Digital Preservation of a Process and its Application to e-Science Experiments

Stephan Strodl
SBA Research
Vienna, Austria
sstrodl@sba-research.org

Rudolf Mayer
SBA Research
Vienna, Austria
rmayer@sba-research.org

Gonçalo Antunes
INESC-ID
Lisbon, Portugal
goncalo.antunes@ist.utl.pt

Daniel Draws
SQS
Cologne, Germany
daniel.draws@sqs.com

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

ABSTRACT

The increase in computationally intensive science (called e-science) drives the need to make scientific processes available for the long term. The current approach is often to archive only the resulting publications, and at very most the data sets, of scientific experiments, which is insufficient in experimental and data intensive science. The preservation of scientific experiments and their results enables others to reproduce and verify the results as well as build on the result of earlier work. The TIMBUS projects aims at preserving processes for the long term. In this paper we present the process framework developed, and apply it to the preservation of a Music Classification evaluation process. This classification experiment represents a typical information retrieval process for classifying music into predefined categories, and evaluating the performance thereof. The paper describes and applies the process steps of the three phases of the TIMBUS approach: plan, preserve and redeploy.

Keywords

Digital Preservation, E-Science

1. INTRODUCTION

Digital preservation ensures the access to digital information objects over time. The main focus of research, so far, has targeted static digital objects such as text and multimedia documents. Recently, however, there is an increasing demand for preservation of dynamic objects and whole workflows and processes. The preservation of workflows and process is driven, besides others, by research institutions that run data intensive experiments. These experiments and their results need to be verifiable to others in the community. They need to be preserved as researchers need to be able to reproduce and build on top of earlier experiments to verify and expand on the results. Current practice in many disciplines is however often restricted to publishing results as a summary in scientific publications without detailed specification of the experiments. This is in some settings augmented by also making the data sets utilised available, but the lack of detailed information about the execution of the experiments, or the availability of the software employed, poses problems for the re-use in the long term. To avoid the loss of scientific results, work on digital preservation has thus expanded from a data centric perspective towards approaches to preserve the process to execute, render and analyse data.

Processes are increasingly supported by service oriented architectures, employing numerous services offered by different, external providers. These dependencies on third party services pose new challenges for the long term usability of processes. Software services are in general not designed for long term availability, as they rely on a number of technologies for execution, for example hardware, file formats, operating systems and other software libraries, which all face the risks of obsolescence. In the long run, the availability of today's technology cannot be guaranteed. The authentic functionality of processes in the long term can therefore be violated in terms of missing software services and outdated and unavailable technology.

The TIMBUS project[1] thus focuses on the preservation of (business) processes. The developed approaches and methods are domain independent and can be applied to different settings (e.g. business settings or E-Science domain). By analysing the execution context and identification of dependencies, the accessibility to processes and the supporting services is maintained over time. In this paper, we present the TIMBUS Preservation Process Framework, which specifies the process steps for the digital preservation of a process. The application of the preservation process is demonstrated on a use case process of a Music Classification experiment. This scientific process evaluates the performance of methods to classify music into sets of predefined genres, and is a typical task in Music Information Retrieval research.

[1] http://timbusproject.net

The remainder of this paper is organised as follows. Section 2 points out related work in the field of digital preservation with focus on holistic life cycle approaches. The Music Classification process is introduced in Section 3. The TIMBUS Preservation Process is explained in Section 4, showing the application of the music classification process. The paper concludes with a summary and outlook provided in Section 5.

2. RELATED WORK

Although digital preservation has been traditionally driven by memory institutions and the cultural heritage sector [18], it is increasingly recognized that it is a problem affecting all organizations that manage information over time, and as such it affects most of contemporary organizations where information systems provide important support to the business. Although the OAIS Reference Model [8] remains an important source of concepts to the field, it lacks directives and guidelines to address complex preservation scenarios with multiple business support systems and complex digital objects in place. In such scenarios, digital preservation requires a holistic view, acting as a combination of organisational and business aspects with system and technological aspects, so that all the contextual aspects surrounding a complex digital object can be captured and the objective of rendering it in the future in the same or in similar conditions can be attained.

With this holistic concern in mind, digital information life cycle models have been designed, of which the DCC Curation Life Cycle Model [6] and the SHAMAN Information Life Cycle [3] are noticeable examples. The DCC Curation Life Cycle Model elongates the traditional scope of preservation to include curation. It addresses two phases: a *Curation* phase, which might involve the creation of new information or the access and reuse of already existing information and its appraisal and selection; and a *Preservation* phase, which involves the ingestion of the information into the archive, the application of preservation actions, and the storing of that information. During the two phases, community watch and participation and preservation planning take place in order to keep descriptive metadata and representation information up to date. The SHAMAN Information Life Cycle, besides including the *Archival* phase already addressed by the OAIS model, suggests two additional pre-ingest phases and two additional post-access phases. The pre-ingest phases *Production* and *Assembly* aim at the capturing of the context of production of the object and its assembly into an information package, respectively. The post-access phases *Adoption* and *Use* concern the preparation of the retrieved package so that its information contents can be used.

