IJDC | *Peer-Reviewed Paper*

Process Management Plans

Tomasz Miksa
SBA Research

Stephan Strodl
SBA Research

Andreas Rauber
SBA Research
and Vienna University of Technology

Abstract

In the era of research infrastructures and big data, sophisticated data management practices are becoming essential building blocks of successful science. Most practices follow a data-centric approach, which does not take into account the processes that created, analysed and presented the data. This fact limits the possibilities for reliable verification of results. Furthermore, it does not guarantee the reuse of research, which is one of the key aspects of credible data-driven science. For that reason, we propose the introduction of the new concept of Process Management Plans, which focus on the identification, description, sharing and preservation of the entire scientific processes. They enable verification and later reuse of result data and processes of scientific experiments. In this paper we describe the structure and explain the novelty of Process Management Plans by showing in what way they complement existing Data Management Plans. We also highlight key differences, major advantages, as well as references to tools and solutions that can facilitate the introduction of Process Management Plans.

Received 17 October 2013 | *Accepted* 26 February 2014

Correspondence should be addressed to Tomasz Miksa, Favoritenstrasse 16, 1040 Wien, Austria. Email: tmiksa@sba-research.at

An earlier version of this paper was presented at the 9th International Digital Curation Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

http://dx.doi.org/10.2218/ijdc.v9i1.303
DOI: 10.2218/ijdc.v9i1.303

Introduction

According to recent estimates there are currently about three hundred Research Infrastructures in Europe (Koski, 2012), which attract a wide range of diverse stakeholders who cooperate and exchange information to look for solutions of important problems of society (European Strategy Forum on Research Infrastructures, 2011). This cooperation is enabled by the increasing computational power of computers and throughput of information systems, which allow exchanging facilities, resources, services and datasets that are "born digital" (National Science and Technology Council, 2009) and, due to their complexity and size, are referred to as Big Data (Thanos et al., 2012). Such cooperation is an excellent example of the implementation of the fourth paradigm of science, also referred to as data-driven science or eScience (Hey et al., 2009). This new paradigm assumes that scientific breakthroughs would not be possible without special tooling, software and processes that allow researchers to transform, visualise and interpret the data (Van der Graaf and Waaijers, 2011). If any of the results obtained are to be validated the process of obtaining the results has to be documented and preserved for further reference.

Currently, there is no common way of documenting and managing research data. There is a serious threat that results cannot be reproduced or re-evaluated (McCullough, 2007). One way of addressing this burning issue is through Data Management Plans (DMPs), which are becoming a standard requirement for scientists applying for research funded by the European Union (2012), the National Science Foundation (2011), or the Australian National Data Service (2011). Their aim is not only to ensure that the scientific data is being properly managed during the lifetime of a project, but also that it is available and preserved in the future (Rice et al., 2013). The recent change of focus in the area of Digital Preservation from the preservation of static information like documents, scans, etc., to the preservation of complete systems and processes should facilitate the documentation and preservation of eScience discoveries made within the research infrastructures. This is reflected by the growing number of projects contributing tools (Miksa et al., in press), processes (Strodl et al., 2012) or systems (Killeen et al., 2012) which ease tackling this challenge. However, a data-centric approach represented by DMPs does not take into account the processes that created and transformed the data. This hinders the possibility of reliable verification and validation of results (National Science Foundation, 2011), as well as the reproducibility of research, which is one of the key assumptions of credible data-driven science (Lesk, 2012; Begley and Ellis, 2012).

For that reason, we proposed to introduce Process Management Plans (PMPs) (Miksa and Rauber, 2013). This new concept complements the description of scientific data by taking a process centric view, viewing data simply as the result of underlying processes such as capture, (pre-) processing, transformation, integration and analyses. The general objective of PMPs is to foster identification, description, sharing and preservation of scientific processes. Similar to the concept of executable papers (Nowakowski et al., 2011) or overlay journals (Hendler, 2007), PMPs see data and their processing/analysis as an integral part of research, but go beyond these concepts in putting focus specifically on the research process, where the data, analysis, interpretation and resulting publication all form equally important elements. PMPs have several advantages. First and foremost, they complement the data-centric approach of

Data Management Plans by providing a process-centric view. This facilitates the exchange and re-use of whole processes or parts of them in other experiments, decreasing development time and lowering costs, especially as PMPs are designed to be machine-actionable.

