2024 Annual ACA Conference
June 12, 2024

# Teachable AI for the Archival Profession

Richard Arias-Hernandez, PI

Kaila Fewster, GAA

# Land Acknowledgement

I am grateful to live, work and play on the traditional, ancestral and unceded lands of the hәńq̓әmińәm̓ (Halkomelem) and Sḵwx̱wú7mesh sníchim (Squamish) speaking xwmәθkwәy̓әm (Musqueam), Sḵwx̱wú7mesh Úxwumixw (Squamish) and sәlilw̓әtaʔɬ (Tsleil-Waututh) Nations, who have stewarded this land since time immemorial. I also want to acknowledge that, although virtual, this conference is taking place on Treaty 7 territory and within the Battle River Territory of the Métis Nation of Alberta, which also includes the traditional lands of the Tsuu T'ina, Siksika (Blackfoot), and Nakota Sioux (Stoney) peoples.

# Acknowledgments:

# Roadmap of Presentation

- Introduction
- Review of literature for AI in archival processing
- Study updates
- Preliminary findings and data
- Next steps

# Review of Literature

Reasons why archives professionals use AI for archival processing

Addressing the "digital sludge" (Rolan et al, 2019); Staying abreast of technological developments;

Processing email collections (Decker et al, 2022; Schneider et al, 2019)

Facets of arrangement and description that AI can facilitate

Processing overly large volumes of digital records; Enhancing descriptive metadata in poorly described records

(Cain, 2016); Identifying and redacting sensitive or personal content in records (Marciano et al, 2018); Linking

creators and subjects and disparate fonds (Anderson, 2021)

Challenges faced in implementation

Translating archival principles into practices; Acquiring knowledge of computational methods (coding languages,

AI tools, etc).

# Is more education/training needed?

- There are lots of training on machine learning/artificial intelligence (ML/AI), but not with a focus on archives needs and workflows or considerations of archivist and records managers as learners (Darby et.al., 2022; Cordell, 2020)

- Evidence-based evaluations of existing data science, machine learning, and AI training opportunities and identification of training gaps within and outside archives communities are needed to develop further training for archivists and records manager (Darby et.al., 2022; Padilla, 2019)

# Research Goal

What are *learning outcomes*, *curriculum structure*, *activities*, and *assessment* that can be designed for archival students and professionals to be able to "leverage"(and possibly "design") AI/ML to support the ongoing processing, availability, and accessibility of trustworthy public records?

