Data Quality for New Science: Process Curation, Curation Evaluation and Curation Capabilities

Andreas Rauber
Vienna University of Technology
Favoritenstrasse 9, 1040 Vienna, Austria
rauber@ifs.tuwien.ac.at

ABSTRACT

In order to fully support the potential of data-driven science, eScience, the 4th Paradigm, and other similar concepts, we face significant challenges in curating the data, ensuring its authenticity, accessibility, proper reusability and repurposing in different contexts. So far, the primary focus in these areas has been on documentation and preserving the actual data. This position paper argues for an approach focusing on the curation of the actual processes involved in the collection, pre-processing and use of data, capturing process contexts and the actual processes together with the data. We further present an approach on how to validate and measure conformance of a re-activation of any such process to ensure and prove authenticity and validity. Last, but not least, we argue in favor of a capability and maturity based view of data and process curation, rather than mere auditing and certification, and the establishment of supporting (IT-)processes.

General Terms

E-Science, Research Infrastructures, Process Preservation, Context Information, Evaluation Framework, Enterprise Architectures, Maturity Model

1. INTRODUCTION

Like all digital data, research data is exposed to threats of digital obsolescence, i.e. when the digital objects become unusable. This may occur on three different levels - the bit level, the logical level, and the semantic level. While a range of solutions and best practice experience exists for bit-level preservation, most of digital preservation research focuses on logical preservation, i.e. ensuring that the file formats that the information is provided in remains accessible by current software versions. For research data, this challenge in some aspects is both harder as well as easier than for many conventional objects: on the one hand, research data is frequently represented in some form of numeric representation that is both rather stable in terms of accessibility, with simpler format specifications, a clearer separation between data and functionality, i.e. no embedded code, and thus simpler transformation settings for data migration. On the other hand, research data preservation at the logical level is more complex, as in many cases both data formats as well as preservation requirements are rather unique to each data set, with characteristics of data sets ranging both from individual data sets with massive volumes of data items to myriads of rather small data sets, each with their own and very specific designated community. Yet, the most serious challenge to data curation arises at the semantic level, ensuring the authenticity and correct interpretability of data. Conventionally, this comprises capturing as much information about the data, its preprocessing and use as well as actions performed on the data during curation activities as possible in order to establish provenance and interpretability.

We claim, however, that several aspects related to data curation, specifically with a focus on ensuring its quality, are not receiving sufficient attention in current R&D. This paper summarizes some of our current considerations and areas of research focus with respect to data curation both at the Vienna University of Technology[1] as well as at Secure Business Austria[2], most prominently in the research projects SCAPE[3], TIMBUS[4], APARSEN[5] as well as some new initiatives on data curation and evaluation to be launched.

First, establishing context of data is focused strongly on documentation, i.e. documenting intention, data capture, and potential processing steps and many others. Yet, specifically with respect to data (pre-)processing, pure documentary approaches are probably not sufficient: as the processing modules and processes become more complex, the risk of either not fully documenting the process or of the process as implemented not perfectly following the intended process grows. As a result, erroneous pre-processing software, processing steps not obeyed due to misunderstanding or lack of diligence etc. may lead to artifacts being introduced into the data, or lead to incoherent results when trying to repeat experiments under identical conditions. We thus argue that capturing and curating the (pre-)processing processes

[1] http://www.ifs.tuwien.ac.at/dp
[2] http://www.sba-research.org/research/data-security-and-privacy/digital-preservation
[3] http://www.scape-project.eu
[4] http://timbusproject.net
[5] http://aparsen.digitalpreservation.eu/

is in many cases an integral part of data curation. It also enables re-running earlier experiments with new data under identical conditions. We thus are currently working on new approaches for process and process context capture, documentation, preservation and re-activation [10, 8, 9].

Second, once the processes are curated as part of the data, mechanisms, strong emphasis must be placed on establishing whether any re-activation of research data is actually faithful to the original with regards to a set of determined significant properties. We feel there is a lack of established mechanisms and frameworks, both at the data/process capture as well as re-activation phases, to determine whether all essential aspects offered by a new viewing application, after a transformation, or even when opening objects in an emulated environment. In fact, it can be shown that both migration as well as emulation approaches are rather identical in character, and need to be evaluated in very similar manners [5]. We thus are currently investigating more formal frameworks for documenting and verifying identity of digital objects on re-use with respect to established properties [3, 4].

