Towards Practical Emulation Tools and Strategies – State of the Art Research Meets Real-World Requirements

Mark Guttenbrunner
Secure Business Austria
Vienna, Austria
mguttenbrunner@sba-research.org

Dirk von Suchodoletz
University of Freiburg
Freiburg i. B., Germany
dsuchod@rz.uni-freiburg.de

Klaus Rechert
University of Freiburg
Freiburg i. B., Germany
klaus.rechert@rz.uni-freiburg.de

ABSTRACT

The workshop features an increasingly relevant topic in DP research – emulation. Memory institutions receive new types of digital objects or getting involved in web archiving. Additionally, business face new challenges to keep their processes accessible and science funding agencies requires reproducible results of experiments and publications. While emulation is now widely accepted as a necessary access strategy, mature software frameworks and workflows are still missing. The workshop brings together practitioners and researchers and is intended to provide an additional forum for experts in emulation alongside the iPres 2012 main conference to discuss current trends, today's strategies and formulate future implementation and research agendas. Content holders will be given a platform to describe their challenges with complex data-objects to align research with practical problems.

1. INTRODUCTION

Through significant research projects on digital preservation and access in the last couple of years emulation has become more and more accepted and adapted. Even non-scientific press recognizes the necessity of emulation and associated workflows to preserve and provide a holistic picture of today's emerging digital culture.[1] Emulation as a preservation strategy for memory institutions has been researched since Jeff Rothenberg's article in Scientific American Magazine (1995). The strategy relies on emulating hard- and/or software that has become obsolete. Rendering a digital work by emulating the original platform for which it was originally created has the preservation benefit of recreating the (near)original look and feel of the work. The better the original rendering platform is emulated the higher the fidelity of the rendering. It has been argued that emulation is specifically suited for dealing with the preservation of complex objects (e.g. works of art) and executable works (e.g. software games). The specification of such works and of their behavior during execution is more complex than that of simple objects, such as a PDF-document for example, which requires relatively simple format and reader specifications. The specification of complex objects and executable works often requires going back to the source code of the work (if it is software) and the application layers on which it runs, in order to recreate it perfectly. As has been discussed extensively in the literature, emulating all the layers of software and hardware on which a work runs is not the most effective and economic strategy. A more effective strategy might be to emulate only the lower layers (hardware and operating systems software) and to preserve the upper layers (specific software applications). Emulators bridge the widening gap between the digital past and todays working environments and thus help the process of becoming independent of further technological developments.

The progress of emulation as a digital preservation strategy is not only driven by large-scale research projects like PLANETS, KEEP, TIMBUS and others, but has independent strong roots in other communities as well. Hobbyists run instances of their old computer systems, electronic media studios sell old versions of electronic games for new platforms, emulation is deployed in the sub-field of virtualization and used for future system software development. Many memory institutions begin to see the proven usefulness of emulation for complex, dynamic, interactive objects and complete original environments. Emulation is an extremely versatile and durable solution for retaining access to any kind of digital content without the need to alter the original and thus threaten the object's authenticity.

While a couple of significant building blocks of emulation based strategies are present, a number of components are still unsatisfactory or missing. The proposed workshop intends to bring together the major actors in the domain of emulation but also practitioners from memory institutions to present the results of projects and initiatives. Furthermore, the workshop provides practitioners a platform to identify missing pieces for productive integration and deployment. Together, both communities will lay out requirements to close the gaps in the existing preservation processes and associated workflows. The goal of this workshop is to push emulation out of it's status as a niche strategy handled only by a few trained experts to the wider community of memory institutions facing the new digital world.

[1] History flushed, Economist Apr 28 2012, `http://www.economist.com/node/21553410`

2. USE CASES AND CHALLENGES FACED BY CONTENT HOLDERS

The following parts were compiled with the help of the workshop's call-for-contribution.

The electronic collections of today's libraries, museums and archives are growing and increasingly have a more relevant role in the holdings. Memory institutions must address users' need to access a widening range of digital artifacts. Often the formats of those artifacts are outdated and they cannot be run or rendered on today's systems any longer. A substantial problem for memory institutions is to provide a wide range of different users with access to ancient environments and to allow using the original environment for a given object. Ways are needed to be provided to allow users to experience original digital material in e.g. emulation. Such functionality can be integrated in a framework of web services for common preservation tasks like viewing or migrating digital objects. A further major challenge arises from the preservation of access to research data. It has always been fundamental to a wide range of domains in science but has grown substantially in complexity and scale in the last couple of years. More and more primary data is generated solely in the digital domain and gets processed in inter-disciplinary projects and complex workflows. The associated costs of archiving activities are consistently rather small in proportion to the initial costs to acquire or generate them. Additionally, with the initiatives of research funding agencies and public bodies, the data is required to remain available of certain periods of time to fulfill the requirements of reproducible science and to follow open data strategies. Especially in the long-run the data can not be processed and interpreted in a meaningful way, if the original set of tools and workflows is not available any more. Proper archival workflows are required to ensure and proof the preservation of all necessary components over the relevant periods. Permanent storage and access is technically feasible but challenging. Planning and continuous curation are necessary to secure long-term access. To face these challenges a holistic approach which considers the complete digital lifecyle is necessary.

