Access IT Training

How to digitize objects?

Digitisation

“Digitisation is the conversion of analogue materials into a digital format for use by software, and decisions made at the time of digitisation have a fundamental impact on the manageability, accessibility and viability of the resources created.”

MINERVA Technical Guidelines for
Digital Cultural Content Creation Programmes

Introduction

- Project Planning
- Preparing for the Digitization Process
  - The selection of materials for digitization
  - The physical preparation of materials for digitization
  - **The digitization process**
- **Storage and Management of Digital Master Material**
- Metadata, standards and resource discovery
- **Delivery formats**
- Publishing on the Web
- Re-use and re-purposing
- Intellectual Property and Copyright

Agenda

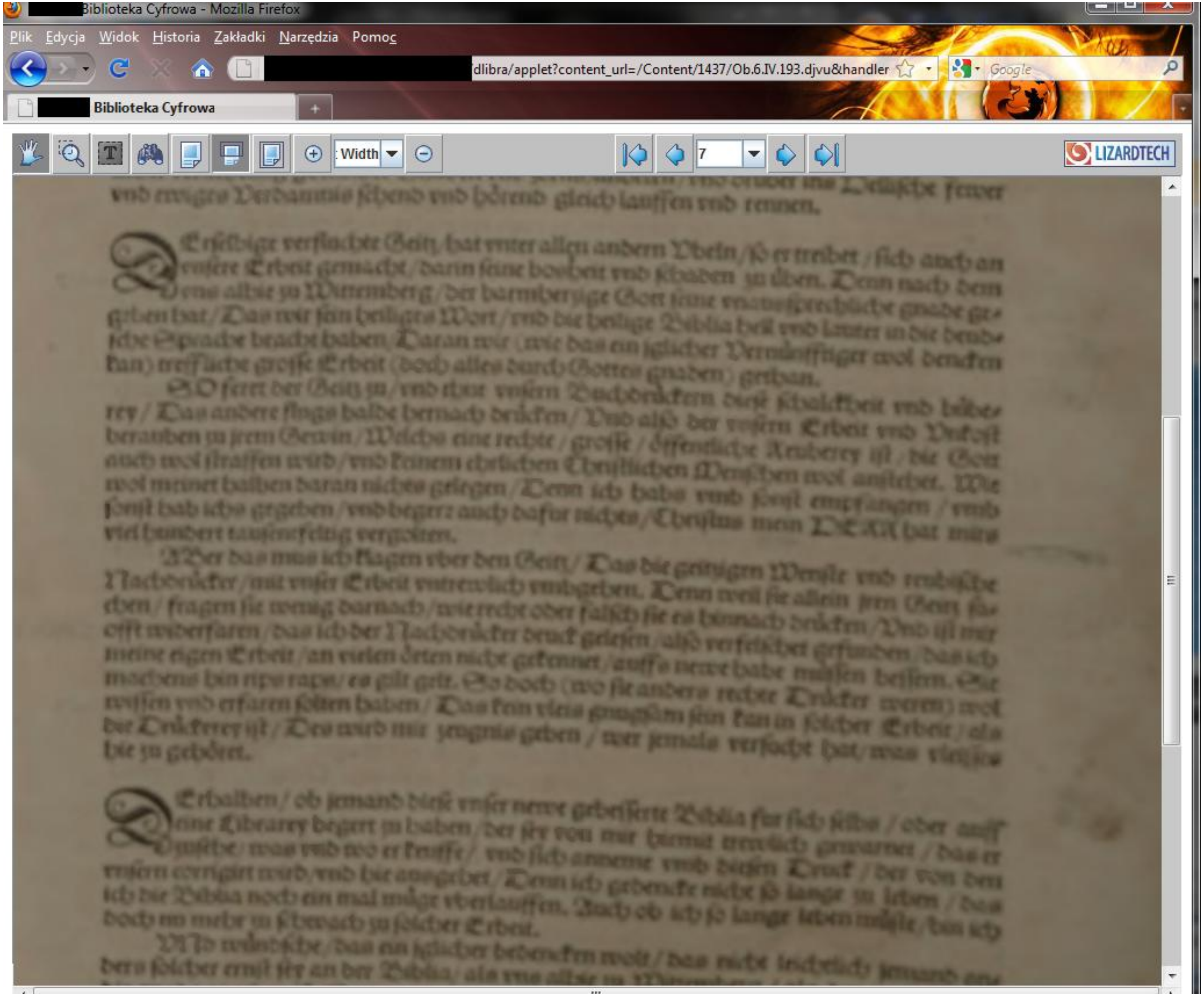
- How bad it can be?
- Digitization strategy
  - How to develop internal digitization strategy?
- Setting technical requirements
- How to handle scanned material?
- Dealing with Digital Master copies
- Scanning equipment

How bad it can be?

- Files are too big
  - Long download time
  - File is too big for browser to handle
    - e.g. huge PDF files
- Using inappropriate file formats for online delivery
  - proprietary/closed/not well known file formats may cause problems for users, webcrawlers, screen readers

How bad it can be?

- Files are too small
  - Unreadable content – to low resolution


[redacted] Biblioteka Cyfrowa - Mozilla Firefox
Plik
Edycja
Widok
Historia
Zakładki
Narzędzia
Pomoc
[redacted] dlibra/applet?content_url=/Content/1437/Ob.6.IV.193.djvu&handler
Google
[redacted] Biblioteka Cyfrowa
Width
7
LIZARDTECH

vnd ewiges Verdamnis ſehend vnd hörend gleich lauffen vnd rennen.

DEr ſelbige verfluchte Geitz / hat vnter allen andern Vbeln / ſo er treibet / ſich auch an vnſere Erbeit gemacht / darin ſeine bosheit vnd ſchaden zu üben. Denn nach dem vns alhie zu Wittemberg / der barmhertzige Gott ſeine vnausſprechliche gnade gegeben hat / Das wir ſein heiliges Wort / vnd die heilige Biblia hell vnd lauter in die deudſche Sprache bracht haben / Daran wir (wie das ein jglicher Vernünfftiger wol dencken kan) treffliche groſſe Erbeit (doch alles durch Gottes gnaden) gethan.

SO fert der Geitz zu / vnd thut vnſern Buchdrückern dieſe ſchalckheit vnd büberey / Das andere flugs balde hernach drücken / Vnd alſo der vnſern Erbeit vnd Vnkoſt berauben zu jrem Gewin / Welchs eine rechte / groſſe / öffentliche Reuberey iſt / die Gott auch wol ſtraffen wird / vnd keinem ehrlichen Chriſtlichen Menſchen wol anſtehet. Wie wol meinet halben daran nichts gelegen / Denn ich habe vmb ſonſt empfangen / vmb ſonſt hab ichs gegeben / vnd begere auch dafur nichts / Chriſtus mein HERR hat mirs viel hundert tauſentfeltig vergolten.

ABer das mus ich klagen vber den Geitz / Das die geitzigen Wenſte vnd reubiſche Nachdrücker / mit vnſer Erbeit vntrewlich vmbgehen. Denn weil ſie allein jren Geitz ſuchen / fragen ſie wenig darnach / wie recht oder falſch ſie es hinnach drücken / Vnd iſt mir offt widerfaren / das ich der Nachdrücker druck geleſen / alſo verfelſchet gefunden / das ich meine eigen Erbeit / an vielen örten nicht gekennet / auffs newe habe müſſen beſſern. Sie machens hin rips raps / es gilt gelt. So doch (wo ſie andere rechte Drücker weren) wol wiſſen vnd erfaren ſolten haben / Das kein vleis genugſam ſein kan in ſolcher Erbeit / als der Drückerey iſt / Des wird mir zeugnis geben / wer jemals verſucht hat / was vleis hie zu gehöret.

DErhalben / ob jemand dieſe vnſer newe gebeſſerte Biblia fur ſich ſelbs / oder auff ſeine Librarey begert zu haben / der ſey von mir hiermit trewlich gewarnet / das er auffſehe / was vnd wo er keuffe / vnd ſich anneme vmb dieſen Druck / der von den vnſern corrigirt wird / vnd hie ausgehet / Denn ich gedencke nicht ſo lange zu leben / das ich die Biblia noch ein mal müge vberlauffen. Auch ob ich ſo lange leben müſte / bin ich doch nu mehr zu ſchwach zu ſolcher Erbeit.

VNd wündſche / das ein jglicher bedencken wolt / das nicht leichtlich jemand eine
ders ſolcher ernſt ſey an der Biblia / als vns alhie zu Wittemberg

How bad it can be?

- Digital Master material is removed after creation of web delivery formats
- Lack of text recognition
  - Even the best metadata is not as useful as properly recognized and indexed text

Digitization 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

Digitization strategy

- Project digitization strategy should reflect specific (long and short term) goals and objectives
- Such a document can be created for one institution/project/country
- What can be inside digitization strategy?

Digitization strategy

- What will be digitized?
  - "In 2009-10 the Library will digitize approximately 14,550 collection items."
  - "Target in 2009-2010: (each resource type has detailed list of objects selected for digitisation)
    - Pictures collection
      - 9 945 items
    - Maps Collection
      - 1 175 items"
    - ….
  - http://www.nla.gov.au/digital/program.html

Digitization strategy

- How those goals will be reached?
  - Source: http://www.nla.gov.au/digital/standards.html
- Image capture standards
  - Define image capture standards used in collection material digitisation
- Images for Web delivery - standards
  - Standards used in production of derivatives for Web delivery

Digitization strategy

- Digital capture equipment
  - Capture devices used and guidelines for selection of a device suitable for capture of particular types of collection materials
- Guidelines for presentation of digitised pictures
  - Guidelines for presentation of digitised pictures on the Web, in particular on image cropping and borders

Digitization strategy

- Metadata for created images
  - Guidelines for extracting or creating metadata describing images created through the digitisation process
  - What kind of metadata should be associated with object at the digitization stage
- Persistent identifier scheme for digital collections
- Care and handling guidelines
  - Guidelines on the care and handling of material to be digitised

How to develop internal strategy?

- Every digitization project is different
- Need to find a balance between speed and quality
- Consider deadlines, number of people involved, available funds
- Digitization strategy should be a part of project planning
  - Directly associated with cost of hardware/software/employment

How to develop internal strategy?

- External (high) requirements and type of digitized material can also lead to outsourcing
- Before you start, check out
  - national digitization strategy (if exist)
  - national standards for electronic documents

How to develop internal strategy?

- What will be the result of digitization?
  - Digital Master Copy of object
    - Object which will be stored for a long, long time
  - Object delivered through the digital library over the web
    - Presentational form of object
    - Some digitization companies can deliver content as ready to deploy digital library
  - Both forms are equally important
  - Depending on the goal of digitization quality requirements may differ

Digitisation objectives

- Library of Congress distinguishes different requirements for different digitization objectives:
  - **Goals for text documents:**
    - Allow users to see the content of document
    - Recognize the document text to allow full- text search
  - **Goals for images and photography**
    - Allow users to see the content of document
    - Allow to prepare reproduction of image

Digitisation objectives

- National Library of New Zealand uses a more general approach
- NLNZ defines two levels of requirements : **minimal** and **recommended** without further explanations or referring to specific goal of digitization
- Those goals/levels are followed by specific technical requirements like resolution, format etc.

