

Digitization

from physical to digital worlds

Tefko Saracevic, Ph.D.

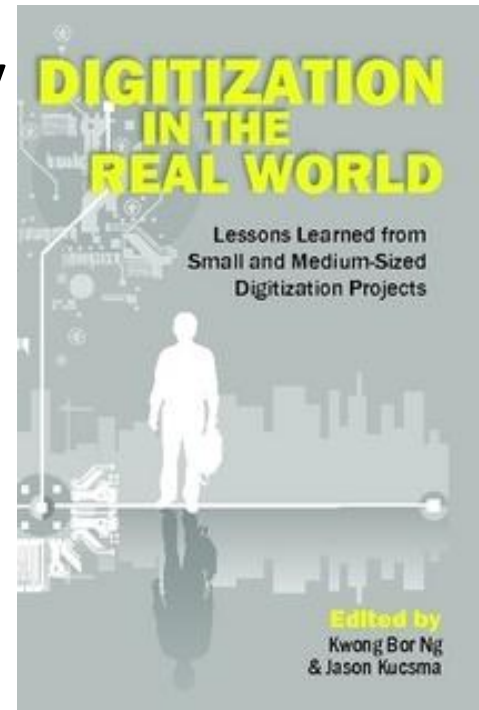
This work is licensed under a
[Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States License](https://creativecommons.org/licenses/by-nc-sa/3.0/)

Digitization is not just a process,
but decisions, decisions, decisions
related among others to
selection
standards
best practices
technology
\$\$\$ \$\$\$ \$\$\$
management
& more

What is Digitization? A glossary

useful to go through, just to clarify in your own mind

Also a book co-edited by Rutgers graduate Kwong Bor Ng:



ToC

- Context & processes
- Selection
- Scanning
- Quality control
- Posting
- Large scale projects

Context:

Classic critical questions revisited

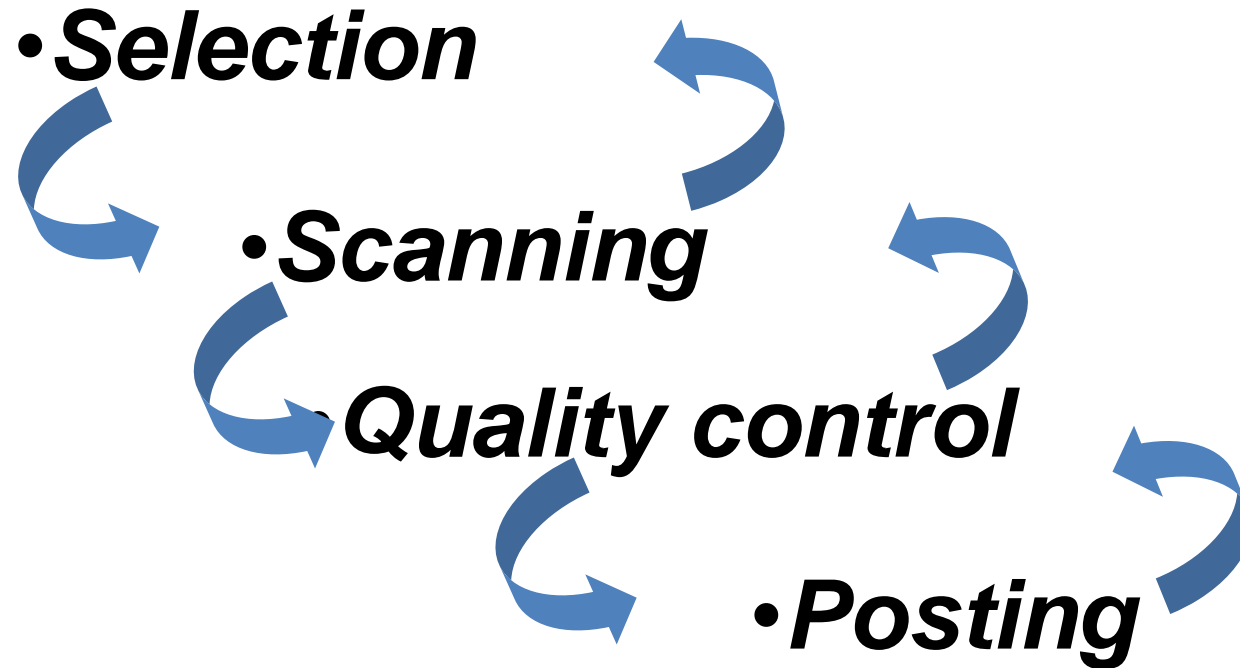
To repeat Michael Lesk's questions:

- Does everyone want to get everything on their screen?
- If so, can we do it technically?
- If so, can we do it legally?
- If so, can we afford to do it?
- If so, should we do it?

