Sharing Data for Handwritten Text Recognition (HTR)

Peter Stokes, Benjamin Kiessling

Sharing Data for Handwritten Text Recognition (HTR)

Peter Stokes and Benjamin Kiessling
École Pratique des Hautes Études – Université PSL
Archéologie et Philologie d'Orient et d'Occident (UMR 8546)

Abstract

Handwritten Text Recognition (HTR) is at present perhaps the principal application of Artificial Intelligence to the Digital Humanities. It falls under the category of supervised machine learning, and this in turn depends almost entirely on the data that is used for training. The consequences of this are numerous: these techniques are therefore positivist, insofar as they are only applicable to cases where the answer can be known and defined in advance; they will necessarily reflect historical practice; and they will also reflect all of the inevitable biases that are present in the data (for just one example of which see Brown et al., 2020, 36–39). The data itself also becomes valuable and so a commodity in its own right and, conversely, the availability or lack of data is itself shaping decisions about applications in machine learning. For HTR, this means that the most progress has been on modern material in widely-used languages and writing systems such as English and others written in the Latin alphabet, while so-called rare and historical scripts have seen much less success. It also suggests many communal benefits in publishing and sharing training data, in order to combine effort and expertise and avoid unhelpful repetition of labor. Such sharing brings many challenges, as it requires standards or at least common practices in transcription (including treatment of abbreviations, punctuation, "non-standard" spelling and capitalization and so on), as well as standards for data sharing that allow for the many important variations in the world's writing systems. It also assumes the willingness (and the possibility) of sharing data openly, including transcriptions and images, but this in turn can depend on many different institutions and interests. Despite these challenges, some initiatives have begun to point the way, and the benefits of this work are already being felt, which is encouraging for the future of the field.

The Challenge of Big Data for "Rare" Languages

Handwritten Text Recognition (HTR) normally refers to the automatic transcription of handwritten texts in manuscripts: in other words, a system that takes as its input digital images of handwritten books or other documents, and gives as its output a transcription of the corresponding text. For this reason HTR is sometimes referred to as "OCR for manuscripts", and in fact at present there is often no technical difference between OCR and HTR since the same methods are normally applied to handwritten and to printed texts.[1] The key point for this article is that the current methods for HTR all rely on machine learning, and this in turn means that "Ground Truth" data is required for training the model. In short, the principle is that the machine must first "learn" to carry out the HTR, and this means that it must be provided with examples of many thousands or millions of images of lines of text, along with the transcription that should be given as a result. The machine is then trained based on these examples,

[1] The literature shows different definitions for OCR and HTR with no clear agreement on the meaning. For this reason, HTR is used throughout to refer in general to the application of machine learning methods to the automatic transcription of images of text, whether handwritten or printed.

creating a model which can then be used to transcribe new images that were not used in the training data.

Even if not Big Data in the strict sense,[2] the discussion above makes clear that many existing methods for machine learning require very large amounts of data at least by the standards of most researchers in the Humanities. For instance, ChatGPT has been receiving a good deal of media coverage at the time of writing and so in some ways represents the State of the Art, but the already outdated GPT 3 model required over 200 billion words of Ground Truth text (Brown et al. 2020, 24 and 46), and this suggests that the cost of training alone is on the order of millions of US dollars. This is challenging for any language or script but is at least possible for modern English and some other modern majority languages. However, it is many times larger than the entire surviving corpus of many so-called "rare" scripts or languages. ChatGPT is of course a "generative" process, namely it is designed to generate new content rather than to read existing texts, and so in this respect is very different from HTR. Nevertheless, the current state of the art for much of machine learning tends to rely on large quantities of data and so the methods do not work for rare languages or scripts. This has several consequences. The first is the risk and indeed reality that only some types of material are being treated and are visible: specifically, machine learning is being applied primarily to contemporary and "mainstream" material that is readily available in large quantities, and this in turn helps to reinforce the dominant position of that material and their languages and scripts. To some extent this has always been the case, with access to material in majority scripts and languages being much easier and more widespread than that in "rare" materials and scripts, but this runs counter to the ideal of the internet and of modern technology as enabling visibility of and access to a much wider range of material. Instead, specific methods are required to address this, some of which are discussed under "Methods to reduce data requirements" below.

