex objects are transformed e.g into a new file format version.

An essential component of the framework is the identification at which levels to measure the faithfulness of property preservation, as depicted in Figure 4. A rendered representation of the digital object has to be extracted on (a) suitable level(s) where the significant properties of the object can be evaluated. For some aspects, the rendering of an object can be performed based on its representation in specific memories (system/graphics/sound card/IO-buffer), for others the respective state changes at the output port have to be considered while for yet others the actual effect of a system on its environment needs to be considered, corresponding to delineating the boundaries of the system to be evaluated. (Note that identity on a lower level does not necessarily correspond to identity at higher levels of the viewpath - in some cases significant effort is required to make up for differences e.g. on the screen level when having to emulate the visual behavior of cathode ray screens on modern LCD screens [15].) An example of applying this framework to the evaluation of preservation actions is provided in [4]

A key challenge in this context will be to come up with a comprehensive model of what information to capture for specific types of processes and preservation requirements, as well as guidelines on how to do this.

5 Conclusions

This paper presents a loose collection of some trends observed in digital preservation research, specifically a shift toward a more risk/cost/benefit based approach to DP, the framing of DP in IT governance principles, and specifically the necessity to preserve entire processes rather than only data. With the growing importance of preserving entire processes rather than sets of homogeneous, static (and usually quite simple) objects, the requirements on techniques to capture, document and reason across increasingly complex sets of context meta-information is growing. This requires new approaches to context capture and representation.

Still, the considerations above cover only a small subset of the quite significant research challenges that continue to emerge in the field of digital curation. We thus strongly encourage the community to contribute to an effort of collecting and discussing these emerging research questions in a loosely organized form. To this end, following the Dagstuhl Seminar on Research Challenges in Digital Preservation[1], a Digital Preservation Challenges Wiki[2] has been created, where we invite contributions and discussion. As a follow-up to the Dagstuhl seminar,

[1] http://www.dagstuhl.de/de/programm/kalender/semhp/?semnr=10291
[2] http://sokrates.ifs.tuwien.ac.at

a workshop on DP Challenges[3] will be held at iPRES 2012 in Toronto focusing on the elicitation and specification of research challenges.

References

1. C. Becker, G. Antunes, J. Barateiro, and R. Vieira. A capability model for digital preservation: Analysing concerns, drivers, constraints, capabilities and maturities. In *8th International Conference on Preservation of Digital Objects (IPRES 2011)*, Singapore, November 2011.
2. C. Becker, G. Antunes, J. Barateiro, and R. Vieira. Control objectives for dp: Digital preservation as an integrated part of it governance. In *Proceedings of the 74th Annual Meeting of the American Society for Information Science and Technology (ASIST)*, New Orleans, Lousiana, US, October 2011.
3. M. Guttenbrunner and A. Rauber. A Measurement Framework for Evaluating Emulators for Digital Preservation. *ACM Transactions on Information Systems (TOIS)*, 30(2), 2012.
4. M. Guttenbrunner and A. Rauber. Evaluating an emulation environment: Automation and significant key characteristics. In *Proceedings of the 9th conference on Preservation of Digital Objects (iPRES2012)*, Toronto, Canada, October 1–5 2012.
5. M. Guttenbrunner and A. Rauber. Evaluating emulation and migration: Birds of a feather? In *Proceedings of the 14th International Conference on Asia-Pacific Digital Libraries*, Taipei, Taiwan, November 12–15 2012.
6. T. Hey, S. Tansley, and K. Tolle, editors. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, 2009.
7. K. Hobel and S. Strodl. Software escrow agreements. In *Tagungsband des 15. Internationalen Rechtsinformatik Symposions IRIS 2012*, pages 603–610, 2012.
8. IT Governance Institute. *COBIT 4.1. Framework – Control Objectives – Management Guidelinces – Maturity Models*. 2007.
9. S. Manegold, M. Kersten, and C. Thanos. Special theme: Big data. *ERCIM News*, (89), January 2012.
10. Y. Marketakis and Y. Tzitzikas. Dependency management for digital preservation using semantic web technologies. *International Journal on Digital Libraries*, 10:159–177, 2009.
11. R. Mayer and A. Rauber. Towards time-resilient mir processes. In *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR 2012)*, Porto, Portugal, October 8-12 2012.
12. R. Mayer, A. Rauber, M. A. Neumann, J. Thomson, and G. Antunes. Preserving scientific processes from design to publication. In *Proceedings of the 15th International Conference on Theory and Practice of Digital Libraries (TPDL 2012)*, LNCS, Cyprus, September 2012. Springer.
13. P. Missier, S. Soiland-Reyes, S. Owen, W. Tan, A. Nenadic, I. Dunlop, A. Williams, T. Oinn, and C. Goble. Taverna, reloaded. In *Proceedings of the 22nd International Conference on Scientific and Statistical Database Management*, SSDBM'10, pages 471–481. Springer, June 2010.
14. L. Moreau, J. Freire, J. Futrelle, R. E. Mcgrath, J. Myers, and P. Paulson. *Provenance and Annotation of Data and Processes*, chapter The Open Provenance Model: An Overview, pages 323–326. Springer, 2008.

[3] `http://digitalpreservationchallenges.wordpress.com/`

15. G. Phillips. Simplicity betrayed. *Communications of the ACM*, 53(6):52–58, 2010.
16. A. Rauber. Data quality for new science: Process curation, curation evaluation and curation capabilities. In *Workshop notes for the UNC/NSF Workshop Curating for Quality*, Arlington, VA, September 10–11 2012.
17. The Open Group. *TOGAF Version 9*. Van Haren Publishing, 2009.