[This a draft text of the talk presented at the 13th Annual Schoenberg Symposium on Manuscript Studies in the Digital Age, November 18 2020. As such, it is a close but not verbatim transcript of the actual presentation.]

Digital and Computational Palaeography: Some Promises and Problems

Peter Stokes, EPHE – PSL

As those of us here today know very well, the last fifteen or twenty years or so have seen an enormous increase in the availability of digital images of manuscripts, as well as tools for working with them. This has been enabled by technological developments, but also by what one might call 'social' changes, such as the increasing prevalence of more or less open access to images and indeed to software; investments at local, national and international level in the acquisition and publication of digital images as well as in open platforms and software; and concerted work on international standards for images, particularly the International Image Interoperability Framework, or IIIF. **[Slide]** Here, for instance we have the Biblissima Portal, which provides searching across digital images of around 75,000 different Western manuscripts, from a dozen or so different libraries, all of which have images that are immediately accessible through IIIF.

I think it's clear that this is (or at least can be) an enormous boon. As noted in the overview to this conference, in many ways we were indeed 'ready' when most of the world's manuscript libraries became inaccessible in March, meaning that I can sit at home during what is now France's second lockdown and consult images of tens of thousands of manuscripts scattered across Europe and beyond. **[Slide]** It also poses interesting questions and challenges, some of which I will just put here for now and will return to them in a minute.

But there are of course limitations here as well, and others (including many in the audience) have spoken repeatedly about this. We know, I think, that availability of digital images is *not* access to the object, however easy it can be to forget this. I think we all agree that to understand manuscripts requires not only the digital but *also* direct physical access to the original objects. There *is* a real risk that more resources in digitisation can mean fewer resources for librarians, cataloguing, conservation and so on (and I'm thinking here particularly of examples like the National Library of Israel which suspended all activities and put 300 staff on temporary leave, citing COVID and budget cuts). I don't think it's a zero-sum game, by the way, as there's no question that greater exposure to these materials has led to greater interest in palaeography, codicology and manuscript studies in general, although just how this translates into funding is much less clear at this point. None of this is new to COVID, of course, but it does seem that the current situation is accelerating this process. One sign of that already is that the new Horizon Europe programme, which sets European funding priorities for the next seven years, is expected to include calls aimed at increasing digital approaches to cultural heritage, with explicit reference to COVID as a motivating factor.

Still, we have long been talking about digitisation as 'democratisation', and this is a question that concerns me a lot. I am very aware that I am extremely privileged, having spent the whole of my career in Cambridge (England), London and Paris, and I am extremely grateful for that. It has meant that I have had relatively easy access to medieval manuscripts, and also that it has been relatively easy for me to meet the 'big names in the field' as I simply had to sit in the manuscript libraries and, in effect, people would come to me. So it's easy for me to say that access to manuscripts is essential, but what does that say about people who don't have this privilege? One of the ideals of the digital humanities and of digitisation has been to try to share this more widely by increasing access, at least in some form. Palaeography teaching is very different from when I started twenty years ago, where now students can go and find hundreds or even thousands of high-quality colour images of manuscripts, a situation very different from the highly selective handful of black and white photocopies that we used to have. I teach Master's students in DH, and **[Slide]** it's well within the capability of good students to write python scripts that harvest hundreds of images of manuscript pages, processes the image to automatically estimate basic statistics such as the dimensions or average number of lines of text on a page, and then look for developments over time, space and so on. There have been very interesting projects around crowdsourcing and gamesourcing to raise interest and teach skills, and one might hope that this in turn leads to more investment in posts and training, but I'm by no means sure about this yet.

At the same time, though, another encouraging trend that I have seen is that there has been more and more interest from the Computer Science community in the analysis of historical documents. **[Slide]** When Arianna Ciula first published her seminal article on digital palaeography about fifteen years ago, there were very few people indeed from Computer Science who were interested in these questions (with a couple of notable exceptions such as Lambert Schomaker from Groningen who is still a leader in the field). Fifteen years ago, many Computer Scientists were telling me that historical documents were not interesting and a 'solved problem', but now the International Conference in Document Analysis and Recognition has a regular dedicated section on historical documents; the International Conference on Frontiers in Handwriting Recognition is growing rapidly; and in these conferences there are very open discussions of how important it is to involve people from the Humanities in this work.

