Signs of Literacy initiative

Help us build a bot

ArchiveBots, Ver 2.0

Tuesday, 18th September, 2018

Help us build a bot

Build a quick and dirty Archival Bot in 50 days

Work has started

Establish a process for machine supported volunteer extraction & synthesis of metadata for the SOLM-2018 machine learning training data set of markes, initials and signatures

Share results at the Transkribus users conference in Vienna, November 8th-9th, 2018

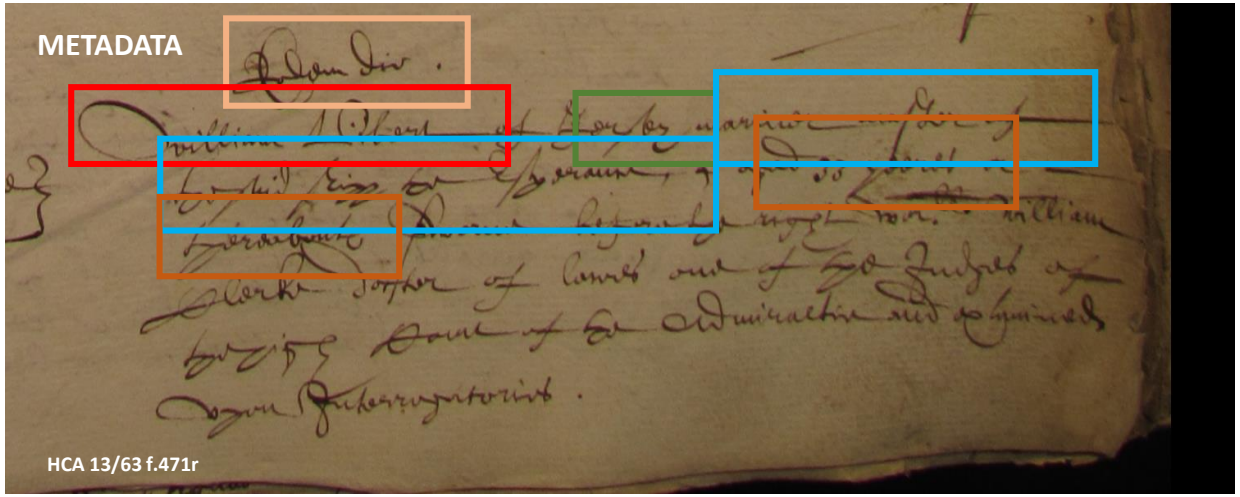
- Identify manuscript page with signoff
- Establish type of signoff [marke, initial(s), signature]
- Establish linked front matter of deposition
- Establish foliation & image record numbers of signoff & front matter
- Establish name of deponent & age
- Establish residence of deponent
- Establish occupation of deponent from front matter
- Establish further occupational data from full text of manuscript

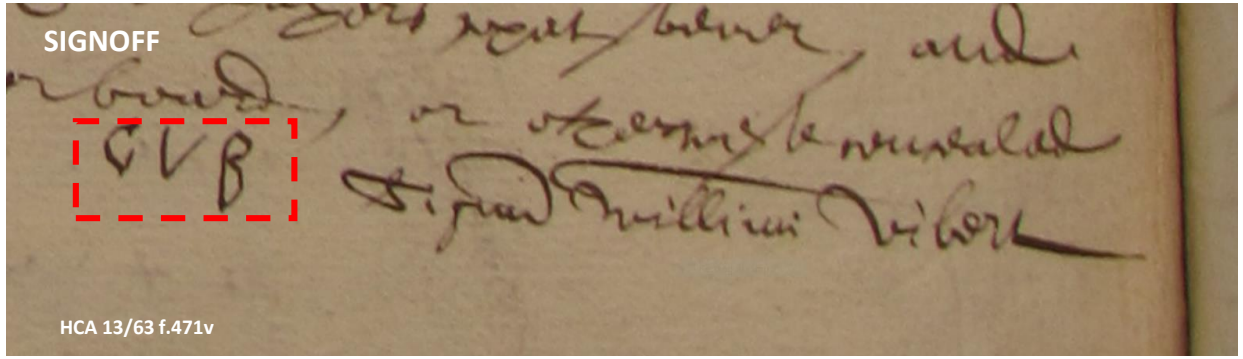
We are looking for a small number of volunteers to work with us as a facilitated remote team – public historians, who can read cursive script, coders & NLP experts - to create a quick and dirty bot as a Proof of Concept to show at the Transkribus Users Conference

Imagine an ArchiveBot

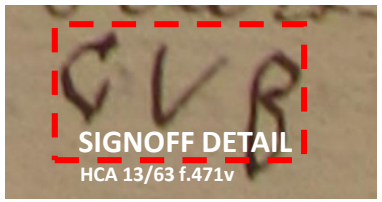
Imagine an ArchiveBot extracting metadata automatically from handwritten manuscripts and working with volunteers to finalise

Signs of Literacy initiative





Date
Name
Residence
Occupation
Age
Signoff



William Vibert (Guillaume Vibert) was a thirty-three year old mariner from the isle of Jersey, who was master of the ship the *Esperansa*.

He signed his deposition in the English High Court of Admiralty on October 29th 1650, using three not two initials.

