If you teach a computer to Read…

Louise Seaward and **Elaine Charwat** talk about a digital text recognition project that has the potential to revolutionise access to handwritten documents, opening up historical records and shining a light on previously understudied sources. But can a machine really 'learn' a task that requires years of specialist training?

Dr Louise Seaward (louise.seaward@ucl.ac.uk) **is Research Associate, Bentham Project, University College London and Elaine Charwat (**elainec@linnean.org**) was until recently Deputy Librarian, The Linnean Society of London.**

DIGITISATION is one of the top priorities for libraries with manuscript collections. Digital images help to preserve precious material and also mean that users can potentially access unique collections from anywhere in the world. A new research project, funded by the EU, is aiming to increase the usability of digital records further still. The Recognition and Enrichment of Archival Documents project (Read)[1] is developing technology which enables computers to read and search handwritten historical documents. As part of the Read project, the Bentham Project at University College London[2] and the Linnean Society of London[3] are applying this technology to their collections of 18th and 19th century manuscripts.

Recognising handwritten text

Computer scientists working on the Read project are developing Handwritten Text Recognition (HTR) technology using thousands of manuscript pages of various dates, styles, languages and layouts. This large data set will make it possible for computers to process any kind of handwritten document, whether it be medieval Latin, old Swedish or modern English. This will allow users to conduct a full-text search of large historical collections and to extract specific information such as names, places and dates. HTR technology thus has the potential to open up historical records and shine a light on previously understudied sources.

We are all familiar with OCR (Optical Character Recognition) technology for printed texts – a page from a book can be scanned and the OCR software will produce the typed text automatically. The once static text is now fully searchable, keywords can be indexed or tagged and it can be edited. Essentially, the Read project aims to achieve similar results for handwritten texts through HTR. Anyone familiar with archival material will recognise the hurdles: the complexity of a dizzying array of individual handwriting styles, eclectic abbreviations, special characters and foreign languages.

Can a machine really 'learn' a task that requires years of specialist training? The answer is that this has become a real possibility. At the Read project conference and kick-off meeting in Marburg (Germany) in January this year, initial scepticism soon turned into

Further information

The READ Project:
http://read.transkribus.eu/

Transkribus:
https://transkribus.eu/Transkribus/

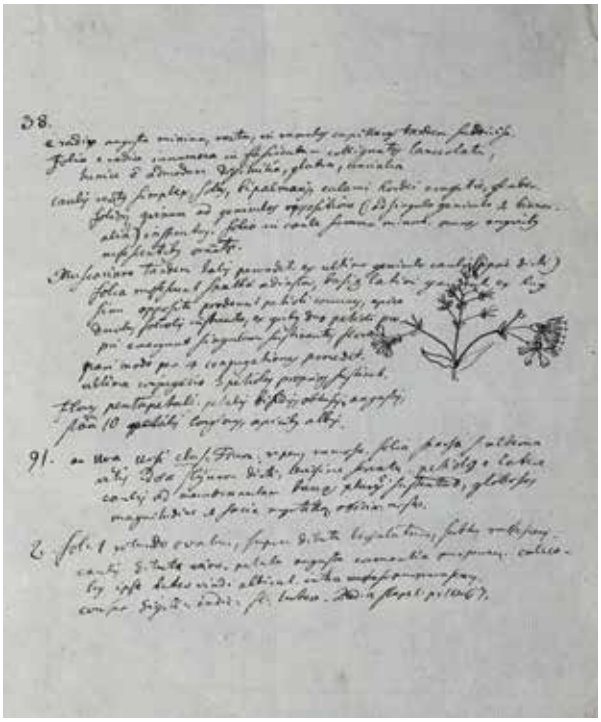
Bentham Project:
http://www.ucl.ac.uk/Bentham-Project

