iPRES 2017 Kyoto

Collaborative Notes

Conference notes by Micky Lindlar, Joshua Ng, William Kilbride, Euan Cochrane[1], Jaye Weatherburn, Rachel Tropea

- 25-29 September 2017

[1] I added very little

Speaker slides will be made available by the end of the conference at:

Twitter Archive: courtesy of @mhawksey's TAGS:

- Explorer: https://hawksey.info/tagsexplorer/?key=1wsl3yX_Eg8Fj6INb3VGGqeWZC78CgUbmuyer437qBkc&gid=400689247
- Archive: ZC78CgUbmuyer437qBkc&gid=400689247https://hawksey.info/tagsexplorer/arc.html?key=1wsl3yX_Eg8Fj6INb3VGGqeW

Storifies

- Jaye Weatherburn's Storify day 1 and 2: https://storify.com/jayechats/ipres2017-day-1
- Jaye Weatherburn's Storify day 3: https://storify.com/jayechats/ipres2017-day-3
- Jaye Weatherburn's Storify day 4: https://storify.com/jayechats/ipres2017-day-4
- Jaye Weatherburn's Storify day 5: https://storify.com/jayechats/ipres2017-day-5

Blogs

- Blogs from DPC members at iPres - Jaye Weatherburn, Anthea Seles, Chris Fryer and Louise Lawson: http://dpconline.org/blog
    - Louise Lawson's blog of Ingrid Dillo's Keynote: http://www.dpconline.org/blog/fair-and-open-data
    - Louise Lawson's blog of the Emulation and Software Preservation session: http://dpconline.org/blog/emulation-and-software-preservation-at-ipres-2017
    - Jaye Weatherburn's blog of the Aquisition and Appraisal Session: http://dpconline.org/blog/acquisition-and-appraisal-at-ipres2017-2
    - Christopher Fryer. "Data management at iPRES 2017"
- David Rosenthal. "iPRES 2017"

Pictures

- iPRES 2017 Flickr: https://www.flickr.com/groups/4139869@N20/

Table of Contents

Day 1: Monday, 25 September

13:00-14:00 @ Main Hall

Japanese Tutorial 1
ディジタルリソースの長期保存に関する概観

***Organizer:** Shigeo Sugimoto*

***Abstract:** ディジタルリソースの長期保存は困難ではあるが取り組まねばならない問題として広く理解されている。ここではディジタルリソースの長期保存に関する基本的な理解を得ることを目的として、ディジタルリソースの長期保存の考え方、ディジタルリソースの長期保存の標準モデルであるOAIS（Open Archival Information System）等を紹介し、技術的な面からディジタルリソースの長期保存を俯瞰する。*

Power Point Slides (.pdf)

13:00-18:00 @ Conference Room 1

METS Editorial Board Annual Face-to-Face Meeting

Organizers: *Betsy Post and Tom Habing*
Abstract: *The METS schema is a standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library. Topics to be discussed at this year's annual board meeting include the development of METS Lite, METS RDF, relationship to other standards, and maintenance of the existing 1.0 schema.*
The meeting is open to the public. Potential attendees include METS users and potential users, adopters of closely allied schemas, such as PREMIS and ALTO, as well as anyone with a general interest in formats that promote standards based description for the exchange of complex digital objects.

Notes

Public METS Editorial Board Meeting - Karin Bredenberg

Introduction to the board meeting and the participants. Special welcome to those not on the board.

Introduction to METS - Tobias Steinke

Presentation based on METS overview available at:
http://www.loc.gov/standards/mets/presentations/METS.ppt

- Question: Why would you want to put binary data directly into the METS file?
  - No one in the room uses it, the problem is that you'd have to be able to decode it down the road, not really digital preservation proof
  - Someone wanted to have it in there when the standard was written
  - Use case: embed thumbnails of data you are referencing
- Structural link section and behavior section are two sections of METS which are not frequently used
  - Behavior section which can record all of the dissemination behaviors that pertain to a digital entity or its parts was originally requested by a US university when the standard was written - turns out they don't use it

- METS profiles are written in XML, should be versioned and describe how the elements within the file are used. Profiles can reference other profiles.

METS XML - Karin Bredenberg

Focus of past Editorial Board Meetings

- F2F Editorial Board Meeting 2014: introduced the idea of METS 2.0
    - The idea is the development of a METS Light version while determining that core functions of METS are OAIS compatible, still able to link to other descriptions and interoperable with other standards like OAI.ORE, BagIT, FOXML and PREMIS
    - For the past 5 years METS has released approx 1 profile a year, however, there are still many profiles out there which are not registered (e.g., Swedish Archives use Base Profile for Swedish agencies, which is not registered because it's a base profile that can be extended; also Finish
- F2F 2015 Editorial Board Meeting @ipres : Mets Lite/Light is definitely needed
    - Also a need for METS RDF
- F2F 2016 Editorial Board Meeting: focus on website and spreadsheet as internal working material for elements

F2F2017 open discussion:

- What is METS Lite/Mets Light/...
    - Starting point might be to identify not used / rarely used / often mis-understood / better covered in other standards element and get rid of those for METS lite
    - Problem is that METS has no underlying data model - unlike PREMIS
    - Other option is to restrict METS to certain use cases, e.g. digital preservation / packaging information for preservation (currently main use case at BnF), metadata for delivery
        - Identify these use cases, optimize METS for those use cases while keeping the existing METS standard
    - METS2 = METS Lite + RDF, while METS 1.xx will still be maintained
    - Many elements need to be explained / understood better - like filegroup. Why do you form filegroups?
- What are your needs?
    - Librarians need data in a certain way, archivist need data in a certain way - but if i work at an aggregator i need a format that i could use to deliver all these objects

    - Structure of METS is very complicated - subelements with subelements with subelements
- Do you need easier reading, e.g. of Primer?
    - Yes, METS is really complicated to learn
    - One suggestion: make files on website PDF (not doc, xls)
- Examples other than the ones in the profiles?
    - Swedish Archives, CSC Finland
- Why don't you register a profile?
        - Sweden: It's not finished yet, still in the phase of discussing it with involved parties

METS RDF - Bertrand Caron:

- METS is really flexible, but a data model is missing - one of the reasons why Fedora moved away from METS to PCDM
    - PCDM is ontology used by Fedora, Hydra, Islandora. It's way simpler than METS with the core classes collections, objects and files and the relationships between them.
- New PREMIS ontology will be published next month - when that happens, the work on the METS ontology can be picked up again

METS in relation to other standards:

Schematron and METS Best Practices - Aaron Elkiss (University of Michigan Library / HathiTrust)

- PREMIS in METS toolkit, originally developed by the Florida Center for Library automation can validate best practices for PREmIS embedded in METS (using Schematron), convert between PREMIS and PREMIS-in-METS and characterize a file with outputs as PREMIS
- PiM Schematron validation checks if PREMIS elements are in the expected METS metadata section types (rightsMD, digiprovMD, etc.), checks that PREMIS IDREFs only reference PREMIS IDs and checks if METS references all embedded PREMIS sections
- PIM conversion between PREMIS-in-METS to PREMIS and PREMIS to PREMIS-in-METS is done via XSLT stylesheets
- Brief introduction to Schematron as ISO standard rule-based validation for XML, based on XSLT / XPath, using xsd
- PiM Tool has gone over to PREMIS editorial board

METS in digital preservation in Finland - Juha Lehtonen

- Recommended file formats, accepted file formats for transfer, administrative and structural metadata, descriptive metadata is described in “standard portfolio”
- National METS profile is based on “standard portfolio”
- Each information package contains a METS document, a digital signature file and digital objects
- Finnish national specification describing mandatory, forbidden and optional elements for national METS profile
- Example of specification for METS Header:
  - Mandatory: CREATEDATE
  - Conditional: LASTMODDATE and RECORDSTATUS must be used, if updating existing data
- <binData> Binary data and <mdRef> in METS is forbidden in profile (amongst others)
- Descriptive MD can be repeated with at least one being from a list of allowed schemas
- At least one techMD per digital object with PREMIS required, using identifier, file format and version, fixity information, creation date of digital object, charset parameter required if file format is text (e.g. text/csv, text/xml,...)
- Separate techMD: NISO MIX (images), AudioMD (audio streams)), VideoMD (video streams), ADDML (csv files) → all those have mandatory elements defined within the profile
- Required provenance MD: creation event (PREMIS Event), Preservation plan (either mdWrap or mdRef → only instance where mdRef is allowed)
  - Preservation plan definition is currently an ongoing process, no specific notation / content defined yet
  - Q: as packages are already in the archive but the preservation plan is currently not available yet, how will you update the packages?
    A: we will create an incremental package, update the data in METS/PREMIS and capture the act as an event. However, some events like checking fixity is not captured in physical metadata (METS), but in database. Focus is on being able to export that information.
    Migration is an example where the information is important to capture, because we need to ask the producer for approval.
    Automated processes which don’t change the file format might not be that relevant to capture physically (in METS).
  - Specification available in English:
    www.kdk.fi/en/digital-preservation/specifications
  - National METS schema, schematron rules and related tools on github:
    https://github.com/Digital-Preservation-Finland

- The Digital Preservation Service in Finland has currently around 1mio information packages in archive, most of them come from web archiving / WARC files. METS captures information on level of WARC file with some minimal techMD information about the WARC file (no information on files within the WARC file)
- Some talk on validation of WARC:
    - JHOVE2 had sophisticated WARC module which allowed recursive capturing of techMD for all files contained within - JHOVE WARC module just captures on level of container
    - CSC used JHOVE2 for validation of WARC containers - problem was that output was bigger than WARC file
    - BnF has dropped JHOVE2 and validated WARC files using JWAT tools

METS Best Practices - Aaron Elkiss

- A number of requirements can't be checked by XML schema, but can be by Schematron rules, e.g.
  DMDID must reference a <dmdSec>
- Other things which only schematron can check are attributes, i.e. in which context attributes should be present or absent
  e.g. fptr element should have a FILEID attribute value (i.e., refer to a specific file) if it does not have a child area, par or seq element
  e.g. BEGIN, (END), BETYPE are optional but all should appear if any do
- In some cases, the schema doesn't give any guidance, these should be discussed.
  E.g. could you have an <area> with no SHAPE or BEGIN attribute? Could you have an <area> with both?
  Bertrand: yes
- GROUPID is another example which is not defined very clearly; one of those problematic elements
  Bertrand: within the profiles, only time i have seen GROUPID used, it was misused
- Start at schematron rule set implementing some best practice METS requirements.
  Includes Tests (in Ruby/RSpec), should validate with the METS XML schema
  https://github.com/mlibrary/mets_best_practices
- Documentation on French METS profiles: http://bibnum.bnf.fr/mets

Conclusion / Maintenance - Karin Bredenberg

- XML will by no means go away - but we do see use cases for RDF
  ➔ from an archival view XML should be there for another 50-100 years ;-)
- METS Editorial Board has begun moving a lot of stuff to github (see links in notes above), including profiles