This renewed understanding of preservation also creates the need for the development of new conceptual models that are able to synthesize this knowledge and make it re-applicable to different scenarios. The SHAMAN Reference Architecture [2] resulted from an infusion of knowledge in the digital preservation field and standards and best practices from the business and IT governance fields. It defines a set of preservation capabilities and their relationships and interaction with other organizational capabilities, so that its integration with the overall capabilities of an organization is facilitated. The overall objective is to promote the alignment between the preservation objectives of the organization, the organization's processes, and the existing technological infrastructure. Additionally, the work done on the CASPAR project on Preservation Networks [5] is a relevant reference on the capturing of the dependencies of complex digital objects through the usage of entity-relationship-like models, although business and organizational aspects are left out of it.

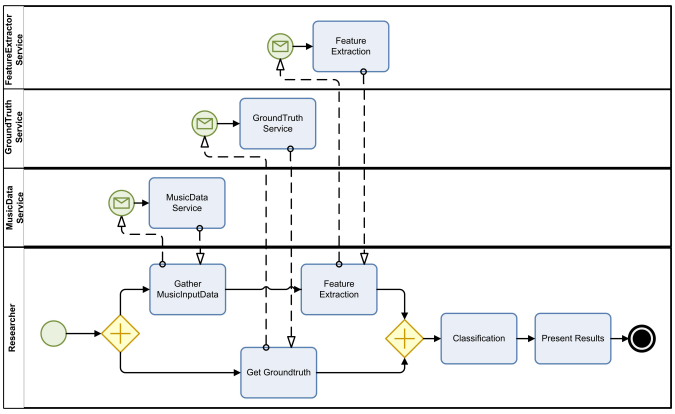

Figure 1: Music Classification experiment

Despite all the works referred to in this section, the preservation of business processes (in the form of their digital representation) along with the surrounding context needed for its long-term understandability is an innovative target being pursued by TIMBUS.

3. MUSIC CLASSIFICATION PROCESS

The process used in our case study is a scientific experiment in the domain of data mining, where the researcher evaluates the performance of an automatic classification of music into a set of predefined categories. This type of experiment is a standard scenario in Music Information Retrieval research, and is used with many slight variations in set-up for numerous evaluation settings, ranging from ad-hoc experiments to benchmark evaluations such as e.g. the MIREX genre classification or artist identification tasks [11].

The experiment involves several steps, which can partially be parallelised. First, music data is acquired from sources such as benchmark repositories or, in more complex settings, online content providers, and in the same time, genre assignments for the pieces of music are obtained from ground truth registries, frequently from websites such as Musicbrainz[2]. Tools are employed to extract numerical features describing certain characteristics of the audio files. In the case of the experimental set-up used for the case study, we employ an external Web service to extract such features. This forms the basis for learning a machine learning model using the WEKA machine learning software, which is finally employed to evaluate the prediction accuracy of genre labels for unknown music. The process is visualised using the BPMN notation in Figure 1.

[2]`http://musicbrainz.org/`

The process described above can be seen as prototypical from a range of e-Science processes, consisting both of external as well as locally available (intermediate) data, external Web services as well as locally installed software used in the processing of the workflow. In the implementation considered in this paper, it primarily consists of the following components:

- The Taverna workflow engine[3] is used to orchestrate the parallel execution and synchronisation of the process steps. Taverna further provides a scripting language based on Java that is employed for the above mentioned script tasks.
- A number of external services, all called from scripts or templates provided by Taverna, are employed:
  - The data source providing the music data is an archive with web interface, and can thus be obtained via HTTP requests.
  - The service offering the ground truth annotations, e.g. the assignment of a piece of music to a genre, is also obtained via HTTP.
  - The web service to extract features is a free service and similar to the one provided e.g. by Echonest[4]. In particular, we use a REST service that takes an MP3 file as input, and provides a vector of floating point values as descriptor.

  These services are provided by third parties, and their availability and similar function is thus not guaranteed in the future.

An illustration of the steps in this implementation of the process is given in Figure 2.

