PROGRAMME

9-10 JUNE 2020

Contents

Thanks

This event would not have been possible without the OPFCON working group. We would like to extend a big thank you to each of you for your time and effort!

Barbara Sierman
David Cirella
Elizabeth Kata
Kati Sein
Peter May
Remco van Veenendaal
Sarah Mason
Yannick Grandcolas

—

OPFCON was originally planned to take place at the Austrian National Library in Vienna. We would like to thank the staff at the library for their generosity and support during the planning for this event, as well as their assistance in transitioning online.

—

We'd like to offer our thanks to those who have volunteered to act as session leads for our collaborative notes:

Sarah Mason
Kati Sein
Ross Spencer
David Underdown
Remco van Veenendaal
Elizabeth Kata

A big thank you also to our live tweeters:

Elizabeth Kata
Micky Lindlar
Sarah Mason
Remco van Veenendaal

Find us on twitter (@openpreserve) to follow along and join the discussion by using #OPFCON!

—

Finally, we would like to thank our sponsors for their generous support. Communications from each of our sponsors can be found in the virtual delegate bag.

Platinum

THE NATIONAL ARCHIVES

Gold

Preservica
Active digital preservation

Silver

ExLibris
a ProQuest Company

Bronze

dutch digital heritage network

Agenda

Tuesday 9 June | 13:00 - 17:00 CEST

13:00 **Introduction**
Becky McGuinness and Charlotte Armstrong (Open Preservation Foundation)

Session 1

13:10 **Welcome**
Martin Wrigley (Open Preservation Foundation)

13:20 **Keynote**
Dr. Adam Farquhar - The Open Preservation Foundation: a decade of digital preservation leadership

13:50 Dr. Monika Zarnitz (ZBW Leibniz Information-Centre for Economics) - That's what we are like: First results of a survey on the worldwide situation of digital preservation communities

14:10 *Break*

Session 2

14:30 Lina Maria Zangerl (Literaturarchiv Salzburg) - Digital Preservation for Everyone: The Personal Digital Archiving Website meindigitalesarchiv.de

14:45 Jon Tilbury and Jack O'Sullivan (Preservica) - Using the Preservation Action Registries API to deliver real user benefits

15:00 **Lightning talks**
Christopher (Cal) Lee & Kam Woods (University of North Carolina, Chapel Hill) - New Tools and Techniques to Analyze and Manage Email in Archives: The Review, Appraisal and Triage of Mail (RATOM) Project

David Cirella (Yale University Library) - Automating 'Inbox Zero': An approach to processing system-generated messages

15:25 *Break*

Session 3

15:40 Elizabeth Kata (International Atomic Energy Agency) - Non-conforming PDF/A Conformance? How to Embed Customized PDF/A-Compliant XMP Metadata into PDFs

15:55 Franz Heinzmann (Arso Project) - Sonar - A peer-to-peer based database for digital archives of media and content

16:10 **Poster session**
David Cirella (Yale University Library) - CD-ROM Preservation: Acquisition, validation, and access by way of proprietary file formats, legacy software, and language support

Felix Bach, (Karlsruhe Institute of Technology) - NFDI4Chem - research data management for chemistry

Elizabeth Kata (International Atomic Energy Agency) & David Clipsham (The National Archives UK) - PRONOM Research Week: What Happened, Where It's Going

Charlotte Armstrong (Open Preservation Foundation) - Revisiting the Planets Time Capsule

Jesse de Vos (Netherlands Institute of Sound and Vision) - AV-archiving at Sound and Vision: how to hit a moving target

Tomasz Parkola & Krystian Minta (Poznan Supercomputing & Networking Center) - Developing sustainable and fit for purpose archiving system

16:35 Martin Wrigley (Open Preservation Foundation) - Closing Remarks

16:45 John Sheridan (The National Archives UK) - Preserving the Present

Wednesday 10 June | 10:00 - 13:30 CEST

Session 4

10:00 **Keynote**
Barbara Sierman - The treasures of the Open Preservation Foundation

10:15 Richard Lehane (International Atomic Energy Agency) - Automated benchmarks for FIDO, DROID and siegfried

10:30 Anni Järvenpää (CSC – IT Center for Science): Tools for Creating and Validating SIPs - Reducing the Technological Know-How of Preparing Content for Digital Preservation

11:00 *Break*

Session 5

11:15 Stefan Szepe & Vitali Bodnar (University of Music and Performing Arts Vienna) - Institutional repository of the mdw - University of Music and Performing Arts Vienna

11:30 Leontien Talboom (University College London & The National Archives UK) - Accessing the intangible: the constraints faced by digital preservation practitioners when making born-digital material accessible

11:45 Thomas Ledoux & Yannick Grandcolas (Bibliothèque nationale de France) - Strengthening the French preservation community around formats and associated tools

12:10 *Break*

Session 6

12:30 **Lightning talks**
David Clipsham (The National Archives UK) - The future of PRONOM - a review

Jeffrey van der Hoeven (National Library of the Netherlands) - The Web Curator Tool renewed

Jochen Stärk (MustangProject) - Technical Insights into European E-invoices

12:50 **Roundtable: PLANETS: Back to the future**
Panellists: Adam Farquhar, Jacqueline Slats (National Archives of the Netherlands), Ross Spencer (Artefactual)

13:20 **Closing remarks**
Martin Wrigley (Open Preservation Foundation)

