Design and Development of an Emulation-Driven Access System for Reading Rooms

Klaus Rechert, Dirk von Suchodoletz, Thomas Liebetraut; Albert-Ludwig University, Freiburg, Germany
Denise de Vries, School of Computer Science, Engineering and Mathematics, Flinders University, Adelaide, South Australia
Tobias Steinke; German National Library, Frankfurt, Germany

Abstract

Archives and libraries around the world are facing the same challenges with regard to accessing digital artefacts that reside on old media and require obsolete hardware, operating systems and applications to access the contents. We discuss a framework that delivers a modular and scalable platform that includes emulation of the original environments, extends existing reading room systems and facilitates appraisal and access of artefacts without undue overheads for institutional staff. The characterization and classification of objects can be automated based on the artefact's own contents. A guided system will ensure that the reading room client will interact with the historical artefact as though he/she were using the original equipment.

A suitable framework should be equally viable as a distributed system where multiple institutions can share their resources and expertise to provide a cost effective system that presents a standard interface to users across organisations. Our paper presents the preliminary survey conducted in the involved libraries, a derived requirements analysis based on selected use-cases as well as the description of ongoing developments. These include the adaptation of Emulation-as-a-Service components to the individual needs of memory institutions.

Introduction

Memory institutions already hold a substantial quantity of digital artefacts and receive an increasing number of complex digital-born objects. These objects require different handling from the traditional linear and static material. They must undergo new treatment with regard to methods and workflows to render them accessible to future users.

The emulation-driven reading room access project focuses on a target group of a broad range of significant digital artefacts for available from floppy, optical media, and other magnetic data carriers, which chart the introduction of the computer age to memory institutions across nearly three decades of the digital revolution – the shift from analog material to increasingly complex digital compilations and applications. These artefacts constitute an important part of the institutions' digital collection, and a vital record of the development of the changing discourse in science, education and entertainment toward a digital society [13, 14]. The projects presented address an increasingly urgent demand for institutional preservation and access strategies for dynamic, complex, often interactive digital objects. There is an increased understanding that these artefacts will be essential for the study, appreciation, and understanding of the history of science but the formation of the digital age.

Many artefacts of the growing electronic collections of libraries and archives are outdated, meaning no modern application or environment can render or run them in a usable and authentic way [16, 15]. At the moment, migration is the method most often deployed and trusted by libraries and archives for long-term preservation of and access to digital objects. This strategy carries along the artefacts through a constantly changing digital landscape, made up of changing hardware and software technology and configurations. It usually requires translation of the artefact's inner structures to an up-to-date schema. Although these translations make it possible to use and render them in actual computer environments, such an approach unnecessarily limits the number of artefact types that can be archived. Moreover, suitable migration tools, are usually not available for dynamic and interactive digital material. A further problem that concerns researchers is the data-centric view of a migration strategy. Modern digital artefacts involve not only data but can also be complete software packages or multimedia objects composed of multiple files. Thus, a pure data-centric strategy misses important pieces that are necessary for a complete and authentic re-enactment, for instance to replicate a complex digital environment with its complete internal context to be preserved.

The classes of artefacts, considered in our paper, require appropriate ingest and access workflows to create proper technical metadata and to experience (render) them on today's systems. Emulation of deprecated digital ecosystems provides an answer for accessing and experiencing an increasing variety of (otherwise inaccessible) object types. Compared to traditional reading room systems, the access to antiquated artefacts involves a more complex workflow. Emulation is usually complex and needs additional software components and configuration, and therefore should be wrapped into a convenient, user-friendly platform. To use emulation one must also make available the original operating system (OS), libraries and applications that replicate the whole environment. For that to occur, each object or artefact must be fully described as to what its requirements are, such as codecs, fonts, and dependencies, and these data stored in the object's metadata.

To design and develop an emulation-driven ingest workflow and access system for reading rooms we identified the following challenges, we will address by two distinct but coordinated research projects:

- How can classify the digital object and create metadata?
- How to co-ordinate the emulation of the hardware, OS and application required?
- How to manage legal requirements such as licenses of OS and proprietary software?

