OUR EXPERIMENTS WITH TRANSKRIBUS ON TABULAR SOURCE MATERIAL (CONSULTANCY WITH AVIVA GORUP ARCHIVE)

Research and text by Dr Oliver Dunn.

The following formed part of a report to Aviva plc on the potential for using Transkribus for digitizing their rich but underused archive. It explains recent research into database creation using the software in conjunction with other tools and methods. Section 1 is an overview of the system, while Section 2 described the steps behind the creation of the data.

Section 1 outline of Transkribus platform:

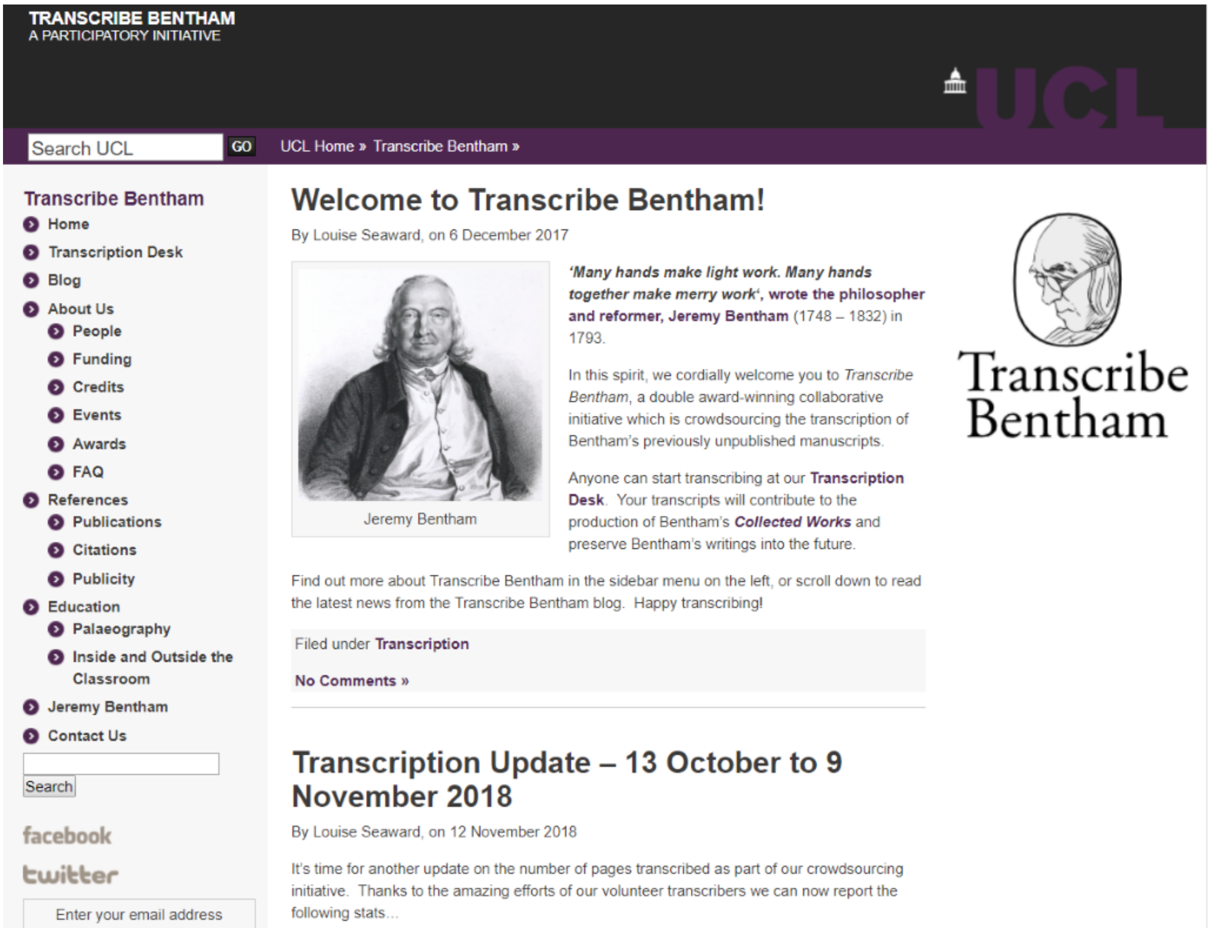
The Transkribus workflow is as follows:

