Sustainable eScience processes and systems

Tomasz Miksa
SBA Research
Vienna, Austria
tmiksa@sba-research.org

ABSTRACT

The advent of new means of performing research and sharing results in eScience settings poses new threats and sets new requirements to research procedures. One of them is providing long-term access and means for ensuring correct interpretation of research artefacts in the future. For this purpose several problems and challenges have to be solved to embody the concept of sustainable eScience. In this proposal we describe challenges concerning process' management, verification and maturity, as well as propose ways of tackling them.

Keywords

digital preservation, eScience, process verification, preservability, process management plans

Categories and Subject Descriptors

H.3.7 [**Information Systems**]: Digital Libraries—*Collection, Dissemination, Standards*; D.2.2 [**Software and its engineering**]: Software organization and properties—*Software design engineering, Software/Program Verification*; K.6.4 [**Management of computing and Information Systems**]: System Management

1. INTRODUCTION

The increasing computational power of computers and throughput of information systems have enabled researchers to make their scientific breakthroughs by processing, linking and exchanging multiple huge datasets. These datasets are very often referred to as "big data" [26] and the methodology is called to be the fourth paradigm of science, namely eScience [12]. Results obtained by research conducted in this way are said to be "born digital" [18]. eScience cooperation is very often realized within the Research Infrastructures (RI), where different stakeholders cooperate and share information to look for solutions of important problems of society. RI relay on information systems and allow exchange of facilities, resources and services.

Digital Preservation (DP) can be defined as a set of actions and efforts whose goal is to maintain digital objects accessible in an authentic manner for a long term into the future. Its focus is not only on ensuring physical preservation of content, but also on ensuring logical and semantic preservation. A wide range of strategies is possible and there is no optimal solution for every case. DP has emerged mainly from memory institutions and cultural heritage sector [17], where static objects (e.g. images, text documents) were in scope of interest. Nowadays, one can observe transition to business sector, where DP has to deal with preserving interactive software, processes or even the whole systems. DP is considered as one of the value adding capabilities, which can be utilized by business institutions, even though their main focus is not DP itself.

This proposal aims at bringing all these disciplines together. eScience conducted within RI needs to be preserved in order to allow reuse and validation of both delivered results, but also processes which created them. For this purpose several problems and challenges have to be solved to embody the concept of sustainable eScience.

2. PROBLEM DESCRIPTION

The research is going to concentrate on enhancing the sustainability of both eScience data and eScience processes by exploring the capabilities which enable creation of preservable systems. This problem consists of several sub problems which have to be tackled separately.

Firstly, Data Management Plans (DMPs) that accompany experiments and scientific investigations that produce "big data" must be a part and parcel of each scientific project. They are recently being enforced on researchers who get the public funding but in most cases they are a set of advices, mainly lists of questions, which researches should consider when developing a DMP. The attention is attracted to what happens with data after it has been created, rather than in what way it was obtained. All the description is provided in a text form and in case of National Science Foundation (NSF) there is a limit of 2 pages. Therefore there is no possibility to reuse or at least reproduce the process which created the data. Furthermore, the correctness of data is taken for granted and thus DMPs do not provide sufficient information that would allow validation of the data. Finally, the quality and the detail of information strongly depends on scientific honesty and a good will of researchers. There is no formal template for specification of DMP which would

ensure that all important information is covered. Therefore there is a need to revise and remodel the current DMPs in order to support information on processes and formalize their structure. Such redesigned DMPs have to be actionable and enforceable. Set of well-defined guidelines which specify what kind of data has to be captured is needed. Information about provenance and possibility to recreate and validate the process have to be granted.

Secondly, in order to ensure authenticity, trustworthiness and credibility of preservation actions, there is a need to prove, that the data captured is not only identical at the physical level, but also sufficient to re-enact the process in an unchanged manner in the future. Therefore, ways to compare the original and the captured process are essential. This has implications on data capturing process, because the data, which facilitates comparison of processes, has to be collected. In case of processes dependent on third party resources, when the access to the preserved content is limited, it is often impossible to access all information needed for preservation of the process. Methods for finding or creating substitutes are one of the key research challenges. Ways to validate if these substitutes conform to the original behaviour are necessary. In addition to this, not only ways to compare the processes but also levels of comparison have to be defined. Compared processes can deliver the same outputs, but the way they were executed may vary significantly. This also has to be detected.