- How to automate the processes required to enable accessibility?
- How to make the framework cost effective?
- How to incorporate the framework into existing appraisal and reading room services?
- How to provide a shareable service across institutions?
- How to provide individual, persistent user session, when working with / exploring a digital artefact?

Related Work

Emulation does not operate on the object directly but rather addresses the environment which was used to create the object. This means, for example, the replication of software and/or hardware through other software. In the best case, it will not make any difference whether the object is handled through an emulated or original environment. Emulators, i.e. specialized software applications running in digital environments, preserve or alternatively replicate original environments. Research on emulation as a long-term archiving strategy has matured since the first reports on archiving of digital information in 1996 [3], and fundamental experiments with emulation executed by Rothenberg [11]. The next phase was reached by the EU projects PLANETS and KEEP. The former looked into the inclusion of emulation into preservation workflows [16, 1], while the latter was focused primarily on media transfer, emulators and emulation frameworks [7, 5].

A more community-centered approach is the Olive platform [1] that is specifically designed to allow collaboration of different curators to a Cloud-based library. Olive also utilizes local emulation using a thin client approach to run virtual machines, but uses its own protocol to stream data necessary to execute the virtual machine over the network. Modifications of a virtual machine, for example newly installed software, can be transferred back to the archive, making derivatives of digital objects possible [12]. Pure web-based approaches like the JavaScript powered emulator service of the Internet Archive project [2] allow convenient access to a large collection of historic computer games but lack portability and generality. While the user only requires a recent web browser, emulated setups are restricted to a specific JavaScript implementation and service offer. The emulation instances remain stateless and accessing user data and customization is not supported.

Problem Definition

Libraries and archives around the world host an increasing variety of diverse multimedia artefacts, such as encyclopedias, electronic learning materials, digital art, and scientific tools like simulations. These artefacts document not only the development of scientific research and teaching in the computer age, they contain as well cultural creations or documents of major social events and therefore play an increasingly important role in the collections of memory institutions.

The projects at the South Australia State Library and the libraries in Germany were created out of the understanding that today's workflows in these institutions are not well prepared and enabled to handle various digital material. The different challenges are manifold starting from material which has not been imaged yet, missing descriptive and technical metadata, to inappropriate workflows for search, retrieval and access. The projects started preliminary phases in 2013 to get a better overview and understanding of the objects in their stacks. The idea was to create input in form of selected use-cases for the reading room ingest and access projects started in the beginning of 2014.

From the analysis of each institution's holdings, possible classification of objects were created. The selection of use-cases has just been completed. They include different types of magnetic media in the case of the South Australia State Library and various optical media in the case of the German libraries. For our project a list of electronic encyclopedia, educational material, software and interactive documents, proprietary database systems, electronic books, electronic guides, digital art and scientific simulation has been compiled.

South Australia State Library

In the South Australia State Library (SLSA) project a variety of use-cases has been selected. These range from artefacts created in 1983 to 2010, covering many different types of files, formats, creating applications and systems platforms. These use-cases are also categorized from "simple" to "complex". A simple one would be a single file of a standard format which was created by a readily available software application, for example a text file containing only ASCII characters which may be accessed by such applications as Notepad, Wordpad, vi, SimpleText and many more. A complex one may be a database system incorporating both database frontend and backend requiring a specific database platform and version or maps that require not only a specific version of the rendering software but also extensions to access all the artefact's features.

German National and Bavarian State Libraries

The German National Library keeps about 500,000 electronic publications on various media. It includes besides other formats numerous artefacts representing a wide range of encyclopedias and tutorials available on standard optical disks like CD-ROM or DVD. From this collection a subset of landmark artefacts was chosen as a use-case for further evaluation and testing of the developments. For many of these artefacts the typical system requirements are not met any more on the modern generation of reading room installations. There are various reasons for this, like incompatible versions of required operating systems or missing crucial components and programs for playback of multimedia elements:

- Class of electronic encyclopedias
  - Space Lexicon, Francis, 1998
  - Compton's Interactive Encyclopedia – Pathfinder 3.0, TLC, 1998
  - Kindler's New Literary Lexicon, Systhema, 1999
  - Compact Dictionary of Biology, Spectrum, 2002
- Class of electronic learning and training programs
  - Viennese – Clear Anyway, Carussell Communications, 1998
  - Multimedia Vocabulary Trainer English, Juncker, 1999

[1] Project homepage, `https://olivearchive.org/`, last retrieved 2/1/2014.