In this paper we present the concept of Process Management Plans. We specify their structure and contents by analysing various templates for Data Management Plans and identifying their deficiencies with special focus on support for processes. We also analyse usage scenarios and consider stakeholders for whom the information provided by the Process Management Plans would be useful. Finally, we provide recommendations on how the concept of Process Management Plans can be realised with use of existing tools and concepts from the area of Digital Preservation, which facilitate the characterisation of processes and technical dependencies, as well as contribute to increasing the sustainability of processes by preserving their execution environment.

Related Work

This section presents related works in the domains of Data Management, Digital Preservation, eScience and Research Infrastructures.

Data Management Plans

In order to identify current support of Data Management Plans with respect to processes, we have defined a set of requirements that should be met by process management activities, these are:

- The way of producing the result should be specified unambiguously;
- The tools, infrastructure and their specific settings used to obtain the results should be fully specified;
- The intermediate data products that lead to final results should be traceable and can be verified.

Then we have evaluated templates and recommendations for DMPs from the Digital Curation Centre (DCC) (Donnelly and Jones, 2011), the Australian National Data Services (ANDS) (2011) and the National Science Foundation (NSF) (2011). The results revealed that the DMPs specified by these institutions have very much in common. They include a set of advice, mainly in the form of checklists, which researches should consider when developing a DMP. The attention is paid to what happens with data after it has been created, rather than the way it was obtained. All of the descriptions are provided in a text form. Therefore, the possibility to reuse or at least reproduce the process that created the data is very unlikely. There are some tools available, like DMPOnline[1] from the DCC or DMPtool[2] from the NSF, which aid the researcher in the process of DMP creation, but these are rather simple interactive questionnaires, which generate a textual document at the end, rather than the complex tools that can validate at least the appropriateness of the provided information.

[1] DMPOnline: https://dmponline.dcc.ac.uk

[2] DMPtool: https://dmp.cdlib.org/

The main conclusion from this analysis is that DMPs focus on describing results of experiments. This is a consequence of their data-centric view, which enforces focus on access and correct interpretation (metadata) of data, and does not pay much attention to the processing of the data. Process Management Plans represent a process-centric view and provide information on the processes that created the data. PMPs also include Data Management Plans to provide information about data products that are used in the process, such as input and output data.

Digital Preservation

The field of Digital Preservation has recently focused on the long term preservation of entire processes and workflows (Strodl et al., 2011). There are a number of research projects addressing the challenges of keeping processes available in the long term. Tools, methods and other research outputs, which may be used to implement Process Management Plans, are developed within those projects.

The Wf4Ever[3] project addresses the challenges of preserving scientific experiments by using abstract workflows that are reusable in different execution environments (Page et al., 2012). The abstract workflow specifies a conceptual and technology-independent representation of the scientific process. They further developed new approaches to sharing workflows by using an RDF repository that makes the workflows and data sets accessible from a SPARQL Endpoint (Garijo and Gil, 2011).

The TIMBUS[4] project researches the preservation of business processes by ensuring continued access to services and software necessary to properly render, validate and transform information. Their approach is to create a context model (Mayer et al., 2012) of the process, which is an ontology based model for describing process components and their dependencies. It allows for the storage of information regarding not only software and hardware, but also organisational and legal aspects of the process. This model can be used to develop preservation strategies and to redeploy the process in a new environment in the future. The project developed a verification and validation method for redeployed processes (Miksa et al., 2013) that evaluates the conformance and performance quality of processes redeployed in new environments. This is especially important when the PMP is used for the purpose of validation (by re-executing the process), or reuse (to build other process).

eScience and Research Infrastructures

Nowadays, several projects benefit from sharing and reusing of data. Success stories have been presented by Darby et al. (2012). De Roure (2011) also discusses the evolution of research practices by sharing of tools, techniques and resources. MyExperiment,[5] is a platform for sharing of scientific workflows (Roure et al., 2010), which represents one step forward beyond just sharing the data, because it enables workflows created and run within the workflow engine Taverna[6] to be published and reused by other researchers. Although the workflows specify information regarding the tools required to run the steps, and provide description of the parameters necessary to re-run the workflow, they do not specify the infrastructure needed to be in place to execute the workflow. However, the data sets that allow users to re-run a particular

[3] Wf4Ever: http://www.wf4ever-project.org/

[4] TIMBUS: http://timbusproject.net/

[5] myExperiment: http://www.myexperiment.org/

[6] Taverna: http://www.taverna.org.uk/

instance of the workflow, and process documentation other than a visualisation of the workflow, are rarely provided together with the workflow. Process Management Plans address this problem and ensure that necessary information is provided by combining tools developed in the aforementioned projects. They also put more focus on the research process than executable papers, which are described in (Nowakowski et al., 2011) and require working in a closed environment.