Third, data curation requires the consistent application of well-defined processes in a highly repeatable, consistent, well-documented manner to ensure trustworthiness. While these may partially be handled by institutions whose primary focus is data curation, we see a shift in such operations occurring as part of other primary business operations. This will result in a shift from current thinking of operational data on the one hand vs. dedicated archival data holdings on the other to a merged operational data repository with integrated preservation capabilities. It will also require an integration of curation activities into standard (IT) operations. Thus, models and standards from data curation will need to be merged with concepts from IT Governance and Enterprise Architectures to allow a consistent view on curation activities as part of a institutions operations. Beyond audit and certification establishing conformance to specific requirements, capabilities and maturity models may offer a more flexible and realistic approach to establishing the competences and improving them, guiding investment and ensuring proper alignment with an institutions objective. We are thus reviewing ways to align the two worlds of IT Governance and Digital Curation, defining capabilities and establishing maturity models to allow for process evaluation and improvement. [1, 2].

The following sections review some of the initial concepts developed clarifying their scope and outlining future directions.

2. FROM DATA PRESERVATION TO PROCESS CURATION

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, preprocessed 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, one needs to go beyond the classical concerns of Digital Preservation research, and consider more

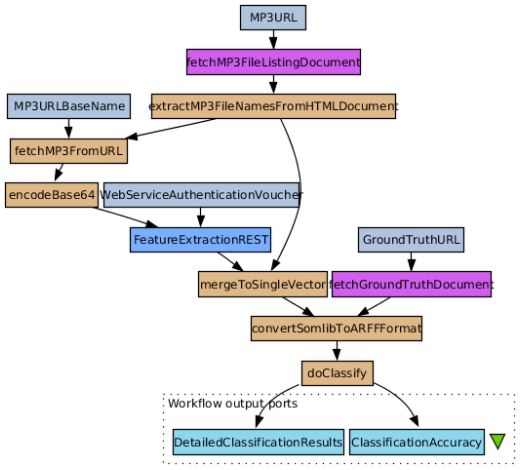

Figure 1: Musical genre classification, including fetching of data, modelled in the Taverna workflow engine

than the preservation of data. The following passages and example are adopted from [9] detailing our approach to process preservation on a simple example from the music retrieval domain.

To move towards more sustainable E-Science processes, we recommend implementing them in workflow execution environments. For example, we are currently using is the Taverna workflow engine [11]. 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 pre-defined scripts (e.g. for encoding binaries via Base64), 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 the classification model or feature extraction, become explicit, either in the form of process input data, or in the script code.

Figure 1 shows an example of a 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-

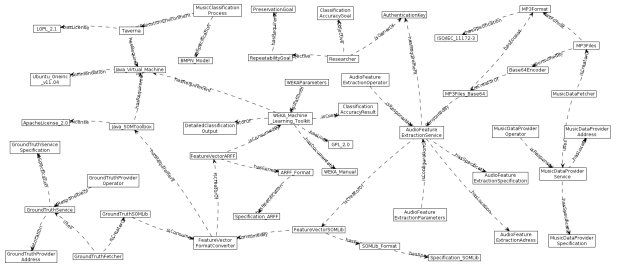

Figure 2: Context Model of musical genre classification process

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[12] 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, thus enabling verification of the results obtained. This is essential for being able to verify system performance upon re-execution, specifically when any component of the process (such as underlying hardware, operating systems, software versions, etc.) have changed.

Curation of business or E-Science processes requires capturing the whole context of the process, including e.g. different or evolved 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 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. The model consists of approximately 240 elements, structured in around 25 major groups. It corresponds to some degree to the representation information network [7], modeling the relationships between an information object and its related objects, be it documentation of the object, constituent parts and other information required to interpret 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. The model is implemented in the form of an ontology, which on the one hand allows

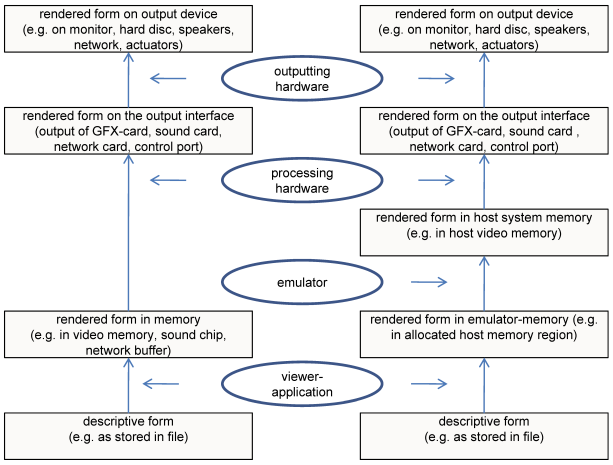

Figure 3: Different forms of a digital object in a system's memory. On the left the layers in an original system are shown, on the right the layers in the system hosting the emulator are shown.

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 gives an overview on the concrete instances and their relations identified as relevant aspects of the process context for the music classification process discussed above.