How to develop internal strategy?

- What kind of IT infrastructure is required?
  - Including storage/network, scanning equipment, software tools
- How big storage media do we need?

How big are Digital Master copies?

- Image type and the dpi ranges alter the file size
- The higher the dpi, the larger the file size
- As an example uncompressed 1" x 1" image in different colour depths and resolutions

Resolution (dpi)	400	300	200	100
black and white	20K	11K	5K	1K
8-bit grayscale or colour	158K	89K	39K	9K
24-bit colour	475K	267K	118K	29K

Source: "Creating and Documenting Electronic Texts", AHDS

How big are Digital Master copies?

- **A4** page, scanned in **300 DPI** and saved as uncompressed **TIFF** weighs more than **26MB**
  - Lossless compression can significantly minimise file size
- In case of microfilms one DVD (**4.7 GB**) can fit **7 000** microfilm frames – **0.6MB** per frame

Technical requirements example 1

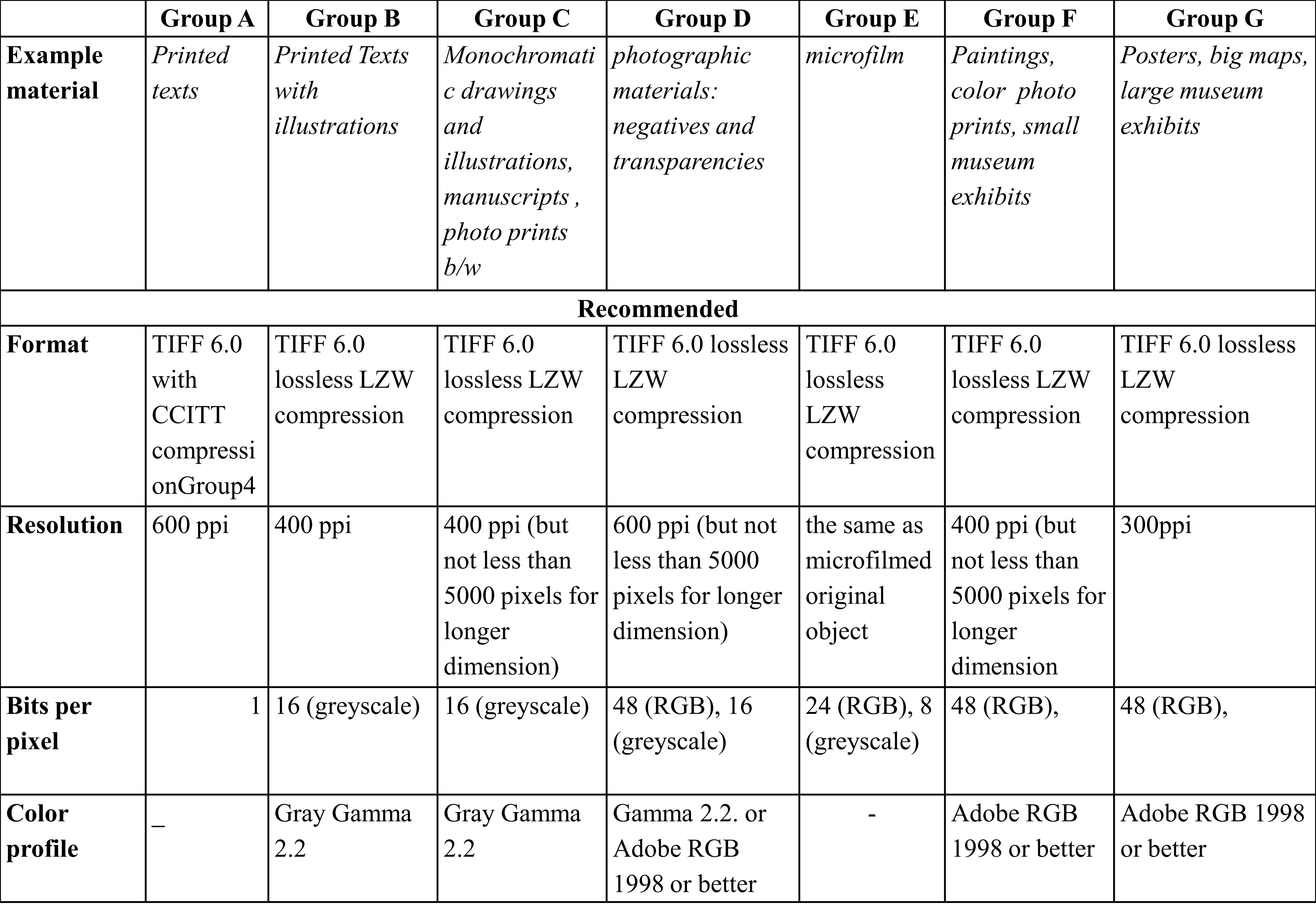
- Requirements proposal coming from committee working on behalf of Polish government
  - http://fbc.pionier.net.pl/id/**oai:bcpw.bg.pw.edu.pl:1262**
- Team classified different type of resources into different groups (from A to G)
  - Each group has two levels of requirements minimal and recommended

Technical requirements example 1

- Requirements consist of
  - an examples of materials from given group
  - Digital Master file format
  - resolution (in ppi or dpi)
  - colour depth
  - used colour space
- At the moment there are no recommendations for web delivery standards

	Group A	Group B	Group C	Group D	Group E	Group F	Group G
Example material	*Printed texts*	*Printed Texts with illustrations*	*Monochromatic drawings and illustrations, manuscripts , photo prints b/w*	*photographic materials: negatives and transparencies*	*microfilm*	*Paintings, color photo prints, small museum exhibits*	*Posters, big maps, large museum exhibits*
Format	TIFF 6.0 with CCITT compressionGroup4	TIFF 6.0 lossless LZW compression	TIFF 6.0 lossless LZW compression	TIFF 6.0 lossless LZW compression	TIFF 6.0 lossless LZW compression	TIFF 6.0 lossless LZW compression	TIFF 6.0 lossless LZW compression
Resolution	400 ppi	300 ppi	300 ppi (but not less than 3000 pixels for longer dimension)	300 ppi (but not less than 3000 pixels for longer dimension)	the same as microfilmed original object	300 ppi (but not less than 3000 pixels for longer dimension)	300ppi
Bits per pixel	1	8 (greyscale)	8 (greyscale)	24 (RGB), 8 (greyscale)	24 (RGB), 8 (greyscale)	24 (RGB)	24 (RGB)
Color profile	-	Gray Gamma 2.2	Gray Gamma 2.2	Adobe RGB 1998	-	Adobe RGB 1998	Adobe RGB 1998

Source: http://fbc.pionier.net.pl/id/**oai:bcpw.bg.pw.edu.pl:1262**



	Group A	Group B	Group C	Group D	Group E	Group F	Group G
Example material	*Printed texts*	*Printed Texts with illustrations*	*Monochromatic drawings and illustrations, manuscripts , photo prints b/w*	*photographic materials: negatives and transparencies*	*microfilm*	*Paintings, color photo prints, small museum exhibits*	*Posters, big maps, large museum exhibits*
Recommended							
Format	TIFF 6.0 with CCITT compressionGroup4	TIFF 6.0 lossless LZW compression	TIFF 6.0 lossless LZW compression	TIFF 6.0 lossless LZW compression	TIFF 6.0 lossless LZW compression	TIFF 6.0 lossless LZW compression	TIFF 6.0 lossless LZW compression
Resolution	600 ppi	400 ppi	400 ppi (but not less than 5000 pixels for longer dimension)	600 ppi (but not less than 5000 pixels for longer dimension)	the same as microfilmed original object	400 ppi (but not less than 5000 pixels for longer dimension	300ppi
Bits per pixel	1	16 (greyscale)	16 (greyscale)	48 (RGB), 16 (greyscale)	24 (RGB), 8 (greyscale)	48 (RGB),	48 (RGB),
Color profile	_	Gray Gamma 2.2	Gray Gamma 2.2	Gamma 2.2. or Adobe RGB 1998 or better	-	Adobe RGB 1998 or better	Adobe RGB 1998 or better

Source: http://fbc.pionier.net.pl/id/**oai:bcpw.bg.pw.edu.pl:1262**

Technical requirements example 1

- Those requirements should be adjusted for very large/very small objects
  - This is particularly an issue for museum exhibits like coins or… pyramids.

Technical requirements example 2

- National Library of Australia technical requirements for **still images**
  - http://www.nla.gov.au/digital/capture.html

Technical requirements example 2

Material type	Tonal resolution (pixel depth)	Spatial resolution*
Colour reflective formats, including: • coloured maps • pencil sketches with wash • sepia or coloured photographic prints • printed music • manuscripts • objects	RGB **24 bits** per pixel	Larger than A5: 300 ppi Smaller than A5 but larger than A6: 600 ppi Smaller than A6: 1200 ppi
Colour transparencies, including 35mm	RGB **24 bits** per pixel	2000 ppi

Technical requirements example 2

Material type	Tonal resolution (pixel depth)	Spatial resolution*
Colour negatives, including 35mm	RGB 48 bits per pixel	2000 ppi Note: two digital masters are created: colour negative and colour positive; derivatives are created from the colour positive.
B&W reflective formats, including: • photographic prints • black and white line art • black and white map	RGB 24 bits per pixel	Larger than A5: 300 ppi Smaller than A5 but larger than A6: 600 ppi Smaller than A6: 1200 ppi

Technical requirements example 2

Material type	Tonal resolution (pixel depth)	Spatial resolution*
B&W negatives 35mm	Greyscale 16 bits per pixel	3000 ppi (TIFF master and derivatives positive)
B&W negatives larger than 35mm	Greyscale 16 bits per pixel	2000 ppi (TIFF master and derivatives positive)
B&W microfilm masters of newspapers	Greyscale and bi-tonal (image optimised for OCR)	400 ppi

Technical requirements example 2

Material type	Tonal resolution (pixel depth)	Spatial resolution*
Print publications	RGB 24 bits per pixel	300 ppi at 100%
Print publications scanned for Copies Direct orders	Bi-tonal (or RGB where the copy would be illegible if scanned as bi-tonal image)	300 ppi at 100% Multi-page TIFF format

How to develop internal strategy?

- What about Web delivery standards?
  - Source: "*MINERVA Technical Guidelines for Digital Cultural Content Creation Programmes*"
- Consideration must be given to the fact that variations exist in:
  - the types of hardware device and client software employed by users
  - the levels of bandwidth restriction within which users operate

Web delivery standards

- To maximize potential audience project should:
  - Make resources available in alternative sizes and/or formats/resolutions/bitrates
  - Project should periodically review the criteria on which decision about delivery format and parameters are based

Web delivery standards

- Web delivery formats should be derived from Digital Master copies
- It is important to automate process of image conversion using software like Image Alchemy or ImageMagick
- Digitized resources should be unambiguously identified and uniquely addressable directly from a user’s Web browser.