The necessity to introduce PMPs is also driven by the rising number of scientific experiments using specialised middleware and infrastructure. One of the efforts aiming to provide such infrastructure is described by Aiftimiei et al. (2012). The authors describe steps towards 'providing a consistent platform, software and infrastructure, for all users in the European Research Area to gain access to suitable and integrated computing resources.' The results may not be reproducible if the settings of the infrastructure used to conduct the experiment are not properly documented. Another reason why PMPs need to be used is the problem of losing the traceability of an experiment, which is a consequence of poor data management. Different data sets scattered around different machines, with no track of dependency between them, are a common landscape for particle physicists, who move quickly from one research activity to another (Curry, 2011). The analysis presented by Gronenschild et al. (2012) demonstrates the impact of a system's software and hardware configuration on the results obtained in neuroanatomical studies. Despite the implementations of algorithms being identical, if the infrastructure used for computation is different the final results may vary, which hinders the reproducibility of the experiment. This could be also one of the reasons why reproducibility in cancer research is low (Van Noorden, 2013).

Lifecycle and Stakeholders

This section presents the lifecycle of Process Management Plans and outlines their key benefits to diverse stakeholders involved by different means in the scientific experiment.

Lifecycle

The PMP is created at the point of process design and is maintained and updated during the lifetime of the process by various stakeholders. The proposed lifecycle of the PMP is shown in Figure 1.

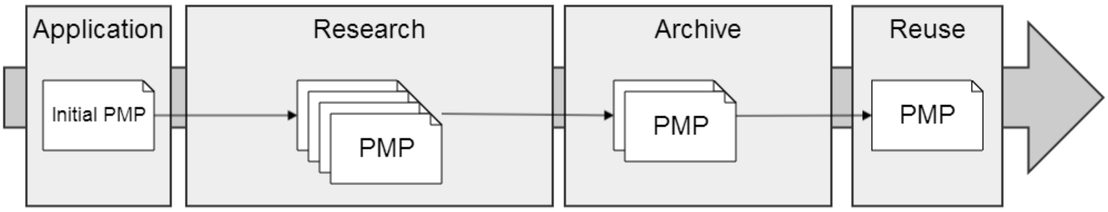


Figure 1. Research process lifecycle and stages when PMP is created or modified.

At the beginning, when the scientist applies for the research grant, an initial version of a PMP is created. It provides a high level overview on the processes being used in the experiment. When the proposal is accepted and the actual research starts, the users work using any tools and methods they prefer, no burden from the PMP is imposed on them. However, when they reach a milestone in the project, for example when they publish some of the results or an intermediate stage result is being handed over to another scientist, then the PMP requires an update, i.e. filling in information that describes the experiment. When the project is finished and the data and processes are being deposited into a repository, all the information must also be provided in the PMP. This does not mean that the lifecycle of the PMP is finished at this stage. When the process data and processes are kept in an archive, several digital preservation actions may be applied to them, such as migration, emulation, and so on. All these actions have to be reflected in the PMP, because they modify the original process. Finally, when the process is redeployed and reused in a new experiment, information stored in the PMP can be transferred to a new PMP created for the new experiment. Thus, the whole lifecycle of the process, beginning with the process design, finishing with the process reuse, is fully documented.

Stakeholders

We have identified five different groups of stakeholders depicted in Figure 2.

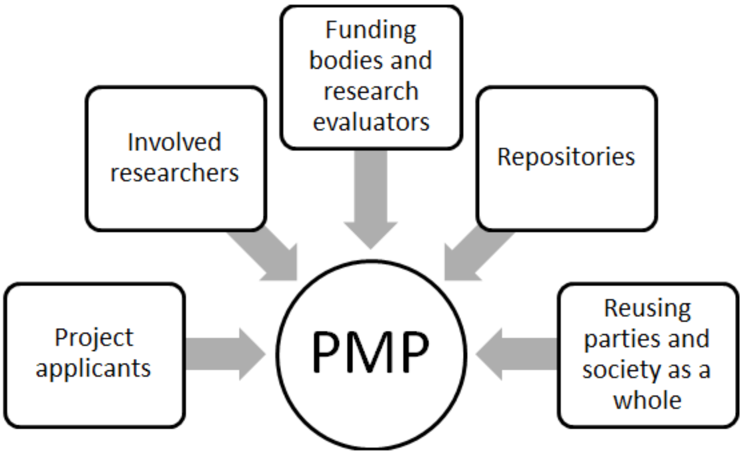

Funding bodies and research evaluators
Involved researchers
Repositories
Project applicants
PMP
Reusing parties and society as a whole

Figure 2. Stakeholders impacted by the Process Management Plan (PMP).

