POZNAŃ
SUPERCOMPUTING AND
NETWORKING
CENTER

„How can the basic digitisation workflow look like?”

Adam Dudczak
Poznań Supercomputing and Networking Center
maneo@man.poznan.pl

Education and Culture DG
Culture Programme

Agenda

- Overview of digitisation workflow
- Born digital documents
- Long term digital preservation
- Discussion

Digitisation strategy

*“A digitization project has many dimensions and **no two digitization projects are identical**. Each project varies according to the type of materials being digitized, the timescale, budget, staff skills and other factors. [...] Each project will need to develop a project plan to fit its particular circumstances.”*

MINERVA Technical Guidelines for
Digital Cultural Content Creation Programmes

Digitisation workflow

- Various goals – various workflows
  - Small library
  - Large institution
- General digitisation workflow
  - Base to handle both simple and advanced scenarios

Digitisation workflow (2)

- Choose objects for digitisation,
- Prepare object's metadata,
- Choose proper equipment to perform digitisation,
- Scan a given object,
- Adjust the scanned image using graphical software,
- Prepare a web delivery version of the digital object,
- Secure the digital master copy,
- Publish the object in a digital library.

Choose object for digitisation

Source: http://www.flickr.com/photos/ol1/4605912815/

Choose object for digitisation (2)

- Choice may depend on:
  - Collection development policy
  - Type and condition of the objects
  - Cost of digitisation
  - Availability of digital version
  - Intelectual Property Rights (IPR)
- Write down a list of items
  - Track progress
  - Estimate amount of work

Prepare object's metadata

Source: http://www.flickr.com/photos/lodz_na_nowo/2356170372/

Prepare object's metadata (2)

- Metadata can be applied to anything
- It helps to find, use and understand the object
- Metadata is machine-readable
- It has well defined structure
- There multiple types of metadata, which might be used for different things

Prepare object's metadata (3)

- There multiple types of metadata, which might be used for different things:
  - Descriptive
  - Administrative
  - Preservation
  - Structural
- In some cases metadata records may already exist for given document

Scan a given object

Source: http://www.flickr.com/photos/hendry/5207841020/

Scan a given object (2)

- Core part of the workflow
- We will refer to various equipment
  - Digital cameras, scanners
- Choose appropriate scanning device

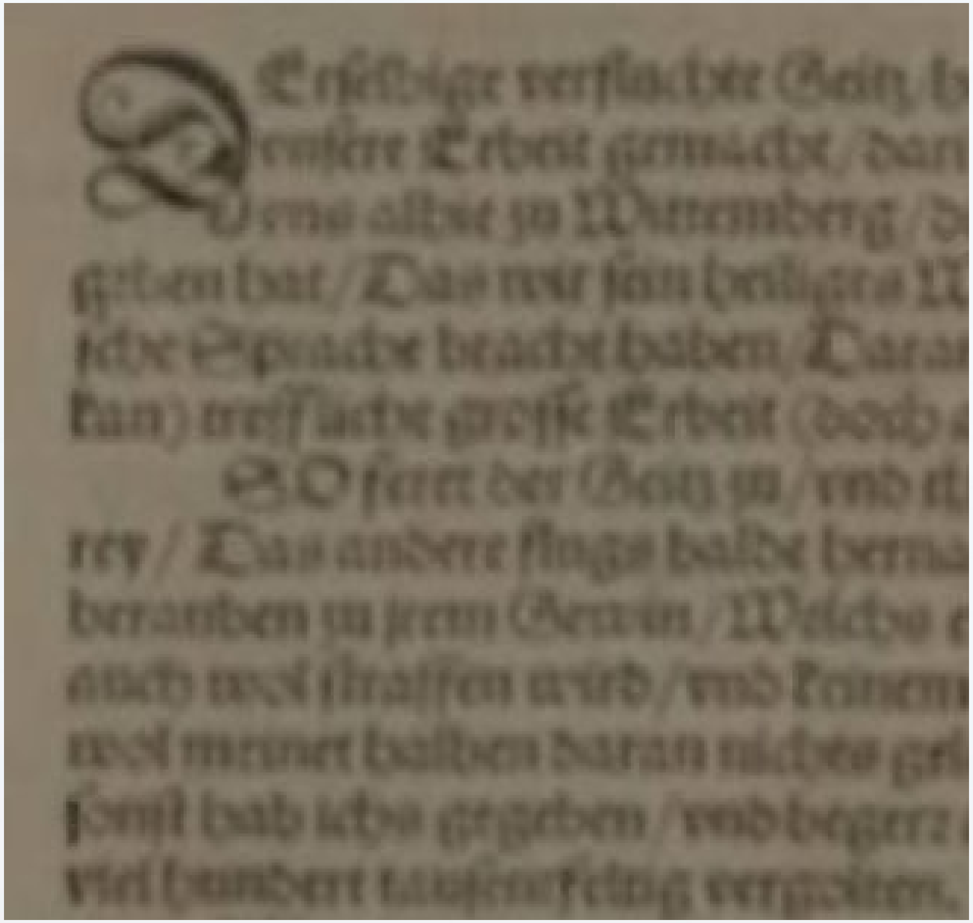
Scan a given object (3)