2.1 National Library of the Netherlands[2]

The National Library of the Netherlands (NLN) was founded in 1798 and is the national deposit library of the Netherlands since 1974. The library located in Den Haag employes 275 FTE and is financed by Ministry of Education, Culture and Science. The NLN has 3.5 million books in its holdings, 110 km of publications stored on shelves and 17 million digital scientific publications. In total the library storage contains more than 400 TByte of digital information. The library offers everyone everywhere access to everything published in and about the Netherlands. It plays a central role in the (scientific) information infrastructure and promotes permanent access to digital information nationally and internationally.

The library started by the end of last century the research of digital preservation. It implemented a market watch to observe new trends in IT and digital materials, it promotes standards and guidelines for government agencies. Additionally, the NLN keeps track of dependencies and characteristics of digital information. As object access strategies Migration and Emulation are considered. The NLA has a long history of research and development in the domain of emulation. Since 1999 it started with first emulation tests together with Jeff Rothenberg of Rand Corporation. In the years 2002 – 2004 it investigated into the Universal Virtual Computer with IBM. In the following years (2005 – 2007) the library started together with the Nationaal Archief the Dioscuri project. Together with the university of Freiburg the NLN researched emulation within PLANETS (2007 – 2010). The latest involvement in emulation was the KEEP project running from 2009 till 2012.

2.2 German National Library and Bavarian State Library

Multimedia objects play an increasing role for scientific research and literature and take over the functions of traditional media. Libraries providing the working material for research, have to be able to provide access to these objects beside the traditional material like books and journals. The preservation of complex digital objects such as multimedia CDs and interactive digital documents is a new challenge, which makes it necessary to develop new methods in order to provide access to these items as part of the scientific information and library services in the long run. To keep them available to the wider public, emulation of their original environments might provide an efficient way of access and therefore the preservation of information. However, this requires a number of organizational and technical solutions that should fit into the existing IT infrastructure and workflows.

Both, the German National Library and the Bavarian State Library plan to investigate into emulation enabled reading room systems for their reading rooms. Both libraries observe an emerging major gap in scientific information and literature provisioning: Scientists as well as users of today's memory organizations are often unable to access complex, characterized by multiple technical dependencies objects such as multimedia CDs, interactive digital documents, databases, digital artwork, computer games, geo-information or scientific simulations. To ensure the long term availability to such objects a project is planned to develop systems and workflows to access a wider range of digital objects in library holdings.

The usability of many objects can only be realized with great effort, or not at all, by migration. Based on concepts and components (frameworks and individual emulators) from different research projects and own developments a reading room system to planned, which executes a wide range of different original environments for end users access. Previous approaches from projects like PLANETS, KEEP or bwFLA will be evaluated as well as the currently available emulators and virtual machines.

2.3 Royal Library of Denmark[3]

The Royal Library of Denmark has a double function as the largest university library in Denmark and as the Danish

[2]Contributed by Jeffrey van der Hoeven, National Library of the Netherlands

[3]Contributed by Eld Zierau

national library which holds and maintains the Danish national collections of published material. The collections are primarily based on legal deposit, a practice that has existed in Denmark since 1697. In 1998 the Danish Legal Deposit Act was changed to include digital materials, including computer games and interactive software on physical media. The library has a computer game collection of about 1.500 games for a variety of platforms on physical media, not including online games. The library is presently engaged in a process of migrating computer game data from original data media to more sustainable long-term preservation platforms. Additionally the national collection of periodicals contains a number of cover discs from various video game magazines. Since 1998 the library has collected online publications, and from 2005 onwards we have been engaged in full-scale web harvesting of Danish web sites to the national web archive, *netarchive.dk*, in collaboration with The State and University Library in Aarhus. The web archive presently comprises some 245 TByte. In 2011 the library began a systematical process of identifying harvested computer games in the web archive. An effort has been made to ensure that as much game data as possible is harvested, however this process is not always adequate, thus an effort is also made to identify and harvest material and web sites related to the games, i.e. community forums, videos, online reviews, blogs, etc. The game data harvested comprises the most common file types for this kind of material: Flash, Java, and – more recently – Unity3d. With the exception of Unity3d these materials are heavily represented in the web archive, not only in relation to games, but as part of the web archive in general. Presently the Library is focusing on extending the computer game collection in two directions: First through the development of coordinated video game documentation practices in order to ensure the relevance of our collection of games related materials. Secondly we are working on clarifying the technical issues involved with the collection and preservation of apps for portable media, primarily iOS and Android.