1. High-standard images are uploaded onto the Transkribus platform and secure servers.
2. These images are scanned, and the manuscript lines and text linked to a blank transcript.
3. Volunteers create an online account with Transkribus to access the scans.
4. Volunteers work through the scanned lines on the image using filling in the blank transcript.
5. Transcriptions are stored automatically and can be exported as various file formats like PDF.

6. Volunteer transcriptions remain linked to original collections and can be shared widely.

Transkribus has been used in a well-regarded project called the Transcribe Bentham project (see Fig 7), which is digitising the manuscript notebooks of the nineteenth century philosopher and reformer Jeremy Bentham and those of his secretaries. This system was first developed with this project at the University of London in partnership with other European universities. It is maintained using public funding provided by the European Commission, which means it is currently free to use. The READ Project manages Transkribus platform, a computing resources, and servers.

There may be specifics to work out with READ, but the website states that they offer a tailored service to archives: Archives and libraries are invited to join the project with a Memorandum of Understanding which will enable them to take part in testing the technology and contributing to the improvement of our tools. We are not currently sure exactly what this memorandum would entail, but it seems likely that the service would be available for this project. Whilst the service is currently free, that may change in the future.



Welcome to Transcribe Bentham!

By Louise Seaward, on 6 December 2017

Jeremy Bentham

'Many hands make light work. Many hands together make merry work', wrote the philosopher and reformer, Jeremy Bentham (1748 – 1832) in 1793.

In this spirit, we cordially welcome you to *Transcribe Bentham*, a double award-winning collaborative initiative which is crowdsourcing the transcription of Bentham's previously unpublished manuscripts.

Anyone can start transcribing at our **Transcription Desk**. Your transcripts will contribute to the production of Bentham's ***Collected Works*** and preserve Bentham's writings into the future.

Find out more about Transcribe Bentham in the sidebar menu on the left, or scroll down to read the latest news from the Transcribe Bentham blog. Happy transcribing!

Filed under **Transcription**

No Comments »

Transcription Update – 13 October to 9 November 2018

By Louise Seaward, on 12 November 2018

It's time for another update on the number of pages transcribed as part of our crowdsourcing initiative. Thanks to the amazing efforts of our volunteer transcribers we can now report the following stats…

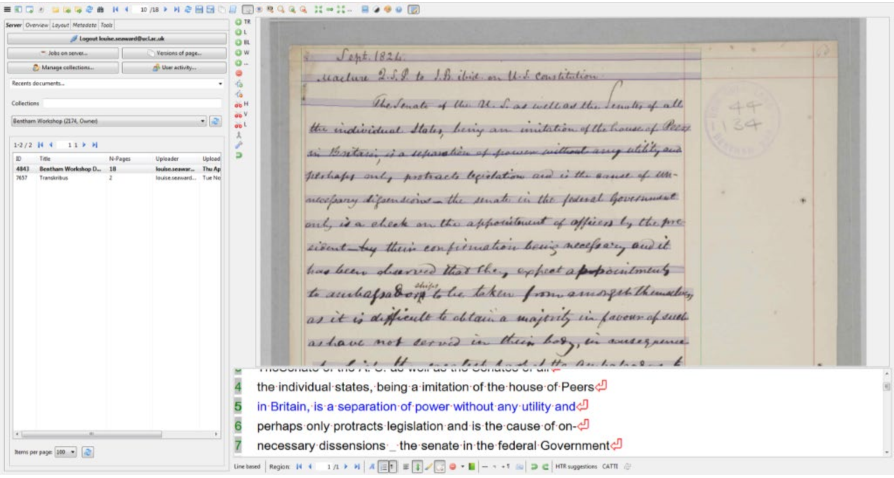
FIGURE 1 SNIPPET FROM THE TRANSCRIBE BENTHAM WEB PAGE. (NOTE SOCIAL MEDIA AND BLOG USES.)