Thirdly, digital preservation has been mainly associated with cultural institutions that aim to keep nations heritage available in the future. However, it has been recently recognized that DP is also a need in companies and enterprises in which main focus is not DP itself. In such cases DP is considered as one of the value adding capabilities, which can be utilized by business. Successful transition of DP towards business can only be achieved when IT Governance methods can be applied to DP in order to evaluate the degree of compliance of actual IT actions to the defined goals and in order to measure their efficiency. Moreover, recently many enterprises have been commonly using Capability Modelling to plan, create and deliver strategic business capabilities to the enterprise. Capability Maturity Models (CMMs) were defined for comparison, verification and assessment of processes which reflect measured capabilities. Therefore there is a need to couple DP standards, frameworks and actions with business approach. Otherwise the concept of digital preservation cannot be fully integrated into business.

Presented problems and challenges require many improvements in several fields of not only eScience, Digital Preservation and Research Infrastructures but also IT Governance, Software Engineering and System Modelling. Ranging from automation of process characterization, extraction, validation and redeployment to redefinition of Data Management Plans, software engineering practices and quality metrics; systems design requirements.

The following points list precise research questions which have to be considered in the first steps in order to address the main research problem.

1. Can we find a way to automatically detect the execution context of an eScience process in order to facilitate the capture of process dependencies and characteristics?
   - What is the minimum set of information needed?
   - What are the ways of storing these information?
   - What are the ways to trace process executed at multiple stages on various computers?
   - What techniques can be applied to trace process executed in a Cloud or Grid environment?
   - What dependencies to external resources must be described?
   - What is the best format to describe detected dependencies?
   - What dependencies are specific for eScience processes?
   - In what way do these specific dependencies influence possible digital preservation actions?
2. Can we find a way to automatically collect sufficient set of information on eScience processes which will allow validation and faithful future redeployment?
   - What are the possible ways to replay a process?
   - In what way can we validate if the mocked, reimplemented or substituted process step complies to the original one?
   - In what way can the original and the preserved process be compared?
   - What are the levels of processes comparison?
   - In what way can we validate processes which require human interaction?
   - Can we identify significant properties special for eScience processes?
3. Can we model the eScience process sustainability as a capability accordingly to Capability Maturity Model?
   - What are the ways to measure sustainability and in what way create key performance indicators (KPIs)?
   - What is the optimal way to conduct assessment process?
   - What criteria should be used for classification?

3. EXPECTED RESULT

In the field of eScience which is conducted within Research Infrastructures there is a big emphasize on reuse of data created by other scientists or experiments. Current IT solutions' design is driven by several different factors, mainly business needs, but surely not mere digital preservation requirements. The thesis is going to examine ways of enabling eScience processes and data to be reliably and authentically preserved and reused. We would like to move towards "process as a fuel" approach, when not only the outcome of the experiment could be reused, but also the methodology to obtain the data (fully or partially). Therefore, this thesis aims at creating a concept of a preservable system, which design is going to follow the needs of digital preservation

as the primary requirement. In order to move towards this vision results in three areas are expected.

To begin with, the concept of DMP is going to be extended by processes. The new concept will allow recreating the process of achieving results. Moreover, some tools which facilitate the capture of processes and allow monitoring for changes are going to be developed. Verification of new DMP guidelines and applicability of proposed tools is going to be evaluated by application to a real eScience scenarios provided by LNEC[1].

Further, a model specifying possible levels of process comparison is going to be specified. It will organize the comparison process and allow to choose the best comparison method depending on requirements. What is more, prototype tools which allow the validation of processes and allow detection of changes in steps of the process are going to be created. Evaluation of the tools and the model is going to be tested in real eScience scenarios.

Last but not least, our work is going to verify applicability of IT Governance methods to digital preservation of eScience data and processes. We are going to analyze and evaluate the usefulness and applicability of Capability Maturity Modelling to actions performed in a designed conceptual system. This will provide evidence if correlation between IT Governance strategies and Digital Preservation guidelines applied for eScience can be achieved.

