AI-Assisted Digitization of Archives + Documentary Heritage Materials

1st Symposium Artificial Intelligence for Records and Archives

Lanzarote, Canary Islands, October 26-27, 2022

RA03: Model for an AI-Assisted Digitization Project, InterPARES AI

Presenter: Eng Sengsavang

Presentation Overview

This topic is from the study "Model for an AI-Assisted Digitization Project" part of interPARES AI.

Participating Researchers:

Eng Sengsavang, UNESCO Archives
Hrvoje Stancic, University of Zagreb
Adam Jansen, Hawaii State Archives
Marta Riess, IAEA
Shadreck Bayane, University of South Africa
Zeljko Trbusic, University of Zagreb (GAA)
Marina de Souza, University of British Columbia (GAA)
Kailey Fukushima, University of British Columbia (GAA)

RA03: Model for an AI-Assisted Digitization Project, InterPARES AI

DIGITIZATION OVERVIEW

Opportunities and Challenges of AI-Assisted Digitization of Cultural Heritage Materials

WHAT IS DIGITIZATION?

	Digitization	Digitalization
Definition	• The conversion of analogue signals conveying information (e.g., sound, image, printed text) to binary bits - OECD • The process of creating a digital image of a physical material (paper, photo, film or video, 3D object, etc.) by scanning or photographing; or by converting analogue signals to digital bits in the case of AV materials; aka imaging, conversion, capture.	• The adoption or increase in the use of digital or computer technology by an organisation, industry, country, etc.; refers also to the way digitisation is affecting economy and society - OECD • Gartner: “The use of digital technologies to change a business model and provide new revenue and value-producing opportunities; it is the process of moving to a digital business.”
Examples	• Scanning or imaging a photograph or text document to create a digital file. • Converting an analog sound recording to an MP3 file. • Recording a presentation or phone call, turning physical sound into a digital file.	• Shifting from paper forms to online forms. • Creating virtual online exhibitions. • Analyzing data collected by online ticketing, enabling user analysis for museums to obtain a set of visitor data and store them within a computer system. • Replacing handwritten signature workflow with an electronic signature workflow.

Sources: OECD, 2017, “Going Digital: Making the Transformation Work for Growth and Well-Being”, p. 9; Gupta, 2020, What is Digitization, Digitalization, and Digital Transformation? ; TruQC, n.d, Digitization vs. digitalization: Differences, definitions and examples

WHY DIGITIZE?