- Can it handle the:
  - range of sizes (e.g. A4, A0),
  - document types (single leaf, bound volume),
  - media (reflective, transparent)?
- Is it possible to scan even an object in a really bad condition?
- Can this scanning equipment produce images of quality which would sufficient for our goal?
- Is it fast and easy enough to operate?

Scan a given object (3)

- Choose appropriate scanning parameters
  - What may go wrong?

Scan a given object (4)



DEr ſelbige verfluchte Geitz h
vnſere Erbeit gemacht / dar
Denn alhie zu Wittemberg / D
geben hat / Das wir ſein heiliges W
ſche Sprache bracht haben / Dara
kan) treffliche groſſe Erbeit (doch

SO feret der Geitz zu / vnd th
rey / Das andere flugs balde herna
berauben zu jrem Gewin / Welchs e
auch wol ſtraffen wird / vnd keinem
wol meiner halben daran nichts ge
ſonſt hab ichs gegeben / vnd begere
viel hundert tauſentfeltig vergolten.

Scan a given object (5)

- Choose apropriate output file formats
  - It is a Digital master copy!

Digital master copy

- Concept of digital master copy
  - High resolution
  - Lossless compression (or none)
  - Suited for long term preservation
  - Well documented
  - Can be used to create other versions of the object i.e. web delivery version
- Helps to avoid situation when it is required to repeat scanning

Post processing

- Some errors made during digitisation can be adjusted without necessity to scan object once again
  - Adjust image rotation
  - Split scanned page
  - Convert to given format

Post processing

- Convert scanned text into digital text
  - OCR
  - Manual rekeying
- Using OCR
  - Use free trial versions to check if it will work with given type of resources.

OCR quality factors

- Type of document
- Quality of print
- Quality of original document
- Scans resolution, number of colours
- OCR software usually works better with high contrast images
- Language of text
- Text layout and formatting

OCR quality factors (2)

- Type of document
- Quality of print
- Quality of original document
- Scans resolution, number of colours
- OCR software usually works better with high contrast images
- Language of text
- Text layout and formatting

Secure the digital master copy

- Assure that Digital Master copy is safe and will be accessible in the future
  - Store metadata next to file
  - Store results of OCR next to file
  - Sometimes it is also required to archive software which is necessary to render content of the DMC

Prepare a web delivery version

- Digital master copy is usually inapropriate for web delivery
  - Too big, TIFF is not suported by web browsers
- Versions for different purposes
  - e.g. Thumbnail copy, view copy, interactive copy
- Copy optimized for web delivery

Prepare a web delivery version (2)

- Depending on type of object various software might be used to create web delivery version.

Born digital objects

- The term born-digital refers to materials that originate in a digital form.
- Do you know any examples of this kind of materials?