3. EVALUATING PROCESS RE-ACTIVATION

A critical aspect of re-using digital information in new settings is its trustworthiness, 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 the this framework include a (1) description of the original environment, (2) the identification of external events influ-

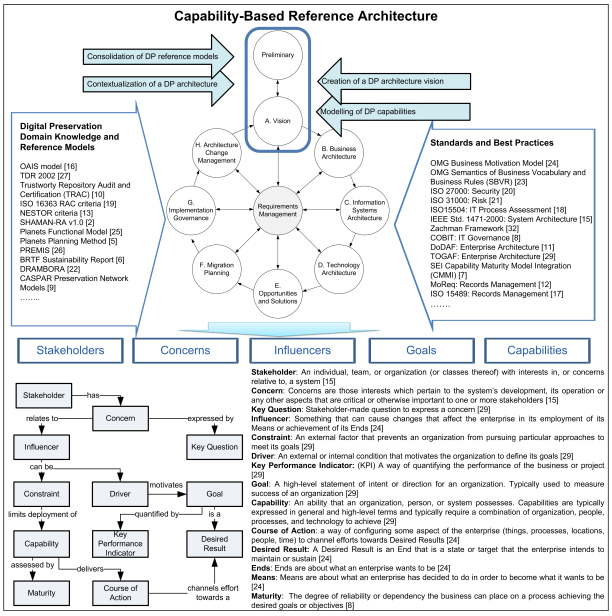

Figure 4: Using TOGAF to integrate reference models creating a uniform view [1]

encing 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. 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.

A further component of the framework is the identification at which levels to measure the faithfulness of property preservation, as depicted in Figure 3. 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 are necessary 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.) [13] An example of applying this framework to evaluation of preservation actions is provided in [4]

4. A CAPABILITY MODEL APPROACH TO DIGITAL CURATION

The types of institutions facing data curation challenges expands beyond the cultural heritage domain to include settings where curation is not the primary business goal. Rather, availability of data and processes is seen as an essential driver, be it due to legal/compliance requirements, as a contribution to business value, or other motivations. In settings where curation is not the main focus, it needs to be aligned with other core activities, integrating smoothly with its primary operations.

Data (and process) curation in research settings may be a typical example when curation is not delegated to a specific institution designated to preserve the data, but when preservation is happening as part of the research (and continued re-use) process. Moving beyond the more traditional data creation and use vs. data archiving approach we may want to aim at integrating all processes that revolve around data smoothly (and transparently for most actors) with curation activities.

To reach this goal, perspectives and approaches from fields such as Enterprise Architectures, Information Systems, Governance, Risk and Compliance may help in achieving a different view on data curation. This will assist in integrating digital curation as part of more generic (IT) operations while also offering a chance to make the needs and benefits of digital curation contributions to the overal value chain of an institution explicit. An overview of such an integrated view based on TOGAF [14], merging different models with the Shaman reference architecture is depicted in Figure 4. We also think that a process-based view on data curation rather than a data-centric view may help to better understand responsibilities, risks and costs involved to meet specific goals. It should also offer a more flexible basis for assessing the capabilities of an institution with respect to data curation, the level of maturity aimed at for specific capabilities, and allow more targeted actions to be planned in order to achieve them.

To this end we have further started modeling curation as a set of capabilities, with a range of maturity levels, as well as a clear specification of drivers and constraints, and their impact on an organization. An example of maturity levels for the capability *Preservation Operation* is depicted in Tab. 1. A detailed discussion of this approach is provided in [1, 2].

5. CONCLUSIONS

Ensuring quality in data curation for research is both simpler as well as more complex than "standard" digital preservation. While it is in many respects similar to any kind of (more traditional, document-centric) data preservation, it raises significant challenges that require solutions going beyond what is currently available as state of the art solutions. While several aspects are predominantly extensions to cover e.g. new/specialized data formats, several challenges are rather unique in their importance to ensure the quality and authenticity of research data.

On the one hand, processes are an essential part of data provenance. Ensuring that any processing steps can be repeated, either on original data for verification and analysis purposes, or on new data to assure identical conditions, poses significant challenges in maintaining entire processing environments available and usable.

With the preservation of more complex environments, particular challenges emerge with respect to verifying the au-