14:00-15:00 @ Main Hall

Japanese Tutorial 2
社会調査個票デジタルデータの収集、保存、二次分析について

Organizer: *Yukio Maeda*
Abstract: *確率標本抽出にもとづく社会調査は1940年代のアメリカ合衆国で始まるが、その初期からデータの共有、保存、二次分析は研究者集団にとって重要な関心事であった。このチュートリアルでは、データの保存と二次分析という観点から、社会調査の歴史を素描した後、日本の現状を東京大学社会科学研究所が行っているデータの収集・保存・提供活動を中心に解説する。また、情報技術の急速な発展に伴って近年生じている変化について紹介する。*

Power Point Slides (.pdf)

15:10-16:40 @ Main Hall

Digital Curation of Historical and Cultural Resources in Japan (1)(in Japanese)
東京大学史料編纂所による前近代日本史史料の調査に基づく史料画像のデジタル化とその保存

Organizer: *Taizo Yamada (東京大学史料編纂所 Historiographical Institute The University of Tokyo)*
Presenter: *Taizo Yamada, Akiyoshi Tani, Toru Hoya*
Abstract: *東京大学史料編纂所における前近代日本史史料に関する調査とそれに基づく史料画像のデジタル化、およびそれの長期保存・長期利用を目指した取り組みに*

ついて報告する。複製史料収集の方法は時代とともに変化してきたが、蓄積した複製史料を永続的に利用するために、撮影・デジタル化・管理などの工程を史料編纂所として画一的に行うための方法を策定し、それを実現するためのシステム構築・運用を行っている。その他、在外日本関係史料の収集とデジタル化およびメタデータ付与についても報告する。

15:10-16:40 @ Conference Room 5

Pre-conference Workshop of Asian Session: 2nd Workshop on Academic Asset Preservation and Sharing in Southeast Asia

Organizer: *Shoichiro Hara*
Abstract: *This is a session to make arrangements for the Asian Session on 26th. It is planned as a semi-closed session, but observers are welcomed.*

IIIF

Mirador

Drupal

Q&A:

Examples of how your collection has been used?

- Researchers and students want to use for their academic papers/projects.
- Museums.
- Publishers.
- Requests for original materials for exhibition.

Any technical difficulties in the 20 years. Backward compatibility problems?

- Some of the TIFF resolution a bit low. Might need to redigitise.

Do you have plans to overlay current maps with older maps?

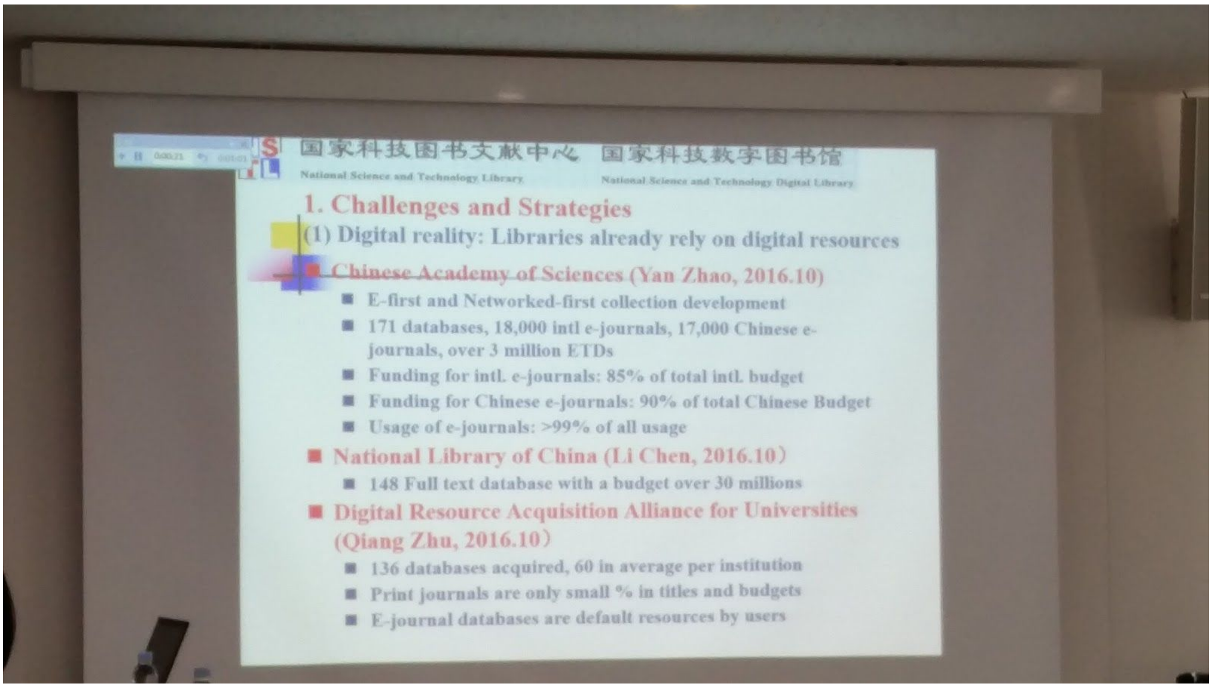
- Very interesting project. We would love to do it if we have the chance to do it. No plans for it for now.

National Digital Preservation Program if China for Scientific Literature

Speaker: Xiaolin Zhang

National Science & Technology Library, China

iPRES 20
14th
International
Conference on
Digital
Preservation
Kyoto, Japan
September
25-29,2017



国家科技图书文献中心 National Science and Technology Library 国家科技数字图书馆 National Science and Technology Digital Library

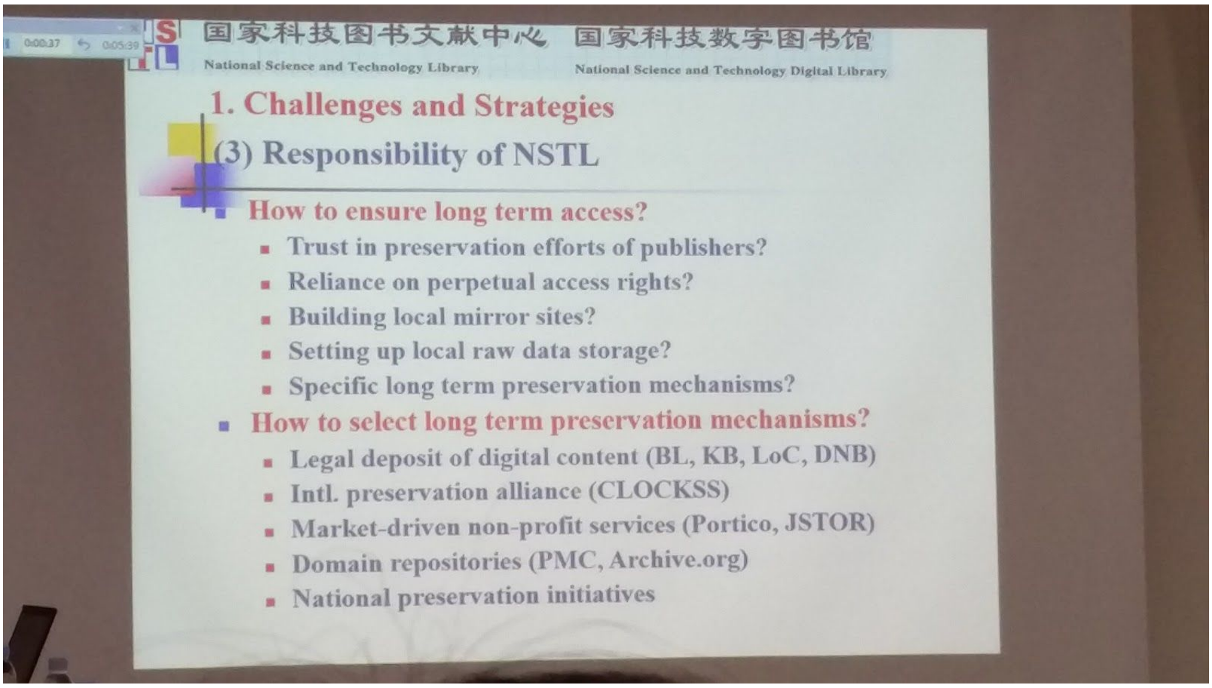
1. Challenges and Strategies

(1) Digital reality: Libraries already rely on digital resources

- Chinese Academy of Sciences (Yan Zhao, 2016.10)
  - E-first and Networked-first collection development
  - 171 databases, 18,000 intl e-journals, 17,000 Chinese e-journals, over 3 million ETDs
  - Funding for intl. e-journals: 85% of total intl. budget
  - Funding for Chinese e-journals: 90% of total Chinese Budget
  - Usage of e-journals: >99% of all usage
- National Library of China (Li Chen, 2016.10)
  - 148 Full text database with a budget over 30 millions
- Digital Resource Acquisition Alliance for Universities (Qiang Zhu, 2016.10)
  - 136 databases acquired, 60 in average per institution
  - Print journals are only small % in titles and budgets
  - E-journal databases are default resources by users

How to ensure long term access?

- Trust in preservation efforts of publishers?
- Reliance on perpetual access rights?
- Building local mirror sites?
- Setting up local raw data storage?
- Specific long term preservation mechanisms?



国家科技图书文献中心 National Science and Technology Library

国家科技数字图书馆 National Science and Technology Digital Library

1. Challenges and Strategies

(3) Responsibility of NSTL

- How to ensure long term access?
  - Trust in preservation efforts of publishers?
  - Reliance on perpetual access rights?
  - Building local mirror sites?
  - Setting up local raw data storage?
  - Specific long term preservation mechanisms?
- How to select long term preservation mechanisms?
  - Legal deposit of digital content (BL, KB, LoC, DNB)
  - Intl. preservation alliance (CLOCKSS)
  - Market-driven non-profit services (Portico, JSTOR)
  - Domain repositories (PMC, Archive.org)
  - National preservation initiatives

Q&A

1. Human Resources? How did you develop?
   - IT department very strong. 30 ppl in Mr Zhang's Institute.
   - IT capabilities != digi pres capabilities
   - Big research and discovery project since 2004.
   - Ministry recognizes this. Develop the human resources.
2. You mentioned mostly scientific data, any humanities materials?
   - There's a division of labor.
   - Ministry of Science and Technology runs NDPP. MoST do it first, hope other ministries will follow suit.
   - If we have access, we'll take it all. We don't purposely exclude humanities journal. If it's part of the publisher's collection, we'll take them.
   - PDF, images and XML are nothing compared to the scale of Astronomical data.
3. Language of the materials preserved. English?
   - We prefer English. Metadata are translated to English.
   - If it's not in English, we will still take it first.

- Chinese journal publication usually also have English title, author and other metadata.