Answers to each of these questions are critical in decisions on digitization:

- What to digitize?
- How technically?
- How legally?
- How economically?
- What policy?

Processes in digitization



More specific - what is involved in digitization

- Project management - to start with
- Selection of materials, based on defined criteria
- Consideration of copyright & intellectual property
- Appropriate treatment for selected materials
- Care and handling of fragile materials, if any
- Image scanning and editing plus quality control
- Storage and retrieval of images
- Indexing and description (metadata)
- Posting: access and delivery via the Web

Selection - what to digitize: *two related processes*

General (setting the stage):



Specific (doing the work):

- Based on institutional mission, community needs
 - Defining domains, (topics, subjects) for selection
 - Specifying policies & criteria, (rules, intellectual value, audience, rights management, ...)
 - similar to library selection policies
 - Considering legal, economic aspects
 - Considering physical size, condition, characteristics
- Selection of resources to digitize
 - documents, artifacts
 - from within - owned, controlled
 - from outside -acquired, borrowed, accessed
 - But both have to pass legal test
 - Care of materials

Legally speaking

(to be covered in more detail in the legal lecture)

- The place to begin
- Resources (books, scores, images, music...) are governed by copyright laws
 - differ somewhat from country to country
- Selection has to consider these aspects
 - but various classes of works have different legal requirements to be followed in selection, digitization and providing access

Legally speaking: Three classes of works

1. **Copyrighted** – in print and being sold
 - generally, access restricted; permission must be sought for digitization; snippets, parts may be shown after a search e.g. as in Amazon
2. **Orphaned** - in copyright but out of print, not sold
 - permissions still must be sought; but legally murky; easier obtained for digitization
3. **Public domain** - but watch: even old works (e.g. when Hamlet is annotated) can be copyrighted
 - if open access is available - easy selection for digitization

Selection: Digitizing heritage

- Many libraries world-wide select own (or available) materials that reflect local, national, cultural, social heritage for digitization – most, if not all, in public domain
 - so do other institutions (museums, archives, historical societies ...)
 - sometimes in cooperative projects with libraries
- Generally, well received & high use; popular; often related to education

Digitizing heritage ...

- Provides open access for works previously with restricted access
 - many were hidden in the bowels of institutions & accessible by selected few only
- Selection criteria vary by institution
- But also sometimes selection seems like a grab bag
 - Marija Dalbello says: “cabinets of curiosities”

Heritage materials

- Wide range of primary sources:
 - Archival material, manuscripts, diaries, personal journals, letters, notes
 - Photographic prints, negatives, movie films
 - Sheet material: handwritten musical scores, maps, posters, drawings, prints
 - Electronic media: video tapes, video disks, computer magnetic tapes, floppy disks
 - Sound recordings: Oral histories, tapes
 - Name it ...

Digitizing heritage: Examples

- A number given in Diversity lecture – three more:
 - [Calisphere](#) (California) - “a world of primary sources and more”
 - [New Jersey Digital Highway](#) – “your “one stop shop” for New Jersey history and culture, from the collections of NJ libraries, museums, archives and historical societies.” RUL a major partner
 - [Gingerbread Castle Digital Library](#) - “In the small town of Hamburg, New Jersey one of America’s earliest theme parks was created in 1930”
 - a project in this course & independent studies by Judith L. Panagokos, LIS student, 2012; see also section [Building a Digital Library – The Challenges](#)
- [Institute of Museum and Library Services](#)
 - provides grants & other activities about heritage through various programs [National Initiatives & Partnerships](#)

World Digital Library

the ultimate heritage example - revisited

- A vision started from Library of Congress now with UNESCO
 - “The WDL makes it possible to discover, study, and enjoy cultural treasures from around the world on one site, in a variety of ways. These cultural treasures include, but are not limited to, manuscripts, maps, rare books, musical scores, recordings, films, prints, photographs, and architectural drawings.”
- Still in development stage - but excellent example
 - more than 14,000 items from all over the globe
 - items can be downloaded; also browsed by place, type of item, time etc.

