Perspectives on Transkribus: MarineLives & Signs of Literacy initiative

Transkribus
Users Conference 2018

Thursday, November 8th, 2018
Technische Universität, Wien

Colin Greenstreet

Slide pack:
https://github.com/Signsofliteracy/ArchiveBots

MarineLives

Maphackathon

Signs of Literacy

SOLM-2018

EM Textiles, Garments & Dyestuffs Glossary

EM Maritime & Mercantile Gazetteer

Chronoscopic Education charitable incorporated organisation

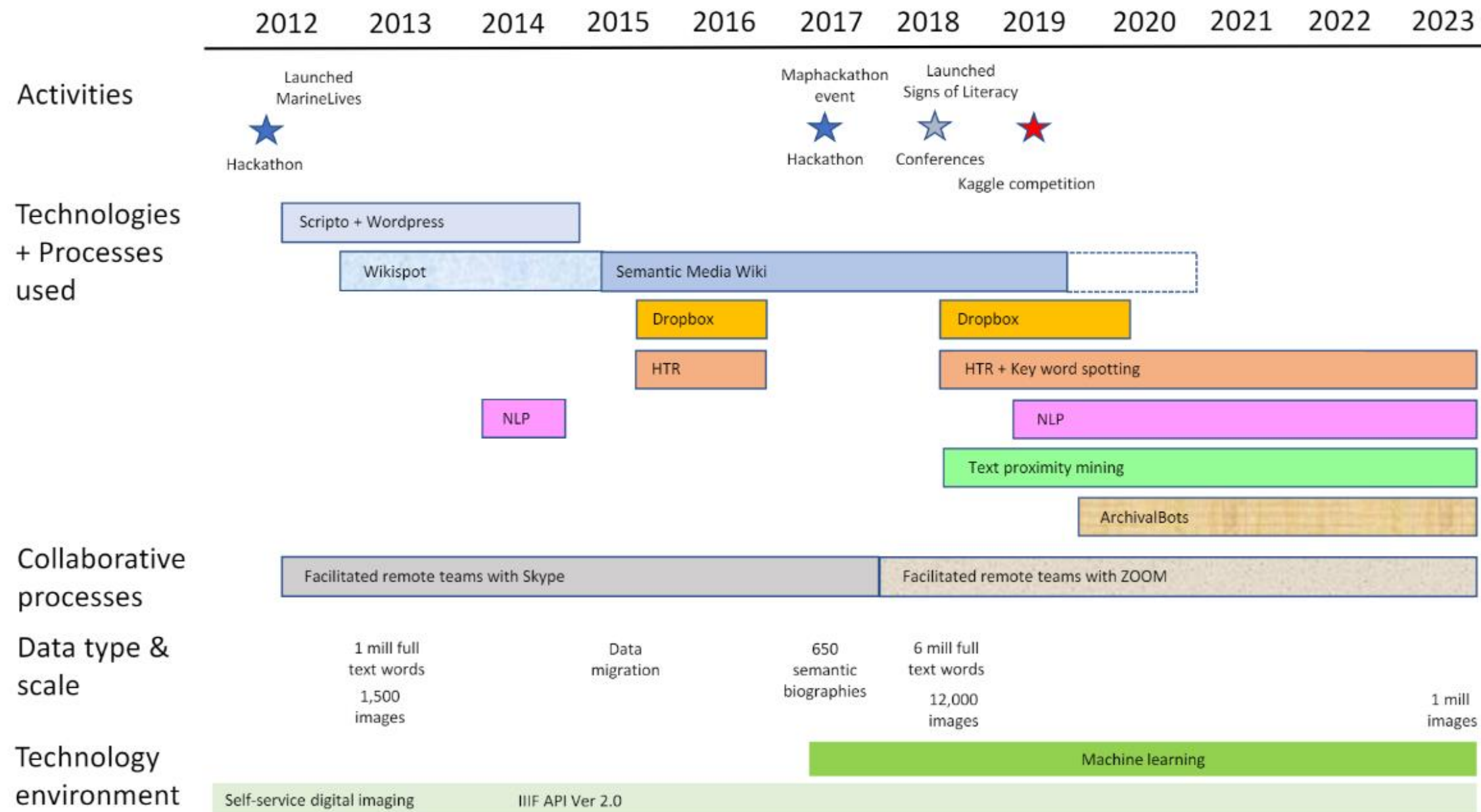FOR INFORMATION ABOUT OUR OUR AIMS AND OUR PROJECTS

We are registering a charitable incorporated organisation under the name of Chronoscopic Education, to be the legal, funding and technology home for MarineLives, our sister project Maphackathon, and our new community, Signs of Literacy.

The aims of Chronoscopic Education are threefold:

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

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

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

MarineLives & its progeny: changing technology and processes

	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023

Activities

Launched MarineLives ★ Hackathon

Maphackathon event ★ Hackathon

Launched Signs of Literacy ★ Conferences

★ Kaggle competition

Technologies + Processes used

- Scripto + Wordpress
- Wikispot
- Semantic Media Wiki
- Dropbox
- Dropbox
- HTR
- HTR + Key word spotting
- NLP
- NLP
- Text proximity mining
- ArchivalBots

Collaborative processes

- Facilitated remote teams with Skype
- Facilitated remote teams with ZOOM

Data type & scale

- 1 mill full text words / 1,500 images
- Data migration
- 650 semantic biographies
- 6 mill full text words / 12,000 images
- 1 mill images

Technology environment

- Machine learning
- Self-service digital imaging
- IIIF API Ver 2.0

Thanks to our one hundred and ninety-six contributors since 2012

We would like to recognise and thank our friends and supporters who have contributed to the MarineLives project, to the Maphackathon event, to the Early Modern Textiles, Garments and Dyestuffs glossary, and to the Signs of Literacy initiative, whether as volunteer transcribers, annotators, commentators, glossary contributors, software developers, advisors, interviewees, workshop participants, workshop and conference speakers, or as PhD Forum participants (in alphabetical order)

Dr Aquiles Alencar-Brayner	Thomas Davies	Viveka Hansen	M. L. Logue	r Benjamin Redding	Dr Samantha Thompson
Dr Kimberly Alexander	Jonathan Dent	Elaine Harrington	Jelle van Lottum	rmila Regulova	Roger Towner
Dr Aaron Allen	Melvyn Dresner	Dr Liam Haydon	Ismail Malik	ethan Reynolds	Dr Alexis Truax
Elizabeth Ames	Dr Stuart Dunn	Phillipa Hellawell	Grace Mallon	aniel Richards	Dr William Tullett
Dr Roberta Anderson	Professor Kai Eckert	Laura Herbert	Paula Marmor	arriet Richardson	Martin Turner
Vicky Annand	Bob Egan	Dr Helmer Helmers	Simon Marsh	ndrew Richens	Oliver Turner
Dr Carolyn Arena	Eglantine	Dave Henderson	Dr Alan Marshall	r Mia Ridge	David Underdown
Deborah Ashby	Dr Charlene Eska	Dr Stefan Hessbrüggen-Walter	Dr Barbara McGillivray	r Dominique Ritze	Rebecca Unsworth
Dr Gary Baker	Louise Falcini	Professor Tim Hitchcock	Dr Angela McShane	r Gavin Robinson	Dr Brodie Waddell
Gabor Bakos	Emilie-Jane Farrimond	Dr Philip Hnatkovich	Angela Middleton	en Robson	Rebecca Want
Iona Bartzi	Dr Karwah Fatah-Black	Rachel E. Holmes	John Miller	r Helen Rogers	Samuel Watson
Rachel Bates	Dr Amelia Fay	Jeffrey Hopper	Anne Mills	r Jill Rogers	Dominic Webb
Rowan Beentje	Dr Janet Few	Dr Jenni Hyde	Kate Morant	el Hee Rømer	Ken Whittaker
Nga Bellis-Phan	Christian Folgar	Dr Tom Irvine	Matthias Müller-Prove	eline Romano	Jill Wilcox
Michael Bennett	Sara Fox	Steve Ives	Professor Steve Murdoch	r Margaret Schotte	Peter Williams
Amber Berkeley	Dr Anthony Firth	Alex Jackson	Dr Shavana Musa	even Schrum	Royline Williams-Fontenelle
Dr Richard Blakemore	Erin French	Stefan Jäggi	Maurice Nicolson	ndrei Seleznov	Ad van der Zee
Lior Blum	Dr Ian Friel	Elin Jones	Stephanie Ostrich	ura Seymour	Dr Kathrin Zickermann
Ffion Boyd	Professor Cheryl Fury	Sue Jones	Frances Owen	hn Sheridan	Dr Suze Zijlstra
Katie Broke	Steve Garnett	Menno Jonker	Gordon O'Sullivan	r Deborah Sherlock	Dr Cäcilia Zirn
Dr James Brown	Dr Perry Gauci	Dr Sam Kaislaniemi	Dr Katherine Parker	r Rainer Simon	and the ever helpful but
Dr Andy Burn	Marja Geesink	Ross Keel	Tim Parry-Williams	a Sjoberg	anonymous @_mapnut
Elio Calcagno	Jaap Geraerts	Dr Patricia Keller	David Pashley	etr Sloup	
Joshua Calloway	Helen Good	William Kellett	Dr Michael Pearce	r Edmond Smith	
Michelle María Early Capistrán	Dr Jamie LH Goodall	Sara J Kerr	Dr Cathryn Pearce	lly Smith	
Rachel Carter	Guy Grannum	Heather Knight	Elizabeth Pimblett	r Stephen Snelders	
Cynthia Chin	Colin Greenstreet	Iris Kramer	Dr Sophie Pittman	ofessor Humphrey Southall	
Nicola Clarke	Francesca Greenstreet	Dr Marcin Krygier	Professor Simone Paolo	r Craig Spence	
Freyalyn Close-Hainsworth	Justin Grierson	John Kuhn	Ponzetto	aniel Stewart-Roberts	
Dalton Coker	Tracey E Griffiths	Dr Craig Lambert	Mark Ponte	r Hannah Steyne	
Dr Giovanni Colavizza	Dr Adam Grimshaw	John Layt	Dr Peter Pridal	n Stoll	
Dr Justin Colson	Dr Adam Grimshaw	Deborah Leem	Dr Jo Pugh	ichael Stoner	
Sam Cottrell	Karen Gunnell	Dr Sjoerd Levelt	Seif Al Rashidi	r Chad Stolper	
Kevin Daniells	Yerevag Hagopian	Amalia Skarlatou Levi	Dr Peter Rauxloh	eter Taylor	
Thierry Daunois	Dr Mark Hailwood	John Levin	Patrizia Rebulla	r Carl Thompson	
Dr John Davies	Finn Halligan	M. L. Logue			