Abstracts and Biographies

Dr. Adam Farquhar
The Open Preservation Foundation: a decade of digital preservation leadership

Dr. Adam Farquhar focuses on digital transformations in library, research, and information sectors. He was founder and first Board Chair of the Open Preservation Foundation as well as Scientific Director of the PLANETS project that established the coalition and impetus to create the OPF. During his career at the British Library, he founded its digital preservation department, established its first research data programme and strategy, led its digital scholarship department and initiatives that connected people to digital collections and data including the British Library Labs, DataCite, the Endangered Archives Programme, the THOR and Living with Machines projects. He has also held responsibility for major collection areas at the British Library including newspapers, photographs, sound, moving image, and maps. With Dr Angela Dappert, Adam is now a partner at Digital Lifecycle Management LLP. This gives him an opportunity to contribute and share some of what he has learned.

Dr. Monika Zarnitz
That‘s what we are like: First results of a survey on the worldwide situation of digital preservation communities

Digital preservation networks developed all over the world from the early 2000s onwards. These networks overcame the traditional borders between libraries, archives, museums and even private enterprises. These different organizations began to work together on the new challenge „digital preservation“ of their digital content. Digital preservation turned out to be a demanding issue concerning the technical and organizational surroundings needed and it appeared early that one important requirement to solve these problems was „cooperation“.

For digital archives, networking is an essential building block in keeping up with the state of the art. While the benefits of networks are clear, not all digital preservation organizations were managed to sustain themselves and several communities had to wind down. Despite the fact that we are aware of this, no thorough overview of global digital preservation networks and communities exists. nestor has set out to close this gap via the community survey, which was addressed directly at the communities, asking them to described against factors such as membership structure, services and financing. These individual descriptions of the communities will be published on nestor’s website. In addition there will be an anonymized analysis of the whole data.

Each community has its own traits. They can be described based on their location - with regional, national or international networks – as well as based on the digital preservation topics a network covers. Some of them cover a wide range of subjects, others target a specific challenge within digital preservation. A third network category is that of distinct service-providing networks. There are many other means of categorizing these networks that can be used in our presentation.

Our study within nestor as one of the key players in the wider landscape itself, serves a threefold purpose:

1. to present a starting point for a closer national and international cooperation with these networks with the potential to support a bottom-up development of closer cooperation of networks by collecting information on these networks
2. to fulfill members' and the wider communities information needs regarding different network activities
3. for nestor it is a means of self-reflection, evaluating where the networks sits within the larger international context

This presentation will describe first results of the analysis of the data and the schemes for the descriptions of the different communities in the internet

Dr. Monika Zarnitz is head of the ZBW department for collection care where a group responsible for digital preservation of holdings of the ZBW is located. Mrs Zarnitz is active in working groups of nestor and is a reviewer for the CoreTrustSeal.

Lina Maria Zangerl (Literaturarchiv Salzburg)
Digital Preservation for Everyone: The Personal Digital Archiving Website meindigitalesarchiv.de

"Personal Digital Archiving" concerns fundamental aspects of responsibly managing one's own digital data. The heterogeneity of digital documents, available in many different formats and distributed via a variety of devices both on and offline, makes their preservation difficult. The inherent complexities of digital preservation – which affect us all in our private and professional lives – are actively being addressed within academic libraries.. However, libraries and cultural heritage institutions outside of academia are increasingly being seen as both partners and experts for questions regarding personal digital archiving in the public domain. This has been the case in the international environment for some time, and more recently also in the German-speaking world. In response, "nestor" – a network of expertise in long-term storage of digital resources – established the "Personal Digital Archiving Working Group" which includes library and archive staff alongside other information experts from Germany and Austria.

Since its formation, the group has sought to present professional insights concerning long-term preservation of personal digital materials in a way that is understandable and achievable for the broader public. Since January 2020, the results of this work are freely available online at meindigitalesarchiv.de. Using a narrative structure, the website highlights fictitious "personas" who have everyday questions and concerns about the use and protection of their digital materials. The characters are then offered practical solutions appropriate to their individual circumstances. Through individual stories, visitors to the website are able to learn about the importance of preserving their digital life based on highly relatable experiences. Embedded "Infos and Tools" offer further opportunity to expand one's knowledge of personal digital archiving and inform visitors about available options, such as cloud storage. My presentation will include interactive elements that reveal the effectiveness of using such method of instruction. In addition, I will show how the information gathered on the website can be utilized to promote and facilitate the exchange of knowledge between curators and information seekers.

Lina Maria Zangerl holds an MA in German Studies from the University of Salzburg and an MA in Library and Information Sciences from Humboldt University in Berlin. From

2010 to 2014 she was archivist for the Salzburg Festival. She has been the archivist at the Literaturarchiv Salzburg since 2015, where she currently focuses on the digital humanities project www.stefanzweig.digital. Her research interests include archival theory, indexing methods and digital literary estates. She is part of Nestor's "Personal Digital Archiving Working Group".

Jon Tilbury, Jack O'Sullivan (Preservica)
Using the Preservation Action Registries API to deliver real user benefits

This presentation will show how the Preservation Action Registries initiative can be implemented to deliver real benefits to real users across the world. It shows how the protocol can lead to information sharing in order to build trusted Preservation Policy best practice and how users can subscribe to this best practice such that this will be automatically enacted on real data. We will cover file identification, property extraction, validation, and migration and show how these activities can be controlled centrally and processed automatically. We will also show how rule changes can be applied retrospectively, so the system reacts to updates and makes sure it is always up to date. We will also explore whether this will be sufficient to open Digital Preservation to a whole new audience.

