AEOLIAN Network’s Online Workshop 1:
Employing Machine Learning and Artificial Intelligence in Cultural Institutions

Programme

Wednesday 7th July from **12:00** to **17:30 BST**

12:00 – 12:10: Welcome from **Dr Lise Jaillant** (Loughborough University) and **Dr Annalina Caputo** (Dublin City University).

12:10 – 13.30: Panel 1. Chair: **Dr Maria Castrillo** (Imperial War Museums)

Dr Giles Bergel (University of Oxford / National Library of Scotland)
Title: *Visual AI and Printed Chapbook Illustrations at the National Library of Scotland*

Einion Gruffudd (National Library of Wales)
Title: *Describing the Welsh National Broadcast Archive*

John Stack (Science Museum)
Title: *Machine Learning and Cultural Heritage: What Is It Good Enough For?*

Followed by Q&A

13:30 – 14:30: Lunch Break (1 hour)

14:30 – 15:10: Panel 2. Chair: **Dr Christopher Loughnane** (Auburn University)

Amanda Henley (University of North Carolina at Chapel Hill Libraries)
Title: *On the Books: Creating and Analyzing Collections as Data*

John McQuaid (Frick Collection), **Dr Vardan Papyan** (University of Toronto), and **X.Y. Han** (Cornell University)
Title: *AI and the Photoarchive*

Followed by Q&A

15:10 – 15:30: Interactive Session

*This session is designed to generate casual discussion, share research interests, and get to know other members of the network. Attendees will have the option to attend one of **four** breakout rooms:*

Room 1: Digital Management in Cultural Organisations.
Chair: **Andrew Woods** (Harvard University)
Room 2: Machine Learning and AI Projects.
Chair: **Leslie Johnston** (National Archives and Records Administration)
Room 3: Working Across Disciplines.
Chair: **Gavin Willshaw** (University of Edinburgh)
Room 4: Developing International Projects.
Chair: **Nora McGregor** (The British Library)

15:30 – 16:00: Comfort Break (30 min)

16:00 – 17:00: Keynote Presentation. Chair: **Nicole Coleman** (Stanford University Libraries)

Thomas Padilla (Center for Research Libraries)
Title: *Keep True: Three Strategies to Guide AI Engagement*

Followed by Q&A.

17:00 – 17:30: Roundtable. Chair: **Dr Katherine Aske** (Loughborough University).

Title: *Roundtable discussion with the AEOLIAN Project Team*: **Dr Lise Jaillant**, **Dr Annalina Caputo**, **Glen Worthey** (University of Illinois), **Prof. Claire Warwick** (Durham University), **Prof. J. Stephen Downie** (University of Illinois), **Dr Paul Gooding** (Glasgow University), and **Ryan Dubnicek** (University of Illinois).

Followed by Q&A.

____

Time Conversions (US)
(BST-4) 8:00–13:30 (break 9:30–10:30)
(BST-6) 6:00–11:30 (7:30–8:30)
(BST-7) 5:00–10:30 (6:30–7:30)

Speaker Abstracts and Biographies

Dr Giles Bergel

University of Oxford / National Library of Scotland

Dr Giles Bergel is based in the Visual Geometry Group in the Department of Engineering Science at the University of Oxford, where he works on the application of visual AI to cultural heritage datasets. He has personal research interests in book history, particularly cheap printed forms such as broadside ballads and chapbooks, and has worked on a number of digitisation and accompanying digital scholarship research projects on these forms. He is also interested in the development of reproducibility standards for AI in cultural heritage.

Title: Visual AI and printed chapbook illustrations at the National Library of Scotland
Abstract: This presentation describes a project undertaken within the National Librarian of Scotland's Fellowship in Digital Scholarship programme for 2020-1.

The National Library of Scotland's Data Foundry repository was created to encourage the application of digital research methods to the collections: it includes a large dataset of images, metadata and transcripts of Chapbooks Printed in Scotland. Chapbooks are small, cheap books sold by travelling pedlars, or chapmen, which comprise one of the most innovative and widely-known forms of popular printed literature of their heyday (c.1700-1900). They are frequently illustrated with relief (woodblock or stereotype) prints, which can aid in printer attribution as well as providing evidence of popular visual culture.