2.4 Austrian National Library[4]

The Austrian National Library is the biggest library in Austria and has the legal obligation to archive Austrian web domains as well as all official publications in Austria. Both for web-archives and document repositories it becomes increasingly clear, that migration is a rather complex task. One of the major problems when migrating images is that the evaluation of the results is very complex. If a migration from TIFF to JPEG2000 is done to save storage space, the original files have to be deleted afterwards. But this means that the results of the migration have to be 100 percent reliable. One problem is that the metadata in scanned documents are not reliably and correctly migrated to the new format, as different scanner manufacturers are storing the metadata differently in the source format. Another important problem is the use of different color profiles by different manufacturers, which are not always properly migrated. One concrete example are LUT-based color profiles that are not supported in JPEG2000. In the context of web archiving migration is not a topic yet, as currently the identification of web content is the more pressing issue. Digital preservation actions can only be taken once the content is reliably identified. Migration would in this case also pose the problem of evaluation of the migration results. The main advantage of emulation in this case is that the original formats and thus the original content is not destroyed.

[4]Contributed by Sven Schlarb

2.5 Austrian State Archive[5]

The Austrian State Archives have the mandate to archive records created in governmental departments. This includes not only data stored in PDF, image formats or other rather static formats, but also data stored in and created by online applications such as the tax reporting system. Not only public online applications, but also programs created in the form of user made MS Access applications, complex Excel sheets and other so called *End User Programming* software poses a threat to data that is not easily migrate-able. An additional problem is that not all of the software is developed in-house. While the source code is available for such software, this is not always the case for individual software that has been ordered from external companies or that has been developed by end users.

2.6 National Library of Australia[6]

The National Library of Australia (NLA) is currently in the process of purchasing a vendor product and building additional preservation functionality. This system will do the usual format identification and characterization of our collection. NLA is also working on preservation knowledgebases about formats,[7] software/dependencies to give a preservation level of support. Until this is done, there could only be a look at a collection in a fairly 'handraulic' way. Eventually, NLA will make decisions based on a preservation intent classification, which may lead to assess the following action options:

1. Take no action; or
2. Replace the failed access software; or
3. Migrate the content to another format that does preserve the presentation of the properties we decide are significant; or
4. Work out some way to emulate the presentation of those significant properties of the content; or
5. Maintain the bits and documentation that will support action in the future when more effective action options may become available [8].

NLA may or may not use emulation. This depends on the preservation intent from each collection. However, for web archives this might be the only sane way forward. The IIPC PWG has been working on a project which records a generic access environments for web archives for each year from (1996–2011). This could probably be used for emulation purposes.

[5]Contributed by Hannes Kulovits

[6]Provided by David Pearson of Digital Preservation team at NLA

[7]See e.g. on identification issues, `http://www.openplanetsfoundation.org/blogs/2012-08-12-file-characterisation-tools-report-testing-project-conducted-national-library`

2.7 Karlsruhe University of Arts and Design[8]

The following section describes the challenges when preserving digital art by the example of "CD-ROM Art". Transmediale is an annual Digital Art festival in Berlin with a history of 25 years. Its collection contains countless pieces, stored on all kinds of media, many of them becoming obsolete. During the period between roughly 1995 and 2002, many artists created art on CD-ROM, so called "CD-ROM art". The pieces were intended to be shown on exhibition computers or to be distributed on CD-ROMs to be consumed and experienced on a private computer. It was a technically surprisingly standardized form: almost everybody employed the software Macromedia Director and connected further packages from Macromedia (SoundEdit, Freehand), Adobe (Photoshop) and Apple (Quicktime). Most CD-ROMs were created for Apple's line of operating systems. The interactive parts always used standard input devices (mouse, keyboard), many pieces offered no interactivity at all. Apple is known for cutting compatibility with old software quite radically after only a short transitional period. On today's systems, the classic CD-ROM art does not perform anymore.

2.7.1 Strategy

Since keeping a huge collection of computer hardware, operating systems and additional software to run any kind of artistic software is potentially a task without borders, the chosen strategy was to set up an emulation system that would allow to run the CDs on simulated computers that are easily to be ported to contemporary and future systems. The emulation environment must be available under a Free Software license to make platform changes sustainable. Since CD-ROM art was a very homogenous genre, this approach, once established, would allow for hundreds of objects to be conserved. First, the CDs need to be put into their native environment to establish and document how they are supposed to look, sound and behave in the emulated environment. Two historic "iMac" computers and a set of operating system install CDs were collected from different institutions dealing with digital art to re-create a typical setup.

Official documentation available online from Apple allowed to quickly narrow down which version of the MacOS (the Apple operating system) were released at what time and would likely be able to run on the hardware. Drawing from many years of experience with digital art it can be stated that lots of artists referenced the contemporary digital work environment in their digital work, for instance by picking up graphical elements from the standard user interface, modifying standard icons or playing with users' expectations about how some standard-looking GUI elements should behave. Therefore it was clear that it is not enough to "make the CD-ROMs work again" but to pay attention to exactly how to make them work again. So several versions of the operating system were installed on the available hardware and their apparent visual and behavioral differences documented.