[2] JMess Internet Archive example, `http://blog.archive.org/2013/10/25/microcomputer-software-lives-again-this-time-in-your-browser/`, last retrieved 2/1/2014.

  - Impulse Physics – Interactive screen experiments / Electricity / 1, Klett, 2001
  - Inky – English for cool kids, Kidoclic, 2001

The Bavarian State Library hosts a continually expanding collection of 60,000 – 70,000 floppy disks and optical media. As part of the preliminary project the library focused its attention on early multimedia presentations, in particular on both a historical and technical perspective. The uses cases were selected from being of special interest to researchers and also for the presentation requirements for the original "Look & Feel" of the artefacts.

- Class of virtual tours and exhibitions
  - Goethe in Weimar – A Virtual Journey into the World of the Great Poet, Navigo multimedia, 1995
  - Deutscher Bundestag Multimedia + Interactive, Deutscher Bundestag, Public Relations Division, 1997
  - Cologne Cathedral – A virtual Tour of 2,000 Years of Art, Culture and History, Dt. Art Verlag, 1998
  - Deutsches Museum – Flight Museum Schleissheim, 1998
  - Virtual dig – A simulated archeological excavation of a Middle Paleolithic site in France, Mayfield, 1999
  - Jewish Life in Baumbach – A Virtual Tour, Search for Clues working group, 2003
- Class of city guides, country guides and maps
  - Munich – Interactively; The unique combination of map and guide on CD-ROM, Grafe and Unzer, 1995
  - Wuerzburg. The Virtual City Guide, 1998
  - Macedonica '99 – Encyclopedia Multimedia, Semos Multimedia, 1999
  - Atlas of Switzerland Interactive – 3D Topography Maps and Statistics, Institute for Cartography ETH Zurich, 2000
- Class of music in multimedia titles
  - The Electronic Collection of Songs – 100 songs with lyrics, melodies, notes and intonations from the PC speaker on diskette, Paul Rossi, 1994
  - Ullstein-Multimedia Encyclopedia of Music – The Interactive Standard Work of Classical Music, Ullstein, Soft-Media, 1996
  - Wolfgang Amadeus Mozart, Fantasy and Sonata in C minor; The original manuscript of Mozart clavier interactively made to sound, Internationale Stiftung Mozarteum, 2006

Requirements Analysis

Each case currently requires a manual process to gather as many features as possible from all accompanying documentation, including manuals, packaging and labels. It is anticipated that this process can be partially automated in the future by identifying key attributes for the classifications.

The preliminary analysis produced a couple of interesting results. An often found problem in every institution was the uncertainty of the size and content of the actual collection. The actual ingest workflows up to now primarily focus on standard material like books and journals and often fail to achieve standardized results for contained digital material like floppy disks or CD-ROMs being part of a book or a journal. The items were often poorly described, the metadata incomplete or of bad quality and thus many artefacts were difficult to find in the catalog. A major point in our project is the definition on how especially technical metadata should be structured. Finally, we identified three main domains of action: imaging of delivered media for bitstream preservation, and proper ingest workflows within the institution and access workflows for future users.

In nearly every case the objects were not imaged yet and were still on the original medium. There are challenges in imaging a range of rare physical formats like 3" or 8" floppies or floppies with a non PC-compatible recording. To enable future access the delivered objects have to be copied from their original data carrier as the media and the reading devices will become obsolete in the future or simply deteriorate. While optical media tend to be much easier to image, the existing workflows usually do not scale to the amount of material in the collection. The sheer amount of storage space required produces peaks in demands for storage space in the institution's digital repositories. The costs of imaging the existing collections can be approximated from the efforts spent during the preliminary projects. They depend on the class of material and the number and size of objects in each class. Depending on the fragility and confidentiality of the objects either in-house operations or contracting it to third party options can be chosen. The latter case makes sense especially for large amounts of standard material without special restrictions as many institutions lack the appropriate machinery for automated operation. Nevertheless, while the object imaging is not directly part of these projects, the results from our findings should provide hints on how to design future ingest workflows.

