Reflections

Melissa Woo

9:30 AM One of my main takeaways from the Terras and Cordell readings for this class is how much the various topics that we've discussed so far all build on and interact with each other. Many of the discussions that we've had are informing broad topics that are integral to these readings and considerations of HTP/OCR in digitization efforts. For example, Terras' analysis of how HTR can and has been adopted into the library digitization effort calls to question how such digitized information should be made available, as well as the benefits and tradeoffs that come with being able to make such information available. The consideration of cultural heritage materials reminds me of the digital records we saw at Mudd Special Collections, where the technology was especially deployed to help preserve the materials from a protected group. As another example, Cordell's piece mentions the collection of metadata – this is so relevant to our conversation of what pieces of information should be collected, and how/why. This has ramifications on how the information is used and analyzed by researchers, and therefore impacts what questions are asked and conclusions are drawn. I wonder how the intentionality of datasheet driven datasets would impact this type of consideration for OCR usage and implementation. Would it change the process or prototyping? In describing how the technology is used and under what conditions, it seems as though a datasheet style description of methodology, decisions made, and motivations behind each one could illuminate the process and make clearer possible limitations or key considerations.

Anya Kalogerakos

11:26 AM Both readings from Terras and Cordell seemed to focus on the necessity of having some sort of record or protocol for the use of text recognition technologies. While Terras seemed to have a more optimistic view of text recognition technologies and their many applications (especially with handwritten text recognition), Cordell seemed to be more suspicious of the use of text recognition and its success. With this wariness, I believe that Cordell raises some important points to consider–namely, treating the digitized text as a new edition. In this, Cordell suggests that digitized text should be seen as a different iteration of the work, best said with the line "Digitization does not remove a historical artifact from material culture, but adds another stratum of computational materiality to its social text." Cordell's way of accounting for this new information (or loss of information) in the process of digitizing text seems to be a more formalized account or bibliography of the process of digitization. In this way, it reminds me of Terras' assertion that libraries will have to establish a recorded protocol for the process of digitizing. I think the strength in both these readings is in recognizing the value of text recognition, but also contextualizing this process in a broader history of the text itself.

Pippa LaMacchia

11:50 AM The most interesting aspect of these readings is the future-facing aspect. In the course thus far, much of our discussion has been theoretical or even engaging in a separation between the digital and the humanities (as if once data or an artifact becomes digitized, it becomes a separate entity or tool). In both of these articles, we really come face to face with the reality and future of digitization and at first I found this notion to be slightly depressing because as a "humanist" I am typically concerned with an artifact or text's tangible form. The conversion to a digital sphere feels like a lack or a loss but Cordell's argument made me consider it differently. These artifacts have gone through dozens of iterations, so why can the transition to

digital be any different? For example, the article explains that “a digitized historical text bears traces of its original, the process of its digitization, and a series of decisions over decades or centuries about documentation, collection, access, and preservation.” This strengthens the argument that if properly managed, digital spaces can increase accessibility to sources and therefore extend scholarly research in fields that may have otherwise been stunted by physical limitations. I am so curious about what traditional humanities will look like once these forms of digitizations become standardized and even more widespread.

Raphaela Gold

3:19 PM Before these readings, I had never heard of HTR or OCR, but I now realize that I must interact with such technologies almost constantly in my daily life, whether I’m researching for a paper or searching The Daily Princetonian digital archives. Reading about these technologies has definitely provided me with new questions that I’ll begin asking myself as I interact with online historical artifacts going forward.

One such question I was left with after these readings was whether things truly need to be digitized. Melissa Terras consistently argued that in modern times, libraries “need” to decide whether or not to use these technologies, and when they do so they “must” determine the proper ways to incorporate technologies like Transkribus into their preexisting methods of research. While Transkribus and similar HTR technologies are increasingly prominent and often very valuable, as they can increase the accessibility of works from Bentham, Foucault, Poe etc., I worry that the importance of handwriting and original documentation might get lost in the process. To me, handwriting is like a human fingerprint, and losing the quality of handwriting would detract from the humanity of original works. Uploading images of handwritten documents alongside their transcribed versions might be a way to preserve this humanity while also increasing access to these important works.

