AI and Historical Literacy – The SOLM-2018 data set

Colin Greenstreet, Signs of Literacy community organiser
The National Archives, Tuesday, September 4th, 2018

Slide number	Main thought	Text
1	Good morning	My name is Colin Greenstreet. I'm the community organiser of an initiative called Signs of Literacy. I am also a director and founder of a collaborative project called MarineLives, which started at a TNA hackathon six years ago. My background is in research intensive business – as a management consultant, as a Vice President in pharmaceutical R&D, and as a pharmaceutical entrepreneur. For the last six years, I have focussed on social ventures.
2	Project portfolio	Today I am going to talk about two of our projects – Signs of literacy and the SOLM-2018 database. The vision behind these closely related projects is to mine archival manuscripts at scale using machine learning and other computational techniques. The SOLM-2018 database, which I am going to be introducing, is a machine learning training data set of 10,000 markes, initials and signatures, extracted and annotated from C16th and C17th digitised mansucripts.
3	Perspective	Without data, machine learners are helpless. Back in 2007, a group of scientists interested in face detection, put together a data set called Labelled Faces in the Wild. A recent paper, from 2016, noted the huge impact that data set has had on facial recognition We want to do the same for the study of historical literacy, using markes, initials and signatures. Of equal importance, we want to stimulate the machine extraction of metadata from digitised manuscripts and its linkage to machine identified markes, initials and signatures.
4	Pattern seeking	We humans are excellent pattern seekers. We can distinguish initials, markes, signatures at speed.
5	Initials	You will have seen a good number of Ts, a few Js, and an odd R and S in the GIF on the last slide.

		Look at the early modern Js on the right hand side of the screen – thick, thin, slanting, squiggly, three bars or two. One on its side. But recognisably the same initial
6	Work together	So why do we need machine learning, if we are so good at pattern recognition? Firstly, there are relatively few people who can read cursive secretarial hands, which dominate Early Modern manucripts Secondly, we are not so good at handling larger quantities of images consistently in terms of ranking and grouping. Our vision is to use computational techniques to open up the archives – locating markes, initials and signatures in digitised manuscripts, which would otherwise be invisible and unsearchable
7	SOLM-2018 database	Our starting point is the creation of a machine learning training set of markes, initial snad signatures. We have named this SOLM-2018 [Signs of Literacy in Manuscripts]. We have 5,000 annotated signoffs in the dataset. Our target is 10,000 signoffs, which is the minimum size to do interesting things with machine learner. The first 5,000 signoffs are drawn from High Court of Admiralty depositions from the C16th and C17th held at the National Archives here in Kew.
8	Our vision	Our vision is a SOLM-2023 database with 1 million markes, initials and signatures from across Europe and North America from the C16th to the C18th. To create this database by hand would require a minimum of 50 person/years and would take ten years. Our challenge to archivists, computer scientists and historians is to do this with five person/years and in half the time.
9	Partners	We are looking for partners in the United Kingdom and internationally, and would love to talk to any archivists and technologists at today's symposium, who would like to learn more. Currently, we are working with the National Archives; with Kaggle, which is the world's largest the machine learning community, now owned by Google; with the IIIF consortium; with the Alle Amsterdamse Akten project team at the

		Staatsarchiv Amsterdxcam; and with Picturae. These are working relationships and are not yet structured as partnerships. We have started to approach local and county archives in England.
10	Kaggle	Kaggle has selected us as one of a small number of probono competitions they support each year. This is on the merits of our proposal and for the potential impact on the research field and archival sector. We are planning to use a Kaggle machine learning competition as a Proof of Concept of our approach to signoff identification in manuscripts. The competition is tentatively scheduled for November this year, but we are considering moving this to the third quarter of next year. This would enable archivists and users of all sorts with technical interests to engage in our initiative and to consider participating in Kaggle competition teams next year.
11	High pixels	Our dataset contains digitised full page manuscript images AND bounded signoff snippets. The snippets are annotated stylistically and semantically. Semantic annotations include the spelling of the signature, and the age, occupation, and residence of the signer, and the date of the signoff. The dataset is definitely a dataset "in the wild" – much of it is high resolution, but we have deliberately included low resolution images. Following advice from machine learners, we have not normalised the images in terms of width, height, or colour balance. We have also included occluded signoffs, and signoffs with underlying manuscript curvature.
12	Low pixels	It will be interesting to see what machines make of low pixel definition signoffs, and how machine learners deal with occlusion and manuscript curvature.
13	Colour analysis	Image processing is key to image based machine learning. There are lots of decisions machine learners will need to make and plenty of image processing suites, such as Open CV

