Creating a technology enabled manuscript community to explore historical literacy

IIIF, Transkribus, & Recogito

Presentation at Washington DC IIIF Conference, Thursday, May 24th, 2018, Morning session: Colin Greenstreet: Founder, Chronoscopic Education; Co-director, MarineLives; Community organiser, Signs of Literacy

SLIDE 1

Good morning.

My name is Colin Greenstreet. I'm here to talk about **creating a technology enabled manuscript community to explore historical literacy.**

As well as addressing the potential of IIIF, I am also going to reference two other interesting technical ecospheres for document users – Transkribus and Recogito.

My interests are in Early Modern historical manuscripts and what they can tell us about the world. And when I speak about "manuscripts", I mean documents which were created by clerks, notaries and ordinary people and are hand written.

They may be loose documents or may have been subsequently bound and foliated.

These documents exist in the tens of thousands, hundreds of thousands and millions in archives, libraries and even museums, unlike medieval manuscripts. Their sheer volume means that my interests in the IIIF context span both the manuscripts and the newspaper community groups, and I have a strong interest in applying AI and Machine Learning techniques to the documents.

SLIDE 2

In my presentation today I would like to do four things:

(1) Introduce Chronoscopic Education, together with two of our five initiatives – MarineLives and the Signs of Literacy community

(2) I want to share our vision of a technology enabled manuscript community to explore historical lteracy

(3) Discuss the relevance of the IIIF, Transkribus and Recogito technical ecosystems to our vision

(4) And call for new content and technical partners to work with us on this vision over the coming years

SLIDE 3

Chronoscopic Education was formed earlier this year and has three trustees – myself an American named Bob Egan, who is a former Chief Information Officer of Boise Inc., and a French legal historian, Nga Bellis-Phan.

It is a not-for-profit social venture, and is intended to be the legal, funding and technology home for community based historical initiatives. We will be registering the organisation with the English Charity Commission as a Charitable Incorporated Organisation.

We have three aims:

(1) To further the teaching of skills at universities and schools, specifically palaeographic, digital research and project management skills

(2) To apply insights drawn from small teams in management consultancy and R&D project management, together with tools from data science, to the subject of history

(3) To build a virtual manuscript-based archive and associated research community, which will foster a culture of collaborative scholarship

SLIDE 4

It all started at a hackathon at the English National Archives in 2012.

I was interested in the dynamics of technology enabled collaboration and decided to recruit a bunch of volunteers to transcribe one volume of legal documents from the English High Court of Admiralty from 1656.

I had a vision of facilitated collaboration, drawing on my experience at McKinsey and Booz.Allen and of pharmaceutical R&D, the industry I know best, having been a Vice President in R&D at GlaxoSmithKline.

Six years later, we have a thriving community and six million words of full text semi-diplomatic transcriptions.

We revised our strategy late last year and have broadened our period and content range and are looking to scale up and partner.

It is our drive for scale and reach which has led to our interest in how different technical ecosystems can support our ambitions and which has led to our interest in IIIF standards for our document images.

SLIDE 5

Here's a typical document image, with its full text transcription

Folio 1 recto from HCA 13/71, the volume we worked on back in 2012

The image is viewable on our semantic media wiki. Click on the image in the wiki page and you will get a zoomable image, adequate for transcription and inspection.

At the moment it is not exposed to IIIF.

SLIDE 6

And here are a list of our MarineLives volunteers.

The left hand "tombstone' lists the **twenty-nine volunteers** who worked on our original volume, HCA 13/71, back in 2012.

Eight of that group of twenty-nine remain involved in MarineLives six years later.

The grey block of names lists **over one hundred and fifty contributors** to MarineLives since we started. They have contributed as transcribers, annotators, commentators, advisors, interviewees and coders.

They live in England, Scotland, Wales, Ireland, France, the Netherlands, Spain, Italy, Sweden, Canada, the United States and Australia.

They include university undergraduates, PhD candidates, early career scholars and full professors, archivists and librarians, archaeologists, digital humanists, corporate linguists, web designers, lawyers, retirees, and many many more wonderful occupations.

Beyond these contributors, we have a rich and vocal Twitter followership numbering – **one thousand eight hundred and seventy six at the last count.**

You can follow us on Twitter at @Marinelivesorg

SLIDE 7

Let me move now to our new Signs of Literacy community initiative and invite you all to consider joining and contributing.

The community is in its infancy, and is being driven by Chronoscopic Education and MarineLives, by the social historian Dr Mark Hailwood, who is at Bristol University, and by Mark Ponte, the acting project leader of the Alle Amsterdamser Akten project at the Amsterdam Municipal Archives.