2.7.2 Evaluation of Rendering Results

Examining the data on the CD-ROMs using the historic Macintosh hardware and by mounting the ISO files with a Linux operating system revealed the production dates of each CD-ROM. Most media contain a text file with some sort of technical requirements, but naturally they were very vague ("Need G3 or G4 PowerMac") or just plain wrong ("QuickTime 4 Pro required"). Apparently most artists just wrote down what they thought is their own working configuration. But typically these "READ ME" files contain valuable instructions on what screen resolution is expected by the software, a note if there is sound and sometimes an artist statement that can help to understand what the artistic focus of a piece might have been. The result is that most "Director Projectors" (the runnable software binaries that Macromedia Director used to generate as an end-product) preferred a very low screen resolution of 800 by 600 pixels and in some cases even 640 by 480 pixels. While typical screen resolutions around the turn of the century would have been 800 by 600, Director's poor performance didn't allow smooth full screen animation in that resolution with 24 bit color depth, so some software requires to set up 640 by 480 pixels in greyscale (8 bit greys). On a typical CRT, low resolutions might look sharper and more defined than hight resolutions, because the holes in the screens aperture mask might better line up with single pixels or because single pixels might even stretch over several holes in the mask. In contrast, resolutions lower than the native one, on contemporary LCD screens *look more blurry*, because the image data is digitally upscaled to fill all available pixels. The upscaling usually happens using a bicubic interpolation, generating gradients in between pixels that otherwise would be sharply divided. Testing the CD-ROMs and different operating systems was summarized into the following requirements for an ideal emulation environment:

- The system must be able to emulate a range of Apple Macintosh operating systems from 7.5 to 10.3, covering the years 1996 to 2004.
- The screen refresh rate should ideally be 60 Hz (pictures per second) – already the mouse cursor's movement is uncomfortably lagging if the refresh rate is less.
- Stereo Sound, in sync with the graphics.
- Meaningful way to display the emulator's arbitrarily sized output full screen without additional blurring and without artificial sharping, for example by using an aperture mask simulation.

Practically, in order to reduce the effort of keeping up dozens of operating system versions and setups, supporting System 7.5 and MacOS 9.1 seems to be sufficient, as the line from 8.0 on offers skinning options via the "Appearance Manager" that are going deep enough to cater for most CD-ROMs' configuration needs. MacOS 10.3, released 2003, is supporting both PPC and Intel processor software plus offers a MacOS "classic" environment to run software for up until MacOS 9.2. CD-ROM pieces sharing certain configuration requirements could even be fitted into one virtual instance together. If the emulator runs on the local machine or on a local network and is providing access through a RDP session, achieving 60 FPS is possible. Practically, non-interactive CD-ROMs some of the time do not require a higher refresh rate than around 24 FPS, because for example Quicktime videos would use this

[8] Contributed by Dragan Espenschied, Karlsruhe University of Arts and Design (HfG), Germany

framerate and computer-generated graphics and animations would not run faster anyway.

In almost the same manner, synchronized image and sound is currently possible if the emulator runs on the local machine, any kind of remote access is still lacking synchronicity or sound in general. This would still allow simple pieces without sound to be shown with a remote emulator. At the moment, only a few specialized emulators for arcade games and MS-DOS feature aperture mask or scanline simulation when upscaling an image. This is not a show-stopper per se, but the problem has to be addressed, ideally by a meta emulation layer. Currently, it seems wise to offer several versions for upscaling (pixel-doubling without interpolation, bicubic interpolation or no upscaling at all), so that it is at least possible to examine a pixel-perfect image in the emulation environment.

Many of these issues can be resolved "automatically" with faster CPUs and faster networks, others require dedicated development. Currently, emulation and virtualization is focused on business cases, on making features of obsolete systems that are generally perceived as desirable – utility software and data sets "trapped in an outdated environment" –, available on contemporary systems. However, emulation (and to a lesser extend virtualization) will have to start caring about features that seem impractical, undesirable or generally be better executed on contemporary systems: low-color and low-resolution displays, slow media access, jerky video playback and low-fidelity sound. As long as these features' availability cannot be guaranteed, it is important to document differences between the original and the emulation. It is to be expected that the conserved versions will become the "originals" very soon, because it will be unfeasible or impossible to run the pieces on decaying or vanished original hardware. We need to start emulating now, before the perfect emulation environment exists, but we also need to keep the original data save and unchanged so that future improvements of the emulation environment can be of benefit to already archived pieces.

2.8 Preservation of Scientific Workflows and Environments[9]

Most of today's scientific workflows are based solely on digital data. Input, output and intermediate results are pure digital objects. Keeping scientific results reproducible, accessible and meaningful is challenging, since w.r.t. digital objects a period of 10 years is a huge timespan. Emulation as a preservation strategy is able to provide a valuable service: by using today's emulators an authentic reproduction of current digital workflows can be verified and long-term access is guaranteed for a defined feature-set. In order to improve the projects understanding of potential users' needs, their workflows and goals a survey has been conducted.