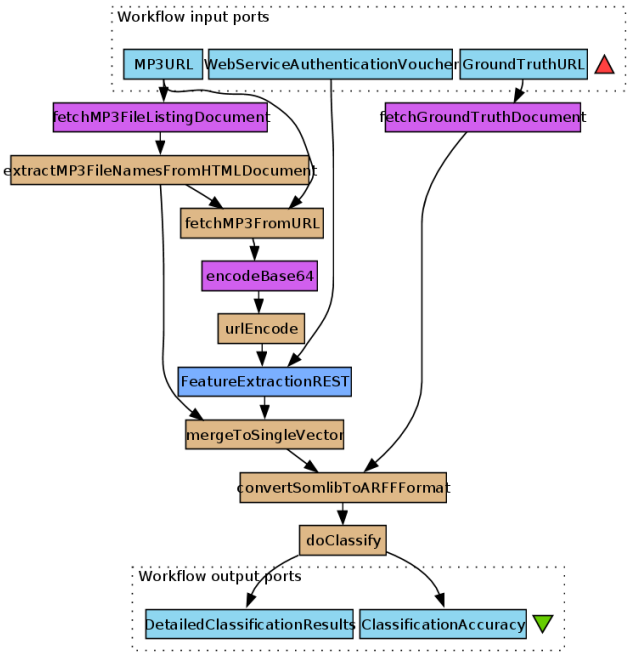

Figure 2: Music Classification experiment implemented in the Taverna workflow engine

4. TIMBUS PRESERVATION PROCESS

The TIMBUS Preservation Process to digital preserve business processes can be divided into three phases: plan, preserve and redeploy. The planning phase concerns the capture of the business process and its context. Risks of the business process are identified by reviewing contractual, policy and legal obligations. Driven from the risk management perspective, digital preservation is considered as a potential mitigation strategy. The assessment of preservation strategies identifies and evaluates different approaches to make the process available in the future. Figure 3 shows the TIMBUS process. Triggered by the risk management, the acquisition of the business process context and the *Assessment of Preservation Approaches* are executed in the planning phase (described in detail in Section 4.1). Within the preservation phase, presented in Section 4.2, the process data from the source environment are captured, preservation actions are executed and the data are prepared for archival storage. The redeployment phase, described in Section 4.3, specifies the re-initiating of the preserved process in a new environment at some point in the future. The fundamental concepts of the TIMBUS Preservation Process are presented in this paper, a detailed specification and description can be found in [16]. The process is domain independent and can be applied to different settings. In this paper we shown the application of the TIMBUS Preservation Process on the Music Classification process, which was introduced in Section 3.

4.1 Planning phase

The planning phase is responsible to capture the process and its context and the assessment of suitable preservation approaches. As shown in Figure 3, the first step is the acquisition of the process context, followed by the risk assessment process. The risk assessment triggers the assessment of preservation approaches sub-process for identification, specification and evaluation of preservation strategies for the process.

4.1.1 Acquisition of the business process context

To successfully capture and archive the context of a business process, we have devised a context meta-model to systematically capture aspects of a process that are essential for its preservation and verification upon later re-execution [1].. This model is in the form of an OWL ontology, which enables checks for conformance and reasoning.

As the context of a process can involve a huge variety of different concepts, it is important to design a meta-model that is on the one hand generic, and on the other hand extensible to cover very specific aspects. We thus utilise a domain-independent ontology (DIO) that provides the generic core concepts, and domain-specific ontologies (DSOs) that are integrated and mapped to the DIO, and can refine its concepts. As the context of a process includes aspects on various different layers, the DIO is based on existing work in

[3] `http://www.taverna.org.uk`
[4] `http://the.echonest.com`

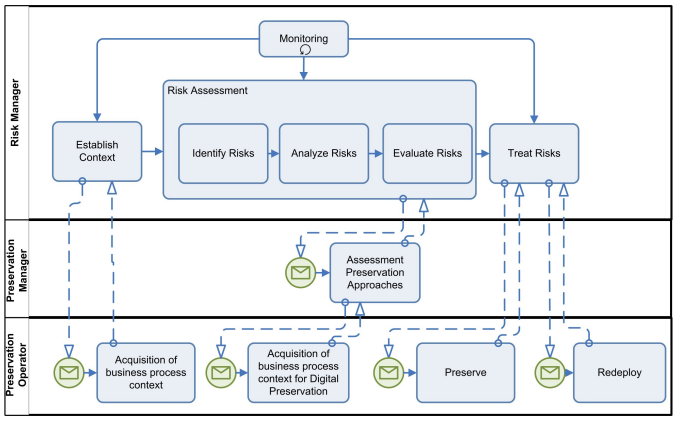

Figure 3: TIMBUS Process

enterprise architecture. Specifically, we adopted the Archimate [15] language, which provides a template to describe a business by around 30 different concepts on the business, application and technology layer.

