

Draft Report on the PERSIST-session at the 4th Annual Conference of the International Council of Archives (ICA) 'Evaluation of Strategies of Digital Preservation & UNESCO's Role in Facing the Technical Challenges', Girona, 13 October 2104

Introduction to PERSIST

PERSIST (*Platform to Enhance the Sustainability of the Information Society Transglobally*) is a collaborative project of UNESCO, IFLA, ICA and other partners to address globally pressing questions on preservation strategies, technique, selection, responsibility and division of labor. It works under the assumption that on these issues a high-level global policy discussion is needed between heritage institutions, industry and government, and that UNESCO's *Memory of the World Program* (MoW) is a unique platform to conduct that discussion.

The idea for PERSIST was born at the Conference *The Memory of the World in the Digital Age: Digitization and Preservation* in Vancouver (September 2012)¹. The Declaration adopted by its participants states that:

'...there is a pressing need to establish a roadmap proposing solutions, agreements and policies, that ensure long term access and trustworthy preservation. This roadmap should address issues like open government, open data, open access and electronic government. It should dovetail with national and international priorities and be in full agreement with human rights.'

PERSIST was launched as a project at the Conference *A Digital Roadmap for Long-Term Access to Digital Heritage* in The Hague in December 2013². In its initial stage the project is coordinated by the Netherlands National Commission for UNESCO and financed by the Netherlands Ministry for Education, Culture and Science.

The work for PERSIST is divided into three task forces: content, technology and policy. The technology task force is in this phase of the project focusing on the different strategies for digital preservation. On the one hand the strategies that give access to and preserve computing systems and applications by emulation and virtualization and on the other hand the strategies that help track the changes in file formats – and thus in digital content - by documentation, validation, testing, and migration. UNESCO tries to identify actions that it could undertake in support of them. The responsibility for the technology task force lies with the President of the ICA. At the Girona Conference, David Fricker (Director-General of the National Archives of Australia) has taken over from Martin Berendse, ICA's resigning president.

Overview of the Session

The position paper for the session 'Evaluation and Strategies of Digital Preservation & UNESCO's Role in facing the Technical Challenges'³ was written by Margriet van Gorsel, Michiel Leennaars, Natasa Milic-Frayling and Jonas Palm.

¹ Vancouver Declaration:
http://www.unesco.org/new/fileadmin/MULTIMEDIA/HQ/CI/CI/pdf/mow/unesco_abc_vancouver_declaration_en.pdf

² PERSIST website: <http://www.unesco.nl/digital-sustainability>

³ Link to the paper:
http://www.unesco.nl/sites/default/files/uploads/Comm_Info/20141011_evaluation_and_strategies_of_digital_preservation_unescos_role_in_facing_the_technical_challenges.pdf

The first outreach activity of the technology task force was organized within the framework of the ICA Annual Conference 'Archives and Cultural Industries' that took place from 11 to 15 October in Girona, Spain. The PERSIST session in that conference was devoted to preservation strategies. Under the chairmanship of **David Fricker**, PERSIST was introduced by **Martin Berendse**.

Iskra Panevska of UNESCO's *MoW Programme* introduced participants to the past and current work of UNESCO in the field of documentary heritage, mentioning particularly the *Recommendation on Preserving and Accessing the Documentary Heritage* that UNESCO hopes to adopt in 2015. PERSIST is an important initiative to implement the needs of the global heritage community as they were formulated in the Vancouver Declaration and to continue a holistic discussion on preservation issues with all stakeholders.

Natasa Milic-Frayling (Microsoft Research). Just as you need light to read text on a piece of paper, so you need software to see and interpret the information contained in a file. The technology in computing devices is very complex. It comprises many different parts, from the computing chips and operating system to the software tools and printer drivers. They are all inter-dependent and a product of dozens of computer businesses, large and small. All these companies are bent on innovation that meet user needs. Thus, to maintain computing systems and applications requires constant updates and investment on the part of the technology providers. If the demand changes, they cannot cover the cost of maintaining the technology because it is economically unsustainable. This, clearly, is detrimental to the long term access to content and data, particularly those that are held in databases and highly interactive multi-media applications.

Due to this close connection between content and software, it is important to focus on preservation of both the software and the document files. Preservation of software can be achieved by providing a computing environment in which the software can function. It is not enough just to save software files. It is critical to enable running of the software. In this way, we can access the original documents, particularly interactive media that cannot be easily migrated to alternative file formats. Furthermore, we can apply migration of documents on demand to any other format that is required for a specific use, without worrying about the possible loss of information since the original application and content are always at the user's reach. Thus, a holistic approach to digital preservation requires both, running of legacy software and migration of file formats when needed to be used with contemporary applications.

The preservation of computing technology can be achieved through emulation and virtualization. Milic-Frayling demonstrates a pragmatic approach where cloud-based virtual machines are used to run old versions of Microsoft Windows. That has a beneficial effect that all past applications can then be run on the virtual machines just as they were used in the past. She demonstrated the use of old version of Microsoft Word and Power Point and brought the digital multimedia encyclopedia Encarta back to life.