Web delivery standards

- End user needs to have the capability to directly and reliably cite an individual resource
- This will allow user to perform basic reuse of digitized objects
- Objects URIs should be reasonably persistent
- Object URIs shouldn’t embed information about
  - file format
  - server technology
  - organization structure of the provider service
  - any other information that is likely to change within the lifetime of the resource

Web delivery standards

- Three examples
  - Example 1 : National Library of Australia - “Images for Web delivery – standards”
    - http://www.nla.gov.au/digital/delivery.html
  - Example 2: MINERVA Technical Guidelines for Digital Cultural Content Creation Programmes
  - Example 3:
    - Why Polish librarians decided to use DjVu?

Derivative and file type	Set of standards used
Thumbnail copy (JPEG)	• Derived for all material formats from TIFF master • Resolution: 72 ppi • Longest dimension: 150 pixels (e.g. 118x150 portrait; 150x118 landscape)
View copy (JPEG)	• Derived from TIFF master using Image Alchemy software • Resolution: 72 ppi • Longest dimension: 600 pixels for pictures and 760 pixels for manuscripts, maps and music (e.g. 590x760 portrait; 760x760 square; 760x590 landscape)

Derivative and file type	Set of standards used
View copy (multi-page PDF)	• Created for print publications scanned for Copies Direct orders • Derived from multi-page TIFF master • Compressed • Resolution: 72 dpi • Longest dimension: 1000 pixels
Examination copy (JPEG)	• Derived for printed music and cartographic materials from TIFF master using Image Alchemy software • Resolution: 72 ppi • Longest dimension: 1000 pixels (e.g. 781x1000 portrait; 1000x1000 square; 1000x781 landscape)

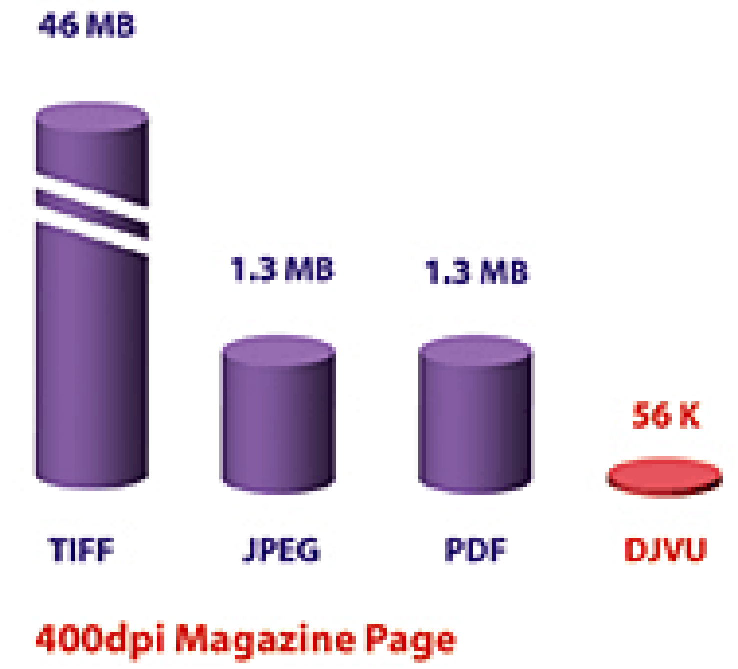
Derivative and file type	Set of standards used
Print copy (PDF)	• Derived for printed music from JPEG examination copies using Image Alchemy software • Compressed • Resolution: 72 dpi • Longest dimension: 1000 pixels
Interactive copy (MrSID)	• Created primarily for cartographic material from TIFF master using MrSID software • Compressed • Resolution: 300 ppi • Longest dimension: as per the TIFF master (varies according to the original physical item)

Example 2

- Delivery of text
  - Selection of character and structure encoding
    - UTF-8
    - XHTML, PDF, DjVU file formats
- Delivery of still images
  - Thumbnail – 72 dpi, 24bpp
    - (8bpp for grayscale) at most 100-200 pixels in longest dimension (Source EMII-DCF)
  - Images for fullscreen presentation – 150 dpi, 24bpp
    - (8bpp for grayscale) at most 600 pixels for longest dimension (Source EMII-DCF)

A few remarks about DjVu format

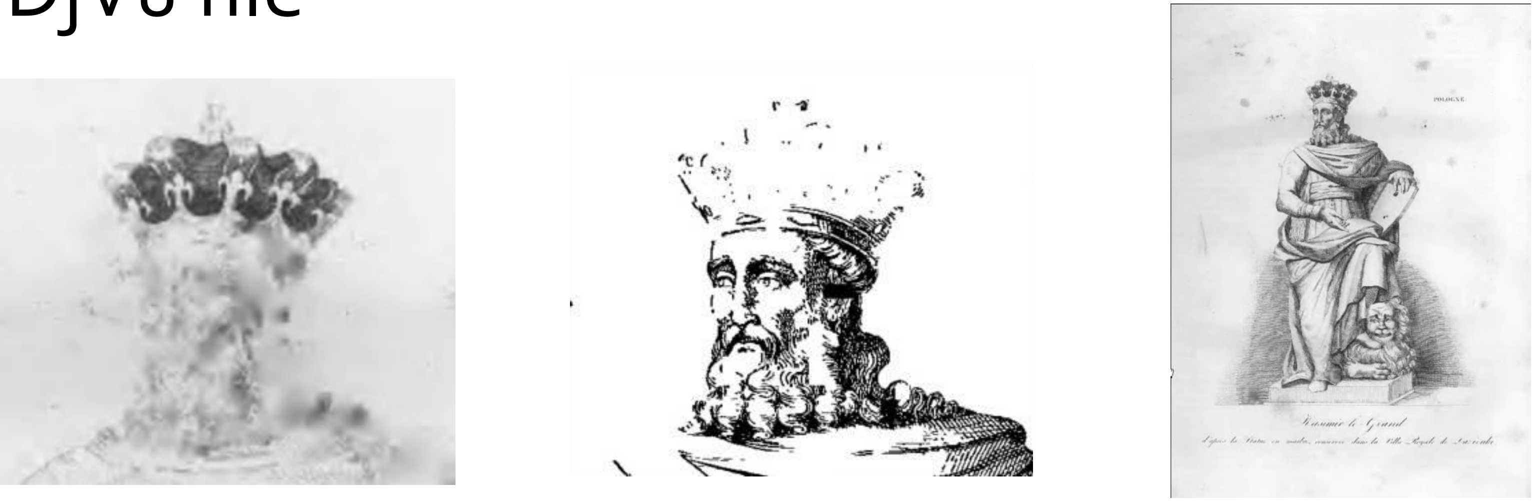
- Most of resources in Polish digital libraries are delivered using **DjVu** file format
- DjVu uses its own compression method which is quite effective and fast
  - typical page scanned in **300 dpi** resolution has from **5 to 30 KB**



Source: http://www.malin.net.pl/wiecej-o-djvu-i-systemie-dlibra in Polish only

A few remarks about DjVu format

- DjVu divides image into layer, there is background layer, foreground layer and text layer
- Metada about object can be embedded inside DjVu file



Source: Silesian Digital Library, http://sbc.org.pl/

A few remarks about DjVu format

- Huge advantaged over PDF is the fact that digital object might be stored in one or in multiple files (e.g. file per book page)
- Although format looks really nice it's not widely recognized and this is a huge barrier for new users
- The biggest drawback is the fact that Google does not support DjVu text content indexing!

Internal digitization guidelines

- How to handle with objects?
  - Example 1 : National Library of Australia - “Care and handling guidelines for digitisation of Library materials”
    - http://www.nla.gov.au/digital/care_handling.html

Care and handling guidelines – general rules

- Wash hands regularly to ensure they are clean at all times
- Cotton gloves may be used, when appropriate, to handle most items
  - This will protect paper and other materials from grease, oils and dirt on bare hands
  - Gloves can make fine work like turning pages more difficult - use discretion
- Always have plenty of room in your workspace to accommodate the material you are working with
- Transport items as much as possible in their folder and on an appropriate trolley

Care and handling guidelines – general rules

- Be careful when removing fragile items from storage enclosures
- Don't lick your fingers prior to handling any collection item
- Make sure items are fully supported at all times.
- Works, particularly damaged ones, should be enclosed in Mylar (polyester) pockets, polyethylene or polypropylene bags or sleeves, archival mounts or folders when in storage or when being transported

Care and handling guidelines – general rules

- Avoid direct handling or touching of surface areas
- Always use pencil when working near collection items
- Never use collection items as a writing surface
- Do not stack different items together e.g. books and artworks
- No food and drinks near collection items; wash hands after eating

Care and handling guidelines – general rules

- Remove paper clips, pins and string carefully
- Replace metal pins and clips with plastic paper clips
- If you need to mark a page use a piece of clean white paper - **do not use a post-it notes** or other adhesive papers or plastics
- Avoid the temptation to repair items;
    - **do not use adhesive tape to repair**, this will eventually discolour and damage the paper

Care and handling guidelines – general rules

- Consult a paper conservator to perform repairs or for further advice

- There is also a number of specific rules for different types of objects
  - Maps, plans, photographic material, artworks, manuscripts and other

Care and handling guidelines - maps

- Be aware that maps and plans are difficult to handle because they are large
  - Make sure there is enough space to handle and work with them
- When not in use store maps in a map cabinet.
- Do not make new folds in maps or plans, as it will damage them

Care and handling guidelines – photographic material

- Ensure glass platen of the flat-bed scanner is clean
  - Wipe with 'screen cleaner' instead of commercial detergents
- Make sure the transparencies are securely in position
- Store photographic items in archival plastic or paper sleeves

Care and handling guidelines – photographic material

- Photographic emulsions are easily scratched and need to be protected when handling more than one photograph at a time
  - You can protect them by separating them or interleaving them - ideally with archival materials
  - **Avoid placing on top of each other**
- Don't mend photographs using self-adhesive sticky tape of any kind
  - These tapes deteriorate and will stain and damage the photograph

Care and handling guidelines – photographic material

- Handle negatives and transparencies by their edges or use gloves
- **Labels and identification stamps** should **not be applied directly** to photographic material
- ID material should be placed on the packaging
    - Stabilo (or similar) or B grade pencils can be used to write on the verso of paper-based prints

Care and handling guidelines – photographic material

- Do not use any water-based solvents such as window cleaner or film cleaner on photographic material
  - Improper cleaning of photographic materials can cause serious damage such as permanent staining, abrasion and loss of binder or image
  - Use a soft brush or photographers blower brush to clean dusty negatives or photographs
  - Consult a photographic conservator to perform repairs or for further advice on cleaning