MarineLives

Legal deposition

Deposition of Mark Harrison; mariner and master; resident in Wapping, Middlesex; age 27;
Dated September 21st 1659 (TNA, HCA 13/68, ff. 1r-3r)

Metadata

Signoff
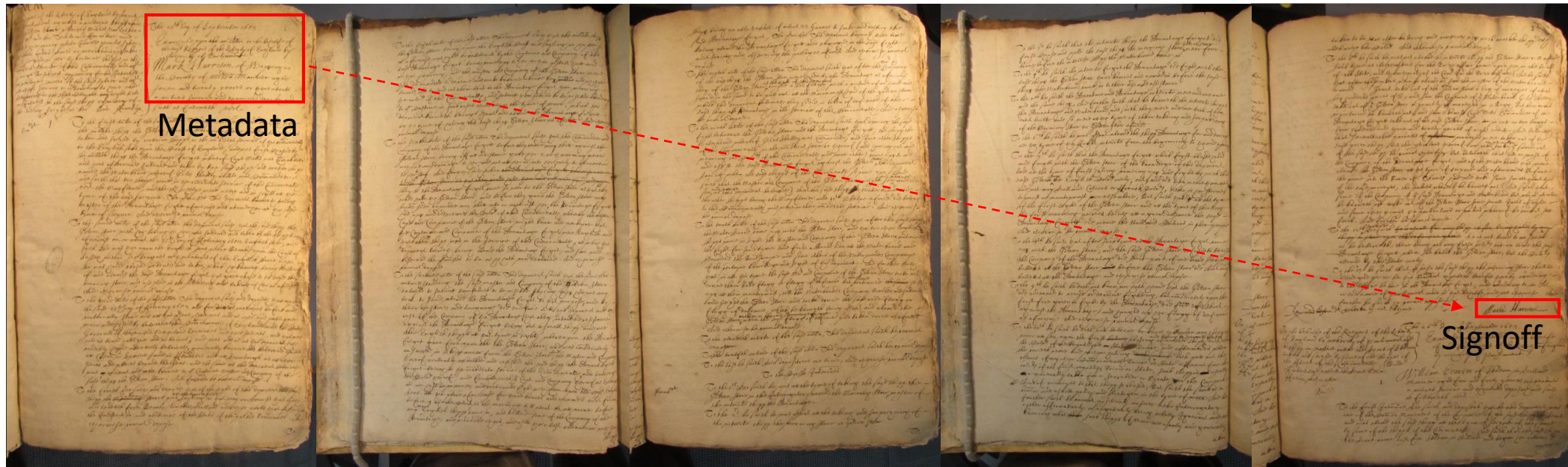
Early adopter of Transkribus

Transkribus recognition engine has been trained on Admiralty Court records

MarineLives experience of Transkribus

- Interface not easy for occasional volunteer users

- User interface challenges can be overcome with good facilitated training and supervision

- In the real world, a volunteer always has the choice of transcribing a page from scratch, for editing by a facilitator, or using Transkribus to generate a first pass, for revision by the volunteer and subsequent editing by the facilitator

- Facilitators go straight to the underlying digital image, not to the raw Transkribus data, and only rarely use statistical predictions by Transkribus of the likelihood of a specific letter or word

MarineLives thoughts on Transkribus

- Working with palaeographically trained volunteers

  - Use of Transkribus needs to be carefully built into volunteer and facilitator work processes
  - MarineLives volunteers need to be recruited and trained differently if using Transkribus
  - Use of Transkribus makes sense if volunteer transcribe is going to do twenty manuscript pages or more (= 10-15 hours work)

- Working with high volume volunteers (public historians, doctoral students)

  - Greater interest in technical support and readiness to learn system
  - But, temptation to transcribe from scratch as own technical competence builds up

MarineLives ideas for Transkribus

- Look at the culture, governance and data structures of other large scale image oriented initiatives, e.g. IIIF; Pelagios/Recogito

- Actively share kernels and algorithms for reuse, e.g. Kaggle

- Publish timely public statistics for Transkribus data universe (manuscripts bu period, by writing style, closed vs open corpora, error rates by HTR engine for different corpora...)

- Provide guidance on planned future development resource & development priorities beyond 2019

- Encourage development of facilitated user groups by period and/or handwriting type

- Address interface issues, e.g. labelling of text areas bty category; naming or numbering of pages to enhance human discoverability

MarineLives plans using Transkribus in 2019 & beyond

- Increase size of ground base to 6 million words and explore impact on accuracy of full text transcription [part of Signs of Literacy initiative]

- Investigate key word spotting for untranscribed pages as a tool to improve search [part of Signs of Literacy initiative]

- Explore the interface being created by the Stadsarchief Amsterdam with their digitisation partner Picturae

Signs of literacy initiative

Pattern seeking

How many different letters can you recognise?
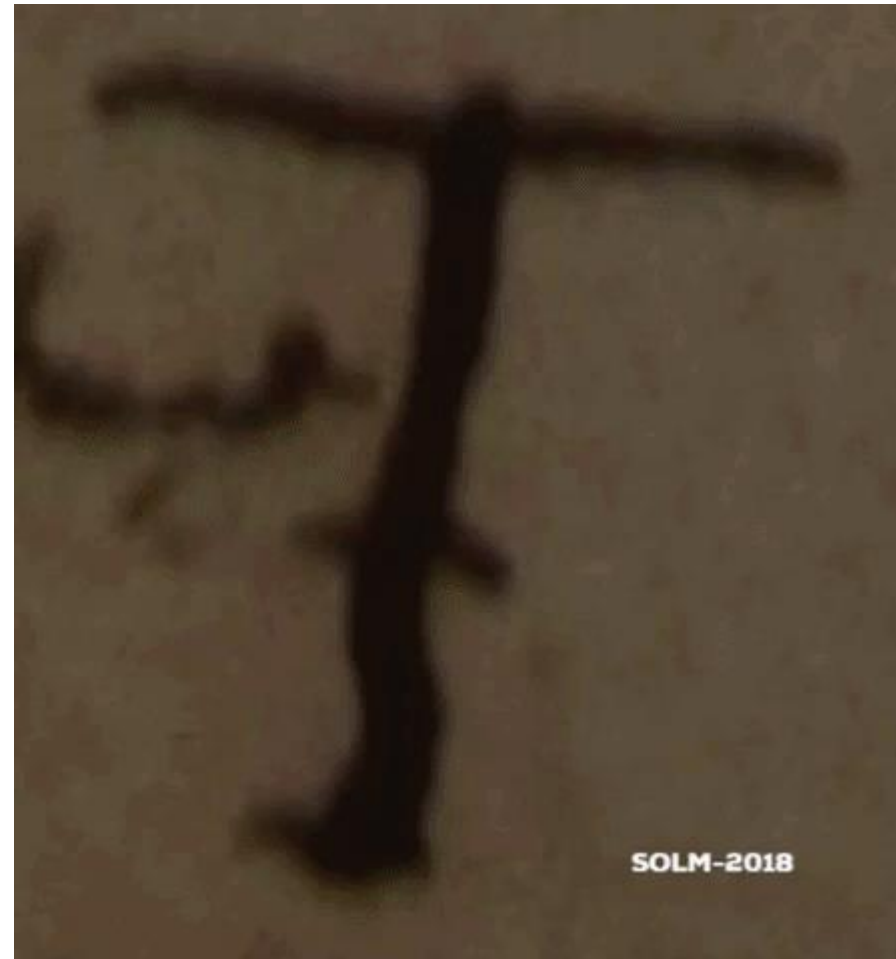
31 images of initials in the SOLM-2018 database
Animated GIF viewable only in Powerpoint

Source: SOLM-2018 database

Initials – Ts and Js, and a few others thrown in

T6.png   T7.png   T8.png   T9.png   T10.png   J5.PNG   J6.PNG   J7.PNG   J8.PNG   J9.PNG

