DIGITAL PRESERVATION CHALLENGES IN LIBRARIES

Alenka Kavčič-Čolić

R&D Division, National and University Library - NUK
alenka.kavcic@nuk.uni-lj.si
1. The National and University Library (NUK)
2. Digital preservation in brief
3. How libraries could contribute?
4. Recommendations
5. Open issues
The National and University Library of Slovenia

Photo: Cvetkovič

Photo M. Štupar
FUNCTIONS OF NUK

- National library – depositary institution, preserves written heritage of Slovenian nation
  - Legislative basis (Legal Deposit Law, Strategic Plan 2015-2019, NUK Digital Preservation Strategy 2012-2020, Slovenian national strategies ...)

- University library – serves the University of Ljubljana (approx. 40,000 students), mainly in the field of human sciences
• Task under Digital Agenda for Europe – Cultural Heritage
• Cooperation between libraries, archives and museums – Social Sciences Data Archive (ADP)
• Digital preservation according to ISO 14721/2003 (2012)
NUK total online digital collection comprises more than 4 million scans, including:

- 99,842 issues and articles of scientific journals,
- 617,857 issues of newspapers and magazines,
- 11,435 books
- 10,703 photographs,
- 112 music records,
- 8,361 posters
- 1,180 theses
- 1,383 manuscripts
- 2,220 research reports
- 734 postcards
- 15 3D objects
- 14 virtual exhibitions
- 693 maps
- 35 films, etc.
Spletni arhiv Narodne in univerzitetne knjižnice vsebuje spletna mesta, ki jih zaradi trajnega ohranjanja slovenske kulturne dediščine na svetovnem spletu periodično shranjujemo od leta 2008. Do arhiva je možno dostopati s pomočjo iskanja po celotnem besedilu ali po URL naslovu ter z brskanjem po tematskih kategorijah ali abecednem seznamu naslovov spletnih mest. Različice vsakega spletnega mesta so urejene po datumu izvodev zajema.

Področja

Družba
Gospodarstvo in industrija
Humanistika
Izobraževanje in raziskovanje

Množični mediji
Narava in okolje
Naravoslovje in tehnologija
Prosti čas, turizem, potovanja

Šport in rekreacija
Umetnost in kultura
Vlada, politika, pravo
Zdravje in medicina

Abecedni seznam

0-9 A B C D E F G H I J K L M N O P R S Š T U V X Y W Z Ž prikaži vse

© Narodna in univerzitetna knjižnica, 2016. Vse pravice pridržane.
SLOVENIAN WEB ARCHIVE (2007 - )
HTTP://ARHIV.NUK.UNI-LJ.SI/

A. Kavčić-Čolić, Staff Training Days, 17 May 2017
IT AND DIGITAL LIBRARY DIVISION

- Small development team
- Developing and maintaining dLib.
- Maintaining Web Archive
  - Started 2007
  - Thematic collection
    - 1500 websites (79 twitter streams)
    - 100 bilion URLs
    - 20TB of data
  - Development
    - SwitchProxy: proxy like tool for interpreting dynamic web pages (Twitter, FB)
    - WaybackAnnotator: portal for using and managing archived sources
- Maintaining and supporting other NUK services

CRAWL OVERVIEW

60 107 035
Number of URL accessed

3 101 707 944 890
Size of harvested material

68 098
Successfull harvested domains

Duration of 112 days 12hours 55minutes 46seconds (and 648ms)

HOW MUCH OF WEB WE SAVED

- Small revisit (August 2016)
- 89,464 domains
- 17,760 not responds
- 386,9292 archived dead URL
- 170 GB of inaccessible content

8% of archived content is not on web anymore

IF WE REALLY WANT TO PRESERVE OUR WRITTEN HERITAGE, WE NEED TO COOPERATE WITH OTHER LIBRARIES AND INSTITUTIONS

WE CANNOT DO IT ALONE!
1. The National and University Library (NUK)
2. Digital preservation in brief
3. How libraries could contribute?
4. Recommendations
5. Open issues
WHY DO WE NEED TO PRESERVE DIGITAL CONTENTS?