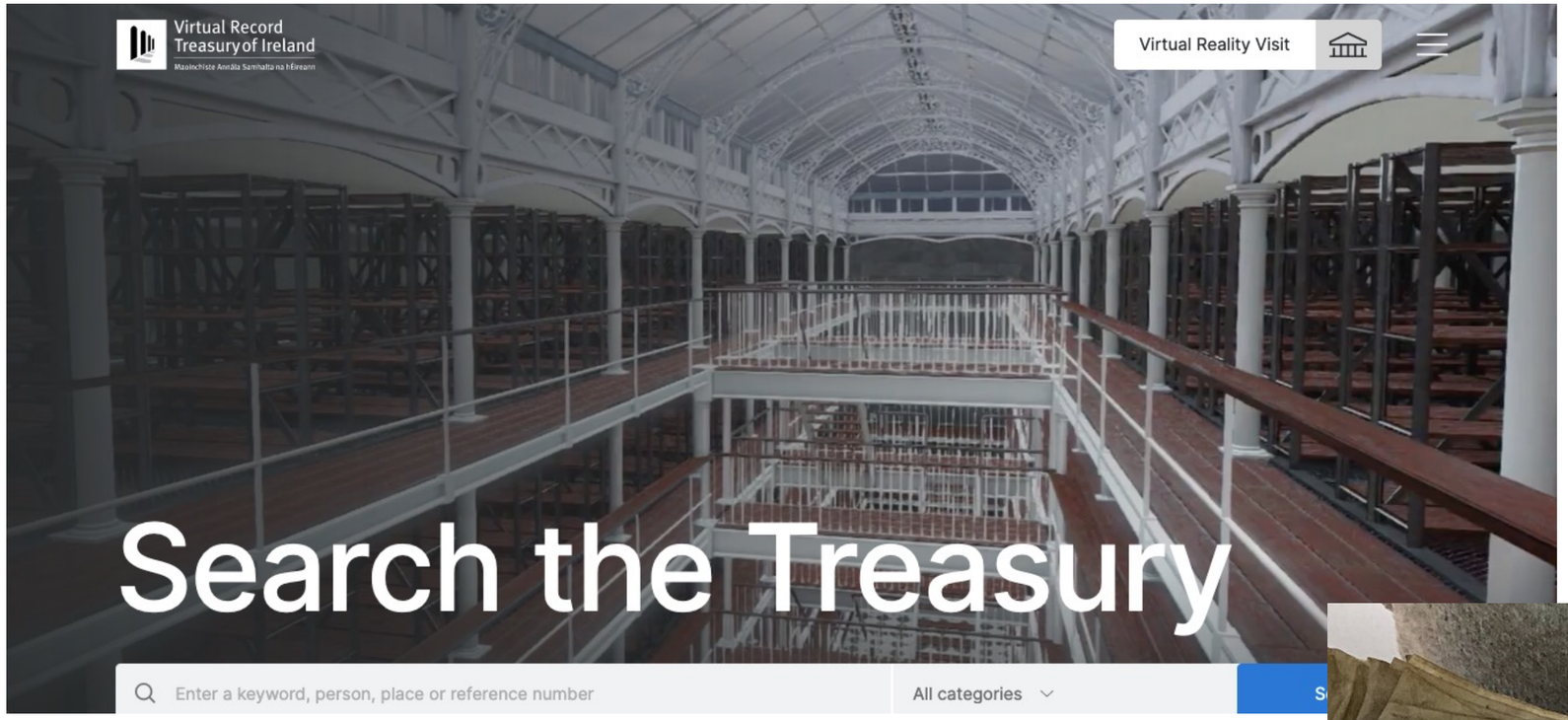
ACCESS

- Multiple users can access digitized collections from any networked location. Reduces geographic and socioeconomic barriers to access.
- Enables access to content on obsolete formats;
- Provides opportunities to improve our knowledge about the collections through metadata collection or creation during digitization. Collections can be discovered or re-discovered during digitization.
- Increased searchability and discoverability of collections published online and through OCR text searching.

PRESERVATION

- Reduces physical manipulation of originals.
- Creation of backup copies of unique materials: “Documentary heritage…is exposed to risks such as natural disasters, poor and inadequate storage, civil disturbance, looting, armed conflict, illicit trafficking and so on…Digitization…sustains the evidential value of records in case of emergency or disaster” (UNESCO, 2020).
- Improved conservation of originals through rehousing during digitization.
- Specialized imaging may reveal previously indiscernible details useful for conservation/analysis of originals.

WHY DIGITIZE?



Beyond 2022: Virtual Record Treasury of Ireland

(Public Records Office)

WHY DIGITIZE?

OUTREACH + COMMUNICATION

- Creation of opportunities for promoting and communicating collections online and ability to reach wider audiences;
- Possibility of linking with related sources across institutions;
- New opportunities for partnerships and projects with other organizations and enterprises;
- Potential to develop and increase the ‘brand’ value of the organization.

STRATEGIC + BUSINESS

- Greater ability to support the current work of the organization through integration in business systems and workflows;
- More effective searching and retrieval of information;
- Greater institutional awareness of collections;
- Development of staff skills.

INNOVATION

- Introduces new possibilities for technical innovation and experimentation;
- Contributes more diverse and historical datasets for AI, open data, etc.

DIGITIZATION PRINCIPLES

Digitization Guidelines, Library and Archives Canada: ”Defensible digitization processes” + “Authoritative digitized record”

- A digitized record must be useable, have integrity, be deemed authentic and reliable, support all business activities, and be able to withstand legal scrutiny;
- A digitized record must be generated under set policies and practices, be fully documented, and be maintained within an official corporate repository.

Source: https://library-archives.canada.ca/eng/services/government-canada/information-disposition/disposition-government-records/multi-institution-disposition-authorizations/pages/digitization-guidelines.aspx#introduction

DIGITIZATION PRINCIPLES

Technical Guidelines for Digitizing Cultural Heritage Materials, Federal Agencies Digital Guidelines Initiative, Code of Ethics:

- A set of professional practices to guide in creating faithful reproductions of historic records held in the public trust;
- The cultural heritage community has a responsibility to produce digital images that look like the original records…and are a “reasonable reproduction” without enhancement [and] without altering the fundamental nature of the historic record;
- Digitized material should document the appearance of the original at the time of capture, not what it may once have looked like if restored to its original condition;
- Certain alterations are acceptable (technical tools that increase access to information of faded text, burnt materials, mold damaged, image stitching, or other physical limitations) but should have metadata or other documentation that describes the image processing.