# Research Goal

What are *learning outcomes*, *curriculum structure*, *activities*, and *assessment* that can be designed for archival students and professionals to be able to "leverage"(and possibly "design") AI/ML to support the ongoing processing, availability, and accessibility of trustworthy public records?
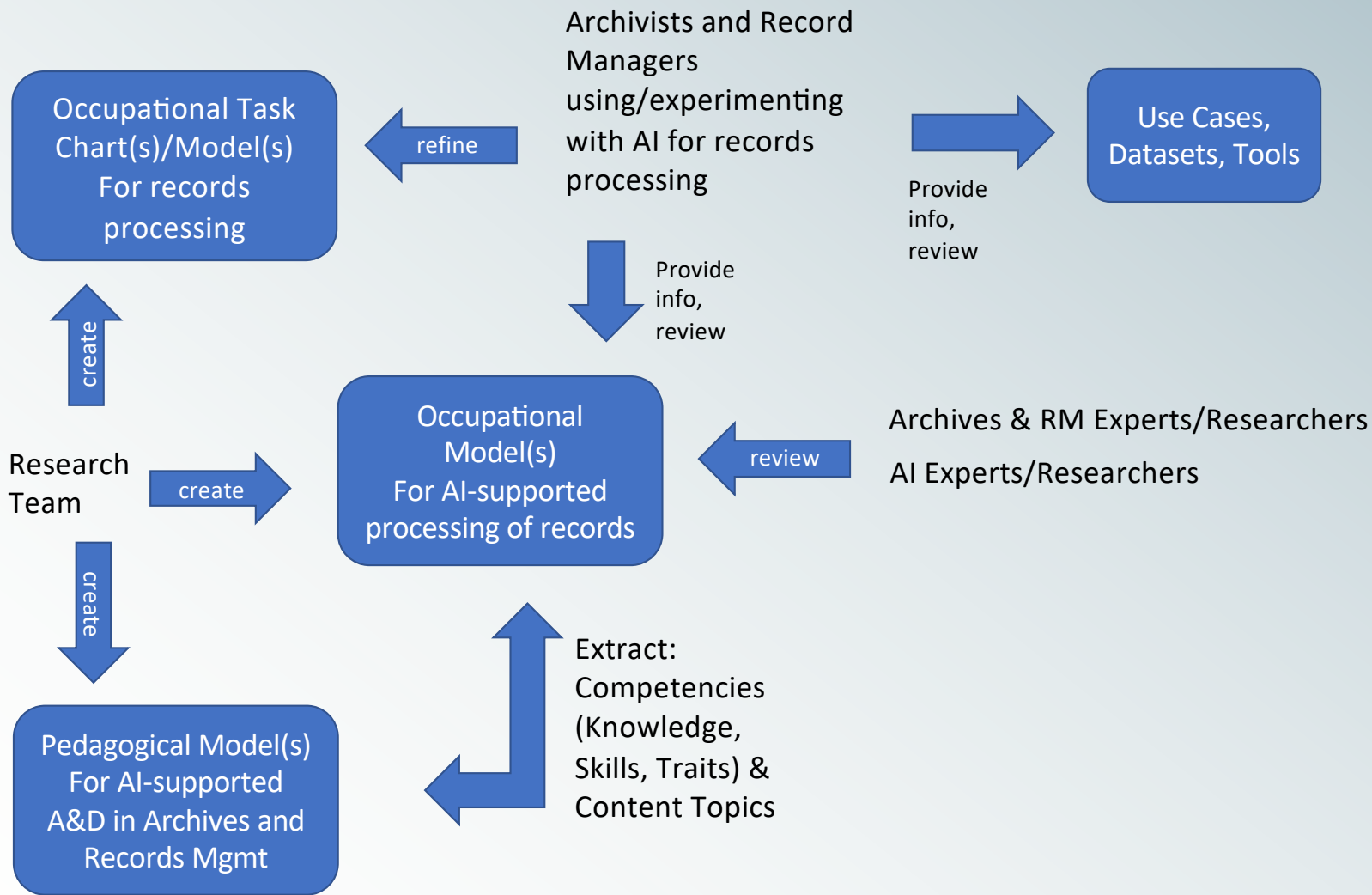
# Literature Review for case studies

In early 2022 we conducted a literature review on 2012-2022 publications with the following goals:

1. Identify use cases of archivists and records professionals using (or testing) AI in their workflows.
2. Identify archival and/or records management practitioners and scholars involved in these cases to involve them in semi-structured interviews.
3. Identify relevant information and didactic materials that could be incorporated into a curriculum on AI for the processing and management of records.

#1: using NLP, ENR, and machine learning to enhance description of multiple personal fonds and reveal connections between them. University Archives.

#2: using NLP, NER, and ML techniques such as topic modeling to appraise and process emails and other born-digital records. University Archives and Special Collections.

#3: using Computer vision and NLP to identify relationships between visual records via identification of subjects depicted in them. College Archives.

#4: using ePADD (NER, NLP of PII, Topic Modeling) to process email series in a personal archives to make it available to researchers. National Library.

**Case Studies (2017-2021)**

#5: using off-the-shelf ML software to classify a corpus of unstructured data against a retention and disposal authority at a State archives

#6: Prototype ML software to open access to email semi-active records from a software company through a ML-enhanced search engine using an attention model (BERT)/deep learning.