Project applicants will benefit by being able to better identify and plan the resources needed for the research. For example, if the process of transforming the experimental data assumes use of proprietary software with an expensive license, this information will be revealed at an early stage and can be taken into the account when applying for a grant.

Researchers will benefit by working with better documented processes. This leverages sharing of results and eases reuse of existing research. Moreover, they will need to spend less time when joining a new project, because useful information will be provided to them in a structured way by the PMP. This is especially important for cooperation within Research Infrastructures, where research challenges are addressed by cooperation of various specialists from different areas contributing to only a specific part of the experiment. The acceptance of PMPs should be comparable to the acceptance of the DMPs or possibly higher, because the comparable effort brings more benefits to the researchers and much of the information will be collected automatically.

From the point of view of funding bodies, PMPs safeguard the investment made into research by ensuring research results are trustable/verifiable, and can be reused at later points in time. Furthermore, PMPs facilitate cooperation between projects because they make it easier to reuse processes used in other projects and facilitate exploitation of results of other funded projects. Thus, PMPs lead to sustainable research and can save funding that can then be allocated to new research projects.

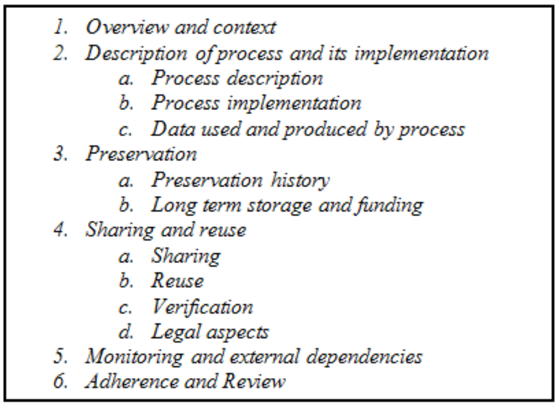
Repositories which keep the deposited processes and data can better estimate the costs of curation and plan actions needed to maintain the deposited resources. PMPs also support long term preservation (keeping processes usable over time) and provide information on possible events triggering necessary digital preservation activities.

PMPs also bring benefits to the reusing parties, as their research can be accelerated by reusable processes. The reusing parties have also higher confidence that they can build on previous work, because the quality is higher due to reproducibility. Furthermore, the scientists whose processes are reused for other experiments can gain higher recognition and credit.

Process Management Plan

This section presents the proposed structure of a Process Management Plan, which is supposed to be a living document. The specification presented in this paper does not assume any particular target implementation of the PMP. Yet, we believe that the PMP has to be more than just a paper report (or a digital version of it). In fact, in order to make the PMPs actionable and enforceable, the automation of its creation and machine readability are required. This helps to ensure higher precision and coherence of information included in a PMP. Moreover, it decreases preparation time, because some of the information is collected automatically. Tools, methods and other solutions that may facilitate the implementation of PMPs are presented in Related Works section.

The proposed structure of the PMP, following and building on guidelines for Data Management Plans, is presented in Figure 3. The PMP sections are described in the consecutive subsections.



1. *Overview and context*
2. *Description of process and its implementation*
   a. *Process description*
   b. *Process implementation*
   c. *Data used and produced by process*
3. *Preservation*
   a. *Preservation history*
   b. *Long term storage and funding*
4. *Sharing and reuse*
   a. *Sharing*
   b. *Reuse*
   c. *Verification*
   d. *Legal aspects*
5. *Monitoring and external dependencies*
6. *Adherence and Review*

Figure 3. The structure of a Process Management Plan.

Overview and Context

This section of the PMP provides a high level overview of the research activity and its context. It allows quick identification of what the project is about, who is involved in it and what requirements and constraints apply to the research project. This information should follow a precisely defined schema for automated analysis and processing. It should cover things like the project name, funding body, budget, duration, research objectives, requirements and policies that influence the creation of the PMP, people and organisations involved, state of the PMP, and so on.