Source: https://www.digitizationguidelines.gov/guidelines/digitize-technical.html

MODEL FOR AN AI-ASSISTED DIGITIZATION PROJECT

Opportunities and Challenges of AI-Assisted Digitization of Cultural Heritage Materials

STUDY PURPOSE

The study explores the relationship between digitization of archives/documentary heritage and AI: how can they be mutually beneficial? Through the study, we intend to:

- Investigate how organizations may benefit from using AI methods and tools during digitization projects and processes;
- Analyze the potential opportunities and challenges of AI for digitization by providing a case study, including investigating potential benefits and biases of using specific AI models;
- Encourage digitization of archival/documentary heritage materials with a view to their long-term preservation, to encourage production of high-quality and diverse historical data sets to enrich machine learning.

STUDY FRAMEWORK

- In-scope: Textual and photographic datasets, applying off-the-shelf AI tools/methods.
- Out-of-scope: Audiovisual materials

Our study is based on the following assumptions:

- Digitization is a process that benefits society and produces valuable public goods.
- Archives and documentary heritage collections should be digitized with digital preservation/long-term sustainability of the resulting digital assets in mind.
- We are for critical approaches to machine learning and support the development of ethical AI and transparency and accountability.

STUDY FRAMEWORK continued...

Archival / Documentary Heritage Digitization: Digitization of records or collections of permanent value, with a view to their long-term preservation, especially digitization by (but not limited to) memory institutions / GLAM institutions (galleries, archives, libraries, and museums).

Memory institution is a collective term which includes, but is not limited to, archives, libraries, museums and other educational, cultural and research organizations.

- UNESCO, 2015, Implementation Guidelines on the Recommendation Concerning the Preservation Of, and Access To, Documentary Heritage Including In Digital Form prepared by Raymond Edmondson, viewed 5 April 2021, https://en.unesco.org/sites/default/files/2015_mow_recommendation_implementation_guidelines_en.pdf.

STUDY FRAMEWORK continued...

Documentary heritage comprises…single…or groups of documents of significant and enduring value to a community, a culture, a country or to humanity generally, and whose deterioration or loss would be a harmful impoverishment…The world's documentary heritage is of global importance and responsibility to all, and should be fully preserved and protected for all, with due respect to and recognition of cultural mores and practicalities…It provides the means for understanding social, political, collective as well as personal history. It can help to underpin good governance and sustainable development.

– *Recommendation Concerning the Preservation of, and Access to, Documentary Heritage Including in Digital Form, UNESCO, 2015*

PROJECT OUTPUTS

1. Survey on Digitization + AI
2. Model of a digitization project: Stages and activities
3. Test one or two AI tools/methods on a digitized archival data set and report on the outcomes

TIMELINE

Phase 1: 2022

Literature Review, Best Practices and Standards for Digitization, Model

Phase 2: 2023 to 2024

Survey and Identification of AI-Supported Digitization Processes

Phase 3: 2025 to 2026

Testing: plan, execute, outcome report

SURVEY ON DIGITIZATION + AI

- A way to share experiences and enable a better understanding of past and current digitization activities as a community; also to capture the knowledge we have gained when doing it;
- Snapshot of digitization activities of memory/GLAM/other institutions (galleries, libraries, archives and museums);
- Identify reasons for digitization, what digitized and how selected, greatest challenges and successes;
- Identify which institutions have used/are intending to use machine learning in digitization processes, what tools/models are being used, challenges and opportunities they experienced.

