Unlocking our Digital Past workshop 2

16 September 2021, 14:00-17:30

This online workshop is the second of two workshops organised as part of the "Unlocking our Digital Past" project. The project aims to bring together government professionals, academics and archive professionals to consider how AI can help improve the preservation, access to and usability of archives. In this workshop, we aim to focus on the potential role of AI in review and access tasks.

More information on the project can be found at www.unlockingourdigitalpast.com.

Feel free to tweet throughout the workshop. We will be using the hashtag #OurDigitalPast

14:00 - 14:15	Introduction from the project team
	Session 1 Selection, review and appraisal of digital archival material
14:15 - 14:30	Christopher (Cal) Lee, Professor at the School of Information and Library Science, University of North Carolina *Incorporating Natural Language Processing and Machine Leaning into Selection and Appraisal Workflows*
14:30 - 14:45	Richard Marciano, Professor at the College of Information Studies, University of Maryland *Automating the Review of Personally Identifiable Information (PII) in Japanese-American WWII Incarceration Camp Records*
14:45 - 15:00	Leslie Johnston, Director of Digital Preservation, US National Archives and Records Administration *Achieving the Promise of AI with Modest Approaches to Machine Learning*
15:00 - 15:20	Questions and discussion facilitated by Clifford Lynch, Executive Director, Coalition for Networked Information

15:20 - 15:45	**Case study session** Andrew Dixon, Managing Director, SVGC *Digital Sensitivity Review at the Foreign, Commonwealth and Development Office*
15:45 - 16:00	Break
	Session 2 Access to digital archival collections
16:00 - 16:15	Leontien Talboom, PhD candidate, The National Archives & University College London *Accessing the Intangible*
16:15 - 16:30	John Sheridan, Digital Director, The National Archives *Archives in the midst of the AI revolution*
16:30 - 16:45	Christopher Day, Head of Modern Domestic Records, The National Archives *'Computing Cholera': topic modelling catalogue entries for the correspondence of the General Board of Health (1848-1871)*
16:45 - 17:05	Questions and discussion facilitated by James Lappin, Digital Era Approach to KIRM, The Cabinet Office.
17:05 - 17:15	**Moving forward** Nicola Welch, Service Owner – Access to Digital Records, The National Archives *Towards a future access service for digital records at The National Archives*
17:15 - 17:30	Closing remarks

Speaker and presentation details

Christopher (Cal) Lee

Incorporating Natural Language Processing and Machine Leaning into Selection and Appraisal Workflows

Abstract: One can encounter digital materials at numerous levels of representation, ranging from raw bitstreams read from physical media to high-level aggregations of digital objects. Selection and appraisal of born-digital materials are iterative processes, with each representation level revealing information that can inform professional judgements. This presentation will discuss several open-source software projects that have incorporated natural language processing and machine learning into selection and appraisal workflows.

Biography: Christopher (Cal) Lee is Professor at the School of Information and Library Science at the University of North Carolina, Chapel Hill. He teaches courses and workshops in archives and records management; understanding information technology for managing digital collections; and adapting digital forensics methods to support curation of born-digital materials. His primary research focus is the long-term curation of digital collections. Cal has served as Principal Investigator and Co-Principal Investigator of numerous digital curation research and education projects. He is a Fellow of the Society of American Archivists and recently completed his term as editor of American Archivist.

Richard Marciano

Automating the Review of Personally Identifiable Information (PII) in Japanese-American WWII Incarceration Camp Records

Abstract: A description of computational treatments of archival collections involving World War II Japanese-American Incarceration Camps. With a focus on automating the review of personally identifiable information or PII and how this approach is related to a taxonomy of Computational Thinking practices.

Biography: Dr. Richard Marciano is a professor at the University of Maryland iSchool and the recipient of the 2017 Emmett Leahy Award for "outstanding accomplishments that have had a major impact on the records and information management profession." He is the founder in 2020, of the Advanced Information Collaboratory, an international research network which explores the opportunities and challenges of disruptive technologies for archives and records management (including CAS – Computational Archival Science, AI, ML, Digital Curation, etc.), while promoting ethical information access and use, with a recent targeted initiative called the Future of Archives and Records Management (FARM).

Leslie Johnston

Achieving the Promise of AI with Modest Approaches to Machine Learning

Abstract: The U.S. National Archives and Records Administration (NARA) has recently been looking at how Artificial Intelligence and Machine Learning might be applied to both processing and public access, consulting with colleagues from several academic and cultural heritage organizations on the state of the art in applied AI; NARA has several medium- and long-term goals. based in part on the need to work more efficiently and improve access at a very large scale. But there are actually short-term goals that NARA or any organization can potentially meet with machine learning tools: every project doesn't need to be about full-fledged AI. Implementation of machine learning as a subset of AI is much more feasible, and organizations may already be using such tools without realizing it: Optical Character Recognition for printed text or handwriting, chatbots, Alexa skills, pattern recognition or pattern matching (used to find PII or other relevant patterns in records), named entity recognition (personal or corporate names, geographic locations, and dates), or computer vision for still images or video such as the controversial use of facial recognition. This talk is focused on the lower barrier to entry types of tools that cultural organizations can use to start their use of AI for processing and access.