T11.PNG   T12.PNG   T13.PNG   T14.PNG   T15.PNG   J10.PNG   J11.PNG   J12.PNG   J13.PNG   J14.PNG

T16.PNG   T17.PNG   T18.PNG   T20.PNG   T21.PNG   J15.PNG   J16.PNG   J17.PNG   J18.PNG   J19.PNG

T22.PNG   T23.PNG   T24.PNG   T25.PNG   T26.PNG   J20.PNG   J21.PNG   J22.PNG   J24.PNG   J25.PNG

T28.PNG   TJ28.PNG   TJ30.PNG   TJ31.PNG   TL29.PNG   J26.PNG   J27.PNG   J28.PNG   J30.PNG   J31.PNG

Source: SOLM-2018 database

Signs of Literacy initiative objectives

- Develop tools and processes for increasingly automated metadata extraction and linkage from digitised handwritten manuscript pages

- Develop the infrastructure to deliver a highly scaleable data set of one million manuscript pages containing signoffs and front matter

- Stimulate the development of machine learning capability in multiple archival centres of excellence

- Build interest amongst historians and linguists in machine learning applications to support close reading and distant reading approaches to large scale data sets

- Answer important research questions in field of historical literacy

Proposed structure of Signs of Literacy initiative workstreams
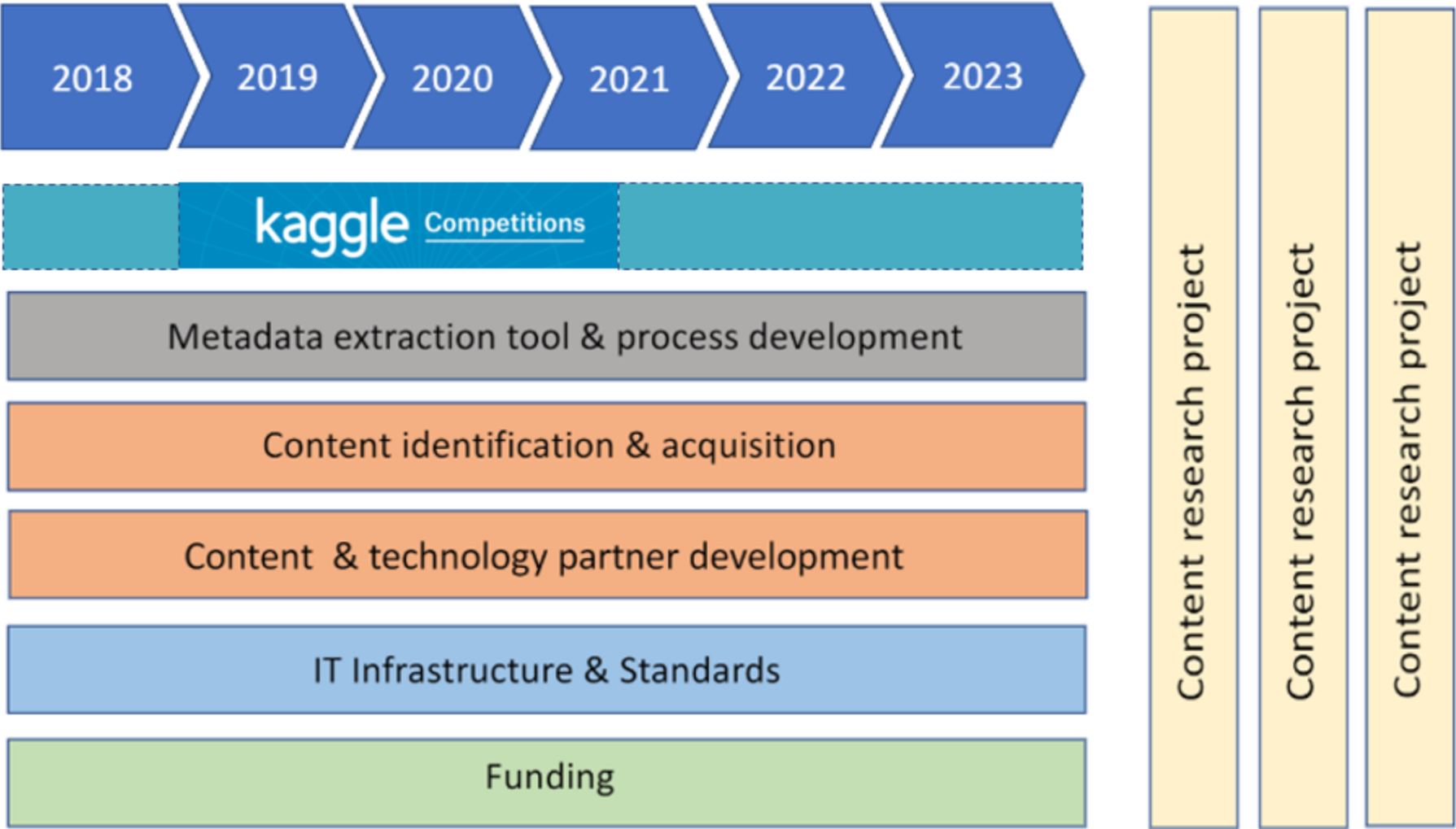
Collective intelligence

Help us build a bot

Volunteers	ArchiveBot
Identify manuscript page with signoff	Machine learning
Establish type of signoff [marke, initial(s), signature]	Machine learning
Establish linked front matter of deposition	Machine learning
Establish foliation & image record numbers of signoff & front matter	Machine learning + database interrogation via API
Establish name of deponent & age	HTR/Key word detection + NLP
Establish residence of deponent	HTR/Key word detection + NLP
Establish occupation of deponent from front matter	HTR/Key word detection + NLP
Establish further occupational data from full text of manuscript	HTR/Key word detection + NLP
Data mine full text for social & economic parameters	Key word/phrase detection + proximity analysis

Facilitated archival volunteer teams

Contact the Signs of Literacy team to learn more

We are exploring a range of potential technical partners

Applying machine learning to Signs of Literacy

There is a better way...