We then further developed a number of domain specific ontologies that refine these concepts, including:

- Software licenses, based on The Software Ontology[5]
- Patents, based on the Patent Metadata Ontology (PMO), developed by the PATEXpert project[6]
- Software application dependencies, based on the CUDF, the Common Upgradeability Description Format[17]
- Digital preservation meta-data, based on the PREMIS data dictionary [13]

Some elements in these domain-specific ontologies are identified as sub-types of concepts defined in the domain-independent ontologies, and are mapped to these respective elements. This allows for a comprehensive description of the domain-specific aspects, while keeping the core ontology minimal.

The meta-model needs then to be instantiated for a specific use case. The context model can be further extended with other DSOs to define domain specific aspects of the processes. Some parts can be acquired automatically, such as the software dependencies on package-based operating systems such as Debian Linux, which also provides means to identify the licenses a certain package is distributed under. Other elements will have to be provided manually, for which we provide a graphical editor, implemented as a plugin to the Protégé ontology editor[7].

4.1.2 *Risk Management*

Risk management is a well establish field with the goal of defined prevention and control mechanism to address risks

[5] `http://theswo.sourceforge.net`
[6] `http://www.patexpert.org`
[7] `http://protege.stanford.edu`

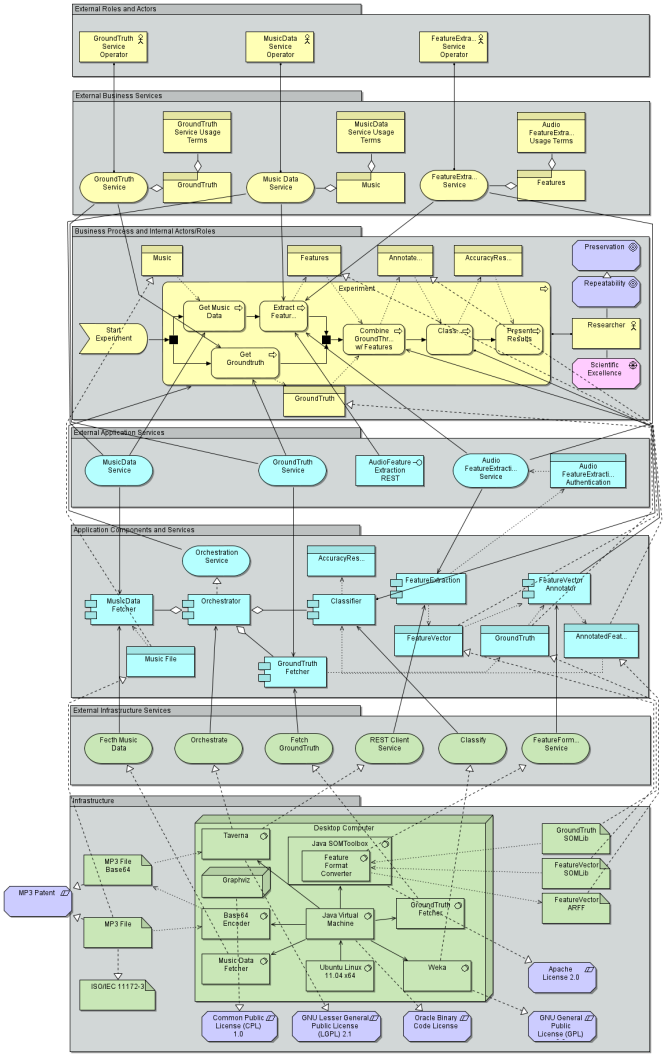

Figure 4: Context Model instance of the Music Classification experiment

related with assets and activities. Preservation can be seen as a potential method to mitigate risks, derived from the potential loss of information over time. The risk management process used in TIMBUS is based on the ISO 31000 standard [7]. Figure 3 shows the risk management process steps in the upper swim lane of the Risk Manager. TIMBUS defines the process-related interfaces to connect digital preservation with risk management. From the TIMBUS perspective, the risk associated to a process can act as a driver prompting its preservation as a way of mitigating the threats that endanger it. Risk management helps to identify and evaluate different risks in a structured and well defined manner. If information related risks have been identified and evaluated, different alternatives to preserve an adequate set of information have to be developed. The risk management process then decides what the best matching solution is. It triggers the preservation process, if a preservation alternative was assessed as a risk treatment.

For the Music Classification process the risk management is motivated by the goal to establish the institution as a sustainable excellence research center. According to the institutional polices, published scientific experiments and results need to be reproducible and verifiable in the future. Another goal of the institution is to foster the reuse and expand previous scientific work. This requires to share and reuse scientific experiments within the institution.