2.8.1 Methodology

The first part dealt with characterization of scientific background and user expectations and potential future-users w.r.t. preservation processes. The questions were grouped in the six following categories:

- *Object of research:* abstract description of the data and the (expected) results of current research activities.
- *Aims and objectives of the preservation process:* specifically asks which objects, results, concepts should be kept available and for which period of time (necessary, relevant, desirable).
- *Re-use of data:* what kind of access is desired, which material available in what kind of form, data still useful without its software environment, re-use of data targeted at what kind of audience?
- *Data protection:* Is the data subject to data protection laws or non disclosure agreement?
- *Formal requirements:* Long-term archival requirement posed by external or internal requirements (funding organizations that support the research)?
- *Existing archiving processes:* Are already preservation processes deployed and used?

The second part of the survey was focused on details of the scientific process and work environment with main emphasis on the following aspects:

- *Data formats (primary data objects):*
  Research data (primary digital objects) use a standardized format representation?
  Is proprietary/restricted software necessary to render these objects? Are alternative open source solutions available?
  What amounts of data are expected and format migration options available/planned?
  What does the scientist know about the longevity of the software / data format?
- *Description of the scientific workflow(s):*
  Identifying what happens to the raw data technically?
  How is data processed?
  Input data? Processing steps? Intermediate results?
  Questions regarding the software environment: operating system, used tools and the software necessary for the processing of the primary research data?
- *External dependencies:*
  Functional external services (e.g. SaaS)?
  External data sources (Big Data / Cloud)?
  In order to gain greater knowledge of hardware dependencies we were asking about dongles, sensors and special hardware.
  Were they aware of any external dependencies, e.g. Cloud, license server, external services or network drives? Which of these can be controlled, internalized, substituted?
  Risk-assessment on external dependencies already done?
- *Virtualization / Emulation:* The users were also asked if they had already worked with virtualized systems such as workstations based on VMWare or VirtualBox?
- *IPR:* The last point was the question about licenses.

[9] Contributed by Annette Strauch, University Ulm, Germany

2.8.2 Use-Case: Embedded Systems

This use-case deals with the scientific research on the development of embedded devices. A first survey was conducted in two parts, namely it asked about primary data first and the background of research. Secondly, it enquired about the working environment and processes as well as the processing of the primary data itself. The subjects of the main research in this field are experiments of simulation data, also tests and knowledge about the real system, including vehicles, machine and plant construction. For this research field no proper preservation and accessibility period could be specified by the researcher. The preservation target was both compiler and toolchains and the access primarily for students and researchers. However, scenarios for (result) verification and embedded development as museum artifacts in the future were mentioned. No explicit preservation processes are employed currently but standard backups are done frequently. Scientific data and environment consists mostly of complex toolchains including MATLAB, MATLAB/Simulink, VHDL, DSO, EDIF. The researchers main goal is to keep toolchains (compiler etc.) functional. Primary objects (source code) not sufficient. Data needs software. Digital preservation should be useful for new students who are conducting research, but also for teaching, new projects, and in order to check if someone has falsified data for example, museum character for the embedded systems in the future.

3. ONGOING RESEARCH

To ensure sustainability both of emulators and access workflow frameworks, future development should be actively supported by a dedicated syndicate, such as a large archiving organization. It shows that large scale challenges can only be solved with wide support from the different communities. Especially in regard to the preservation of a wide know-how, a distributed approach should be chosen in which single institutions specialize on one area but an intensive exchange and shared access to the repositories remains possible. With the emergence of cloud-offerings a re-centralization of services takes place. Compute power and storage capacity becoming less relevant as end users interact with remote services through standardized client applications on their various devices. This offers the chance to use partially the same concepts and methods to access obsolete computer environments and allows for more sustainable business processes. In order to provide a large variety of user-friendly remote emulation services, especially in combination with authentic performance and user-experience, a distributed system model and architecture is required, suitable to create a digital preservation and accessible cloud service: Emulation-as-a-Service. The shift of the usually non-trivial task of the emulation of obsolete software environments from the end user to specialized providers can help to simplify digital preservation and access strategies.

3.1 Research on Special Issues with Emulation

The research group of Geoffrey Brown at Indiana university explores the use of off-the-shelf emulation and virtual machine environments to support the digital archiving of documents that require their original (obsolete) software for access.[9, 4] For example the US Government Printing Office began distributing vital statistics on CDROM with embedded proprietary software about 15 years ago – accessing these data requires running the embedded software. Another use case are the born digital materials for which emulation is the only viable preservation strategy (e.g. Classic Mac hypercard stacks), there are a number of practical hurdles to overcome. These include

1. Capturing any auxiliary programs necessary to access the artifact
2. Creating an environment that a typical patron could use to access the artifact
3. Validating that the object's significant properties are correctly rendered in the emulation environment
4. Ensuring long-term viability of emulation environment

In our experience with more than 1200 CD-ROMs both (1) and (2) can be challenging. In case (1), it is frequently difficult to find the correct versions of required auxiliary programs. In case (2), it was found that emulation environments may need considerable customization for a given artifact and this customization depends upon technical proficiency with the emulated environment. For (3) there is to make sure that the object is properly rendered in the emulated environment, not only when planning for a digital preservation action but also when taking the action (i.e. rendering the object) at a later point in time. Case (4) is especially problematical with emulation environments that are built and supported by enthusiasts because these communities appear to lose interest over time. There are also significant legal hurdles to allow access to an emulation environment [6]. For example, in the case of Classic Mac, the machine ROMs necessary for emulation cannot legally be distributed.