The average user of a memory institution is not necessarily familiar with past computer systems and setup. The goal of a future reading room system is to create ways to automate the access to the desired digital artefact. The concept is to have different original environments available to allow an object to be loaded into its proper environment. The automation needs to assist the average user of a reading room system, who is not necessarily familiar with past computer interaction and GUI concepts.

Emulation as a Service

Until now emulation has been seen as domain reserved for technical experts. Furthermore, emulation did not scale well due to the laborious preparation and technical setup procedures. Driven by the principles of division of labor but also based on the observations on potential stakeholders a scalable service model has been developed – Emulation as a Service (EaaS) [10].

EaaS provides a modular set of technical building blocks (*emulation components*) to standardize deployment and to hide individual emulator complexity. Each emulation component encapsulates a specific emulator type as an abstract component with a unified set of software interfaces (API). This way, different classes of emulators become interoperable, e.g. could be interchanged or interconnected. Furthermore, single emulation components can be efficiently deployed in a large-scale cluster or Cloud infrastructure. An EaaS service-provider then is responsible for efficient hardware utilization and concentration of technical expertise and thus lighten the memory institutions' technical workload and

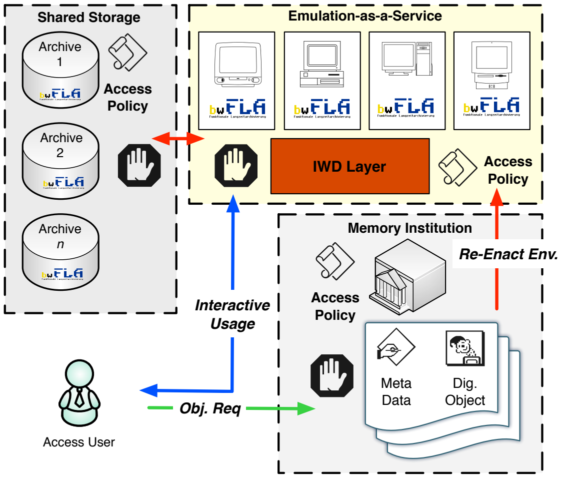
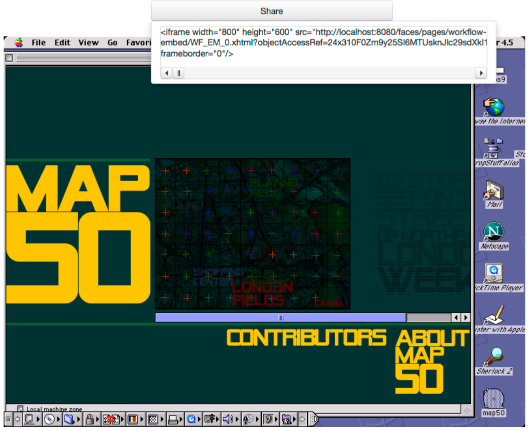


Figure 1. Distributed EaaS access workflows.



```
<iframe width="800" height="600" src="http://localhost:8080/faces/pages/workflow-
embed/WF_EM_0.xhtml?objectAccessRef=24x310F0Zm9y25SI6MTUsknJlc29sdXkI1
frameborder="0"/>
```

Figure 2. Digital art object rendered in a prepared environment.

requirements on necessary infrastructure.

While there has been research into archiving metadata [2, 6, 8] more information and thus metadata must be obtained for emulation as a cloud service (EaaS). When using EaaS consideration must be made of the possible distribution of the whole system. Not only is the virtual computer remote but the other individual components may also be distributed as shown in Fig. 1 where the different resources required for accessing an artefact may be physically held in separate places. Typically, local memory institutions are safekeeping individual and potentially unique objects as well as are the starting point for a user request. Standard software components for instance can be shared among various institutions or can be outsourced to specialized providers.