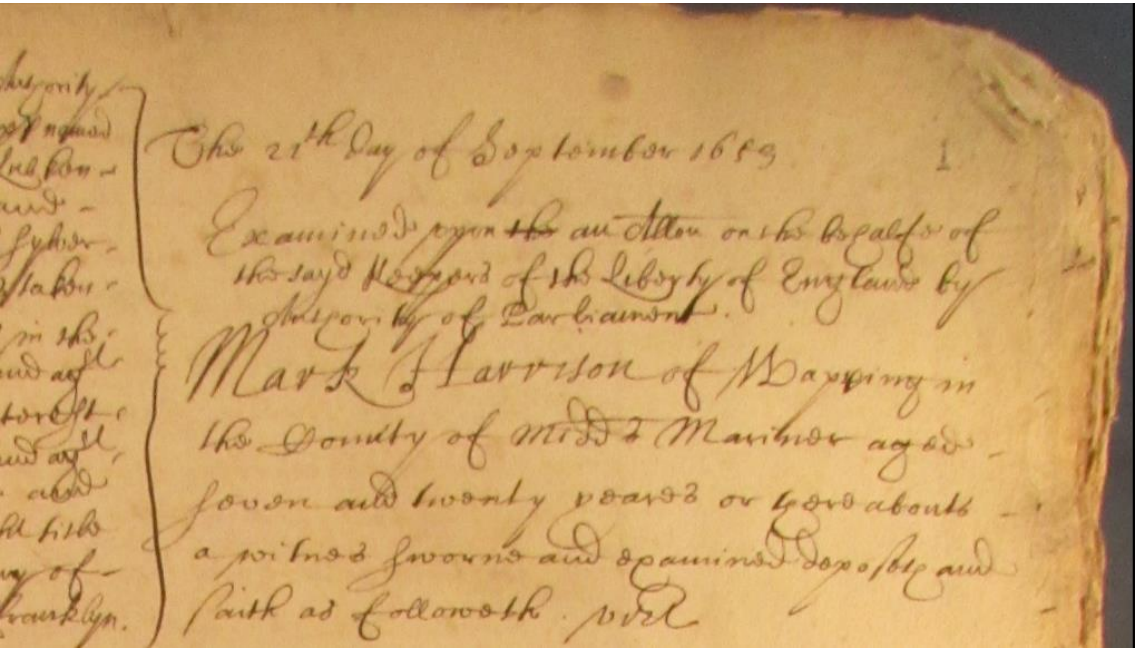
The choice of "G", "V" and "B" for initials suggests he thought of himself and pronounced his name as "Guillaume Vee Bert" (though he could not write a full signature)

Challenge for a Bot

(1) Find the date of the deposition (it is at the top of the same manuscript page above the preceding deposition)
(2) Deal with the interlineation in the metadata
(3) On the next manuscript page, recognise "GVB" are initials not a marke. The "G" is quite tough (easy to mistake for "C"). Yes, they are recognisable letters, but the three don't match the two names "William Vibert". Need to know William – Guillaume in French, and that mariners in Jersey may be born in France and French speaking.

Source: http://www.marinelives.org/wiki/HCA_13/63_f.471r_Annotate; http://www.marinelives.org/wiki/HCA_13/63_f.471v_Annotate

Speech to text recognition



Watson Speech to Text / Speech to Text Demo

Speech to Text

The IBM Watson Speech to Text service uses speech recognition capabilities to convert Arabic, English, Spanish, French, Brazilian Portuguese, Japanese, Korean, and Mandarin speech into text.

Get Started | API Reference | Documentation | Fork on GitHub | Start for free in IBM Cloud

Voice Model:

GB English broadband model (16KHz)

mark Harris[2] and of[4] what happened[7] in[2] the county of Middlesex mariner[8] aged seven and twenty years

mark Harrison[3] of[2] walking[8] in[2] the county of Middlesex mariner[15] aged seven and twenty years or[2] there[4] about

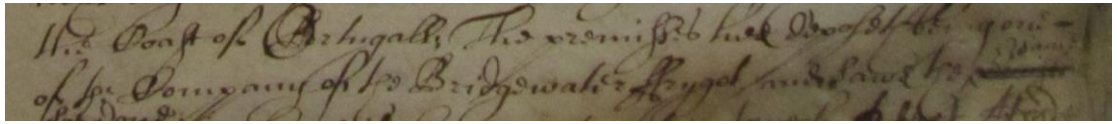
mark[2] Harrison[3] of[2] what[4] happened in the county of Middlesex mariner[10] aged seven and twenty years or[4] there[4]

mark[2] Harrison[3] or[3] walking in the county of Middlesex mariner[8] aged[2] seven and twenty years[3] or[2] the[3] about[5]

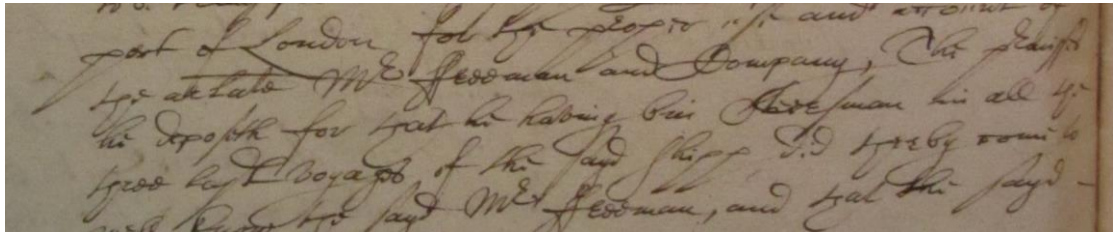
Can we use key word spotting to excavate raw metadata?

Signs of Literacy initiative

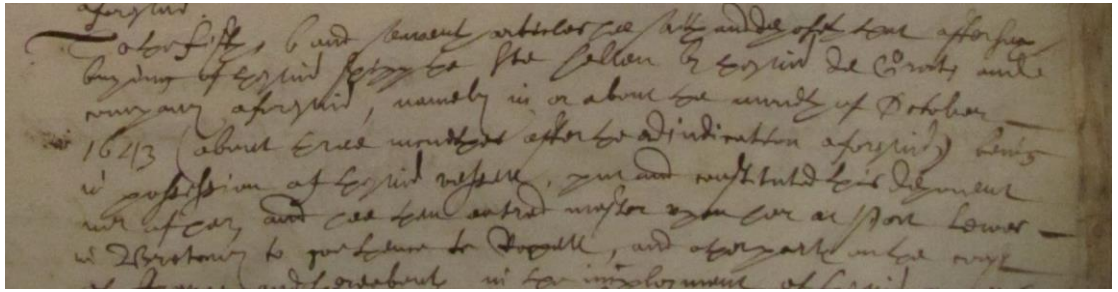
LANGUAGE DENOTING OCCUPATION



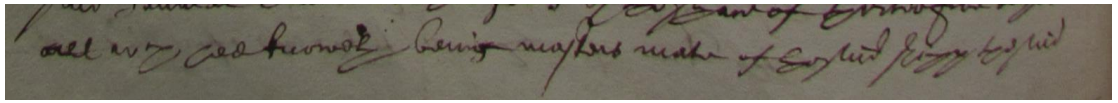
"The premisses hee deposeth being one of the company of the *Bridgewater ffrygott*, and sawe the same soe done" [HCA 13/72 f.90r] [CONCLUSION: One of the company]