Other selections

- A variety of other materials, beside heritage, selected for digitization
 - datasets
 - scholarly reports, papers
 - public domain documents
 - books
 - soundtracks, videotracks
 - dissertations
 - name it ...
- Same selection principles apply

Scanning preliminaries

- What digital imaging technology & software to choose?
 - depends (among others) on answers to questions below & available \$\$\$\$
- What features must be captured?
- What is essential in respect to:
 - resolution (how high?)*
 - accuracy of rendition, including colors
 - seamless combination of texts & images
 - other qualities
 - original sources to be retained?
 - protection of integrity

*The higher the resolution, the smaller the pixels, the more detail and clarity an image will have, the larger the file size!

What is essential?

Leads to scanning standards

- Scanning, among others, involves selection of standards & guidelines
 - a number exist, many similar, but no universal standards
 - some examples:
- [New Jersey Digital Highway](#) – “Digital Imaging Basics and Standards”
- [Oviatt Library Digital Collections](#) (California State U, Northridge)
“Digital Collections Scanning Standards”
- [California Digital Library](#) “Digital File Format Recommendations.”
Graphics, texts, audio, video
- [U.S. National Archives and Records Administration](#) – “Technical Guidelines for Digitizing Archival Materials for Electronic Access”

Examples of best practices

- Numerous best practices for various processes in digitalization have been published
 - some examples:
- [LYRASIS Best Practices and Publications](#) – includes: Digital Imaging, Digital Audio; Dublin Core Metadata
- [U of C, Berkeley Digital Collections](#) – Best practices for a number of process, including [imaging](#)
- And of course, there is a blog: [Digitization 101](#) – “The place for staying up-to-date on issues, topics, lessons learned and events surrounding the creation, management, marketing and preservation of digital assets.”

Scanning also involves selection of technology

- Many technological marvels for scanning on the market
- From sophisticated and specialized
 - for texts & images Optical Character Readers (OCRs)
 - doing Optical Character Recognition (OCR)
 - OK, “OCR” meaning two related but different things
- To simple – a digital camera – you can do it!
 - Digital Camera OCR – “Where can I use my digital camera to capture text?”

... and software for recognition

- Examples:
 - [Abby FineReader](#) – “helps **individuals** turn scans, PDFs and digital photographs into searchable and editable documents.”
 - [OmniPage](#) – more expensive & elaborate; “the fastest, most precise way to convert paper documents into editable digital content”
 - And a few [scanning tips](#) – “... the fundamentals of digital images, about the basics to help you get the most from your scanner. **How it works, for those that want to know.**”

Imaging equipment examples

- Flatbed scanners
 - many varieties & sizes; some specialized e.g. for large maps
- Sheet feed or document scanners
- Slide/film scanners
- Digital cameras

And an example of the whole process

- International Children's Digital Library – (ICDL)
revisited:
 - Digitization process
 - involves description in a nutshell of:
Choosing books; Scanning books; Getting books in ICDL;
and more detailed: Collection development policy;
Scanning procedures
 - topics covered in this lecture

Quality control

- OCR & associated software do not always make perfect scans
 - error rates vary, depending on materials scanned, hardware & software
 - constantly improving
 - but still human control of accuracy & quality needed - require human review for errors
 - sometimes require specialists
 - as in scanning Hebrew texts

Quality control ...

- Measuring technical qualities
 - of a digital master
 - in relation to the original from which it is reproduced
- Criteria:
 - for images: how well reproduced?
 - for texts: how accurately?
 - problem acute with old & non-latin scripts
 - for sounds: how faithfully?

Posting

- Means & ways of providing public access
 - could be also treated as separate from digitizing & part of all access provisions of a library
- Involves
 - policies – e.g. on presentation, access ...
 - site - development, maintenance ...
 - technology – servers, lines ...

Mass digitization of books

- A number of large efforts undertaken to digitize books – few examples here
 - some commercial, other non-profit
 - some involve partnership with libraries to various degree or with OCLC
 - some involve a lot of legal entanglements
 - because of that some do not provide a whole book
 - but have significant impact on numbers & variety of books accessed & even on book sales

Google books Library Project

- Agreement with a number of libraries
 - Harvard, Cornell, NYPL, Princeton, few National libraries
- Digitizing books & providing a book search engine
 - “... Our ultimate goal is to work with publishers and libraries to create a comprehensive, searchable, virtual card catalog of all books in all languages that helps users discover new books and publishers discover new readers.”
- Complex history (from Wikipedia)

- Generally, well received by librarians (they get the digitized books back), not so by publishers (they fear lost sales)
- Controversial from the start
 - strong reaction in Europe
 - Serious legal entanglements
 - Personal experience:
 - I looked for an out-of-print book *“Moving theory into practice : digital imaging for libraries and archives”*
 - found 10 libraries within 15 miles where I live that have it



Project Gutenberg

“the first producer of free ebooks.”