I was also struck by the “Conflict of Interest” statement Terras provided at the end of her piece, and I wonder how I might have approached the text differently had I been aware of this conflict of interest from the outset. While it is important to note that Terras would not financially benefit from this work, I think her position at READ COOP helps to explain the strength of her views around libraries “needing” to begin adopting this technology. In my mind, doing so might certainly be a strategic option, and it may very well be in the interests of that library or archive, but it is not a need.

On a similar note, my main takeaway from the Cordell reading was the concept that digitization can be seen merely as another option for scholars and researchers, rather than the only step forwards. I appreciated Cordell’s assessment in his conclusion that, “Digitization does not remove a historical artifact from material culture, but adds another stratum of computational materiality to its social text” (P. 214). This sentiment aligned closely with the concept of the text as a living document which Cordell raised earlier in the text. Some of our readings so far, and arguably the Terras reading, have positioned digitization as the next step in a text’s life as an organism, marking progress forward. I would like to consider the life of a text in a less linear fashion. If a text is fluid, I believe that it can live an almost horizontal life: it can be digitized while simultaneously living on in its previous iterations, bringing its past along with it into its present. HTR and OCR provide new and exciting options for scholars, but they are not the only options. For example, the mistake “Q I-JTB the Raven” rather than “quoth the Raven” is not an insignificant one. In that case, a better option might be to scan microfilm images of documents clearly into online archives, as Cordell provided examples of, in order to preserve the correct and original language of a text. (edited)

Helen Gao

6:41 PM I was really impressed by the high accuracy rate of Transkribus – even though the 98% accuracy mentioned in “The Role of the Library When Computers Can Read” is in the best-case scenario and not an average, it’s still strikes me as very high, especially considering that even I, as a human, struggled to read some of the handwriting on the Princeton postcards that I looked at online. The article also mentioned several criticisms of HCR, and I noticed that HCR seems to suffer from many of the same issues as ML models as a whole, such as a lack of transparency and environmental considerations when using high-performance computing to train models. Interestingly, the article didn’t really talk about bias in HCR models, which was surprising to me, since I think the topic of bias is usually brought up in discussions about transparency. The article also left me with some questions about the types of mistakes that HCR tends to make, but “Q I-Jtb the Raven” then answered them. I appreciated that “Q I-Jtb the Raven” explained some technical details of OCR processing for newspapers and walked through a specific example to explain how an OCR system turned “Quoth the Raven” to “Q i-jtb the Raven”. Additionally, this article made many connections between the topic of OCR to greater problems in digitization as a whole that we’ve discussed before, such as the adequacy of digital surrogates and the process of choosing what gets digitized.

Colin Brown

9:57 PM I found the readings this week to be full of insightful looks under the hood of OCR and many of the aspects of its use that I hadn’t previously thought about. The big theme that I seemed to see is the myriad of barriers of entry for using OCR technology in digitization projects. First, Terras points out how many of these softwares are privately developed; as such, I could see there being situations where the the development companies prefer to license their products to projects that digitize more widely-used texts so that their product gets wider exposure. Second, Terras also notes that machine learning algorithms, the basis of most OCR software, requires content to be trained on in the first place. This requires a) digital copies of many texts of the same kind, and in turn b) the resources to manually digitize those texts. I would imagine that these resources are allotted on the basis of demand for mass-producing certain digital copies since that is an expensive process. I’m sure, therefore, that smaller but valuable bodies of texts may be left out of the digitization process because the demand isn’t there to justify the libraries and software companies putting their resources towards them. Finally, Cordell discusses OCR errors extensively. It seems like older texts with different writing styles or characters are prone to more errors, which in turn require more manual fixing, which requires the demand to justify. I could see this being another barrier to entry for old and rare works to the digitization sphere.