DIGITIZATION STATISTICS

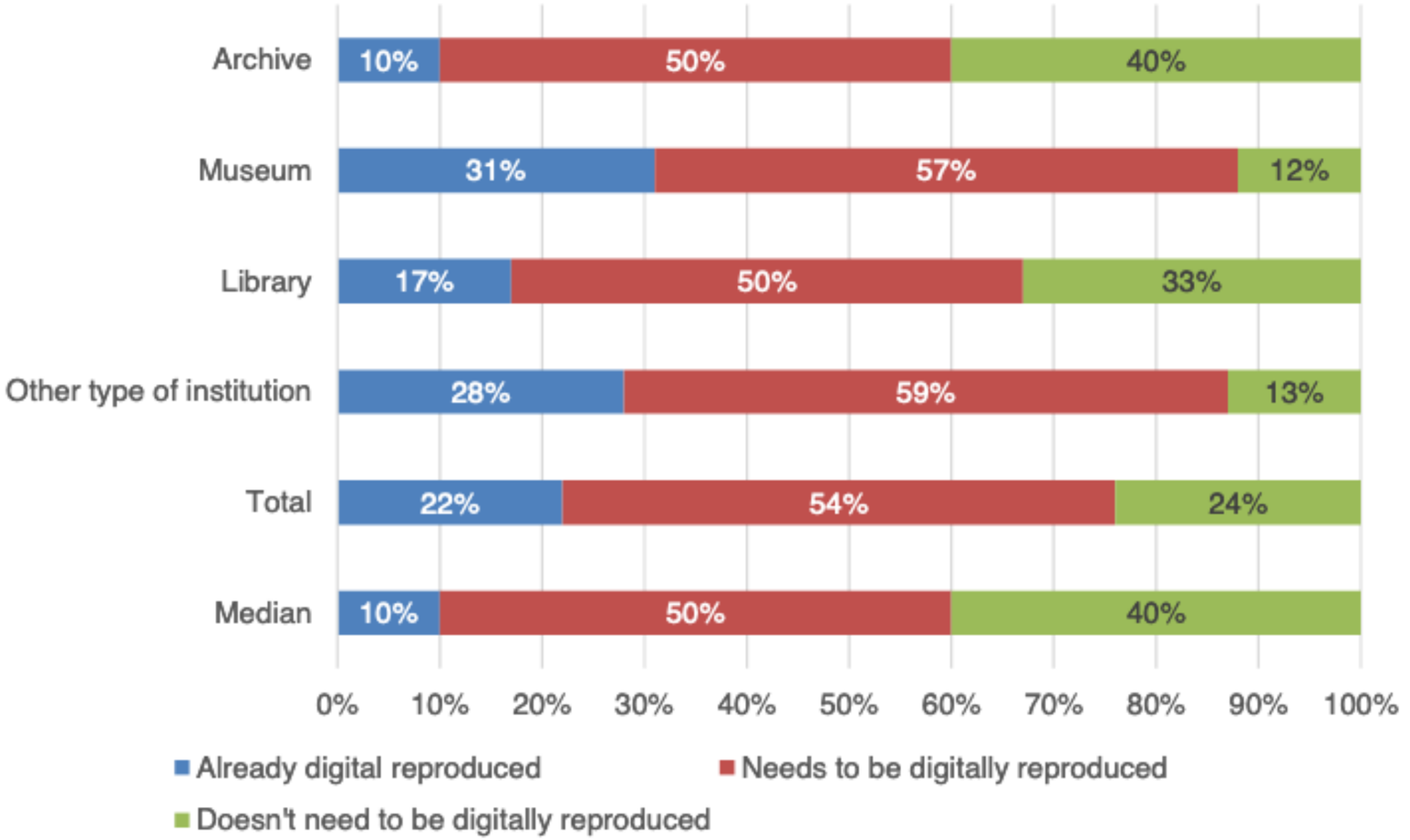
ENUMERATE Observatory for Europeana provides a reliable baseline of statistical data about digitisation, digital preservation, and online access to cultural heritage in Europe**.**

Findings of Survey Conducted in 2017:

- 82% of cultural heritage institutions (Archives, Libraries, Museums, Other) have a digital collection or are engaged in digitization activities (p. 15);
- 42% have a written digital strategy (p. 16);
- 59% have born-digital collections (p. 18);
- Percentage of institutions that have the following types of collections: Text-based - 89% analogue, 55% digital; Visual 2D - 89% analogue, 66% digital; Archival records - 74% analogue, 45% digital; Time-based - 67% analogue, 62% digital; 3D Man-made material - 63% analogue, 29% digital (p. 22-25);
- On average 3.3% of paid staff in all cultural heritage institutions are working full time on digitisation.

CULTURAL HERITAGE DIGITIZATION

Figure 3.7: Estimated percentage of analogue collection that has been digitally reproduced (n=757) or still needs to be reproduced (n=765)



Source: Europeana, Report on ENUMERATE Core Survey 4, 2017, p. 28

MODELLING DIGITIZATION

Opportunities and Challenges of AI-Assisted Digitization of Cultural Heritage Materials

DIGITIZATION PROJECT MODEL

- Concept: Identify main stages and activities in a digitization project and represent it in a visual model.
- The model is not prescriptive: there is no one-size-fits-all approach for every digitization project, however, they all share common activities and stages.
- Cross-reference the activities and stages with various digitization standards and best practices (e.g., FADGI, LAC, etc.).
- Digitization projects are complex and require a series of decisions. We may create other tools such as checklists and decision trees.