This project employed both a variety of computer vision methods to aid in the analysis of the chapbook illustrations. Object detection, using a pretrained classifier retrained on a small sample of the chapbooks, was employed to identify the illustrated pages and to extract the illustrations from the corpus. Next, the illustrations were matched using a visual search algorithm, clustered, and made browsable per visual match and by means of the Library's structured metadata. Candidate matches were registered to provide a means of verification of the closeness of the match, providing also a means of sequencing the printed impressions, and chronological order of publication. Last, an image classifier was applied to the extracted illustrations in order to explore intra-class relationships and similarity to other relevant data.

The presentation will describe several forthcoming outcomes of the research, including a methodological article; a machine learning model; and a dataset of annotated images to encourage improvement of image-detection classifiers. Last, the presentation will offer some reflections on the value of curated data within AI workflows in cultural heritage, and the necessity of further curatorial oversight of their outputs.

Einion Gruffudd

National Library of Wales

Einion Gruffudd started his career as a video librarian at Barcud television resources company in north Wales, before returning to Aberystwyth in 1992 to work at the National Library of Wales where he has served in the Manuscripts, IT and Unique Collections departments. His work has included managing Library systems, business continuity, setting up NLW's digital archive, and successfully leading a HLF funded project to digitise all 1,200 tithe maps of Wales. He has been managing the NLHF funded project to establish a Broadcast Archive at NLW since 2017.

Title: Describing the Welsh National Broadcast Archive

Abstract: This talk will describe how the National Library of Wales is establishing a National Broadcast Archive, a National Lottery Heritage Fund supported project involving acquiring a large corpus of digitised audiovisual material from Welsh broadcasters. This collection which will be made available to the people of Wales for research purposes at various locations across the country.

The project includes a focus on making the collection more discoverable, applying Artificial Intelligence technologies to Welsh Language voice2text and keyword generation. These activities to improve how the collection is described will include volunteer participation in the correction of machine learning output, among many other activities to promote the use of the archive. Issues raised by the ownership and clearance of rights affect all activities including AI activities, and the project's approach to these obstacles will be explained.

The talk will examine how the location of the Broadcast Archive within NLW brings different use cases for archive use, and opportunities to take advantage of other digitisation activities and technologies developed at the Library. A key focus for the end of the project, which will be described, is to develop a "linked data experience" to help people understand the relationships between broadcasting and other historical sources from the wide range of holdings at NLW.

John Stack

Science Museum

John Stack is Digital Director of the Science Museum Group. The Science Museum Group encompasses five museums: Science Museum, London; National Science and Media Museum, Bradford; National Railway Museum, York; Science and Industry Museum, Manchester; and Locomotion, Shildon. He joined in 2015 and is responsible for setting and delivering the Group's digital strategy. He manages the Digital department which encompasses the museums' websites, digitised collections, apps, games and on gallery digital media. Prior to joining the Science Museum Group, he was Head of Digital at Tate for ten years.

Title: Machine Learning and Cultural Heritage: What Is It Good Enough For?

Abstract: Funded through the AHRC's Towards a National Collection Programme, the Science Museum Group (SMG) is collaborating with the V&A and School of Advanced Study, University of London, on a two-year project entitled "Heritage Connector: Transforming text into data to extract meaning and make connections".

As with almost all data, museum collection catalogues are largely unstructured, variable in consistency and overwhelmingly composed of thin records. The form of these catalogues means that the potential for new forms of research, access and scholarly enquiry that range across multiple collections and related datasets remains dormant.

The Heritage Connector project is deploying a range of machine learning-based techniques to extract information from the SMG collection catalogue, link it to third-party sources – primarily Wikidata and the V&A's collection – will then create a set of prototypes that demonstrate and explore the affordances of the resulting dataset.

Rather than attempting to deploy machine learning to create a perfect linked data model, Heritage Connector asks what's "good enough" to provide useful functionality to different audiences.