Jon Tilbury is the founder and CTO of Preservica. He has been working in Digital Preservation in many roles for 20 years.

Jack O'Sullivan is a Senior Engineer in the Innovation Team at Preservica and is member of the PREMIS editorial board. He is a critical part of the team of delivering new Digital Preservation approaches in practical situations.

Christopher (Cal) Lee & Kam Woods (University of North Carolina, Chapel Hil)
New Tools and Techniques to Analyze and Manage Email in Archives: The Review, Appraisal and Triage of Mail (RATOM) Project

The Review, Appraisal, and Triage of Mail (RATOM) project is a two-year effort funded by the Andrew W. Mellon Foundation to develop and test software to assist in the analysis, selection, and appraisal of email held in collecting institutions.

In this talk we present open source software tools developed for the RATOM project to assist archivists and other data management professionals analyzing legacy email collections stored in PST, OST, and mbox formats. Supported tasks include identifying and reporting on entities present within emails and email attachments; identifying materials requiring redaction or review due to the presence of potentially sensitive information; and modules to assist with preparation of materials for release or public access.

Our integrated Python 3 library and API (libratom - https://github.com/libratom/libratom) uses existing open source toolkits to parse content and metadata from PST and mbox sources at scale, along with a production-quality NLP toolkit (spaCy) to identify entities such as people, places, and organizations within message content. Our framework and user-facing tools are designed to make processing large collections simple and efficient. The tools use multiprocessing to leverage the power of modern multi-core systems and minimize the

time required to process large collections. Pre-trained language models provided by spaCy currently support processing collections in English, German, French, Spanish, Portuguese, Italian, Dutch, Greek, Norwegian, and Lithuanian. Results are presented as a sqlite3 database that may be queried manually or by additional tools in an automated workflow. The schema focuses on initial triage and assessment, providing a map of 18 unique entity types to their locations (message and file) within a corpus, along with file metadata for attachments discovered in a collection.

Our processing interface consists of a web application that allows processing archivists to browse email sources and mark messages as suitable for retention. The UI provides faceted search, allowing human experts to narrow the scope of the materials being examined and annotate specific features within a message that are correlated with decisions regarding retention or sensitivity. We expect this tool to be used as part of an iterative processing workflow, in which the annotations produced by processing archivists are used to train a machine learning model to automatically predict classifications of future messages. The processing archivist may subsequently view these classifications as part of the faceted search, marking them as correct or incorrect for a future training cycle.

Christopher (Cal) Lee is a Professor at the School of Information and Library Science at the University of North Carolina, Chapel Hill. He teaches archival administration; records management; digital curation; understanding information technology for managing digital collections; and digital forensics.

Kam Woods is a Research Scientist UNC SILS. He previously served as technical lead for the BitCurator projects, developing techniques and tools to assist in analyzing and providing access to born-digital materials.

David Cirella (Yale University Library)
Automating 'Inbox Zero': An approach to processing system-generated messages

How do you get to 'Inbox Zero' when your systems are sending thousands of notification messages a day? The lightning talk will detail our approach for checking 10,000s of system-generated messages from our digital preservation system. Written in Python, our solution leverages APIs from our third-party ticketing system and digital preservation system to process messages, download objects, and check fixity. This solution allows for the automated and exhaustive verification of system messages, with minimal human interaction.

David Cirella is a Digital Preservation Librarian at Yale University Library. In this role he works with stakeholders from around the institution towards the long-term preservation of their digital content.

Elizabeth Kata (International Atomic Energy Agency)
Non-conforming PDF/A Conformance? How to Embed Customized PDF/A-Compliant XMP Metadata into PDFs

PDF/A is considered an archival standard and widely used. Conformance to the various levels of PDF/A standard can be validated using VeraPDF, a tool maintained by the veraPDF consortium, including the OPF, the PDF Association and the Digital Preservation Coalition. The PDF/A standard allows for the embedding of metadata using specific defined XMP namespaces, such as Dublin Core (dc). It allows for custom XMP

Extension Schemas, which can be useful when metadata requirements expand beyond the pre-defined properties. In this demonstration, I will show how to define an XMP-Extension Schema and embed it into PDFs using ExifTool. I will then show how you can bulk update metadata, including to newly defined extension schema tags, from CSV files using ExifTool. Lastly, I will validate updated files with VeraPDF. I will provide examples that do and do not pass PDF/A conformance, going through some basic troubleshooting of how to identify and solve issues. After correcting the issues, I will validate again with VeraPDF, showing that it is possible to create and add custom XMP-Extension Schema while maintaining PDF/A-compliance.

Elizabeth Kata is a Digital Archives Assistant at the International Atomic Energy Agency in Vienna, Austria. She has a Master's degree in Historical Research, Historical Auxiliary Sciences, and Archival Science from the Austrian Institute of Historical Research at the University of Vienna. Her interests include technical metadata and its potential for linked archival descriptions and 20th and 21st century diplomatics.