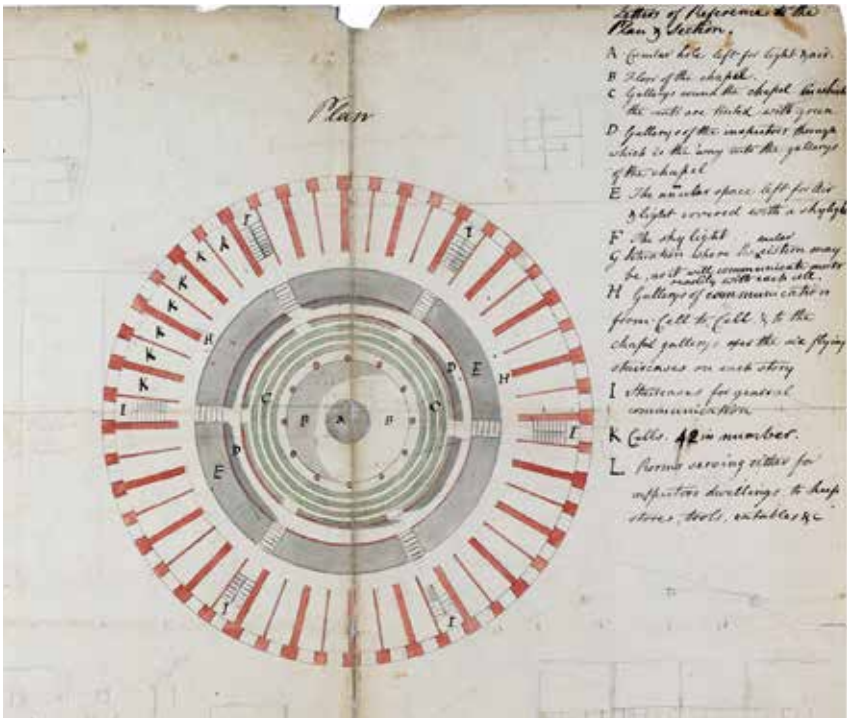
The Linnean Society of London:
https://www.linnean.org/

The LINNEAN SOCIETY of London
A living forum for biology

READ

> A new research project, funded by the EU, is aiming to increase the usability of digital records further still.





Manuscripts often include drawings and illustrations, which also need to be captured and indexed.

amazement. A digital image of a scanned page of handwritten text was enhanced by applying an automatic stain remover, colour binarisation (conversion from colour or greyscale), skew and warp correction, baseline detection, slant correction and size normalisation. The page layout can then be examined in more detail, and text blocks, lines and words are determined.

The software analyses the pixel values of characters and words in a digital image, scanning them in several directions. The researcher will be presented with a hit-list of likely matches for the word he or she is looking for, and is now able to tell the machine which hits are correct, and which are not. The computer will remember this, and fine-tune its performance accordingly. A complex neural network supports this process – it can contain over 106 trainable parameters.

Training the technology

A prototype of this technology is already being made available through the Transkribus transcription platform, which can be downloaded for free from the Transkribus website[4] Unlike Optical Character Recognition, HTR technology cannot recognise words right away. The technology needs to be trained to understand each style of writing by being shown examples of documents that have been correctly transcribed. Machine learning then allows HTR engines to generate a model which is capable of reading documents written in a specific hand.

Anyone who wants to work with HTR technology must go through this training process, although it can be undertaken with as little as 30 manuscript pages. Users upload digital images to Transkribus, transcribe a sample of material and then contact the Transkribus team who will generate a HTR model for their collection. While totally accurate HTR has not yet been possible, the latest experiments show results with a Character Error Rate of around only five per cent. So, around 95 per cent of the characters in an automatically generated transcript would be correct. As the Read project progresses, more data will be collected and the results of the HTR will become increasingly accurate.