4. Metadata schema. How are you managing schema conflict?
    - Request in JHS format. Most international journal publish to PubMed, already have JHS.
    - We ask them to define in XML, so we can resolve it.
    - If don't have XML, we will still take it. Then we have a team to resolve that case by case.
    - To be honest, to run a trusted repository, disaster recovery kind of national digital preservation, you don't need very complicated metadata.

5. How did you convince publishers to oblige? How did you convince them that it won't eat into their revenue, reduce subscriptions?
    - Ours is a dark archive.
    - Most libraries want the latest journals, not necessarily older journals.
    - 90% of publisher's revenue from research heavy institutes. Statistics to convince and persuade them.
6. Data analytics of user behavior of using the archive?
    - Exploring, not doing it yet.
    - Gray legal area.
    - Need to set limits on what we can do.
    -

Day 2: Tuesday 26 September

10:00-12:00 @ Main Hall

Digital Curation of Historical and Cultural Resources in Japan (2)

歴史資料デジタル記録として何を記述すべきか–日本とアジアと世界– (with translation)

Organizer: *Makoto Goto (国立歴史民俗博物館 National Museum of Japanese History)*
Presenter: *John Ertl, Yoshiko Shimadzu, Shigeki Moro*
Abstract: *日本における歴史資料保存の手法には様々なあり方がある。究極的に「何を未来に伝えるべきか」という論点は、現物だけではなくデジタルデータでも同様の課題を抱える。そこで、本セッションでは、デジタルデータ記録の前提となる「何を記録すべきか」「そのための歴史資料情報のデジタルでの記録の課題は何か」という根本の部分を考古学・文化財科学・デジタルの分野から考えたい。*
Abstract: *There are a variety of methodologies for preserving historical resources in Japan. Ultimately, the main point “What and how can we predict the future?” matters not only the original resources but also digital data. This session addresses underlying problems, “What do we record?” and “What are the outstanding issues for information record from historical resources” from the scope of archaeology, scientific studies on cultural properties, and digital.*

Introduction

- Session will not only cover digital material, but also the related cultural heritage objects objects

Shigeki Moro - Hanazono University (@moroshigeki)

- Specialist in Buddhist studies, focus on Buddhist logic and consciousness.
- Use of computers and technology has reached this area as well. Now working with a variety of digital content
- Example: 3D computer graphics restoration of former Buddhistas well as residential structures; several projects often upon request of municipality governments
- 3D structures can be fully navigated using VR technology
- 2009 project of digitally restoring town houses
  Some old houses still exist, however, these have been renovated into cafes or stores. But because of the major fire 150 years ago only very few fully intact houses remain. So, it was difficult to show what townhouses actually looked like.

Documents on layouts of townhouses were found and uses as a basis for the computer graphics generated 3D version.
In 2017 a whole block of buildings was digitally restored (as of yet rough rendering without colors and windows / details).

- Why do we digitally restore townhouses?
  - to contribute to the study of history; digital restoration based on documents lead to results which were different from those of traditional research
  - created materials play a critical role in counteracting revisionist history / term "history wars" in Japan

Reconstructed Buildings as Archaeological Archives - John Ertl, Kanazawa University, Institute for Liberal Arts and Science

- Archaeology is about data & physical objects
- Archaeologists due a fabulous job in preserving the (meta)data about the archaeological artifacts / objects and the (analogue) objects it self
- However, much less care / thought / effort / money is put on data created today. E.g. - what was displayed, how, why? How has the display of a museum changed over time?
- Use case: Reconstructed Buildings
  - based on remains, which can be up to 10000 yrs old
  - While Ertl has been researching these objects for years, he has just started recording them digitally last year
  - Problem: public usually has no idea on how much research and data the reconstructed building is based
    → this is problematic, when designs are based on compromises; need more data to understand the construction correctly
- Metadata to collect about sites (e.g. location, historic data, ...), buildings (e.g. Location - GIS Location, Builder, budget,...) and references (e.g. type of reference, title,...) was defined
- Comment by @mickylindlar: Seems like no checking with existing standards, best practices was given (?) e.g. MIDAS, CARARE → see http://www.dcc.ac.uk/resources/metadata-standards/disciplinary/archaeology
- "I know that after I have written a couple of journal articles I will get bored with this database and probably not update anymore. And I don't know how to digitally preserve it. But I'm hoping to find out more at this conference"

Digital Data in Conservation of Cultural Properties - National Museum of Japanese History, Yoshiko SHIMADZU

- Cultural property protection law in Japan, applies to tangible objects like archaeological objects but also to intangible objects like dance
- Intangible is e.g. dance, theater, local crafts
    - Capturing this requires mainly digital techniques these days - e.g. AV
    - This automatically leads to digital preservation
- Objects in museum can only be interpreted correctly if we know which part is original (e.g. ancient) and which part is not original (e.g. 18th century restoration)

Discussion (all participants):

- Q: What is being done to ensure that the data created is still accessible in the future?
  A: good point. Making it open source as much as possible or using standardized approach is favored. Data should be refreshable / available / not bound to proprietary systems.

13:00-13:30 @ Main Hall

Opening

- Rechert: two strands can be observed in this years program. (1) fast growing field of born-digital content, also especially personal & research data output (2) preservation as output of commercial processes, etc.
- Number of submissions received this year not as high as in the previous year, but of very high quality. Short-list of the best paper awards:
    - Marco Klindt - PDF considered harmful for digital preservation
    - Bertrand Caron et al. - Life and Death of an Information Package
    - Sara Tarkani et al. - Trustworthy Emulation Platform for Digital Preservation

13:30-14:30 @ Main Hall

Keynote (1): FAIR Data in Trustworthy Data Repositories

Speaker: *Ingrid Dillo (Data Archiving and Networked Services, Netherlands)*
Title: *FAIR Data in Trustworthy Data Repositories*
Chair: *Klaus Rechert*
Abstract: *National and international funders are increasingly likely to mandate open data and data management policies that call for the long-term storage and accessibility of data. Open data and data sharing can only become a success if we*

put the concept of trust central stage. The certification of digital repositories is an important means to provide this trust to the different stakeholders involved. In this keynote 1 will talk about data sharing, repository certification and the concept of FAIR data.

Full Abstract: [PDF]

(see also Louise Lawson's Blog on this session: http://www.dpconline.org/blog/fair-and-open-data)

Ingrid notes that iPres is happening alongside meeting of the WDS - World Data System - Asia Pacific conference. Research data invites conversation about data sharing and trust, and certification. This allows us to link the discussions on digital preservation to current topics like FAIR and Open Data.

Links between Japan and Netherlands long established for trade but also learning. The shared history between the Dutch and the Japanese is based on trust, which should also the basis for scientific & data exchange.

Data Sharing

Research Data Management (RDM) in the era of Open Science. Data sharing is important for transparency; replication of research; facilitates re-use of data, and in turn efficiency, return on investment etc. 19th c. Whaling logs used by climate scientists today are an example of the reuse of records outside the original domain for which they were created.

The Concept of Trust

What do researchers thing of open data? Most are supportive in a survey. However most researchers do not make their data available in a way in can be used by others. Intellectual property, confidentiality, misinterpreted, ethical concerns, attribution etc. Trust issues. If you create a system…"will [my data] be lost garbled stolen or misused?"

Data sharing is key to reducing data fraud - prominent cases in the Netherlands (worldwide).

As per survey, majority of researchers seem to approve data sharing … however, in reality the majority of them don't. They keep data for future use on their computer at work or on portable storage carriers and are worried with intellectual property concerns, misuse concerns, etc. Data that is not shared typically sits on private disks or portable storage, immediately raises a derived preservation issue.

What can be data sharing incentives?

Sharing needs to become norm and there have to be professional rewards for data sharing. External drivers in this context are funder policies and publisher requirements.

Incentives for researchers – culture. Mostly affected by their peers/research circle, if sharing is the norm. Getting academic credits – reward for investment.

Pillars of trust - integrity, transparency, competence, predictability, guarantees, positive intentions. External acknowledgements.

Repository Certification

Global certification landscape

There are two standards for repository certification ISO16363, DIN 31644; and there is the Data seal of approval.

DSA and WDS are lightweight, self-assessment, community review. Disciplinary & Geographical spread. Now replaced two with one set of requirements using best of both, two became one under the auspice of the Research Data Alliance => *Core Trust Seal*

DSA and WDS recently joined forces to form partnership for assessment - goal for partnership was to simplify assessment options and to stimulate more certifications. CoreTrustSeal is the result - now contains common catalogue of requirements between DSA & WDS. Using best of both, two became one under the auspice of the Research Data Alliance.

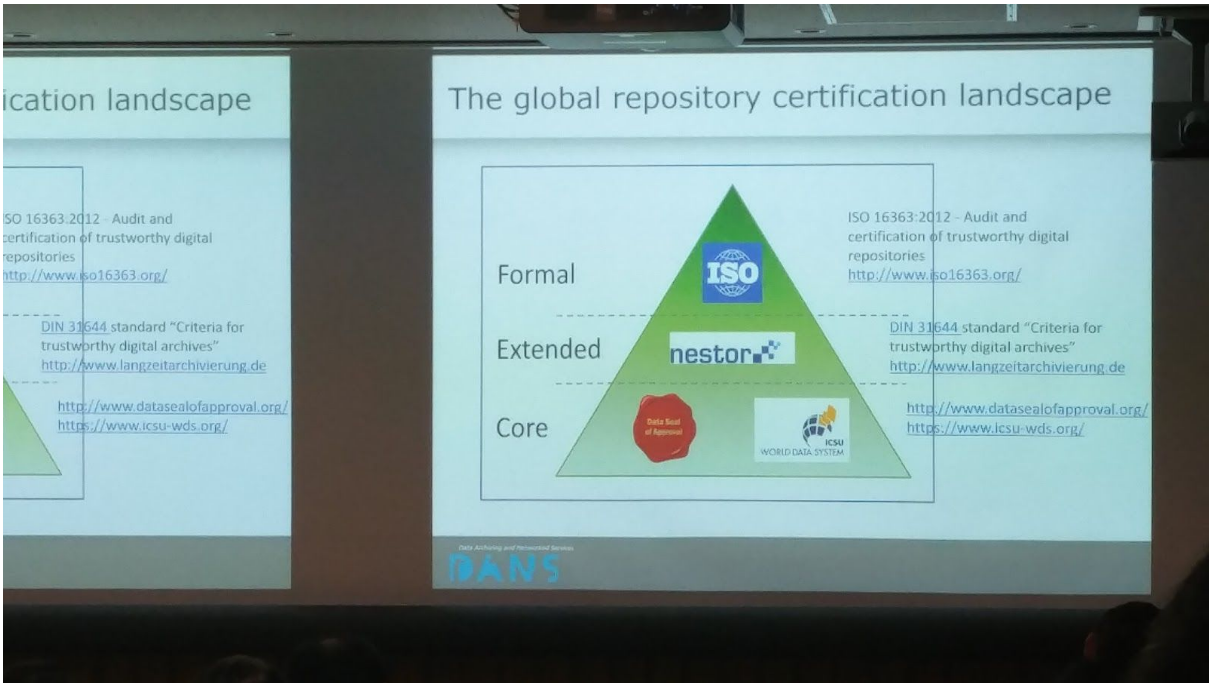
Core Trust Seal:

Core Trust Seal is based on self-assessment. Peer reviewed.