Franz Heinzmann (Arso Project)
Sonar - A peer-to-peer based database for digital archives of media and content

We'll present Sonar, an in-development toolset for digital media archives. With Sonar we're building a foundation for decentralized, yet easily explorable and searchable archives of media and other content. At its core, Sonar is a peer-to-peer database and search engine. Sonar is based on the Dat protocol for the peer-to-peer exchange of data (think BitTorrent and Git being merged) and has a built-in full text search engine. It's very easy to run both on desktop computers, laptops and servers. We provide a simple interface to upload media and also have an easy API to write custom importers. The peer-to-peer base allows to easily mirror archives from device to device. The lean, modern software stack is an interesting alternative to the current heavy stacks usually used in digital preservation settings. The current intention is not to replace such workflows, but augment them and make them accessibly to grassroots organizations, low-resourced groups and initiatives and end users. We'll do a quick presentation of the state of things, where we're coming from and heading to, and do a quick demo of the live replication of archival data possible with Sonar.

Franz Heinzmann is an open source software developer with a strong focus on peer-to-peer technology. He studies Political Science originally, but then took a turn towards solving practical problems for the technical foundations for the open preservation of history. He co-founded arso, a project to research and build tools for the digital preservation of media archives.

David Cirella (Yale University Library)
CD-ROM Preservation: Acquisition, validation, and access by way of proprietary file formats, legacy software, and language support

Digital preservation requires that we take actions to account for the acquisition, validation, and method of future access to those objects we seek to preserve. A preservationist's interaction with an item most often begins mid-life cycle and involves employing the best of our tools and practices to mitigate the issues caused by the technical decisions made around the object's production. The chosen file format, required operating environment, and choice of physical media contribute to the circumstances that must be actively engaged with to avoid the item's end of life. The

presentation will present a case study of the preservation of multi-disk CD-ROM sets of historical newspapers that were processed by the Digital Preservation Unit for long-term preservation.

Our approaches and actions at each step of the preservation process, and the analysis that informed them, will be explored with a focus on the techniques employed at each stage. Necessary to the successful preservation of this collection, issues around proprietary file formats, legacy software and hardware, and language support will be discussed, followed by an exploration of the various methods employed to move through the step-by-step sequence of actions. Bit level preservation media imaging techniques, validation via the creation and configuration of virtualized computing environments, and access via our emulation service will be explored in detail. A presentation on this project was given at the NDSA Digital Preservation 2019 conference.

David Cirella is a Digital Preservation Librarian at Yale University Library. In this role he works with stakeholders from around the institution towards the long-term preservation of their digital content.

Felix Bach (Karlsruhe Institute of Technology)
NFDI4Chem - research data management for chemistry

NFDI4Chem is a grassroots movement of researchers, infrastructure experts and actors of learned societies from all disciplines of chemistry. The vision of NFDI4Chem is the application of digitisation principles to all key steps of research in chemistry. In the initial phase, NFDI4Chem focuses on molecules and data for their characterisation and reactions, both experimental and theoretical.

Felix Bach has been a researcher at KIT since 2010. His topics are data science, data analytics and research data management.

Elizabeth Kata (International Atomic Energy Agency) & David Clipsham (The National Archives UK)
PRONOM Research Week: What Happened, Where It's Going

This poster will present information on PRONOM Research Week 2019 as well as plans for the next PRONOM Research Week. The poster will quantify the contribution of the Research Week to PRONOM Release v96 and detail what will can be expected for the upcoming Research Week.

Elizabeth Kata is a Digital Archives Assistant at the IAEA. She has a Master's degree in Archival Science from the Austrian Institute of Historical Research. Her interests include technical metadata and its potential for linked archival descriptions.

David Clipsham is the Digital Archiving Technical Architect at The National Archives in the UK, and is product lead for both PRONOM and the DROID file format identification utility. His interests include the 80s era of home computing and the impact of emerging technologies on the nature and approaches to digital preservation.

Charlotte Armstrong (Open Preservation Foundation)
Revisiting the Planets Time Capsule

What does a JPEG photograph, a message in Java source code, a short film in .MOV format, a web-page in HTML, and a brochure in PDF have in common? Ten years ago, the Planets project joined forces with the high-security data storage vault, Swiss Fort Knox (SFK) to deposit a TimeCapsule containing these items. A decade on, how many of these items are still accessible and usable? Has the digital universe really expanded at the rate predicted? What became of the TimeCapsule team? This poster will provide answers to these mysteries and more...

Charlotte Armstrong is Project Officer at the Open Preservation Foundation. She manages the office, coordinates our outreach and community activities, and supports the Foundation's internal and external projects.

Jesse de Vos (Netherlands Institute of Sound and Vision)
AV-archiving at Sound and Vision: how to hit a moving target
The Netherlands Institute for sound and vision, as a national archive for AV-materials and a museum for media, continuously works to bridge the gap between the fast-moving world of media production and technology on the one side and the marathon that is digital preservation on the other. In this poster some highlights of Sound and Vision's preservation efforts and plenty of pointers to other sources of information for those who want to know more.

Jesse de Vos is a product manager at the Netherlands Institute for Sound and Vision. He is responsible for developing preservation strategies for complex digital objects in collaboration with other heritage institutions. He also works on projects using linked data to enrich vocabularies and collections.