Care and handling guidelines – artworks

- Artworks with friable media (pastels, charcoal and pencil drawings) are easily smudged
  - They should always be in a window mount with a cover over the window and stored in a poly bag
  - **Never place these items in a polyester** (Mylar) sleeve because static attraction may lift any looses particles from the surface

Care and handling guidelines – artworks

- Paper from the late 1800s will often be of poor quality and brittle
  - Please consult a conservator for further advice
- **Do not stack unenclosed items**
- Items without friable media and of a suitable size can be scanned on a flat bed scanner

Care and handling guidelines – artworks

- Use an overhead camera for pastels, graphite and charcoal drawings or oil paintings
- Gently place items face down onto the platen and avoid moving items around while they are face down

Care and handling guidelines – manuscripts

- A manuscript collection can include :
  - sheet music, note books, diaries, correspondence, reports, drafts, maps, plans, charts, photographs, x-rays, pamphlets, forms and faxes
- If you come across material with a friable surface such as pastel, watercolour, and graphite or charcoal - **use an overhead camera**

Care and handling guidelines – manuscripts

- Bound material should also be digitised under an overhead camera using a supporting cradle
- Please refer to a conservator if you are unsure

Care and handling guidelines – manuscripts

- Remove any staples or pins before scanning
- Wrinkled or folded items should be smoothed out by gentle pressure with your fingers and palms
  - If items are severely crumpled and damaged a conservator should treat the item
- Make sure items are fully supported at all times

Care and handling guidelines – manuscripts

- Bound sheet music items will be too large for the platen of a flatbed scanner
- Place a support stand around the sides of the scanner to hold open sheets to prevent sagging and dangling while scanning

How to develop internal strategy?

- How to cope with Digital Master copies?
  - Source: "*MINERVA Technical Guidelines for Digital Cultural Content Creation Programmes*"

- Projects must consider the value in creating a fully documented high-quality 'digital master'
  - All other versions (e.g. compressed versions for access via the Web) can be derived from this one

How to develop internal strategy?

- Network capacity and user needs may change over time – periodical migration of web data might be necessary
  - Consider automatic conversion!
- Full documentation includes additional metadata which may be useful in the future

How to develop internal strategy?

- How/Where to store Digital Master copies?
  - Source: "*MINERVA Technical Guidelines for Digital Cultural Content Creation Programmes*"
  - Digital storage media have different software and hardware requirements for access
  - Different media present different storage and management challenges

How to develop internal strategy?

- The resources generated during digitisation project will typically be stored:
  - on the hard disks of one or more file servers
  - and also on portable storage media
- Consider creation of file server for project

Selection of storage media

- Minerva technical guidelines states that the most commonly used types of portable medium are :
  - Magnetic tapes
  - Optical media (CD-R/DVD)
- Portable media chosen should be of good quality and purchased from reputable brands and suppliers
  - new instances should always be checked for faults

Selection of storage media

- The threats to continued access to digital media are two-fold:
  - The physical deterioration of or damage to, the medium itself
  - Technological change resulting in the obsolescence of the hardware and software infrastructure required to access the medium

Selection of storage media

- Many factors can influence the selection of media for long-term digital storage
- Weighing those factors against each other for the great variety of media available today can be a complex task
- In the following example, each medium is scored against the criteria
  - on a scale of 1 (does not meet the criterion) to 3 (fully meets the criterion)
  - and a minimum total score of 12 is recommended for consideration

Selection of storage media

- Longevity
  - Media storage option chosen should have a proven life span of at least 10 years
- Capacity
  - Minimizing the number of actual media to be managed will generally be more efficient and cost effective

Selection of storage media

- Viability
  - Media and drives should support robust error-detection
  - Proven data recovery techniques should also be available in case of data loss
  - Media should be write-once or have write protection mechanism
    - To avoid accidental erasure

Selection of storage media

- Obsolescence
  - Media and supporting software should be based on mature technology
  - Technology should be widely available
  - Open standards for both media and software are generally preferable
- Cost
  - Total cost = media cost + cost of ownership
  - Media cost should be expressed as a price per GB
  - Cost of ownership include cost of software and equipment

Selection of storage media

- Susceptibility
  - The media should have low susceptibility to physical damage
  - and be tolerant of a wide range of environmental conditions without data loss

Selection of storage media

Media	CD-R	DVD-R	Zip Disk	3.5” Magnetic Disk	DLT	DAT
Longevity	3	3	1	1	2	1
Capacity	2	2	1	1	3	3
Viability	2	2	1	1	3	3
Obsolescence	3	2	2	3	2	2
Cost	3	2	1	1	3	3
Susceptibility	3	3	1	1	3	2
Total	16	14	7	8	16	14

Source: http://www.nationalarchives.gov.uk/documents/selecting-storage-media.pdf

Portable storage media handling

- Media should be handled, used and stored in accordance with their suppliers’ instructions
  - http://www.icpsr.umich.edu/dpm/dpm-eng/oldmedia/threats.html
- Media should be refreshed on a regular cycle within the lifetime of the medium
- Refreshment activity should be logged

Portable storage media handling

- Take a look at:
  - „Care and Handling of CDs and DVDs: A Guide for Librarians and Archivists”
    - http://www.clir.org/pubs/reports/pub121/contents.html

Choosing proper hardware for digitization

- When resolutions, formats, color parameters are set for each type of resources you can choose right digitization equipment
- NLA lists all the equipment used in their digitization lab:
  - http://www.nla.gov.au/digital/capturedevice.html

Type of device	Devices used
Digital cameras	- x1 Canon 1D mkIII - small format camera - x2 Canon 1Ds mkIII - small format camera - x1 PhaseOne P65+ - medium format capture back mounted on PhaseOne camera - x1 Betterlight Super 10K-HS scanback - large format scanback mounted on a Sinar P2 large format camera - x2 Sinar eVolution 75H - medium format capture back mounted on a Sinar P3 camera
Film & slide scanners	- x1 Hasselblad Flextight X1 film scanner - x1 Nikon SuperCool Scan 4000ED 35mm/IX240 film scanner
Flat bed scanners	- x2 Creo iQsmart1 - colour, up to A3 (transmissive + reflective) - x1 Creo iQsmart2 - colour, up to A3 (transmissive + reflective) - x2 Creo iQsmart3 - colour, up to A3 (transmissive + reflective) - Ricoh IS330DC – used to scan print publications for Copies Direct orders

Type of device	Devices used
Overhead scanners	- SMA 21 –used to scan print publications for Copies Direct orders
Microfilm/microfiche scanners	- Used by external agencies for digital capture of the newspapers

Choosing proper hardware for digitization

- The nature of the material to be digitised, particularly its **condition** and **construction**, determines the type of equipment used for digital capture
- All collections are surveyed prior to digitisation and minimal handling during the digital capture process is preferred

Choosing proper hardware for digitization

- **Flat bed scanners** are used for items **less than A3 in size (420x297 mm)**, such as:
  - **printed music**: unbound sheet music
  - **printed material**, including
    - single sheets (e.g. handbills, newspapers, posters),
    - ephemera (e.g. brochures, tickets),
    - single section pamphlets or other bound items that can be opened with ease

Choosing proper hardware for digitization

- **Flat bed scanners** are used for items **less than A3 in size (420x297 mm)**, such as:
  - **manuscripts**: single sheets in pencil or ink (e.g. letters), small printed items
  - **maps** in good condition
  - artworks on paper
  - hand and mechanically produced prints in black ink with no added colour or extreme relief
    - e.g. engravings, etchings, lithographs

Choosing proper hardware for digitization

- **Flat bed scanners** are used for items **less than A3 in size (420x297 mm)**, such as:
  - pen and ink drawings without watercolour (e.g. cartoons)
  - photographic material: film negatives, transparencies, microforms, glass negatives and lantern slides, B&W and colour gelatin silver prints, albumen prints

Choosing proper hardware for digitization

- Digital cameras are used for:
  - fragile items
  - oil paintings
  - most original artworks on paper
    - e.g. watercolours, drawings
  - artwork with loose, friable media
    - e.g. pastels, charcoal, crayons, soft pencil
  - watercolours with thick paint, gouache or glazes
  - loose pages, sheets from albums of printed items

Choosing proper hardware for digitization

- Digital cameras are used for:
  - items larger than A3 or oversize
    - e.g. posters
  - bound volumes:
    - books, albums, printed music, atlases
  - maps:
    - large or fragile maps, river maps, maps with colour
  - manuscripts:
    - bound diaries, letter books, folded items, large items

Choosing proper hardware for digitization

- Digital cameras are used for:
  - documents on parchment and vellum
  - photographic items:
    - oversize B&W and colour prints, historic process photographs (e.g. daguerreotypes, ambrotypes), platinum prints and other non-silver prints, glass negatives
  - three dimensional material:
    - textiles, sculpture, objects, wax seals

Choosing proper hardware for digitization

- Film/slide scanners are used for:
  - strip films, negatives and transparencies
  - mounted slides
- More information about scanning equipment can be found in :
  - **6B. Technical Infrastructure: Image creation** chapter of "Moving Theory into Practice - Digital Imaging Tutorial"
  - http://www.library.cornell.edu/preservation/tutorial/technical/technicalB-03.html

After a short break…

- Digitization workflows
- Cutting cost of digitization
- Remarks about resolution for scanning text document
- OCR
- Formats for non-still images digitization
  - Preservation and web delivery
- Useful tools

Q&A

Access IT Training

How to digitize objects? (2)

Agenda

- Digitization workflows
- Cutting cost of digitization
- Remarks about resolution for scanning text document
- OCR
- Formats for non-still images digitization
  - Preservation and web delivery
- Useful tools

Agenda

- **Digitization workflows**
- Cutting cost of digitization
- Remarks about resolution for scanning text document
- OCR
- Formats for non-still images digitization
  - Preservation and web delivery
- Useful tools

Workflows for still images

- All-in-one approach or technological line?
- Three examples
  - NLA – based on document: "Workflows for still image digitization"
    - http://www.nla.gov.au/digital/stillimagedigitisationandworkflows.html
  - Wroclaw University Library
    - http://bibliotekacyfrowa.pl/
  - Silesian Digital Library
    - http://sbc.org.pl/dlibra

Workflows for still images

- Other resources:
  - “GREENSTONE DIGITAL LIBRARY FROM PAPER TO COLLECTION”
    - http://www.greenstone.org/manuals/gsdl2/en/html/Chapter_three_examples.htm
  - “The Great War Poetry Archive”
    - http://www.thegreatwatarchive.org/

Workflows for still image digitization - 1

- The following workflows are used by the NLA in its collection digitization program
- **Selection**
  - Collection managers select the collection material for digitization
- **Preservation assessment**
  - The condition of all collection materials is considered prior to digitization
  - Preservation treatments are undertaken by the Preservation Branch staff as required