An important caveat should be added about Transkribus: the accuracy and overall usefulness of handwritten text recognition models depends on the number of different handwriting styles found in runs of documents in series. Because there is a range of different handwriting found across AGA documents, some shorter or more diverse series of manuscripts may be unsuited to automated text recognition.

However, just 20,000 words (around 40 pages) of training input from similar handwriting can produce accuracy rates of about 80%. With 50,000 words input, rates increase up to 95%. We do not know which, or how many of Aviva's record collections contain as much as 50,000 words in the same hand. Some research in the archive would be necessary to identify promising record series. Resources permitting, we think it would be worth experimenting with the machine learning element of Transkribus in the AGA Archive. That said, whatever can be done today, can probably be done better and with lower labour inputs in the future as the technology evolves.

Transkribus would be a good platform to use for transcription work regardless of the efficacy of the machine learning auto-transcription, because of the benefits of the platform it provides for manual transcription. As a bonus, once an initial input text has been keyed of around 10-20,000 words (around 20-40 pages), is that the search function can then be enabled for the entire photographed collection (if the same hand continues) since accuracy rates need not be high. This is because models with lower success rates can still be used to search through documents, if not to produce highly accurate full transitions from scanned images. Fuzzy searches can also be used to improve search results. Searches could presumably be enabled across AGA digitised collections allowing searches of surnames to find potential ancestors, for example. We think that with limited training data AGA collections could benefit from key-word search functionality even with lower accuracy rates based on heterogeneous handwriting across documents. Such systems, as they evolve, are bound to represent the future for archives. A further reason for using Transkribus is that it is currently free to use. No doubt there are other platforms, but this is the only one we are familiar with.

For printed materials, the process of optical character recognition is much simpler and more advanced in terms of technology. AGA printed records from the nineteenth century could be scanned and transcribed again using Transkribus, or equivalent tools from Adobe, or ABBYY Finereader.



4 the individual states, being a imitation of the house of Peers

5 in Britain, is a separation of power without any utility and

6 perhaps only protracts legislation and is the cause of on-

7 necessary dissensions _ the senate in the federal Government

FIGURE 2 SCREENSHOT FROM TRANSKRIBUS. THE IMAGE IS A DIGITAL IMAGE OF A NOTEBOOK ONCE BELONGING TO THE PHILOSOPHER JEREMY BENTHAM.

Section 2 Transcription of tabulated data:

Much of Aviva's Archive collection consists of tabular material organised in columns. This material is more challenging than continual texts, like the Bentham notebooks, for OCR and machine-learning transcription. Appendix Two details our experiments with Transkribus with such material.

With training, Transkribus can read large amounts of standard manuscript text of the kind found in letters and notes such as those of Bentham. The fact the handwriting is in the same hand for thousands of pages is a considerable advantage, as is the format of a continuous stream of text going from left to right, with successive lines going down the page. Tabular records and records with changes of hand are considerably more difficult. However, Transkribus is also useful when transcribing data from tabular record types of a kind typically found in records in the AGA – such as the thousands of lists of policyholders of fire and life insurance. Transkribus and similar software promises to revolutionise access to historical documents by its ability to automatically read old texts. Here we give details of our experiments with Transkribus using manuscript tables and lists. The sources we used were significantly more challenging to read than any of the records likely to be in the Aviva Archive.