Alison Fortenberry

3:57 PM This week’s readings reminded me of a project I worked on recently trying to identify every instance of a specific person being mentioned in a newspaper’s digital archive. When I went through the archive’s search tool, only about half of the total instances came up because, about half of the time, the OCR registered characters incorrectly. While the search tool certainly took a large portion of searching time out, because it was verifiably inaccurate, I still had to go through large portions of the digital archive searching for the person. What level of inaccuracy is justifiable/ worth looking over to use a largely successful tool? Like the errors described in the Cordell reading, the errors were understandable, but they still existed, and someone who stumbled upon the archive may not know the search engine has some inaccuracies. A person likely wouldn't have made the same mistakes, but would a person have been able to sort through every newspaper page and transcribe every instance of that name by the time I used the archive? While OCR systems still have inaccuracies that there is not enough human manpower to correct, how should we navigate digital archives? As Cordell discusses, digitized materials are not just surrogates of

original material, but have been transformed throughout the digitization process. Still, I think many people approaching digital archives (myself included) tend to think of them as a different platform to present the same information. I wonder if education about the digitization process and how that should inform interactions with digitized material could help smooth over some of the inaccuracies of OCR in digital archives. Giving users an insight into how the information was collected and sorted, and informing them of why they may need to take a deeper look at the archive could help close some of the gaps formed through inaccuracies as OCR is still being perfected.

James Sowerby

4:26 PM These were really interesting readings and made me reflect a lot about how I conduct research and interact with digital surrogates. In particular, I was reminded of a lecture I went to last semester that highlighted a professor's paleographic research on one particular German codex from the Middle Ages. The lecture presented the question of authorship—for this unsigned manuscript of the Gospels, there was a departure between the handwriting and style between Mark to John. A few members of the audience actually started participating in the discussion by offering their interpretation of the scripts and was able to come to a better consensus on these outstanding questions. I wonder, however, whether OCR would have a better time of comparing the two. Terras's article mentioned the potential use in analyzing damaged Greek manuscripts, which seems to strike at the heart of this paleographic/codicological question: is there a right way to go about interpreting these faulty manuscripts? This also hangs in conjunction with Terras's considerations—how do you convey through the digital surrogate the relative confidences of recognition or mention that some things are disputed? Of course, Cordell's article also dealt with this. By thinking of digitization projects as another "edition" of a text, it's a lot easier to see how the ethical questions (themselves a frequent topic in this class) come to the fore. For example, the case of Dicken's Bleak House corrections is very similar to the Ulysses controversy that I alluded to in a past post. Just as various publications of books can radically alter the meaning and interpretation of a text, so too can the digital version encode various biases and errors without full acknowledgment of the system's deficiencies

Andrew Huo

8:52 PM Reading Terras' article "The Role of the Library When Computers Can Read" was fascinating as I have never really payed attention to HTR or OCR before. My only experience has been the Google Translate text recognition that can translate menus or posters from a phone's camera or Apple's recent IOS software update that can recognize and copy texts from photos. It is interesting that in one piece of new technology that I very easily took for granted has a whole world of of questions, issues, and advancements that can help other areas of academic research and digitization. I was amazed by what HTR could do and its 98 percent accuracy and below five percent character error rate (bare in mind I know nothing about machine learning). But one of the things I thought about the most was — what is the current state of HTR now? The article written in 2022 talks about its potential benefits but also listed a bunch of issues such as still needing a lot of resources to engage with feedback loops, no good way of reporting errors and possible improvements, little transparency that goes against open science, and a lack of research. HTR's relationship with Library archives and digitization seemed to be at its infancy so I am curious about how the process has developed – are there new guidelines and regulations? Or more openness about its tools? Cordell's article was a bit too dense and specific for me but what I found interesting was his point that computers and their softwares are not treated as essential parts of bibliography and are mostly seen as digital surrogate for the original physical text.

Pia Bhatia

4:35 PM Both the Terras and the Correll articles are making me think of this course's structure more broadly – many of the readings have followed the format of introducing a new way for the humanities to interact with their inevitable and at times necessary digitization in some way, which is followed by an evaluation of the many concerns that are connected to the given advent. These, too, follow what is emerging to seem like a very streamlined line of thought: the same considerations are given about access, misuse, and the humanities content or artwork that is produced by or covers groups that are susceptible to misrepresentation. I wonder how these arguments will apply to the projects we look at later in the semester that are digital humanities projects from their inception. Another broader and more 'thematic' aspect of the texts I want to discuss more in class is the bibliographic accounts of work that are being digitized. Some of the bio-datas that were presented in class pointed out where researchers themselves had included what they considered to be shortcomings of the digitized versions. Cordell's piece echoed this, urging that "a digitized text must concern itself with the institutional, financial, social and governmental structures that lead one historical object to be digitized, while another is not." In this spirit, Terras included a conflict of interest statement regarding the paper itself. I would love to discuss the efficacy of doing this, because it relates to a debate about displaying problematic art (and whether that's OK to do with bibliographic context). New