Description of Process and Implementation

Each process used during the course of the research must be described. This description should consist of three main parts presented below.

Process description

The process description included in a PMP should be given at different levels of detail. It should provide not only an executive summary, which allows quick understanding of the purpose of the process, but also a more detailed description of the steps involved, including the data, research methods and approaches used. These processes may be described in a wide range of ways, such as textual description, UML diagrams, workflows, and so on. The description used should follow the best practices of the given scientific community. Moreover, it should include a specification of both functional and non-functional characteristics of each process, as well as any auxiliary resources that may help future users to understand the process, such as publications, slides, tutorials, etc. A list of software and hardware requirements, but also legal restrictions (such as licenses, obligations regarding used research data like personal data) will be required. All the information could be grouped into a Research Object (Page et al., 2012), which is one of the latest community efforts to provide a container for artefacts describing scientific experiment.[7]

Process implementation

In order to analyse and reuse a process, its structure must be understood and documented. This implies that all components that are used within the process implementation, their dependencies and relations between them have to be discovered. The comprehensive process description may require different kinds of process analysis. The high level process description as a sequence of generic actions described in the previous section needs to be accompanied by a description at the technological level, which can be viewed as a sequence of inputs and outputs from software and hardware components. Therefore, the infrastructure used to run the experiment and specific software and hardware needed to run the process – such as special database software, libraries, software device drivers, fonts, codecs – have to be covered in this PMP section. It is essential to capture the full process context, including all the dependencies and relations between them, because this information is crucial for reusing the process, as well as ensuring its continuity by applying digital preservation actions. The ontology based context model (Mayer et al., 2012) can be used to store this information in a structured way, creating the possibility to automate its creation using software tools.

[7] Research Object for Scholarly Communication Community Group: http://www.w3.org/community/rosc/

Data used and process by process

Reference to data sets that are used in a process have to be provided. This part links to an accompanying Data Management Plan for more information on used and produced data sets. Intermediate and result data of the process needs to be described in the DMP. For example, information on the data formats used in the process may help researchers to take decisions based on whether they can easily reuse the process with their own datasets. Existing templates for DMP specification can be reused to provide this information. Furthermore, techniques providing unambiguous identification of data sets and allowing data citation (Proell and Rauber, 2013) can complement the description.

Preservation

PMPs have a special focus on preservation, which is considered from the very beginning of the research process. This influences the design of processes and selection of tools for their implementation, and improves the sustainability of the entire process. We believe that careful process documentation will significantly improve the sustainability of research, as more information allowing the analysis, maintenance, porting or modification the process will be available. Furthermore, planning for long term storage and securing funding for this purpose in advance increases the confidence that the research results will be available in the future. Two kinds of information concerning preservation and required by PMPs are discussed below.

Preservation history

A PMP is a living document edited by multiple stakeholders, which collects information on actions that are performed to maintain the process over time. This information can be provided by the repository that takes care of the long term availability of the process. For example, outdated or obsolete hardware may be emulated; the process data may be migrated to a new format; or the part of the process in which data is read may be newly implemented. A full track of changes to the original process implementation and evidence that these actions were performed correctly is necessary to maintain the authenticity of the process. Such information can be automatically obtained from tools that assist preservation planning (Kulovits et al., 2008).

Long term storage and funding

The sustainability of research results is increased by depositing the process in a trusted repository. The information on how long the research object will be kept is specified in this section of the PMP. Some parts of process may be discarded after certain time periods, keeping only some artefacts beyond certain time periods. People or institutions responsible for taking decisions about the deposited process after the end of a research project have to be assigned. Finally, information on the funding of actions to ensure the sustainability of processes (such as preservation actions or costs of storage) is specified.

Sharing and Reuse

It is important for the modern science to share and reuse results of other experiments. PMPs foster this by making the processes and their data easily available. They also list potential purposes for reuse, as well as providing verification methods and data which can be used to verify if the reused process behaves like the original process. Finally, PMPs provide information on the legal regulations and ethical issues related to the process.

Sharing

At this point, a location where the process, its implementation and documentation about it will be stored, and means of providing access, are specified. The conditions on which the resource can be accessed are also provided, such as whether the access is free or paid. Besides this, information on where the research results are published and how the location of the process is disseminated (e.g. scientific paper, blog, presentations, etc.) has to be given. If the process cannot be shared (due to non-disclosure agreement), then this information has to be provided here.

