Informational Session

So, at the end of this week's presentation, we're going to be looking at preparing documents for Hypothesis. So, I'm Andrew Wilson. I'm the new director of research computing and digital scholarship over in ITS. And I'll be running you through how to use obviously our, one software package, in particular, Adobe on how to change your PDFs or images into human-readable documents, which will allow them to be used in Hypothesis. So, before we get started, just some housekeeping. If you don't mind just muting your mics, but I think you've already done, all done that already, which is fantastic. And if you've got any questions, just pop them in the chat and I will make sure that at the end of the presentation that I'll go with, we can go through the questions and ask them personally, if you want, as well at the very end.

So just a quick overview, I'm going to talk a little bit about what OCR is, optical character recognition, how it works. I'm going to be brief on how it works and just talk you around the points of how it could, it generates these types of documents, different applications you've probably seen already, what, how to prepare the documents for Hypothesis. Then I'll give you a quick demo on using Adobe Acrobat DS. I'll show you how to download it, how it's going to go on your computer, and then also how to use that for different types of documents. And then at the very end, I will talk about best practices and what works, what doesn't, and how to get around some of the weirdness with old languages and things like that.

So, OCR is called, is optical character recognition. And it's basically a method of digitally converting images, either typed, handwritten, printed into machine-encoded text. So that allows you to do is when you get a document, you can scan it in, and then you can actually select the different sections of the text once you've run it through this process. Most PDFs you get nowadays are already, have already had this process put in or done to them so you can get access to the documents and access to the texts and said documents. But if you scan them yourself or you have your own documents that you want to scan, then you'll have to run them through some sorts of OCR, this process, using this process. It also makes them searchable as well. So, you can search the PDF for things you're looking for.

So, the technical aspect of it, how it works. So, it works, there's four different sections to OCR. There's a pre-processing section where the computer will readjust the screen or readjust the document or, and then it will look at the document itself and see how the document looks. It will do segmentation, where it will look, look for the spacing and look for the different font types, the characters, the characters. It will then try and recognize the documents, which you can see here. I think with the annotation tool, which is this section here, where it'll recognize the letters and the words, and then it will go through and predict the actual word itself. And then the processing itself will just spit out the document again, with the correct text in place. Clean it up.

So, you've probably all seen OCR, in fact, a lot. It's used pretty much every day, nowadays. Anything from data entry forms, when you get the chance to fill out a form

and you can send it in and the computer will read the form itself. Whoops, sorry, I'll just admit this person. And automatic number plate recognition is another technology that, that uses optical character recognition edition, authentic information extraction. So, you can basically import the document with text and a figure, or a table and the computer can figure out that it's a table. It will put the text in the correct locations in the table. And it will understand that I taught you all a full-sized table. You can also use it for different versions of printed materials or printed documents. In theory, it also works for handwritten documents, but as everybody's handwriting is very different, this technology is, works some, some of the time. It's getting much better. And it has got a million times better in the last couple of years, but obviously, your handwriting, especially cursive text, is also very difficult for them to sort of recognize or calculate what the, what that is. It's also a really, really good tool for assisted technology. If you have a picture or an image of text, it's a very good idea to OCR it. The OCR will then make it available to screen readers. So, anybody who's partially sighted or struggles with large amounts of text, you can put that into a screen-reader, and it will read out the document for them. So, it's a really good tool for assist technology. Especially if you have a lot of documents that are just pages and pages of documents, without any ability to select those scans, or search in that document. OCR is a very good tool for that type of thing.

So, so the pre-, to create a document itself, I wait on the principal, uh, most of the documents that will require OCR are going to be something that you scanned in either with a scanner or you taking a picture with your cell phone of a document. Most, or the PDFs that you get nowadays... I can show you, just about a minute, and show you. Most PDFs you see will already have selectable text in them. We download them from journals. They will already be pre-prepared for tax selection. They'll be prepared for searches, so you can search this document for any, anything you want. And then it will find it in the document, but not all texts like that. Here's a document from a journal I read called Antiquity, which is a psychological journal. And you get documents like this or journals like this, where you can't select the text at all. It's not searchable, so I can't search or anything in here.

It just won't let me search for that document. So, and the computer just recognizes this as a flat image. It doesn't recognize it as a document with text inside of it. Where this one, which is from a computer science journal, it recognizes as having documents inside and texts inside. So, you can select it all. One of the pe-, one of the, cause Hypothesis is an annotation tool, it needs to be able to do this type of selection before you can annotate the sections of texts that you're interested in, which is why OCR becomes vitally important for using it with, ah, Hypothesis. But it's also a good practice to have, your text OCR, because then it means that you can obviously make it available to the students. The students can search for things inside that document. But also, as I said before, it'll also allow accessibility and screen viewers and things for students to read. The presentation.