iiif

Recogito | AN INITIATIVE OF Pelagios commons

Signs of Literacy Kaggle Research Competition, 2018
Colin Greenstreet on LinkedIn
April 30, 2018
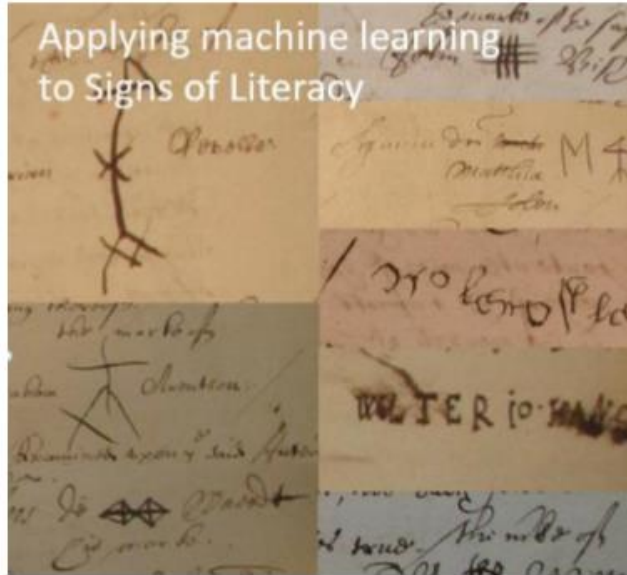
Feature extraction: blots, smudges, stylistic features, & deletions
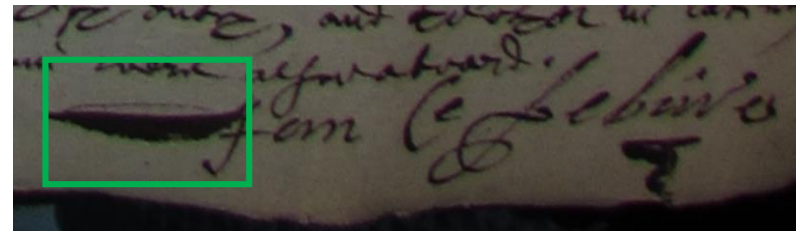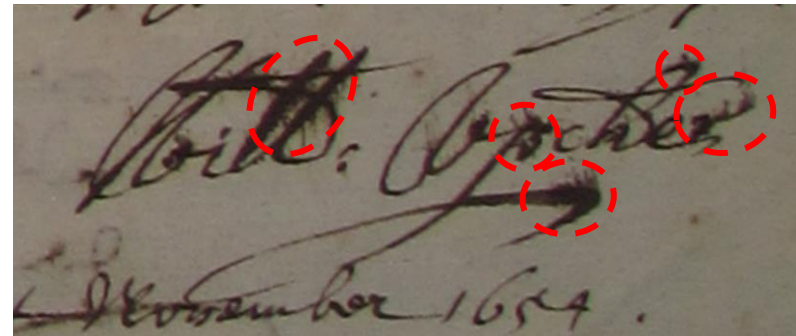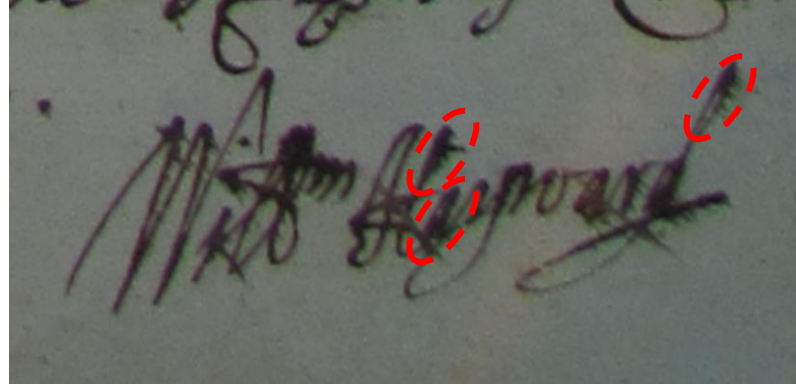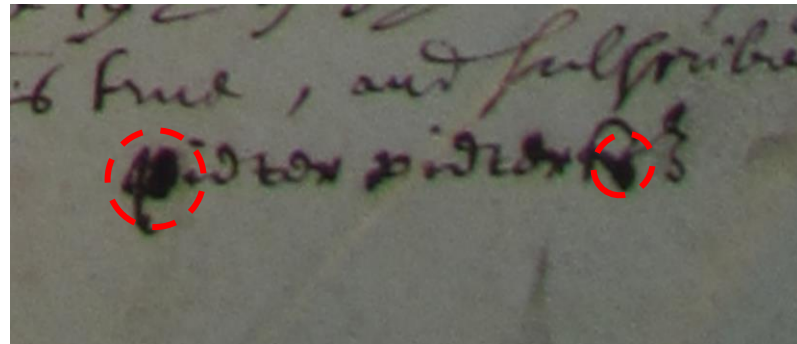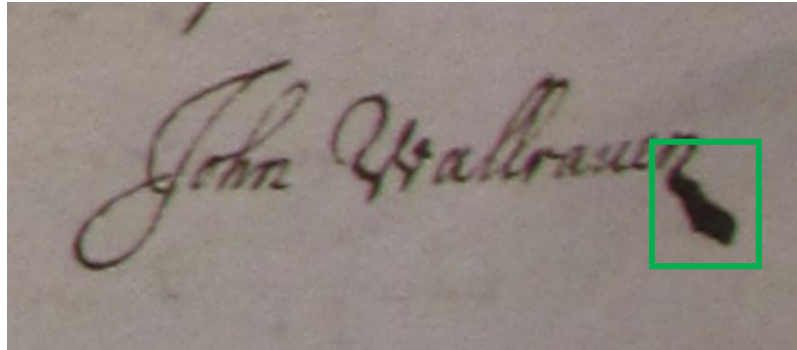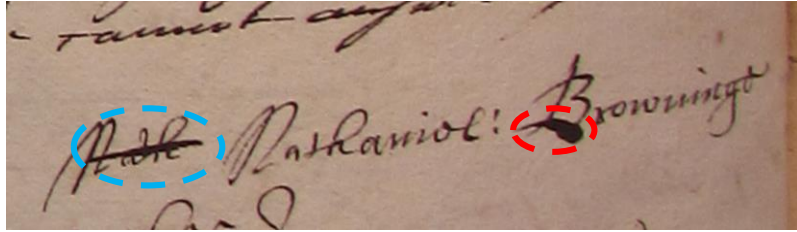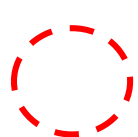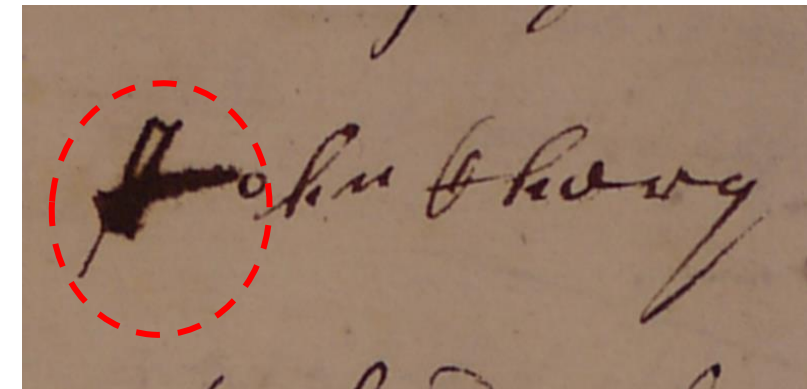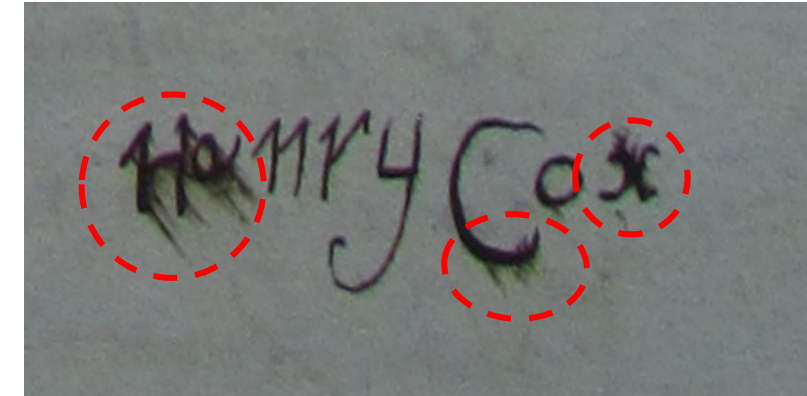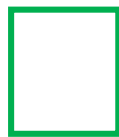
Ink blots or smudges | Stylistic feature or smudge? | Deletion

Feature extraction: Detection of "shake" in straight and curved lines
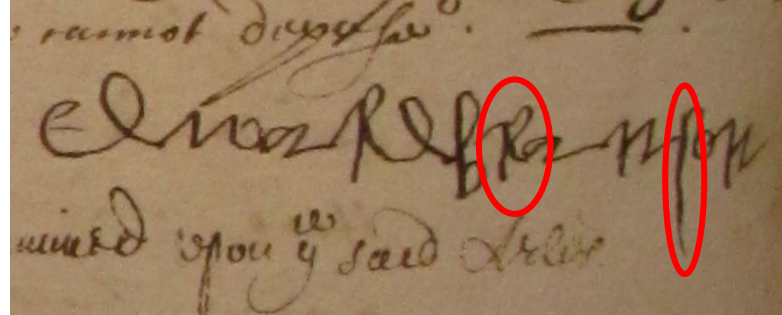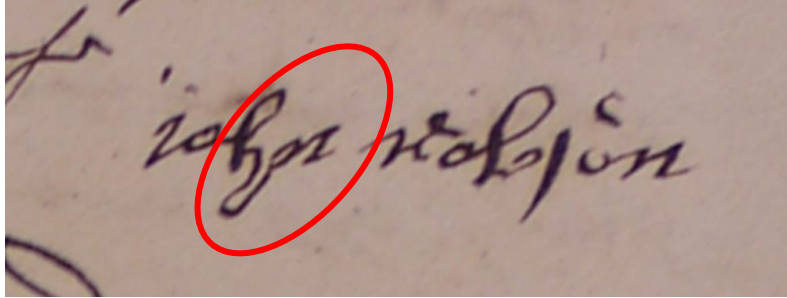
HYPOTHESIS: Shaky lines may be a sign of poor signature execution (and by inference, poor handwriting execution) suggesting either lower level of literacy than smooth executed lines, or the effect of illness or age
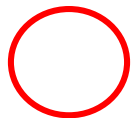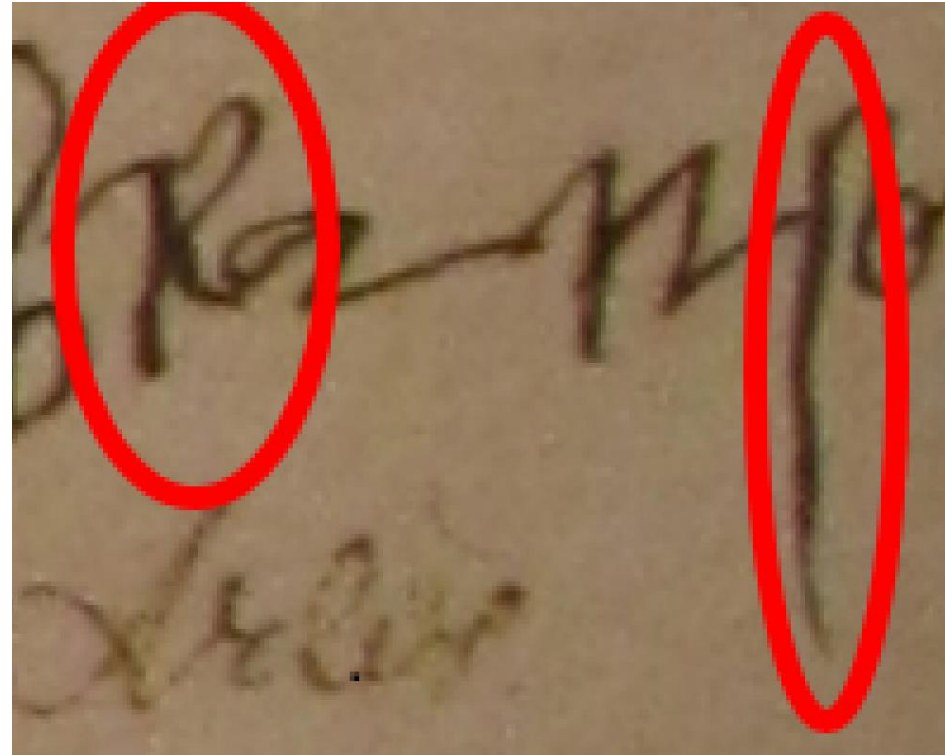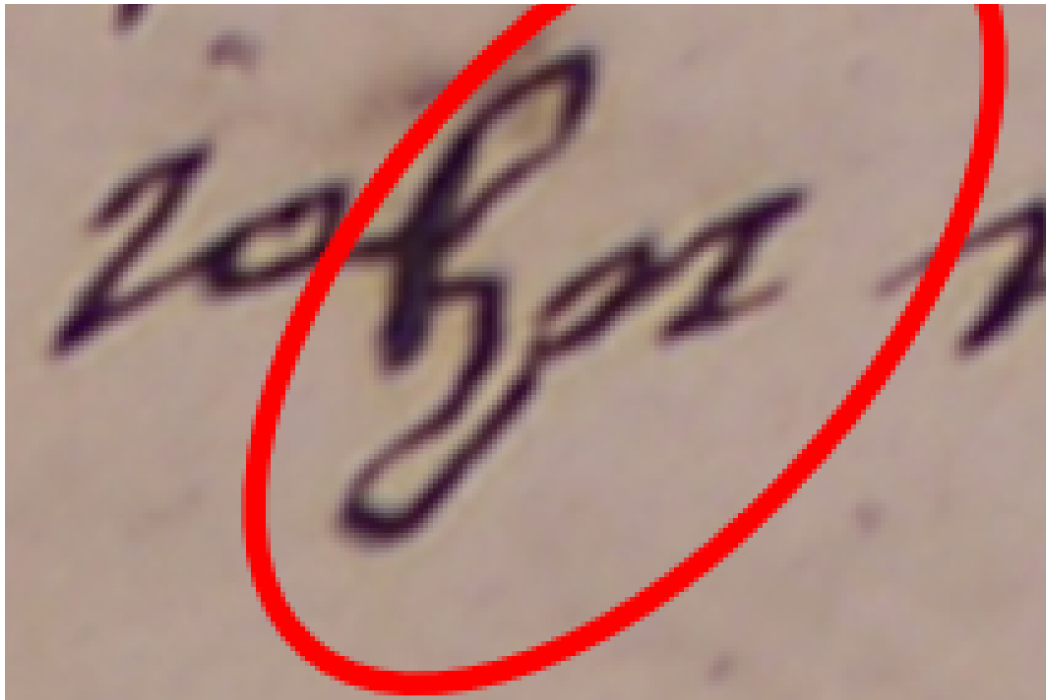
Shaky straight lines and/or loops