Workflows for still image digitization - 1

- **Bibliographic description**
  - All material to be digitized is catalogued using Voyager, the local Integrated Library Management System (ILMS)
  - Then the records are uploaded to the Australian National Bibliographic Database (NBD)
- **Persistent Identifiers**
  - Persistent identifier (PI) is devised for each digital file prior to digital image capture
  - It will remain constant throughout the life of the file

Workflows for still image digitization - 1

- **Digital image capture**
  - Digital images are created by either scanning the material on a flat bed scanner or photographing it with a digital camera
- **Image processing**
  - The resulting images are cropped and, when required, rotated in Adobe Photoshop

Workflows for still image digitization - 1

- **Image upload**
  - The master images are loaded to the Library's Digital Object Storage System (DOSS) as uncompressed TIFF files
  - Metadata is extracted from TIFF headers using tifftool

Workflows for still image digitization - 1

- **Online delivery**
  - derivative images (e.g. thumbnail, examination copies) are automatically generated by Image Alchemy for delivery to users through
    - The Library's catalogue - delivery systems
    - Federated resource discovery services
- **Quality assurance** (QA)
  - QA is undertaken at several stages during the digitisation process, i.e.
    - at digital image capture, image upload, image acquittal and when the images are first made available online

Wroclaw University Library – example 2

- WUL digitization department is divided into three parts :
  - Digitization section
  - Archive and delivery
  - Metadata preparation
- Results of digitization process are published in Wroclaw University Digital Library:
  - http://www.bibliotekacyfrowa.pl/

Wroclaw University Library – example 2

- Each scanner type is handled by a dedicated person
- Scanner operator should take care of *:
  - Assuring that proper initial parameters are set for the scanner
  - Setting proper scanning resolution depending on type of resources
  - Setting the most accurate speed of scanning head
    - For low resolution speed should be set to 0.5
    - for higher 0.25
    - In special cases (manuscripts) – 0.125

* This is just an example, not all scanners require such a handling

Wroclaw University Library – example 2

- Scanner operator should take care of *:
  - Settings proper filters
  - Adding basic metadata to files obtained from scanners (owner, signature, short description)
  - Other operations necessary to assure high quality of results

* This is just an example, not all scanners require such a handling

Wroclaw University Library – example 2

- After scanning process is finished, information about number of scanned pages and object's signature should be **logged in shared spreadsheet** (docs.google.com)
- Example of how scanner operator should deal with scanned materials

Wroclaw University Library – example 2

- Initial overview:
  - Book is complete, with cover, there are some illustrations at centrefolds
- Work plan:
  - Scan covers:
    - 600 DPI, head speed 1/8, without pressing scanner glass plate
  - Scan pages without graphics
    - 300 DPI, head speed ¼
  - Scan centrefolds
    - 600 DPI, head speed ¼

Wroclaw University Library – example 2

- Additional remarks:
  - To reduce visibility of traces of text from preceding page put cream piece of paper under the centrefolds
- Total number of scans:
  - 300 (27 pages in 600 DPI, 273 pages in 300 DPI)
- Total working time: 4 days
  - work was interrupted by other orders from commercial clients

Wroclaw University Library – example 2

- Only some scanners require/allow to manipulate head speed
- In this particular example Zeutschel 10 000 TT was used
- This is just an example of how book may be scanned
- After scanning, images are checked for completeness and prepared for further treatment using software like Gimp or Photoshop
    - Both of them may be used for the same purpose : cropping, rotation, batch treatment (using Photoshop macros)

Wroclaw University Library – example 2

- Next stage includes conversion to delivery format and preparing a batch bundle for dLibra
- WUL delivers objects in one of formats:
  - DjVu, HTML, PDF, JPEG
- Software used:
  - Document Express Profesional 5.0
  - Fine Reader 7.0
  - Fine Reader XIX

Wroclaw University Library – example 2

- Other software used:
  - Total commander - multiple files rename, putting files in proper directories
  - Virtual Printer – conversion between PDF and DjVu
- Result of this stage:
  - DjVu/PDF file, with text layer (after OCR)
  - Thumbnail of the first page
  - digital master copy compressed using lossless LZW

Wroclaw University Library – example 2

- Such a set of files is once again validated, PDF files are optimized for the web.
- FineReader is usually better in case of OCR quality, it is often used instead of Iris OCR engine used in Document Express
- This requires one additional step for DjVu files because FR produces output in PDF files
- So one additional conversion is done to get back to DjVu

Wroclaw University Library – example 2

- **Metadata preparation** is done in different departments of library, depending on the origin of digitized resources
- Usually curators are describing objects from their own collection
- Object is usually catalogued before the whole digitization process begins.

Wroclaw University Library – example 2

- Web optimized version of object is uploaded to digital library
- dLibra software allows for creation of planned publications which consist only from metadata
- When files of one or more planned publications are ready – batch upload can be started
- dLibra will convert planned publications to normal publications and associate metadata with uploaded files

Wroclaw University Library – example 2

- They are also trying to scan every item only once
- So in some cases when object was microfilmed , digitization is done from microfilms
- Hardware used :
  - Zeutschel OS 10000 TT, Zeutschel OM 1600
  - Kyocera KM 4035
  - Nikon D200
  - Scamax

Wroclaw University Library – example 2

- There is also a dedicated person who records DVDs with Digital Master Copies
- It is a good practice to record a checksum (MD5, SHA1) of stored files
- Thanks to this it would be possible to verify the consistency of file loaded from DVD

Wroclaw University Library – example 2

- Each step in the workflow is documented using Google Docs
- This allows to monitor performance and notify about finished jobs

Social Digitisation Lab – example 3

- Social Digitisation Lab is operating in Silesian Library in Katowice since October 2007
- The project was launched mainly to satisfy different operational and educational objectives of Silesian Library
- Social Digitisation Lab is a place where people (in particular seniors and trainees) are digitizing books and newspapers

Social Digitisation Lab – example 3

- Such an inexperienced stuff can cope only with resources which are less sensitive for damage
- Library offers trainings and assures assistance for volunteers
- During the first year **23 volunteers** managed to digitize **101 705 pages**

Social Digitisation Lab – example 3

- This project has also other social benefits
  - e.g. Seniors learn how to use computers and internet
- All the equipment comes from national funding.
- Results of volunteer work are published in Silesian Digital Library
  - http://sbc.org.pl/

Agenda

- Digitization workflows
- **Cutting cost of digitization**
- Remarks about resolution for scanning text document
- OCR
- Formats for non-still images digitization
  - Preservation and web delivery
- Useful tools

Cutting cost of digitization

- This is an excerpt from “Minerva: Handbook on cost reduction in digitization”
  - http://www.minervaeurope.org/publications/CostReductioninDigitisation_v1_0610.pdf
- The most frequently used methods for reducing costs in digitisation are:
  - Reduce the cost of labour
  - Automate to reduce levels of human intervention in digital conversion and metadata creation

Cutting cost of digitization

- The most frequently used methods for reducing costs in digitization are:
  - Select and prepare originals to enable higher volumes and reduce variation in the workflow
  - Increase overall performance and throughput to make the most efficient use of capital expenditure
  - Continuous improvement and optimisation through rigorous quality assurance

Reducing the cost of workforce

- The higher the level of human intervention the greater costs associated to digitization process
- Basic means to reduce the level of intervention may be achieved by:
  - automating the scanning mechanism or
  - metadata creation process
- The general rule for cost reduction is to look at every stage that requires human intervention and either **remove it**, **reduce it** or **make it** as **efficient** as possible

Reducing the cost of labour

- Salary reduction usually means that the task is made easier thanks to that lower skilled and less expensive staff are needed
- The objective is to redesign the activity so that highly skilled staff are only used when absolutely necessary
- Lower qualified (lower paid) staff can carry out repetitive or straightforward tasks

Reducing the cost of labour

- In many digitisation projects:
  - 85% of work is repetitive or routine
  - 12-15% - difficult
  - <1% - very difficult
- Employ high paid staff only when needed
  - For solving very difficult problems

Reducing the cost of labour

- This can be achieved by:
  - Breaking the activity down into modules
  - Provide better tools and guidance
    - Use of better tools and proper guidance can significantly reduce the skills needed
  - Invest in training
    - lower paid member of staff with appropriately focused training may be able to achieve the same performance in a narrowly defined activity as a more costly member of staff

Outsourcing?

- For large volumes, outsourcing will generally be cheaper than setting up in-house digitisation processes.
- But…

In-house digitization

- There are some circumstances in which considering in-house digitisation will continue to offer advantages for practical and cost reasons, such as:
  - the collection is difficult to move or cannot be moved outside of the institution
  - the collection is badly organised, not inventoried to the item level and needs skilled reorganisation as an integral part of the process

In-house digitization

- the digitisation needs to be phased in relatively small amounts over a long period
- 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

Automation

- There are a number of ways in which automation can significantly reduce the costs of digitisation
  - Mechanistic automation:
    - replacing or reducing the human handling of original materials.
  - Software based automation:
    - speeding processes, replacing human intervention, or enabling end user interaction that means less effort at creation

Automation

- Automation is important but it should always be noted that automation is not a panacea for reducing costs

Mechanistic automation

- When we focus narrowly on the pure costs of image scanning as opposed to the whole workflow there are two cost elements:
  - **The cost of handling** or otherwise moving the original material through the scanning process
  - The cost of **writing an output image file** to the required resolution, bit depth and quality
- Technology costs have been reduced for everything but very large files
- This has been achieved through improvements in computer hardware - the use of removable storage, Storage Area Networks and optical fibre networks

Mechanistic automation

- The cost of handling is mainly related to the amount of automation that is possible in a process. For example:
  - A4 laser printed sheets can be passed automatically through a scanner using sheet feeders
  - Bound volumes need every page individually turned on a bookscanner
  - Photographic prints cannot easily be fed automatically through a scanner because they will be damaged and jam the mechanism
  - 35mm mounted photographic slides can be loaded into a carousel for automated batch scanning

Mechanistic automation

- The cost of handling is mainly related to the amount of automation that is possible in a process. For example:
  - Microfilm rolls can be scanned automatically, but microfiche and jacketed film tend to still need human supervision
  - Glass plate photographs can take up to **3-5 minutes each** just to **:**
    - **get the plate out** of its enclosure and
    - put onto the scan mechanism
    - take a scan
    - and then remove the plate back to its enclosure.
  - Automation of the transit of materials through a scanner will reduce unit costs – but equip. It is very expensive.