Biography: Leslie Johnston is the Director of Digital Preservation for the National Archives and Records Administration (NARA), responsible for developing and executing their digital preservation strategy. Ms. Johnston has worked in the cultural heritage, higher education, and federal communities; her expertise includes system design and implementation, setting and applying content and metadata standards, and the preservation of born-digital and digitized collections.. She has managed large-scale initiatives to develop new digital and web-based resources for teaching and learning, and led the integration of these resources into organizational use. She has a B.A. and an M.A., both from UCLA.

Andrew Dixon

Digital Sensitivity Review at the Foreign, Commonwealth and Development Office

Abstract: This case study session looks at the implementation of AI tools in the digital sensitivity review process at the Foreign, Commonwealth and Development Office (FCDO). The process started in 2018 and is the result of a collaboration between the FCDO, academic researchers and consultancy firm SVGC.

Biography: Andrew is the Managing Director of SVGC, a Defence Consultancy providing Evidence Based Decision Support to the Defence and Security Environment. Andrew has been working with the FCDO and other Government departments on their digital sensitivity review process since 2018.

Leontien Talboom

Accessing the Intangible

Abstract: This talk will outline the initial theoretical framework from a collaborative doctoral project which focusses on the constraints that digital preservation practitioners face when making born-digital material accessible. The framework has been created with the help of semi-structured interviews with digital preservation practitioners and an extensive literature review. Topics that will be discussed focus on the digital environment where this material is made available in, the processable nature of digital material and the changing expectations of users when accessing this material.

Biography: Leontien Talboom is a collaborative PhD student at The National Archives, UK and University College London. Her research focuses on the constraints faced by digital preservation practitioners when making digital material accessible.

John Sheridan

Archives in the midst of the AI revolution

Abstract: Digital archives hold important evidence. With the advance of machine learning techniques, the opportunities to gain new insights from data in the archive are rapidly expanding. Similarly, digital archives might also be used to synthesise new content. We do archives stand, as we find ourselves in the midst of the fourth — and perhaps beginning the fifth — industrial revolution, driven by AI? This presentation will address three challenges, from the perspective of The National Archives:

- What kinds of insights and value can be derived from digital archives, using AI based approaches? Where are the greatest opportunities? What potential harms should the archive protect against?
- How should AI approaches best be deployed to aid information management and the selection of records for the archive? What's possible and where are the limits? In particular, how do we make sure that the value of records in aggregate is properly accounted for when making selection decisions?
- How do we select and preserve the AIs that are now such an important part of our digital fabric?

Biography: John Sheridan is the Digital Director at The National Archives, with overall responsibility for the organisation's digital services and digital archiving capability. His role is to provide strategic direction, developing the people and capability needed for The National Archives to become a disruptive digital archive. John's academic background is in mathematics and information technology, with a degree in Mathematics and Computer Science from the University of Southampton and a Master's Degree in Information Technology from the University of Liverpool. Prior to his current role, John was the Head of Legislation Services at The National Archives where he led the team responsible for creating legislation.gov.uk, as well overseeing the operation of the official Gazette. John recently led, as Principal Investigator, an Arts and Humanities Research Council funded project, 'big data for law', exploring the application of data analytics to the statute book, winning the Halsbury Legal Award for Innovation. John has a strong interest in the web and data standards and is a former co-chair of the W3C e-

Government Interest Group. He serves on the UK Government's Data Leaders group and Open Standards Board which sets data standards for use across government. John was an early pioneer of open data and remains active in that community.

Christopher Day

'Computing Cholera': topic modelling catalogue entries for the correspondence of the General Board of Health (1848-1871)

Abstract: The correspondence of the General Board of Health (1848-1871) documents the work of a body set up to deal with cholera epidemics in a period where some English homes were so filthy as to be described as 'mere pigholes not fit for human beings'. Individual descriptions for each of these over 89,000 letters are available on Discovery, The National Archives (UK)'s catalogue. This presentation will examine how data science can be used repurpose archival catalogue descriptions, initially created to enhance the 'human findability' of records (and favoured by many UK archives due to high digitisation costs), for large scale computational analysis. The records of the General Board will be used as a case study: their catalogue descriptions topic modelled using a latent Dirichlet allocation model, visualised, and analysed – giving an insight into how new sanitary regulations were negotiated with a divided public during an epidemic. Questions of the validity and utility of using the descriptions of archival material, as opposed to the records themselves, will also be discussed.

Biography: Chris Day is Head of Modern Domestic Records at The National Archives, the official archive of the UK Government. Chris specialises in the records of 19th and 20th century public health and social security bodies, the records of the Home Office, and of cats on the British government's payroll. He is also interested in digital research, particularly how archival metadata created for one purpose can be reused for computational analysis.
Twitter: @Dentiloquy

Nicola Welch

Towards a future access service for digital records at The National Archives

Abstract: To round off the day, Nicola will talk through some of their plans for developing greater access services for the digital archives at The National Archives

Biography: As the newly appointed Service Owner for Access to Digital Records, Nicola is tasked with scoping and building this service to sit alongside the preservation and transfer services currently offered by the digital archiving department at The National Archives (TNA). Prior to this Nicola was Head of Cross-Government Engagement for TNA, working with the UK government KIM and DDAT professions to tackle the challenges and opportunities posed by the digital transformation of the public record. As a qualified archivist and records manager, Nicola started her career at TNA in the web archiving team which led to a continued interest in working with born digital records.