- revisited

- A volunteer effort to digitize, archive and distribute cultural works – out of copyright
 - “... over 50,000 free ebooks: choose among free epub books, free kindle books, download them or read them online.”
 - “Over 100,000 free ebooks are available through our [Partners, Affiliates and Resources](#).”
 - fun to explore “[Top 100 downloads](#)” or “[Pretty pictures](#)” – download graphs by day
 - Gutenberg is used a lot!



Million Book Project *also called Universal Library*

- Project started in 2001 by Carnegie Mellon University & partners in China, India & Egypt
- Now at Internet Archive
 - funding: \$3M NSF starting grant in 2001, over \$100M from other sources for over a decade
 - digitization done in India, China & Egypt with mirror sites in China & India
 - over 1.4M books scanned, but the project stopped & absorbed in Internet Archive
 - major digitization of books in a number of languages



Internet Archive texts

“... building a digital library of Internet sites and other cultural artifacts in digital form. Like a paper library, we provide free access to researchers, historians, scholars, the print disabled, and the general public.”

E.g. [Children's Library](#), [Genealogy](#)

- [Internet Archive](#) is a non-profit started by [Brewster Kahle](#) – an Internet entrepreneur
 - est. 1996 to maintain Web's historical record
 - includes archives for [Moving images](#); [Audio](#); [Software](#); [the Web](#); [TV News](#) etc.
 - each includes many sub-collections from others
 - partners with many institutions



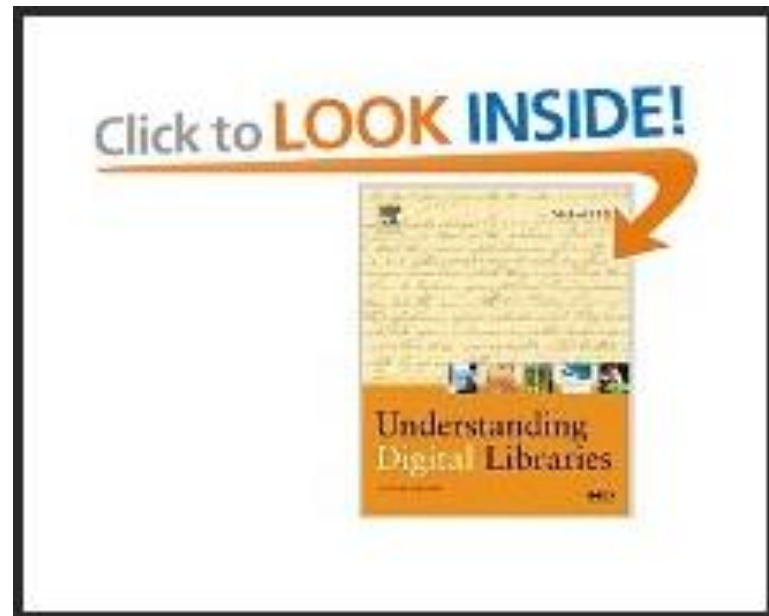
Amazon Books

- Amazon??? it is a bookstore!
 - well, Google is a search engine & went into digitization business
- New service: “Search inside a book”
 - search the full text of books & read a few pages
 - publishers agreed to have it included – voluntary
 - limits no. of pages you can read
- Profound changes in bookselling business
 - readers can sample
 - helps increase sales; sells the long tail end

Example

- Search inside the second edition of Michael Lesk's - a classic

Understanding Digital Libraries



Another example

- Search inside a provocative library management book



On conclusion

- Digitization is certainly a technical process
- But it is also a process involving
 - making many decisions
 - deciding on standards & then adhering to them
 - considering the law
 - very often, involving partners & cooperation
 - & learning from experiences & mistakes of others
- It seems to be a constant learning process
- *And it is changing things at a Gutenberg level*