14	Detection and analysis (1)	It will be exciting to see what we can do with the images We are interested in machine detection and analysis of blots, smudges, stylistic features, and deletions
15	Detection and analysis (2)	We are also interested in algorithmic detection of "shake" in straight and curved lines. Shaky lines may be a sign of poor signature execution. This could be due to relatively low levels of literacy, or due to the effect of illness or age. With large quantities of data, we will be able to explore this.
16	Porters	We are particularly interested in exploring literacy and sophistication of signoff execution by occupation. This type of analysis requires excellent metadata Here you see nine signoffs by porters, handling coals, whale oil, ginger and corn. Four use markes. One uses initials, nicely formed. There are four signatures of varying quality – one is smudged, one is somewhat uncertain. Dig into the metadata, and you will see that the porters with signatures include two citizens, with two further citizen porters using a marke and initials. The range of portering activities is from heaving and labouring on lighters to higher status work in yards and warehouses.
17	IIIF viewers	Once the snippets are annotated and linked to metadata there are interesting things we can do to make them available to users for close reading. Here you see a mockup of anchor markes drawn from the SOLM-2018 database displayed as a manifest in a IIIF-compliant Mirador viewer.
18	Boundary boxes	There's a lot which can be done with simple boundary boxes. Here we are using boundary boxes to display and quantify the visual geometry of a signature.
19	Simple signatures	These are simple signatures with no flourishes.
20	Flourishes	These are complex signatures of merchants. The Hebrew signature is that of David Ben Mordeccai, a forty-one year old Moroccan Jewish merchant, who was travelling on the ship the *Prophet Elias* from Hamburg, when the ship ws

		seized by ships of the English Commonwealth. It is dated October 28th 1653. Note the flourish is from the left, not from the right. This reflects Hebrew being written, like Arabic, from right to left.
21	Legal deposition	One big challenge is the automatic identification, mapping and linking of metadata and signoffs, particularly when they are several pages apart.
22	Machine recognition	We want to work on machine recognition of metadata – both as a block or blocks of text, and to get to the semantic components of that metadata.
23	Speech to text recognition	Speech to text recognition has become increasingly familiar in the last three years.
24	Key word spotting	We see the potential to use handwriting text recognition and key word spotting techniques to excavate raw metadata
25	Raw to refined metadata	The real challenge, we believe, is to refine raw machine generated metadata using a combination of NPL, controlled vocabularies, and programmable decision rules. That challenge requires archivists to work closely with computer scientists and users if it is ever going to be taken out of the lab and put into practice.
26	Visual metadata	We need visual metadata, which can be machine processed Maths 101. You can add them, subtract them, and otherwise manipulate them.
27	Work together	Archivists, computer scientists and users of all sorts need to work together in the creation of these visual metadata maps and techniques Historians, in particular, are far too distant from archivists, despite depending on archives for their livelihood.
28	Early results (1)	Three slides now to show the power of working with signoffs at scale. These data are NOT machine generated. But they have been extracted from the SOLM-2018 machine learning training data set. This first slide shows signatures by occupation. 60 signoffs for each occupation displayed, with the exception of a few, like fishermen and labourers, where we are still accumulating data.

		At the top, in yellow, pursers – all sixty use a signature. At the bottom, fishermen in orange, with between zero and 20% using signatures – the rest use markes and initials. The standard deviation on these data ranges between roughly 10 and 20%, though of course with pursers it is zero. Basically, there is a high literacy group on this slide of masters, masters' mates, gunners and boatswains. A middle literacy group of common seamen, described in the manuscripts as "one of the company", "common men", "foremastmen" and "sailors". A lower literacy group of watermen and men on coal ships. And finally, the lowest literacy group of fishermen and labourers. These data supplement the work of Cressy, Earle, and Hubbard. They promise, when scaled up, to offer powerful insights into literacy by occupation, age, and location over time.
29	Early results (2)	This second slide is a novel one, since no significant work has been done on literacy at a comparative parish level on a large scale. The data show early to mid-C17th London as a linear maritime city, as seen in the location of High Court of Admiralty deponents, largely from the 1637 to 1667 period
30	Early results (3)	The third slide shows in detail the area to the east and south-east of the city of London. The north bank super-parish of Stepney dominates this map showing the east and south-east of the city of London Seventy-five percent of all the Admiralty Court deponents outside the city of London in the counties of Middlesex, Surrey and North Western Kent are north of the River Thames A stunning sixty-three percent of these deponents outside London are in the many hamlets of Stepney – Wapping, Wapping Wall, Shadwell, Ratcliff, Limehouse, Poplar and Blackwall.
31	Contact details	That's me with my dog up a mountain That's Dr Mark Hailwood, chilling somewhere nice I have included our email addresses and some web links Email: colin.greenstreet@gmail.com m.hailwood@bristol.ac.uk

		Weblinks: http://signsofliteracy.org http://marinelives.org http://chronoscopic.org GitHub: https://github/Signsofliteracy/Signoff Twitter: Marinelivesorg Our GitHub site has a repository called Signoff, where you will find a PDF of today's presentation and plenty more of interest. Do join the organisation, to get full access to our material and to our GitHub project groups. I would like to thank John Sheridan and the National Archives for the invitation to today's symposium, and also David Underdown of the National Archives, who is on our Kaggle competition organising committee. Maggie Demkin at Kaggle, for recognising the potential of a approach to archival manuscripts. Glen Robson, technical coordinator of the IIIF consortium, for his support. And finally, Mark Lindeman, Chief Executive of Picturae, who has offered us server space and software development resource.
32	Our contributors	Finally, finally, a very big thank you to our contributors. Academics, students, archivists, librarians, software developers, and many many others. From England, Scotland, Wales, Ireland, France, Germany, Sweden, Italy, Spain, Canada, the United States and Australia. They are volunteer transcribers, annotators, commentators, glossary contributors, developers of our semantic media wiki, academic advisors, interviewees, workshop participants and speakers
33	Discussion	We have some time for discussion
34	Framework for discusion	Here's a simple timeline with our goal for 2023 at the far right and a lot of blank space