Title: AI and ML for Catalogue Conversion

Project summary

This project aims to provide proof-of-concept for using artificial intelligence and machine learning to support the transition from traditional ISAD(G) compliant catalogues to Ric-compliant catalogues and linking archival metadata to the semantic web via Linked Open Data.

Project description

The potential for artificial intelligence (AI) exploitation of digital and digitised archives via machine learning (ML) has long been trumpeted and, in particular in relation to description of image archives, automated transcription of archival texts and identifying names of people, places and organisations via named entity recognition (NER), has already been demonstrated to greater or lesser degrees of success. Risks and professional concerns have also been identified, along with potential ways by which these could be mitigated (Bunn, 2020; Jaillant and Rees, 2023; Lee, 2023). The value of linked data (LD) for archival projects has also been explored. To date, however, connections have not been drawn between the two areas of research, although this was identified as an important area for future research (Schreur 2020, Hawkins 2022 and Jaillant 2022a and 2022b). These issues are particularly pertinent at a time when the new international standard for archival cataloguing, Records in Contexts – RiC (https://www.ica.org/ica-network/expert-groups/egad/records-in-contexts-ric/) is in the process of being introduced. RiC was designed to be inter-operable with LD but the resource implications of converting existing ISAD(G)-compliant catalogues to RiC and incorporating LD have neither been researched nor addressed. There is therefore a risk that the anticipated benefits of both RiC and LD will not materialise.

As Rolan et al (2019) observe, there is 'A lack of compelling case studies - maybe this is the hype-cycle effect, but while commentary abounds, there are not many real-world examples within the academic or professional literature'. All the examples of the use of AI in recordkeeping analysed by Rolan in 2019 related to retention scheduling and appraisal; records management has also been the focus to date of the LUSTRE network (https://lustre-network.net/outputs/) and although catalogue data has been mined by NER projects, there has been less interest in using the results to enhance cataloguing. The proposed project aims to fill this gap.

A key difference between RiC and traditional indexing is that the relationship between the entities is characterised – for example an agent (e.g. a person) whose name might previously have been indexed (i.e. used as an entry point into the resource), through RiC can be linked to a resource (e.g. a letter) by being its author, its recipient, its subject, its collector, its owner or its donor etc. Whilst NER can be used to find entities, it cannot be used to characterise their relationships. This project therefore aims to identify the potential for AI in identifying different types of relationships expressed within existing catalogue data or discoverable via AI from other resources.

The aim of this project would be to demonstrate proof of concept for the use of AI and ML to support the conversion of existing catalogue data from ISAD(G) to RiC via the incorporation of linked open data (LOD) resources, with a particular focus on identifying materials of potential relevance to the subject of enslavement. This is any area of huge current interest, both for academic and community researchers and a topic for which legacy finding aids are often inadequate (see e.g. Buncombe and Prest 2021; Smallwood 2016; Thomas 2013). If successful, the tools thus created could be developed for professional application within the archive sector and to support research projects using archival data (e.g. research using prosopographical and topographical methods).

The project would work with existing catalogue data supplied by Liverpool Record Office. This currently exists in three formats:

- Digital data within an Axiell CALM database (structured in accordance with ISAD(G), at various levels of detail, from fonds-level descriptions to item-level calendars)
- Digital data in word-processed format (containing ISAD(G) compatible information and visually structured but without machine-readable metadata connecting content with ISAD(G) components)
- Analogue data in typescript format (containing ISAD(G) compatible information and visually structured but without metadata connecting content with ISAD(G) components).

The project would create a dataset consisting of catalogue data from these sources and interrogate it both using human intelligence and AI.

The project is a collaboration between the Liverpool University Centre for Archive Studies (primarily co-directors Dr Alexandrina Buchanan and Dr Victoria Stobo) and the Liverpool University Digital Innovation Facility, which houses the UKRI National Centre for Digital Heritage. LUCAS has extensive experience in archival cataloguing and metadata, whilst Victoria Stobo has specific expertise in NER. The DIF has extensive experience in AI, including AI in relation to heritage projects. This is an established partnership: Alex Buchanan is already working with the DIF on other projects involving AI.

Project benefits

The project would have four main (1-4) and two supplementary (5-6) benefits:

(1) It would provide initial proof of concept for AI input to catalogue conversion. If successful, this would support the profession in moving from ISAD(G) to RiCs without significant additional staff resources. It would be our ambition for any algorithms developed to be made open source, or incorporated into existing software packages. Having undertaken this project, we would be in a good position to apply for further funding to develop this work (e.g. via Innovate UK).

(2) It would provide initial proof of concept for using AI to connect archival data to LOD. This has the potential to save work (individual repositories would not have to construct their own authority records) and would improve findability of resources. This would support enhanced collaboration between archival institutions holding related materials and cross-sectoral working between the GLAM professions.

(3) The tools developed would support research projects in providing tools for identifying and using the wealth of archival catalogue data for research purposes, particularly prosopographical and topographical research. If RiCs compatibility were built into

future research projects, this would enable academic research to have more real-world impact and contribute to benefit (1) above.

(4) The project and its findings would contribute to professional education. The project and its findings would inform teaching on the Liverpool University Masters in Archives and Records Management. If successful, the tools would form part of a workshop style teaching session for MARM students and would be disseminated to the profession via the Liverpool University Centre for Archive Studies.