3.2 Evaluation of Rendering

The usual process of comparing the result of a migration as a digital preservation action is to compare stored object properties before and after the migration. In emulation the object stays unchanged. This makes it necessary to compare the actual rendering of the object. As the process of emulating a digital object is usually continuous, we have specific requirements to the digital preservation action of emulating an object. Some of these requirements have been formulated in [2], which should in addition to the steps outlined in [7] lead to reliable emulation strategies. The rendering process of the object has to be comparable. This means, that the digital object has to be rendered exactly the same under the same conditions in the original environment and in the emulated environment. We call this "deterministic behavior" of an object. Usually the rendering of an object depends on external events. These events change the behavior of an object making the rendering no longer deterministic. For a video game this can be control input by the player of the game. When running a business process, this can be an external service providing data that influences the results of the process. For a successful evaluation we need to be able to provide external data in a unified way to the emulated environment to deterministically render the digital object. But not only for the evaluation of a rendering external data

has to be applied to the process. Even when using the emulator to actually rerun a specific setting, external data has to be applied to the emulation environment. This can be input as player actions in a video game, which has to be recorded from the host system and passed over to the emulation environment. But it also can be data from a no longer existing web service supplying input to a business process. In this case a mock-up of the original service has to be provided to the emulation environment, to make sure that the process can be rerun. Any kind of interface between the object to be emulated and the surrounding environment is what influences the rendering and what has to be either supplied by simulated data (for evaluation purposes) or by an interface between the host system and the emulated environment. To actually compare the rendering of a digital object, emulation environments have to be able to provide information about the emulation process. Depending on the significant properties of the emulated object, the requirements to this information can have different forms. Generally speaking any kind of output from the object to be emulated to the environment surrounding it has to be captured for comparison. This can be either output on a screen in the form of screenshots (if a single rendered image is significant) to a recording of a stream of images (if a continuous rendering is considered important), but also all other kind of results of a rendering, e.g., sound waves, network activity, signals to actors. If the rendering of a digital object is deterministic, then the recorded data is comparable. Usually not only the data itself, but also the temporal component of the data is important. E.g. in a video game this means that the game is running at the same speed as in the original environment, not too fast but also not too slow. In a business process it could be that the answer of an emulated process to a request has to be fast enough so the requesting system is not running into a time-out reporting the process to fail.

3.3 TIMBUS Project

TIMBUS is a European research project that deals with the preservation of business processes. To address this, the project extends its research focus beyond the area of data preservation. TIMBUS considers the future executability of these processes including the dependence on third-party services, information and capabilities. One of the main tasks of the project in the first year was the creation of a context model that defines technical, organizational and legal aspects of a process [3]. Other work included a study on legal requirements and the extraction of software dependency information from running systems and their recreation in a new environment. Current research is done on refinement of the concepts and tool support, as well as their validation on business and scientific use cases. The view-path of an object in its environment is one of the things that is described in the technical aspects of the context model of TIMBUS. Tools developed in the project allow an automated identification of all necessary secondary objects in the view-path on a Linux system. This information can then be used to create a "clean" system from scratch that rebuilds the view-path, allowing objects to run in a future environment. But not only the view-path options are identified in the context model, also external dependencies that influence the rendering of a digital object are documented. This information is necessary for evaluating the rendering of a business process in a future environment. How the evaluation if a digital object's significant properties are rendered correctly in a new environment is performed is described in a validation framework also developed in TIMBUS [2]. Based on the context model, risk can be identified and analyzed, thus aligning preservation actions more fully with enterprise risk management (ERM) and business continuity management (BCM) [1, 5].

3.4 bwFLA Project

The Baden-Württemberg Functional Longterm Archiving and Access (bwFLA) [10] is a two-year state sponsored project transporting the results of ongoing digital preservation research into the practitioners communities. Primarily, bwFLA creates tools and workflows to ensure long-term access to digital cultural and scientific assets held by the state's university libraries and archives. The project builds on existing digital preservation knowledge by using and extending existing preservation frameworks. It will define and provide a practical implementation of archival workflows for rendering digital objects (user access) in their original environment (i.e. application) with no suitable migration strategies available, like interactive software, scientific tool-chains and databases, as well as digital art. Thereby, the project focuses on supporting the user during object ingest to identify and describe all secondary objects required. This way technical meta-data will be created describing a suitable rendering environment for a given digital object. The technical meta-data will serve as a base for *long-term access through emulation*.