Take a look at some of the image snippets on this slide.

You will see a mixture of markes, initials and signatures.

Some are simple, some quite complex.

They are the work of mariners, coopers, fishmongers, fishermen, merchants and many many more marine and shore based occupations.

They exist in the hundreds of thousands and millions in archvies, libraries and museums all over the world, and **we want to collect these and work on them at scale.**

SLIDE 8

Historian David Cressy had the insight forty years ago that markes, initials and signatures could be used as a surrogate marker for literacy.

His approach was simple – if you could sign your name, you were literate, I you left a mark you were not literate.

His work was innovative and spawned studies of literacy across different languages and periods, but the scale of the work was at the tens of thousands.

Forty years later, the time is right to work at much greater scale and to apply more sophisticated approaches to measuring literacy and to do this in a collaborative, comparative and international manner.

We want to build a diverse community – historians, linguists, modern literacy researchers, digital humanists, developers and machine learners

We are taking the rest of this year to form the community and to refine its aims, culminating in an exciting machine learning competition at the end of the year.

Then in 2019 we want to put together a series of grant bids to pursue different strands of research under the signs of literacy umbrella. We are organic, not hierachical, and that will be reflected in our funding strategy.

SLIDE 9

Let's look at some of the raw material and surface some of the technical challenges.

I'm going to use High Court of Admiralty documents as the use case.

Here you see the deposition or witness statement of Mark Harrison. He was a mariner and master of a ship. He lived in Wapping on the Thames, and was twenty seven when he made his deposition on September 21st 1659. This deposition comes from HCA 13/68 and spans folio 1r through to 3r.

The deposition is five pages long

The document itself is written in the hand of a notary or clerk.

The only portion of the document which gives direct evidence of literacy is the signoff, at the end of the deposition.

SLIDE 10

The start of the deposition contains useful metadata

The date of the deposition, the name of the deponent, his place of residence, his occupation and his age.

We can hand extract and transcribe these metadata, but at scale we would like to do this automatically.

SLIDE 11

The end of the deposition contains the sign off.

Here you see a signature “Marke Harrison”

Whereas the notary spelled Mark’s name without an “e”, Marke himself signs with an “e”.

We can hand extract and classify these signoffs, but at scale we want to do this automatically.

Can any one see a problem?

There is a cross to the left of Marke’s apparent signature.

Could Marke’s actual signoff be the cross and the apparent signature simply be written by the notary?

The answer is no, because the handwriting of the signature is different from the body of the deposition, and the spelling of the first name differ’s from the notary’s spelling in the metadata. And finally, when a deponent

SLIDE 12

Once you have the metadata AND the signoffs, and can link them, you can start to do some powerful things.

On this slide you see the signoffs of dockyard and warehouse porters handling different commodities

Some of them sign with very crude markes, some with initials, and some with very well executed signatures.

Dive deeper into their places of residence, their ages, and their detailed activities and we soon see that quite a wide range of types of people are being captured under the occupational descriptor of “porter”.

Richard Wincles, top left, is described in the notary’s metadata as a porter, and in the text of the deposition describes being employed as a labourer with fifteen other men to unload coals from a ship into lighters at Wapping. His signoff is a crude marke.

George Greenwood, the one with the nice signature, on the right, half way down, is a citizen and vintner, but self-described in the text of the deposition as a porter employed by the Commissioners for Prize Goods to deliver ginger to a London warehouse

SLIDE 13

There are many patterns to identify and explore in the data.

This slide shows that anchors were a popular marke amongst mariners, and former mariners.

Take the looking glass maker Andrew Beake, who provides two slightly different anchor marks three weeks apart in 1655 – he was a former seaman.

We are keen to use machine learning to examine sub-groups of signoffs, within the three main classes of markes, initials and signatures.

SLIDE 14

The best way to learn more about the Signs of Literacy initiative is to take a look at our GitHub organisation.

We have an active wiki, listing planned events and research issues, and a number of teams.

The next community event will be a four hour workshop at the Stadsarchief Amsterdam on Tuesday June 5th, hosted by my colleague Mark Ponte.

The subject is Technology Tools to Explore Historical Literacy and it will bring together for the first time archivist and technologists at the Stadsarchief Amsterdam, with the Chronoscopic/MarineLives team

There will be both technical and non-technical participants from England, the Netherlands, and France.

Glen Robson is planning to join in remotely by ZOOM, as is Rainer Simon, the technology lead for Recogito.

In parallel, we are starting discussions with John Sheridan, Digital Director of the National Archives in England, about potential National Archvies participation in the machine learning aspects of what we are doing.