To digitise tabular information is considerably more difficult than a stream of text as found in notes and letters because of the need to organise the information into the correct boxes or cells. Policyholders' names, often with addresses, occupations, place of residence, sums paid, are typical forms of data found in AGA records. A ball-park estimate suggests there may be 50 million lines of such data in the AGA records going back to 1697. To transcribe even a fraction of such a collection manually would take a lifetime, but we think this nominal "big-data" source can be more rapidly digitised using machine reading combined with data cleaning methods outlined now. The approach we have been experimenting with is a semi-automated one because final manual correction of transcriptions remains necessary given the current state of technology and normal requirements of data entry precision. We share it here in simple form because we think that with tasks such as indexing policyholder names it is possible to cut transcription and data entry time by 50-70% by employing it.

First, we will outline the procedure and then provide some more detail before giving a concrete example of results.

The method involves the following steps:

1. High-definition document photography of AGA documents;
2. Image pre-processing using off-the-shelf scanning software to enhance images;
3. Uploading these pre-processed images into Transkribus;
4. Manual transcription of a sub-set of records to 'teach' the software to automatically recognise handwriting in each collection;
5. Manual selection of the lines of text that Transkribus will process (termed 'segmentation') in step 6.
6. Automated recognition of text in images using the training model;
7. Exportation of results as raw automatically-transcribed text;
8. Global correction of the raw text (using a combination of Word and Excel functions);
9. Filtering and cutting out individual lines and words (in Excel);
10. Arrangement of data into columns and rows in a database;
11. Final manual correction of the data by comparing with the original source.

Step 1 is vitally important to get fully focussed images to enable effective scanning and recognition of text and text fields. This makes the quality of document photography even more important than is normally the case. It is important to use very high-quality lenses if the software is to function well.

Step 2 is also important to add definition and bring out manuscript text from the page for better results later.

Step 3 is simply to upload pre-processed images to remote servers operated by Transkribus.

Step 4 involves manual transcription input from anyone who can read the handwriting found in the chosen source. This input is then used to 'train' 'models' for handwritten text recognition (HTR). This job can take time. However, records that contain lists and tables, such as insurance policy registers, contain lots of repeating text because the data of interest includes names like Thomas, John, and Smith, and other common names and terms. The repetitiveness of such data forms speeds up the process of HTR training in Transkribus and increases accuracy rates simply because fewer words require learning by the system. The same goes for repeating number values, for example birth years, which are another data form that are often good to record but very time-consuming to enter manually.

Note that Transkribus benefits from data sharing from other users, and there may already be a 'model' out there that will work for some Aviva collections without any training even being necessary. Note that Aviva's registers and ledgers are likely to overwhelmingly feature legible and clear handwriting. People with messy handwriting (like Bentham) are unlikely to have been employed.

Steps 5-7 can be done quickly, although manual selection of text fields using drag and select tools in Transkribus is a bit more involved. Currently, it takes about 30 seconds to do this manually per document page. Therefore, a typical insurance policy register of 300 pages might take 2.5 hours to process for this step.

Step 8 involves semi-automated methods of correction of raw text output from Transkribus. Much text, like names, even if they come out mis-transcribed, which they will, can be easily corrected throughout the scanned text because transcription errors are easy to spot and can be fixed globally in standard packages such as Word or Excel. A common name like William, for example, will only suffer from limited errors of a kind that are correctable globally across millions of words in a transcript. In some cases, it will be necessary to refer to the source to correct difficult errors, which is more time-consuming. This is done in the final stages of the process by manual data correction work using the original source and after the main body of partially-corrected data is arranged in Excel (see step 11). Getting the best transcription output possible from Transkribus is important because less time will be needed for correction later.

Transkribus exports raw text as lines matching the layout of the original text. Step 9 involves isolating the lines that contain the chosen data, for example those lines that contain names, and then filtering out any unwanted lines. Words in the filtered lines are 'parsed' from the raw text lines to create new cells and columns of data. Filtering and parsing this way are skilled jobs, but they are automatable to a large degree using Excel. This step can be considered a second cleaning stage after global corrections in step 8 and before the final corrections undertaken in reference to the original source in step 11.

Step 10 involves making sure the chosen data is all arranged in tabular columns in Excel with appropriate headings, for example forename, surname, or address. Because the data at this stage is already parsed and filtered, it should not take long to realign stray data into the correct chosen columns.