16 requirements. Self-assessment (publicly available) . 3 year seal period.

When you receive a seal, info you’ve provided becomes public. This encourages ‘trustworthiness’ and common knowledge, shared learnings.

New requirements – need to have many people involved, insight into organisational structure, staff, funding.



The global repository certification landscape

Formal

Extended

Core

ISO 16363:2012 - Audit and certification of trustworthy digital repositories http://www.iso16363.org/

DIN 31644 standard "Criteria for trustworthy digital archives" http://www.langzeitarchivierung.de

http://www.datasealofapproval.org/
https://www.icsu-wds.org/

Before:

Core level: World data system or Data Seal of Approval.

Similarities:

Lightweight standards. Self-eval, reviewed by committee. Originate in Europe.

One new certification body to replace WDS and DSA. New body is called Core Trust Seal (launched 11 Sep). (Perhaps unprecedented since the norm is for standards to replicate rather than be reduced coherently. A significant accomplishment in community cohesion)

Research Data Alliance (RDA). 23 things: Libraries for Research Data. Great resource. https://www.rd-alliance.org/system/files/documents/23Things_Libraries_For_Data_RDA.pdf

Deeper dive into the new CoreTrustSeal standard. Three broad components of the standard - organizational infrastructure, digital object management and technology.

Why do institutions undergo certification?

· Why do repositories invest in certification efforts?

Builds stakeholder confidence/trust in the repository, raises awareness, improve communication within the repository, improve repository processes, and, to differentiate yourself from other repositories.

Importance of certification – seal holders & very god reward for investment. Many think 'Core level' is enough.

NCDD report on perceived benefits of data seal of approval shows that there was a strong ratio between effort and rewards for the seal of approval. Almost no one aimed for certification at the higher level so the core trust requirements seemed to be sufficient for most participants. Report availble here: http://www.ncdd.nl/wp-content/uploads/2016/10/201611_DE_Houdbaar_Report_DSA-survey_2016.pdf

Certification like CoreTrustSeal says something about the quality of the repository, FAIR data principles say something about the datasets contained within.
Matrix was developed with 5 criterias / stars for each category (Findable, Accessible, Interoperable). Stars in the criteria shall automatically lead to R = Re-use. This is the first step towards an assessment tool.

FAIR and Open

FAIR data - **F**indable, **A**ccessible, **I**nteroperable, **R**e-usable

Guarantees of technical quality. Need licences, continuity plan, integrity checks, long term storage, sufficient metadata, IDs are mandatory.

A lot depends on depositor, eg metadata attached

FAIR Badge Scheme / Data Assessment Tool:

Core certification does not differ from concept of FAIR – tried to operationalize concept of FAIR.

Tried to come up with metrics for FAIR. Prototype being tested. Hard to define metrics for R, so we decided to add up FAI to get the R.

Have Survey Monkey questionnaire, user tested.

Some people say that FAIR data has nothing to do with OPEN data - however, Dillo says that there is a relationship because accessibility has a lot to do with openness. However, some data just can't be made openly available so they won't be able to achieve 100% It's unfair! But that’s the way it is. Also Open is not necessarily the key to fairness. It’s a combination of Open and Fair that we should be aiming for.

Levels of openness – Datatags

New European laws around personal data (similar in other geo-locations). Sweeney & Crosas introduced the notion of datatags. Working on system of levels based on GDPR – GDPR DataTags. Researchers complete questionnaire (it’s currently in draft form).

Comment from William McBride: regarding GDPR (data protection regulations). If risk averse organisations hear a message that they should be reducing the amount of data they have, a strange consequence of this is that may be that we have less or no data to preserve. Lessons to take home - preservation is not an unrealistic thing to attain [great message Ingrid is giving]; and, data protection regulations – take account of that work. Strengthen relationship between data protection and digital preservation.

14:40-16:40 @ Main Hall

Asian Session: Reports and Discussion

Abstract: *Countries in the Asia-Pacific region are very diverse in terms of culture, language and economic environments. Long-term management, keeping and use of digital resources is a common and pressing concern for many of these countries which are producing more and more digital resources. However, reports on digital preservation activities are lacking, especially from the countries in East and South-East Asia. This session is aimed to share up-to-date information about developments of digital archives and digital preservation in East and South-East Asia and to discuss issues on digital preservation in this region with the audience from other parts of the world.*
This session will first present five talks by invited speakers from Japan, Taiwan, Philippines, Thailand and Singapore about digital preservation at the speakers’ institutions and/or countries. Then, we solicit voluntary reports from other Asian countries followed by general discussions with the audience.
Moderator: *Natalie Pang (Nanyang Technological University)*

Invited Speakers: *Shuji Kamitsuna (National Diet Library, Japan), Sophy Shu-Jiun CHEN (Academia Sinica, Taiwan), Lee Kee Siang (National Library Board, Singapore), Wararak Pattanakiatpong (Chiang Mai University, Thailand), Chito Angeles (University of the Philippines Diliman, Philippines)*

Digitally endangered species - #bitlist

Philippines - digital archives from newspapers. Hard to track down, not a clear responsibility. Delighted to find materials archived in other countries

Japan - intangible cultural heritage preserved in local villages. Elderly populations.

Singapore - Social media, snapchat in particular for young people. Snapchat is a subculture so even when captured the context is lost.

Singapore New technology allows publishing to be fast but not good at capturing content for archives. These are our new real challenges

Japan - context for the photographs; games are lost too quickly, but Japan has a lot of expertise on this. Nintendo.

How to capture emotional response and experience of digital artefacts. Anyone doing this? Generational gap.

Disconnect between generation. How do we address this issue?

Cambridge, UK: depends on what kind of materials formative years. Gen X = old computer games, emotional responses. Material changes over time. I might not have the same response to an app, but my peers would. More pertinent, are the younger gen taught to look under the hood? They might know how to use tech superficially. They might not know the nuts and bolts.

Kee Siang (SG): Citizen archivist programme. Encourage citizens to contribute. Old gen and their kids input metadata related to the photo or manuscript. Aging population -> oral history, capture their memories. Training institutions to conduct oral history. Audio form. Audio to text conversion.

Taiwan: a lot of old photos. Historical maps. Digitise only historical maps with GIS. Combine photo and maps. Produce modern publication, walking into old taipei, old tainan. Young ppl can use the app, enjoy the digitised materials.

Dr Natalie Pang: Bukit Brown cemetery. Document tombs and Graves. Old cemetery, belongs to generation before Singapore was formed. But the idea was to layer maps (GIS) with tomb inscriptions and story of the descendents. Project didn't come to fruition but this is one idea to bridge the gap.

KAMITSUNA, Sjuki - National Diet Library, JAPAN
How can we use we archives? A brief overview of WARP and how it is used

WARP is the web archive of the National Diet Library
Harvesting at National Library - current size: 1 PB, 5 bio files, 130,000 captures.
85% of that is open access / freely available websites to the public via the internet.

Use Case for the WARP archive:
Use Case 1: Linking from Live websites
WARP can be used to push data into the web harvesting process of the National Diet Library. Many public agencies use this as a form of data back-up, where websites are pushed into WARP prior to update.

Use Case 2: Analysis and visualization
WARP data is exploited to create aggregate and visualize information on Japanese websites, e.g. for relative size of data accumulated from each of the 10.000 websites archived in WARP.

Use Case 3: Curation
Creating a specific collection - e.g. around the earthquake

Use Case 4: Uncovering PDF Documents
WARP uncovers PDF files of book and periodical articles that are contained in websites - 1.5 mio PDF records were pushed out to the catalogue that way through WARP

Future challenges:
Data mining
Current indexes comprise 2.5 billion files in 17 TB of data - results contain much duplicate material archived at different times and other forms of “noise”. What is needed is a robust and accurate search engine specialized for web archive which must implement temporal elements.

Website of iPRES2017 is already archived in WARP

A Review of Current State of Digital Preservation in the Academia Sinica Center for Digital Cultures / Taiwan

Large scale Digital Library Initiatives in Taiwan data back to 1998. 100+ institutions (GLAM, Academia) have contributed to digital archive with over 5 mio digitized objects and over 750 websites & databases now in the archive.

Digitization Guideline Series has been put forth by Academia Sinica Center.
Linked Open Data principles have now been aplied to some websites (e.g. “Fishing in the Data Ocean Project” - https://summit2017.lodlam.net/2017/04/12/fishing-in-the-data-ocean/

Curating Tool: “DIGIMUSE System”. Includes Temporal and Spatial module for easy discovery along map / timelines.
Redundant storage - > 100 km apart, on file basis weekly.
Currently running pilot project on emulation.
Refreshment and Migration: on demand transfer of data between two types of same storage medium so there are no bitrate changes or alternation of data, esp. For digital objects of audiovisual, video, etc.

Overview of Digital Preservation National Library Board Singapore - Lee Kee Siang

Singapore is a very “wired” community - the #1 smart phone users in the world, where the average person carries 3.3 devices. Very high rate of internet access.

NLB oversees the Public Libraries (26), National Library (1.9 million visitors yearly), National archives (>40,000 visitors yearly).

“Digital Preservation” is a dedicated focus area within the national digital strategy. The goal is for every citizen to be able to access and personalize the data.

“Everything in the NLB Collection that is precious will be ingested and preserved beyond this generation”. This is achieved using:

- Security systems for secret documents
- System & processes for mass ingestion of born digitals (eg. digital legal deposit, website archivals, post-NLB Act Amendments, National government email records)
- High-res audiovisual materials
- Digital storage space

NLB’s digital preservation journey started in 2005, moving things to different storage.
2011 - Rosetta was implemented as a digital preservation System.
2012 - the National Library and the National Archives of Singapore merged.
2018 - the Digital Act will be enacted which will clearly differentiate between mandatory and voluntary deposits. It will also allow web harvesting.

The Singapore webarchive started in 2006. Collection principle: websites about Singapore, from Singapore.
Web archiving Curation system includes pre-harvesting, harvesting and post-harvesting functions.

Ongoing efforts include streamlining preservation policies and strategies. Challenges include balancing between overwhelming content and limited resources. The lack of organizational responsibility, resources and infrastructure supporting active preservation activities.

Q/A: What about dynamic websites?
Audience recommendation: Webrecorder from RHIZOME - see webrecorder.io
Audience feedback: webrecorder is a great tool, but will fall short for use cases like NLB Singapore because it doesn’t scale as well. We need to fundamentally address that problem as a community where we are still working with tools and processes that are decards year all.