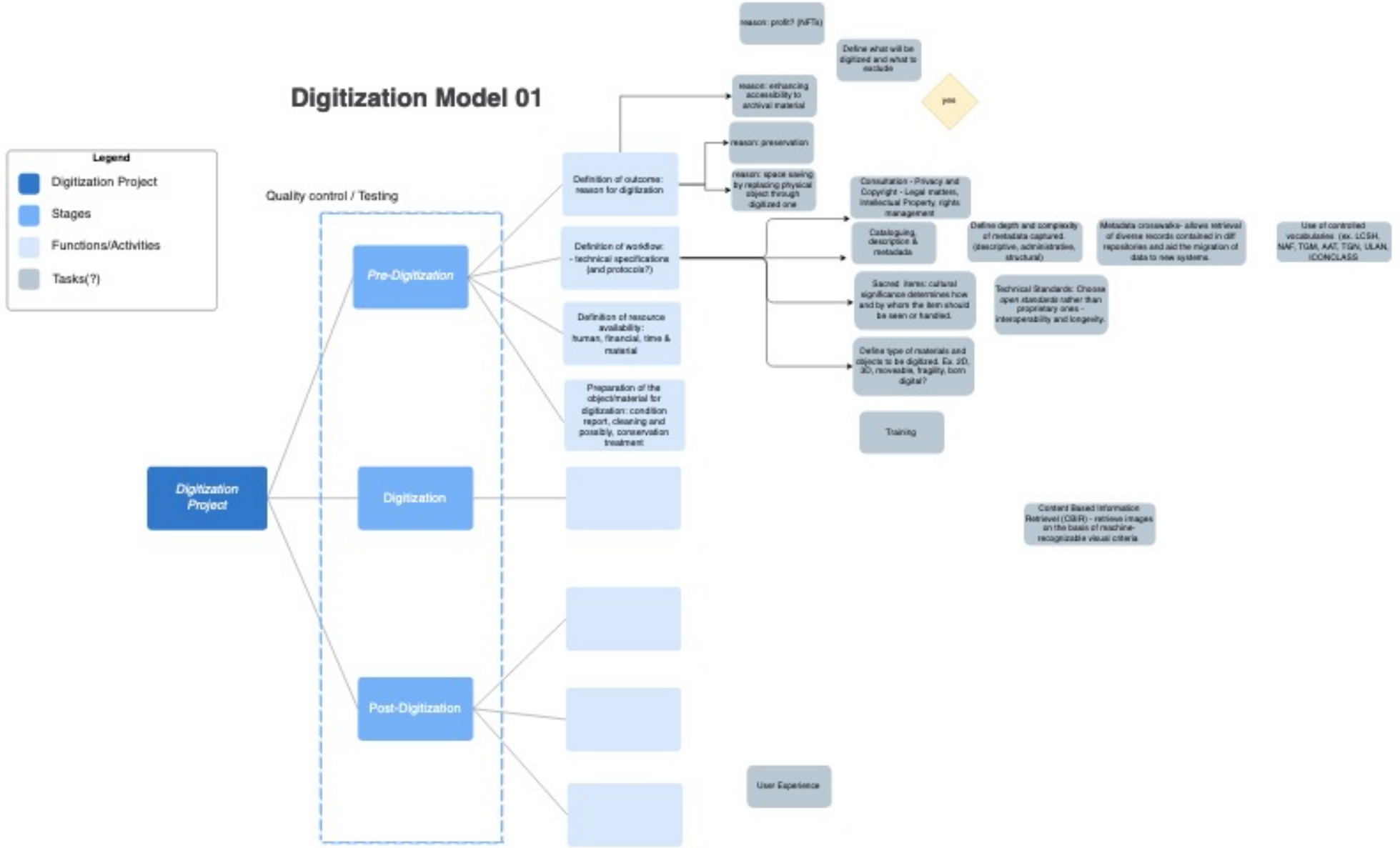
DIGITIZATION PROJECT ACTIVITIES + STAGES