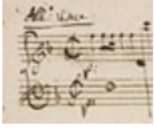




Magnificent example - technology at its best

British Library
[Turning the Pages](#)™
system

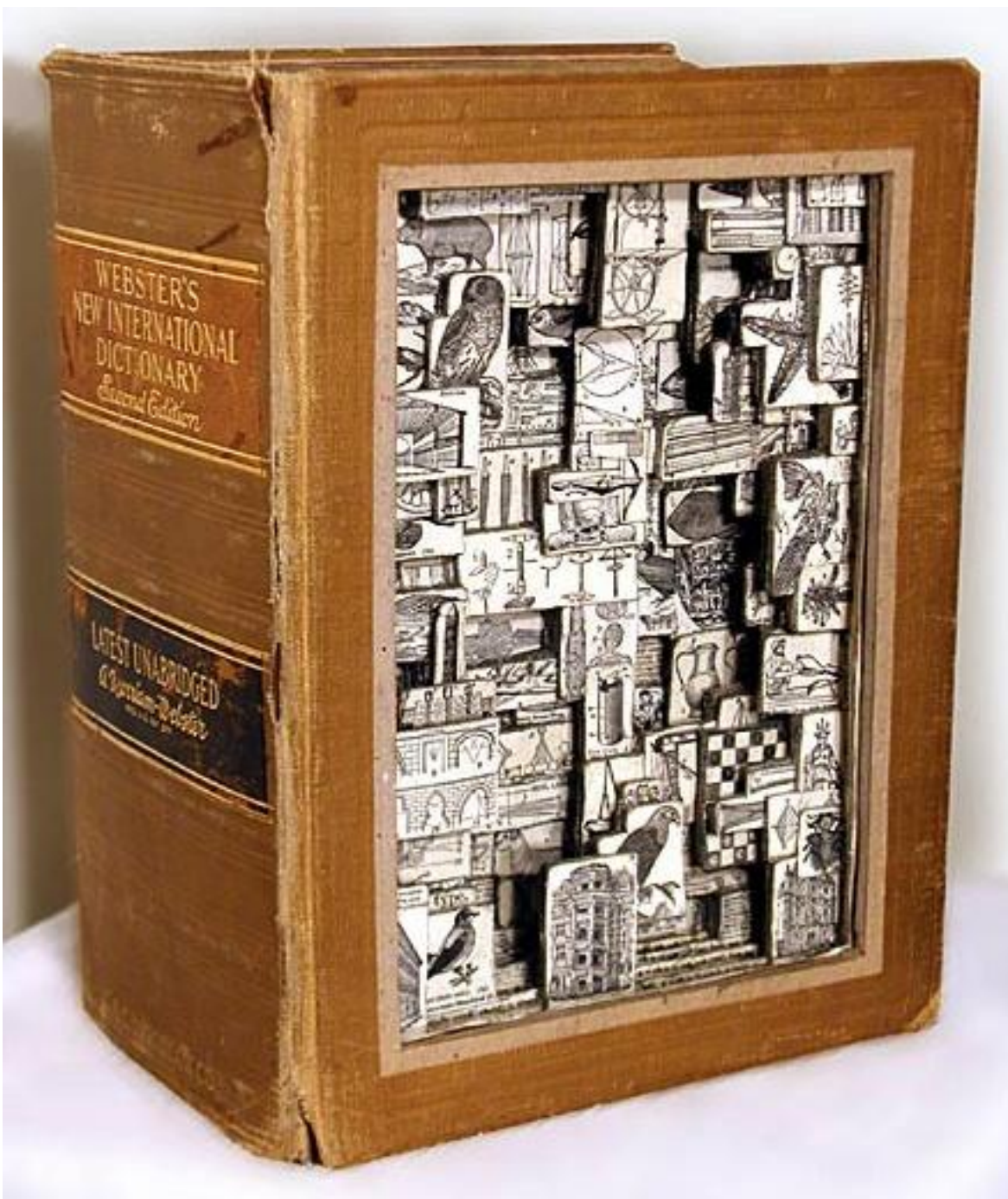
Viewing a number of
rare & old books by
turning pages,
listening to audio,
magnifying and
appreciating the
beauty and art.

next
slide



	<p>Mozart's musical diary</p> <p>The composer's own notes, from 1784 until his death</p>
	<p>My Ladye Nevells Booke</p> <p>Complete manuscript of William Byrd's keyboard music</p>
	<p>My Ladye Nevells Booke selections</p> <p>From the manuscript of William Byrd's keyboard music</p>
	<p>Outstanding 15th-century church book</p> <p>The wonderful, and weighty, Sherborne Missal</p>
	<p>Pinnacle of Anglo-Saxon Art</p> <p>The priceless Lindisfarne Gospels</p>





Brian Dettmer:
Sculptural Book Art

no, this is
not from the
British
Library