This leads to another well-known difficulty in machine learning, that of bias in training, particularly in the case of rare and historical documents, and recognizing and overcoming these biases is central to success in the field. Machine learning models are deeply influenced by selection in the digitization and publication of accessible data, and so if we use only data that is already easily accessible then we will continue to reproduce existing biases. This is well known, but other problems are much more subtle. For instance, one may think that the transcription of historical texts is entirely objective and free from ethical questions. However, as discussed further below, transcription always requires choices, and these choices must necessarily be conveyed to the HTR engine via the training data. It is therefore not difficult to imagine certain cases where the systematic choice of one form over another could alter the sense of the text, and this in turn could potentially have a real impact on people's lives particularly for religious or historical material.

A further complication is that of engineering. This may not seem relevant to data at first, but generating and using data in practice requires not only human effort, expertise and standards, but also algorithms implemented in workable tools, and this in turn requires effort and investment in engineering. A large part of modern computing software and hardware was first developed in the context of modern English, typically from the United States. One consequence of this is that modern computers still rely on

[2] There is no clear definition of "Big Data", but it is normally taken to mean very large amounts of data (in the order of petabytes), which is highly heterogeneous, and which arrives very rapidly in real time: in other words, it is "big" insofar as the combination of volume, variety and velocity make it impossible to treat with traditional methods. Datasets for historical HTR rarely reach the size of petabytes, and are usually homogeneous and limited in velocity, but they are too large to be treated by traditional methods (transcription by hand).

assumptions that lie very deeply in the system, such as that writing necessarily comprises a single sequence of elements that progress in a single direction across the page, and that these elements are drawn from a defined and relatively small set of possible signs. However, these assumptions do not necessarily apply to all possible languages and writing-systems, particularly those that are historical or "other" from a Western point of view. One result of this is that the effort required for developing tools that are truly capable of managing all possible current and historical languages and scripts is enormously more complex and demanding than for any one system. These problems are increasingly recognized, and significant and important movements have helped to reduce the difficulties that are found here, perhaps the best-known being the Unicode standard which is a far cry from the American Standard Code for Information Interchange (ASCII). Indeed, a great deal of discussion and tools can be found to meet the challenges of what is now called internationalization, that is, of developing software that is so-called "world ready", including that it should work seamlessly for any language or script, without requiring manual intervention to correct for less common situations that do not fit developers' expectations (see, for instance, Stokes et al. 2021, §4). In practice, however, software libraries and tools and their underlying algorithms are still almost always developed with inbuilt assumptions that limit the usefulness to well-known majority languages and scripts and which therefore can require substantial manual intervention to allow for other cases for which these do not hold. These interventions are not necessarily difficult in themselves, but each one requires further engineering effort, and while individually they may be small the combination can require very significant financial and human resources. These resources are often difficult to find, because on the one the high degree of specialization can make them of little interest to a wider public, and on the other because the work often falls under the rubric of "resource development" or "software development," and so it is considered outside the remit of most research funding agencies.

Openness of data and models[3]

The discussion above suggests strongly that an effective way forward is to do as much as possible to combine efforts and share the data that we have available as Ground Truth for training models, and indeed to share the models themselves. The advantages seem obvious, particularly for so-called "rare" scripts and languages, given the difficulties in accessing the corpora, the high level of expertise required to prepare the data, the very limited financial and human resources that are normally available for such work, and other challenges discussed above. This suggests that we should be doing all we can to reduce duplication of effort, not only in human terms but also in electricity and computational resources, since training Deep Learning models is extremely intensive and so retraining multiple models for the same data is an enormous waste that is increasingly difficult to justify. In practice, however, many systems for HTR and indeed for Deep Learning in general do not allow this, and in fact they often actively disallow it. As our society is finally beginning to notice, data is extremely valuable, with our personal data worth billions of dollars a year as demonstrated by the success of the companies such as Meta and Alphabet, and although it is by no means the same scale, even data for HTR can and has been monetized. The most obvious example in this context is the READ Coop which changed the funding model for the Transkribus HTR platform in 2020, such that the models are trained by data provided by users, but the users must now pay to use those models for automatic transcription, without a way of exporting the models for use outside the system. The principle of "user-pays" is understandable – someone must pay at some point if

[3] Some of the points in this section have also been discussed in shorter form by Stokes et al. 2021, §3.

the service is to be sustainable – but the fact remains that the models are monetized based on freely-contributed labor, much of which was done while the platform was freely accessible and was not changed with the consent of those contributing. Furthermore, this approach means that users are largely locked into the platform, leaving them vulnerable to the inevitable point in future when they no longer have access to the platform, whether because of increases in access fees or because the software itself is no longer supported. Users can – and should – export and publish their training data, but many do not, and even doing so still requires that the models must be retrained on other systems, with resulting duplication of time, effort, and energy.