CMU’s LIbrary’s Digital Archives and Digital Preservation
CHiang Mai University Library, Thailand
Wararak Pattanakiatpong
CMD Digital Archives holdings:
Etds, e-heritage manuscripts from library’s holdings, microfilm digitization results, e-rare books, local newspaper and thai newspapers (digitized directly from 2015-present, digitized from microfilms for 1953-2014), e-commerce archive.
Academic records = approx. 27k, stored in DSpace repository with Dublin Core metadata.

Enruring Long-term Access to and Preserving the Cultural Heritage of the Philippines and the Institutional Memory of its National Univeristy
Digital Preservation Initiative at the University of the Philippines (P) and Collaborating Institutions

National University in the Philippines is only university in the country, founded in 1908 - divided into 8 campuses

22k students, 1.5k faculty at main campus

Material is collected based on “Decree on Legal and Cultural Deposit”, signed in 1975
“Within one month from the data of any printed book … is first delivered out to the press, the publisher of such book shall furnish, free of charge and in the same finish as the best copies of the same are produced, two copies thereof to the National Library, and a copy each to the UP Main Library, the UP Library at Cebu City, the MSU Library, …”

Additionally: Executive Order No. 13 which establishes University Archives to collect and maintain archival materials.

Milestones of digitization and preservation activities (2005 - present):
2005 - outsourced digitization project (microfilm)
2008 - first in-house digitization with hw purchased
2016 - digitization services expanded
Early 2017 - digitization and digital preservation services were expanded with the acquisition of new scanners and storage devices; also Digital Archives @UPD with ETD submission launched.

- The eLibrary Project
- UP’s institutional Repository Project - eprints to DSpace, then in-house development
- Digital Archives @ UDP containing thesis, records, personal papers, UP Presidents’ papers
- Digitization hardware: various scanners, AV converter
- Software
- Formats: JPEG for image, MP3 for audio, MP4, FLV for video
- Infrastructure, eg SAN

Challenges include IPR, technical obsolescence, data migration, interoperability, sustainability

Question to the panel:
William Kilbride, the Digital Preservation Coalition

“What is the most digital endangered material which the participants in the panel worry about?”

- Philippine newspapers - the National Library currently does not have a project to convert newspapers to digital form. Only a few of the publishers in Philippines maintain their newspaper archives
- Japanese intangible cultures which are disappearing from the countryside
- In Japan there is also a big gaming culture with a lot of different games being produced, a content type which heritage institutions struggle to preserve appropriately
- Young Singaporean social media users who socialise on platforms like Snapchat and Facebook on a regular basis. Young people are losing many of the social moments

which they experience on Snapchat. The context is often lost even if the content is captured

- Dynamic websites and other technologies where there are no good solutions for capture and issues with security

16:40-17:30 @ Main Hall

Posters/Demos Lightning talk

Chair: Unmil Karadkar

17:30-19:00 @ Foyer

Welcome reception

Day 3: Wednesday, 27 September

Katherine Thornton, Euan Cochrane, Thomas Ledoux, Bertrand Caron and Carl Wilson. Modeling the Domain of Digital Preservation in Wikidata

- https://www.wikidata.org
- http://www.twitter.com/wikidigi
- Google group: https://groups.google.com/forum/#!forum/wikidata-for-digital-preservation
- OPF blog post: http://openpreservation.org/blog/2016/09/30/wikidata-as-a-digital-preservation-knowledgebase/

Chunqiu Li and Shigeo Sugimoto. Metadata-Driven Approach for Keeping Interpretability of Digital Objects through Formal Provenance Description

Metadata longevity should be ensured as well for future use. Provenance is crucial component of PDI defined in OAIS. Provenance of metadata describes change history, responsible agents, activities occurred on metadata objects. The changes of metadata definitions should be traced to prevent inconsistencies in the future use of metadata. Proposal of a model to describe provenance of metadata application profiles based on W3C PROV and Singapore Framework for Dublin Core Application Profile.

Provenance description should be machine-readable, traceable, and interoperable in the Web environment.

Digital Art ity - Building a Data Model for Digital Art Corpora
Celine Thomas, Bertrand Caron

3 year French research project, started in 2015, combining skills of digital arts & digital preservation
Includes interactive art exhibitions, with almost autonomous pieces; e.g. artworks based on algorithms displaying light cubes based on human interaction / movement; robot dogs fighting against each other based on random algorithms
Often involved custom made software / hardware
In France digital art is the most used use case for preservation of interactive objects
Creating entity map of (...some info missing…). Event / experienced artwork

How to get information on art works development / intention /experiences?
Interviews, Questionnaires
E.g. taping artist while interacting with their art work and explaining the work
If done with the captured environment (e.g. at BnF), artist con confirm whether preserved art work is working as expected or not (interview with artist as QA)

Project transformed entity map into ontology

How to capture information on art work in BnF's catalog?
Intermarc format used: pros - very granular structure & evolving taxonomis; cons - fixed fields and not designed for the specificities of digital art description.
Challenging, but possible.

In the digital repository digital art objects are structured in METS files. Stored in different systems (on the agenda is migration of art works into the SPAR system).
SPAR system has little experience with preserving software - art work media img will be a first for img within SPAR.
Packages described in 2 METS files will include:

METS file 1
- Disc image
- Install info (installation plan, installation manual, launcher)
- Reference information on package virtualization
- Reference information on package hardware
- Reference information on package software

METS file 2
- Documentation on art work
- Doc / work → data and user guide

Two METS files are linked together

Q: impressive METS diagram - but what are your plans for how to get it all back out if you want to move it to a different installation?

A: we are just at the beginning of the reflection on emulation. The work we need to do is gather experience from the IT department, from the AV department that has a lot of experience with manual emulation. When currently some kind of multimedia document is requested by the reader in the reading room, the package was disseminated from the audiovisual-system and there was a totally manual operation by the engineer preparing the standard computer setup / virtual machine / emulator required. Due to this there is currently a large percentage of work that cannot be displayed to the public. Not sure what degree of automation we can reach.

Q: when you add the environments to your schema, do you use some kind of schema or are you looking at developing your own?
A: it's based on PREMIS3 specification to add environments and the link / relationships between object & environment.

Q: with the installation plan you add to the plan - what do you add to that? How do you keep track of the installation? Capture all lines for example? Is there an automated process to confirm the installation instructions you captured?
A: There are standard processes for each package that we have, which will be included in the documentation. These are not really machine readable yet.

Q: Description - descriptive models we are using are just not concise enough (yet). If you would start from point 0, how would you build an access system that is flexible enough to allow for all access requirements to our digital objects for our users? E.g. in case of uncertainty about data, about provenance, about software, about hardware. Descriptive properties need to be flexible enough.
A: Big question, will have to think about that.

Portable emulation

Problem: providing a secure remote access to restricted born-digital content
Solution: secure, trustworthy and portable emulation architecture for digital preservation

- protects the confidentiality and integrity the sensitive content against malicious IS
- Has performance overhead (highly depends on system design and can be improved)
- Can be used to build non-emulation trustworthy systems

Trustworthy emulator-GameBoy prototype
Connection to emulator between user & server might be done using secure encrypted load cache
Goal: to seal emulator being used from end-user platform to run in a secure way so it can't be compromised by malicious software etc.
Requirement for this: service providers trust secure hardware (in this case Intel SGX)

What the solution doesn’t protect against: side-channel attacks, displayed output could be captured via screenshots or audio recordings

Problem:
How can you establish trust in a remote system?
How do you verifiably execute a program on a remote host?
Attestation

In progress: checking into trustworthy non-emulation platforms
E.g for for PDF →
Xpdf and XpdfReader use following libraries: Qt (for UI), FreeType (for font rendering), libpng (for handling png images), Little CMS

In progress: More general purpose application, i.e. trusted full system emulator “Basilisk MacOS emulator”. Goal is to run encrypted content as VHD disk on Basilisk emulator

Q: Is Intel SGX needed on the client side? If so, it will really limit the usage on the user side.
A: Completly right, currently it is limited because it’s expected on the user side. But we think that this will be different in the future.

Software Heritage - Why and How to Preserve Software Code
Software Heritage - Roberto Di Cosmo, Stefano Zacchiroli

“The source code for a work means the preferred form of the work for making modifications to it” - GPL license
In the future, software code will be the only place to find all information about the software and it’s intention / structure.
In a sense, open software is common material / information. Are we as a cultural heritage community taking good care of this?

Where software is published on the internet is flux - many “fashion victims” exist (like Sourceforge). Projects tend to migrate from one place to another over time.
Like all digital information,FOSS is fragile - due to inconsiderate / malicious code loss (e.g. Code Spaces) or business-driven code loss (e.g., Gitorious, Google Code) or obsolete code / physical media decay

Data structure of archive - a giant Merkle DAG with no loops

Archive is live - currently containing around 4 billion files (unique), 900 mio commits, 65 mio projects
Current sources:
GitHub, Debian, GNU, WIP: Gitorious, Google Code, Bitbucket
150 TB blobs, 5 TB database (as a graph 7 billion nodes + 60 billion edges

We believe this is the richest source code archive already

How to use the archive:

- Lookup by content hash (done)
- Browsing: wayback machine for archived code (done via Web API, soon via Web UI)
- Download: wget / git clone from the archive (todo)
- Deposit of source code bundles directly to the archive (todo)
- Provenance lookup for all archived content (todo)
- Full-text search on all archived source code files (todo)

It is now urgent to preserve software source code itself

Software heritage is taking a very systematic approach, has synergies with cultural, research and industry needs
SW heritage is a shared infrastructure that can benefit us all … we should collaborate and pool resources to make it so.

www.softwareheritage.org

Q: Are you planning to include virus code as well?
A: I'm sure we already have viruses in there - we're archiving github after all. We're complying with local regulations.

Q: Are you collaborating with blackduck?
A: we are not, they are aware of what we're doing. They are more for a commercial use case in connection with license information. We are after an open provenance db approach. We would like to build shared data and infrastructure that these companies can build on.

Q: As source code is mostly text - if you are crawling repositories, are you just harvesting text or also the binary? And have you encountered any text encoding issues, e.g. by grabbing tar balls from old projects or for non-UTF-8 stuff.
A: No, we're not discriminative against files, if you have binaries we take them. We haven't find any encoding issues in archival, because what we are archiving is just bytes - of course in displaying them, there will be files that we cannot display on your browser and in that case we will just let you download that!

Q: What is your preservation part of what you are doing?
A: Right now 3 copies - 2 on premise, 1 with cloud provider. What we stored once is stored forever, if we are forced to download something it will only be taken down for download not for archive. Persistent internal identifiers. Healing copy mechanism is in place.