Reuse

There are many possible process reuse scenarios. For example: re-running the original experiment for a litigation case, applying a process to new data, reproducing the experiment with improved computation algorithms or tools, reusing part of the experiment to build a new experiment, and so on. In every case, the process must be ported, installed and configured in a new environment. Although, the comprehensive description provided in previous sections provides exhaustive information on the process and its dependencies, it may still not be sufficient for setting up and reusing the process. Therefore, a list of actions which help to port, install and configure the process on a new platform is needed.

Verification

The PMP provides information and data that allows the verification of preserved and re-run processes. Processes are implemented in complex IT environments in which many indirect elements may impact the process' results or performance. For example, the process may have a different execution time on a different processor, or peripheral devices may deliver results to the process with different precision which leads to different values on the outputs of the computational process. Before any reasoning on the process can be made, the process has to be verified for its conformance to the original behaviour. Therefore, a set of precisely described tests showing process conformance is described in this section. Framework presented by Miksa et al. (in press) can be used to drive the verification process.

Legal aspects

While the Share and Reuse sections of this paper focus on the technical means of providing the access to the process, this section focuses on the legal aspects of working with the process. At this point, all necessary information on licenses, copyrights of data and software are specified. Any legal regulations affecting the reuse of the process or any ethical or privacy issues (e.g. confidentiality of data used) that may restrict the use or distribution of the entire process or its parts should be described here.

Monitoring and External Dependencies

Processes are implemented in a specific environment, which must be available in order to run the processes. PMPs specify the process components needed to run the process and aim to ensure that information about them is available. For example, if the process uses an external web service to import some data, then such a web service has to be monitored for its availability. Otherwise, if the process is run again and the web service is no longer available, then the process is not operable and very likely useless. Therefore, the PMPs should specify a list of critical dependencies that should be periodically monitored for their availability. Such monitoring can be a part of preservation infrastructure, as specified by the SCAPE project[8]. More examples of threats to processes can be found in Zhao et al. (2012). Only active monitoring of process dependencies, accompanied by the application of mitigation strategies ensuring availability of a required resource, can ensure reusability of processes.

Adherence and Review

Due to the fact that the PMP is a living document, it must be kept up to date and must reflect the actual actions that took place in each of its lifecycle phases. In order to ensure adherence, a person has to be assigned responsibility for the process. Furthermore, the reviews and methods applied to ensure that the PMP reflects reality have to be in place. This information needs to be specified at a very early stage of PMP development course and is highly dependent on the implementation of the PMP. As it was already mentioned, the actionability of the PMP can foster its enforceability. This can be achieved when the PMP is not a static textual report, but a structured document interpreted by the machine, parts of which can be generated automatically. In such case, the adherence to the PMP can be relatively easy checked and confirmed. Otherwise, manual reviews conducted, for example, by auditors representing the funding agency are needed and have to be planned in advance. The outcome of inspections has to be included in the PMP as well.

Conclusions and Future Work

In this paper a novel concept of Process Management Plans was presented. It aims to describe scientific experiments taking a process-centric view, viewing data as the result of underlying processes, such as capture, (pre-) processing, transformation, integration analyses and presentation. PMPs foster identification, description, sharing and preservation of scientific processes.

We specified the structure and content of Process Management Plans by analysing existing templates for Data Management Plans and extended them to provide support for processes. We considered also potential stakeholders for whom the information provided by the Process Management Plans would be useful (project applicants, involved researchers, funding bodies and research evaluators, repositories, reusing parties and society as a whole) and presented the lifecycle of the Process Management Plan. Finally, we provided recommendations on how the concept of Process Management Plans could be embodied using existing tools and concepts from the field of Process Management and Digital Preservation.

We are currently working on automation of PMP creation and verification by extracting process characteristics automatically from its environment. Our specific focus is on tool support to automate many of the various documentation steps, specifically capturing and monitoring of process components used in the process implementation. Moreover, we are currently evaluating the PMP with stakeholders from different scientific communities.

[8] SCAPE: http://www.scape-project.eu

Acknowledgments

This research was co-funded by COMET K1, FFG - Austrian Research Promotion Agency and by the European Commission under the IST Programme of the 7th FP for RTD - Project ICT 269940/TIMBUS.