Clockwise, from top LH: KaggleTestSnippet_HCA_1371_f.435v.PNG, KaggleTestSnippet_HCA_1368_f.483v.PNG, KaggleTestSnippet_HCA_1368_f.483v_PIXELS.PNG, KaggleTestSnippet_HCA_1371_f.435v.PNG_PIXELS.PNG

Close reading using IIIF anchors manifest in Mirador viewer
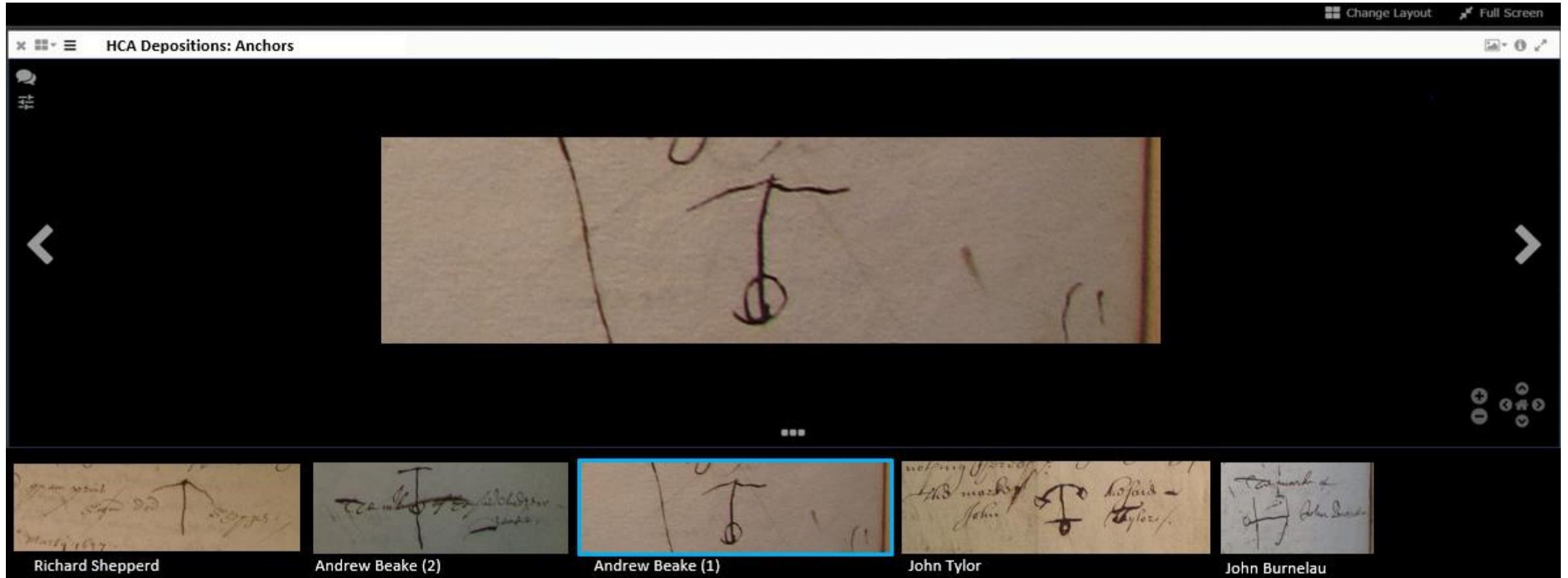
Mockup of a IIIF manifest in Mirador viewer, using http://projectmirador.org/demo/ ;
http://www.marinelives.org/wiki/HCA_13/70_f.252v_Annotate

Planned use of AI

- Scale of targeted Signs of Literacy data set is 1 million manuscript pages vs. current 12,000 MarineLives manuscript pages

- Designed from get go as an exercise in collaborative intelligence, with strong focus on machine-human interface, expecting the role of machine and human to change over time

- Volunteers will be recruited explicitly for their interest in working with machines, rather than for their interests in palaeography and/or history

- Explicit measurement of user satisfaction & productivity as we introduce new tools to support volunteers in their process steps and/or automate specific process steps, allowing volunteers to concentrate on other parts of the overall metadata creation and linkage process

Our interest in Transkribus tools

- Text area identification

- Lineation and interlineation

- Key word spotting

- [HTR]

- [Upload of phone images with hand input document metadata to a IIIF database]

Contact details

Colin Greenstreet
Community organiser,
Signs of Literacy

Dr Mark Hailwood
Lecturer in History, 1400-1700,
University of Bristol

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

Working with: The National Archives   kaggle Competitions   iiif   PICTURAE LET'S UNLOCK OUR HERITAGE TO THE WORLD
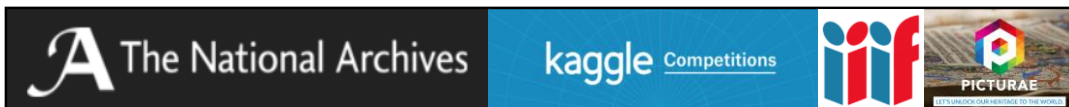
BACKUP
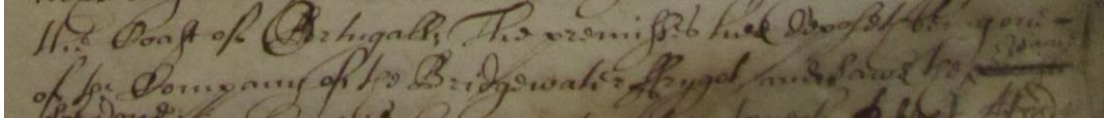
Imagine an ArchiveBot extracting metadata automatically from handwritten manuscripts and working with volunteers to finalise