So from here I want to change direction a little and consider some new areas that are emerging in 'digital palaeography', in part again as a result of increasing 'access' or 'democratisation', if you will. I think the area where this is most immediately evident, and changing extremely quickly, is in artificial intelligence and Deep Learning. Those of you who know me well may be surprised by this, as I consider myself relatively averse to fashionable buzzwords, and indeed I spent many years arguing *against* many uses of AI in palaeography. To be clear, I still do: but not against *all* uses, and things are changing so fast that I do think now that we are already at a time where we have clear examples that work extremely well. In fact, AI and palaeography have in a sense had a long history together: Douglas Hofstadter, a famous writer and researcher in AI, was very much preoccupied by writing right back in the 80s, and he once stated that [**Slide**] 'the central problem of AI is the question: what is the letter *a*?'. (In fact this is a question which has obsessed me for many years now. What *is* the letter a? *Really*? And what does that say for palaeography that we can't even answer a question as fundamental as this? Today, though, people are throwing AI at these questions in their spare time. [**Slide**] Here, Erik Bernhardsson apparently decided more or less on a whim to download 50,000 computer fonts, train a neural network on them, carry out various statistical analyses, and then program the network to generate new characters in different styles. Now, Mr Bernhardsson is clearly a skilled programmer, but still, this is something he did at home in his spare time.

Another example is a project that my team in Paris is working on, called Kraken/eScriptorium, **[Slide]** which uses machine learning to automatically transcribe manuscripts. This software is available online, and you can download it and use it yourself (if you have the necessary technical skills). As you probably know, this isn't magic and someone needs to teach the computer how to do the transcription, by showing it many lines of material that has already been transcribed. This requires a lot of work, as someone needs to sit down and prepare the transcription, so an active area of research is how to reduce this as we will see in a minute. Now, one of the advantages of Kraken/eScriptorium is that the trained models themselves are also open and can be freely exported and shared, and as far as I know this is pretty much the only platform where it's the case. So I can train a model on my manuscripts from the eleventh century, say, and then export that model, and give it to you, and you can then retrain it for your manuscripts which are similar to mine but not quite the same. This saves a huge amount of time, and indeed energy and environmental resources, compared to us all retraining our models from scratch.

This again opens up all sorts of interesting opportunities. Think about it: we now have access to tens of thousands (probably hundreds of thousands, if not millions) of images of manuscripts. We are starting to get trained models capable of transcribing those images. It won't be perfect, sure, but even if it's, say, 80% correct, that's more than enough to automatically identify the texts, and even probably which variants or versions of texts there are. It may be enough to do some basic linguistic analysis, such as identification of dialect. It's certainly more than enough to do things like check the number of lines per page (a useful complement to data in the Schoenberg database, for instance). It may be enough to identify ownership inscriptions, glosses, and other information. It's not quite true that we can do all this on our home computers in practice, because for the moment AI requires fairly high-powered computing to run in a reasonable time, but even this is changing rapidly as well (NVIDIA for instance has released an ultra-cheap hobbyist version of a GPU as a sort of Rasberry Pi for AI).

In fact I mentioned that there is active work on speeding up this process, and, believe it or not, this leads to another area of 'democratisation' in AI: Deep Fakes. I'm sure you've heard of this: it’s a fairly new technology based again on neural networks, that these days seems mostly used to put your face onto that of a famous celebrity in ways that can be genuinely difficult to detect. But this technology is also being used by Computer Scientists for document analysis. Yes, you hard that right, [**Slide**] people are Deep Faking manuscripts!

This raises an obvious question: why would you want to do that? Well, I can think of very interesting if disturbing applications for outreach and engagement, though as far as I know this hasn’t been done in practice. But there is also a very valid reason for Computer Science. I mentioned that we need examples of transcribed manuscripts in order to teach the computer to do the transcription itself, and obviously this transcription takes a lot of work to produce. But another option, which works surprisingly well, is that we can Deep Fake images of handwriting in the style of our document, and in this case we already know what the text says so we don’t need to prepare our own transcription for training purposes. Instead, we can ‘teach’ our machine to read the Deep Faked manuscripts, and then apply it to real manuscripts where we don’t know already know the text, and it turns out that the results are surprisingly good!

