Functional Access to Electronic Media Collections using Emulation-as-a-Service

Thomas Bähr, Michelle Lindlar
Technische Informationsbibliothek
Hannover, Germany
firstname.lastname@tib.uni-hannover.de

Klaus Rechert and Thomas Liebetraut
University of Freiburg
Freiburg, Germany
firstname.lastname@rz.uni-freiburg.de

ABSTRACT

Over the last 30 years the German National Library of Science and Technology (TIB) accumulated a large collection of various electronic media, such as floppies or CD-ROMs. This poster describes both practical workflows as well as technical infrastructure to provide authentic and interactive access to the TIB's large electronic media collection.

General Terms

Case Studies and Best Practice

Keywords

Emulation, Access, Media Collection

1. CURATION CHALLENGES

The TIB – German National Library of Science and Technology – is the national subject library for all areas of engineering, architecture, chemistry, information technology, mathematics and physics. The library provides national and international research and industry with information regardless of the information's language or material type. It furthermore functions as a "library of last resort" for the specified subject areas and has a legal mandate for archiving. As an archival library, the TIB has dedicated staff and resources for digital preservation activities.

As part of the digital preservation activities, the TIB is currently analyzing it's holdings on removable data carriers. Here, the cataloguing practice of the past 30 years proves to be problematic when investigating exact numbers of items by carrier type per collection, due to the fact that any electronic source was often described simply as "electronic media" in the catalogue, thus lacking a distinction between, e.g., CD-ROM, CD-R, DVD, floppy or online source. A first analysis of a few selected collections has brought forth estimates for optical data carriers (CD-ROM, CD-R, CD-RW, DVD) as summarized in Table 1. A sampling of the evaluated collections established that the content of the data carriers is often complex: often, a carrier may contain a combination of software needed to render the file as well as the digital object itself.

Table 1: TIB optical media inventory

Collections	Media Items
Supplements to Monographs	∼ 20.000
Patents, Rules and Standards	∼ 18.000
Conference Proceedings	∼ 14.000
Serials	∼ 13.000
Lexica, Dictionaries, Databases	∼ 300

The optical carrier is still a requested medium – the supplements to monographs, for example, which can be only viewed within the library's reading rooms, booked around 1.500 requests in 2013. A different statistic showed that in the month of January of 2014 alone, a total of 2.819 pages of inter-lending requests fulfilled by TIB were generated from information that the library held only on CD carriers.

In the light of deteriorating data carriers as well as hardware and software dependencies of the materials contained on the data carriers, preservation and continuous access strategies for this material type need to be developed and implemented. While bit preservation issues can be addressed by moving the information off of the original carrier – either in form of ISO9660 images and/or in form of bitwise copies – the logical preservation activities, especially for complex objects held on the data carriers, is a different matter. One potential way to preserve the accessibility of the content with a high level of authenticity and utility is by using emulation. To test the feasibility of this approach, the TIB is currently implementing a pilot workflow with the University of Freiburg based on the "Emulation-as-a-Service" infrastructure.

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 and based on the observations on potential stakeholders a scalable service model has been developed – Emulation as a Service (EaaS) [1]. 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, i.e. an emulator capable of replicating a certain system architecture, as an abstract component with a unified set of software

interfaces (API). This way, different emulators can be integrated and can be used in dedicated archival workflows. For this purpose, EaaS offers users various options to interact interactively with an emulated environment, e.g. through remote desktop protocols (VNC, RDP) or, more conveniently, through a HTML5-enabled web browser. Furthermore, emulation components can be dynamically deployed in a large-scale cluster or Cloud infrastructure upon request. Hence, no spare computing resources have to kept available. 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 requirements on necessary infrastructure.

2. WORKFLOWS

On the library side of the workflow, a number of pre-ingest steps need to be conducted. The legacy CD collection needs to be evaluated to be able to prioritize data carrier migration. This may include a number of factors, such as evaluating the uniqueness of the collection, the age of the data carrier or its risk of being damaged and copyright clearance. In a next step, the content of the carrier needs to be replicated, e.g. by creating an ISO9660 image. The TIB will ingest the images into its digital preservation system, which functions as a bit preservation layer and keeps necessary metadata.