Activity	Sub-activity	Pre-Digitization	Digitization	Post-Digitization	Project Management	Cited In	Decision Tree?
Physical collections selection		✓					
Physical collections preparation		✓	✓				
	Condition assessment	✓					
	Risks assessment	✓				02	
	Cleaning	✓	✓				
	Physical reconditioning and	✓	✓				
Creation of technical specifications		✓					
	File format standards	✓					Y
	File versions/resolutions	✓					Y
	OCR or Non-OCR	✓					Y
Establishment of file naming		✓					
Training		✓					
Metadata management		✓	✓	✓			
	Creation of metadata models (descriptive, technical, structural, administrative)	✓					
	Metadata gathering	✓	✓	✓			
	Metadata creation	✓	✓	✓			
	Cataloguing and description	✓	✓	✓			
Imaging			✓				
	In case of in-house digitization: check existing hardware / potential procurement of hardware	✓					
	Imaging		✓				
	OCR		✓			02	
Quality control							

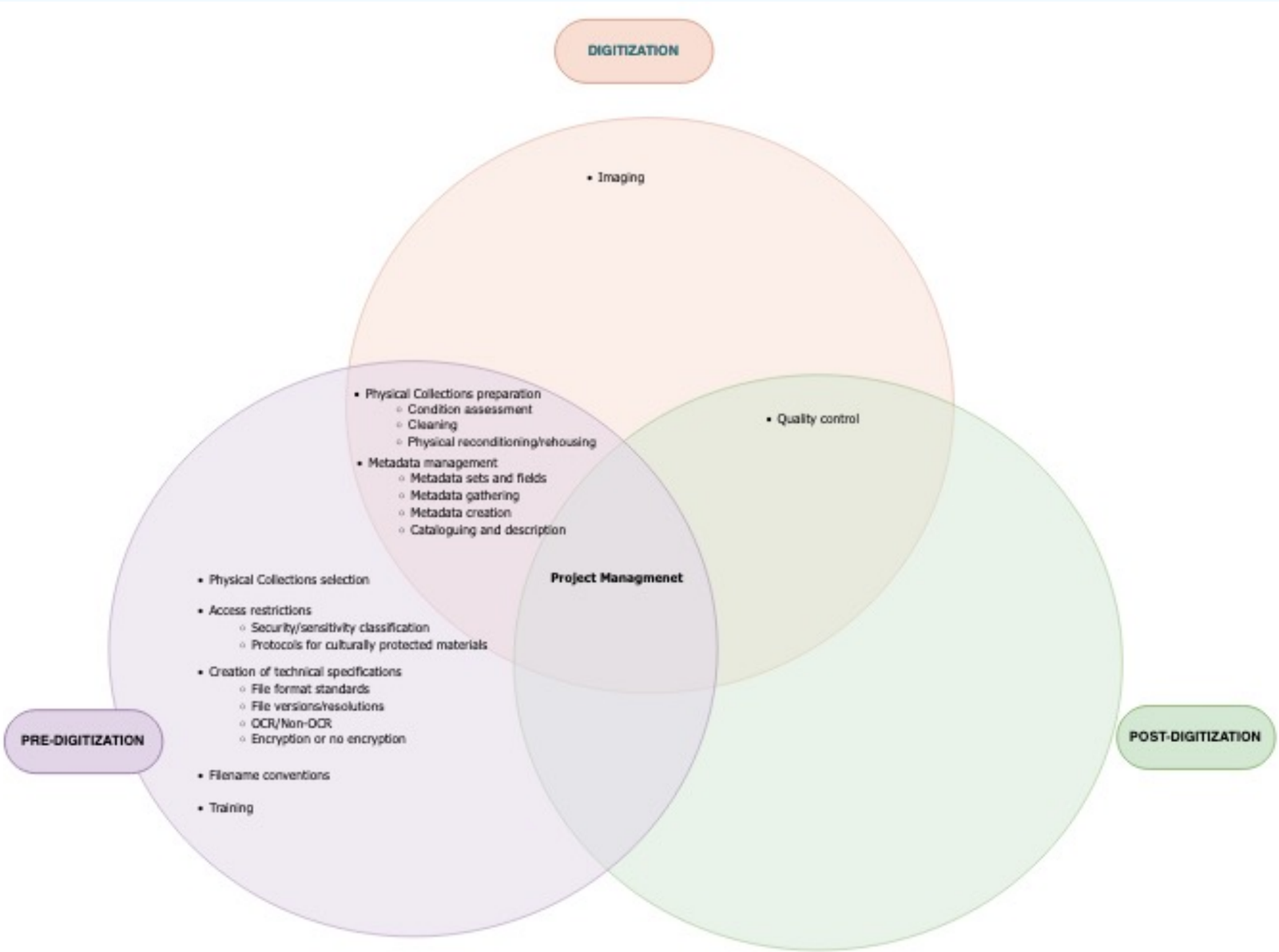
DIGITIZATION PROJECT ACTIVITIES + STAGES