Ethan Haque

10:08 PM Between the two articles, I liked the Terras one more than the Cordell article. At the end of the Terras article, I felt like I disagreed with some of the technical claims and claims about where the world is at with HTR, so I thought I might like the Cordell article more, but that one was sorta boring. I used to work for a lab doing machine learning research to improve HTR performance for primarily handwritten Judeo-Arabic documents and the state of HTR feels so uneven. Uneven in the sense that there has been a ton of research into the field, but so many fundamental problems haven't been solved. Rotation invariance for example. Most ML/AI based HTR transcription technology breaks down when the inputs are rotated by some arbitrary amount. How can a field be considered mature when such glaring fundamental problems still exist? At any rate, I like the combination of talking about the role of HTR in research and how libraries now have some responsibility to make ensure that digital content is created with consideration for how it's going to be used and who it is going to benefit. As for Cordell, I felt like they wax on about some philosophical notion of HTR and text and digital surrogates and ends up saying very little with a whole lot of words. One thing I really did like was the part where do they a bit of digital forensics to figure out how many times a physical object has been transformed into another medium via some metadata contained in their digital objects. (edited)

Talia Goldman

10:45 PM The two readings for this week got me thinking about the digitization of text as objects themselves, a theme running through several of our course readings this far. Both Terras and Cordell frame digitized texts as distinct scholarly sources and new objects capable of evolving through time and subject to careful bibliographic categorization. Terras writes that "If we are to see HTR-generated datasets used as new source material to underpin novel research, we must be able to explore their provenance so that the resulting datasets can be trusted as a scholarly source" (Terras 143). In evoking the concept of provenance for digital humanities, Terras concurs that to be used as valid sources in academia, datasets must be seen as their own, original sources with unique stories of their creation and ownership. In framing HTR-generated datasets as new works, Terras' view appears consistent with Cordell's assertion that "The first challenge for

a serious bibliography of digitized material, then, is one of apprehending: of seeing the digital object as such, as an artifact with a distinct materiality and sociology" (Cordell 192). In thinking about how digitized text came into being, who was behind certain projects, and the errors of technology and from with the original sources, HTR and OCR appear more concrete. Thus far, thinking about data, metadata, and datasets as new objects as made some intuitive sense, but I think framing this idea in text recognition technology provides more solid ways to consider some nuances of how these recognition tools, with varying degrees of human intervention, affect the final product of these new objects for scholarly use and (at some point in the future) primary sources for historical archives.

Clay Glovier

10:47 PM I enjoyed reading both the Terras and Correll articles. It would be very powerful if the capabilities of digital humanities technologies continue to progress and researchers gain the ability to search entire library collections for certain terms. However, it would be very difficult to achieve this capability. Firstly, it would be difficult to have all the texts from a library digitized and uploaded to an online collection given the sheer number of texts. Moreover, even if each work was able to be uploaded, searching older texts would be very difficult given the variations in phrases (and translations). Words could be spelled difficulty, and the differences in structure could be tricky for search engines to decipher. Thus, engines would have to be flexible in order to account for variations in the terms they are searching for. If they only provide a couple values for each search they will not work as effectively. It is also important that institutions that develop the capability to incorporate more advanced searching and digitization technologies into their library collections allow these features to be accessed from other groups not part of their own programs. If only the wealthiest and most prestigious educational institutions have access to such powerful search functions, it would only exacerbate the inequality present in academia, and give researchers at the best universities an even more pronounced advantage over their peers. While digitization projects are certainly beneficial, their implementation must be pursued in such a way that they do not strengthen pre-existing inequalities. (edited)

