Cambridge Cultural Heritage Data School

17 – 30 March 2021
Online delivery

Cambridge Data School contacts

Dr Anne Alexander, CDH Learning Director (E: raa43@cam.ac.uk)
Huw Jones, Library DH Coordinator, CDH Labs (E: hej23@cam.ac.uk)
Chiara Capulli, Methods Fellow, CDH Learning (E: cc826@cam.ac.uk)
Dr Mary Chester-Kadwell, Senior Software Developer (E: mec31@cam.ac.uk)
Karen Herbane, DH Learning and Events Coordinator (E: admin@cdh.cam.ac.uk)
## Timetable

Dates and provisional timings (all timings in GMT)

*Please note that because of the fast moving nature of Coronavirus, this timetable and content may be subject to change because of staff availability.*

<table>
<thead>
<tr>
<th>Date</th>
<th>Time</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wednesday 17 March 2021</td>
<td>1.30pm – 2.15pm</td>
<td>Introduction and welcome</td>
</tr>
<tr>
<td>Wednesday 17 March 2021</td>
<td>2.30pm – 3.30pm</td>
<td>Session 1: Digital Research Design and the Project lifecycle</td>
</tr>
<tr>
<td>Thursday 18 March 2021</td>
<td>1.30 – 2.30pm</td>
<td>Session 2: Digital text mark-up and TEI I</td>
</tr>
<tr>
<td>Thursday 18 March 2021</td>
<td>3.00 – 4.00pm</td>
<td>Session 3: Geodata collection and cleaning I</td>
</tr>
<tr>
<td>Friday 19 March 2021</td>
<td></td>
<td>Self-paced study day</td>
</tr>
<tr>
<td>Monday 22 March 2021</td>
<td>1.30pm – 2.30pm</td>
<td>Session 4: Semantic data modelling I</td>
</tr>
<tr>
<td>Monday 22 March 2021</td>
<td>3.00 – 4.00pm</td>
<td>Session 5: Named Entity Recognition I</td>
</tr>
<tr>
<td>Tuesday 23 March 2021</td>
<td></td>
<td>Self-paced study day</td>
</tr>
<tr>
<td>Wednesday 24 March 2021</td>
<td>1.30 – 2.30pm</td>
<td>Session 6: Machine Learning and image collections</td>
</tr>
<tr>
<td>Wednesday 24 March 2021</td>
<td>3.00 — 4.00pm</td>
<td>Session 7: Digital text mark-up and TEI II</td>
</tr>
<tr>
<td>Thursday 25 March 2021</td>
<td>1.30 – 2.30pm</td>
<td>Session 8: Geodata collection and cleaning II</td>
</tr>
<tr>
<td>Thursday 25 March 2021</td>
<td>3.00 — 4.00pm</td>
<td>Session 9: Semantic data modelling II</td>
</tr>
<tr>
<td>Friday 26 March 2021</td>
<td></td>
<td>Self-paced study day</td>
</tr>
</tbody>
</table>
Data School live pre-session; Dr Anne Alexander to facilitate session

17 March 2021
This session will include a short presentation about CDH combined with opportunities for a Q and A as well as introductions by participants.

We will organise online drop-in sessions for participants needing remote help with software installation and other technical problems, these will be advertised closer to the date of the school.

Module 1: Digital Research Design and the Project lifecycle

17 March 2021
This introductory session explores the lifecycle of a digital research project across the stages of design – data capture, transformation, analysis presentation and preservation. It also introduces tactics for embedding ethical research principles and practices at each stage of the research process.

Module 2: Digital text markup and TEI with Huw Jones

18 March and 24 March
The TEI (Text Encoding Initiative https://tei-c.org/) is a standard for the transcription and description of text bearing objects, and is very widely used in the digital humanities – from digital editions and manuscript catalogues to text mining and linguistic analysis. This module will take you through the basics of the TEI – what it is and what it can be used for – with a particular focus on uses in research, paths to publication (both web and print) and the use of TEI documents as a dataset for analysis. There will be a chance to create some TEI yourself as well as looking at existing projects and examples. The module will take place over two sessions – with an introductory taught session, then a chance to work on TEI records yourself, followed by a review and discussion session.