The use of external services represents a risk for the re-execution and verification of the Music Classification process, as the availability of the external services used cannot be guaranteed in the long term. Another risk identified is the lack of the documentation of the executed experiments. The results of experiments are published in conference paper, journals or reports, but these information are not sufficient to re-produce the experiment. Input data set, used software and parameter settings are often not specified in detail or not available any more. Moreover the technological dependencies of the experiment setting and a potential technical obsolesces were identified as potential risks for the Music Classification workflow.

A more detailed analysis of the risk assessment with respect to digital preservation for e-science processes is presented in [4]. After the risks are assessed, mitigation strategies are requested from the *Assessment of Preservation Approaches* process by the risk management as shown in Figure 3.

4.1.3 Assessment of Preservation Approaches

The *Assessment of Preservation Approaches* process is responsible for the identification and evaluation of different preservation approaches for the process. It starts with the refinement of the context model. In the first iteration the context model was created for the use by risk management. More detailed information about the technical implementation of the process is required for the planning of preservation strategies. The preservation requirements of the process are specified and documented for the evaluation and comparison of preservation approaches. The requirements specify the significant properties, describing functional and non-functional requirements of the process that need to be maintained over time. Redeployment scenarios support the specification of the significant properties regarding the preservation of artefact and execution of the process in terms of performance and behaviour. Different redeployment scenarios for future usage can be considered, e.g. execution of the original process with original data for confirmation of documented outcomes, execution of the original process with new data, or to modify parts of the process but using the original data e.g. for scientific workflows to evaluate improvement of new methods or models on the experiment results. For verification of the results from the Music Classification workflow, the original process needs to be re-executed with the original data from the executed experiments. Other preservation requirements can include amongst orders compliances to standards, institutional policies or legal obligations. The requirements are later used to evaluate and compare different preservation approaches.

The context model describes the implementation of the process. The primary goal is the preservation of the business logic of the process, not all implementation aspects are relevant for the future. Thus some implementation details can be abstracted and replaced to higher concepts. The abstraction of technical details can facilitate the preservation of the components (e.g. replacement through alternative implementations, use of emulators, or encapsulation). The level of abstraction and the aspects that can be generalised depend on the specific setting and the preservation requirements. As the abstraction causes loss of information, it is vital to ensure that no relevant information that is required in the future is lost during this step. An example for the music workflow is the operating system that can be generalised. As the workflow runs within the Taverna workflow engine, the underlying operating system does not represent a significant property of the experiment.

A process is an orchestration of tasks that are executed in a particular sequence, it can be complex, involving different services from various systems. For this reason, a combination of different preservation actions can be applied to preserve a process for the long term. Examples are virtualisation and emulation approaches for preserving functionality of services, and migration for documents.

A challenging task for the preservation of complex processes is the preservation of relationships and dependencies between components over time. Knowledge of the dependencies is important for maintaining the functionality of the components. Broken dependencies can prevent the redeployment of the process in the future. Examples are manifold, such as missing libraries for software execution, missing databases for data input, incompatible hardware for operating systems or missing credentials for encrypted data. The dependencies need to be considered whenever changes are applied to components. Modification of components for preservation purposes for example can have undesired side effects on other components. Examples are the migration of data into other formats that cannot be processed further by other software components, or the replacement of software components by new versions that offer different interfaces for interaction. Reasoning and queries based on the context model can help to identify decencies and further try to determine feasible preservation approaches [12].

While strategies for digital preservation, so far, mainly focus on data migration and emulation, the preservation of processes need further approaches, especially with respect to external dependencies. Different strategies can be used to maintain the significant properties of the process over time. Examples of strategies that support preservation and archival storage of processes are:

Metadata/Documentation
In order to maintain the usability, interpretability, accessibility and understandability of the process, additional metadata of its components are required. Understandability involves providing sufficient information so the component can be interpreted and understood in the future. Manual steps of the process that are not implemented by information systems require sufficient documentation and description for later redeployment. Furthermore logging and tracking functionalities of SW components (such as workflow engines) can be used to document the process execution and provide provenance information for the future [9].

Migration
Migration can be seen as the copying or conversion of digital objects from one technology to another. It is a widely adopted strategy for storage media and data formats. Besides that, the migration to alternative software services or components can be a vital approach for processes. For example in terms of licences, the use of alternative open source resources that provide the same functionality can be suitable strategy to overcome legal conflicts. Another aspect of migration can raise from the use of external services. As the availability of external services cannot be assured, a potential strategy is to transfer external services into the own system (in-housing). The strategy requires access to the implementation and data of the service as well as the licences and rights to operate the service. An example is cloud storage services that are operated by third parties.