Having investigated these three areas we should be able to create a set of guidelines for preservable systems. Their usefulness and correctness is going to be supported by proofs of concept and evaluation of case studies specified above.

4. METHODOLOGICAL APPROACH

Due to the complex and interdisciplinary nature of the problem, there is a wide range of actions and methods that have to be applied in order to tackle the problem successfully. At the current stage of research, it is hard to foresee which of them will prove themselves to be more useful than the others. Moreover, one should be aware that only diverse combinations of them will enable success. Therefore the purpose of this section is to provide a list of methods, which will be considered not only as finite and unique way to address the problem but also as a set of actions allowing to build up ways of dealing with a considered problem.

The context and background of eScience processes have to be deeply understood. A set of use cases allowing to test, verify and validate the hypotheses as well as prototype tools needs to be created. This can be achieved by establishing contact and close cooperation with scientific communities, who are understanding the need and urgency to introduce digital preservation actions into the world of eScience. To the best of our knowledge in fields of astronomy, physics, geo-sciences and social sciences [20] these concerns are gaining understanding. Hence discussions, talks and meetings with representatives of these communities should enable us to progress with research more efficiently. The feedback received during these events cannot be overestimated because it will enable us to quickly validate correctness of our theories and usability of proposed tools. A cooperation with LNEC and scientists from TU Wien are the most likely cooperation, which will provide us with real eScience scenarios.

Further, we will investigate existing recommendations concerning data and process management, digital preservation, IT governance and other kind of recommendations which apply to the investigated domain. In many of these areas there are some rules and recommendations already in use but they are very often inconsistent and are partial solutions. We are going to analyze them and incorporate best of them in order to create an updated version of DMPs extended by processes, as well as, define CMM for sustainable eScience. By cooperation with institutions, which are involved in these areas, we will participate in shaping real solutions, which have to obey many legal and financial restrictions. Thus we will have many occasions to confirm our findings and recommendations against the board of other experts working in problem's area as well as possibilities to apply them to real actions implemented at a broad scale.

We also cooperate closely with Secure Business Austria (SBA) which is participating in TIMBUS Project[2]. This is an European Union project funded within Seventh Framework Program and its focus is to "deliver activities, processes and tools that ensure continued access to services and software to produce the context within which information can be accessed, properly rendered, validated and transformed into knowledge". This cooperation will allow investigation of more practical aspects of the problem like automation, dependency analysis, assessment and validation. By participation in design, implementation and testing of prototype tools or proofs of concept, we will broaden the horizons of problem understanding and also contribute directly to enhancing sustainability of eScience by developing new tools and techniques. Cooperation is going to focus on building monitoring tools that monitor running processes in order to extract and capture data and feed them into the context model which will be also specified during this cooperation. Finally, application of these tools and models should result in identification of process comparison levels and therefore lead to improved organization of process comaprison process.

Summing up, tackling such a broad problem of enhancing sustainability of eScience can only be dealt with by a broad spectrum of actions. Hence the methodological approach ranges from top-view actions like investigation and reshaping of recommendations to bottom-view actions like direct interviews with researchers and implementation of tools.

5. STATE OF THE ART

This section presents related works in the domains of data management, digital preservation, eScience and research infrastructures.

5.1 Digital Preservation

Digital Preservation has emerged mainly from memory institutions and the cultural heritage sector [16]. The part and parcel of Digital Preservation is Preservation Planning. It

[1]Laboratório Nacional de Engenharia Civil

[2]`http://timbusproject.net/`

specifies actions plans, fosters decision making process and provides evidence for any DP operation. This broad topic has been investigated in [4], [5]. A framework which was created for multi level comaprison of emulation effects was presented in [11]. In case of migration its effects can be validated in an automatic manner with a use of eXtensible Characterisation Languages (XCL) [6]. There is a number of research projects addressing the challenges of keeping processes available in the long term. WF4Ever[3] addresses the challenges of preserving scientific experiments by using abstract workflows that are reusable in different execution environments [23]. The TIMBUS[4] project researches preservation of business processes. The approach is to create context model [14] of the process, which is an ontology based model for description the process component and their dependencies. It allows to store information on not only software and hardware, but also organisational and legal aspects of the process.