Adding Emulation to existing Digital Preservation Infrastructure

Emulation as a Service
EaaS - number of "base environments" (Win 95, 98, etc) were created and made available via the service

The METS record is created describing media installation and usage order (e.g. for multi-part discs)
Metadata gets embedded into Preservica which makes it available via its REST API

The EMIL Characterization tool which comes as part of the EaaS package is run for gathering the technical environment requirements

What are the current Barriers to widespread use?
IT's currently not embedded in existing digital preservation off-the-shelf systems, because …
…. Technical: shifting lots of data around, big disks & big emulators. Ingest throughput. Validating less common platforms. Server based systems. Continuous updates. Future server based systems?
… Management: organizational limitations in personnel
… commercial barriers: protection by copyright law on so many levels (not just microsoft, but even the obscurest little package in Windows)

Phantoms of the Digital Opera: The need for long term preservation of born-digital actors and multimedia objects using methods that permit ongoing new creations
Dena Strong / U of Illinois

Problem: researchers on decade-long or multi-decade projects need to be able to preserve objects while still trying to work on them
→ is this a digital preservation problem? If not, who is going to take care of it?

Creators need to whole "opera house" - actors, sets, costumes, lights, sound … and each component needs to remain editable and "active"

SHINKAI Makoto: Hoshi no Koe (Voices of a Distant Star) - created in 2002, if re-released today, this would have to be recreated from scratch

Nina PALEY, Sita Sings the Blues / Book of Exodus - chose fossizilation and a digital divide within her own body of work; split her systems between first interview (2014) and second (2017)

David Fleming, professional translation for anniversary editions, DVD to Blu-ray series conversion. SubTitling requires sub-section exact work. SW/HW used is SubStation Alpha last updated in 2001, Excel plug-in compatibility needs Office 2003, Video converted from VHS no longer accessible in Win 9.1, Frame rate mismatches
What's the path forward?

Four primary tactics and a dream:
Emulation - promising, but currently incomplete and difficult for users
Migration - rarely 1:1, significance depends on project.
Re-creation - redoing old work takes time away from new work

Fossilization: many creators only option, but fragile
My dream: easy-to-create, easy-to-use portable cloud-based emulation …. But how do we get there? (codecs, drivers, licenses, etc.). And what about software that now only runs in the cloud which you can't own? (like Photoshop 2017)

Dream world includes easy to use, hardware independent bt hardware-clone-compatible containerized emulation; grandfathered licensing when old versions were owned but new ones are "rented"

The dream: "See this system I'm typing on at the moment? Make a containerized cloud copy of it for me. Oh, and do it in just a couple clicks. Then I can get on with my actual creative work / research work"

Q (from presenter): Is this preservation?
A (from audience): yes. At least personal preservation. Much content being preserved must still be re-used.

A (from audience): regarding dream world - we're almost there

Digital Preservation Training Needs Assessment

Sarah MASON and Edith HALVARSSON. Designing and Implementing a Digital Preservation Training Needs Assessment: Findings from the Bodleian Libraries' Institutional Repository (S)

Sarah background on DPOC, the training pilot:
Training pilot background:
Two rounds of semi-structured interviews
Round 1 Winter 2016, Round 2, Spring 2017-09-27

What do we mean by digital preservation skills?
Sarah & Edith did an extensive literature review
Selected DigCurV framework from lit review: it has three lenses, Executive, Manager, Practitioner with progression pathways. Also provides 110 different skills and attributes – used these to customise final list.
Next step: map DigCurV to other frameworks.
Discovered during literature review: core DP skills – "technical skills", metadata standards, communication skills, domain specific and digital preservation knowledge, project management and preservation planning skills, Understanding the "designated community's access and research needs, legal frameworks

Round 1 Interviews: Findings & Trends

Edith on the first round of interviews:
Interview various staff from the Oxford University Research Archive (ORA) (Repository for scholarly outputs)
Practitioner questions used, and manager questions used for different staff levels
Findings: all staff were strong in traditional library skills (metadata editing, communication, legal frameworks, self-directed learning
Gaps: understanding how digital preservation fits into the ORA service, but had a good grasp of types of digital preservation risks.
On digital preservation specific knowledge: what this actually is is up for debate! The approach we took was random selection of terms – asked staff do you recognise these terms? Do you know what they mean? Only one member of staff recognised all of the terms, this person was the only one who had completed a post-graduate qualification in digital curation.

Sarah on round two of interviews:
Interviewed developers: 6 software developers from Bodleian repository. They had a good understanding of metadata standards, data models
None were familiar with digital preservation language - but knew it in different terms. Highlighted need for common language.
Key finding: communication has become a key skill across all roles, even more than technical skills.
Future work: developing an in-house training program based on the findings. Will give staff common language, and is a starting point for future education offerings.
Sarah: I’ve run some classes on personal digital archiving - has been a good starting point for awareness raising among staff.
Questions for the audience: what do you find is the most important digital preservation skill that you have? What digital preservation skill is lacking in your organisation?

Audience Q&A: Digital literacy is still a big problem, digital preservation is so far away from the basic upskilling for librarians etc. in some contexts.

Digital Dunhuang

DAM System put in place
Facilitates asset creation & cataloging of image & video & text files
Manages high-resolution master files & original documents
Tracks preservation actions
Has version control

Was hard to convince people that digital preservation starts at creation

PDF/A
Slides: https://speakerdeck.com/mklindt/a-considered-harmful-for-digital-preservation
Let’s look at PDF/A-a

Constrained version of PDF/A
We have to create a second structure tree parallel to the tree for drawing the pages in order to create structure
→ add 11 new objects (and altered 3) to our PDF document just to add the logical structure
Clearly shows that accessibility was specified as an afterthought for PDF

Creation vs Conversion
PDF/A A-level conformance depend on (structural) information available in creation context
→ if not present, the information CANNOT be generated

PDF2 - tries to remove ambiguities in spec, tagging support aligned with PDF/UA for better accessibility → will most likely result in PDF/A-4

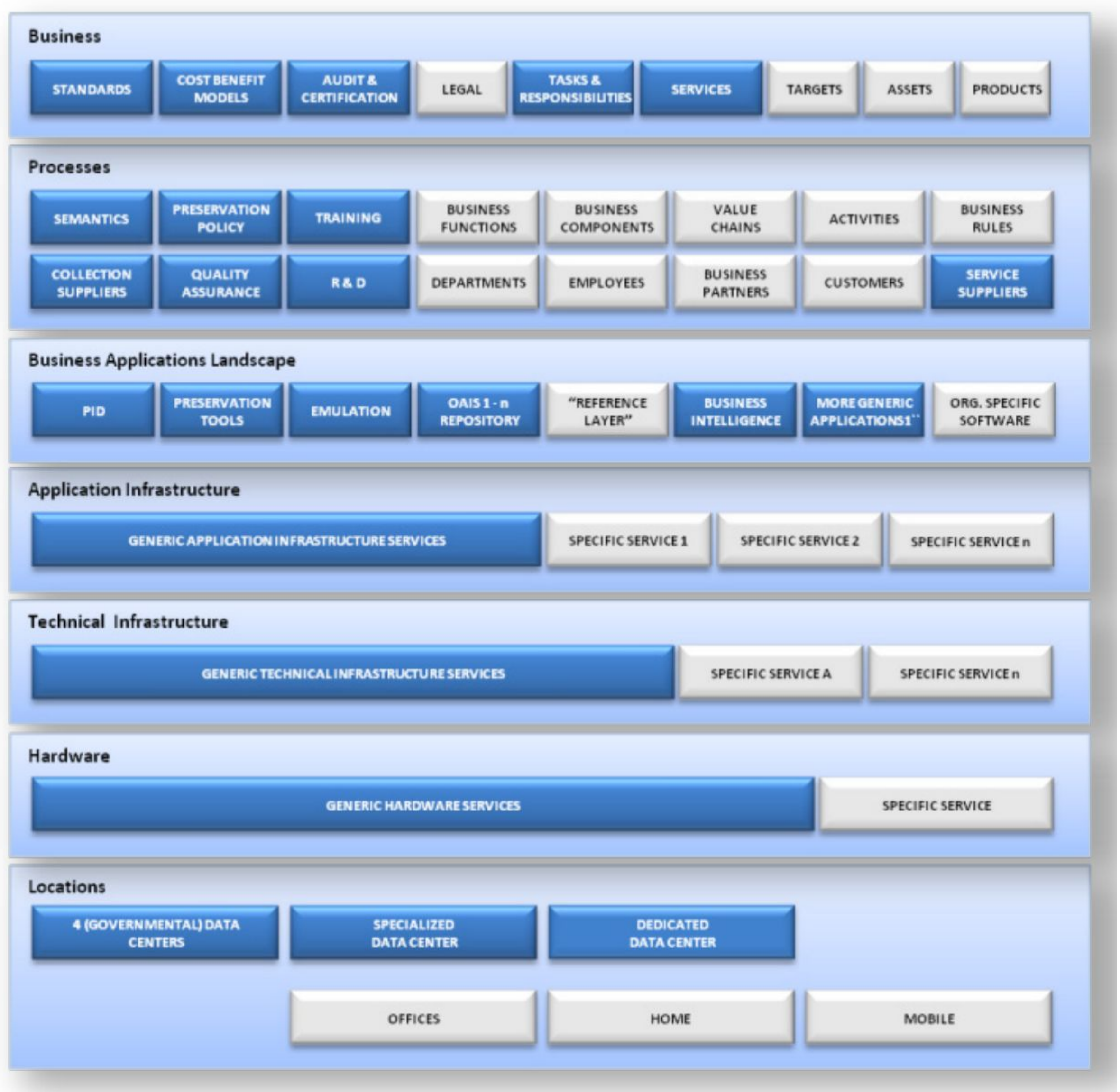
Nationwide facilities

Joost van der Nat and Marcel RAS. A Dutch approach in constructing a network of nationwide facilities for digital preservation together
Linked open data is likely a mechanism that can connect silos.

Goal of the project to increase effectiveness. Defining “completely effective” in a scenario where smaller institutions can keep digital materials for as long as required.

What is infrastructure when you’re talking about digital preservation? ICT component, OAIS box needs to be embedded in archive management – if you don’t start at the beginning, applying the correct metadata, you won’t find it easy to digitally preserve.

