On the Complexity of Process Preservation: A Case Study on an E-Science Experiment

Rudolf Mayer
Secure Business Austria
Vienna, Austria
rmayer@sba-research.at

Stephan Strodl
Secure Business Austria
Vienna, Austria
sstrodl@sba-research.at

Andreas Rauber
Secure Business Austria
Vienna, Austria
arauber@sba-research.at

ABSTRACT

Digital preservation of (business) processes is an emerging topic in Digital Preservation research. Information technology driven processes are complex digital objects, living in an broad context of aspects relevant to their preservation. In this poster, we detail the broad environment of one sample process from the domain of E-Science, a genre classification experiment in the domain of Music Information Retrieval. We show the magnitude of aspects involved, on technology as well as organisational, legal and other aspects.

General Terms

Process Preservation, Case Study, E-Science

1. INTRODUCTION

Preservation of information technology driven business and scientific processes is an emerging topic in Digital preservation research. These processes are complex digital objects, themselves including and using many other digital objects along the process execution. In this poster, we want to demonstrate on how complex the context of an even rather simple scientific workflow with a limited number of processing steps may become. We show tool support for defining and visualising this context.

2. MUSIC CLASSIFICATION PROCESS

The specific process used in our case study is a scientific experiment in the domain of Music Information Retrieval, where the researcher performs 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 [1].

The experiment involves several steps; a model of the process in BPMN 2.0, is depicted in Figure 1. First, music data is acquired from sources such as benchmark repositories or, in more complex settings, online content providers. In parallel, genre assignments for the pieces of music are obtained from ground truth registries, frequently from websites such as Musicbrainz.org. Tools are employed to extract numerical features describing certain characteristics of the audio files. In the case of the experimental set-up used in this example E-Science process, we assume a more complex scenario where an external web service is used 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 predict genre labels for unknown music. Further, several scripts are used to convert data formats and other similar tasks. 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, with several dependencies between the various components.

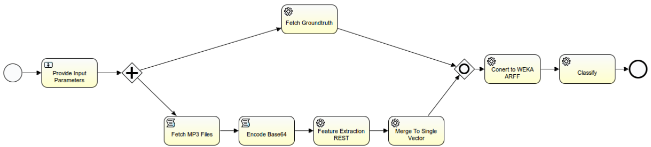

Figure 1: The Music Process Workflow (BMPN 2.0)

Figure 2 gives an overview on the elements identified as relevant aspects of the business process context, and their relations to each other; we will describe some of these elements below. As the scientific experiment is a process mostly focusing on data processing, a significant amount of the identified aspects are in the technical domain – software components directly used in the processing steps (and their dependencies), external systems such as the web service to extract the numerical audio features from, or data exchanged and their format and specification. However, also *goals* and *motivations* are important aspects, as they might heavily influence the process. As such, the motivation for the providers of the external systems is relevant, as it might determine the future availability of these services. Commercial systems might be more likely to sustain than services operated by a single person for free. Another important aspect in this process are *licenses* – depending on which license terms the components of our process are released under, different options of preservation actions might be available or not. For closed-source, proprietary software, migration to a new execution platform might be prohibited.