Currently, emulation components for all major past and present desktop systems, e.g. PPC, Sparc, M68k, Intel-based x86, etc., and major operation systems, e.g. OS/2, MS Windows, Mac OS 7, Linux, and newer, etc., are available for deployment. A detailed technical description of an EaaS framework and its workflows can be found in earlier work [10].

As a next step tailored workflows have to be developed to fit the needs of the individual institutions. These should safeguard quality assurance of content, as well as generating technical metadata to assure completeness and later re-enactment for access.

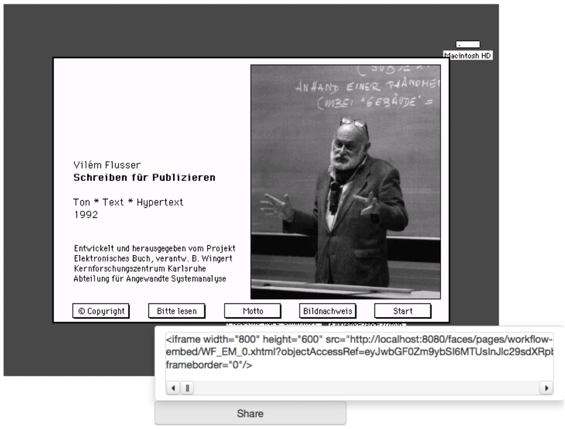
Preliminary Results

As a result of previous work prototypical workflows for object preparation, i.e. to create or to modify system environments and to link a digital object to a specific rendering environment were created and evaluated. While preparing a rendering environment through a guided process is optional, this however, results in technical metadata with an exact description of the environments view-path [16] and its configuration. This information then can be (re-)used to classify and index rendering environments, such that they provide a base for other objects and a starting point for the creation of new derivatives. For instance, Fig. 2 shows a rendered CD-ROM from the Transmediale Archive, Berlin in a prepared emulated environment. A detailed description on the workflow and metadata generated was presented in earlier work [9].

In some cases however, there is no isolated artefact available, instead the object is connected to its rendering environment, or the environment itself qualifies as a valuable object, e.g. an image of the hard-disk of a famous person donated to a memory institution. For instance, Fig. 3 shows an emulated version of a hard-disk image taken from an original Apple Macintosh computer held by the Flusser Archive, Berlin.

Furthermore, the user is able to evaluate the performance of the environment and object. Fig. 4 shows the UI for the art domain. The rendering of digital art and evaluation of their performance have been evaluated using a large collection of CD-ROM art[4].

The outcome of these workflows is technical metadata describing the runtime configuration of the chosen setup as well as descriptive metadata describing individual performance features or the lack thereof. The outcome can further be published and used to replicate exactly the same environment by invoking an emulation component, given a publicly available object, e.g. as embeddable HTML-tag rendering the output as an interactive HTML5 canvas (cf. Fig. 3[3] and 2).



Vilém Flusser
Schreiben für Publizieren

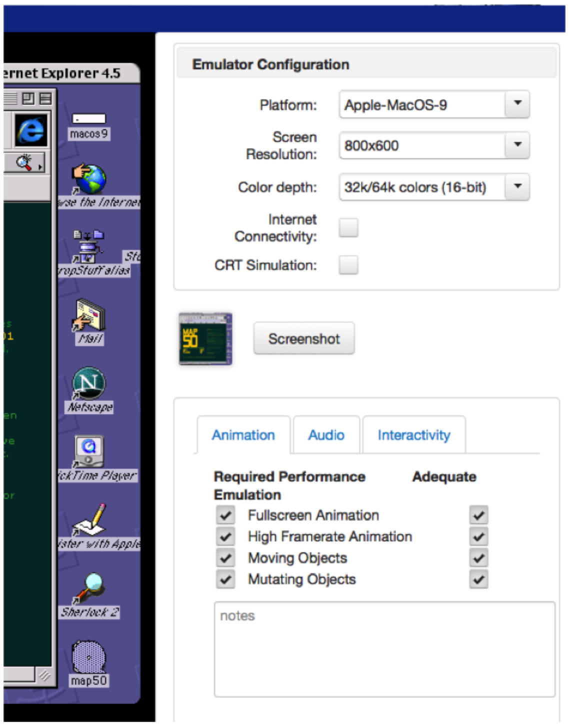
Ton * Text * Hypertext
1992

Entwickelt und herausgegeben vom Projekt
Elektronisches Buch, verantw. B. Wingert
Kernforschungszentrum Karlsruhe
Abteilung für Angewandte Systemanalyse

```
<iframe width="800" height="600" src="http://localhost:8080/faces/pages/workflow-
embed/WF_EM_0.xhtml?objectAccessRef=eyJwbGF0Zm9ybSI6MTUsInJlc29sdXRpt
frameborder="0"/>
```

Figure 3. Rendered original environment of Mac OS 7 with Hypercard.