Tomasz Parkola & Krystian Minta (Poznan Supercomputing & Networking Center)
Developing sustainable and fit for purpose archiving system

Poznań Supercomputing and Networking Centre (PSNC) is an ICT-focused R&D institution located in Poznań, Poland. One of its developments is a "Digitise and Go" toolset (DInGO) composed of online presentation platform (dLibra), digitisation pipeline management system (dLab) as well as long-term preservation system (dArceo). In 2019 DInGO toolset reached more that 140 deployments in Poland and abroad with many off-the-shelf solutions as well as several dedicated ones. DInGO has been developed to support cultural heritage institutions (CHIs) with their daily activities related to (large-scale and distributed) digitsation. One of its components, dArceo, is a long-term preservation tool for graphic, textual and audiovisual content. dArceo is a simple, yet powerful tool to create and store/retrieve archival packages aligned with OAIS model.

The archiving package itself is a composition of master files, descriptive metadata, technical metadata, administrative as well as life-cycle related ones. dArceo uses METS, MIX, PREMIS and alike standards to compose an archival package and then submit it to a storage system. To extract this information it uses state of the art tools such as FITS or ffmpeg. Initially, dArceo has been equipped with many features, including conversion and advanced delivery, SPARQL endpoint and alike. During 10-years operational and maintenance activities it started to be more and more visible that there is a crucial set of features expected by CHIs and other stakeholders. The relevant set of features is focused on proper archival package and automated pipelines in the digitisation workflow. Therefore lately, PSNC has been working on simplification, fit for purpose and low-cost

maintenance of the system. In the presentation we would like to share our findings and express conclusions, hoping to get relevant feedback from the audience.

Tomasz Parkoła is the Head of Digital Libraries and Knowledge Platforms Department at Poznań Supercomputing and Networking Center (Poznań, Poland) where he manages research & development teams responsible for digital humanities infrastructure (http://ehum.psnc.pl/en/main-page/), products and services for digital libraries (https://dingo.psnc.pl/) as well as Polish Federation of Digital Libraries (https://fbc.pionier.net.pl/). He was involved in national and international research and development projects with main themes on data access & processing, long-term preservation, digitization workflows as well as data aggregation & interoperability. He was a national coordinator for the EIFL-FOSS programme (2010-2013). Since 2011 he is a board member in the IMPACT Centre of Competence and since 2016 he acts as a scientific and technology director. He was programme committee member for iPRES 2014, iPRES 2016, DATeCH 2017 and DARIAH AE 2019. He is an author or co-author of several dozens of scientific and popular science publications. Since 2019 he acts as a Product Board member of the Open Preservation Foundation. He also has PMP and UX-PM certificates.

Krystian Minta is a graduate of computer science at the Poznań University of Technology. Sine 2018 works at Poznań Supercomputing and Networking Center in the Department of Digital Libraries and Knowledge Platforms. One of the software developers responsible for the development of the DInGO toolset, including dArceo system for long-term preservation. He was involved in the development of the data conversion and ingestion pipelines in the context of digitisation workflows, preservation routines as well as online access to digital resources.

John Sheridan (The National Archives, UK)
Preserving the Present

John Sheridan is the Digital Director at The National Archives, with overall responsibility for its work as a digital archive. He also serves on the Executive Board of Digital Preservation Coalition and the Executive Committee of the DLM Forum. John is a long standing civil servant, first joining the Cabinet Office in 2004. A former co-chair of the W3C e-Government Interest Group, he has a strong interest in web and data standards. John also has a strong background in legal informatics and legal information systems, leading the team that developed legislation.gov.uk and, as Principal Investigator, an Arts and Humanities Research Council funded project, 'big data for law', which explored the application of data analytics to the statute book. John's academic background is in mathematics and information technology, with a degree in Mathematics and Computer Science from the University of Southampton and a Master's Degree in Information Technology from the University of Liverpool.

Barbara Sierman
Keynote: The treasures of the Open Preservation Foundation

Barbara Sierman worked from 2005 – 2020 as digital preservation manager at the Research Department of the KB National Library of the Netherlands. In this position she advised the KB on all matters related to digital preservation. Over the years she participated in European projects Planets, SCAPE, APARSEN and KEEPS. She published

several articles about a wide range of topics with an emphasis on preservation policies, OAIS and audit and certification. She was a co-author of the ISO 16363 Certification of Trustworthy Digital Repositories. From 2015-2020 she joined the Board of the Open Preservation Foundation and was Chair of the Board for 3 years. She was also a member of the Steering Committee of the International Internet Preservation Consortium and chaired the User Engagement Group. She is also actively participating in initiatives related to the Dutch Digital Heritage Network, and was vice chair of the IPRES 2019 conference. Recently she was involved in the launch of the TRUST principles.

In 2018 she was awarded the Digital Preservation Coalition Fellowship for her contribution to the field of digital preservation. After her retirement in June 2020 she will continue working as a consultant in digital preservation at DigitalPreservation.nl and will continue there writing her blog posts on www.digitalpreservation.nl

---

Richard Lehane (International Atomic Energy Agency)
Automated benchmarks for FIDO, DROID and siegfried
One of the tools that supports my development of siegfried is its automated benchmarks (https://www.itforarchivists.com/siegfried/benchmarks).

Every time I push a new release of siegfried to Github, a Travis CI job runs. This job purchases a server with Packet.com (a cloud infrastructure provider); installs tools on that server; copies a set of corpora – including the OPF's GovDocs Selected corpus - from Backblaze.com (a cloud storage provider); runs a series of tests; and, finally, posts the results to my website, itforarchivists.com.

The benefits of this approach to benchmarking are that it is cheap, transparent, reproducible, and, most importantly, doesn't require *me* to do anything (so I can spend that time working on siegfried).