	Awareness and Communication	Policies, Plans and Procedures	Tools and Automation	Skills and Expertise	Responsibility and Accountability	Goal Setting and Measurement
1	Management recognizes the need for preservation operations. There is inconsistent and sporadic communication.	Some operations are carried out, but they are not controlled. No useful documentation is produced about procedures and actions.	Some tools may be employed by individuals in an unsystematic ad-hoc manner.	There is no common awareness of which skills and expertise are required for which tasks.	There is no common awareness of responsibilities.	There is no clear awareness of goals; operations solely react to incidents and are not tracked.
2	Management is aware of the role of operations for authenticity and provenance. No formal reporting process exists, but there is some documentation about process results. Reports are delivered by individuals.	Some operational procedures emerge, but they are informal and intuitive. Operations rely on individuals; different procedures are followed within the organization. QA is recognized as a process, but mostly carried out ad-hoc and manual.	Automated tools are beginning to be employed by individuals based on arising needs and availability. Their usage is unsystematic and incoherent.	Staff obtain their operational skills through hands-on experience, repeated application of techniques and informal training by their peers.	Responsibility for operations emerges, but is not documented. Accountability is not defined.	There is individual awareness of short-term goals to achieve in operations, but no consistent goal definition or measurement.
3	Management understands the role of operations for authenticity and provenance. There are guidelines about statistics and reporting procedures, but they are not consistently enforced.	There is a defined process for all operations that relies on standardized plans. The processes and rules used are defined by available components, services and skills. QA and metadata management are not driven by business goals.	Plans are deployed according to specifications, but the process of initiating operations is mostly manual. No integrated system exists for tracking the state and results of operations.	A formal training plan has been developed that defines roles and skills for the different sets of operations, but formalized training is still based on individual initiatives.	Responsibility for operations is assigned, but accountability is not provided for all operations.	Operational goals are specified, but no formal metrics are defined. Measurements take place, but are not aligned to goals. Assessment of goal achievement is subjective and inconsistent.
4	Management fully understands the role of operations for authenticity and provenance and how they relate to business goals in the organization. Reporting processes are fully specified and adhered to.	Plans are fully deployed as operational activities, and the compliance of all operations to goals and constraints specified in plans is fully monitored. All Operations are actively monitoring state of operations.	An automated system exists to control automated operations, and automated components are widespread, yet not fully integrated.	Required skills and expertise are defined for all roles, and formal training is in place.	Responsibility and accountability for all operations is clearly defined and enforced.	A measurement system is in place and metrics are aligned with goals. Compliance monitoring is supported and compliance enforced in all operations.
5	Operations are continuously improving. An integrated communication and reporting system is fully transparent and operates in real time.	Extensive use is being made of industry good practices in plan deployment, analysis, actions, metadata, QA, and reporting.	All operations are fully integrated, status is constantly available in real-time.	Operators have the expertise, skills and means to conduct all operations. Continuous skills and expertise assessment ensures systematic improvement.	A formal responsibility and accountability plan is fully traceable to all operations.	Compliance is constantly measured automatically on all levels. Continuous assessment drives the optimization of measurement techniques.

Levels: 1: Initial/Ad-Hoc, 2: Repeatable but Intuitive, 3: Defined, 4: Managed and Measurable, 5: Optimized [6]

Table 1: Maturity Levels for the capability *Preservation Operation* [1]

thenticity of the performance/rendering of a process or data object in such a preserved environment. Formal models for these, as well as assistance in identifying and capturing the essential aspects needed for subsequent verification still represents a significant hurdle, with even more severe difficulties emerging from the need of automating any such validation in more generic settings.

Last, but not least, we feel that a shift from data-centric views of traditional approaches to depositing data somewhere for long-term curation needs to be superseeded by a view where curation processes are integrated into the operational environments. Furthermore, rather than auditing whether a specific sets of requirements is met by an institution tasked with curation we feel that a capability and maturity model based approach offers more flexibility to focus on essential aspects of data curation for a wide set of institutions.

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[6], a Digital Preservation Challenges Wiki[7] has been created, where we invite contributions and discussion. As a follow-up to the Dagstuhl seminar, a workshop on DP Challenges[8] will be held at iPRES 2012 in Toronto focusing on the elicitation and specification of research challenges.

[6] `http://www.dagstuhl.de/de/programm/kalender/semhp/?semnr=10291`

[7] `http://sokrates.ifs.tuwien.ac.at`

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

Acknowledgements

Part of this work has been supported by TIMBUS project, co-funded by the European Union under the 7th Framework Programme for research and technological development and demonstration activities (FP7/2007-2013) under grant agreement no. 269940, and the SCAPE project, under grant agreement no. 270137, and APARSEN, grant agreement no. 269977.

6. 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] IT Governance Institute. *COBIT 4.1. Framework – Control Objectives – Management Guidelinces – Maturity Models*. 2007.

[7] Y. Marketakis and Y. Tzitzikas. Dependency management for digital preservation using semantic web technologies. *International Journal on Digital Libraries*, 10:159–177, 2009.

[8] R. Mayer, S. Pröll, and A. Rauber. On the applicability of workflow management systems for the preservation of business processes. In *Proceedings of the 9th conference on Preservation of Digital Objects (iPRES2012)*, Toronto, Canada, October 1–5 2012.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] G. Phillips. Simplicity betrayed. *Communications of the ACM*, 53(6):52–58, 2010.

[14] The open Group. *TOGAF Version 9*. Van Haren Publishing, 2009.