Emulation
An emulator software mimics the behaviour and functionality of components, hardware or software. Emulation is a widely adopted strategy to preserve older computer platforms (e.g. video game console systems) and operating systems.

Virtualisation
Virtualisation (most common hardware virtualisation) has become a common business practice for server management. Virtualisation software provides a separation layer between the application services and the underlying hardware resources. The separation from actual hardware provides an abstraction layer of the physical environment, such as network, storage and display. It increases the robustness of virtual machines (VM) against changes of the underlying hardware. The virtualisation is a practical approach to capture complex systems to maintain the dependencies within VMs.

Mock-up of SW Services
A special problem for preservation represents the use of third party software services (e.g. Web services) within a process. A potential solution can be a mock-up of the services in form of a simulation of the original service. The basic principle is to intercept and record messages from the original system between process and service, which the simulation can then use to respond to request that have been captured previously. The approach is limited, as it can only be used for deterministic services (i.e. services for which the request and response pair always match, and which themselves are not dependent on any external state), and the mock-up can only respond to messages that have been recorded in the original system. For simple services and for the preservation of particular instances of a process, the mock up can provide a suitable solution if no other possibilities are given. An analysis of mock-up strategies for Web services, and recommendations to make Web services more resilient in general, can be found [10].

Software Escrow
Processes are often using proprietary and customised software application and services. The software is in many cases delivered as closed source to the customer that means the source code remains at the vendor and only the binaries of the software are delivered to the customer. From the preservation perspective, this scenario limits the potential preservation strategies for the software, as the software cannot be adapted to changes in the execution environment in the future. Software Escrow offers a mitigation as it places a trustable third party between the developer and the customer. All artefacts relevant to the software development are deposited at the escrow agent and released to the customer in case of predefined events (e.g. when the vendor goes out of business, or does not want to further maintain the software).

Different approaches can be used to preserve a process, using different strategies or tools. Each approach is specified in a *Process Preservation Plan*. The plan also defines procedures for capturing the process data and later redeploying and verifying the process. In order to preserve the process, the components and process data need to be captured from the source systems. The acquired data need to be in a consistent state that redeployment leads to a valid state of the process (e.g. all database transaction are closed). The redeployment procedure defines the execution of the preserved process in a new environment in the future. In order to ensure that the process is redeployed correctly, a verification and validation procedure is required. It defines measurement points to check that the redeployed process shows the same significant properties as the original process.

The proposed planes are evaluated against the previous specified preservation requirements. The evaluation includes the assessment whether the proposed models and procedures are complete and correct and that all significant properties of the process are preserved. In case the evaluation shows that relevant aspects of the process are missing or requirements are not fulfilled, a feedback loop of the *Assessment of Preservation Approaches* process allows the refinement of the context model or the preservation plan specification. Different preservation plans can be evaluated, and the evaluation results are submitted to the risk management for decision making. The impact of the different strategies on identified risks is assessed and the best matching solution is selected for treatment.

For the music workflow a combination of strategies was identified as most suitable preservation approach. The client side including the workflow, the workflow engine and the classification engine is captured in a virtual machine using Virtual Box[8]. As a underlying operating system Linux is used, because the licensing and activation methods of current Windows release can cause interferences in the future. The Music Classification workflow uses three external services, the music data, ground-truth and the feature extractor service. As the music and ground-truth data are free available, we can deploy the service on the client side. For the feature extractor we need another approach as we have no access to the implementation. In order to verify the experiments in the future, a mock up of the feature extraction service can be used to capture and replay the communication to the Web service of the process execution. Publications and documentation are migrated in standardized formats, PDF and Word documents are migrated to PDF/A by using Adobe Acrobe Distiller. The available software documentation in

[8] `https://www.virtualbox.org`

HTML format remains in the format.

4.2 Preservation phase

The acquisition and preservation procedures of the *Process Preservation Plan* are applied to the business process in the preservation phase. The software and data of the process are captured from the source environment. Preservation actions are executed and the process is prepared for archival storage. Validation and verification data are captured from the source system for redeployment. For the Music Classification process a sample input set and expected output can be used to validate the redeployment. Other measurement points can be logging information created from the workflow engine during the process execution that can be used for verification in the future.