The Linnean Society's role

The Linnean Society of London plays a supporting, but important, role in working towards this shared goal – to unlock complex handwritten material in archival collections, to automatically index digital images of text and to teach computers how to transcribe handwritten text. The society holds a great number of very complex archival items, which are hugely important historically and scientifically. The collections of Carl Linnaeus (1707-1778) are some of the foundation stones of modern biology, and are still crucial today for naming, describing and protecting biodiversity. This material, often written in complicated styles of handwriting and in languages like Latin or Swedish, sadly remains out of reach for all but the most intrepid researchers. Likewise, much of the handwritten information relating to specimens in the society's collections (plants, insects, etc.) is waiting to be transcribed and tagged to the digital image of the specimen so that it becomes finally searchable.

As a Read Project Partner with a 'Memorandum of Understanding', the society is providing some of this very complex material as digital images, along with the expertise of staff, Fellows and researchers, to help fine-tune the HTR software. Andrea Deneau, the society's Digital Assets Manager, has uploaded, processed and transcribed an impressive number of manuscript pages on the Transkribus platform. She pointedly summarised the many challenges faced when dealing with this complex material: 'Several of the manuscripts I did manage to transcribe are littered with errors and "unclears" (Linnaeus's handwriting is notoriously difficult and is in poor Latin, which makes it all the worse) but I am hoping that there will be enough there for patterns to be recognised by the HTR software.'

Transcribing Bentham at UCL

HTR technology could potentially revolutionise the way members of the public access historical records. The Bentham Project at UCL aims to test this hypothesis by integrating HTR into its crowdsourcing initiative, Transcribe Bentham (http://blogs.ucl.ac.uk/transcribe-bentham/). The Bentham Project is working on the scholarly edition of the writings of the British philosopher and founder of utilitarianism, Jeremy Bentham (1748–1832). With a total of around 75,000 Bentham manuscripts to sift through and edit, the Bentham Project faces a formidable task. In 2010, the Transcribe Bentham initiative was launched in the hope that members of the public could be encouraged to contribute to the project. Volunteers have surpassed all expectations, having accurately transcribed over 16,000 manuscript pages and counting. These transcripts are used by Bentham Project researchers in their editorial work and are also hugely important in spreading awareness of Bentham's philosophy.

Simplifying transcription

The Bentham Project is hopeful that HTR technology could make the task of transcription simpler, thereby encouraging new volunteers to take part and increasing the productivity of existing participants. The Read project is building an e-learning tool, which will allow users to train themselves to read Bentham's rather indecipherable handwriting. This should give new users a boost of confidence before they get to work. Read will also be creating a new version of the Transcribe

Bentham crowdsourcing platform, where users can transcribe with the assistance of HTR. If a volunteer comes across a word they cannot read, they will be able to ask the computer to suggest what it might be. HTR technology could help to expand on the important work of Transcribe Bentham's volunteers.

Spreading awareness

The Read project is committed to spreading awareness of these innovations and ensuring that they are used by researchers, collection holders and members of the public. Tools developed by the project are open source[5] and research data and publications will be open access. Interested institutions can even become part of the Read project network by signing a 'Memorandum of Understanding', like the Linnean Society of London.

What should be in your digital toolbox?

As part of Read's outreach, the Bentham Project and the Linnean Society organised a conference which took place on 10 October. We asked archivists, librarians, curators, scientists and researchers 'What should be in your Digital Toolbox?'[6]. In addition to presentations by Read project managers and developers about the mind-boggling programming that underpins the project's HTR functionality, the conference talks showcased an interdisciplinary cross-section of projects from the Digital Humanities, the sciences and curatorial practice, which have one thing in common – their aim to unlock stubborn data and make it accessible to an array of disciplines and users. Projects included 'Asymmetrical Encounters'[7], a large-scale EU-funded project that uses text – and sentiment-mining in long runs of historical newspapers, as well as the metadata extraction and full text transcription of the extraordinary Zooniverse (https://www.zooniverse.org/), with its multitude of individual projects and disciplines – the world's largest and most popular platform for people-powered research.