Publication of object in digital library

- Create new or use existing digital library
  - More about this in upcoming session

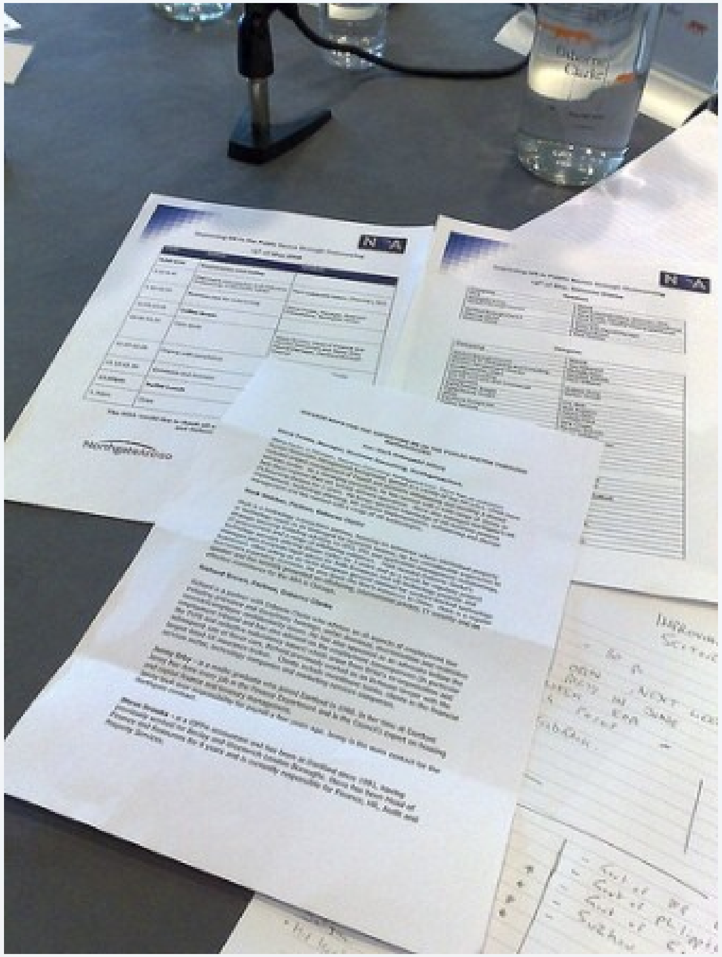
General conclusions

- Sometimes there is more than one right tool for the job,
- You need to know when to digitise using a given tool,
- Remember that you are dealing with old/valuable objects, so be careful,
- This might be the last chance to digitise a given object, so you have to assure quality during and after the scanning process,

General conclusions (2)

- Conversion from a digital image with text to a digital text can be a really tedious task but it is worth the effort,
- Digital libraries are created for users; keep that in mind while choosing the format for online publication.

Outsourcing



Source: http://www.flickr.com/photos/markhillary/2494078253/

Outsourcing (2)

- Definition:„Outsourcing is often viewed as involving the contracting out of a business function to an external provider”
- Outsourcing may genuinely reduce digitisation costs and provide strategic advantage

Outsourcing (3)

- When to outsource?
  - a large volume of work to be done in a short period of time
  - excessive cost of specialist equipment (such as bound volume or microfilm scanners)
  - lack of capability - unable to deliver the quality needed in-house due to lack of skills and experience;
  - The project has space, infrastructure or staffing constraints which preclude in-house digitisation

When should I avoid outsourcing?

- The collection is difficult to move or cannot be moved outside of the institution;
- The collection is badly organised, not inventoried or un-catalogued to the item level and needs skilled reorganisation as an integral part of the process;
- The digitisation needs to be phased in relatively small amounts over a long period;

When should I avoid outsourcing? (2)

- The preservation handling of the originals cannot be satisfactorily achieved in the outsourced environment;
- The digitisation tasks and goals are very complex and varied; and/or
- The volume of work is very small.

What can be outsourced?

- Check out: „When should I create my own digitisation lab?”
  - http://dl.psnc.pl/moodle/mod/lesson/view.php?id=89