SLIDE 15

We are interested working with three technology ecosystems as we build a technology enabled manuscript community to explore historical literacy.

We are clear that IIIF should be at the center, but also have strong interest in working with Transkribus to utilise their handwriting text recognition capabilities of Transkribus, together with their document layout recognition capabilities. We are also looking to collaborate with Recogito, for its semantic annotation capabilities, which are expanding beyond geodata into other aspects of semantic annotation.

The cultures, organisation, funding and scale of these three ecosystems are quite different. IIIF is an organic, dynamic collaborative community, embedded in GLAM. Transkribus is vertical, with a greater separation between the technologists and the users, and comes out of academia. Recogito is an initiative of the Mellon funded Pelagios commons, and shares a lot of the community spirit and collegiality of IIIF, though at a smaller scale and coming out of academia.

We hope that Chronoscopic Education and the Signs of Literacy initiative can help promote dialogue between these three systems, and that some of the funding we intend to raise in 2019 and beyond can help develop specific interactions to support the study of literacy. Transkribus has a lot to learn, I believe, from the community spirit and organisation of IIIF, and IIIF has much to gain from reaching out to the archival user community, which is at the core of Transkribus members.

I will be representing Chronsocopic Education at the next Transkribus users conference in Vienna in November, and am also concluding a Memorandum of Understanding with Transkribus to become a formal member. We are also making a big commitment to getting 3 million words of our full text transcription together with 4,000 images into the Transkribus training database, which will make us the biggest training base in their system. It would be great if all those images could also be made IIIF compliant.

SLIDE 16

Our experience of running MarineLives as a low cost highly collaborative community has taught us a lot about working with technology and technologists.

Six years is long enough to have changed platforms from a Wordpress Scrripto implementation, to a series of wikis, to an integrated semantic media wiki

It is also long enough to have experienced a forced data migration and the loss of support for specific plugins

As the power of our data has become clearer to a wide range of users, it is become clearer that we need interoperability both for researchers and for technical providers

So we have some design principles as we think about the technology ecosystems and communities we want to work with.

SLIDE 17

We are going to take the rest of this year and into early next year to develop and sequence our user requirements.

Some of them will be experimental and quick and dirty tool based.

Others will be requirements for a robust sustainable system

SLIDE 18

Two of the requirements are around machine learning, and it is these that we wish to test in a Proof of Concept at the end of this year

SLIDE 19

We are delighted that we have been selected by Google owned Kaggle as one of a small number of pro bono competitions they support each year.

This is on the merits of our proposal and the potential impact on the research field of running the competition.

Kaggle will cover the running costs of the competition. We will provide the prize pool and are now out money raising from potential sponsors and partners.

The Signs of Literacy Kaggle Research Competition will launch in November 2018 and run till early January 2019.[1]

The competition is a Proof of Concept and will contain two parts

(1) Algorithmic identification of markes, initials and signatures

(2) Algorithmic discrimination between degrees of "sophistication" within the three categories of "marke", "initial" and "signature"

After proving the concept, we will be looking for an image or vision oriented computational laboratory with which to develop a grant funded collaboartion to take the work further in 2019 and beyond.

SLIDE 20

A big requirement of our Kaggle competition will be to create a training data set of approximately 10,000 images.

Ideally we would get these into IIIF and then annotate them, before hauling them into a database to expose to the Kaggle machine learners.

We plan to use a team of volunteers to annotate this training dataset, including classifying the signoffs by degree of sophistication of execution.

One way to do this would be to use conjoint analysis to force a binary choice by volunteers as to which of two images is more sophisticated in execution.

[1] https://www.linkedin.com/pulse/proposed-signs-literacy-kaggle-research-competition-2018-greenstreet/, viewed 27/05/2018

I'm a big fan of Jack Reed's implementation of MapTab as a chrome extension to display David Rumsey's IIIF map collection. So imagine a Conjoint analysis plugin, which popped up when you launched your laptop, and invited you to click on one of the images for a certain attribute.

SLIDE 21

Lots to do and lots of partner opportunities.

Specifically, we are looking for one or more additional content partners in the US, to complement material from the National Archvies in London and the Stadsarchief in Amsterdam.

We are also looking for technical partners to work with us in four areas

- IIIF crowdsourcing
- IIIF annotation
- Image based machine learning
- Metadata extraction and analysis using HTR

SLIDE 22

If you are interested, or know of someone who may be interested, we are easy to find

Here are our contact details.

My long hair days are over, unlike my colleagues, but I do have a hairy dog

SLIDE 23

And some further reading if you want to learn more on the content side