References

Aiftimiei, C., Aimar, A., Ceccanti, A., Cecchi, M., Meglio, A. D., Estrella, F., Fuhrmam, P., Giorgio, E., Knya, B., Field, L., Nilsen, J. K., Riedel, M., & White, J. (2012). Towards next generations of software for distributed infrastructures: The European middleware initiative. In *Proceedings of the 2012 IEEE 8th International Conference on e-Science*. IEEE Computer Society. doi:10.1109/eScience.2012.6404415

Australian National Data Service. (2011). *ANDS guides - awareness level: Data management planning.* Retrieved from http://ands.org.au/guides/data-management-planning-awareness.pdf

Begley, C.G., & Ellis, L.M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature, 483*(7391), 531–533. doi:10.1038/483531a

Curry, A. (2011). Rescue of old data offers lesson for particle physicists. *Science, 331*(6018), 694–695. doi:10.1126/science.331.6018.694

Darby, R., Lambert, S., Matthews, B.,Wilson, M., Gitmans, K., Dallmeier-Tiessen, S., Mele, S., & Suhonen, J. (2012). Enabling scientific data sharing and re-use. In *Proceedings of the 2012 IEEE 8th International Conference on e-Science*. IEEE Computer Society. doi:10.1109/eScience.2012.6404476

De Roure, D. (2011). Machines, methods and music: On the evolution of e-research. In *2011 International Conference on High Performance Computing and Simulation (HPCS)*. doi:10.1109/HPCSim.2011.5999801

Donnelly, M. & Jones, S. (2011). *Checklist for a Data Management Plan.* Retrieved from http://www.dcc.ac.uk/sites/default/files/documents/data-forum/documents/docs/DCC_Checklist_DMP_v3.pdf

European Commission. (2012). *Commission recommendation of 17.7.2012 on access to and preservation of scientific information.* Retrieved from http://ec.europa.eu/research/science-society/document_library/pdf_06/recommendation-access-and-preservation-scientific-information_en.pdf

European Strategy Forum on Research Infrastructures. (2011). *Strategy report on Research Infrastructures: Roadmap 2010.* Retrieved from http://ec.europa.eu/research/infrastructures/pdf/esfri-strategy_report_and_roadmap.pdf

Garijo, D., & Gil, Y. (2011). A new approach for publishing workflows: Abstractions, standards, and linked data. In *Proceedings of the 6th Workshop on Workflows in Support of Large-Scale Science*. New York, NY: ACM. doi:10.1145/2110497.2110504

Gronenschild, E.H.B.M., Habets, P., Jacobs, H.I.L., Mengelers, R., Rozendaal, N., van Os, J., & Marcelis, M. (2012). The effects of FreeSurfer Version, Workstation Type, and Macintosh Operating System Version on anatomical volume and cortical thickness measurements. *PLoS ONE 7*(6): e38234. doi:10.1371/journal.pone.0038234

Hendler, J. (2007). Reinventing academic publishing, part 2. *IEEE Intelligent Systems, 22*(6), 2–3. doi:10.1109/MIS.2007.116

Hey, T., Tansley, S., & Tolle, K. (Eds.). (2009). *The fourth paradigm: Data-intensive scientific discovery.* Microsoft Research.

Killeen, N., Lohrey, J., Farrell, M., Liu, W., Garic, S., Abramson, D., Nguyen, H., & Egan, G. (2012). Integration of modern data management practice with scientific workflows. In *Proceedings of the 2012 IEEE 8th International Conference on E-Science*. IEEE Computer Society. doi:10.1109/eScience.2012.6404426

Koski, K. (2012). *EUDAT: Towards a pan-European collaborative data infrastructure.* Retrieved from http://eudat.eu/system/files/EUDAT_conference 23_10_2012.pdf

Kulovits, H., Becker, C., Kraxner, M., Motlik, F., Stadler, K., & Rauber, A. (2008). Plato: A preservation planning tool integrating preservation action services. In *Lecture Notes in Computer Science, Vol. 5173. Research and Advanced Technology for Digital Libraries* (pp. 413–414). doi:10.1007/978-3-540-87599-4_49

Lesk, M. (2012). Scientific data quality: Openness, provenance, and replication. In G. Marchionini, C.A. Lee, H. Bowden, & M. Lesk (Eds.), *Curating for Quality: Ensuring Data Quality to Enable New Science* (pp. 42–44). Chapel Hill, NC: University of North Carolina. Retrieved from http://datacuration.web.unc.edu/files/2012/10/NSF_Data_Curation_Workshop_Report.pdf

Mayer, R., Rauber, A., Neumann, M.A., Thomson, J., & Antunes, G. (2012). Preserving scientific processes from design to publication. In *Lecture Notes in Computer Science, Volume 7489. Theory and Practice of Digital Libraries* (pp. 113–124). doi:10.1007/978-3-642-33290-6_13

McCullough, B.D. (2007). Got replicability? The Journal of Money, Credit, and Banking Archive. *Econ Journal Watch, 4*(3), 326–337. Retrieved from http://econjwatch.org/file_download/170/2007-09-mccullough-econ_practice.pdf

Miksa, T., Mayer, R., & Rauber, A. (in press). Ensuring sustainability of web services dependent processes. *International Journal of Computational Science and Engineering*.