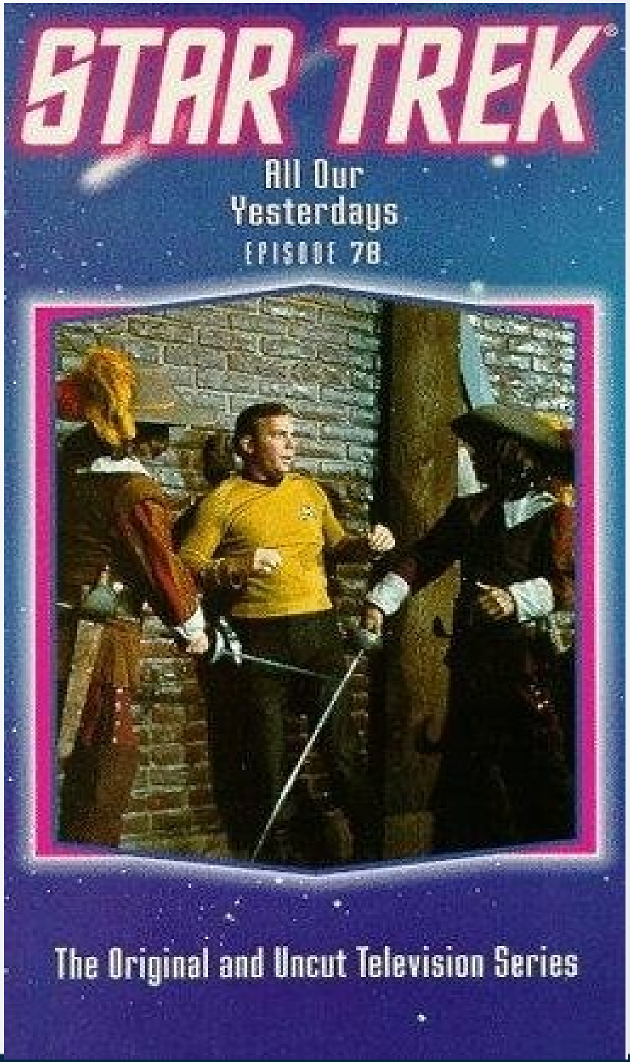
Long term preservation

„The problem of digital preservation is one of the most challenging research problems faced by the community of digital libraries today receiving growing interest by researchers and practitioners alike”

„Some Preliminary Ideas Towards a Theory of Digital Presevation“
Giorgos Flouris, Carlo Meghini
CASPAR project

A little bit of context

Star trek
- „All our yesterdays”



The Past is Prologue

- First digital preservation programs were developed in **1960s**
- Each generation of technology brings changes in potential capabilities to
  - creation
  - preservation of digital content
- Each new technology goes through a similar path
  - From idea to **implementation** than to **mainstream use** and in most cases to **obsolescence**

The Past is Prologue (2)

- Digital Preservation encompasses a broad range of activities designed to:
  - extend the usable life of computer files
  - protecting files from media failure
  - physical loss, and obsolescence.
- Information must be intact and readable whenever user needs it!

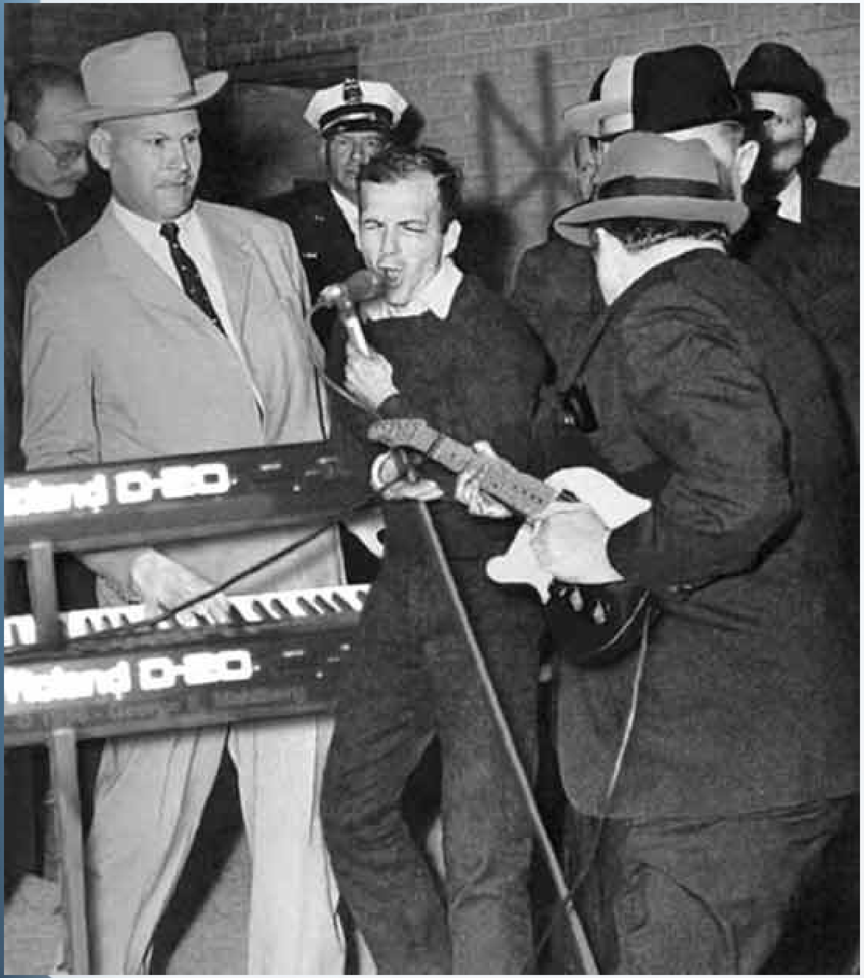
The Past is Prologue (3)