Amanda Henley

The University of North Carolina at Chapel Hill Libraries

Amanda Henley is Head of Digital Research Services at the University of North Carolina at Chapel Hill University Libraries. Ms. Henley leads a team of data and technology experts in supporting research that uses geospatial technology, spatial and numeric data, statistics, data visualization, text analysis, and other digital scholarship methods. Ms. Henley has extensive experience initiating new, technology-driven services and integrating them into library workflows. She has been principal investigator of a collections as data project since 2019. Prior to that, she was co-PI on a project to train librarians in text analysis.

Title: On the Books: Creating and Analyzing Collections as Data
Abstract: On the Books: Jim Crow and Algorithms of Resistance is a collections as data and machine learning project that uses text analysis to identify racist laws enacted in North Carolina during the Jim Crow era. The project has created two corpora: one contains all North Carolina laws from 1866-1967, and the other contains a subset of laws that were identified as likely Jim Crow laws. The products are publicly available and include the corpora, the scripts written to create the corpora, and a white paper describing workflows. This presentation will provide an overview of the project and share lessons learned.

John McQuaid

Frick Collection

John McQuaid is Photoarchive Lead at the Frick Art Reference Library. He received a BA in Art History and Classics from Case Western Reserve and a MA in the History of Art from The Ohio State University.

Vardan Papyan

University of Toronto

Vardan Papyan is an assistant professor in the department of Mathematics at the University of Toronto, cross-appointed with the department of Computer Science. He received his BSc, MSc, and PhD at the Technion and was a postdoctoral researcher at Stanford University.

X.Y. Han

Cornell University

X.Y. Han is a PhD student in the department of Operations Research and Information Engineering at Cornell University. He received his BSE in Operations Research and Financial Engineering from Princeton University, and MS in Statistics from Stanford University.

Title: AI and the Photoarchive
Abstract: In this talk, we describe a collaborative project between art historians and staff at the Frick Art Reference Library (FARL) and researchers at Cornell, Stanford, and the University of Toronto to develop an algorithm that will apply a local classification system based on visual elements to

the Library's digitized Photoarchive—a study collection of 1.2 million reproductions of works of art. We leverage state-of-the-art artificial intelligence (AI) systems to develop a classifier for the automatic annotation of digitized but not-yet-catalogued images in the FARL's Photoarchive. This was achieved by engineering the syntax of the classification system into the training and predictive process of deep convolutional neural networks, the cornerstone of modern AI advancements.

The classifier is integrated into a mobile and desktop application that allows Photoarchive staff to quickly validate or correct the decisions of the networks. We demonstrate promising performance metrics and offer informative scientific insights that have the potential to create a valuable tool for metadata creation and image retrieval. This project offers a useful model for effective interdisciplinary interaction.

Thomas Padilla

Director of Information Systems and Technology Strategy
at the Center for Research Libraries

Thomas Padilla is Director of Information Systems and Technology Strategy at the Center for Research Libraries. He is the author of the library community research agenda, *Responsible Operations: Data Science Machine Learning, and AI in Libraries*, Principal Investigator of *Collections as Data: Part to Whole*, and past Principal Investigator of *Always Already Computational: Collections as Data*. Thomas is Vice Chair, ACRL Research and Scholarly Enviaronment Committee; Executive Committee Member, Association for Computers and the Humanities; and Technical Advisory Board Member, Linked Infrastructure for Networked Cultural Scholarship.

Title: Keep True: Three Strategies to Guide AI Engagement

Abstract: Recurrent bouts of AI enthusiasm over decades suggest no sector is immune to losing itself in the face of potential. In the archipelago of varied sector actors implementing AI, GLAMs have an opportunity to distinguish themselves. While the component parts of this community are quite different and sometimes functionally opposed in approaches to similar work, we share in common a set of contemporary commitments that seek to advance equity in the communities we serve. In what follows I will present three strategies I believe strengthen our ability to realize these commitments: nonscalability imperative, avoiding neoliberal traps, and seeing maintenance as innovation.