- Is part of the scientific/organizational/cultural heritage
- Its production grows exponentially
- Based on IT and networks
- Needs adequate resources for access (IT generation/experts & constant investments ...)
- IT obsolescence and fragility ...
McLeod (British Library) identified risks to digital objects across the collections which could be classified in 8 categories:

- Physical deterioration of media  ➔ **most critical!**
- Technical obsolescence of media;
- Technical obsolescence of file formats;
- HW obsolescence;
- Operating system’s obsolescence;
- Application SW obsolescence;
- Lack of institution’s policy;
- Unsufficient metadata.

(McLeod, R. (2013). Risk Assessment; using a risk based approach to prioritise handheld digital information - [https://fedora.phaidra.univie.ac.at/fedora/get/o:294139/bdef:Content/get](https://fedora.phaidra.univie.ac.at/fedora/get/o:294139/bdef:Content/get))
DIGITAL PRESERVATION

➢ „... activities necessary to ensure continued access to digital materials for as long as necessary.“

(Digital Preservation Coalition - http://handbook.dpconline.org/glossary)

➢ „Digital preservation combines policies, strategies and actions to ensure access to reformatted and born digital content regardless of the challenges of media failure and technological change. The goal of digital preservation is the accurate rendering of authenticated content over time.“

(ALA -http://www.ala.org/alcts/resources/preserv/defdigpres0408)
DEFINITIONS

DIGITAL CURATION

“... maintaining, preserving and adding value to digital research data throughout its research lifecycle.” – ONGOING PROCESS! (Digital Curation Centre - http://www.dcc.ac.uk/digital-curation/what-digital-curation)

DIGITAL ARCHIVING

LONG-TERM PRESERVATION ...
MAIN FUNCTIONS CONCERNING DIGITAL PRESERVATION

- **COLLECTION** (gathering, harvesting, reformating/digitisation, etc.)
- **PRESERVATION / CURATION** (policies, strategies & actions through the entire lifecycle to ensure future access)
- **ACCESS** (usage, reuse)

→ **ISO 14721** (rev. 2012) OAIS
Figure 2: OAIS Functional Model

Source: DPC Technology Watch report: http://dx.doi.org/10.7207/twr14-02)
COLLECTION LEVEL:

- Acquisition of:
  - Born-digital contents – no analogue equivalent
  - Digital material originated from analogue originals (through digitisation)
  - Digital materials originated from a digital source but have been printed to paper (e.g. electronic records)
  - Digital material collected through harvesting (web pages and other web contents)
- Collecting relevant information for preservation activities
- Quality control

Figure 2: OAIS Functional Model

Source: DPC Technology Watch report: http://dx.doi.org/10.7207/twr14-02
PRESERVATION/ CURATION ACTIVITIES:

- Production of reliable archival/master files
- Managing different versions and formats
- Enriching contents with descriptive, administrative and structural metadata to ensure future access
- Documentation of all policies, strategies and procedures
- Use of persistent identifiers
- Record provenance and change history for all objects
- Verification mechanisms

• Attention to security requirements
• Routine audits
• Trusted digital repository system
• Storage and synchronization of files at multiple sites
• Continuous monitoring and management of files
• Programs for refreshing, migration and emulation
• Creation and testing of disaster prevention and recovery plans
• Periodic review and updating of policies and procedures

ACCESS LEVEL:

- DRM (CC, moratorium, standard publisher’s copyright ...)
- Descriptive metadata for search and browsing interface functionalities and indexing
- Access files and technology
- Exchange of metadata and contents with other external systems
- Statistics production ...

Source: DPC Technology Watch report: [http://dx.doi.org/10.7207/twr14-02](http://dx.doi.org/10.7207/twr14-02)
THE SELECTION OF PRESERVATION STRATEGY/METHODS DEPENDS ON WHAT WE CONSIDER IS IMPORTANT TO PRESERVE!!!
STRUCTURE OF DIGITAL OBJECT

INFORMATION SPECIFIC PROPERTIES
(way of access, functionalities, colour, sound, etc.)