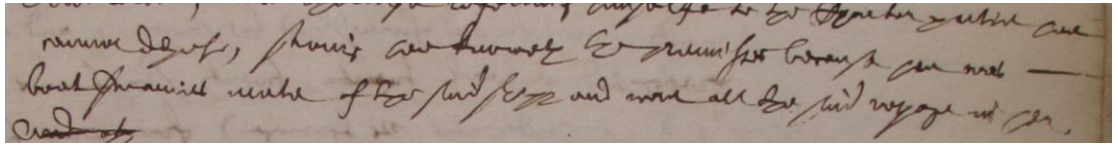
"The premisses he deposeth for that he the deponent was not onely for the voyage arlate wherein she was stranded, but in two former voyages stiersman of the sayd ship" [HCA 13/72 f.90v] [CONCLUSION: Steersman]



"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?

Signs of Literacy initiative

LANGUAGE DENOTING OCCUPATION

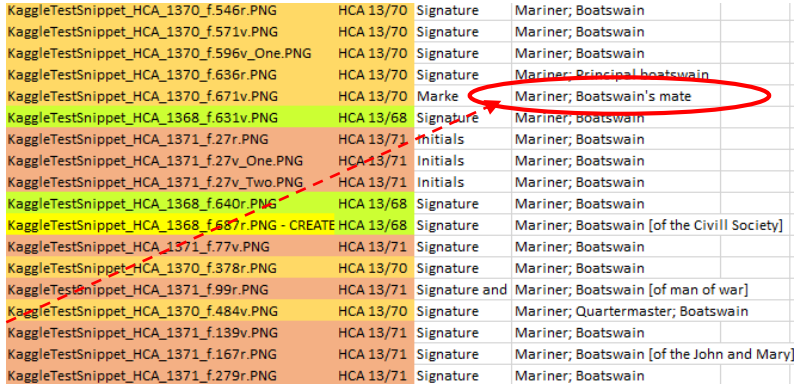
"The premisses hee deposeth being one of the company of the *Bridgewater ffrygott*, and sawe the same soe done" [HCA 13/72 f.90r] [CONCLUSION: One of the company]

"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]

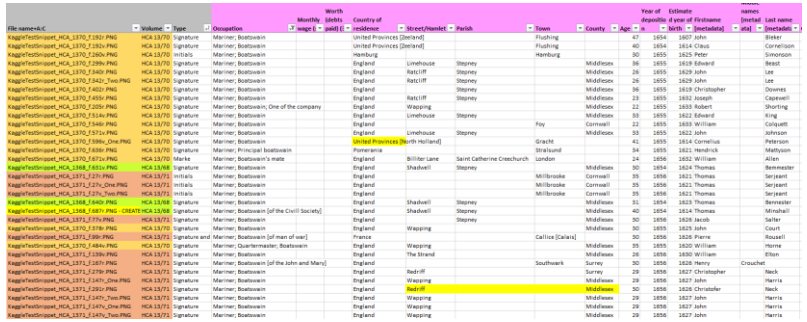
"after such buying of the said shipp the *Santa Hellen* by the said da [?Groots] and company aforesaid, namely in or about the moneth of October 1643 (about three monethes after the adiudication aforesaid) being in possession of the said vessell, put and constituted this deponent master of her, and hee 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]



KaggleTestSnippet_HCA_1370_f.546r.PNG	HCA 13/70	Signature	Mariner; Boatswain	
KaggleTestSnippet_HCA_1370_f.571v.PNG	HCA 13/70	Signature	Mariner; Boatswain	
KaggleTestSnippet_HCA_1370_f.596v_One.PNG	HCA 13/70	Signature	Mariner; Boatswain	
KaggleTestSnippet_HCA_1370_f.636r.PNG	HCA 13/70	Signature	Mariner; Principal boatswain	
KaggleTestSnippet_HCA_1370_f.671v.PNG	HCA 13/70	Marke	Mariner; Boatswain's mate	
KaggleTestSnippet_HCA_1368_f.631v.PNG	HCA 13/68	Signature	Mariner; Boatswain	
KaggleTestSnippet_HCA_1371_f.27r.PNG	HCA 13/71	Initials	Mariner; Boatswain	
KaggleTestSnippet_HCA_1371_f.27v_One.PNG	HCA 13/71	Initials	Mariner; Boatswain	
KaggleTestSnippet_HCA_1371_f.27v_Two.PNG	HCA 13/71	Initials	Mariner; Boatswain	
KaggleTestSnippet_HCA_1368_f.640r.PNG	HCA 13/68	Signature	Mariner; Boatswain	
KaggleTestSnippet_HCA_1368_f.687r.PNG - CREATE	HCA 13/68	Signature	Mariner; Boatswain [of the Civill Society]	
KaggleTestSnippet_HCA_1371_f.77v.PNG	HCA 13/71	Signature	Mariner; Boatswain	
KaggleTestSnippet_HCA_1370_f.378r.PNG	HCA 13/70	Signature	Mariner; Boatswain	
KaggleTestSnippet_HCA_1371_f.99r.PNG	HCA 13/71	Signature and	Mariner; Boatswain [of man of war]	
KaggleTestSnippet_HCA_1370_f.484v.PNG	HCA 13/70	Signature	Mariner; Quartermaster; Boatswain	
KaggleTestSnippet_HCA_1371_f.139v.PNG	HCA 13/71	Signature	Mariner; Boatswain	
KaggleTestSnippet_HCA_1371_f.167r.PNG	HCA 13/71	Signature	Mariner; Boatswain [of the John and Mary]	
KaggleTestSnippet_HCA_1371_f.279r.PNG	HCA 13/71	Signature	Mariner; Boatswain	



An ArchiveBot would need to be able to read manuscript text from the main body of legal depositions, not just the front matter at the start of a deposition