An alternative model is that the trained models themselves can also be exported, published and reused, alongside the Ground Truth data for training. The approach here is for truly open software that can be downloaded and installed by different teams on different servers, with the teams publishing and openly sharing their trained models on platforms such as Zenodo and GitHub. This approach is used by software such as kraken/eScriptorium, developed by the authors of this chapter, and collections of trained models can already be found online such as those in the OCR/HTR Model Repository Community on Zenodo.[4] This approach has the advantage of reducing the waste of human and electric energy, as well as helping sustainability on the principle that, if any one instance of eScriptorium is no longer accessible for any reason, other instances can continue, and users can easily move their data and models from one instance to another. There is no obligation to share data, and indeed there are reasons why one might not be able to do so, for instance if the images are restricted due to issues of copyright or the sensitive nature of the content, but we argue that the possibility to do so is important, and that the publishing and finding of models and data should be as easy as possible for users. This also has scientific benefits in addition to the financial, environmental and ethical ones listed above, since it allows for greater transparency and reproducibility in transcriptions, helping to reduce at least some of the "black box" effect that is often cited as a difficulty in machine learning methods, both to help identify biases (for which see further below) and to help ensure scientific transparency and accountability (Stokes 2020b, 39–40, as well as Kestemont et al. 2017, 105–108, among many others). In principle, open publication of Ground Truth and models can also allow for more attribution of credit, since Ground Truth and models can and should be cited in publications like all other datasets, even if this is difficult in practice for models that are the result of many different stages of training on many different and potentially heterogeneous datasets. The ideal here is that the models and data are published on research data infrastructures such as Zenodo or other infrastructures for research data, since this then provides not only long-term accessible storage, but it also attributes a persistent identifier to each version which allows one to be very precise when citing and solves many of the difficulties in keeping track of the augmentations and ameliorations of training data, the retraining of models, and of understanding exactly which data was used to train each specific version of a model. This is why kraken includes a tool to publish models directly to Zenodo, and to import existing models from the repository if they have been published in such a way that the tool can find them. Other similar initiatives include OCR-D and HTR-United. The first provides detailed guidelines for transcription and Ground Truth, as well as specifications for interchange formats, models, modules and so on (OCR-D, n.d.). HTR-United is similar but on a much smaller scale, encouraging the publication of Ground Truth data (and, ideally, also trained models) and providing tools to help ensure consistency of the data and accurate and relatively straightforward recording of metadata to help findability and reuse (Chagué and Clérice 2022). Neither

[4] https://zenodo.org/communities/ocr_models

HTR-United nor OCR-D is specific to kraken models, but in practice most of the datasets in HTR-United in particular are used in kraken since it is one of the very few effective HTR engines that allows such ease of import and export.

Standards and practices in transcription

As we have seen, at least on the face of it, the most efficient means of collecting Ground Truth data for training models is to share the data and indeed the trained models as much as possible. However, this approach depends on the implicit assumption that transcriptions are interchangeable, and this in turn leads directly to an ongoing discussion in philology. Although some have suggested that transcription is the purely objective recording of "what is on the page", it seems clear that this applies at best to printed books and little – if at all – to manuscripts. As scholars such as Peter Shillingsburg (2006, 151–160), Michael Sperberg-McQueen (2009), and others have repeatedly reminded us, a manuscript (and indeed a printed book) is a complex object that can be represented in an infinite number of different ways, and so any transcription is necessarily a selection, with all the assumptions and biases that this implies. Indeed, in practice even so-called "type facsimile" or "graphematic" editions are often implicitly driven by technological limitations: historically by that which can be printed on the page, and more recently by that which can be represented in Unicode (Pierazzo, 2011). Furthermore, both editions and transcriptions are not all the same but are produced following different principles according to their different purposes and different types of texts. We therefore find critical editions which seek to reconstruct the author's hypothetical original work, diplomatic editions that seek to accurately represent a specific document, graphematic or type-facsimile editions that seek to represent at least partly the layout and different scripts on the page, facsimile editions that combine text and image, and so on. While these too are partly driven by technological limitations, they also depend on the editor's view of the text and physical object. For instance, texts in medieval European vernacular languages are often very unstable, existing in many copies that are very different from each other and sometimes reflecting modifications and additions from different people, meaning that the very idea of an authorial original does not apply. In these cases, the interest is often rather in identifying a specific version of the work, for instance one that was read by a specific group or person, or at a certain period. Similarly, transcriptions for linguistic analysis may be in order to study the language of the author, but they may also be to study the orthography of a given scribe or a given region, and the choices made when transcribing will necessarily vary accordingly (for which see especially Pierazzo 2015, 85–101, as well as Stokes et al. 2021, §5, Stokes 2020a, 50–54, Robinson 2013, 116, and Sperberg-McQueen 1991, among others).