The goal of this talk will be to explain how siegfried's automated benchmarking works in order to share what I've learned through this process and, hopefully, to inspire other open source digital preservation projects to try out similar approaches.

Richard Lehane is an archivist with the International Atomic Energy Agency in Vienna. In his part time he develops siegfried, a file format identification tool that implements various signature libraries including PRONOM, the MIME-Info spec and the Library of Congress's FDDs.

Anni Järvenpää (CSC – IT Center for Science)
Tools for Creating and Validating SIPs - Reducing the Technological Know-How of Preparing Content for Digital Preservation

The Finnish National Digital Preservation Services provide preservation services for the cultural heritage and research data sectors. The services are owned by the Ministry of Education and Culture of Finland, and both managed and developed by CSC – IT Center for Science Ltd. Currently, we have preserved more than 1.4 million AIPs amounting to more than 600 terabytes.

Having detailed specifications for both file formats and metadata is a precondition for a fully automated ingest process. However, preparing and ingesting digital assets in an appropriate format according to our requirements can be a demanding task, as the

process can be very time consuming and requires detailed know-how of metadata formats. The growing demand for making the preparation process easier for our partner organizations is the reason we have developed tools to decrease the burden of creating valid SIPs from scratch. In this presentation, we introduce two of our tools.

Firstly, the Pre-Ingest Tool, which aids in creating SIPs programmatically. It is a set of modular software components that produces a METS document containing all the necessary metadata conforming to our national preservation specifications. It can automatically extract technical metadata from files into the PREMIS metadata format and digitally sign the SIP. The pre-ingest tool is in production in several of our partner organizations, helping them integrate their back-end systems with the National Digital Preservation Service. When new partner organizations deploy our digital preservation service they almost invariably integrate the tool to their own systems, as opposed to creating their own solutions.

Secondly, the File Scraper can identify files, collect metadata from them and check their well-formedness. The tool uses third party software to validate and extract metadata from files and then normalizes the results into a uniform structure. It produces a python dictionary of technical metadata, a python dictionary about the used software including their outputs, and the validation result as a boolean value. The tool can also be used without validation in order to just identify the file and collect the technical metadata. The tool uses for example FFMpeg, FIDO, file, GhostScript, ImageMagick, JHOVE, LibreOffice, MediaInfo, Pillow, pngcheck, v.Nu, veraPDF, and warc-tools to validate files and extract metadata from them.

Both tools are available under LGPLv3 license at GitHub among other our tools and libraries (https://github.com/Digital-Preservation-Finland). Although these tools are developed specifically for our digital preservation service, we firmly believe that these tools can beneficial for digital preservation community in general.

Anni Järvenpää's main responsibilities include application level design and implementation of digital preservation services.

Stefan Szepe & Vitali Bodnar (University of Music and Performing Arts Vienna)
Institutional repository of the mdw - University of Music and Performing Arts Vienna

The mdwRepository is a registered institutional repository of the mdw – University of Music and Performing Arts Vienna, one of the most prestigious universities of the arts in the world (ranked No. 1 by QS Ranking 2019).

Data Management Policies at the mdw - University of Music and Performing Arts Vienna
Managing the content of the mdwMediathek and other university assets were the driving forces to implement a repository at the mdw. Soon digital collections in the areas of historical musicology, musical instruments and other research data extended the content scope. Technical infrastructure and policies have been developed in parallel. The mdw was among the first Austrian universities who adopted its Data Management Policies in December 2017. The key points of the policies are empowerment of the researchers as owner of their data and embedding of FAIR principles.

Technical features of the mdwRepository
The mdwRepository consists of a content-agnostic internal backend and public touchpoints that are operated in-house. Its core part is the digital asset management

system (DAMS), nuxeo, integrated with python based frameworks. The mdw focus on leveraging open source software from

- internal databases (mainly PostgreSQL),
- ElasticSearch Cluster with Kibana,
- Apache Jena Fuseki and Blazegraph for triple store and
- perl based OAI-PMH endpoint for data publication and distribution.
- All services undergo daily backups / snapshots and weekly data accuracy checks (based on checksums for files).

Case study: Archives of the Film Academy Vienna
The Film Academy Vienna (mdw's Department of Film and Television) is Austria's only place of university-level training for film professionals. 2015 the Film Academy Vienna (FAV) started to analyse its archives in order to store and to preserve their contents sustainably. Most of the analog audio & video works of the FAV are stored at the premises of the Academy itself, as well as in the archives of the Austrian Film Museum and the Film Archive Austria while some works remained in private collections. Since 2015 the works from these locations have been registered in a central database. It is the first archiving solution for video material enabling cooperation of the involved film & video archives within a single database and even enables co-working on the same film records. ICA's Archival Standard ISAD(G) is used for managing the content (from fonds to item). Additionally, the FAV has developed a metadata set with the partner institutions that enables a detailed description of the works and their manifestations on technical, legal, and film-study levels. As the database grows, searching and analysing the FAV's archival content is becoming easier.
Next steps, challenges

- CoreTrustSeal certification (by mid-2020)
- establishing a middleware and APIs that enable easier data presentation for projects
- data publishing, clearing of rights and legal status of the data
- optimized data management planning

Stefan Szepe is Digital Asset Management and AV-Production at the mdw – University of Music and Performing Arts Vienna; studied communication science, political science and history of arts at the University of Salzburg; as the digital asset manager (DAM) at the mdw Stefan acts as listener & communicator, problem solver, and tech solution provider overseeing all DAM related projects of the university and implementing the local institutional repository on OpenSource technologies (mdwRepository) since 2014; prior to joining mdw worked as consultant for Linz based software vendor celum gmbh.