Signs of Literacy initiative

Table 21.1: Extract from occupational data sub-set in SOLM-2018 data set			
Data as of 11/09/2018			
Occupations [Synthesised]	**Occupations stated in main metadata**	**Additional occupations & titles in body of deposition**	**Stated activities in body of text**
BOATSWAINS			
Seaman; Commonman [outwards]; Boatswain [return]	Seaman	Commonman [outwards]; Boatswain [return]	
Nauta [mariner]; Coxwain [outwards] [on Thomas and John]; Boatswain [return] [on Thomas and John]	Nauta [mariner]	Coxwain [outwards] [on Thomas and John]; Boatswain [return] [on Thomas and John]	
Nauta [mariner]; Boatswain's mate [of Thomas and John]; [?Part-owner of ship Thomas and John]	Nauta [mariner]; Boatswain's mate [of Thomas and John]	[?Part-owner of ship Thomas and John]	
Nauta [mariner]; Boatswain [of the Samuel and Mary]	Nauta [mariner]; Boatswain [of the Samuel and Mary]		
Nauta [mariner]; Boatswain [of the Mathewe]	Nauta [mariner]; Boatswain [of the Mathewe]		
Mariner; Boatswain [of the Hopewell]; One of the company [of the Hopewell]	Mariner; Boatswain [of the Hopewell]	One of the company [of the Hopewell]	
Mariner; Boatswain [of the Hawke of Stockholm]	Mariner	Boatswain [of the Hawke]	
Mariner; Boatswain [of the herring busse the Liesda]	Mariner; Boatswain [of the herring busse the Liesda]		
Mariner; Boatswain [of the Pearce]	Mariner; Boatswain [of the Peace]		
Mariner; Boatswain [of the Saint John]	Mariner; Boatswain [of the Saint John]		
Mariner; Boatswain [of the Liesde]; One of the company [of the Liesde]	Mariner	Boatswain [of the Liesde]; One of the company [of the Liesde]	
Mariner; Boatswain [of the Recovery]	Mariner; Boatswain [of the Recovery]		
Boatswain [of the Arke]	Boatswain [of the Noahs Arke or Arke]		
RANDOM SELECTION OF OCCUPATIONS			
Waterman	Waterman		Being aboard the Saint Christofer
?	?		Bound in a bond
Mariner; Sopra cargo; Steward [of unnamed ship]	Mariner	Sopra cargo; Steward [of unnamed ship]	Has used the Virginia trade for 16 years
Haberdasher; Citizen	Haberdasher; Citizen		Involved in arbitration
Merchant	Merchant		Kept accounts for Mr William Bewly, merchant of London
Merchant; Servant [to William Brugg]	Merchant	Servant	Kept his bookes [of William Brugg]
Merchant cashier [to merchant]	Merchant cashier		Made entries in books of accounts
Mariner; Master's mate [of the Advantage frigot]	Mariner; Master's mate [of the Advantage frigot]		Present at taking of the Golden Starr
Sailor	Sailor		Put on board captured ship
Cooper [on shore]	Cooper [on shore]		Taking up oiles
Sailor; Worked on building of ship [the Great Christofer]	Sailor		Worked on building of ship [the Great Christofer]

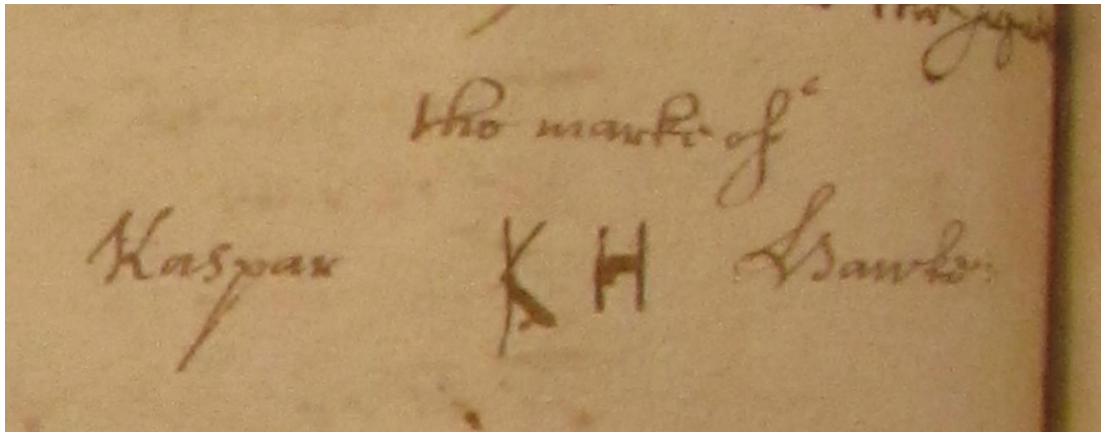
Note four out of thirteen boatswains identified here are described simply as seaman, nauta or mariner in the legal front matter

Note information on activities of deponents in main body of text helps clarify and refine the main occupational descriptions

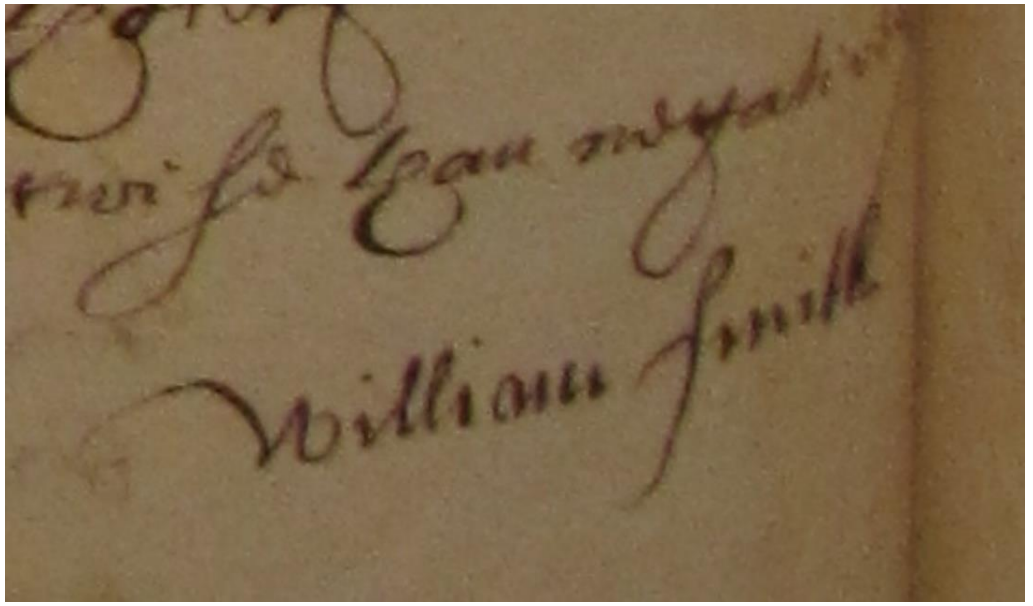
Source: Extract from SOLM-2018 database, as of 11/09/2018