Software based automation

- Batch image manipulation
  - Cropping, surrogate creation etc.
  - Character recognition of textual content

Selection and preparation

- By far the greatest of these costs is preparation, including activities like
  - Transportation planning
  - Time taken to assign unique identifiers to originals
  - Preservation risk management
  - Preparing the physical item
    - Removing the object from its enclosure, removing staples

Selection and preparation

  - The cost of clearing copyright
  - Inventory check on all original items returned to ensure everything has been returned
- It is possible for selection and preparation could account for **20-30% of the total digitisation unit cost**
- This cost can be reduced by applying some simple “tricks”

Reducing cost of preparation

- Select whole collections where possible
- Batch originals by their physical nature for scanning
- Organise the workflow so that low cost labour is used and that preparation is appropriate for the scan mechanism
- Inventories may be automatically created from existing catalogues and other indexes

Reducing cost of preparation

- Copyright clearance is relatively expensive in time and effort, but cheaper than:
  - litigation; loss of reputation; or having to remove the digital version
  - A clear procedure and records will make this easier and cheaper
- If each original is given a clear and unique identifier then this
  - speeds inventory
  - makes scanning and metadata capture easier and faster

Quality Assurance

- Quality Assurance (QA) that is embedded and a natural part of the activity is **a major source of cost reduction for digitisation**
- Done well it provides opportunities for **continuous improvement** and optimisation **of processes**
- It needs to be systematic, focussed and **proactive not passive**
- One mistake often made with QA is to assume it is purely about finding and correcting errors

Agenda

- Digitization workflows
- Cutting cost of digitization
- **Remarks about resolution for scanning text document**
- OCR
- Formats for non-still images digitization
  - Preservation and web delivery
- Useful tools

Remarks about text documents

- In case of **really small fonts**, consider using **Quality Index** (QI)
- It is a base of formal approach to determine resolution used at Cornell University
  - Based on formula for microfilmed text that was developed by the C10 Standards Committee of AIIM

Remarks about text documents

- The QI formula relates quality (QI) to smallest character height (h) in mm and resolution (dpi).
- As in the preservation microfilming standard, the digital QI formula forecasts levels of image quality:
  - barely legible (3.0),
  - marginal (3.6),
  - good (5.0),
  - and excellent (8.0).

Remarks about text documents

- There are two different formulas:
  - One for B/W images
  - And one for grayscale/color images
- The formula for B/W scanning provides some oversampling to compensate for **misregistration** and **reduced quality** (due to thresholding information to black and white pixels)

Remarks about text documents

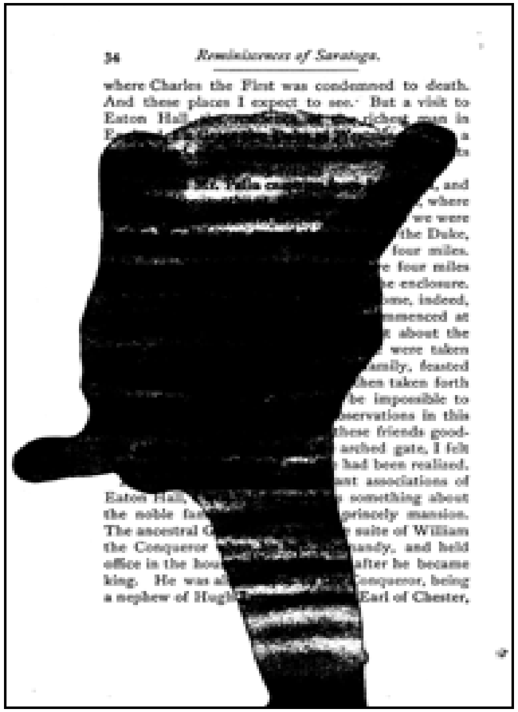
- B/W QI Formula for Printed Text
  QI = (dpi x 0.039h)/3
  h = 3QI/0.039dpi
  dpi = 3QI/0.039h
- Note: if the measurement of h is expressed in inches, omit the 0.039
- Assuming **QI=8** and **h=1mm** we will have to scan B/W content in **600 dpi**

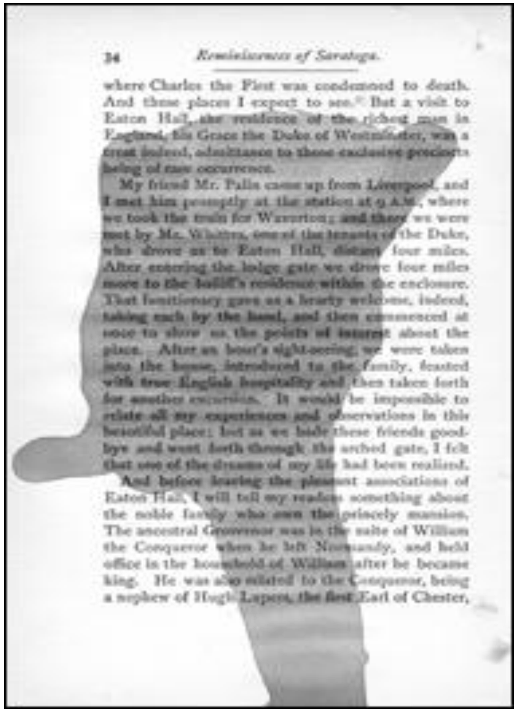
Remarks about text documents

- Some printed text will require greyscale or colour scanning for the following reasons:
  - Pages are badly stained
  - Paper has darkened to the extent that it is difficult to threshold the information to pure black and white pixels
  - Pages contain complex graphics or important contextual information
    - e.g. embossments, annotations
  - Pages contain colour information
    - e.g. different colours inks

Remarks about text documents

- Some printed text will require grayscale or colour scanning





34 *Reminiscences of Saratoga.*

where Charles the First was condemned to death. And these places I expect to see. But a visit to Eaton Hall, the residence of the richest man in England, His Grace the Duke of Westminster, was a treat indeed, admittance to these exclusive precincts being of rare occurrence.

My friend Mr. Pallis came up from Liverpool, and I met him promptly at the station at 9 A.M., where we took the train for Warrington; and there we were met by Mr. Whitten, one of the tenants of the Duke, who drove us to Eaton Hall, distant four miles. After entering the lodge gate we drove four miles more to the bailiff's residence within the enclosure. That functionary gave us a hearty welcome, indeed, taking each by the hand, and then commenced at once to show us the points of interest about the place. After an hour's sight-seeing we were taken into the house, introduced to the family, feasted with true English hospitality and then taken forth for another excursion. It would be impossible to relate all my experiences and observations in this beautiful place; but as we bade these friends good-bye and went forth through the arched gate, I felt that one of the dreams of my life had been realized.

And before leaving the pleasant associations of Eaton Hall, I will tell my readers something about the noble family who own the princely mansion. The ancestral Grosvenor was in the suite of William the Conqueror when he left Normandy, and held office in the household of William after he became king. He was also related to the Conqueror, being a nephew of Hugh Lupus, the first Earl of Chester,

Remarks about text documents

- Grayscale/Color QI Formula for Printed Text
  QI = (dpi x 0.039h)/2
  h = 2QI/0.039dpi
  dpi = 2QI/0.039h
- Note: if the measurement of h is expressed in inches, omit the 0.039
- Assuming **QI=8** and **h=1mm** we will have to scan greyscale content in **400 dpi**

Remarks about text documents

- Usually "e" can be treated as the smallest character
    - Source: "Digitisation as a Method of Preservation? Final report of a working group of the Deutsche Forschungsgemeinschaft"
        - European Commission on Preservation and Access,
        - http://www.knaw.nl/ecpa/publ/weber.html
- Even when using such a formal approach it is always worth to look at the results of digitization and verify correctness of assumptions

Agenda

- Digitization workflows
- Cutting cost of digitization
- Remarks about resolution for scanning text document
- **OCR**
- Formats for non-still images digitization
  - Preservation and web delivery
- Useful tools

Text capture: Optical Character Recognition and Rekeying

"Much work has gone into creating web-accessible electronic catalogues, as a logical first step, but these are only the foothills of digital librarianship compared with the bigger challenge of creating full-text digital content."

*"**Digitisation: do we have a strategy?**"*

David Pearson, 2001

Text capture: Optical Character Recognition and Rekeying

- What is important while capturing text from scans?
  - Whether it is a scanned printed text or manual handwriting
  - Quality of print
  - Quality of original document
  - Scans resolution, colour resolution
    - B/W is not always optimal in case of old prints
    - OCR software works better with high contrast images

Text capture: Optical Character Recognition and Rekeying

- What is important while capturing text from scans?
  - Language of text
    - Availability of dictionaries and language models
  - Text layout
  - Text formatting
    - Adequate white space between lines, columns and at edge of page so that text boundaries can be identified

Text capture: Optical Character Recognition and Rekeying

- When OCR quality is acceptable?
  - Quality evaluation in most cases is done by humans
  - It can be done for a sample part of text
  - It can be measured at word and character level
  - OCR accuracy can vary very broadly depending on mentioned factors

Text capture: Optical Character Recognition and Rekeying

- NLA published some indicators of OCR quality for historical newspaper
  - http://www.dlib.org/dlib/march09/holley/03holley.html
- OCR accuracy is
  - Good = when 98%-99% is accurate
  - Average = when 90%-98% is accurate
  - Poor = below 90% accuracy

Text capture: Optical Character Recognition and Rekeying

- In case of OCR, the question is exactly the same as in case of digitization
- We can spent our own workforce for this or outsource this activity
- National Library of the Netherlands performed a study where they surveyed some OCR services providers, detailed description can be found at:
  - http://www.dlib.org/dlib/january08/klijn/01klijn.html

Text capture: Optical Character Recognition and Rekeying

- When developing OCR in-house, one should consider using one of widely used OCR engines like:
  - ABBYY Finereader (http://www.abbyy.com/)
    - offers a vast selection of modules for recognizing specific types of content (XIX, gothic fonts)
    - proprietary software
  - Document Express (http://lizardtech.com/)
    - with its Iris OCR engine
    - proprietary software

Text capture: Optical Character Recognition and Rekeying

- When developing OCR in-house, one should consider using one of widely used OCR engines like:
  - *Omnipage* (http://www.omnipage.com/)
  - Tesseract/OCRpus
    - http://code.google.com/p/tesseract-ocr/
    - Tesseract is an OpenSource OCR engine originated from Hewlet - Packard Lab.
    - http://code.google.com/p/ocropus/
    - OCRpus is layout analyzer, which is necessary in case of newspapers digitization
    - At the moment funding for development of these projects is coming from Google
    - There is no paid support

Text capture: Optical Character Recognition and Rekeying

- As was said OCR quality can be very different depending on various factors
- There is no clear statement about publishing of dirty (not fully corrected) results of OCR
- Some misspelled words can create false search results
- What can be done?
  - Correct OCR manually
  - Crowdsource OCR correction
    - e.g. NLA's Newspapers Digitisation Programme
  - Rekeying