Module 3: Geodata, controlled vocabularies and principles of semantic data modelling with Chiara Capulli

Geodata – 18 and 25 March
Semantic data modelling – 22 and 25 March
This module will focus on techniques to enrich historical data about artworks in a way that they can be mapped, standardised, and structured against machine-readable schemas. Using information from the 16th-century ‘The Lives of the Most Excellent Painters, Sculptors, and Architects’ by Giorgio Vasari, we will collect and map geospatial information; we will then clean and reconcile the data against controlled vocabularies with OpenRefine. The second part of the module will introduce the CIDOC CRM, the semantic data standard for Cultural Heritage designed to ensure longevity and interoperability to digital documentation.

Module 4: Named Entity Recognition with Python with Dr Mary Chester-Kadwell
22 and 29 March

Text-mining is extracting information from unstructured text, in other words, text that has not been encoded with semantic markup. In this module we will look at one way of extracting information from unstructured text by recognising named entities automatically. A named entity is any type of real-world object or concept, such as, a person, organisation, location or date. Using the example of letters from the 19th-century botanist John Stevens Henslow, we will introduce how to: recognise and visualise named entities using machine learning; create training data for improving the results; and link named entities to existing knowledge bases.

Participants will be able to choose either a ‘no code’ or a ‘Python’ track for this module. Everyone will join the same virtual sessions, and have access to the same self-paced study materials and exercises, but the suggested directions given will be different depending on which track you choose to follow. For those with experience in Python, the materials include a set of Jupyter notebooks using the spaCy NLP library, but prior knowledge of Python is not required in order to complete the module.

Module 5: Using machine learning to work with large-scale image collections with Dr Anne Alexander
24 and 29 March

Machine vision systems can potentially help users and curators of historical and cultural image collections see them differently, providing both tools for research and supporting new modes of public engagement and discovery. These sessions provide an introductory overview of basic tasks in machine vision, such as Image Classification, Object Detection and Image Captioning within a critical framework highlighting the problems of algorithmic bias which emerge when applying machine learning-based tools to understanding and interpreting data from different periods of history and across different cultures.
Biographies

Anne Alexander
Director of CDH Learning
Anne Alexander is Director of the Learning Programme at Cambridge Digital Humanities. Her research interests include ethics of big data, activist media in the Middle East and the political economy of the Internet. She is a member of the Data Ethics Group at the Alan Turing Institute and a member of the Steering Group of the Trust and Technology Strategic Research Initiative.

Caroline Bassett
Director, Cambridge Digital Humanities
Caroline Bassett is Professor of Digital Humanities, and Director of Cambridge Digital Humanities. Her research explores digital technologies in relation to questions of knowledge production and epistemology (how does ‘the digital’ change scholarship, transform understanding, produce new scales or perspectives?) and in relation to cultural forms, practices, and ways of being (how can we understand the stakes of informational capitalism, what are its symptoms, how can we understand its temporalities, the forms of life it enables, and those it forecloses?).

Chiara Capulli
CDH Data School Fellow
Chiara Capulli is an AHRC-funded PhD student in the Department of History of Art, with a research project that examines the consequences brought by the 1529 Guasto of Florence to the artistic and architectural heritage of a number of suburban religious houses. In her PhD, she tackles the themes of displacement, artistic identity and networks of artistic patronage by integrating traditional art historical research with the spatial methodologies of GIS and 3D reconstruction. As a researcher on the Getty-funded Florence 4D project (University of Exeter – University of Cambridge) over the past two years, she has addressed the problems of reconstructing and re-contextualising altarpieces on a large, city-wide scale; and has helped develop a pipeline for research-based 3D modelling, metadata-mapping and sharing using the CIDOC CRM ontology and IIIF manifests.
Mary Chester-Kadwell
Digital Humanities Developer (CDH Lab)

Mary Chester-Kadwell is a Senior Software Developer at Cambridge University Library and Digital Humanities Developer for CDH Lab, where she advises researchers on the technical aspects of their projects. Mary has a PhD in the landscape archaeology and material culture of early medieval England, which focussed on data-driven spatial analysis with GIS and the methodology of metal detecting. Mary is interested in the pedagogy of coding, and teaches text-mining and the programming language Python.

Huw Jones
Library Digital Humanities Coordinator (CDH Lab)

Huw Jones is Head of the Digital Library Unit and Digital Humanities Coordinator at Cambridge University Library, working with researchers, curators, and technical staff to make special collections accessible online. Cambridge Digital Library is our main platform for the digital humanities, freely accessible to anyone with an internet connection. It contains more than 35,000 items, ranging from the papers of Isaac Newton and Charles Darwin, to manuscript and photograph collections representing the global scope of the Cambridge's physical collections.