Mismatch between English rendition of names in front matter and initials and signatures made by non-native English speakers at end of deposition

Signs of Literacy initiative



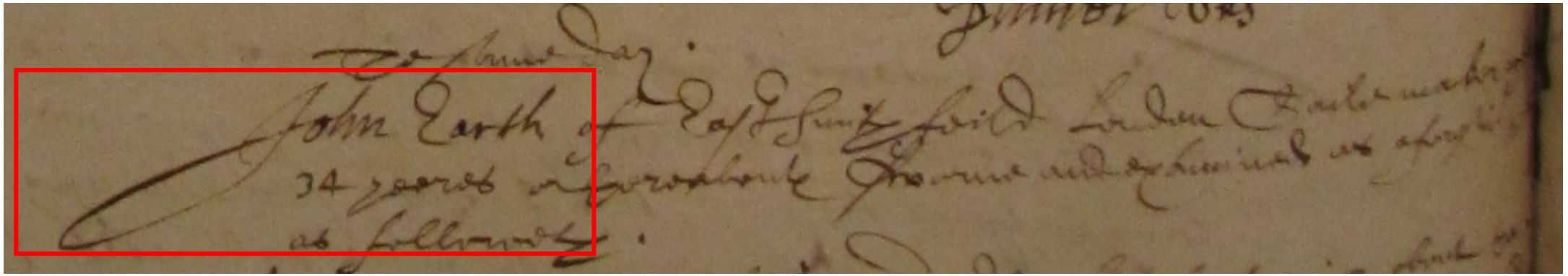
Front matter: Gaspar Hawke
Initials: KH [Kaspar Hawke]

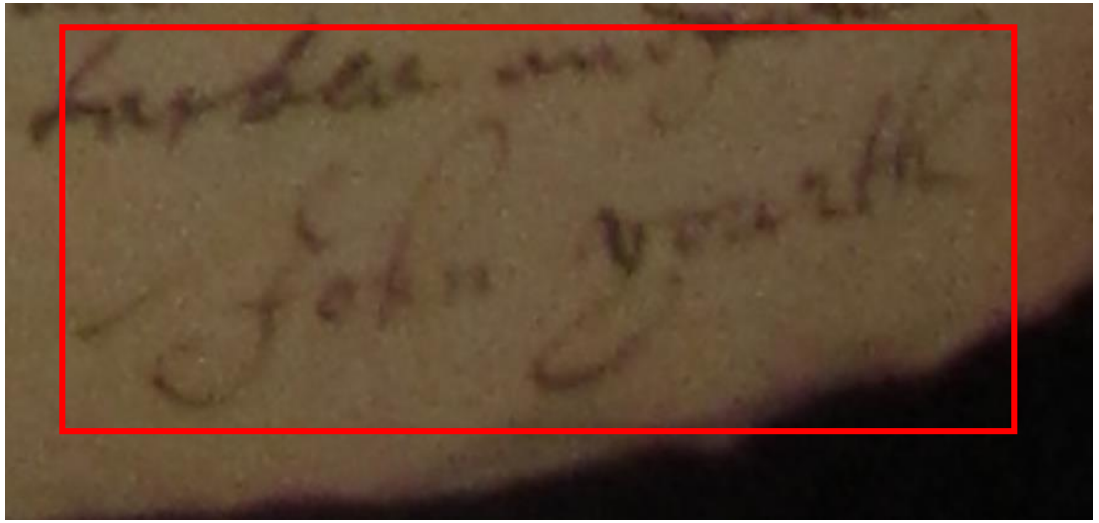


Front matter: William de Smitt
Signature: William Smith

Source: Clockwise from LH corner: KaggleTestSnippet_HCA_1368_f.97v.PNG, KaggleTestSnippet_HCA_1368_f.337v.PNG

Phonetic difference between versions of signatures in legal documents





Front matter: “John Earth”
Signature: “[?John Yearth]”

Source: Extracts from HCA 13/68 f.543v

Occupations in front matter associated with “One of the company” in body of text

Signs of Literacy initiative

Table 21.2: Extract from occupational data sub-set in SOLM-2018 data set	
Occupations in front matter associated with “One of the company” in body of text	
Data as of 12/09/2018	
Occupations stated in main metadata	**Additional occupations & titles in body of deposition**
Carpenter [on ship] [of the Great Christopher]	One of the company [of the Great Christofer]
Mariner	Boatswain [of the Liesde]; One of the company [of the Liesde]
Mariner	Foremastman; One of the company
Mariner	One of the company
Mariner	One of the company [of the Peter and Jane, burthen = 250 tons]; Prisoner [in Lisbon]
Mariner	One of the company [of the ffreetrade]
Mariner	One of the company
Mariner	One of the company; Cooke [cook]
Mariner	One of the company; Common man of the company
Mariner	One of the company [of the Olive Branch]; Master [of the Olive Branch, on death of its master during voyage]
Mariner	One of the company [of the John]; Steersman [of a Spanish man of war, the Angell, under a Claes Johnson]; Appointed acting master [of the Angell, by Claes Johnson, while he went overland to Saint Domingo]
Mariner	Coxen [of the ffrancis and John]; One of the company [of the ffrancis and John]
Mariner; Boatswain [of the Hopewell]	One of the company
Mariner; Boatswain [of the Seven Sisters, a ship granted a Commission]	One of the company [of the Seven Sisters, a ship granted a Commission]
Sailor	One of the company
Sailor	Cook [of the Saint Peter]; One of the company [of the Saint Peter]
Sailor	One of the company [of the Endeavour of Arundell]
Sailor	One of the company [of the Maidstone frigot in the immediate servce of the English Commonwealth]