For the preservation, a empty VM image is created and Ubuntu 13.04[9] is installed as operating system. The Taverna workflow engine is set up for the Music Classification workflow. The data source and ground-truth are migrated to local services by using Apache HTTP Server[10]. A mock-up software is installed to capture the traffic between the workflow engine and the external Web service of the feature extraction. The scientific experiments are executed in order to capture the responses from the Web services for the music files used. The implementation of the feature extractor cannot be preserved, but the behaviour of the service is documented through capturing of the traffic for later verification. A replay software that mimics the web services including the captured data set is installed on the VM system. Documentation and publications of the process and its component are stored within the VM. Viewer applications for the documents are installed as well. This strategies allows to bundle all required software and information in a single VM image. It reduces the technical dependencies for a re-execution of the process to a compatible VM player. The correct execution of the preservation phase is verified by test wise re-instantiation of the VM and the execution of the process. The validation and verification procedure is applied and the results can be compared to the original process. The preserved process needs to implement all signification properties of the original process as defined by the preservation requirements. In a last step of the preservation phase the process data are stored in the archive. In order to ensure that the process can be executed in the future, monitoring criteria have to be defined. The monitoring includes the external dependencies of the preserved process, e.g. technical requirements to redeploy. But also the requirements and polices of the organisation for the preservation of the process needs to be observed.

4.3 Redeployment phase

The redeployment phase defines the reactivation of a preserved process in a new environment at some point in time. The key characteristics of new environment are captured including the available technical components, organisational and legal aspects. A gap analysis between the requirements of the preserved process for redeployment and the available environment is performed. The technical infrastructure

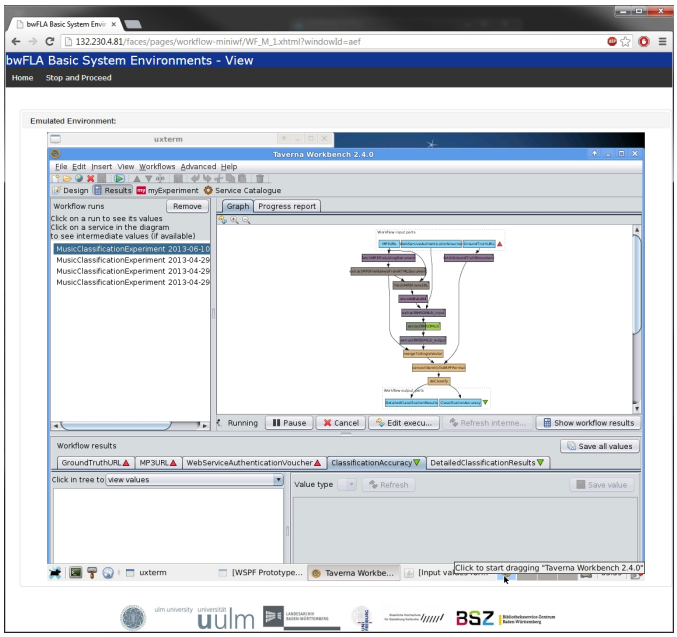

Figure 5: Emulation-as-a-Service of the Music Classification experiment

needs to be adjusted and prepared for the redeployment, different approaches can be used to overcome identified gaps, e.g. tools to emulate components, or migration of data formats. Required software and data are installed according to the redeployment procedure defined in the *Process Preservation Plan*. Tools and components for validation and verification are also set up in the new environment. As a final step the process can be re-executed and the taken measurements can be validated.

The archived Music Classification workflow requires a compatible VM player for the redeployment. While currently enough players for the format are available, the format needs constant monitoring to ensure the availability in the future. The use of external services for the redeployment can help to reduce investment and effort for the hosting institution. Emulation-as-a-Service can provide an interface to render preserved virtualised system. An example is the bwFLA project[11] that provides a web-based access to different rendering environment by using an emulation approach [14]. Figure 5 shows the Music Classification workflow that can be accessed via a Web browser interface and re-executed the captured process instances of the experiment. Captured validation and verification data can be used to check the correct behaviour of the redeployed process. Published results of the classifier can easily be verified while the reuse of the workflow is limited due to mock-up of the feature extraction. The mock-up service can only provide feature sets for music files that have been captured and preserved in the preservation phase. But experiments with new or modified classifiers can be done with the same music data set in the future, for example determining performance improvements or new classification approaches.

[9] `http://www.ubuntu.com/`
[10] `http://httpd.apache.org`
[11] `http://bw-fla.uni-freiburg.de`

5. CONCLUSION

This paper presents the TIMBUS Preservation Process to preserve processes for the long term. The process provides a guideline for the required steps to plan and perform the preservation and the later redeployment of processes. Driven from a risk management perspective, Digital Preservation is considered as mitigation strategy to address potential loss of information over time.