5.2 eScience and Research Infrastructures

Due to the increasing computational power of computers and throughput of information systems researchers are making use of them and make their scientific breakthroughs by working with immensly huge datasets. These data sets are very often referred to as "big data" [26]. Results obtained by research conducted in this way are said to be "born digital" [18]. Pioneering computer scientist Jim Gray does not hesitate to name such research the Fourth Paradigm [12]. He places it in a row with theoretical, experimentational and simulational paradigms.

Quick progress in any scientific area which is taking advantage of eScience can only be reached by a close cooperation of not only scientists, but also IT specialists. Several tools and improvements have to be created in order to introduce enhancements in different areas ranging from databases and workflow management to graphics and representation of results. Calls for action and requirements of successful eScience are presented in [22], [17], [12]. Many projects has been funded by European Union in order to fulfill this vision. APARSEN[5] is an example of Network of Excellence established in order to foster "information creation, curation and re-use". It consists of diverse projects. For example ODE[6] aims to "engage and raise the profile of data sharing, re-use and preservation". While SCIDIP-ES [7] main focus is on delivering services and infrastructure that allows to preserve eScience data for a long term having Earth Science data as a case study [1].

Research Infrastructures (RI) are being established in order to facilitate the cooperation and information sharing between different stakeholders, who look for solutions of important problems of society. These problems span across several fields, for example: Energy, Food, Transport, Climate, Societies, Health [8]. RI relay on information systems and allow to exchange facilities, resources and services. Accordingly to EU strategies the RI must be opened to all interested researchers [10], [21].

[3] http://www.wf4ever-project.org/
[4] http://timbusproject.net/
[5] http://www.alliancepermanentaccess.org/
[6] http://www.alliancepermanentaccess.org/index.php/community/current-projects/ode
[7] http://www.scidip-es.eu/science-data-infrastructure-for-preservation-with-focus-on-earth-science/

5.3 Data Management and Open Access

A Data Management Plan is a formal document that describes in what way is the data going to be handled during the research as well as after the project is completed [27]. The main aim of DMP is to plan data management actions and define practices which are going to be executed in the project. Several aspects like for example collection of metadata, storage of the data in easily preservable formats, etc. have to be considered in order to facilitate future reuse of preserved contents. Agencies which fund scientific projects are recently enforcing DMP to be a part of a research proposal. However, there is no common definition of what has to be included in such a plan. National Science Foundation (NSF), Digital Curation Centre (DCC), Australian National Data Service (ANDS) are the world leading institutions which have constructed their own DMPs. In most cases plans consist of a set of advices [19] and checklists [9] that help the researchers to create their own plans, rather than formal and structured documents. DMPs does not allow to validate the obtained results. Furthermore, the accuracy of information provided about data depends mainly on honesty, awareness and experience of a person creating the plan. Finally, all of these plans does not preserve processes which created the data. These issues may be addressed by a novel concept of Process Management Plans [15]. It complements the description of scientific data taking a process centric view, viewing data simply as the result of underlying processes such as capture, (pre-) processing, transformation, integration and analyses.

Several projects benefit nowadays from sharing and reusing the data. Success stories have been presented in [20]. Another example of successful sharing of data is Economic and Social Data Service (ESDS) provided by Economic and Social Research Council in Great Britain. Recent study [7] has proved that the value of the shared data to the researchers is "£25 million per annum at 2010 prices and levels of activity use". This confirms that properly managed and shared data can result in major benefits.

The reuse of data and process is fostered by Open Access initiatives which support access not only to scientific publications but also to data [30]. There is also a number of projects like RECODE[8] which aim to unify approaches of different institutions by producing policy recommendations for open access to research data based on existing good practice.

5.4 IT Governance and Enterprise Architecture

Enterprise Architecture (EA) is a concept that allows to take control of fragmented applications, organizational structures and processes by creating an environment with optimized processes which could respond to changes coming from both external factors as well as from changes in business strategy [28]. The Zachman framework is one of the first EA approaches which is often referred to by other EA frameworks.