Building blocks for digital preservation



Business

STANDARDS | COST BENEFIT MODELS | AUDIT & CERTIFICATION | LEGAL | TASKS & RESPONSIBILITIES | SERVICES | TARGETS | ASSETS | PRODUCTS

Processes

SEMANTICS | PRESERVATION POLICY | TRAINING | BUSINESS FUNCTIONS | BUSINESS COMPONENTS | VALUE CHAINS | ACTIVITIES | BUSINESS RULES

COLLECTION SUPPLIERS | QUALITY ASSURANCE | R & D | DEPARTMENTS | EMPLOYEES | BUSINESS PARTNERS | CUSTOMERS | SERVICE SUPPLIERS

Business Applications Landscape

PID | PRESERVATION TOOLS | EMULATION | OAIS 1 - n REPOSITORY | "REFERENCE LAYER" | BUSINESS INTELLIGENCE | MORE GENERIC APPLICATIONS1** | ORG. SPECIFIC SOFTWARE

Application Infrastructure

GENERIC APPLICATION INFRASTRUCTURE SERVICES | SPECIFIC SERVICE 1 | SPECIFIC SERVICE 2 | SPECIFIC SERVICE n

Technical Infrastructure

GENERIC TECHNICAL INFRASTRUCTURE SERVICES | SPECIFIC SERVICE A | SPECIFIC SERVICE n

Hardware

GENERIC HARDWARE SERVICES | SPECIFIC SERVICE

Locations

4 (GOVERNMENTAL) DATA CENTERS | SPECIALIZED DATA CENTER | DEDICATED DATA CENTER

OFFICES | HOME | MOBILE

Day 4: Thursday, 28 September

PIDs and digital preservation

Angela DAPPERT and Adam FARQUHAR. Permanence of the Scholarly Record: Persistent Identification and Digital Preservation – A Roadmap
Reuse over space and intent has the same problem shape as reuse over time: message is we need to use the lessons that we have learned in digital preservation – we've learned a lot in 15 years, and PID systems should take these lessons in board as there are gaps in current practice.
Pid services came to be with a clear purpose, when they reached a goal they expanded their practice/scope. – in digital preservation we've done the opposite – lots of thinking – what's

the concepts behind what we do – and because PID systems grow organically, the data models are not as extensible and interoperable as they should be.
Examples
We use premis as a main data model – agent, rights, events. We never talk about an object without clarifying what it is we're talking about.
PIDs – don't do this. Don't do machine harvesting or machine interpretation.
Recommendation: PId services need to rethink the underlying data models especially if we want to harvest the research output.
Technical metadata – tech metadata needs to be created as early as possible in the information network eg file types, creating software, computing platform
Another example
PIDs concerned with partial and dynamic datasets – researchers work with cleaned subsets of base data – if we use premis it can identify clearly the derived dataset – so that it can be computed on demand from the original data set.
Provenance information in the scholarly record: crossref and datacite events are no based on a common data model like we have in digital preservation
Lesson: need to translate digital preservation data modelling into deliberate design for PID services – locally, in apis, in web environment

Operational Pragmatism in Digital Preservation - Establishing context-aware minimum viable baselines (PANEL)

The panel consists of Anthea Seles (US), Andrea Byrne (UK), Jones Lukose (Netherlands), Bertrand Caron (France), Dr. Xiaolin Zhang (China)

Question 1: What does digital preservation looks like if you do not have the skills available in your organisation? How can you do something, and what is good enough?

Andrea S.: Need to be clever around how we use our skills and decide what we will/will not support. These are hard questions which we need to ask ourselves.

In terms of my experience in Africa, and what my African colleagues are facing, there is no skills training. There is no infrastructure to train staff.

Andrea B: DP is a learning curve - you never know enough and there is always something new. The important thing is that we need to inspire staff to continue learning and being curious. Project management training with staff is the kind of thing we may want to focus on.

Jones: Grew up in Kenya (left 12 years ago), and there has been a large digital revolution at home in the last five years. Businesses etc. are moving completely away from paper. This is the right time to go back to Kenya to try and change the mindset of this new generation.

Bertrand: For us at the BnF, we need to become better at sharing. It would be good to be able to give staff something similar to a "competency map" to display what people's' skills are. We also need basic training with all staff, go give them a minimum level of DP skills.

Xiaolin: Sometimes we are approaching this lack of skills problem in the wrong way. I am running a national DP project. Staff want simple, quick, hassle free training - they do not want to “become us” [as DP professionals]. Two hours maximum training is probably fine. Then when you put processes in place, you will end up forcing that skills up as well.

Comments from the audience:

- Jay G. NZ - agrees that everyone does not need to know everything, they are instead focusing on basic digital literacy. They need to know enough that they do not get in my way, and I do not get in their way. Also has a monthly digital talk at NZ, where DP is sometimes covered.

Question 2:

OAIS and Distributed Digital Preservation in Practice

Eld Zierau

Why was IO-OO model necessary in Denmark? Denmark wanted one central bit preservation system used by several organizations spread across Denmark. The joint project started in 2009. Goals was full independence within the system for the different organizations, multiple copies, etc.

OAIS was chosen as the common language to model the system - however, it became confusing when processes relating to the OAIS functional entities take place either within the shared bit storage system or within the repository system in use at the respective organization (e.g, where migration takes place).

Time has come now to audit the Bit Repository (IO level).

Examples of the audit.

IO should be independent personal - e.g. if someone has too many administration rights, this needs to be fixed.

Independent Operating systems - Windows and Linux servers

Independent Organisation - currently looking at the implication of libraries merging

Other examples of checks:

Preservation Planning: including media

Audit of Bit Repository - IIO level

Bit integrity challenge: Is the trigger for changing tapes set right?

Specific confidentiality challenge - e.g. external service provider has direct access to system, this is a no-go.

Other cases who use the OO-IO model / have been informed by DDP project: MetaArchive, BitRepository.org, DuraSpace, Chronopolis, etc.

IO-OO can be used for other processes as well, e.g. ingest

Preliminary Ingest

“Minimal effort ingest” (see iPRES2015) used in first step to be able to ingest quickly instead of having to wait for long curation time, etc.

This can be modelled as OO-IO as well - this can be modelled via IO-OO as well. When SIP is ingested via minimal effort ingest, is becomes an AIP in the Bit Repository but is still being worked on outside (e.g., to finish curation processes, gather metadata).

Helen Hockx-Yu - University of Notre Dame

Superb Stewardship of Digital Assets

Hockx-Yu job: based on the expectation that stewardship requires close collab between IT and library

Gap analysis has been done looking at the current status of data stewardship at the university

Digital assets are records, data, resources typically owned by the univerity & thought of as having value - no discrimination between born-digital & digitized

3 categories: uni records, research data, resource for teaching/learning/research

Gap analysis put forth that UND is in good position to address gaps but there's currently a siloed approach with a lack of coordinated, cohesive view of digital assets and uneven use of tech across orga. There's a strong focus on "now" and some assets have already been lost or are at risk.

Recordings at risks are things like student radio recordings on open reel tapes including performances from Alan GInsberg etc.; 1967 Sister Survey where coded responses from 130,000+ sisters were in an inaccessible file format.

Key challenges identified:

- Lifecycle management
- Digital archiving: collect and retain the assets (of critical importance to the future)
- Digital preservation: maintain these assets so that they remain accessible and usable

Recommendations based on analysis:

-Strategy, policy and organization

- add "digital" as new category of assets to University's Strategic Plan for which superb stewardship is required
- move away from task-force or project based approach
- embed archiving and preservation considerations in business processes

-Storage and Cloud services

- immediate goal: reduce the use of direct-attached devices as a long term storage solution
- explicitly define treatment of data for services in the Cloud

LIbrary and Archive specific

- establish a digitisation and preservation centre, based at Hesburgh Libraries, to coordinate and serve campus needs
- make informed decision on archiving assets residing on the web & social networks
- build capability and transition to electronic records management

Research

Recommendations were then prioritized using Prioritisation techniques

Now in the planning phase

Q: How are your relationships with various parts of the organization? Library? IT?

A: Nobody has been that rude to me … just yet. It's important that people see me as part of the library AND part of the IT. Very important that I have support from top-down. And I've really taken some time to dig into the resources. I've also come across issues - digital requires cultural change.

Certification
Barbara Sierman

Sustainability Program - is the Digital Preservation Program, but it was felt that "Sustainability" is more understandable

Starting point is understanding that Certification is beneficial. It's rewarding for your own organisation. But it's also beneficial for the preservation activities in the Netherlands in general. If you want to have a large cross-institutional infrastructure, you need to trust each other and it's better if you can proof the trustworthiness. While Certification is currently voluntary, we except it might be required in the future, e.g. by research funders. Most Sustainability program partners are currently preparing themselves for the CTS (formerly DSA) or nestor Seal.

We have a roadmap to certification:
1. Selfassment score model (tool based, developed in conjunction with Flemish colleagues - tool asks basic questions like "how many objects do you have?", "do you have two copies?" and also explains why these things are important - so in a way it's a handbook as well)
2. Exploratory phase
3. CoreTrstSeal
4. Nestor / DIN
5. ISO 16363 Trustworthy repositories

DSA Survey was done asking those already certified how much time it took for them to receive the seal, what the hurdles were, if they intend on going for the other levels, …
Link to survey:
http://www.ncdd.nl/wp-content/uploads/2016/10/201611_DE_Houdbaar_Report_DSA-survey_2016.pdf
Few people said they would go for the nestor seal, no one said they would go for ISO.

A huge achievement of this Program was that now a large digital preservation network in the Netherland exists with over 100 people.

13:00-14:00 @ Main Hall

Keynote (3): Endeavors of Digital Game Preservation in Japan- A Case of Ritsumeikan Game Archive Project

Initially started in April 1998 as a personal project of Professor Koichi Hosoi

Objective was to create a research platform which allow scholars from multiple disciplines to study video games (e.g. economics, sociology, computer science, …) and to collect all types of resource about video games from hardware to software.

First Task completed in 2004: create a digital database for 1769 Famicom titles (Loaned from Nintendo) → no longer exist

3 fold approach:
1. Preservation of actual software and hardware
2. Preservation using emulator system
3. Preservation of playing screen images / filming

2003: Famicon Digital Library (FDL) - developed under the permission from Nintendo - the only outside party to do so (only 2 titles: Donkey Kong and Mario Brothers)

Launched Industry-Academic Collaborative Conference: Game++ Digital Ingeractive Entertainment Conference 2005 (brought together developers / creators from Atari, NES, Pacman, Super Mario, Zelda, Metal Gear Solid series, Half Life 2, etc.

Recording Play Images
In order to distinguish one title to the next, including the traits of each, recording visual image considered to be of significance. Aim at reusability for research, the game play control along with images of game players and the game play are recorded simultaneously and archived as one set of the game play record.

Current situation:
6263 Titles from approx. 14 consoles (PlayStation, SEGA Saturn, Super Famicom, PC Engine, FamilyComputer, PlayStations2, Dream Cast, Play Statio Portable, Nintendo DS, SEGA Mega Drive, Gameboy Advanced, Nintendo GameCube, Game Boy).

Status of data carrier/hw Video Game Preservation at Ritsumeikan:
Not emulated due to legal restrictions
- all products located at RCGS (3 rooms)
- one room designated for dedicated use of storage, 24C humidity 50% - this is for high priority preservation
- QR code placed on each item.