**Case Studies (2017-2021)**

# Development of competencies

- "Digital competencies at work are a set of basic knowledge, skills, abilities, and other characteristics that enable people at work to efficiently and successfully accomplish their job tasks regarding digital media at work." (Oberländer et.al., 2020)
- 10 semi-structured interviews completed between January and August 2023
- Participants identified via literature review and recruited via direct email
- Primary goal of interviews was to identify key competencies

# Learners should know about:

| | |
|---|---|
| Archival principles/practices | e.g., archival theory & functions, legal aspects, etc. |
| Basic algorithmic thinking | e.g., computational thinking, using algorithms for problem solving and pattern recognition, etc. |
| Basic data analysis | e.g., descriptive statistics, data operations, data exploration, etc. |
| AI/ML fundamentals & techniques relevant to archives | e.g., AI/ML types of models & pipeline, most commonly applied ML models to archival processing: NER, NLP, Topic Modelling, etc. |
| Human-Computer/ Human-Information Interaction | e.g., basic principles of user experience/user interaction with AI/ML apps, tools, data, and interactions with digital records, etc. |
| Critical data and critical AI/ML for archives and RM | e.g., Indigenous data sovereignty, data governance, data biases in AI/ML, data provenance/paradata, social implication and ethics, etc. |

# Learners should acquire:

| | |
|---|---|
| Basic programming skills | e.g., Scripting, basic coding, tweaking code, using ML libraries, etc. |
| Data-related skills | e.g., collecting/selecting data, datafying digitized records, using data transformation tools, etc. |
| AI/ML skills | e.g., labelling data, training/refining models, selecting and evaluating ML models, etc. |
| Tool testing/adaptation to workflow skills | e.g., experimenting/tweaking will apps and tools, adapting tools (off-the-shelf or not) to archival workflows, etc. |
| Management/Collaboration skills | e.g., engaging, teaming-up, and collaborating effectively with SMEs, technical experts, and data stakeholders, project management, etc. |

# Learners should develop attitudes around:

| |
|---|
| Intellectual curiosity towards AI/ML for archives and RM |
| Self-regulated/independent learning of digital tools |
| Critical thinker |
| Team player |

# What is next …

- Development of open and free educational modules for the curriculum (i.e. module learning outcomes, learning activities, and assessment of learning)
- Testing and refinement of learning modules in the classroom

Thank you!

# References

Anderson, B. G. (2021). On Constructing a Scientific Archives Network Exploring Computational Approaches to the Cybernetics Thought Collective. *Archivaria*, *91*, 104–147.

Cain, J. O. (2016). Using Topic Modeling to Enhance Access to Library Digital Collections. *Journal of Web Librarianship*, *10*(3), 210–225.

Cordell, R. (2020). Machine Learning + Libraries: A Report on the State of the Field. Library of Congress: Washington, DC.

Darby, A. et.al. (2022). AI training resources for GLAM: a snapshot. arXiv:2205.04738. DOI: 10.48550/arXiv.2205.04738

Decker, S., Kirsch, D. A., Kuppili Venkata, S., & Nix, A. (2022). Finding light in dark archives: Using AI to connect context and content in email. *AI & SOCIETY*, *37*(3), 859–872.

Marciano, R., Underwood, W., Hanaee, M., Mullane, C., Singh, A., & Tethong, Z. (2018). Automating the Detection of Personally Identifiable Information (PII) in Japanese-American WWII Incarceration Camp Records. *2018 IEEE International Conference on Big Data (Big Data)*, 2725–2732.

Padilla, T. (2019). Responsible Operations:  Data Science, Machine Learning, and AI in Libraries. OCLC: Dublin, OH.

Rolan, G., Humphries, G., Jeffrey, L., Samaras, E., Antsoupova, T., & Stuart, K. (2019). More human than human? Artificial intelligence in the archive. *Archives and Manuscripts*, *47*(2), 179–203.

Schneider, J., Adams, C., DeBauche, S., Echols, R., McKean, C., Moran, J., & Waugh, D. (2019). Appraising, processing, and providing access to email in contemporary literary archives. *Archives and Manuscripts*, *47*(3), 305–326.