James Sowerby

11:08 PM The DSC article on TEI was very interesting because it made me think more about my own annotation practices more. It's not something I usually do, since I almost exclusively work with print materials and write in margins with pens, but reading Beshero-Bondar's article once again referenced the question of accessibility: how do we encode our own thoughts such that others can read them later and reproduce our train of thought or research? I, for one, would expect very few people to be able to follow my train of thought in annotations. I tend to write in cursive, which is itself a form of encryption these days, and often make allusions to hyper specific references from my own life or from other classes. Sometimes I can't even remember the point I was trying to make when I go back. But I remain convinced, sort of along the DSC article's methodology, that it is incredible important to mark your thinking at the time to track changes and understand your research narrative. I have never used XML or any of these tools, but I clearly see how complicated it is to standardize methods. That also makes sense in the context of the authors' comment on funding—were they to be funded by grants, they would have to use XML to encode things as a matter of transmission and proof of work. As a side note, I thought it was really cool to see the emails reproduced in the article that show us the evolution of Quinn's work as they are trying to figure it out. Very cool for a peek into the process. Lastly, I thought the Shakespeare and Company article was super interesting because I'm always curious about how bookstores and libraries work/have changed over time. It also is relevant to a project I did last year with the Nassau Literary Review that wanted to highlight underrepresented authors in

their archives. It was a hard thing to do, since to ascertain demographic markers we really could only go off of names or yearbook photos, which is, at best, an unreliable method, so I appreciate the difficulty of their project.

Emanuelle Sippy

1:17 AM Cordell's piece compellingly positions digital texts as "descendants" of textual "antecedents," about which we need information to accurately analyze the text over time (Cordell 201). He suggests that: "One of the most compelling reasons to take bibliography seriously for digitized historical texts is that doing so forefronts their createdness: the chain of human labor that led to their present existence" (Cordell 214). In this way, current digitization practices lack not only individual pieces of important information but also obscure the lineage of the text and the work that has been done on it over time. Moreover, Cordell writes that the practices he envisions and advocates for recognize the "iterative realities of digitized text" (Cordell 216). So too Terras asserts that this "iterative" quality must be accounted for and the work at each stage must be properly cited and credited (Terras 143). Both pieces made me recall what the archivist who spoke to us at Mudd said about how digitization produces new objects that are also in need of preservation. When we do not hold the object in our hands or see the physical space it takes up in storage, it is easy to overlook its materiality. Cordell warns against this, writing: "Digitization does not remove a historical artifact from material culture, but adds another stratum of computational materiality to its social text" (Cordell 214). I find this assertion of digital artifacts' materiality and relationality to be a very helpful framework and one that might lead to more buy-in for reaching the goals Terras and Cordell articulate for moving towards the "transparency" and contextualization of digital texts (Terras 138, 143).

Yaashree Himatsingka

2:50 AM I appreciated Terras' call for a critical framework in adopting HTR technologies, especially when considering its implications for scholarship. She explains how HTR, which allows for the digitization of handwritten texts, unlocks previously inaccessible archives, manuscripts, and documents. However, while HTR significantly enhances research capabilities by making vast amounts of textual data searchable and analyzable, it needs to be implemented thoughtfully. She underscored the importance of libraries in mediating the relationship between digital tools and researchers. In addition to this, I think that interdisciplinary collaboration is necessary to effectively integrate HTR into scholarship. This might involve partnerships between technologists, librarians, historians, linguists, and other domain experts to ensure that the contexts and nuances of the original sources are preserved and understood. Relatedly, Cordell considered the challenges of working with 'dirty' OCR. It was interesting to think about OCR as not merely a technical hurdle but a kind of mediator that affects the interpretation of digitized texts. Like Terras, Cordell seems to suggest that information literacy and continuing deliberation on ethical standards of transcription technologies are important next steps. I think critical engagement and bibliographic transparency will probably be most important in enabling the responsible adoption of digital tools like HTR and OCR. (edited)

Layla Williams

8:59 AM The HTR systems are interesting developments within the digital humanities, especially because I think it has been one of the systems that I personally have interacted with the most before considering it a part of the digital humanities. I have mostly seen the side of misread handwritten texts that causes the transcription itself to be unreliable. For example, in a workshop for my junior seminar we were looking through old African American archival material, such as handwritten entries of market stock. The most common misinterpretation was the difference between the "s" and the "f". In addition, I found the teaching

of the HTR systems to recognize Japanese characters being another system in itself interesting as well. Here, I would imagine the HTR seeing the composition of the character as a full image and seeing the strokes in relation to each other. I do not know if it was because gaming systems were mentioned shortly after, but I wonder if we can use this technology to possibly recognize other forms of compositions as well, such as artworks? Could a HTR recognize the strokes from a specific artist? Could it recognize the stills from a particular director? What information would we need to try to achieve this?