3.5 Meta-Data Describing Software Environments[11]

While the EC Planets and KEEP projects[12] made significant strides towards bringing emulation into the mainstream of digital preservation, it is important to highlight the fact that this prototype work must be built on in order to ensure that the tools and services created in these projects are maintained and developed. It is especially important that the tools should be accessible to a wide range of stakeholders, who have been able to provide input to their development. This extended abstract will start by showing how the TOTEM registry[13] was developed in KEEP to support the KEEP EF, and how it is also now used within the bwFLA project. We then move on to show how TOTEM was originally developed within KEEP as a database, but that colleagues at the University of Cologne have recently updated this work by mapping the data models behind TOTEM onto the Planets software and hardware ontologies to produce a TOTEM RDF[14] version for use with Linked Data, which many stakeholders prefer. This RDF initiative would also fit in well with the notion of the OPF eco registry. Finally,

[10] bwFLA homepage, `http://bw-fla.uni-freiburg.de`.

[11] Contributed by Janet Delve and David Anderson, Future Proof Computing Group, University of Portsmouth, UK & Johanna Puhl, University of Cologne & Tatiana Jimenez Cardenas, University of Freiburg

[12] Project homepages, `http://www.planets-project.eu/`, `http://www.keep-project.eu/`

[13] Available at: `http://keep-totem.co.uk/` and `http://www.verlagdrkovac.de/3-8300-6418-7.htm`

[14] Resource Description Framework, see: `http://www.w3.org/RDF/`

we demonstrate how external registries such as TOTEM and IIPC are playing a vital role in helping develop the PREMIS environment descriptions involving emulation, as part of the work carried out by the PREMIS Environment Working Group.

3.5.1 TOTEM and Emulation Frameworks

The KEEP EF[15] has been developed to call on external registries such as TOTEM in order to determine suitable emulation pathways for particular digital objects, or ranges of thereof. Here TOTEM's function is to locate compatible elements for versions of: software and libraries; operating systems, and hardware; which can then form the parts of an emulation pathway. Beyond KEEP, in the bwFLA project, the TOTEM data model is used as a base for a temporary bwFLA database with additional new entities Viewpath and Emulator, where a so-called viewpath describes an instance of the current environment in which a specific digital object runs successfully. The description of this environment is described by concrete and abstract elements where concrete elements are known dependencies that can be retrieved from an archive. These are represented with a set of pointers to TOTEM entities that build a suitable emulation pathway containing elements such as operating system, software, libraries, etc. On the other hand, abstract elements aim to preserve the knowledge of an expert regarding the context of the digital object. This knowledge is stored as descriptions, feedback, issues and configurations during the construction of a viewpath. Finally the last element of a viewpath is an emulator with the capabilities to emulate the environment described. An emulator is a specialization of a software in a generalization relation. A software entity contains the basic data regarding an emulator and within the TOTEM entity hierarchy it also indirectly contains data that fulfill a "runs on" relation. The aim of having a specialization of a software entity is to describe its capabilities, that in this particular case include the operating systems an emulator is able to emulate. Moreover this link to operating systems adds an implicit mapping of emulators to hardware requirements. Hardware and related entities are still a weak connection between bwFLA and TOTEM models, this is because from bwFLA's perspective, information hardware requirements is only necessary on a high level representation which is hard to map to a more structured and detailed model in TOTEM's side. For instance, the model should be able to map quite detailed hardware requirements derived from a concrete software/operation system description to a generic hardware/system-emulator subsuming a wide range of potential hardware configurations. Particularly the bwFLA data model requires information about platforms; this need has been handled temporarily by adding a property (platform) in the HardwareArchitectureType TOTEM entity. The inclusion or relocation of such a field is to be analyzed between both teams. Currently the implementation is under preparation to actually populate the database with test data. This stage will be followed by defining functions and operations on such data

3.5.2 Mapping the TOTEM Data Model in RDF

There exists a large body of work aiming to define and collect significant properties for different aspects of long-term preservation and emulation. Between the TOTEM team and the University of Cologne we have shown in a proof-of-concept that emulation metadata from the TOTEM database can be mapped into an OWL/RDF-presentation of preservation metadata resulting from the Planets-project and related projects.

The Planets Ontology is extensible and has already been extended following the outcomes of several European preservation projects. In essence, it is modelled out of six different sub-ontologies that all follow the same structure and are then integrated via namespaces into one overall RDF-file – the Planets ontology. The resulting design and structure can be easily queried via SPARQ[16] or similar techniques. It can also help to gain a lot of cognition into the different intellectual approaches on metadata and how they are related.