Source: Extract from SOLM-2018 database, as of 12/09/2018

Processes for automation

Help us build a bot

Build a quick and dirty Archival Bot in 50 days

Work has started

Establish a process for machine supported volunteer extraction & synthesis of metadata for the SOLM-2018 machine learning training data set of markes, initials and signatures

Share results at the Transkribus users conference in Vienna, November 8th-9th, 2018

- Identify manuscript page with signoff
- Establish type of signoff [marke, initial(s), signature]
- Establish linked front matter of deposition
- Establish foliation & image record numbers of signoff & front matter
- Establish name of deponent & age
- Establish residence of deponent
- Establish occupation of deponent from front matter
- Establish further occupational data from full text of manuscript

We are looking for a small number of volunteers to work with us as a facilitated remote team – public historians, who can read cursive script, coders & NLP experts - to create a quick and dirty bot as a Proof of Concept to show at the Transkribus Users Conference

Identify manuscript page with signoff

Establish type of signoff
[marke, initial(s), signature]

Establish linked front matter of deposition

Establish foliation & image record numbers of signoff & front matter

Establish name of deponent & age

Establish residence of deponent

Establish occupation of deponent from front matter

Establish further occupational data from full text of manuscript

Visualising metadata created with aid of ArchiveBots

We need visual metadata, which can be machine processed

Signs of Literacy initiative

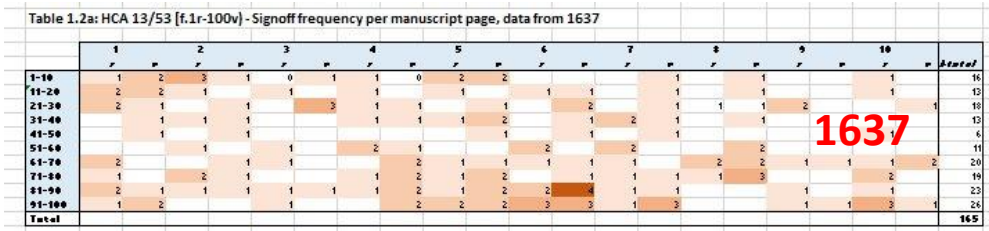
Table 1.2a EXPANDED: HCA 13/53 [f.1r-340v) - Signoff frequency per manuscript page, data from 1637

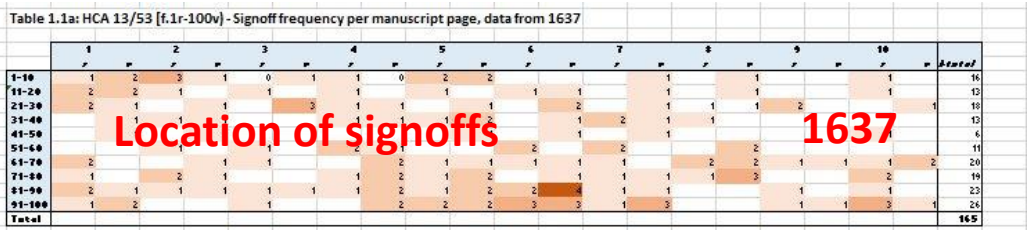
	1		2		3		4		5		6		7		8		9		10		
	r	*v*	*r*	*v*	*r*	*v*	*r*	*v*	*r*	*v*	*r*	*v*	*r*	*v*	*r*	*v*	*r*	*v*	*r*	*v*	*Subtotal*
1-10	1	2	3	1	0	1	1	0	2	2				1		1			1		16
11-20	2	2	1		1		1		1		1	1		1		1			1		13
21-30	2	1		1		3	1	1		1		2		1			2			1	16
31-40		1	1	1			1	1	1	2		1	2	1		1					13
41-50		1		1						1		1		1					1		6
51-60			1		1		2	1			2		2			2					11
61-70	2			1	1			2	1	1	1	1	1		2	2	1	1	1		18
71-80	1		2	1			1	2	1	2		1	1	1	1	3			2		19
81-90	2	1	1	1	1	1	1	2	1	2	2	4	1	1			1		1		23
91-100	1	2			1			2	2	2	3	3	1	3			1	1	3	1	26
101-110	2	1	2	2	1	1	1	2	2	1	1	1	1		2	1	2				23
111-120	1			1				1	1	2	1		2	1	2	2	1	1			16
121-130			1					2			1	1	2			3	1		1		12
131-140	2	3	2	1	1	2				1		1	1	1	2	2	1	2		1	23
141-150	1	1	2	1	2	2	1	2		1	2	1	1	1	1		2		1		22
151-160		1					2		1	1	2	1	2	2	1	1	2		1	1	18
161-170		1		2	2	1					1	2	1	1	2	1	1		2		17
171-180	1	2			2	1	1		1	1		1						1			11
181-190					2							3	1	2	1	4	4	2	1	1	21
191-200		1		1		1	1	1	1		1	1		1	1	2	1	2	2		17
201-210	2	2				1	1			2	4	3		1		2	1	2	1	2	24
211-220	1			2	1			1		1		1	4	1	2	3	4	2	2		25
221-230	2	2	1		3		1	2		1	1	1	1	2	2	2	2			2	25
231-240		1	1		1		2	1	1			3		2	2					1	15
241-250	2										2	1	1	1	2	2	1		2	1	15
251-260	2		2	2	1		1	1		1	1	1			1		1			1	15
261-270		1	1		1		1	1			1	1		1		1				2	11
271-280	2			1			1		1	1	1			1		2	1		1		12
281-290	1			1		2	1		1	1	1		1	1	1	2			1		14
291-300		1		1	1	1	1	2		1		1	1	1	1		1	1		1	15
301-310		1		2						2	1		2	1	1		1			1	12
311-320			1					1						1	2	1					6
321-330				1					1				1					1	3		7
331-340	1	2	2	2				2		1	1	1	1	1	1	1			1		18
Total	31	30	24	27	23	17	23	30	19	31	31	39	31	33	30	42	33	16	29	16	**555**