HCA 13/63 f.471r

METADATA
HCA 13/63 f.471r

SIGNOFF
HCA 13/63 f.471v

SIGNOFF DETAIL
HCA 13/63 f.471v
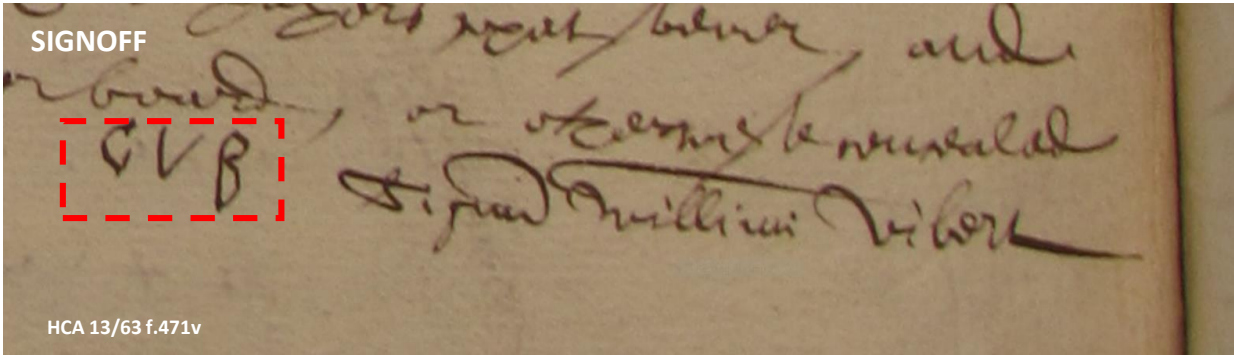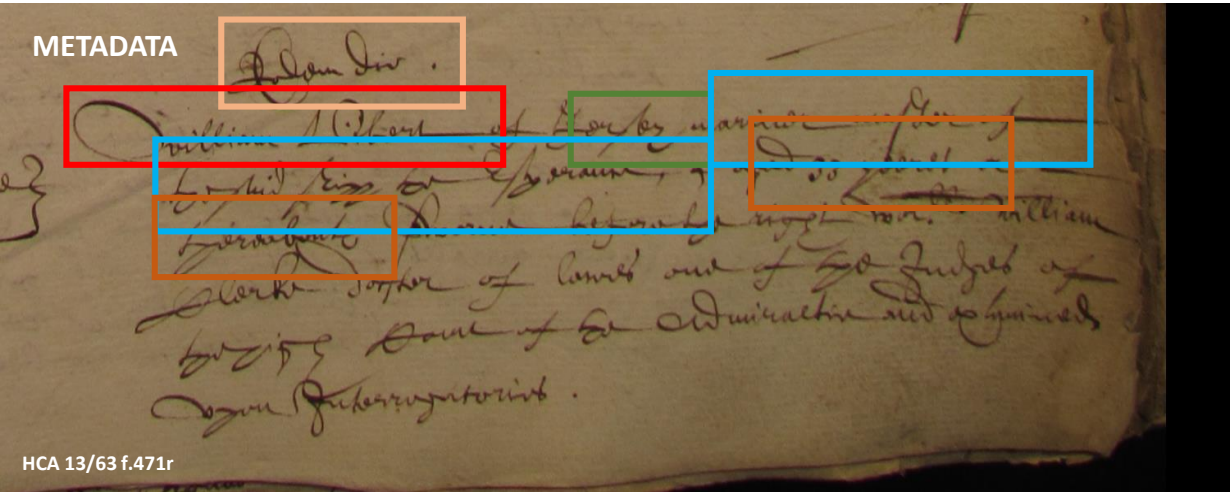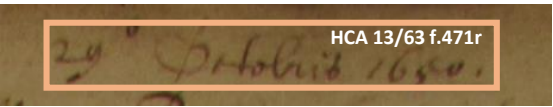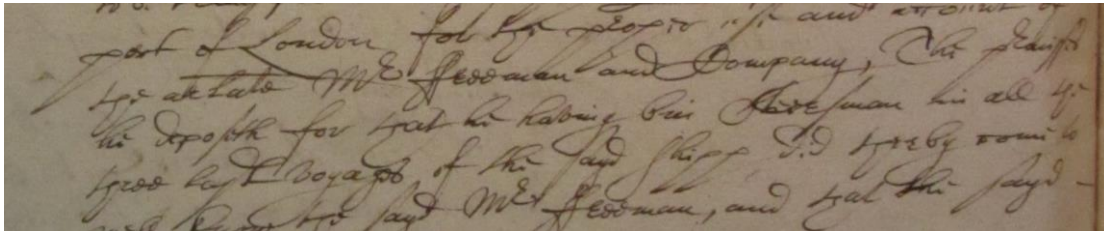
	Date		Occupation
	Name		Age
	Residence		Signoff

Can we use key word spotting to excavate raw metadata?

LANGUAGE DENOTING OCCUPATION

"The premisses hee deposeth being one of the company of the *Bridgewater ffrygott*, and sawe the same soe done" [HCA 13/72 f.90r] [CONCLUSION: One of the company]

"The premisses he deposeth for that he the deponent was not onely for the voyage arlate wherein she was stranded, but in two former voyages stiersman of the sayd ship" [HCA 13/72 f.90v] [CONCLUSION: Steersman]

"after such buying of the said shipp the *Santa Hellen* by the said da [?Groots] and company aforesaid, namely in or about the moneth of October 1643 (about three monethes after the adiudication aforesaid) being in possession of the said vessell, put and constituted this deponent master of her, and hee then entred master upon her at Port Lewes in Bretany" [HCA 13/72 f.95r] [CONCLUSION: Master]

"all which hee knoweth being masters mate of the said shipp the said voyage" [HCA 13/70 f.669v] [CONCLUSION: Master's mate]

"hee knoweth the premisses because hee was boatswaines mate of the said shipp and went all the said voyage in her" [HCA 13/70 f.671r] [CONCLUSION: Boatswain's mate]

Can we refine raw machine generated metadata using a combination of NPL, controlled vocabularies, and programmable decision rules?

LANGUAGE DENOTING OCCUPATION

"The premisses hee deposeth being one of the company of the *Bridgewater ffrygott*, and sawe the same soe done" [HCA 13/72 f.90r] [CONCLUSION: One of the company]

"The premisses he deposeth for that he the deponent was not onely for the voyage arlate wherein she was stranded, but in two former voyages stiersman of the sayd ship" [HCA 13/72 f.90v] [CONCLUSION: Steersman]

"after such buying of the said shipp the *Santa Hellen* by the said da [?Groots] and company aforesaid, namely in or about the moneth of October 1643 (about three monethes after the adiudication aforesaid) being in possession of the said vessell, put and constituted this deponent master of her, and hee then entred master upon her at Port Lewes in Bretany" [HCA 13/72 f.95r] [CONCLUSION: Master]

"all which hee knoweth being masters mate of the said shipp the said voyage" [HCA 13/70 f.669v] [CONCLUSION: Master's mate]

"the premisses because hee was boatswaines mate of the said shipp and went all the said voyage in her" [HCA 13/70 f.671r] [CONCLUSION: Boatswain's mate]

# Folio 4 recto in volume HCA 13/124 of personal answers has a signature as a "signoff"

Signature of 'Wm Ashwell' (William Ashwell), who gave personal answer in the High Court of Admiralty, Nov. 1650, (HCA 13/124 f.4r)