So, where does this leave us? In fact, where in all this is the palaeography? Donald Knuth once wrote that **[Slide]** ‘the best way to understand something is to know it so well that you can teach it to a computer’ (and he added that ‘the process of seeking such explanations will surely be instructive for all concerned’). But have we not solved what Hofstadter identified as ‘the central problem of AI’: have we not taught a computer to successfully identify a in thousands of different fonts and handwriting styles? **[Slide]** If this commentary on Reddit is anything to go by then yes: according to them, we are done, the problem is solved. But what does this tell us really? Do we now understand this so well that we can teach it to a computer? The answer in fact is clearly ‘no’, I think: we have shown the computer a series of images and told it what we think, but it hasn’t yet done much for our understanding, and *this* Is one of the main reasons why I have always been very wary of highly complex algorithmic analyses such as Deep Learning: because of the enormous problem of algorithmic transparency and understanding. This is now, finally, largely recognised among the Computer Science community [**Slide x 2**], and there are now very good people working very hard on these questions. But I still think these questions apply just as much to the

Digital Humanities, and that the real question from our point of view is not so much 'what is the answer?' as 'what does this mean?'.

Indeed, these systems can give us ***an*** answer, but we will always need *many*, *different* answers. As Collette Sirat has noted, Carl Popper's point is very relevant also to palaeography: **[Slide]** 'Two things which are similar are always similar in certain respects. … Generally, similarity, and with it repetition, always presupposes the adoption of a *point of view*: some similarities or repetitions will strike us if we are interested in one problem or another'. In order for our digital methods to be useful, then, we need to find ways of allowing for these different points of view, and this includes contexts different from our own.

One of the many privileges of being at the EPHE is that I have colleagues working on pretty much every historical script you can imagine, and many more that you can't. At the moment, though, the vast majority of online tools and models are designed at least implicitly and often explicitly for Western, often English, documents. So when we're talking about wider access in a world of COVID lockdown, how are those working on less representative materials managing? The answer is often 'not very well'. We need to allow for different characters, different scripts, different directions of writing, different conventions for transcription and scholarly presentation. We need to be able to handle cases where we don't already have millions of words already available, where there may not already be trained models at all, or where the corpus may be very small. We need to be able to treat pages like this **[Slide]**, where the writing is in Arabic, so from right to left, but with glosses radiating out in this star-like form.

To take another example, for sharing transcription data, you need to specify the baseline. But what is the baseline, exactly? **[Slide]** And these are the easy cases: we then have ones like this **[Slide]**. Yes, I *could* model Hebrew as being right to left written on a baseline, but if I then share my data with yours, did you model it this way, or did you do it right to left from a topline? How can I tell?

Even very basic definitions start to fall apart very quickly once you move outside an Anglo-European context: what exactly is a grapheme, for instance? What exactly is a word?

I talked a bit already about sharing trained models, but sharing any models and software (whether trained AI, or hand-built databases, or whatever) requires a huge amount of communication, understanding, and making knowledge explicit, and this is still one of the real challenges, I think. *This*, I think, is what Knuth meant when he wrote about teaching the computer. **[Slide]** (For what it's worth, here is the start of my answer to the question 'what is the letter *a*'.) There are a couple of small groups working on these questions, and here is one of my contributions to the subject, but relatively speaking there is much less going on here, I think.

So COVID has accelerated a process of digitisation, and this with the current trends look like more funds for digital work on cultural heritage, at least in Europe, but with the economic crisis likely meaning less for so-called 'traditional' palaeography, librarianship and so on. So I think the old questions still apply: We need to think how to be relevant without oversimplifying, how to engage but in the right way, how to pool our resources and be as efficient as possible without privileging a few points of view that are already dominant. These challenges are not new, but they look likely to become more urgent than ever.

[Slide] Thank you