Vitali Bodnar is Project Officer in the Research Support Unit of the mdw - University of Music and Performing Arts Vienna since 2011. Vitali is the first point of contact for general data management planning in research projects and recently co-ordinated development of the mdw's Data Management Policies. He holds a Diploma in German Studies and a Master of Advanced Studies European Integration.

Leontien Talboom (University College London & The National Archives UK)
Accessing the intangible: the constraints faced by digital preservation practitioners when making born-digital material accessible

This paper will outline the initial theoretical framework from a collaborative doctoral project with University College London and The National Archives which focuses on the

constraints that digital preservation practitioners face when making born-digital material accessible. The framework has been created with the help of semi-structured interviews with a wide range of digital preservation practitioners from different institutions and an extensive literature review. The main topics discussed during this paper will be:

The Digital Environment – This archival material is no longer only being made available in a physical space, such as a reading room, but a lot of archives and other memory institutions are opening up their collections in a digital environment. This has led to material being available at any time, without the physical constraint of going to these institutions, however it has also opened many new questions surrounding copyright, but also the community that this material is made accessible to.

The material itself – Digital material is not only human readable, but also machine readable. This opens up new opportunities for memory institutions, such as the use of artificial intelligence to help create or enhance associated metadata. But this also opens up questions surrounding how these computational methods should be applied in archives and what an ethical approach is for these methods. Also, questions arise around what should be preserved to make this material accessible, not only the material itself, but the contextualisation associated with it.

The processes used to make this material accessible – Processes surrounding the acquisitions, archiving and making accessible this material have been around for a few decades now, one of the most popular models being the OAIS model. Although, archivists have slowly gotten to grips with preserving this material, providing access still seems to be problematic. With a focus mainly on archiving, what does this mean for future accessibility and re-use of this material?

Leontien Talboom is a PhD student on a collaborative project with University College London and the National Archives in the UK. Her thesis is looking at the constraints that digital preservation practitioners are facing when making born-digital material accessible.

Thomas Ledoux & Yannick Grandcolas (Bibliothèque nationale de France)

Strengthening the French preservation community around formats and associated tools

In September 2019, within ARISTOTE (Association de Réseaux Interconnectés en Systèmes Totalement Ouverts et Très Elaborés / Association of interconnected networks in open source and highly developed systems), the “Format Watch Unit” has been established with the aim of pooling know-how and expertise on file formats between institutions.

Its ambition is to produce concrete, disseminable and reusable outputs in order to serve as decision support in the implementation of policies for the preservation of national establishments and the territorial network of public archives, and in order to raise awareness among professionals in digital archiving.

It aims to contribute to the knowledge and improvement of existing tools in terms of sustainability. It wishes to become an international interlocutor through its participation

in adequate bodies and initiate the translation of documents on sustainability, useful to professionals in digital archiving.

In order to proceed with the development of the community, internal regulations have been written and sub-groups have been created since :

- Experts directory
- Knowledge of formats
- Tools and corpus

We will present the progress of this network and how we use the expertise gained through OPF workshops and exchanges to accelerate the sharing of knowledge.

Last but not least, we considered how we will expose OPF assets (like the COPTR wiki) to a more international community in overcoming the cultural barriers.

Thomas Ledoux has a coordination mandate at the Information Technology Department of the Bibliothèque nationale de France.

Yannick Grandcolas is a curator and expert in digital preservation at the Department of Conservation of the Bibliothèque nationale de France.

David Clipsham (The National Archives UK)
The future of PRONOM - a review
In November 2019, The National Archives began work on the next version of PRONOM, the popular technical registry of file format information that underpins file format identification tools such as DROID, FIDO and Siegfried. In January 2020, a virtual workshop was held in which PRONOM users from around the world contributed a set of functional requirements to help drive the direction of this development work. This presentation will review the work completed to date and will discuss how PRONOM users can continue to contribute to the future success of the registry.

David Clipsham is the Digital Archiving Technical Architect at The National Archives in the United Kingdom, and is product lead for both PRONOM and the DROID file format identification utility. His interests include the 80s era of home computing and the impact of emerging technologies on the nature and approaches to digital preservation. David is also an OPF board member.

Jeffrey van der Hoeven & Trienka Rohrbach (National Library of the Netherlands)
The Web Curator Tool renewed

As online presence is indispensable in our volatile world, web archives are an invaluable source of factual information about what was online at a certain moment. As online content is short lived it is important to gain the highest quality when the site is crawled or days thereafter before the content has changed or disappeared. It is therefore that work on the Web Curator Tool (WCT) is currently focused on the area of quality management.

The WCT is an open source workflow management tool (available since 2006) for selecting, crawling websites, performing QA and preparing websites for ingest. Through close collaboration between the National Library of New Zealand (NLNZ) and the National Library of the Netherlands (KBNL) the WCT has already undergone several

uplifts in the past two years. Version 2 added support for Heritrix 3 and improved the project documentation adding new tutorials, installation and administration guides. Version 3 addressed a large volume of technical debt, in which the underpinning frameworks were upgraded, creating a stable foundation to take the development of WCT forward. With that work done, the project team began to focus on functional enhancements desired not only by the KBNL and NLNZ, but also by the wider community.