By the advent of cloud computing virtualization can be used to the benefit of the preservation of software because virtualization is a fundamental element of cloud computing. Cloud providers can offer infrastructure, platforms or software as a service. With regards to the management of cloud resources, there are different infrastructure models. They primarily vary based on the control and exclusivity that an organization or a user may want with regards to the infrastructure, the services and the communication protocols between the devices and the cloud. Typically, there is a trade-off

between the level of control and the investment required to uphold that control by the organization, considering the ever changing computing ecosystem. The control often includes the concerns related to the trustworthiness and stability of the infrastructure as well as the security, data protection, and privacy of the access and communication protocols. There are already a number of models ranging from public and private clouds to community clouds and hybrid clouds. Thus, it is now practical to devise and implement digital preservation strategies that take a holistic approach and focus on both software and files produced by digital technologies.

UNESCO can play an important role in stimulating digital preservation by mobilizing content holders and IT providers to set up a repository of legacy software that would be readily available to heritage institutions.

Jos van den Oever (KO GmbH, Ministry of the Interior and Kingdom Relations) presents the part of the position paper written by **Michiel Leenaars** (Stichting NLnet). A repository of legacy software is a real contribution to the sustainability problem. The recreation of the computational ecosystem in which a file functioned originally is good but it doesn't cover all preservation issues.

Modern software is often more complicated than it could be, because it is not designed from scratch, but builds on older layers of code. Developers typically live from patch to patch and see only part of the entire application. One is never sure whether a combination of a file of a certain type with a programme of a certain type will yield all the information the file contains or not. Understanding the original file format completely is essential. With knowledge of the original file format, the document can be migrated to a modern format as needed. Exhaustive documentation, maintenance of modern software that understands the old formats and systematic testing can bring these problems to light, and open the possibility to write converters: programmes that transform the content of a file in such a way that it will become usable in a different computational environment.

This approach is validated by the history of web browsers. In the past, tests of the behavior of web-content in different browsers have played a major role in increasing the consistent performance of browsers. This, in its turn, brought about the current healthy diversity of browser available to internet-users. The dominance of Microsoft's Internet Explorer is now checked by the advent of Safari, Firefox or Google Chrome. Similar tests for example for office and spreadsheets are needed, so that these files can be examined and used reliably in modern software that passes these tests.

Interoperability testing and the making of converters are excellent ways by which heritage institutions and other users can increase the command they have on their digital collections. UNESCO can help by supporting (morally and financially) organisations who can do this.

When a technology is new, standardization would go against the grain of innovation, but in a later stage it is possible and necessary. The typesetting system for mathematical formulas TEX shows clearly that continual change is not an inherent necessity of the digital world. Improving our current software practices so that future archivists can reliably work with files from the current era is essential to help the software industry mature. UNESCO should use its influence to stimulate standardization in areas where this is possible and needed. Institutional and private users of software should be steered towards focusing on standardized file formats with comprehensive test suites. Software is forever changing but the files that are created will remain with us and need to remain available to us in the software of the future.

Kuldar Aas (National Archives of Estonia)

TREASURES (Technical Registry Enabling Access to Services Used in Research E-infrastructures) is an idea for a technical registry for which a proposal has been made in the framework of the Horizon2020 program of the European Commission. Similar initiatives exist, but these are often too small in compass, covering typically only the technical needs of the participating organizations and the information they gather lack in depth and breadth: the links between the information collected is often insufficient and so is the information on hardware and peripherals.

TREASURES would like to capture and relate all aspects of technical information relating to digital preservation, including, but not necessarily limited to: formats, environments, applications, risks, plans, etc. The initiative is supported by a wide range of stakeholders; coordination lies with the University of Portsmouth (UPHEC) and the Austrian Institute of Technology (AIT)

The goal of TREASURES is threefold:

- support the digital preservation community in understanding, characterizing, validating, risk identification, and preservation of digital objects;
- provide a resource for organizations and individuals becoming involved in, or learning about, digital preservation;
- provide a comprehensive, , consolidated, accurate, and open information source that can be used in conjunction with any digital preservation repository.

The technical registry of TREASURES and the idea for a global repository of software that is currently developing in the PERSIST project could both gain by close cooperation.

Conclusions

- Virtualization of software provides access to original content that can be in form of complex and highly interactive multi-media documents.
- Standardization on file formats with unambiguous specifications and comprehensive interoperability tests will accelerate software diversity and reduce long term costs.
- The two strategies for digital preservation sustainable computation and interoperability are not mutually exclusive. For both strategies, a global repository for legacy software would be an important asset. UNESCO should support sustainable computation and interoperability testing by mobilizing the different stakeholders needed to install and maintain such a repository. In conjunction with the registry, UNESCO should likewise mobilize moral support (via its Logo) and financial support (via Memory of the World, IFAP or another donor) to the Document Liberation Project to facilitate the writing of test sets for important formats like ODF and OOXML and the writing of convertors.
- PERSIST should keep the connection with TREASURES that will be launched in the framework of the Horizon2020 program and should broaden the participation of the ICT industry community in the register project, as the registry should contain software from various companies.
- The PERSIST policy task force scheduled for April 2015 should convene high-level representatives of the relevant global consortia for heritage, government and ICT-industry together to decide on the creation of this registry and the organization interoperability testing that will be carried out on the files it will collect. In the intervening period, the technology task force should provide the policy taskforce with a detailed study of the modalities of the registry.