Since 2012 Game Archive Project has been selected as the official partner for creating game section of Media Arts Database, which currently includes:
38.042 console games
  5.018 arcade games
  1.623 PC games
In total: 44.683 titles
Link: https://mediaarts-db.bunka.go.jp/help/gm/help.html?locale=en

Current research:
- project evaluating who collects Video games in Japan and beyond and of what years (RCGS, Leipzig Univ., Meiji Univ., Strong Museum, NDL, Japan Game Museum)
- developing and analyzing ontology in Japanese video games titles (e.g. length of titles, number of hiragana, number of katakana, etc.)
- regional differences in perception of Japanese game titles

Q: what kind of strategy do you have for preservation of online games and apps like PokemonGo?
A: we have a master student currently doing some work on online games. Problem is version changes frequently, sometimes daily and you're not even informed (e.g. patches). This is impossible to track and that's a huge issue. Apps are probably even more difficult because there's a higher server client dependence. Maybe for the online games we need to preserve

them from a journalistic point of view - rather record and document experience etc. and not actual preservation of all version-ups. That's most likely not possible unless the company will contribute to that.

Q: Do you see more commercial cooperations in the future?
A: Several years ago they gave us Donkey Kong because they thought it was old and there was no market - but since there have been re-releases and every game is still commercially available now. So no, I don't see any more cooperations.

Workshop:

What is Preservation Storage?

Shared notes at https://goo.gl/qMEoZH

Preservation Storage Criteria v2 at https://docs.google.com/document/d/1Ko7JwgNFf5KCnyQJ3sSY2d1T2wfSoRLIQYn-AjNiO-E/edit

Day 5: Friday, 29 September

Curating Digital Content with Fedora
by David WILCOX

Fedora is a flexible, extensible, open source repository platform for managing, preserving, and providing access to digital content. This workshop will provide an introduction to Fedora 4, including a feature overview, data modelling best practices, and a tour of the import/export utility.

Preparation instructions Here.
Related info Here.

Events info
https://wiki.duraspace.org/display/FF/2017-09-29+iPRES+Curating+Digital+Content+with+Fedora+Tutorial

Agenda/Presentations

Time	Topic
9:00 - 9:20	Welcome, introductions, VM setup
9:20 - 10:20	Introduction to Fedora
10:20 - 10:30	Break
10:30 - 11:20	Linked Data Basics and Data Modelling with PCDM
11:20 - 11:50	Digital Preservation Services
11:50 - 12:00	Wrap-up and Discussion

Q: Anyone uses wikipedia on top of Fedora?
A: Wikipedia pulling data, yes. As far as I know, no one has wikipedia reading and writing into Fedora. Theoretically it's possible. Might need to implement locking to prevent multiple people editing at the same time.

Working with WARCs

New Tools for Harvesting, Accessing, and Researching Web Archives
Vinay Goel,
Jefferson Bailey

Working with WARCs
New Tools for Harvesting, Accessing, and Researching Web Archives
Vinay Goel | Senior Data Engineer | vinay@archive.org
Jefferson Bailey | Director, Web Archiving | jefferson@ar...org
Web Archiving Group | Internet Archive | IPres 2017
INTERNET ARCHIVE
ARCHIVE-IT

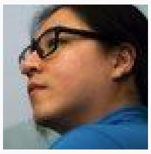
Slides at http://bit.ly/ipres-warcs
API part at http://bit.ly/wa-apis
ArchiveIT Research http://bit.ly/ait-research
Archive Research Services workshop https://github.com/vinaygoel/ars-workshop

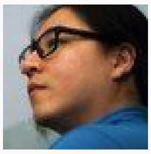
If you have the chance before the workshops, please visit the Archive Research Services workshop repo and complete the "Initial Setup" to install the core pieces (Git, Docker, the repo) for that portion of the workshop: https://github.com/vinaygoel/ars-workshop#initial-setup

Application Programming Interface (API)

API part at http://bit.ly/wa-apis

Joshua Ng (@joshuatj)	29-Sep-2017 04:22

Not many digital humanities research interest on web archives. Why is it so? #iPRES2017 https://t.co/NYjDPckVoz	

Joshua Ng (@joshuatj)	29-Sep-2017 04:26
.@vinaygo:"Just looking at header information in WARCs has helped longitudinal studies e.g. #webhistory #fileformat" #iPRES2017	

WAT record is a WARC derivative, contains only metadata, stripped of main content.
A WAT file is a derivative file created from a WARC but more lightweight and containing only key data and not the full resource information (such as page text).

Now that I have a WARC file, what do I do with it?

Archives Research Services Workshop https://github.com/vinaygoel/ars-workshop

1. Install Git
2. Install and Run Docker
3. Download Workshop

What kinds of derivatives are there?

- `WARC`
  - `CDX`: Plain Text file with one metadata line per WARC record
  - `WAT`: Metadata WARC file with one metadata JSON record per WARC record
  - `Parsed Text`: Plain Text file with one *<Key, Value>* line per HTML document in the WARC. *Key=<URL, Digest>*, *Value=JSON* containing the parsed out plain text and hyperlinks
    - `LGA`: Linked Graph Data. Plain Text files with one JSON line per HTML document. *Map* file that maps URLs to unique identifiers and *Graph* file that lists the set of URLs (identifiers) that a HTML document links to.
    - `WANE`: Web Archiving Named Entity. Plain Text files with one JSON line per HTML document. JSON contains the set of named entities extracted from the document.

Text analysis, use `Parsed Text` derivative.
LGA, Linked Graph Analysis.

PREMIS tutorial

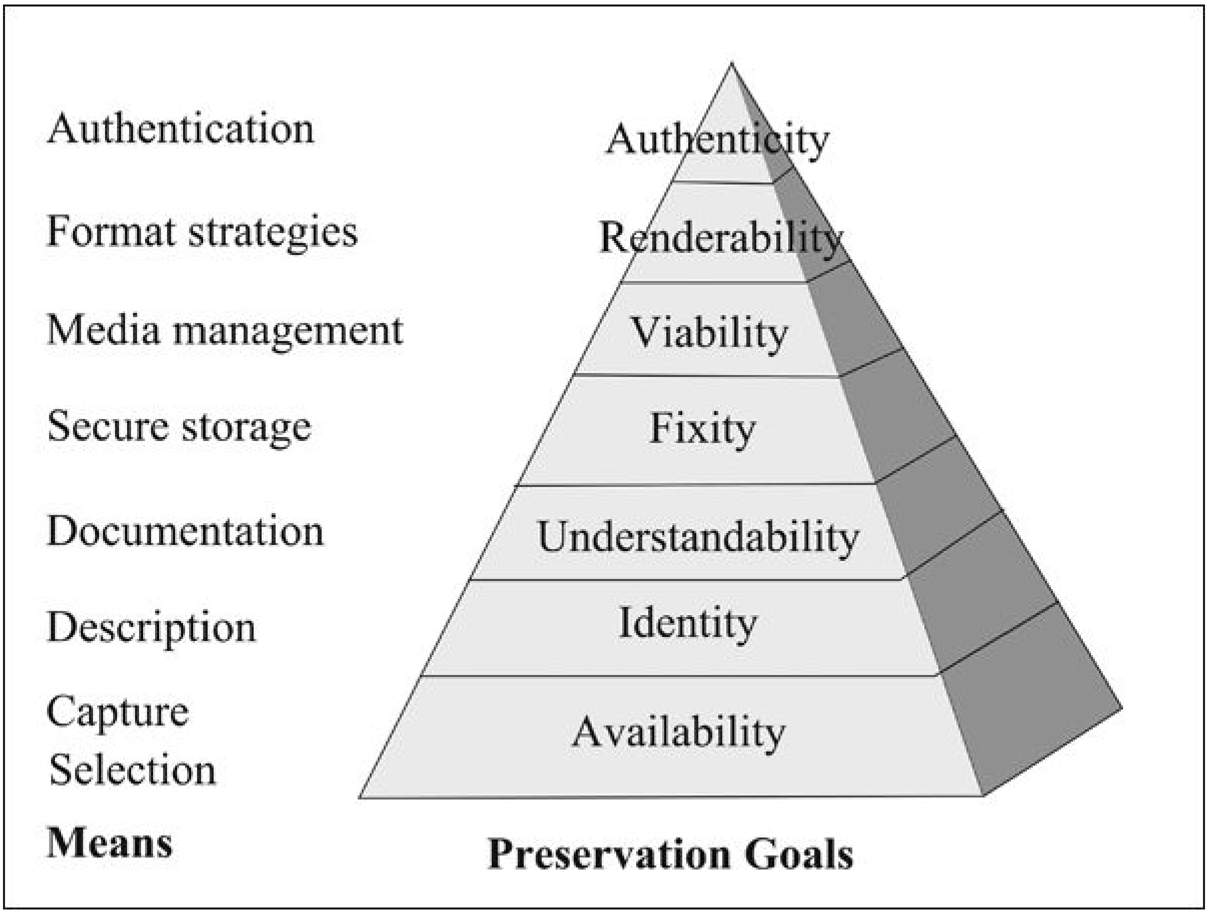
13:00-16:00 Tutorial 3 Understanding and Implementing PREMIS Karin BREDENBERG, Angela DAPPERT and Eld ZIERAU

Background and what is PREMIS
What is premis?
International de-facto standard for metadata to support the preservation of digital objects and ensure their long-term usability – if you want to support more services you may need more metadata. Implemented in digital preservation projects around the world, in commercial and open source digital preservation tools and systems.

PREMIS data dictionary: http://www.loc.gov/standards/premis/v3/

Digital objects must be self-describing.
Important questions: when you do your preservation metadata: where do put it and how do I encode it?

Using preservation pyramid from Priscilla Caplan



Availability: basis of being able to preserve. Object is in our control or in control of trusted accessible repository – cloud use depends on policy, legislation.

Identity: each relevant entity is persistently uniquely identified – file, work, person, organisation, licence – identifier type and identifier value

Understandability: file names not enough, files might not be readable –PREMIS offers ability to record physical structure, so does METS – can capture info like reading order to make sense, whether files are embedded in other files. Knowing about the logic to make sense – knowing title, etc to make sense. Need to know the context: where it came from – original source, related items.

Fixity: the object is unchanged – the order of 0s and 1s are the same, no bit rot over time or that it has been transferred properly.

Viability: the object remains readable. What data carrier is it stored on? Age of medium, date of recording, did I read and write on it a lot so it will decay sooner?

Renderability: able to render or execute the object – barrier of technology, lots of dependencies and need format information, rendering information (computing environment hardware software) – either you want to preserve the environment or you need enough information to see what the original was to see how it can be migrated to in a new tech situation.

Authenticity: digital objects are always undergoing object transformations and there is danger with changes like bit migration, content migration, replacing part of the rendering stack, forensic transformation actions.

Using PREMIS as a checklist: be able to think about what metadata you need for your repository to check if potential vendors can satisfy those needs. Understanding what your needs are according to policy, org context is very useful before checking vendors’ offerings.