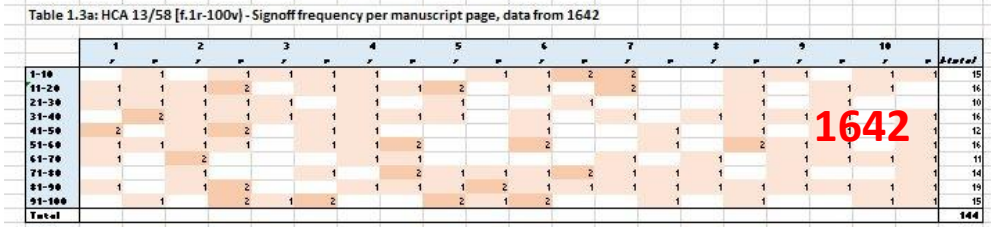
Source: SOLM-2018 database

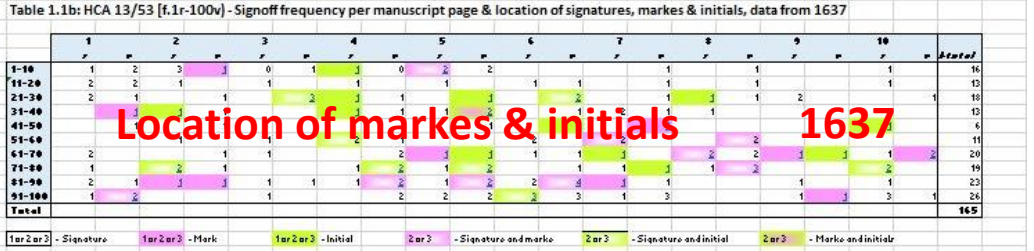
Archivists, computer scientists and users of all sorts need to work together

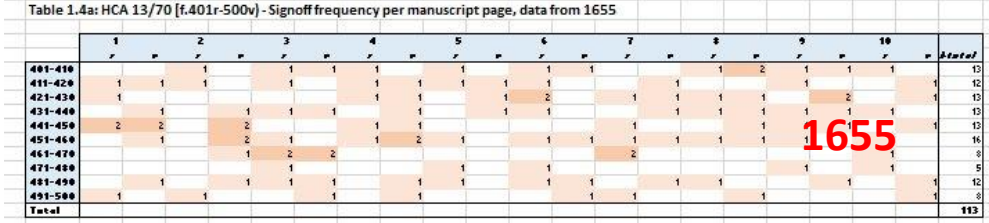
Signs of Literacy initiative

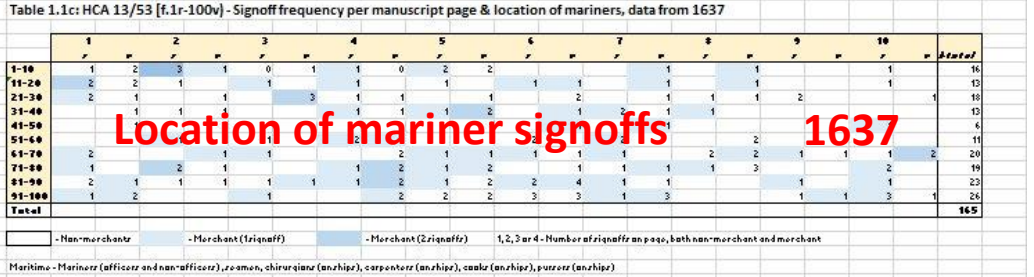


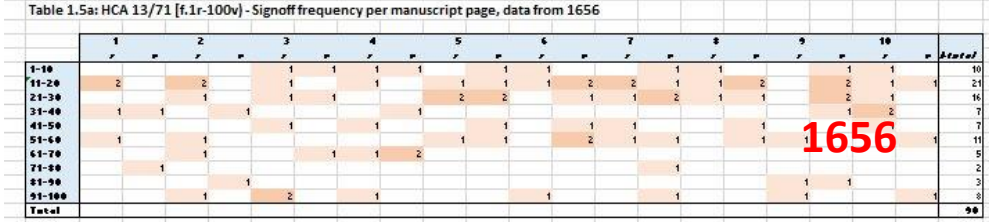


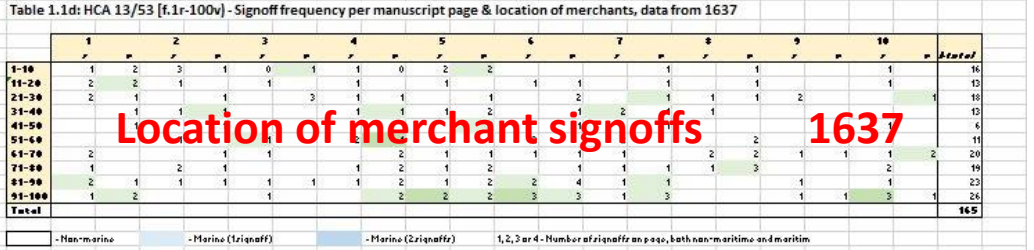












Source: SOLM-2018 database

SOLM-2018 machine learning training data set

Signs of Literacy initiative

The SOLM-2018 data set - nearly 5,000 markes, initials and signatures for individuals born between 1520 and 1669, aged between 15 and 86 years