It is possible to complete steps 1-4 and 6-10 relatively rapidly by a skilled person for a single record series containing the same or similar handwriting. We estimate it would take such a person five working days to carry out these steps for a series of documents with similar handwriting. The data cleaning, filtering and parsing is equally scalable. However, step 5, involving manual selection of text, does slow things down because it needs to be done individually for each page. Transkribus are working on an automated segmentation tool for converting manuscript tables automatically into modern spreadsheet forms. As and when this functionality becomes available, this will further increase the time savings of using Transkribus.

Step 11 involves fine correction, and this takes more time because it involves manual checking against sources to ensure accuracy. It is normally necessary to spend most time correcting small isolated sections of text, like numbers, that are more often missed by Transkribus. This step ensures the data produced is highly accurate. The amount of work is required will vary in part on the completeness of any transcription. To create an index on key fields is less time consuming than creating a full transcript – as is shown in our example below where we only extracted a fraction of the available data.

The final correction against the original document, as per step 11, takes more time than the preceding more automated stages. This is done at the end when the data has been automatically processed as far as is possible. At this stage there should be a sequence of data that resembles the original layout in the source. Matching the original arrangement will aid the person who does the final corrections because he or she can go through the data methodically by following the original order of information in the original document. We found during recent experiments that such page by page manual corrections took approximately 10 minutes on average per page, which is at least three times faster than entering the same data manually.

Step 11 is the most time-consuming step. At this stage, providing a person who will undertake the final corrections with an almost correct version can lead to faster total turnaround times because it is far easier to correct large amounts of data compared to manually entering it from scratch; in fact, we estimate that it takes at least half the time. This is because by having the text already mostly laid out and inputted aids the data enterer, who simply is required to go through the source image page by page, line by line, to correct any errors caused by Transkribus.

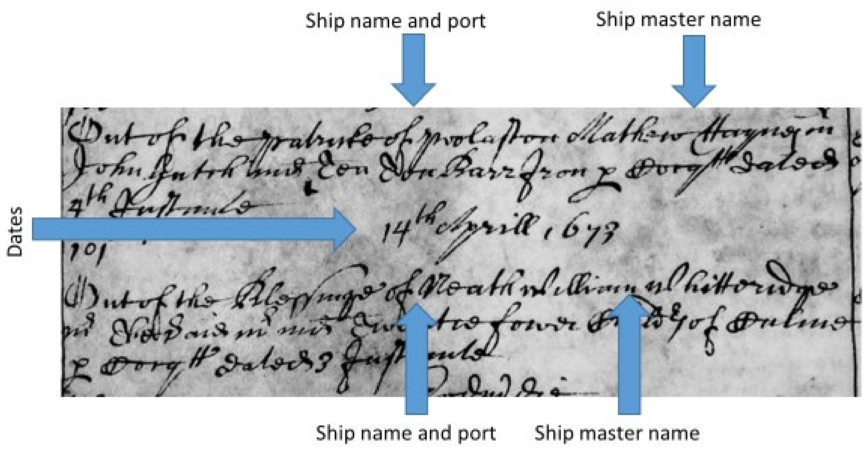
To give an example, to manually transcribe data from a 300-page policy register of the kind viewed for this report, we estimate it might take about 45 minutes per page to extract data such as names and addresses and other personal information. 30 working days would be required for transcription of the whole volume. Using our method, we estimate the same work would take 16 days including training for the first volume, and then any subsequent volumes would likely take around 13 days since the system would have already been trained.

Another consideration is that manual transcription and data entry can be an unpleasant job and can put volunteers off to begin with, but also it can delay projects because working time is normally restricted to just a few hours per day because of the difficulty of the work. Cases of Repetitive Strain Injury do occur if people spend too long doing this kind of work manually, and eyesight can suffer. The proposed method, then, is not only faster and thus potentially cheaper, but also kinder on the transcriber. This point becomes even more important when volunteers are being asked to do transcription work. The transcription and data entry tasks become less arduous, and people will likely volunteer more time as a result.