[3] Currently the Flusser image is available online at the bwFLA project website, `http://bw-fla.uni-freiburg.de/demo-flusser.html`, last retrieved 2/3/2014.



Emulator Configuration

Platform: Apple-MacOS-9

Screen Resolution: 800x600

Color depth: 32k/64k colors (16-bit)

Internet Connectivity: □

CRT Simulation: □

Screenshot

Animation | Audio | Interactivity

Required Performance Emulation | **Adequate**

☑ Fullscreen Animation ☑

☑ High Framerate Animation ☑

☑ Moving Objects ☑

☑ Mutating Objects ☑

notes

Figure 4. *Evaluation of performance properties of a rendered digital object.*

Conclusion and Outlook

We hope to broaden the scope of digital objects that can be actively used in reading rooms and also provide a readily usable platform for the scientific community so that artefacts that are currently difficult to access can be made as available as artefacts that were created on current systems. In later stages of these projects necessary training material for the operators in the memory institutions and on-line help and guidance for the users will be created.

These may be held by different collaborating institutions such as libraries sharing database platforms, or operating system images. The metadata must be able to capture the necessary parameters to assemble all requirements to render the artefact accessible. More investigation is needed to specify the full list of requirements beyond OS versions and their deltas, and software dependencies. Licensing of software applications is a complex problem not yet resolved. Over the past four decades there have been many different models for licensing, these include:

- A node locked license or authorized user license
- A mobile compute license
- A floating license or Token license
- A site license
- A limited site license
- Date-based versioning
- License activation kits

A usable solution for not only capturing, classifying and storing this information is needed as is a way to be certain that the institution can ensure that licenses will continue to be valid.

Beside technical solutions, sound cost models need to be developed to support decision making about which preservation steps and strategies should be applied to the different object classes. Deploying the ingest and access workflows as suggested in our paper requires sufficient funds for both technical systems and personnel to administer them.

Nevertheless, the reading room system suggested needs to be sustainable over time and must be able to adapt to technological change. Thus, the development and deployment of a test suite should accompany the scheduled re-evaluation of the reading room system as a part of an ongoing surveillance ensuring accessibility to current technologies.

References

[1] C. Becker, H. Kulovits, M. Kraxner, R. Gottardi, A. Rauber, and R. Welte. Adding quality-awareness to evaluate migration web-services and remote emulation for digital preservation. In *Proceedings of the 13th European Conference on Digital Libraries (ECDL09)*, 2009.

[2] L. Carroll, E. Farr, P. Hornsby, and B. Ranker. A comprehensive approach to born-digital archives. *Archivaria*, 2011.

[3] Commission on Preservation and Access and The Research Libraries Group. Report of the Taskforce on Archiving of Digital Information. http://www.clir.org/pubs/reports/pub63watersgarrett.pdf, 1996.

[4] D. Espenschied, K. Rechert, I. Valizada, D. von Suchodoletz, and N. Russler. Large-Scale Curation and Presentation of CD-ROM Art. In *iPres 2013 10th International Conference on Preservation of Digital Objects*. Biblioteca Nacional de Portugal, 2013.

[5] B. Lohman, B. Kiers, D. Michel, and J. van der Hoeven. Emulation as a business solution: The emulation framework. In *8th International Conference on Preservation of Digital Objects (iPRES2011)*, pages 425–428. National Library Board Singapore and Nanyang Technology University, 2011.

[6] B. Lohman and E. Noordermeer. Keep architectural design document emulation framework. Release 2.0.0 (February 2012), 2012.