Chart 1.0: Year of birth of deponent
SOLM-2018, ver.1.29, n = 4772, 12/09/2018

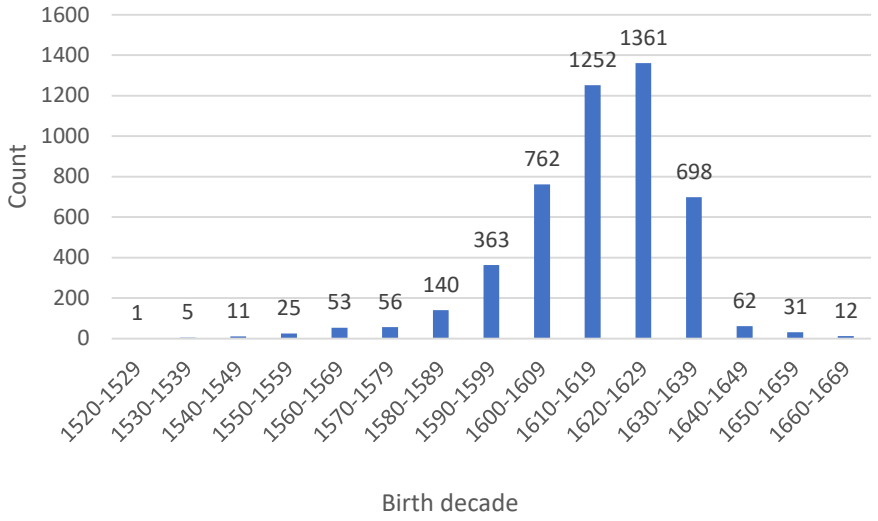
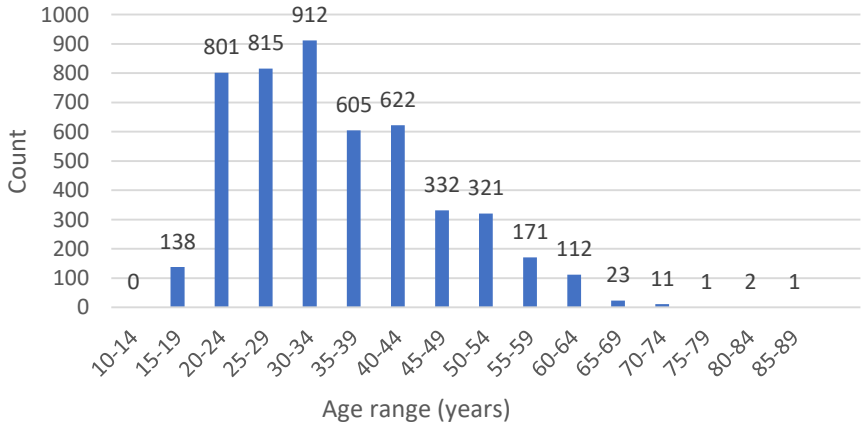


Chart 2.0: Age of deponent in year of deposition
SOLM-2018, ver. 1.29, n = 4,867, 13/09/2018



Source: SOLM-2018 machine learning traing data set

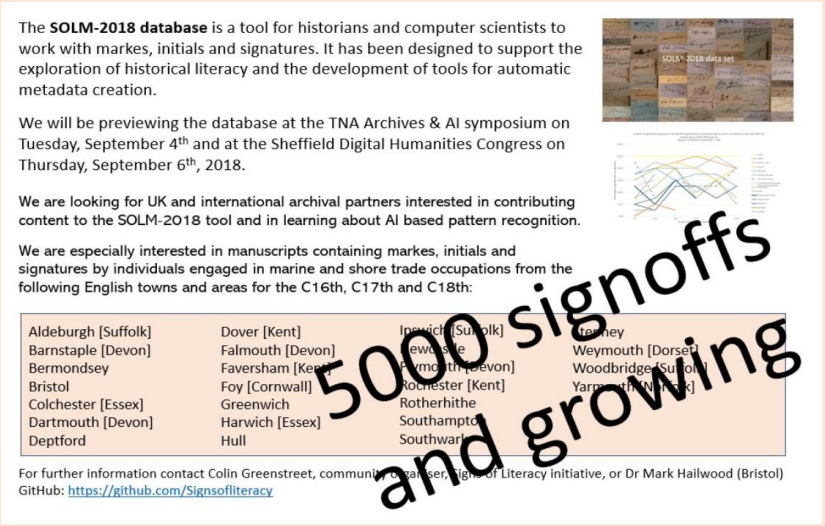
Our vision is a SOLM-2023 database with 1 million markes, initials & signatures from across Europe & North America from the C16th to C18th

Signs of Literacy initiative

The maths

- 3 person/months to create 5,000 signoff SOLM-2018 database consisting of image snippets; boundary boxed snippets on full page images; 5,000 lines x 25 rows of metadata
- 6 person/months to create our targeted 10,000 SOLM-2018 training database
- 20,000 signoff processing per person year
- Target of 1 million signoffs in our database
- 100,000 signoffs per year with 5 people working full time

That's TEN YEARS to achieve our vision
with 50 person years to do it!!!!!


The **SOLM-2018 database** is a tool for historians and computer scientists to work with markes, initials and signatures. It has been designed to support the exploration of historical literacy and the development of tools for automatic metadata creation.

We will be previewing the database at the TNA Archives & AI symposium on Tuesday, September 4th and at the Sheffield Digital Humanities Congress on Thursday, September 6th, 2018.

We are looking for UK and international archival partners interested in contributing content to the SOLM-2018 tool and in learning about AI based pattern recognition.

We are especially interested in manuscripts containing markes, initials and signatures by individuals engaged in marine and shore trade occupations from the following English towns and areas for the C16th, C17th and C18th:

Aldeburgh [Suffolk]	Dover [Kent]	Ipswich [Suffolk]	Walmey
Barnstaple [Devon]	Falmouth [Devon]	Lewes [Sussex]	Weymouth [Dorset]
Bermondsey	Faversham [Kent]	Plymouth [Devon]	Woodbridge [Suffolk]
Bristol	Foy [Cornwall]	Rochester [Kent]	Yarmouth [Norfolk]
Colchester [Essex]	Greenwich	Rotherhithe	
Dartmouth [Devon]	Harwich [Essex]	Southampton	
Deptford	Hull	Southwark	

50,000 signoffs and growing

For further information contact Colin Greenstreet, community organiser, Signs of Literacy initiative, or Dr Mark Hailwood (Bristol)
GitHub: https://github.com/Signsofliteracy

Our challenge to archivists, computer scientists and historians: Help us develop the tools to create a SOLM-2023 database of 1 mill signoffs with a productivity rate of ten times today's best, at a resource cost of 5 person/years, not 50 person/years, and in half the time

More generally, we need to work together, if we are going to make sense of our digitised manuscript archives – **developing AI tools to process archival images and to identify, extract, read and record metadata**

For more information contact Colin Greenstreet, community organiser of the Signs of Literacy initiative, and Dr Mark Hailwood (Bristol)
https://github.com/Signsofliteracy

We are looking for international partners – archives & archivists, digital publishers,
image & text oriented computer scientists & machine learners,
corpus & historical linguists, and historians.
United Kingdom, Netherlands, Sweden, Poland,
Germany, France, Spain & North America

Contact:
Colin Greenstreet (community organiser, Signs of Literacy initiative)
or Dr Mark Hailwood (Bristol)
GitHub: https://github.com/Signsofliteracy/
Web: http://signsofliteracy.org