The method as described so far is somewhat convoluted, so the following worked example illustrates how it works in practice. The following illustrates the method based on very recent work using Transkribus to create large quantities of data from historical lists of shipping dating from the seventeenth century. We created thousands of lines of data using this method. This was done much more rapidly than was possible before.

Figure 1A is a snippet from a volume of seventeenth-century shipping registers, originally created by customs officers at that time. We experimented with extracting specific text as data from this source using Transkribus. The following section illustrates how we did this.

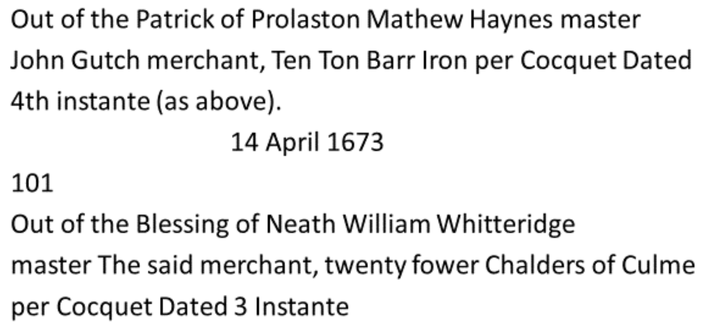
Data to be extracted in image scan


Ship name and port
Ship master name
Dates
Ship name and port
Ship master name

Out of the Patricke of Woolaston Mathew Harrington
John Patricke mr. Ten Ton Barr Iron p Docqt dated
4th Instante
14th Aprill 1673
101
Out of the Blessinge of Neath William Whittoridge
mr. [illegible] two foure [illegible] of Culme
p Docqt dated 3 Instante

Figure 1A image scan of customs list of two English ships sailing into Bridgewater in 1673: the *Patrick* and the *Blessing*

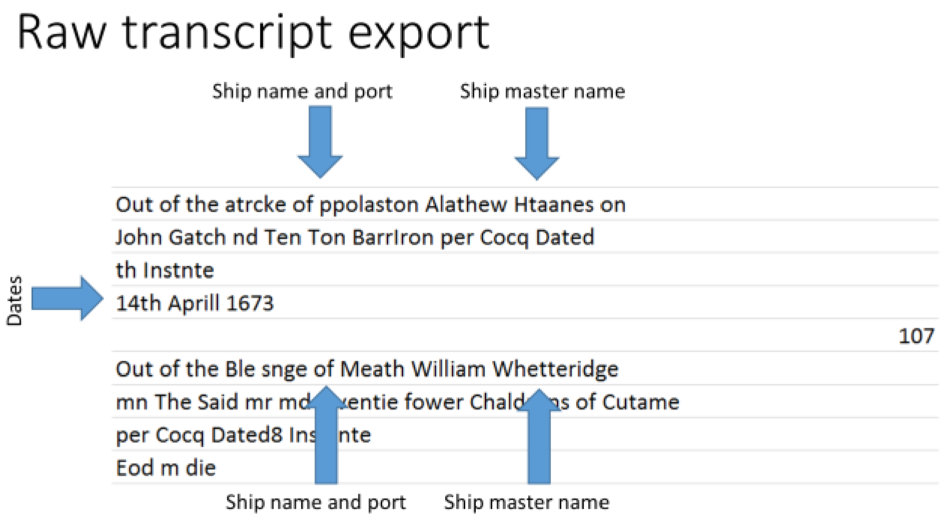
Our own (skilled human) manual transcript of Figure 1A is as follows:



Out of the Patrick of Prolaston Mathew Haynes master
John Gutch merchant, Ten Ton Barr Iron per Cocquet Dated
4th instante (as above).
14 April 1673
101
Out of the Blessing of Neath William Whitteridge
master The said merchant, twenty fower Chalders of Culme
per Cocquet Dated 3 Instante

Figure 2A manual transcription of figure 1 A

Steps 1-7, as outlined above, were performed first to extract an automated transcription from the image scan uploaded to Transkribus and shown in Figure 1A using a bespoke HTR model. Figure 3A below is an image of the raw text output results. These can be compared with our manual, correct transcription in Figure 2A. Note in Figure 3A how the original line breaks are maintained. This eased the filtering for chosen lines with relevant data to be parsed later because lines did not get rearranged and mixed-up during export. Next, the raw text from Figure 3A was corrected globally, filtered, and parsed into columns following steps 8-10.