CONCEPTUAL LEVEL
(meaningful unit of inf.: image, table, text ...)

LOGICAL LEVEL
(application softwer)

PHYSICAL LEVEL
(Signs Inscribed on a Medium)

EXAMPLE OF TWO DIFFERENT LOGICAL LEVELS

Hexadecimal Dump of MS Word

Hexadecimal Dump of PDF

PLANNING THE PRESERVATION OF THE:

➤ LOGICAL LEVEL (eg. emulation, museum approach, reincarnation of old HW ...)

➤ CONCEPTUAL LEVEL (eg. migration of versions, conversions etc.)
Are the content and functionalities more important than the look like?

Or we should consider both?
• Archival master files – shown in cont.
• Metadata – all relevant information for accessing the digital objects in the future
• Unique Identifiers – URN, DOI, handle … comparable to bibl. ident.
• Deposit in a trusted digital repository system
1. **Access file formats** – thumbnail & better quality user access format

2. **Archival formats** – for the master files – could be or not the original formats!
<table>
<thead>
<tr>
<th>Type of content</th>
<th>RECOMMENDED ARCHIVAL FILE FORMATS</th>
</tr>
</thead>
</table>
| **TEXT**        | **PDF/A-1** Part 1 = PDF 1.4 2005-10-01 ISO 19005-1 PDF 1.4 (Adobe Systems, PDF Reference, third edition)  
**PDF/A-3** Part 3 – use of ISO 32000-1 with support for embedded files 2012-10-15 ISO 19005-3 PDF 1.7 (ISO 32000-1:2008)  
JPEG 2000 (ISO 15444-1) not more than 1% of compression |
| **IMAGES**      | **PNG** (ISO/IEC 15948:2003)  
**JPEG 2000** (ISO 15444-1)  
**TIFF 6.0** (ISO 12639) |
| **AUDIO**       | **MPEG-2** layer 3 (MP3) (preferred fidelity: 192 kbit/s @ 44.1 KHz; min. fidelity: 128 kbit/s @ 44.1 KHz) |
| **VIDEO**       | **JPEG 2000 MXF** (Motion JPEG 2000) in 4K (35mm), 2K (16mm), 16 bit in each color (=4TB/h) |
OTHER RECOMMENDED FILE FORMATS

Are not standards, but have rich documentation:

**WAVE** (Broadcast WAVE Format – BWF)

**ICC.1** – International Color Consortium color profile format

**GeoTIFF**

**SPIFF** – Still Picture Image File Format (IDO/IEC 10918-1)

**JFIF** (JPEG File Interchange Format)
JPEG 2000 (JP2) (ISO 15444-1)

- Good alternative to TIFF for large images (saves 50% of storage) - it can manage very large images in different resolutions
- It can implement very high resolutions in some parts of the image (ROI – region of interest)
- JP2 can store color information at 48-bit color depth
- It can store metadata inside the file format
- We can manage IPR (prevent downloading of the whole image)
- If corruptions in the header – freeware available for its fixing.
Progressive subband decomposition avails „multi-resolution“ image:

„Visually lossless“ mode feature of JP2 – potential losses in RGB spectrum are not visible to human eyes

(Source: http://upload.wikimedia.org/wikipedia/commons/7/78/JPEG_2000_Artifacts_Demonstration.png)
Buonora & Liberaty tested robustness of JPEG 2000

Possible errors appeared in:

- Bit errors
- Packet losses (during transmission)

Error 0.01%  Error 0.1%  Error 1%

A corrupted JPEG 2000 image, compression ration 1:20 (1KB of loss)

JP2000 header: highlighted values represent image height

Impact on the image of corrupted header height values

FixIt! JPEG 2000 is a freeware and shareware utility implemented by the Laboratorio Digitale of the Centro di Fotoriproduzione, Legatoria e Restauro of the Italian State Archives, and can be downloaded from the website: http://softwaretopic.informer.com/fix-it-2000/