Within this session we will delineate how entities from the TOTEM database were converted into the Planets / RDF ontology-structure by conceptually mapping these two data models. We will therefore describe the existing RDF-model and explain its concept. From there we will show a prototypical bridging between a TOTEM entity and an OWL/RDF entity in the Planets ontology-structure. Finally we will discuss the resulting TOTEM OWL/RDF file as a whole and future visions:

- This work can help to compare the different foci from different projects;
- later enable a community to automatically query the RDF-files and gain new cognition by that and
- be extended by collecting further work into the already existing RDF structure

3.5.3 TOTEM and the PREMIS Environment Working Group

The PREMIS Environment Working Group has developed use cases to support PREMIS descriptions and relating documentation for the new PREMIS Environment Entity: an essential component to support emulation as a preservation action. The working group has included representatives from the PREMIS Editorial committee, the TOTEM technical registry, the IIPC, DAITSS and the TIMBUS project,[17] and has received user requirements from New York University. This part of the session will focus on the recent emulation use case discussed by the PREMIS Environment Working Group. This emulation use case, based on creating an emulation environment for a digital object in a library with limited technical information from the catalogue, makes varied and frequent calls on the TOTEM registry within the PREMIS description.

[15] Homepage and software download, `http://emuframework.sourceforge.net/`

[16] Please refer to `http://www.w3.org/TR/rdf-sparql-query/`

[17] Please see the project homepages for the IIPC Preservation Working Group, `http://netpreserve.org/about/pwg.php`, `http://daitss.fcla.edu/` and the TIMBUS project `http://timbusproject.net/`

The session will conclude by drawing attention to the need for future collaboration in all the above areas, with the help of such organizations as the DPC[18] and the OPF.[19]

4. OUTLOOK

As there is a rising awareness of the strategic importance of data, future research environments, scientific desktops and general business processes should consider preservation and long-term data management activities from the beginning. Future users of digital assets significantly benefit from accessible data and user-friendly functional toolsets both in the scientific and cultural heritage as well as in the commercial domain. Emulation services for digital preservation can help to bridge outdated working environments for a wide range of objects and original environments onto todays' devices. Nevertheless, a long-term emulation-based archiving strategy can not be achieved by a single organization, even of the size of a national library, since specific technical knowledge will be spread between the archive and science communities depending on available digital artifacts and archival focus. Only the collaborative effort of archivist, collection managers and researchers from memory institutions combined with developers can lead to technical and organizational solutions securing our digital cultural heritage and research data assets. Several working groups and small to large scale projects are active to push the research in the domain of emulation in digital preservation.

5. ACKNOWLEDGEMENTS

The workshop organizers thank the iPres 2012, Toronto, Canada coordinators for offering us the possibility of an emulation workshop. They also want to thank all contributors to this introductory paper. A wide range of content holders and research projects have contributed to the topics cited and presented here.

6. REFERENCES

[1] J. Barateiro, G. Antunes, and J. L. Borbinha. Manage risks through the enterprise architecture. *HICSS*, pages 3297–3306, 2012.

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

[3] R. Mayer, A. Rauber, M. A. Neumann, J. Thomson, and G. Antunes. Preserving scientific processes from design to publication. In G. Buchanan, E. Rasmussen, and F. Loizides, editors, *Proceedings of the 15th International Conference on Theory and Practice of Digital Libraries (TPDL 2012)*, pages 23–29. Springer, 2012.

[4] T. Reichherzer and G. Brown. Quantifying software requirements for supporting archived office documents using emulation. In *Digital Libraries, 2006. JCDL '06. Proceedings of the 6th ACM/IEEE-CS Joint Conference on*, pages 86–94, june 2006.

[5] D. Simon, F. Simon, D. Draws, and S. Euteneuer. Short term preservation for software industry. In *8th International Conference on Preservation of Digital Objects (iPRES2011)*, pages 167–170. National Library Board Singapore and Nanyang Technology University, 2011.

[6] J. van der Hoeven, S. Sepetjan, and M. Dindorf. Legal aspects of emulation. In A. Rauber, M. Kaiser, R. Guenther, and P. Constantopoulos, editors, *7th International Conference on Preservation of Digital Objects (iPRES2010) September 19 - 24, 2010, Vienna, Austria*, volume 262, pages 113–120. Austrian Computer Society, 2010.

[7] D. von Suchodoletz, K. Rechert, J. van der Hoeven, and J. Schroder. Seven Steps for Reliable Emulation Strategies – Solved Problems and Open Issues. In A. Rauber, M. Kaiser, R. Guenther, and P. Constantopoulos, editors, *7th International Conference on Preservation of Digital Objects (iPRES2010) September 19 - 24, 2010, Vienna, Austria*, volume 262, pages 373–381. Austrian Computer Society, 2010.

[8] C. Webb, D. Pearson, and P. Koerbin. 'oh, you wanted us to preserve that?!' statements of preservation intent for the national library of australiaâĂŹs digital collections. *D-Lib Magazine*, 2012.

[9] K. Woods and G. Brown. Assisted emulation for legacy executables. *International Journal of Digital Curation*, 5(1), 2010.

[18] Digital Preservation Coalition, `http://www.dpconline.org/`

[19] Open Planets Foundation, `http://www.openplanetsfoundation.org/`