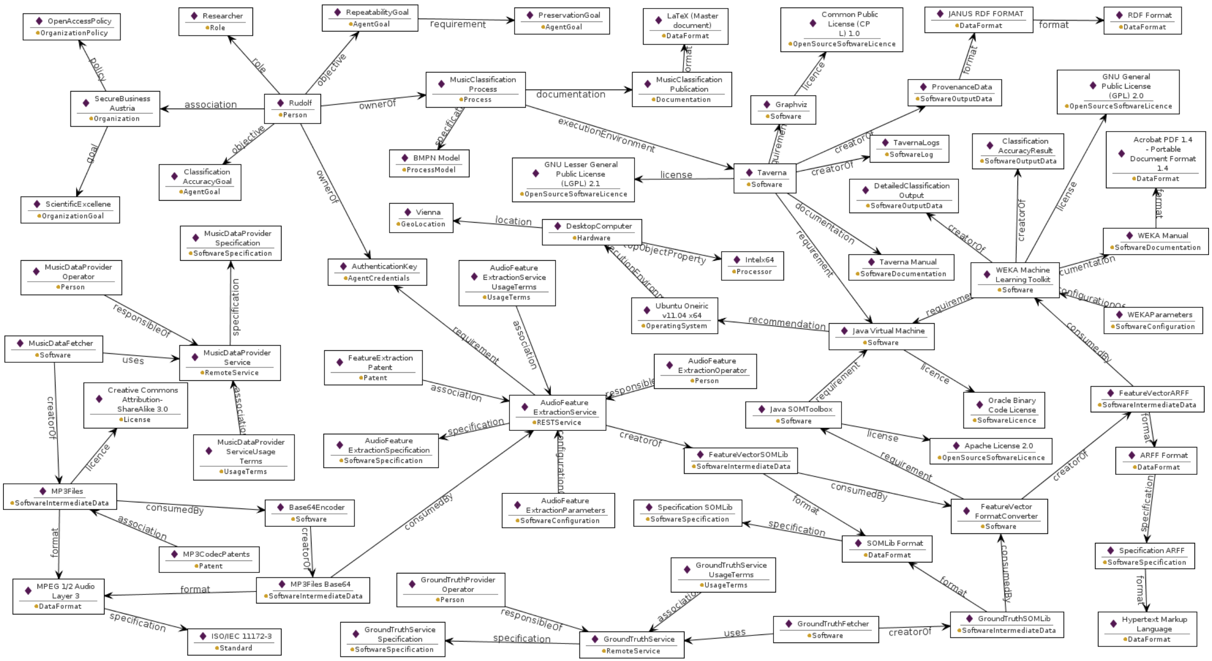

Figure 2: Relevant aspects identified in the scientific workflow

A central aspect in the scientific process is the *AudioFeatureExtractionService*, i.e. the remote web-service that provides the numeric representation for audio files. The service needs as input files encoded in the *MP3 format* (specified by the *ISO standard 11172-3*). More specifically, as they are binary files, they need to be further encoded with *Base64*, to allow for a data exchange over the HTTP protocol. The web-service accepts a number of parameters that control the exact information captured in the numeric representation; they are specified in the *AudioFeatureExtractionSpecification*, which is authored as a PDF document. The specification further provides information on how the extraction works (i.e. a coarse *documentation* of the signal processing steps applied to obtain the final result). The operator of the web-service provides the service for free, but requires *authorization via a key* that, which is granted to a *person*, and can't be shared under the *usage terms*. The service returns the numeric description as ASCII file, following the *SOMLib format specification*, which is authored in *HTML*.

As a software component used locally, the *WEKA* machine learning toolkit requires a *Java Virtual Machine* (JVM) platform to execute. The JVM in turn is available for many operating systems, but has been specifically tested on a Linux distribution, *Ubuntu, version "Oneiric" 11.04*. WEKA requires as input a feature vector in the *ARFF Format*, and a set of *parameters* controlling the learning algorithm. These parameters are specified in the WEKA manual, available in *PDF Format*. As output result, the numeric performance metric "accuracy" is provided, as well as a textual, detailed description of the result. WEKA is distributed under the terms of the open-source GNU Public License (GPL) 2.0, which allows for source code modifications.

After this experimentation process, a subsequent process of result analysis and distillation is normally performed, taking input from the experiment outcomes, and finally leading to a publication of the research in the form of e.g. a conference or journal *publication*. Here it is modelled as a single information object (the paper written in *LaTeX*) connected to the process, and thus to all data and processing steps that led to the results published. It might also be modelled as a process in its own, specifically if a paper reports on meta-studies across several experiment runs.

Tool support for automatically extracting, manually creating and viewing such process context has been implemented, and will be demonstrated.

3. ACKNOWLEDGMENTS

Part of this work was supported by the FP7 project TIMBUS, partially funded by the EU under the FP7 contract 269940.

4. REFERENCES

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