To secure long-term access to the media's content, its information object is to be prepared by using EaaS ingest workflows. Through these workflows it is possible to create or to modify emulation environments, i.e. an emulated hardware system, an operating system and software required to render the digital object. In a second step a digital object then is linked to a specific environment that is able to render this specific object. While preparing a rendering environment process is optional (a ready-made standard environment could be used), linking an environment and a digital object results in technical meta-data with an exact description of the environment's view-path and its configuration such that a deterministic re-enactment of the system and object becomes possible. During ingest, the EaaS workflow allows to test-run the environment. This allows for an evaluation of the rendering quality and performance of object and environment. Fig. 1 shows an example of this process.

In many cases, pre-configured standard environments are not sufficient to render a digital object. A preliminary evaluation of a sample set of digital objects provided by the TIB showed that almost all of the tested objects require proprietary multimedia frameworks that were usually not included with the operating system. Therefore, before these objects can be used, these frameworks or respective viewer applications have to be installed.

For instance, the object shown in Fig. 1 is an interactive training program for Microsoft Excel set in a "futuristic virtual teaching room." This training program requires the Video for Windows multimedia framework and uses a proprietary viewer application to render the interactive content. In order to make this object accessible, a base system has been selected (Windows 98 SE) and additional software installation steps were performed using the "setup.exe" installer program provided on the object's CD-ROM. After these installation steps, the modified environment has been archived, creating a derivative of the base system specifically designed for this specific object. The derivative consists only of the actual changes to the installation medium (the system hard disk) together with a stable reference (HDL) to the original base image, both to save storage space and to allow for distributed data management, i.e. store derivatives together with the object in a single repository. Output of the EaaS ingest workflow is an *emulation environment* description, defining the EC configuration and referencing both the prepared rendering environment and digital object. In our prototypical implementation we make use of the Handle system as persistent identifiers for both environments and objects.

To access the prepared object, the *emulation environment* description is used by an EaaS provider to allocate and setup a suitable EC. External data-sources, i.e. a specific software environment and the digital object itself, are resolved and attached to the EC. The user is then able to connect to and interact with the environment using an HTML5-enabled web browser. This allows for instant rendering of complex objects by clicking on the corresponding link when browsing the library's web-based catalogue. On the end library user layer, the link to the emulated object in its respective environment will be linked from the library catalogue. Access modalities will follow the traditional CD workflow and will only be available from within the library's reading room.

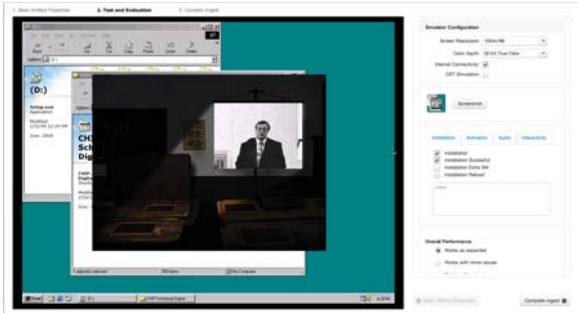

Figure 1: Assessment of a CD-Rom's content.

3. OUTLOOK

The current prototypical workflows address current curation challenges. Using EaaS provides novel access options to a significant part of the libraries electronic media collection. While the current state is not production ready yet, practical experience has been collected. The next steps, focus on further automation, e.g. automatically determine a suitable rendering environment for uniform object classes. For this, the sample will be significantly extended by TIB. On the library side, a workflow will be established to check the question of intellectual property rights for different subcollections in order to decide whether the EaaS for a specific group of works may only be offered within the reading room. Furthermore, the library plans to extend the EaaS workflow to the emulation of other data carriers, such as for example USB sticks, which are now starting to enter the collection especially in the area of congress and grey literature deposits.

4. REFERENCES

[1] T. Liebetraut, K. Rechert, I. Valizada, K. Meier, and D. von Suchodoloetz. Emulation-as-a-Service – The Past in the Cloud. In *7th IEEE International Conference on Cloud Computing (IEEE CLOUD)*, page to appear, 2014.