- Mentioned accessibility can be divided to:
  - Content renderability
    - Content can be viewed by humans or processed by computers
  - Understandability
  - Content can by interpreted by humans
- This implies main issues:
  - Bitstream preservation
  - Preservation of content, form, style and functionality

The Past is Prologue (4)

- What may happen with the information over time?
- Interpretation requires additional knowledge – a context
  - Take a look at this what is going with Flickr:The Commons
- Can we do something to avoid such a situation in the future?
- There is also an issue of authenticity of information

Information authenticity



Source:

“Digital History: A guide to gathering, preserving, and presenting the past on the web.”
by Daniel J. Cohen and Roy Rosenzeig

http://chnm.gmu.edu/digitalhistory/

Information authenticity (2)

© AP

What may go wrong?

- Format obsolescence,
  - Associated with the lack of software which is capable of displaying the content of a given file properly.
  - Data are imprisoned in the unsupported file format.
- Physical deterioration of the carrier,
  - If a copy of a digital object is stored on a DVD disc and this disc is scratched, it may have a terrible impact on accessing the content of the object.
- Carrier obsolescence
  - Similarly to file format obsolescence, also a carrier may become unreadable because of the lack of appropriate hardware. Do you remember a 3.5 inch floppy disks? Availability of hardware which is able to read its content is very limited.

What may go wrong? (2)

- Lack of information to properly interpret the intellectual content of the object,
  - Over the time the knowledge about interpreting information in a given object may disappear. Imagine a picture showing a building which does not exist anymore; over the time the knowledge about its existence may vanish and it would be impossible to identify the content of that particular picture.
- Institutional sustainability
  - If the institution or company which is driving a given repository went bankrupt, what would happen with its digital assets?

Means of digital preservation

- There is no universal solution which could be used for all data types and situations
- There are many different content preservation elements
- The most important includes:
  - Bitstream refreshing
  - Technology preservation (technology museum)
  - Analog backups
  - Migration
  - Reliance on Standards
  - Replication
  - Emulation

What are the features of good file format?

- Open specification is a minimal requirement
- Wide adoption
- Flexibility
  - e.g. TIFF
- History of backward compatibility
- Good metadata support
- Good range of functionality, but not overly complex
- Available interchange format with usable target
- Built-in error checking
- Reasonable upgrade cycle

What are the features of good file format? (2)

- When choosing proprietary format ensure that migration path exist
- PRONOM database
  - http://www.nationalarchives.gov.uk/PRONOM/Default.aspx
  - British National Archive database of file formats
- In case of some materials properietary format is the only solution i.e. video
  - Solution: store software next the object

Team digital preservation

- Series of cartoons showing what may go wrong and how to fix it
- Team Digital Preservation and the Aeroplane Disaster
  - http://www.youtube.com/watch?v=EKnsZZzuUr4

Team Digital Preservation and the Aeroplane Disaster

Discussion

- Organization of digitisation workflows in small memory institutions

Discussion

- Role of/experience with outsourcing of digitisation.

Discussion

- Issues related to Long term preservation

References

- Lessons from the DRMSI course:
  - „How can the basic digitisation workflow look like?”
  - „When should I create my own digitisation lab?”
  - „What is a Digital Master Copy and why it is so important?”