PDF/A

- PDF – *Portable Document File* – developed since 1993 by Adobe Systems
- Open format – rich documentation
- PDF/A - archival version (ISO 19005).
- Stores the content as it was created, together with its structure
- Permits full-text searches
- Different font could be embedded (excellent for OCR).
- Device independent
PDF/A

- **PDF/A-1** (ISO 19005-1:2005)
  - **PDF-A-1a**: includes features of **PDF/A-1b** and features intended to improve a document’s accessibility (language specification, hierarchical document structure, tagged text spans, character mapping to Unicode ...)
  - **PDF/A-1b**: includes reliable reproduction of a document’s visual appearance
- **PDF/A-2** (19005-2:2011) – includes JPEG 2000 image compression, support for transparency effects and layers, embedding of Open Type fonts, provisions for digital signatures, option of embedding PDF/A files to facilitate archiving of sets of documents with a single file.
- **PDF/A-3** (19005-3:2012) – same as **PDF/A-2** + it allows embedding of arbitrary file formats (such as XML, CSV, CAD, word-processing documents, spreadsheet documents, and others) into PDF/A conforming documents

(Source: https://en.wikipedia.org/wiki/PDF/A)
The **PREMIS Data Dictionary for Preservation Metadata** is the international standard for metadata to support the preservation of digital objects and ensure their long-term usability.

Developed by the Library of Congress (US), RLG, and OCLC. Maintained by the Library of Congress (http://www.loc.gov/standards/premis/)
1. The National and University Library
2. Digital preservation in brief
3. How libraries could contribute?
4. Recommendations
5. Open issues
3. HOW LIBRARIES COULD CONTRIBUTE?

When developing our digital collections ...

- Do we take care to get quality originals / master files from producers?
- Do we collect all relevant information on the digital contents collected?
- Do we know what are the specific characteristics of the contents that should be preserved?
- Do we deposit these collections in trusted digital repository systems?
- Do we produce quality scans when digitising?
MAIN PROCESSES IN LIBRARIES

- ACQUISITION (PRODUCTION/DIGITISATION)
  - DIGITAL PRESERVATION STARTS HERE!!!

- COLLECTION MANAGEMENT

- SERVICES PROVISION (ACCESS)
• We create a completely new and unique collection
• This collection contributes to the image and realiability of our organization
• Users trust that the this collection will be online available in the future
• We invest a lot of resources in digitisation ...
WE INVEST TOO MUCH IN DIGITISATION!

CATEGORIES OF DIGITISATION COSTS:

- digitisation planning
- staff training
- material’s selection
- scanning, image processing
- metadata production
- OCR, SGML or TEI encoding
- identification, collation, repair,
- handling of original material
- digital material management ...

A. Kavčič-Čolić, Staff Training Days, 17 May 2017
COST OF DIGITIZATION

SOURCES OF FINANCING IN LIBRARIES

- 62% - own resources
- 30% - national or int’l program
- 8% - sponsors

(Poll, 2010)

Sources: Assessing the costs of digitisation, 2001; Bia, A., Munoz, R., and Gómez, J., 2010; Hammond, M., and Davies, C., 2009; Poole, N., 2010; Puglia, S., 1999)
• Digitisation often used for preservation of original physical material by keeping them safe from damage and degradation produced by its use or exposure.

• Usually digital masterfiles are not properly preserved!
SOME OF THE COLLECTIONS CAN SURVIVE IN DIGITAL FORM ONLY!
KEY DIGITISATION FACTORS IMPORTANT FOR PRESERVATION OF MASTERFILES

1. DIGITISATION PLAN
2. SCANNING TOOLS
3. METADATA
4. RESOLUTION AND BIT DEPTH

The longiness of the masterfile life cycle will depends on its production quality
DIGITISATION OF BOOKS
SPECIFICATIONS

➢ Resolution: min. 300 dpi – max. 600 dpi
   ➢ For lower formats (eg. A6) higher resolution (600 dpi)

➢ Bit depth: 1-8 scale for black & white texts; 24 bit depth scale for coloured images. All books before 19th century should be scanned in 24 bit depth scale.