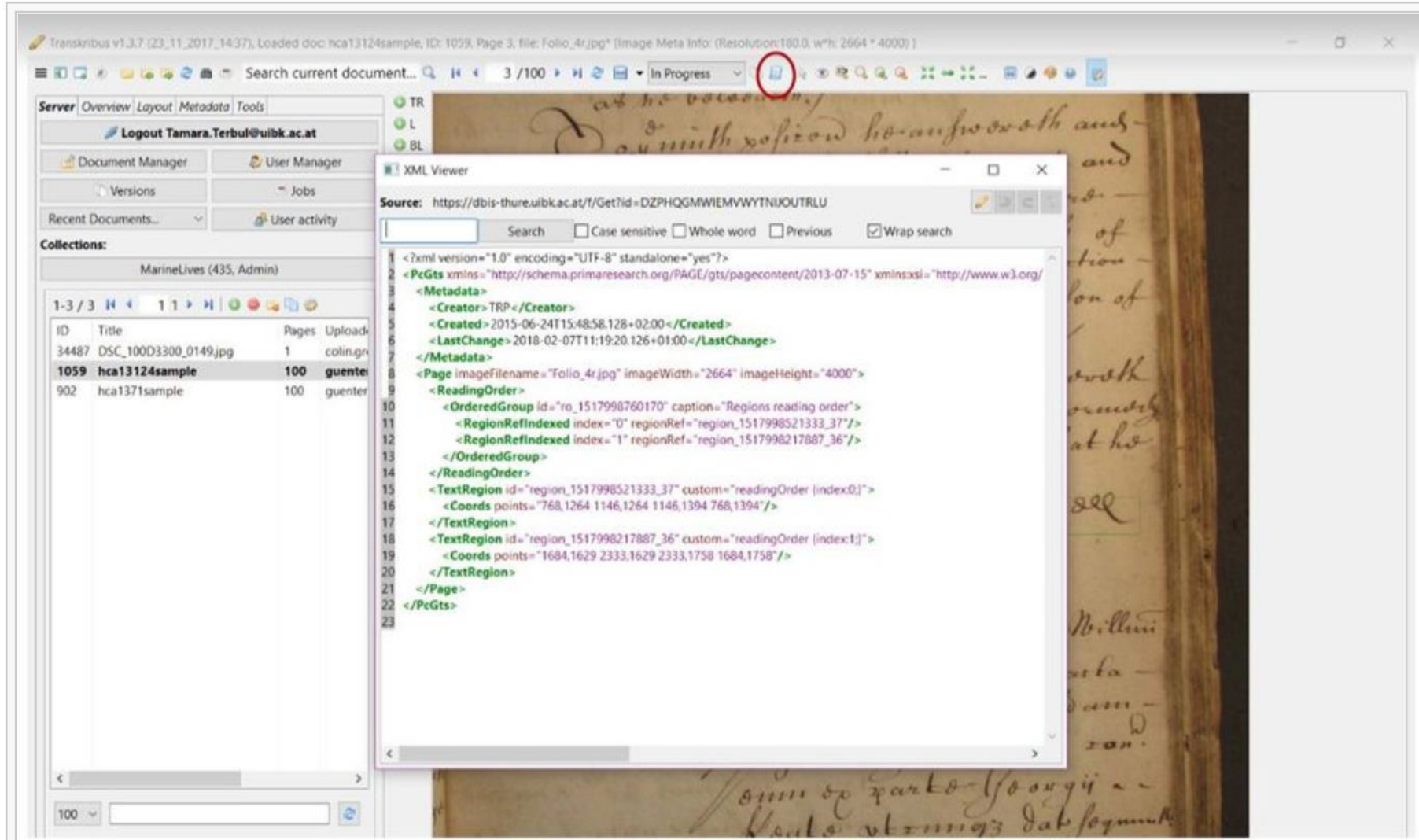# This folio is located within Transkribus document ID 1059, labelled as page 3 (out of 100)
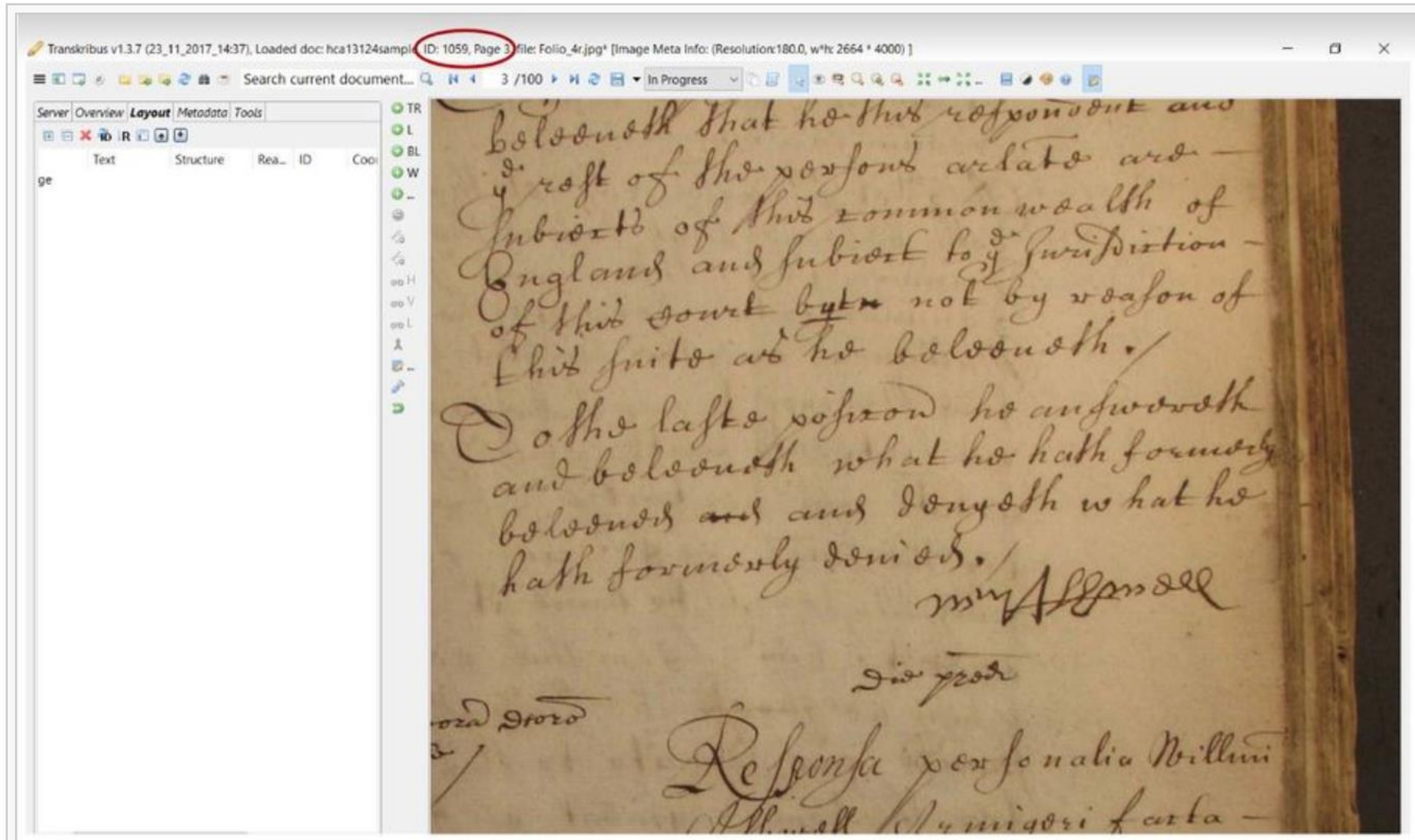
HCA 13/124 f.4r is located within a Transkribus Document ID 1059, and has been labelled page 3 by Transkribus out of 100 pages in this Transkribus document

Hand or automatic identification and labelling of signoff with a boundary box
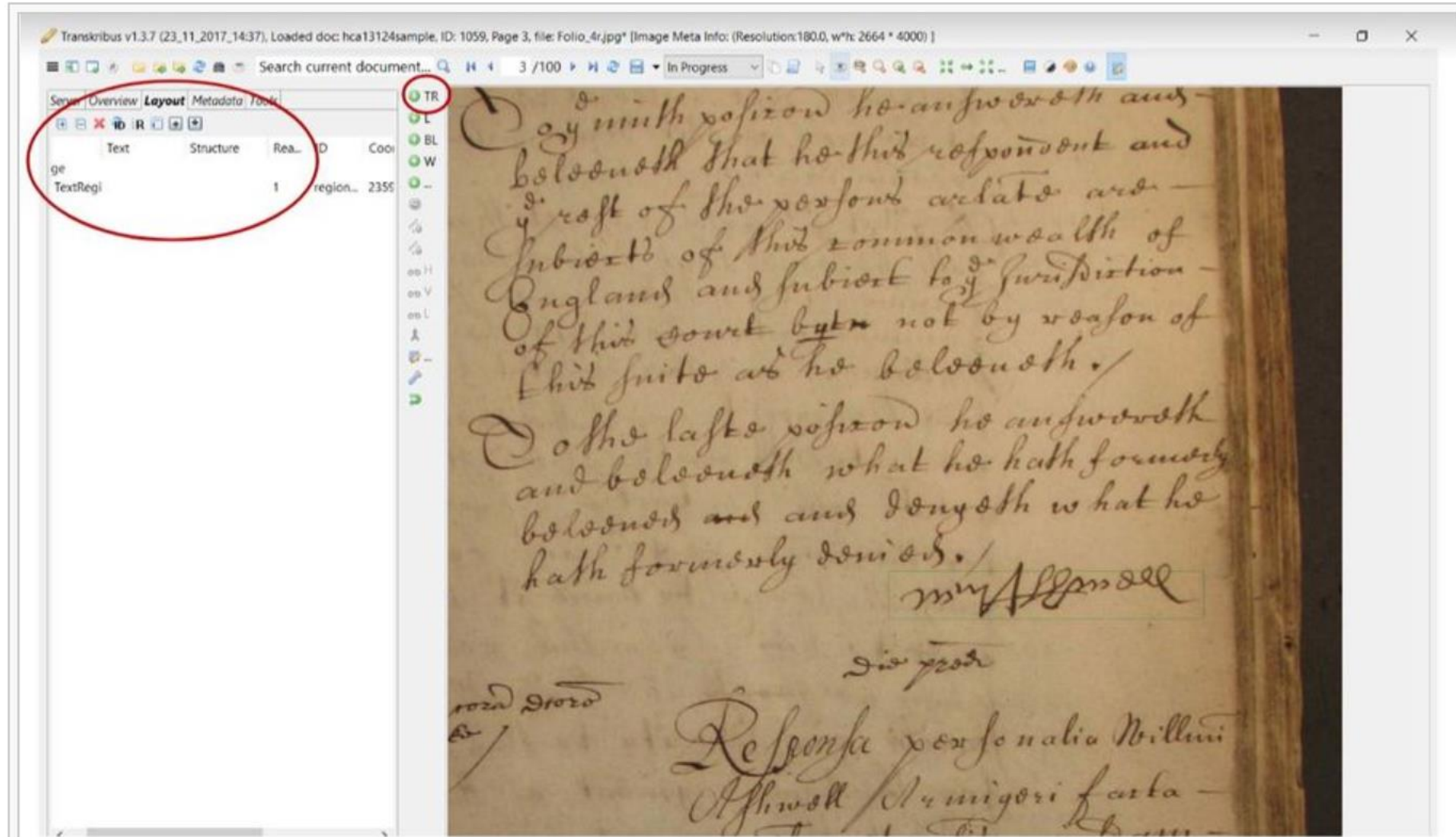
MarineLives will hand identify one or more signatures and/or markes on a page with one or more rectangular 'Text Region' boxes

XML data on location of signoff

XML data for each page containing a signature or marke can be inspected and processed by researchers, and linked (if desired) to metadata on name, place of residence, occupation, age, and date of deposition, through the Excel spreadsheet described above