Activity	Sub-activity	Pre-Digitization	Digitization	Post-Digitization	Project Management	Cited In	Decision Tree?
File transfer			✓	✓			
File validation							
Digital preservation				✓			
Communication and outreach				✓			
	Access and rights management			✓			
	Management/publication of			✓			
	Promotion of digitized collections			✓			
Copyright and licensing				✓			
Objective setting					✓		Y
	Determination of digitization project type (see J38)	✓			✓	02	
	Determination of originals	✓			✓	02	
Budgeting/Resource analysis					✓		Y
Project planning					✓		
	In-house or external digitization				✓		Y
Workflow creation					✓		
Vendor selection					✓		
Creation of digitization policy and					✓		
Documentation					✓	02	
Project evaluation					✓		

DIGITIZATION MODEL 01



DIGITIZATION MODEL 02



OPPORTUNITIES & CHALLENGES OF AI-ASSISTED DIGITIZATION

Opportunities and Challenges of AI-Assisted Digitization of Cultural Heritage Materials

AI-ASSISTED DIGITIZATION

ORIGINAL FORMAT	DIGITIZATION ACTIVITY	AI TOOL/METHOD	OPPORTUNITY	CHALLENGE
Photos - prints, negatives	Metadata creation - image captioning	Image classification, NER, Topic modeling	Potentially save time; apply batch descriptions	Accuracy, ensuring un-biased language
Glass lantern slides	Image correction - reconstruction	GAN – Generative Adversarial Network, Data augmentation?	Fill in gaps left by deteriorating/peeling slides	Accuracy, Ethical challenges

AI-ASSISTED DIGITIZATION

ORIGINAL FORMAT	DIGITIZATION ACTIVITY	AI TOOL/METHOD	OPPORTUNITY	CHALLENGE
Textual documents	Metadata generation for arrangement and description - Summarization and Translation	Machine translation, NER, Topic modeling	Assisted archival description; increase global access to digitized records	Accuracy
Textual documents	Tagging/indexing using controlled vocabularies	Text classification	Automate keyword indexing	Is it faster to have humans index or to teach a machine to do it accurately?

CONCLUDING REMARKS

- Digitization is and will continue to be a major activity for all types of organizations, and will continue to have an impact on our access to, and preservation of, archives and documentary heritage.
- Can AI tools and methods actually help save time and resources and improve processes and outcomes for organizations undertaking digitization? Especially off-the-shelf tools?
- Machine learning is time-consuming and seems to require a substantial investment in time and resources up front. But it could have huge and powerful benefits for archival institutions in the long term.

FEEDBACK

Please send us any suggestions for out-of-the-box AI tools, methods, models.
AI experts who wish to join our study are welcome!

For any questions or suggestions, please contact:

Eng Sengsavang
e.sengsavang@unesco.org

THANK YOU!

REFERENCES

- Besser, H., Hubbard, S., Lenert, D., Getty Publications Virtual Library, & Google Books - Getty Publication Virtual Library. (2003). Introduction to imaging (Revis ed.). Oxford University Press, Incorporated.
- Bevans, D. (2021, March) *Digitization and Automation: What Are They, and How Do They Relate?* https://www.mendix.com/blog/digitization-and-automation-what-are-they-and-how-do-they-relat
- *Digitization vs. digitalization: Differences, definitions and examples.* (2022). https://www.truqcapp.com/digitization-vs-digitalization-differences-definitions-and-examples/
- Gupta, M. (2020, March). *What is Digitization, Digitalization, and Digital Transformation?* ARC Advisory Group. https://www.arcweb.com/blog/what-digitization-digitalization-digital-transformation
- Imran, M. (2022, June)). 5 AI Tools For Image to Text Conversion. Folio 3. https://www.folio3.ai/blog/image-to-text-conversion-ai-tools/
- Inkcapture (2022) https://www.inkcapture.com/en/inkcapture-english/
- Nauta et al. (2017) Report on ENUMERATE Core Survey 4. Europeana. https://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/ENUMERATE/deliverables/DSI-2_Deliverable%20D4.4_Europeana_Report%20on%20ENUMERATE%20Core%20Survey%204.pdf
- Raimo, N., De Turi, I., Ricciardelli, A., & Vitolla, F. (2021). Digitalization in the cultural industry: Evidence from italian museums. International Journal of Entrepreneurial Behaviour & Research (https://doi.org/10.1108/IJEBR-01-2021-0082)
- UNESCO. Managing low-cost digitization projects in least developed countries and small island developing states: a manual (2021). https://unesdoc.unesco.org/ark:/48223/pf0000380165.locale=fr
- UNESCO Institute for Statistics, 2009 UNESCO Framework for Cultural Statistics http://uis.unesco.org/en/glossary-term/cultural-heritage?wbdisable=true