The project team demonstrated the new versions of WCT and asked for feedback on the features other organisations would like to see incorporated. The Hungarian National Library in particular contributed many enhancement ideas, several related to improving the WCT's quality control features. These dovetailed nicely with the WCT QA improvements already planned for the next release which will cover three core areas:

- Crawl patching using Webrecorder

Integrate Webrecorder into the WCT QA workflow by making use of its new "patch" capability. This will add the ability to repair missing content in addition to the existing WCT import and prune functionality. The integration will transfer newly patched content back into the WCT, incorporating it into the original web harvest.

- Screenshot generation

The WCT previously contained limited functionality to capture screenshots of a web harvest. Realising the potential QA benefit, we are enhancing this tool to capture screenshots of live websites being crawled and the resulting web harvest for comparison. The integration of the screenshot software will be configurable, allowing for the use of 3rd party tools. We also intend to leverage the advancements in screenshot comparison metrics within the web archiving community.

- Integration with Pywb viewer

The WCT already provides a WARC viewer and OpenWayback integration to browse harvests, but recently Pywb has become the benchmark for web archive replay. This integration will provide WCT users with the best options available for web harvest replay and review.

The KBNL is currently performing a retrospective assessment of their web archives using the WCT that will identify additional QA features to include."

Jeffrey van der Hoeven is head of the Digital Preservation department of KBNL and is in this role responsible for the active preservation of digital collections at the library. He has been involved with various national and international projects in the field of digital preservation, science data infrastructures and research.

Trienka Rohrbach is service administrator of the WCT at KBNL and works closely together with our colleagues in New Zealand. She is responsible for the day to day operation of the WCT in our production environment, selectively crawling thousands of websites a year.

Jochen Stärk (MustangProject)
Technical Insights into European E-invoices

Recent legal changes in the B2G sector (Directive 2014/55EU and the EN16931 standard) include the optional use of structured electronic invoices vis á vis European authorities. Some nations, like Germany, even abolished paper invoices.

This talk will distinguish between structured and unstructured e-invoices and present the two eligible XML formats UN/CEFACT and UBL for structued invoices. Within the strutured invoices it will also tackle hybrid e-invoices: hybrid standards like ZUGFeRD/Factur-X embed XML files into PDF files.

For ZUGFeRD/Factur-X the key role of OPF's VeraPDF will also be reflected.

All invoices are subject to short term retention (6-10 years) but this talk proposes synergies of e-invoice retention to long term preservation, e.g. why the use of PDF/A may solve PDF issues for hybrid (and unstructured) invoices, why, how and when a validation by the sender and by the recipient can be beneficial for the sender and recipient and how hybrid invoices might be used for a better metadata management of unstructured and scanned paper invoices.

Jochen is a bore and hobby bureaucrat and failed his way into electronic invoices since 2014. In his spare time he publishes open-source software for electronic invoices like mustangproject.org and https://github.com/ZUGFeRD/ZUV/. He likes to call himself "Mustangproject.org Chief ZUGFeRD Amatuer" and is in no way qualified for any speaking engagement whatsoever.

Adam Farquhar, Jaqueline Slats (National Archives of the Netherlands), Ross Spencer (Artefactual)
Roundtable: PLANETS: Back to the future

Dr. Adam Farquhar focuses on digital transformations in library, research, and information sectors. He was founder and first Board Chair of the Open Preservation Foundation as well as Scientific Director of the PLANETS project that established the coalition and impetus to create the OPF. During his career at the British Library, he founded its digital preservation department, established its first research data programme and strategy, led its digital scholarship department and initiatives that connected people to digital collections and data including the British Library Labs, DataCite, the Endangered Archives Programme, the THOR and Living with Machines projects. He has also held responsibility for major collection areas at the British Library including newspapers, photographs, sound, and moving image. With Dr Angela Dappert, Adam is now a partner at Digital Lifecycle Management LLP. This gives him an opportunity to contribute and share some of what he has learned.

After studying information & communication technology, **Jacqueline Slats** worked for the computer center of the Ministry of Watermanagement for 7 years. In 1994 she joined the then State Archives, where she was responsible for various automation projects and collaborated in the Functional Design of the Digital Depot.

From 2000 to 2004 she was the program manager of the Digital Preservation Testbed at ICTU. In addition, in 2002 she was responsible for the program management of the Taskforce Digital Preservation.

When she returned to the National Archives, she became head of the new Digital Preservation department. In addition, she fulfilled the role of program manager of the

strategic e-projects within the National Archives. In the European collaboration project Planets, she was the work package leader of Testbed Management and Execution of Tests and Experiments. After the Planets project, she became one of the founding directors of the Open Planets Foundation, later the Open Preservation Foundation. She was the quartermaster for the development of the e-Depot at the National Archives. She is now head of the Infrastructure and Services department.

Ross Spencer is a Software Developer at Artefactual Systems Inc. and has previously worked at the national archives of the United Kingdom and New Zealand. Ross has been an open source advocate for as long has he worked in archives releasing all of his code in the open. Ross believes in being open source by default. He remains staunchly committed to that cause and democratizing the work being done in technology across the GLAM sector.

SPONSORED BY

WWW.OPENPRESERVATION.ORG
@OPENPRESERVE