Some early results

Analysis of signatures by groups of 10 signoffs organised by occupational type to assess consistency of data and effect of sample size in SOLM-2018 data set
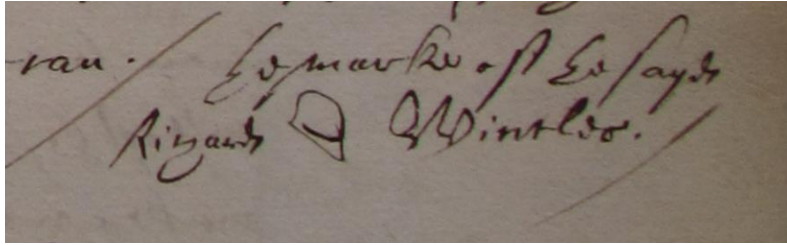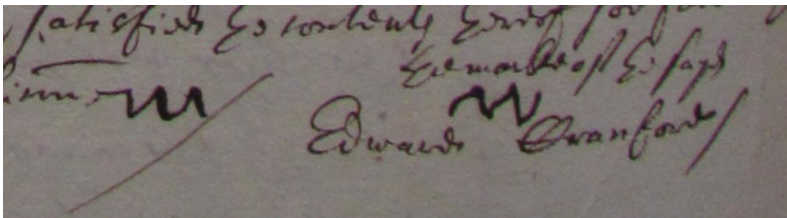Data as of 27/08/2018 & 29/08/2018, n = 760

Source: SOLM-2018 database

SOLM-2018 data set: Number of signoffs by hamlet & parish to east & south east of city of London (1592-1683)

Source: SOLM database
n = 1545
Data set as of 03/09/2018

X% — Hamlets & parishes with percentage using signature

Y% — Hamlets & parishes with lower percentage using signature

Some early results

35

Bubble labels:

- 100% — 1 — Bethnal Green
- 100% — 2 — Mile End
- 78% — Saint Mary Matsellon Whitechapel — 54
- East Smithfield — 27 — 67%
- Saint Botolph without Algate — 8 — 63%
- 77% — Wapping — 346
- Precinct of Saint Catherine near the Tower of London — 85 — 78%
- 80% — Ratcliff — 170
- 80% — Shadwell — 113
- 100% — 2 — Saint Pauls Shadwell — 67%
- 80% — Limehouse — 162
- 100% — 2 — Bromley near Bowe
- 1 — 0% — Saint Leonard Bromley
- Stepney [no hamlet details] — 110 — 64% — 64%
- Blackwall — 14
- Wapping Wall — 26 — 69%
- 73% — Saint Saviours Southwark — 22
- 1 — Saint George Southwark
- Saint Olave Southwark — 72 — 71%
- 72% — Horsey Downe — 29
- 71% — Saint Mary Magdalen Bermondsey — 48
- Redrith [Rotherhithe] — 153 — 86%
- 89% — Poplar — 27
- Deptford — 40 — 77%
- 45% — Greenwich — 33

75% of Admiralty Court deponents outside City of London to south and east are north of the river

North of river =1147
South of river = 398
Parish of Stepney = 63%
[assigning all Wapping to Stepney]

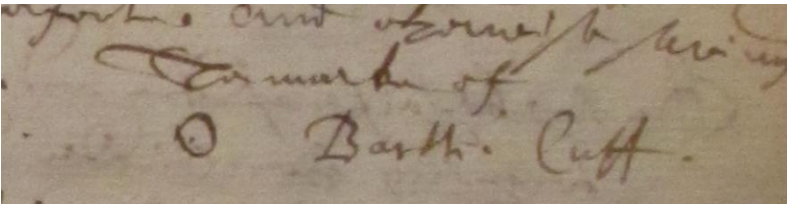# Porters handling coals, whale oil, ginger & corn

Richard Wincles, thirty-three year old porter, of the parish of Stepney, Middlesex, Dec. 15, 1656; employed as a labourer with fifteen other men to unload coals from the *Imployment* moored near Execution Dock, Wapping, into lighters for fixed rate of 12 *s* per man (HCA 13/70 f.554r)
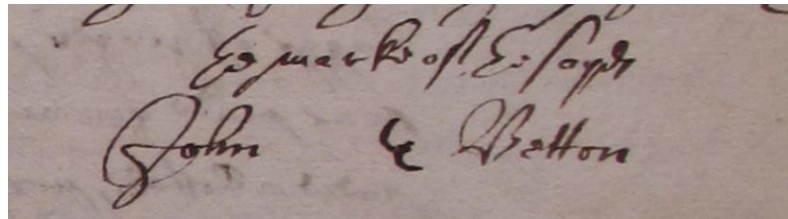

Edward Cranford, forty-four year old coale heaver or porter, of the parish of Stepney, Middlesex, Dec. 15, 1656; employed as a labourer with fifteen other men to unload coals from the *Imployment* moored near Execution Dock, Wapping, into lighters for fixed rate of 12 *s* per man (HCA 13/70 f.555v)

Bartholomew Cuff, sixty year old porter of the Stillyard, of the parish of Allhallowes the Greate, London, May 15, 1658; assisted in the landing of whale oil from lighters at the Stillyard Key and loading them away into a warehouse (HCA 13/70 f.555v)
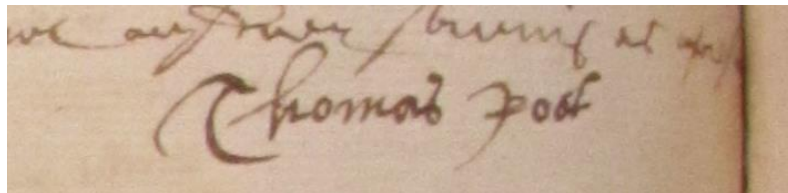
John Betton, fifty-four year old citizen and white baker of London, of the parish of Saint Buttolph Algate, London, Jul. 31, 1655; self-described as a porter employed by the Commissioners for Prize Goods to deliver ginger from a warehouse at Ralphes Key (HCA 13/70 f.449r)
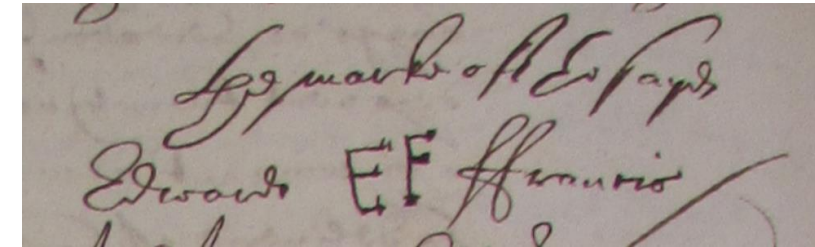

Edward Sherwin, fifty-six year old cittizen and leatherseller, of the parish of Little Allhallowes, London, Jul. 31, 1655; self-described as a porter employed by the Commissioners for Prize Goods to deliver ginger from a warehouse at Ralphes Key (HCA 13/70 f.449v)
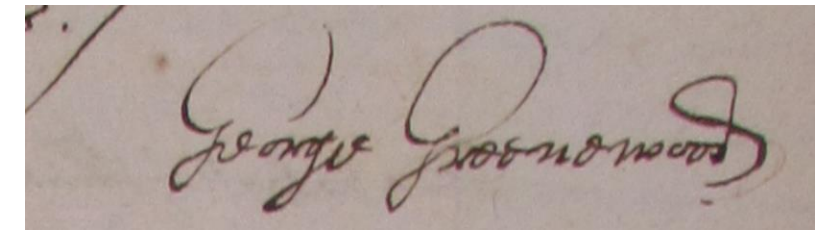

Thomas Roots, twenty-nine year old porter, of the parish of Greate Allhallowes, London, May 15, 1658; assisted in the landing of whale oil from lighters at the Stillyard Key, as one of the Stillyard porters, and loading them away into a warehouse (HCA 13/72 f.330v)
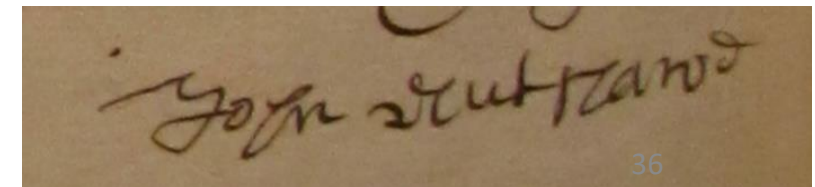

Edward ffrancis, citizen and merchant taylor of London, of the parish of Saint Olave in Southwarke, Jul. 31, 1655; self-described as a porter employed by the Commissioners for Prize Goods to deliver ginger from a warehouse at Ralphes Key (HCA 13/70 f.450v)

George Greenwood, thirty year old citizen and vintner of London, of the parish of Saint Buttolph Bishopsgate, London, Jul. 31, 1655; self-described as a porter employed by the Commissioners for Prize Goods to deliver ginger from a warehouse at Ralphes Key (HCA 13/70 f.454r)

John Nutshall, fifty-five year old corne porter, of the parish of Saint Saviours Southwarke, Nov. 19, 1653; employed with a barber chyrurgeon/corne meter, an additional corne-meter, and other labourers to unlade a cargo of what in the *ffortune* of Stettin, moored against Limehouse; eight years of experience as a corne porter (HCA 13/70 f.352v)

We need visual metadata, which can be machine processed

Table 1.2a EXPANDED: HCA 13/53 [f.1r-340v] - Signoff frequency per manuscript page, data from 1637

Source: SOLM-2018 database