AI-ASSISTED DIGITIZATION: TOOLS AND PROCESSES

Some AI tools for image-to-text conversion	
Product	Description
inkCapture	Tool designed specifically for handwriting recognition. It will offer not only the extraction of text from a document, but also advanced search in documents, where it will not only search for an exact match, but also similar words. Specialized focus on the Czech language. Mobile app will be available soon
Ocr.best	Conversion platform that offers to transform scanned document images and photos into editable text formats
Ocr2edit.com	Platform for converting images to text with extensive language support. Allows you to upload images from various sources, including cloud and offline system storage.
Onlineocr.net	Online application that gives you multiple conversion options for optimized PDF format with optimized support. It even supports the basic conversions like PDF to Word and Excel to make them editable. Supports up to 46 languages, including Chinese and Korean.
Text-image.com	This online tool is enabled by an AI-powered algorithm that accurately reads and extracts text from an image. This website supports multiple image uploading formats. It has a mobile-friendly interface which means you can access it easily on a smartphone or tablet.

Sources: https://www.folio3.ai/blog/image-to-text-conversion-ai-tools/; https://www.inkcapture.com/en/inkcapture-english/

DIGITIZATION VS. AUTOMATION

	Digitization	Automation
Definition	• The conversion of an analogue signal conveying information (e.g., sound, image, printed text) to binary bits - OECD • Refers to creating a digital representation of physical objects or attributes. • Converting data, documents and processes from analog to digital.	• The use of technology to carry out repetitive tasks systematically. • A process is considered automated if it was once handled manually but is now executed without human intervention (or decreased human intervention). • A series of rules written by business subject matter experts to accomplish tasks without any human intervention.
Examples	• Scanning or imaging a photograph or text document to create a digital file. • Converting an analog sound recording to an MP3 file. • Recording a presentation or phone call, turning physical sound into a digital file.	• Automatically send users/clients reminders, discount codes (e.g., for museum tickets). • Generate automatic reports rather than manual reports (for example, through a centralized database). • Helps organizations save time, enables people to focus on other core or value-added work activities.

Bevans, 2021, "Digitization and Automation: What Are They, and How Do They Relate?"

SURVEY ON DIGITIZATION + AI

Demographics

- Disseminate survey globally
- Role and seniority of respondent
- Organization + size
- No. of staff in archives/records unit
- Location

Digitization Activities

- Have undertaken digitization activities?
- What , why, how digitized?
- On-premises or off-site?
- In-house or hired vendor?
- Digitization policy?
- What main activities did it involve? A
- Greatest challenges/successes?

Digitization + AI

- Have you used AI for digitization?
- If yes, for what processes/activities?
- Have you used AI elsewhere in your work?
- If yes, for what?
- What tools/methods/models used?
- What benefits?
- Challenges?

Questions

- Should survey be translated and what languages?
- What type of organizations are being targeted/what is scope - only archives, GLAM, any type??

CULTURAL HERITAGE?

Cultural heritage: “Cultural heritage includes artefacts, monuments, a group of buildings and sites, museums that have a diversity of values including symbolic, historic, artistic, aesthetic, ethnological or anthropological, scientific and social significance. It includes tangible heritage , intangible cultural heritage (ICH) embedded into cultural, and natural heritage artefacts, sites or monuments. The definition excludes ICH related to other cultural domains such as festivals, celebration etc.” (UNESCO, 2022)

Cultural heritage digitization: “Digitization of cultural heritage refers to the dynamic and evolving interdisciplinary domain that encompasses philosophical, social, cultural, economic and managerial aspects and consequences of management of cultural heritage in the technological environment.” (Manžuch et al, 2005).