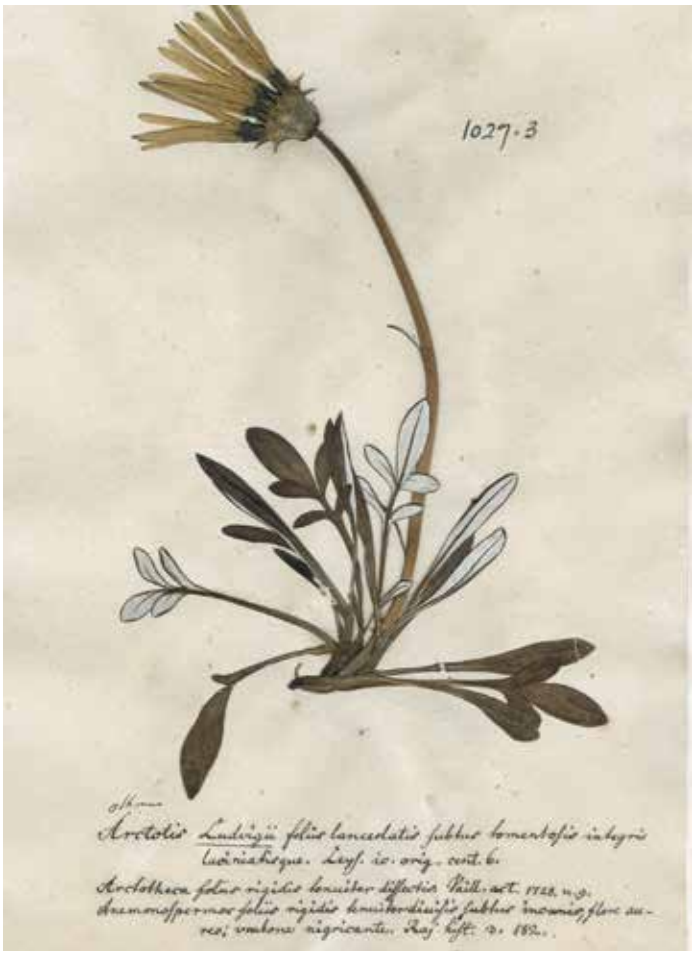
It was a great privilege to have Melissa Terras, Professor of Digital Humanities at UCL, as the keynote speaker. She deftly pulled together the various topics discussed at the conference – the tools used to extract and visualise data, and how we should use these tools for innovation. Crowdsourcing, citizen science, outreach and learning are all part of the package. Collaboration is key, establishing interdisciplinary partnerships between researchers and enthusiasts with holding institutions, matching curatorial requirements with human expertise and curiosity, and, last but not least, programming power. Sources for funding are equally important. The Read project and its European/UK partner institutions have received generous funding from the European Union's Horizon 2020 research and innovation programme, which ensures that the outputs of the project will be open to all, stimulating research and innovation beyond corporate interests and restrictions, and enabling universities and holding institutions like the Linnean Society to be part of a technological revolution. The future of UK institutions to get involved and to continue to benefit from projects like Read is currently unclear.

At the conference, the librarians, archivists and curators amongst us got truly excited about the possibility of using digitised images of large and complex collections and HTR for initial sorting and cataloguing. This will enable institutions, even in the current difficult funding climate, to start tackling collections which would otherwise take decades to process. If the prospect of searching handwritten documents appeals to you, why not download Transkribus and try it out? By feeding data into the platform, you are helping to strengthen the power of HTR technology and making it easier for everyone to access manuscript material. The Digital Toolbox improves the more it is used – the future is all yours. [T]

References

1 http://read.transkribus.eu
2 UCL www.ucl.ac.uk/Bentham-Project
3 https://www.linnean.org
4 https://transkribus.eu/Transkribus
5 https://github.com/Transkribus
6 https://www.linnean.org/meetings-and-events/events/what-should-be-in-your-digital-toolbox
7 http://www.ucl.ac.uk/asymmetrical-encounters

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 674943.



The Manuscript Collection of Carl Linnaeus includes a wealth of complex material – the layout is often more of a challenge than the handwriting.