Miksa, T., Pröll, S., Mayer, R., Strodl, S., Vieira, R., Barateiro. J., & Rauber A. (2013). Framework for verification of preserved and redeployed processes. In J. Borbinha, M. Nelson, & S. Knight (Eds.), *Proceedings of the 10th International Conference on Digital Preservation*. Retrieved from http://purl.pt/24107/1/iPres2013_PDF /Framework for Verification of Preserved and Redeployed Processes.pdf

Miksa, T., & Rauber, A. (2013). Increasing preservability of research by process management plans. In *Proceedings of the 1st International Workshop on Digital Preservation of Research Methods and Artefacts (DPRMA '13)*. New York, NY: ACM. doi:10.1145/2499583.2499591

National Science and Technology Council, Committee on Science, Interagency Working Group on Digital Data. (2009). *Harnessing the power of digital data for science and society.* Retrieved from http://www.nitrd.gov/About/Harnessing_Power_Web.pdf

National Science Foundation. (2011). *Data management for NSF EHR Directorate.* Retrieved from http://www.nsf.gov/bfa/dias/policy/dmpdocs/ehr.pdf

Nowakowski, P., Ciepiela, E., Harezlak, D., Kocot, J., Kasztelnik, M., Bartynski, T., Meizner, J., Dyk, G., & Malawski, M. (2011). The collage authoring environment. *Procedia CS, 4,* 608–617.

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

Proell, S., & Rauber, A. (2013). Scalable data citation in dynamic large databases: Model and reference implementation. In *2013 IEEE International Conference on Big Data* (pp. 307-312). doi:10.1109/BigData.2013.6691588

Rice, R., Ekmekcioglu, C., Haywood, J., Jones, S., Lewis, S., Macdonald, S., & Weir, T. (2013). *Implementing the research data management policy: University of Edinburgh roadmap.* Paper presented at the 8th International Digital Curation Conference, Amsterdam.

Roure, D., Goble, C., Aleksejevs, S., Bechhofer, S., Bhagat, J., Cruickshank, D., … Zhao, J. (2010). The evolution of myExperiment. In *2010 IEEE Sixth International Conference on e-Science* (pp.153–160). IEEE Computer Society. doi:10.1109/eScience.2010.59

Strodl, S., Draws, D., Antunes, G., & Rauber, A. (2012). Business process preservation, how to capture, document and evaluate. In *Proceedings of the 9th International Conference on Preservation of Digital Objects (IPRES2012)* (pp. 117-125).

Strodl, S., Petrov, P., & Rauber, A. (2011). *Research on digital preservation within projects co-funded by the European Union in the ICT programme.* Vienna University of Technology. Retrieved from http://cordis.europa.eu/fp7/ict/creativity/ report-research-digital-preservation_en.pdf

Thanos, C., Manegold, S., & Kersten, M.L. (2012). Big data: Introduction to the special theme. *ERCIM News, 89.* Retrieved from http://ercim-news.ercim.eu/en89/special/big-data-introduction-to-the-special-theme

Van der Graaf, M. & Waaijers, L. (2011). A surfboard for riding the wave: Towards a four country action programme on research data. *Knowledge Exchange.* Retrieved from http://www.knowledge-exchange.info/surfboard

Van Noorden, R. (2013). *Initiative gets $1.3 million to verify findings of 50 high-profile cancer papers.* Nature News Blog. Retrieved from http://blogs.nature.com/news/2013/10/initiative-gets-1-3-million-to-verify-findings-of-50-high-profile-cancer-papers.html

Zhao, J., Gmez-Prez, J.M., Belhajjame, K., Klyne, G., Garca- Cuesta, E., Garrido, A., … Goble, C.A. (2012). Why workflows break: Understanding and combating decay in Taverna workflows. In *Proceedings of the 2012 IEEE 8th International Conference on e-Science*. IEEE Computer Society. doi:10.1109/eScience.2012.6404482