One direct consequence of this multiplicity is that there will necessarily be different ways of transcribing a given manuscript. Given that there cannot ever be one transcription that suits all purposes, it follows also that there can never be one model for automatic transcription, or one Ground Truth transcription for training models. This necessarily limits the degree to which we can share our data in practice, since my data is only useful to you if it meets your requirements in a transcription, or at least if it can be transformed into something that works for you. This might suggest that transcriptions should include as much information as possible, for instance including both normalized and literal transcriptions of texts, with markup indicating which is which, presumably using a standard format such as XML TEI. To some extent this would indeed be helpful but only to a point: such markup will also make the transcription

more difficult to use, since it would need to be first transformed into the required state for training, and this in turn means closely studying the file(s) to understand exactly how they were produced and exactly what would be needed. It also would significantly increase the time and effort to produce the transcription to begin with, since in effect the transcriber(s) would be producing multiple versions of the text, adding significantly to the complexity of the task and to the risk of error. Finally, even this would be limited, since, as noted above, the range of facts that could be recorded of a text are infinite, and so even a multifaceted transcription would still necessarily constitute a choice which cannot ever meet all possible needs. In this respect, then, there is a tension or even a contradiction: on the one hand, many purposes require only the "pure" text without any markup, but at the same time the "pure" text is a fantasy that does not and cannot exist. Adding XML markup does not mean that the text is no longer free of interpretation as some have suggested (Schmid 2009, 349; for arguments against see Rovira et al. 2009 among others) but it does add a dimension of information which cannot be taken into account by many tasks such as automatic transcription. Instead, a more feasible approach is to admit the plurality of texts and so agree within communities of practice on a set of different transcription principles for specific purposes, for instance at different levels of normalization from the more diplomatic to the more edited (for an example see OCR-D n.d). It is inevitable that these will vary in the details from one community to another, but it is feasible that each community for a specific domain or field of research could develop its own similar definitions to enable sharing data.

As this discussion suggests, sharing data in a multiplicity of possible transcriptions in turn requires sharing standards in transcription, and being very explicit in those standards. In order to achieve this, we need to be very clear and explicit in the principles that we use for transcription, being sure to record these and publish them for other users, something that many of us are not very practiced in doing. Indeed, it is often not at all evident the degree to which one intervenes in the text, as many "obvious" emendations are applied without even noticing. A striking example of this is given by Elena Pierazzo's edition of the *Stufaiuolo* of Anton Francesco Doni (Pierazzo, 2015b). As part of this work, she tried as far as possible to make explicit every intervention in the text, categorizing each one such as adding or removing accents, adding or removing apostrophes, regularizing graphemes (i and j, u and v, ß and s and so on), modernizing word spacing, modernizing the use of upper and lower case, and so on. The text is short, filling approximately 20 manuscript pages with around 11,000 words, but she counted over 4,500 editorial interventions in each of the two manuscripts, so around 9,000 interventions in total, meaning almost one intervention in every word. Although this may sound extreme, as she shows these interventions are "obvious" ones that are often made silently. It is also clear that this work is an edition and not a transcription, and some normalizations such as spelling might not normally be applied to the latter, but the line between the two is blurry at best, and users of HTR engines can exploit this by training models to carry out some of this normalization automatically in order to reduce manual labor in future steps (see, for example, Camps et al. 2021). This is pragmatically valuable and scientifically justifiable, but it almost inevitably results in hundreds of decisions such as word separation and interpretation of accents, all of which need to be clearly documented if the material is to be used as Ground Truth.