[8] http://recodeproject.eu/

Nowadays The Open Group Architecture Framework (TOGAF) and the Department of Defense Architecture Framework (DODAF) [29] are the leading frameworks.

"IT Governance focuses on the leadership, organizational structures and processes that ensure that the enterprise's IT sustains and extends the organization's strategies and objectives." [2] Frequently used framework in this area is COBIT [13]. It assists managers to identify and decrease the gap between control requirements, technical issues and business risks. COBIT uses the concept of maturity which is also being used in Capability Maturity Model Integration (CMMI) [24] for software engineering. Maturity modeling allows to identify and classify gaps in capabilities in such a way that management can develop action plans, which will bring the capabilities to the desired level [13].

Some attempts have already been taken to integrate frameworks from IT Governance and Enterprise Architecture with Digital Preservation. [2] describes capability model for DP, [3] is an attempt to extend COBIT to cover DP as part of IT Governance and [25] defines the preservability as a set of system capabilities originating from a combination of system/software qualities.

6. WORK PLAN

Problem defined in Section 2 was divided in three areas. The work plan assumes tackling all these three areas in parallel. For this reason each of these three areas have their own work plan presented below.

Firstly, we are going to start with a thorough study of DMPs, process characterization languages and process runtime environments in order to define concept of DMP extended by processes. Then we are going to create tools which are going to be able to extract the process and collect automatically their full context including external dependencies. At the last step the evaluation of the new concept and created tools is going to be conducted with an application of real eScience use cases.

Secondly, we are going to investigate existing frameworks and approaches allowing validation of captured processes. On this basis, a model for process comparison is going to be specified. Further, the prototype tools which are going to validate captured process are going to be created. Application of tools and model to eScience scenarios will validate their correctness.

Thirdly, IT Governance and Business Enterprise Modeling concepts are going to be investigated in order to verify their suitability for sustainable eScience. Results of this analysis are going to be compared against DP guidelines. The cohesive vision of CMM is going to be created by taking into consideration conducted investigations.

7. REFERENCES

[1] Arif Sharon et al. Towards a long-term preservation infrastructure for earth science data. Proceedings of the 9th International Conference on Preservation of Digital Objects (iPres 2012). (to appear).

[2] Becker, C., Antunes, G., Barateiro, J., and Vieira, R. A capability model for digital preservation: Analysing concerns, drivers, constraints, capabilities and maturities. In *Proceedings of the 8th International Conference on Preservation of Digital Objects (iPres 2011)* (2011). Vortrag: iPres 2011 - 8th International Conference on Preservation of Digital Objects, Singapore; 2011-11-01 – 2011-11-04.

[3] Becker, C., Antunes, G., Barateiro, J., Vieira, R., and Borbinha, J. Control objectives for dp: Digital preservation as an integrated part of it governance. In *ASIST 2011* (2011), American Society for Information Science and Technology (ASIST). Vortrag: 74th Annual Meeting of the American Society for Information Science and Technology (ASIST), New Orleans, Louisiana, USA; 2011-10-09.

[4] Becker, C., Kulovits, H., Guttenbrunner, M., Strodl, S., Rauber, A., and Hofman, H. Systematic planning for digital preservation: Evaluating potential strategies and building preservation plans. *International Journal on Digital Libraries (IJDL)* (December 2009). http://dx.doi.org/10.1007/s00799-009-0057-1.

[5] Becker, C., and Rauber, A. Preservation decisions: Terms and Conditions Apply. Challenges, Misperceptions and Lessons Learned in Preservation Planning. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL 2011)* (June 2011).

[6] Becker, C., Rauber, A., Heydegger, V., Schnasse, J., and Thaller, M. A generic xml language for characterising objects to support digital preservation. In *Proceedings of the 2008 ACM symposium on Applied computing* (New York, NY, USA, 2008), SAC '08, ACM, pp. 402–406.

[7] Charles Beagrie Ltd and The Centre for Strategic Economic Studies (CSES)University of Victoria. Economic impact evaluation of the economic and social data service, March 2011.