To preserve a process for the future, its influence factors and implementation need to be understood. Hence the context of the process is acquired, relevant aspects are identified that need to be maintained for the future. Potential preservation strategies are identified and tested considering the specific conditions, requirements, and goals of a setting. Different approaches can be combined to maintain the process over time. Processes are often implemented by using a service-oriented architecture implemented on a distributed infrastructure. The paper presents preservation approaches that address the specific needs of process preservation such as the preservation of external services. The redeployment reruns the archived process in a new environment at some time in the future. The process needs to be adapted according to the conditions of the new environment. The behaviour of the redeployed process is verified by comparing its measurements with measurements taken from the original process.

The application of the framework was presented in this paper by the preservation of a scientific workflow for music classification. The Taverna workflow engine was used to design and execute the Music Classification process. The process uses three external services that provide music files, ground-truth data and a feature extraction service. For the preservation of the process two of them were migrated to local alternative services implementing the same functionality. For the feature extraction a mock-up services was used to record the communication between the process and the Web service. The records captured can be used to mock up the service in the future and replay the communication. All software components used by the process were set up on a virtual machine. Documentation and publications about the scientific workflow were also stored on the VM. The result is an encapsulated process in a VM that can be archived. A potential redeployment strategy represents Emulation-as-a-Service where emulators for rendering environments are provided as Web service.

Acknowledgments

This work has been co-funded by COMET K1, FFG - Austrian Research Promotion Agency and by the TIMBUS project, co-funded by the European Union under the 7th Framework Programme (FP7/2007-2013) under grant agreement no. 269940. The authors are solely responsible for the content of this paper.

6. REFERENCES

[1] G. Antunes, A. Caetano, M. Bakhshandeh, R. Mayer, and J. Borbinha. Using ontologies for enterprise architecture model alignment. In *Proceedings of the 4th Workshop on Business and IT Alignment (BITA 2013)*, Poznan, Poland, June 19 2013.

[2] C. Becker, G. Antunes, J. Barateiro, and R. Vieira. A capability model for digital preservation - analyzing concerns, drivers, constraints, capabilities and maturities. In *Proceedings of the 8th Int. Conf. on Preservation of Digital Objects (iPRES 2011)*, 2011.

[3] H. Brocks, A. Kranstedt, G. Jaschke, and M. Hemmje. *Smart Information and Knowledge Management*, chapter Modeling Context for Digital Preservation, pages 197–226. Springer Berlin/Heidelberg, 2010.

[4] S. Canteiro and J. Barateiro. Risk assessment in digital preservation of e-science data and processes. In *Proceedings of the 8th International Conference on Preservation of Digital Objects (iPRES 2011)*, 2011.

[5] E. Conway, B. Matthews, D. Giaretta, S. Lambert, and M. Wilson. Managing risks in the preservation of research data with preservation network. *The International Journal of Digital Curation*, 7:3–15, 2012.

[6] S. Higgins. The DCC curation lifecycle model. *The International Journal of Digital Curation*, 3:134–140, 2008.

[7] ISO. *ISO 31000: 2009 Risk management – Principles and Guidelines.*

[8] ISO. *Space data and information transfer systems – Open archival information system – Reference model (ISO 14721:2003)*, 2003.

[9] R. Mayer, S. Proell, and A. Rauber. On the applicability of workflow management systems for the preservation of business processes. In *Proceedings of the 9th Int. Conf. on Digital Preservation (iPres 2012)*, pages 58–65, Toronto, Canada, October 1-5 2012.

[10] T. Miksa, R. Mayer, and A. Rauber. Ensuring sustainability of web services dependent processes. *International Journal of Computational Science and Engineering (IJCSE)*, 2013. Accepted for publication.

[11] Music Information Retrieval Evaluation eXchange (MIREX). Website. `http://www.music-ir.org/mirex`.

[12] M. A. Neumann, H. Miri, J. Thomson, G. Antunes, R. Mayer, and M. Beigl. Towards a decision support architecture for digital preservation of business processes. In *Proceedings of the 9th Int. Conf. on Digital Preservation (iPres 2012)*, 2012.

[13] PREMIS Editorial Committee. Premis data dictionary for preservation metadata. Technical report, March 2008.

[14] K. Rechert, D. von Suchodoletz, and I. Valizada. bwFLA – practical approach to functional access strategies. In *Proceedings of the 9th Int. Conf. on Preservation of Digital Objects (iPRES2012)*, 2012.

[15] The Open Group. *ArchiMate 2.0 Specification.* 2012.

[16] TIMBUS consortium. D4.6: Use Case Specific DP & Holistic Escrow. Technical report, 2013.

[17] R. Treinen and S. Zacchiroli. Description of the CUDF Format. Technical report, 2008. http://arxiv.org/abs/0811.3621.

[18] C. Webb. *Guidelines for the Preservation of Digital Heritage.* National Library of Australia, 2005.