The problem of standards is even greater when considering the full variety of the world's writing systems, and particularly when different systems are combined into a single document. For instance, many systems for identifying page layout for HTR require that the line of writing be identified in the data. This line is typically assumed to be the baseline, with the writing on top of it, since this is

normative for most European writing-systems and therefore also for those used in the United States and across the Anglophone world. However, this is by no means universal: Hebrew, for instance, is (or at least can be) written beneath a topline, while in Mongolian the "baseline" is oriented vertically, from top to bottom (Ishida 2019), whereas scripts written vertically in columns such as some Japanese and Chinese have no line to speak of and so a center line can be added if required for processing purposes. Similarly, if a baseline of English text runs horizontally across the page, then the normal assumption is that the writing should be above this line and running from left to right. However, this ignores the possibility that the line of writing might be rotated 180° with respect to the remainder of the page: in this case, relative to the orientation of the image the writing will be upside-down, below the baseline and running from right to left. The situation becomes even more complex with some historical scripts which survive from before the writing direction became standardized for that culture. For instance, Greek inscriptions can be found which run not only from left to right and top to bottom, but also right to left then bottom to top, horizontally boustrophedon (left to right then right to left, with lines from top to bottom), and vertically boustrophedon (so top to bottom then bottom to top, with lines proceeding from right to left). Furthermore, the orientation of the characters changes to face the line of writing in these inscriptions, just as it does in other writing systems such as Egyptian hieroglyphics. This means that, for example, a K (Greek letter kappa) is written in the way we would expect it when the text runs from left to right, but is written "backwards," that is reflected horizontally, when in a context from right to left. This relatively lengthy and detailed discussion demonstrates again that specifying a line of writing is not sufficient: instead, a suitably general standard must specify the line, which type of line it is (baseline, topline, centerline) and its orientation relative to the image (upside down, right way up, rotated) in order to be unambiguous. While existing standards such as ALTO and PageXML provide for more or less sophisticated mechanism to encode such writing, the rarity with which these features are encountered often results in confusion on how the standards should be employed, as well as a general lack of support in the software. For instance, even if international standards such as Unicode and the W3C allow for mixing both left to right, right to left script, and top to bottom scripts (Davis, Lanin and Glass 2020), in practice these often fail in specific libraries that are used as the building-blocks in software design.

Methods to reduce data requirements

An ostensibly attractive way to circumvent the issues mentioned above when integrating heterogenous training data for HTR is to reduce the amount of training data that is required, such that producing models "from scratch" becomes feasible for all but the lowest-resourced material. This in fact is a thriving field of research, and so researchers in machine learning have developed a range of methods to reduce the amount and complexity of training data necessary to produce models, while still retaining the generalization that is essential to an effective process. These methods range from almost universally deployed approaches like pretraining on surrogate tasks (the most widely used probably being to train backbone networks on ImageNet classification), through augmentation and outright synthetic generation of datasets, to various ways to reduce the level of supervision required such as semi- or self-supervision.

While these methods are sometimes capable of achieving impressive results from very minimal datasets, their practical application often suffers from the same issues that leads one to seek them in

the first place. In the context of HTR, one such example is self-supervision. On the most basic level, this is the automatic distillation of new training data from previously unlabeled data with the machine learning model itself, and it is most often built around a language model scoring the quality of the recognition model's output on images that lack corresponding transcriptions. Only output fulfilling certain thresholds of quality, such as lexicality or word distribution, is then added to the pool of existing training data. A standard approach is to go through multiple iterations of this procedure, each time producing more powerful recognition models from a limited number of initial manual transcriptions. In practical applications, these methods exhibit multiple sources for unwanted normalizing drift, which principally stems from the tendency of HTR models to learn not only a purely visual model of the handwriting on a page but also basic relations between the individual graphemes. To identify which automatic transcriptions are usable and sufficiently accurate for training purposes, we most often compare the results to typical language use, resulting in unusual samples being sifted out with a higher probability than more typical, higher frequency samples. However, depending on the capabilities of the system employed for this filtering, this may mean that the system rejects results that show "non-standard" characteristics in orthography, use of abbreviations and so on. This in turn can mean that the model learns a version of the text that is too regularized, replacing the idiosyncrasies of individual copies and scribes with that of a normalized "standard" form. This may be useful in some cases, but researchers in the humanities are often interested particularly in the exceptions, and so an automatic and largely invisible normalization in these cases is clearly undesirable.