[7] D. Pinchbeck, D. Anderson, J. Delve, G. Alemu, A. Ciuffreda, and A. Lange. Emulation as a strategy for the preservation of games: the keep project. In *DiGRA 2009 – Breaking New Ground: Innovation in Games, Play, Practice and Theory*, 2009.

[8] PREMIS Editorial Committee. Premis architectural design document. http://www.loc.gov/standards/premis/v2/premis-2-2.pdf, 2012.

[9] K. Rechert, I. Valizada, and D. von Suchodoletz. Future-proof preservation of complex software environments. In *Proceedings of the 9th International Conference on Preservation of Digital Objects (iPRES2012)*, pages 179–183. University of Toronto Faculty of Information, 2012.

[10] K. Rechert, I. Valizada, D. von Suchodoletz, and J. Latocha. bwFLA – A Functional Approach to Digital Preservation. *PIK – Praxis der Informationsverarbeitung und Kommunikation*, 35(4):259–267, 2012.

[11] J. Rothenberg. Preserving authentic digital information. *Authenticity in a digital environment*, pages 51–68, 2000.

[12] M. Satyanarayanan, V. Bala, G. St.Clair, and E. Linke. Collaborating with executable content across space and time. *7th International Conference on Collaborative Computing:*

Networking, Applications and Worksharing (Collaborate-Com), (October):528–537, 2011.

[13] H. Stuckey, M. Swalwell, and A. Ndalianis. The popular memory archive: Collecting and exhibiting player culture from the 1980s. In *Making the History of Computing Relevant*, pages 215–225. Springer, 2013.

[14] H. Stuckey, M. Swalwell, A. Ndalianis, and D. de Vries. Remembrance of games past: the popular memory archive. In *Proceedings of The 9th Australasian Conference on Interactive Entertainment: Matters of Life and Death*, page 11. ACM, 2013.

[15] M. Swalwell and D. de Vries. Collecting and conserving code: Challenges and strategies, 2013. The Invisibility of Code; accepted, to be published.

[16] J. van der Hoeven and D. von Suchodoletz. Emulation: From digital artefact to remotely rendered environments. *International Journal of Digital Curation*, 4(3), 2009.

Author Biography

Dr Klaus Rechert studied Computer Science and Economics at the University of Freiburg and received a Diploma in Computer Science in 2005 and is currently the project manager of the bwFLA project, a two-year project leveraging emulation for access and migration tasks in digital preservation. He was a visiting researcher at the National Institute of Informatics (NII) in Tokyo, Japan and has worked a software engineer on the PLANETS EU-FP6 project. Klaus is currently preparing the upcoming reading room project and is project manager of a joint state project on research data management.

Dr Dirk von Suchodoletz is currently holding a position as a lecturer and principal researcher at the chair in Communication Systems at the Institute for Computer Science at Freiburg University. Dirk received his Ph.D. in computer science at Albert Ludwigs University of Freiburg in 2008 on "Requirements for emulation as a long-term preservation strategy". Through 2006 till 2010 he was involved in the EU-funded project PLANETS. Recently he cooperated with the National Archives of New Zealand in a full system preservation project. At the moment he participates in the "bwFLA" project on emulation based workflows with the goal of defining and providing a practical implementation of archival workflows for the rendering of digital objects in its original environment.

Dr Denise de Vries has, since the early 1980s, developed commercial complex database systems on a variety of platforms from mainframes to a range of personal computers. She is currently a lecturer of computer science in the School of Computer Science, Engineering and Mathematics at Flinders University. Denise's current research is on techniques to preserve digital history and data semantics including techniques to deal with changes to information in a database such as structural change, semantic change and constraint change. She is a Chief Investigator of the multi-disciplinary Linkage project "Play It Again" and developed the "Australasian Heritage Software Database" which is co-managed with Melanie Swalwell.

Tobias Steinke is a digital preservation and web archiving specialist at the ICT department of German National Library since 2003. He represents the library in several national and European projects and is member of national committees for standards for technical approaches. Born in 1971 in Berlin, Germany, Tobias studied Computer Science at the University of Technology Darmstadt and worked for three years for a software development company.