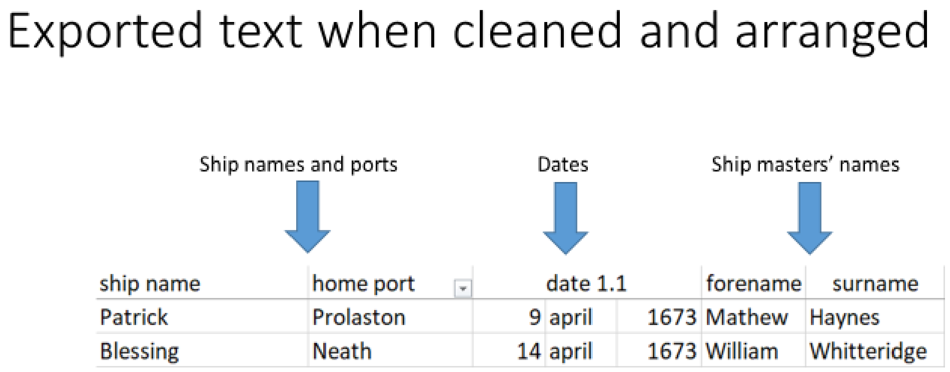
Raw transcript export

Out of the atrcke of ppolaston Alathew Htaanes on
John Gatch nd Ten Ton Barrlron per Cocq Dated
th Instnte
14th Aprill 1673
107
Out of the Ble snge of Meath William Whetteridge
mn The Said mr md entie fower Chald s of Cutame
per Cocq Dated8 Ins nte
Eod m die

Figure 3A Raw text from transkribus based on a 10,000-word training model

Figure 4A shows data extracted from the raw text in Figure 3A after processing in steps 8-10. During the global correction step, most errors that are clear in Figure 3A were easily spotted and corrected. We could see immediately, for example, that 'atrcke' was Patrick without time-consuming referral to the source, and likewise that 'Ble snge' was Blessing, and 'Alathew', Mathew. Almost all names were corrected globally, and hundreds of corrections were made in seconds where they occur in the rest of the text (not shown). A little knowledge of the actual content and nature of the source enables rapid recognition of name spelling errors.

In Figure 4A, specific data of interest was extracted from the two ships' entries seen in the raw text in Figure 3A and originally in the image in Figure 1A. The new data in Figure 4A now covers ship names, ports, arrival dates, and ship-masters' names. These data have been globally corrected, filtered from the main raw text in Figure 3A, parsed and arranged into columns in Figure 4A. Other data, such as for ship cargoes, seen in Figure 3A, were filtered out. These data can be reprocessed and added later to create new data if desired. The selected data is now ready for the final step which is to check it for errors using the source images.



Exported text when cleaned and arranged

ship name	home port	date 1.1		forename	surname
Patrick	Prolaston	9 april	1673	Mathew	Haynes
Blessing	Neath	14 april	1673	William	Whitteridge

Figure 4 A historical data corrected and arranged in Excel after all steps

One final important point to make is that this information and the time estimates given are based on our first experiments with using this method. We would expect to get better results as we get more experience. The software will also keep improving. Also, we predict that Aviva's records could portably be processed much more accurately than the very difficult seventeenth-century handwriting we have used so far in our test projects, firstly because of neater writing hands that are more consistent, but also because most of these hands date from later times, meaning Transkribus will be more familiar with them and so will likely need less training and will be more accurate as a result.