Even when some kind of regularization is acceptable or desired, a second major obstacle exists for low resource material. Powerful language models exist for modern high resource languages, with the most recent ones using a substantial proportion of all the text that is available in digital form: in practice, this normally means harvesting a large percentage of all material available on the Internet. However, assembling sufficiently large quantities of ground truth can be challenging for historical languages, since in many cases there simply is not enough content that survives, let alone the relatively small percentage that is digitized and available in accessible form. Furthermore, the cost of training exceptionally large language models such as BERT and GPT can also be punitively expensive for historical and "rare" languages, particularly as the commercial interests are very limited indeed and so the costs must often be borne by public institutions and research grants. Even if fundamentally feasible, the training pipelines for language models typically include steps for extensive normalization of texts, and this again makes them unsuitable for historical and minority languages for reasons noted above. In these cases, extensive adaptation of the pipelines is required, and this again requires financial and human resources which may well not be available. The task of reconciling HTR datasets with potentially conflicting transcription standards therefore becomes a task of reconciling the different vocabularies of HTR systems and language models, as well as accounting for the biases induced by them. In addition to the increased complexity of analyzing the behavior of such a multi-faceted system, the activities required are also significantly less rooted in existing scholarly practice: most humanities scholars are acquainted with the kinds of normalization their colleagues might perform on a text for a particular task, as well as how to detect this normalization when looking at an existing transcription. However, it is safe to say that even skilled digital humanists are rarely able to dissect, understand and compensate for the biases and limitations of complex training pipelines containing multiple powerful machine learning models.

Paradoxically, less powerful methods that only offer modest reductions of required training data often reduce the risk of inadvertently training for an undesired target, for instance one that is overly

normalized. Two examples here are transfer learning, namely developing a new recognition model from an existing one trained on closely related writing, and another is synthesizing training data through augmentation. Both these approaches tend towards training characteristics where undesirable normalization is not hidden but is more transparent, because the overall accuracy of the final model correlates closely with proximity to the target domain. In simple terms, an existing model that has been fine-tuned with new training data on another handwriting style will perform quite poorly on this style overall if the amount of training data is insufficient to fully learn the transcription standards of the new training dataset. In these cases, deficiencies are easily detected, and they can be corrected through simply increasing the amount of manually prepared training data.

Conclusion

As this discussion has shown, the question of data for training automatic transcription is by no means straightforward, and – perhaps unsurprisingly – it is subject to many if not all of the same issues as other forms of machine learning. These challenges include the influences of implicit and explicit bias, and these in turn impact everything including the availability and selection of training data, which languages and writing systems are digitized, which are used for training, which are feasible given assumptions which are deeply imbedded into the technologies, and so on. There are no simple answers to any of these questions, but there are measures that can help to reduce the difficulties that they pose, in particular by ensuring that models are published and openly shared, along with the training data that was used to train them, all with properly controlled versioning and stable identifiers such as those provided by Zenodo or other data repositories. Also essential is further work on establishing standards for transcription and greater transparency in their use, and again one single standard will never work (in our opinion), but a viable compromise may well be to define a set of options for each domain depending on the type of transcription and the reasons for which it is produced.

These recommendations are not new, and indeed are consistent with principles of best practice for artificial intelligence more generally, as expressed (for instance) by the ACM's Statement on Principles for Responsible Algorithmic Systems (2022), or the Assessment List for Trustworthy Artificial Intelligence published by the Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission (2020). Both of these include transparency and reproducibility among the core requirements for trustworthy AI, including the ability to trace the data that was used for training, as well as the importance of diversity in training data. Automatic transcription of handwritten documents may seem narrow and much less significant than other applications of AI in our society, but issues of cultural hegemony and the representation and valorization of so-called "minority" languages and scripts are nevertheless important and not to be ignored.

Finally, OCR and even HTR may seem to be largely a "solved problem", and one dominated by very large companies with resources far beyond those of any research team in a university, library or other such institution. However, as this discussion as shown, this perception is valid only for certain languages and scripts, namely those that are relatively available, in very large quantities, and with sufficient commercial interest to justify the investment. However, this leaves all the other cases that require much more effort in engineering, much more specialized knowledge in reading and transcribing (thereby also meaning that they are often less amenable to crowdsourcing), and of an interest that is much less tangible and less directly financial. We cannot ignore the work that is being done in industry, and we

should take advantage of it when it meets our needs, but at the same time we must also lead the way in other directions, by cooperating and coordinating in order to ensure that other needs are also met.

Works Cited

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. "Language Models Are Few-Shot Learners." https://doi.org/10.48550/ARXIV.2005.14165.