Text capture: Optical Character Recognition and Rekeying

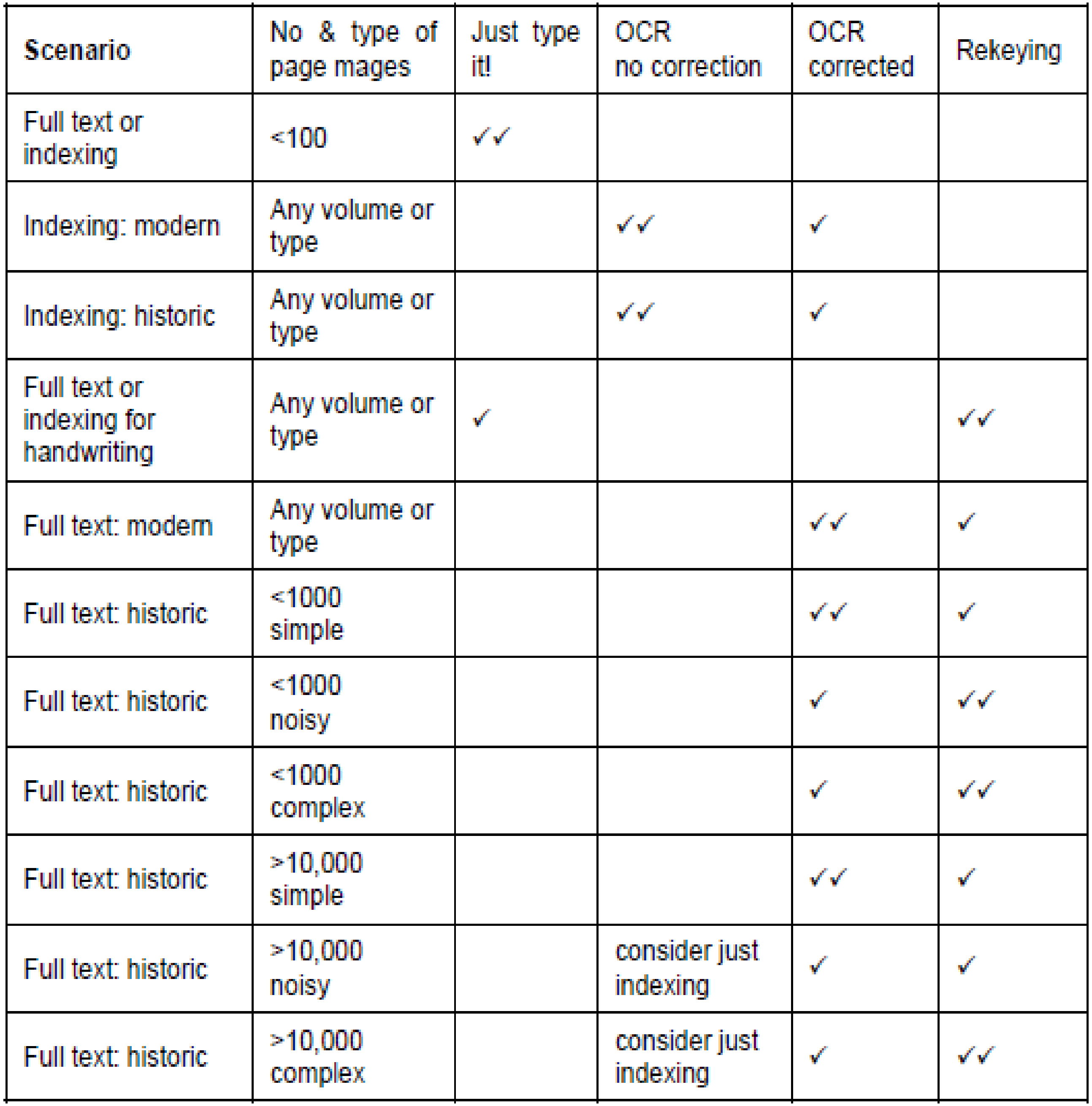
- What is the difference between rekeying and simple OCR correction or typing text in?
  - "rekeyed" — means, manually typed in without the help of OCR software
  - The example of such a process is the **triple keying procedure** used when we need documents with close to **100 percent accuracy**
  - Two people type the same document; then a third person reviews the discrepancies identified by a computer
  - **This is expensive but when OCR have to be 100% correct this seems to be good choice**

Text capture: Optical Character Recognition and Rekeying

- Which option should be chosen, see the following table :
  - Simple = Very clear
    - Cleanly printed text in a single column
    - One language only with no scientific notation, small font sizes, unusual characters/words, tables, graphics or illustrations
  - Noisy = Same as Simple
    - Except the printed text is not clear or clean because of factors such as dirt, tears, foxing, other marks, creases or show through

Text capture: Optical Character Recognition and Rekeying

- Which option should be chosen, see the following table :
  - Complex = As Simple but includes either multiple columns, multiple languages, scientific notation, small font sizes, unusual characters/words, tables, graphics or illustrations
  - Modern = Post 1950's printed text from a book or journal
    - content is mainly black and white with some grayscale or color
  - Historic = Pre 1900's printed text from a book or journal
    - content is mainly black and white with some grayscale



Scenario	No & type of page mages	Just type it!	OCR no correction	OCR corrected	Rekeying
Full text or indexing	<100	✓✓			
Indexing: modern	Any volume or type		✓✓	✓	
Indexing: historic	Any volume or type		✓✓	✓	
Full text or indexing for handwriting	Any volume or type	✓			✓✓
Full text: modern	Any volume or type			✓✓	✓
Full text: historic	<1000 simple			✓✓	✓
Full text: historic	<1000 noisy			✓	✓✓
Full text: historic	<1000 complex			✓	✓✓
Full text: historic	>10,000 simple			✓✓	✓
Full text: historic	>10,000 noisy		consider just indexing	✓	✓
Full text: historic	>10,000 complex		consider just indexing	✓	✓✓

Source: „Minerva: Handbook on cost reduction in digitization”

Text capture: Optical Character Recognition and Rekeying

- What for manual handwriting and really poor prints?
  - Create transcription manually
  - Crowdsource transcription creation
  - Rekeying
- The same is for voice and video text capture, there are no well-known tools to cope with this
- You can always employ “manual” techniques

Text capture - References

- "How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs"
  - http://www.dlib.org/dlib/march09/holley/03holley.html
- „Crowdsourcing Improves Historical Newspapers”
  - http://www.nla.gov.au/pub/gateways/issues/102/story06.html
- "Going Grey? Comparing the OCR Accuracy Levels of Bitonal and Greyscale Images"
  - http://www.dlib.org/dlib/march09/powell/03powell.html

Text capture - References

- “GREENSTONE DIGITAL LIBRARY FROM PAPER TO COLLECTION
  Chapter 3 OCR: Optical Character Recognition”
  - http://www.greenstone.org/manuals/gsdl2/en/html/Chapter_ocr.htm
  - Contains also some remarks about OCR correction productivity and tools which might be used for this

Agenda

- Digitization workflows
- Cutting cost of digitization
- Remarks about resolution for scanning text document
- OCR
- **Formats for non-still images digitization**
  - Preservation and web delivery
- Useful tools

Formats for non-still images digitization

- Source: “*MINERVA Technical Guidelines for Digital Cultural Content Creation Programmes*”
- By non-still images we particularly mean :
  - Audio and video
  - 3d worlds
  - Vector based graphics
- Digitisation of audio/video content can be quite expensive and requires additional knowledge
- Think twice before you start on your own

Digitisation of video content

- Video should be stored using
  - the uncompressed **RAW AVI** format
  - **without the use of any codec**
  - at a frame size of **720x576** pixels
  - a frame rate of **25 frames per second**
  - using **24-bit color**
  - **PAL colour encoding** should be used
- Video may be created and stored using
  - appropriate MPEG format (MPEG-1, MPEG-2 or MPEG-4)
  - or the proprietary formats Microsoft WMF, ASF or Quicktime

Digitisation of video content

- Analog Video was stored on a huge number of different carriers
  - Betacam, VHS, Hi8, SVHS
- The proper digitization equipment may be hard to obtain
- Because of huge size of digitized video files it is worth to consider storing videos of different kind in different resolution
  - e.g. use higher bitrates for movies and lower for everyday TV auditions (e.g. News)

Digitisation of video content

- Web version can have lower resolution
  - Video may be delivered as a **FLV** (Flash Video), **WMV** or open standards like **OGG** format (using Vorbis and Theora codecs)
  - It is worth to consider publishing more than one resolution/bitrate
- For watching video content on the web, users often need to install additional software like
  - Adobe Flash, Windows Media Player (or compatible)
- It is worth to consider which technology allows to reach greater audience

Digitisation of audio content

- Audio should usually be stored in the uncompressed form obtained from the recording device without the application of any subsequent processing such as noise reduction
- Audio Master Copy should be created and stored as
  - an uncompressed format such as Microsoft WAV or Apple AIFF
  - **24-bit stereo** sound at **48/96 KHz sample rate**
- This sampling rate is suggested by
  - Audio Engineering Society (AES)
  - International Association of Sound and Audiovisual Archives (IASA).