[8] Copenhagen Research Forum. Visions for horizon 2020, 2012.

[9] Digital Curation Centre. *Checklist for a Data Management Plan*, 3 ed., March 2011.

[10] ESFRI. *Strategy Report on Research Infrastructures*, March 2011.

[11] Guttenbrunner, M., and Rauber, A. A measurement framework for evaluating emulators for digital preservation. *ACM Trans. Inf. Syst. 30*, 2 (May 2012), 14:1–14:28.

[12] Hey, T., Tansley, S., and Tolle, K., Eds. *The Fourth Paradigm: Data-Intensive Scientific Discovery.* Microsoft Research, 2009.

[13] IT Governance Institute. *Cobit 4.1.* ISA, 2007.

[14] Mayer, R., Rauber, A., Neumann, M. A., Thomson, J., and Antunes, G. Preserving scientific processes from design to publication. In *Proceedings of the 16th International Conference on Theory and Practice of Digital Libraries (TPDL 2012)* (Cyprus, September 23–29 2012), P. Zaphiris, G. Buchanan, E. Rasmussen, and F. Loizides, Eds., vol. 7489 of *Lecture Notes in Computer Science*, Springer, pp. 113–124.

[15] Miksa, T., and Rauber, A. Increasing preservability of research by process management plans. In *Proceedings of the 1st International Workshop on Digital Preservation of Research Methods and Artefacts* (New York, NY, USA, 2013), DPRMA '13, ACM, pp. 20–20.

[16] National Library of Australia, and Unesco. *Guidelines for the preservation of digital heritage / Prepared by the National Library of Australia.* National Library of Australia ; Information Society Division, United Nations Educational, Scientific and Cultural Organization, Canberra : Paris, France :, 2003.

[17] National Science Foundation. *It's About Time: Research Challenges in Digital Archiving and Long-term Preservation. Final Report of the workshop on Research Challenges in Digital Archiving and Long-Term Preservation*, 2002.

[18] National Science Foundation. Harnessing the Power of Digital Data for Science and Society. Report of the Interagency Working Group on Digital Data to the Committee on Science of the National Science and Technology Council. Tech. rep., 2009.

[19] National Science Foundation. *"Writing an NSF Data Management Plan"*, 2012.

[20] ODE Opportunities for Data Exchange. 10 tales of Drivers and Barriers to Data Sharing, 2012.

[21] OECD Global Science Forum. Large research infrastructures, 2008.

[22] Organization for Economic Co-operation and Development (OECD). Principles and Guidelines for Access to Research Data from Public Funding. Paris, France, 2007.

[23] Page, K., Palma, R., Holubowicz, P., Klyne, G., Soiland-Reyes, S., Cruickshank, D., Cabero, R. G., Garc'ia, E., Cuesta, D. D. R., and Zhao, J. From workflows to research objects: an architecture for preserving the semantics of science. In *2nd International Workshop on Linked Science* (2012).

[24] Paulk, M. C., Weber, C. V., Curtis, B., and Chrissis, M. B., Eds. *The capability maturity model: guidelines for improving the software process.* Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1995.

[25] Proenca, D., Antunes, G., and Miksa, T. On the assessment of preservability: Method and application. In *Proceedings of the 10th International Conference on Digital Preservation* (2013), Biblioteca Nacional de Portugal, pp. 187–196.

[26] Thanos, C., Manegold, S., and Kersten, M. L. Big data - introduction to the special theme. *ERCIM News 2012*, 89 (2012).

[27] The Australian National University. Anu data management manual, 2012.

[28] The Open Group. TOGAF 9 - The Open Group Architecture Framework Version 9, 2009.

[29] US Department of Defense. DoD Architecture Framework Version 2.0: Volume 1 (Manager's Guide - Introduction, Overview, and Concepts). Tech. rep., Pentagon, Washington DC, May 2009.

[30] Zuiderwijk, A., and Janssen, M. A comparison of open data policies and their implementation in two dutch ministries. In *Proceedings of the 13th Annual International Conference on Digital Government Research* (New York, NY, USA, 2012), dg.o '12, ACM, pp. 84–89.