Camps, Jean-Baptiste, Chahan Vidal-Gorène, and Marguerite Vernet. 2021. "Handling Heavily Abbreviated Manuscripts: HTR Engines vs Text Normalisation Approaches." In *Document Analysis and Recognition – ICDAR 2021 Workshops*, edited by Elisa H. Barney Smith and Umapada Pal, 12917:306–16. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-86159-9_21.

Chagué, Alix and Thibaut Clérice, ed. n.d. HTR-United. Accessed 16 March 2023. https://htr-united.github.io/

Davis, M., A. Lanin, and A. Glass. 2020. "Unicode Bidirectional Algorithm [Ref.42]." In *Unicode Standard 13.0*. Unicode Consortium. https://www.unicode.org/reports/tr9/tr9-42.html.

Kestemont, Mike, Vincent Christlein, and Dominique Stutzmann. 2017. "Artificial Paleography: Computational Approaches to Identifying Script Types in Medieval Manuscripts." *Speculum* 92 (S1): S86–109. https://doi.org/10.1086/694112.

OCR-D. n.d. "The Ground-Truth-Guidelines: Ground Truth Level Overview." Wolfenbüttel. Accessed 16 March 2023. https://ocr-d.de/en/gt-guidelines/trans/trLevels.html.

Pierazzo, Elena. 2011. "A Rationale of Digital Documentary Editions." *Literary and Linguistic Computing* 26 (4): 463–77. https://doi.org/10.1093/llc/fqr033.

Pierazzo, Elena. 2015. Digital Scholarly Editing: Theories, Models and Methods. Farnham: Ashgate.

Pierazzo, Elena. 2015b. "Lo Stufaiuolo by Anton Francesco Doni: A Scholarly Edition." *Scholarly Editing* 36. https://scholarlyediting.org/2015/editions/intro.stufaiuolo.html

Robinson, Peter M. W. 2009. "What Text Really Is Not, and Why Editors Have to Learn to Swim." *Literary and Linguistic Computing* 24 (1): 41–52. https://doi.org/10.1093/llc/fqn030.

Robinson, P. R. 2013. "Towards a Theory of Digital Editions." *Variants: The Journal of the European Society for Textual Scholarship* 10: 105–31. https://doi.org/10.1163/9789401209021_009.

Rovira, J. et al. 2010. "The Inadequacy of Embedded Markup". *Humanist* 23:776. Archived at https://humanist.kdl.kcl.ac.uk/Archives/Current/Humanist.vol23.txt

Schmidt, Desmond. 2010. "The Inadequacy of Embedded Markup for Cultural Heritage Texts." *Literary and Linguistic Computing* 25 (3): 337–56. https://doi.org/10.1093/llc/fqq007.

Shillingsburg, Peter L. 2006. *From Gutenberg to Google: Electronic Representations of Literary Texts*. Cambridge University Press. https://doi.org/10.1017/CBO9780511617942.

Sperberg-McQueen, C. Michael. 1991. “Text in the Electronic Age: Textual Study and Text Encoding with Examples from Medieval Text.” *Literary and Linguistic Computing* 6 (1): 32–46. https://doi.org/10.1093/llc/6.1.34.

Sperberg-McQueen, C. Michael. 2009. “How to Teach Your Edition How to Swim.” *Literary and Linguistic Computing* 24 (1): 27–52. https://doi.org/10.1093/llc/fqn034

Stokes, Peter A. 2020a. “Palaeography, Codicology and Stemmatology”. In *Handbook of Stemmatology: History, Methodology, Digital Approaches*, ed. by P. Roelli., 46–56. De Gruyter https://doi.org/10.1515/9783110684384-002

Stokes, Peter A. 2020b. “On Digital and Computational Humanities for Manuscript Studies: Where Have we Been, Where are we Going?” *Manuscript Cultures* 15: 37–46. https://www.csmc.uni-hamburg.de/publications/mc/files/articles/mc15-04-stokes.pdf

Stokes, Peter A., Benjamin Kiessling, Daniel Stökl Ben Ezra, Robin Tissot, and El Hassane Gargem. 2021. “The EScriptorium VRE for Manuscript Cultures.” In *Ancient Manuscripts and Virtual Research Environments*, edited by Claire Clivaz and Garrick V Allen. *Classics@* 18. https://classics-at.chs.harvard.edu/the-escriptorium-vre-for-manuscript-cultures/.