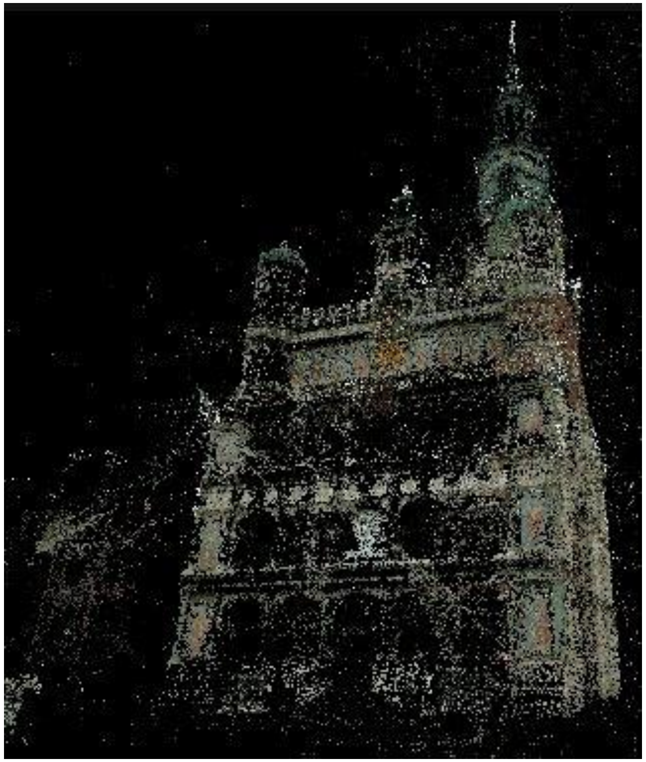
Digitisation of audio content

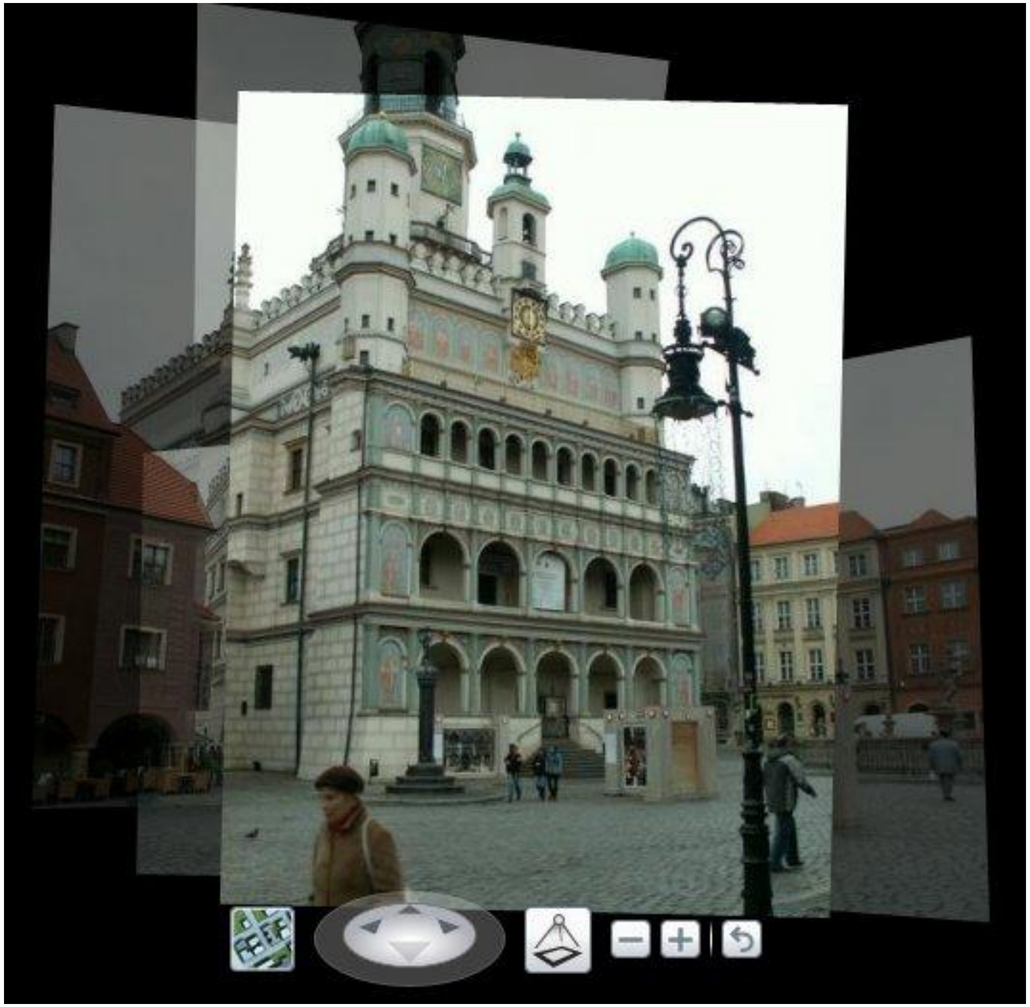
- Audio may be created and stored using compressed formats such as **MP3, WMA, RealAudio** formats
- For web delivery
  - Sampling rate can be adjusted depending on type of recording
    - e.g. higher should be used for music track
  - For good quality (CD quality) **256 Kbps**, but also 160 Kbps gives decent quality

Digitisation of 3D content

- 3D scanners are still very expensive and can cope only with quite small objects.
- 3D reconstruction can be done in different ways
  - e.g. from huge picture corpora : http://photosynth.com

Digitisation of 3D content





http://photosynth.net/view.aspx?cid=1c851c1b-f11b-44bd-9696-901a565c7fd5

Digitisation of 3D content

- Many authorities recommends utilisation of X3D which is a widely accessible open standard
- Formerly the most widely known standard for encoding 3D Word was VRML
- Both of them requires that users will install additional software – browser plugins
- Exposing large 3D words may require quite huge computational power

Digitisation of 3D content

- Projects may wish to consider using managed 3-D virtual worlds (such as **Second Life**, **OpenSimulator**) to engage new audiences and display digital assets.

VRML-based virtual exhibition
http://dlibra.psnc.pl/biblioteka/publication/114

Digitisation of 3D content

- There are also other ways to expose semi-3d experience
- Can be done using picture stitching software (e.g. Hugin) and some additional Javascript/Flash/Java tools
    - http://www.mathieusavard.info/threesixty/demo.html
    - http://www.openstudio.fr/jquery-virtual-tour/
    - http://www.all-in-one.ee/~dersch/StBp_ptvj.html

Formats for vector graphics

- The most widely used open standard for encoding vector based graphics is a Scalable Vector Graphics (SVG) format.
  - http://www.w3.org/Graphics/SVG/
- It is XML-based and most of modern web browser knows how to display such a graphics

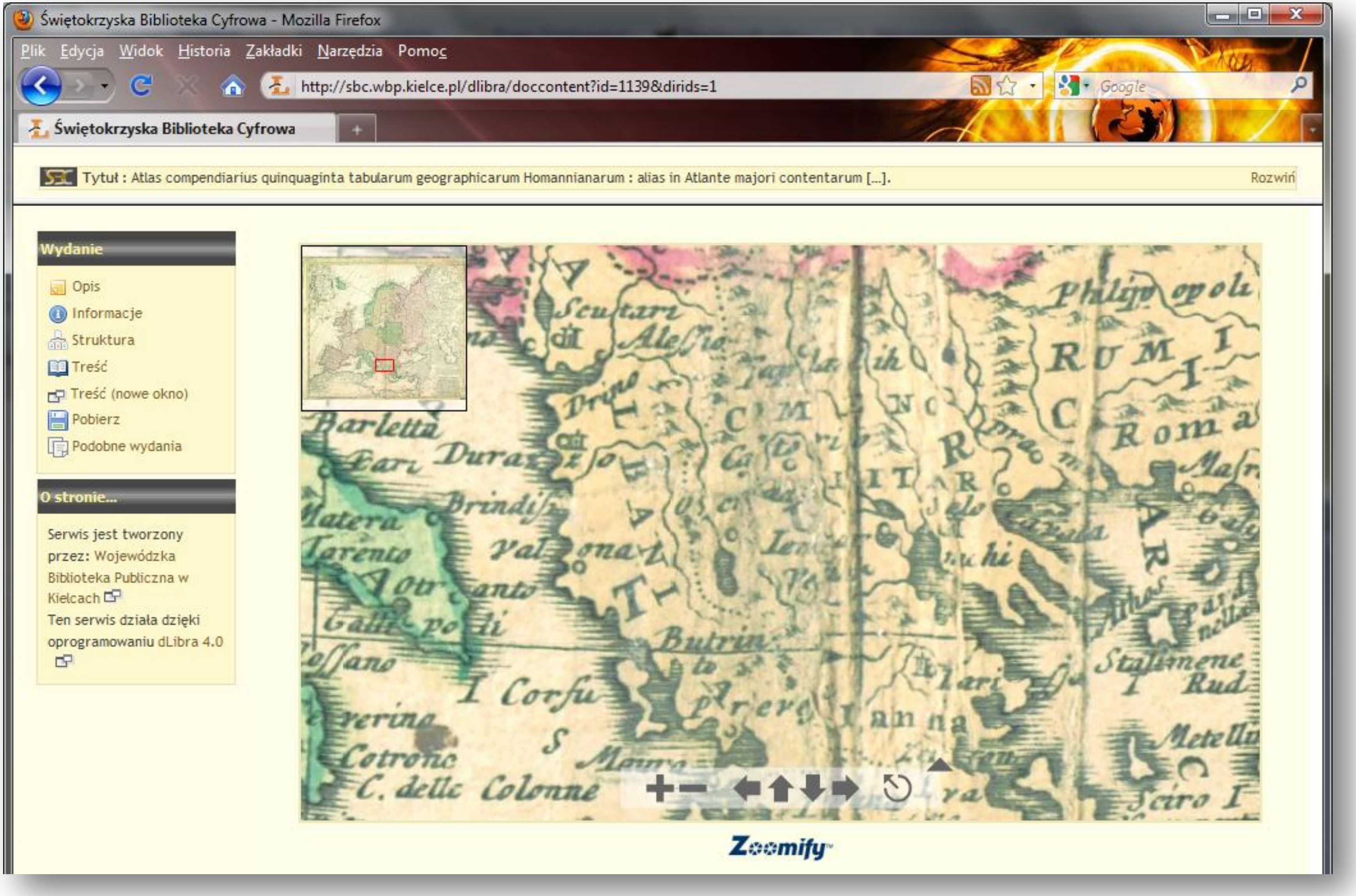
Agenda

- Digitization workflows
- Cutting cost of digitization
- Remarks about resolution for scanning text document
- OCR
- Formats for non-still images digitization
  - Preservation and web delivery
- **Useful tools**

Useful tools

- Zoomify format
  - http://www.zoomify.com/
  - Good for providing high quality huge images e.g.maps

http://fbc.pionier.net.pl/id/oai:sbc.wbp.kielce.pl:1139



Świętokrzyska Biblioteka Cyfrowa - Mozilla Firefox
Plik Edycja Widok Historia Zakładki Narzędzia Pomoc
http://sbc.wbp.kielce.pl/dlibra/doccontent?id=1139&dirids=1
Google
Świętokrzyska Biblioteka Cyfrowa
Rozwiń
Wydanie
Opis
Informacje
Struktura
Treść
Treść (nowe okno)
Pobierz
Podobne wydania
O stronie...
Zoomify

Tytuł : Atlas compendiarius quinquaginta tabularum geographicarum Homannianarum : alias in Atlante majori contentarum [...].

Serwis jest tworzony przez: Wojewódzka Biblioteka Publiczna w Kielcach

Ten serwis działa dzięki oprogramowaniu dLibra 4.0

Useful tools

- Free software
  - Graphical tools :
    - Gimp
      - http://gimp.org/
    - Picasa
      - http://picasa.google.com
    - Hugin
      - http://hugin.sourceforge.net/
    - ImageMagick
      - http://www.imagemagick.org/script/index.php
    - PDF2DJVU
      - http://code.google.com/p/pdf2djvu/
    - BullZip free PDF printer
      - http://www.bullzip.com/products/pdf/info.php

Useful tools

- Free software
  - Graphical tools :
    - Zoomify Express
      - http://www.zoomify.com/express.htm
    - Scribus
      - http://www.scribus.net/
    - Gsview
      - http://www.ghostscript.com/

Useful tools

- Free software
  - Other tools:
    - Total Commander
      - http://www.ghisler.com/
    - Google Docs
      - http://docs.google.com
    - Open Office
      - http://openoffice.org
    - Sun PDF Import Open Office Extension
      - http://extensions.services.openoffice.org/project/pdfimport
    - PDFedit
      - http://pdfedit.petricek.net/en/index.html
  - OCR :
    - Tesseract/OCRpus
      - http://code.google.com/p/ocropus/
      - http://code.google.com/p/tesseract-ocr/

Useful tools

- Commercial software
  - Graphical tools :
    - Adobe Photoshop
      - Zoomify Export in Adobe Photoshop
    - Paint Shop Pro
    - Image Alchemy
    - Adobe Acrobat
  - Other tools:
    - Microsoft Word - for OCR correction
  - OCR :
    - ABBYY FineReader
    - Document Express (DjVu)

Q&A