(5) The outputs of the project would support Liverpool Record Office in identifying and providing better access to resources relevant to the involvement of the city in enslavement, supporting both academic and community researchers. Having undertaken this project, we would be in a strong position to apply for further funding to develop this work, making activity in this area more sustainable.

(6) The outputs of the project would support Liverpool University's Centre for the Study of International Slavery, its partnership with National Museums' Liverpool International Slavery Museum and their work with communities locally and internationally affected by the impacts of enslavement.

Operational Plan

The project would have the following work packages and outputs:

WP1: Data ingest and mark-up (8 days of student intern)
Outputs: (1) Dataset for AI and ML
(2) human-generated list of names (of people, places and organisations)
(3) human-generated list of dates
(4) human-generated list of identifiable relationships (initially to train the AI, then to be use to test it for completeness and accuracy).

WP2: Data mining via NER and date recognition (2 weeks technician; 2 days of student intern). This would use Stanford Named Entity Recognizer (potentially with Batchner to convert to CSV format); SpaCy; Natural Language Toolkit or similar tools.
Output: (5) a list of names (of people, places and organisations)
(6) list of dates.
These will be compared with Outputs 1 and 2 to test for completeness and accuracy.

WP3: AI data analysis (2 weeks technician; 2 days of student intern).
This would use Wmatrix; CQPweb or similar tools for existing corpus query and SERP AI; Infranodus; Microsoft Azure; Label Studio (Open source) or similar tools for relationship analysis.
Output (7): identification of relationships within the catalogue data. This will be compared with Output 4 for completeness and accuracy.

WP4: Links with LOD. (2 weeks technician; 3 days of student intern)
We will test the congruence between names identified via WP1-2 and the LOD resource 'Enslaved: Peoples of the Historical Slave Trade' (https://enslaved.org) to identify whether/how much additional work would be needed to enable the names to be associated with existing LOD.

Output (8) Interim report on results to be published on the LUCAS website. An academic article would not form part of the proposed project as it could not be completed within the timeframe but would be an essential future output. We would aim for this to be published in *Archives and Records*.

Risk assessment: Should any of the personnel initially involved with the project become unavailable, there are additional interns available via LUCAS and staff at the Digital Innovation Facility who could take over. Either Alex Buchanan or Victoria Stobo would be able to provide expert input/oversight should the other be unavailable and either/both would be able to write up the proposed article. Should the AI prove unable to perform the tasks requested to the standard required, this would in itself be a useful finding and the possible reasons for this failure would be interrogated and reported.

Budget

Student intern (15 days, MARM student, grade 5 spine point 20)	£2,500
Alex Buchanan, supervision of intern (1 day, spine point 54)	£500
Technician (1.5 months, 1.0 FTE, spine point 25)	£5,250
Supervision of technician by DIF (spine point 33, 6 days)	£1,500
Victoria Stobo, specialist archive input (1 day, spine point 46)	£250

Total: £10,000

References

Buncombe, M., and Prest, J. (2021). Making ‘slave ownership’ visible in the archival catalogue: findings from a pilot project. *Archives and Records*, 42(3), 228-247.

Bunn, J. (2020) “Working in contexts for which transparency is important: A recordkeeping view of explainable artificial intelligence (XAI)”, *Records Management Journal* 30, no. 2 (2020): 143-153.

Hawkins, Ashleigh (2022) "Archives, linked data and the digital humanities: increasing access to digitised and born-digital archives via the semantic web", *Archival Science* 22.3 (2022): 319-344.

Jaillant, Lise (2019) “After the digital revolution: working with emails and born-digital records in literary and publishers’ archives”, *Archives and Manuscripts*, 47(3): 285–304.

Jaillant, Lise (2022a) “How can we make born-digital and digitised archives more accessible? Identifying obstacles and solutions”, *Archival Science*. https://doi.org/10.1007/s10502-022-09390-7

Jaillant, Lise (ed.) (2022b) *Archives, Access and Artificial Intelligence: Working with Born-Digital and Digitized Archival Collections*. Bielefeld, Germany: Transcript. https://doi.org/10.14361/9783839455845.

Jaillant, Lise, and Arran Rees (2023) "Applying AI to digital archives: trust, collaboration and shared professional ethics." Digital Scholarship in the Humanities 38.2: 571-585.

Lee, Benjamin Charles Germain (2023) "The “Collections as ML Data” checklist for machine learning and cultural heritage." *Journal of the Association for Information Science and Technology*.

Rolan, G. et al. (2019) “More human than human? Artificial Intelligence in the archive”, *Archives and Manuscripts* 47, no. 2: 179-203.

Schreur, Philip E. (2020) "The use of linked data and artificial intelligence as key elements in the transformation of technical services." *Cataloging & Classification Quarterly* 58.5: 473-485.

Smallwood, Stephanie E. (2016) “The Politics of the Archive and History’s Accountability to the Enslaved.” *History of the Present* 6, no. 2: 117–32.

Thomas, Deborah A. (2013) "Caribbean studies, archive building, and the problem of violence." *Small Axe: A Caribbean Journal of Criticism* 17, no. 2: 27-42.