➢ Compression: without or lossless compression
General specifications:

- Resolution of the master file: 300 dpi, TIFF, JPEG 2000;
- Resolution of access files: 120 dpi; 72 dpi JPEG 2000.


A. Kavčič-Čolič, Staff Training Days, 17 May 2017
• Resolution is determined by the number of pixels used to present the image, expressed in dots per inch (dpi) or pixels per inch (ppi)

• Input / output resolution

Original size

50 % of normal size
= 600 dpi

300 dpi

Twice enlarged - 150 dpi
“(...) amount of information (stored as bits) that is represented by a single pixel.” (Zhang & Gourley, 2009, p. 58)

<table>
<thead>
<tr>
<th>Bit depth</th>
<th>Number of colours available</th>
<th>Example of digital values of each colour</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>1, 0</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>11, 00, 01, 10</td>
</tr>
<tr>
<td>4</td>
<td>16</td>
<td>000, 0001, 0010, ..</td>
</tr>
<tr>
<td>8</td>
<td>256</td>
<td>00000000, 00000001, ...</td>
</tr>
<tr>
<td>16</td>
<td>65.536</td>
<td>0000000000000000, 0000000000000001, ...</td>
</tr>
<tr>
<td>24</td>
<td>16.777.216</td>
<td>00000000000000000000000000000001, 00000000000000000000000000000011, 0000000000000000000000111...</td>
</tr>
<tr>
<td>48</td>
<td>281 trillion</td>
<td>000000000000000000000000000000000000000000...</td>
</tr>
</tbody>
</table>

Black and white
Grey scale image
‘true colour’

R -> 8 bits -> $2^8 = 256$
G -> 8 bits -> $2^8 = 256$
B -> 8 bits -> $2^8 = 256$

$2^{8 \times 3}$ or 16.777.216 different colours = 'true colour' (24 bits per pixel)
DIGITISATION

WORKFLOW

DIGITISATION PLANNING

SCAN/CAPTURE SET-UP SETTINGS

BIT DEPTH

RESOLUTION

FILE FORMATS

FILE NAMING

SCANNING / IMAGE CAPTURE

ARCHIVAL FILES

IMAGE PROCESSING

ORIENTATION

CROPPING (1 cm »BORDER ZONE«)

SKewing

STITCHING (not recommended)

BLANK LEAVES (SHOULD BE SCANNED!!)

OCR PROCESSING WITH POSTCORRECTION

METADATA

DIGITAL COLLECTION MANAGEMENT
1. The National and University Library
2. Digital preservation in brief
3. How libraries could contribute?
4. Recommendations
5. Open issues
RECOMMENDATIONS

- Start thinking on your digital collections like digitally born materials
- Integrate the digitisation workflow with the digital preservation workflow
- When planning the overall digitization workflow and requirements should be very well understood by the Head of the digitization project!
RECOMMENDATIONS IN THE ACQUISITION PHASE

- Check the integrity and completeness of the acquired digital publication
- Get and define digital rights from all authors
- Get all contact addresses of authors and publishers
- Be sure that you have the necessary software for accessing the content of the digital publication
- Collect all relevant information on the content of the digital publ., used technology for its creation, IT environment for access, identifiers, etc.
OTHER RECOMMENDATIONS

- When scanning use optimal master file with minimal compression
- Store a backup of the master file in a depository organization (e.g., national library or archive) or store it in a distant location together with the original files
- Choose optimal carrier for storing your digital publication
- Do regular check outs of your digital files
- Do migration to new carriers every two years
- Secure optimal temperature and humidity in storage facilities
OUTLINE

1. The National and University Library
2. Digital preservation in brief
3. How libraries could contribute?
4. Recommendations
5. Open issues
OPEN ISSUES

- RESEARCH DATA MANAGEMENT
- INTEGRATION OF PRESERVATION WORKFLOWS IN DIFFERENT CULTURAL SECTORS
- BALANCE BETWEEN OPEN ACCESS AND PRESERVATION
- MORE COOPERATION
THANK YOU FOR YOUR ATTENTION!

QUESTIONS?

alenka.kavcic@nuk.uni-lj.si