So, you give OSR on a document and you go using a flatbed scanner. Then for the best results, you're looking for 300 dots per inch. So, when you start, when you come up with a scan document it will ask you how, what resolution do you want to scan it in? If you use a photocopier, which I know a lot of photocopies nowadays also have scanning facilities, you can set it up so it will scan in 300 dots per inch. You also want to make sure that the brightness is not too high as well, and not too low. So, you're not getting weird edges, especially if you used the photocopier, you get that, strange black edges around, if you've got too much light into the print bed or scanning bed, sorry. It

also goes the same for the scanners, as well, just make sure you don't have it, too much light in there.

If you're taking a photograph with a cell phone and you know, you're using a flash, just make sure the flash doesn't wash out the text before you move it into the software. Also, the how, the way the software, the way the OCR works, it works on straight lines. So, if you do have a text document that is a bit wonky, then a little bit crooked, then the software won't work as well for OCR because it basically reads as a human would read in straight lines. So, we have this slight, it's slightly skewed. It will struggle to recognize, and I'll show you an example in a minute. Also, older documents and discolored documents, you might need to scan them in RGB, which is in full color rather than just black and white. As people, people are shooting, when you, when you scan documents, it's just a black and white document.

But if you have old yellowed, I'll give you an example of this, like really old, yellowed documents. Like they switched out the Canterbury Tales with the yellow background. You'll have to scan it in full color. So, make sure that the computer can actually read the text inside of the... Some text considerations as well. So, language really matters in this. The OCR software has to be able to support the language that you're interested in capturing. I, I'll, I'll run through in a minute into the, into software and I'll run through with most of the languages. I've got a couple of examples of Chinese, Japanese, and English and different styles of English as well. So old English and modern English as well. Just also be careful as well of any text published before type funds, like maybe 1850. It might not be compatible with the OCR. Most of the software nowadays does a very good job at recognizing the different texts, but it might struggle with some of the old English.

And I'll show you another example of that in a minute of a really old English text, where it's got different font sets and the software doesn't recognize that. Documents with low contrast as well, also really affects the software, but also the inconsistent use of font types. So, then another example of this is when you've got the really capitalized big letter at the very beginning of the paragraph or the first section of the book, and that will rethrow the OCR software off for the rest of that page, because the way it works with the straight lines, it doesn't recognize that the I am up in the corner and that there's a run-on text. I'll show you another example of that once I get to the demonstration side of this. Then some examples of just the English languages texts. So, something like this on the far side, where you've got really old English.

The software will really struggle to use that type of text. But you can get around that by going through and editing some of the words yourself, if the software will do an okay job at recognizing some of it, but then it will struggle with the rest of it. So, the software we're going to use, or on this demo today is Adobe Acrobat DS. You can download Adobe from creative cloud and this creativecloud.adobe.com. You can then log in with your Pomona account. You just log in, we just undid login account, and that will allow you to download Adobe. I can bring it through the process, if that makes sense, that helps. I can just type it into here.

I'll share my screen now. So, once you logged into Adobe, you've come up with this screen. We just go to an Acrobat DC, here. And if you click on this one and click, get up, that'll download the latest version of Adobe, to your computer, it's Mac and PC. So, it will work for both sides of the system. And it, once you, once you log in with your Pomona account, it will do all the logins for you automatically. You're already logged in to my.Pomona. It would log you in automatically to the creative cloud as well once you clicked on the link. The software.

So, once you have Adobe open door, it will look something similar to this, which is just all the different text files. All the different PDF files you have. You can create PDFs, combine them, organize them, share them and export them. You can also edit them in Adobe Acrobat. So, you can go through and edit the PDFs. If you have a PDF that you've created yourself, you can go through and edit them. The thing you're looking for is you click on this, see all tools. You wanna click on that and I'll bring up an extended section in Adobe. And the section you're looking for now is this scan and OCR.

So, once you click on scan, click on open, it'll ask you to open up a file. I've already got the Canterbury Tales opened up. Then at the top here, you can see recognize text. Once you click on that, and then you click on in this file, you can also do multiple files as well. Maybe I'll do a bunch of PDFs all at once. You don't have to select the language you're looking forward to. This is where the language is really important. And it matters of what language the text is. All the languages supported right now are in this list. So, the big ones are with non-Latin characters, are the Chinese, Japanese, and I think, the Russian as well. And there is a few other sort of other languages, unfortunately, Arabic is not supported with Adobe Acrobat. I was going to try it with an Arabic script, but it's not, unfortunately, supported.

So, once you select the language, I'm going to select UK English, because I'm assuming that Canterbury Tales is in UK English and not UK American. And then you click recognize. And what the software will do now is it'll just go through the text, and it will recognize everything and make it into a readable PDF. So now you can actually select the different sections of the text. So, one problem with this document, you can see is it has the really big w at the very start of the text. So, what happens there is it will break the way the OCR works, so that the OCR has not worked on this text document because of this w. It assumes that the text runs on. So, it doesn't actually work very well for this type of text.

And you can go through and edit the text. So, you can click on the recognized text again, and click correct recognized text. Then that will show you all of the words that the software has recognized in that. So that's what I mean by the text consideration. You need to consider what texts you have available. Canterbury Tales is a bad example because it's an old English example of how it works. So, what I'll do is I'll jump to my Stonehenge paper, and this is a regular, this is a good example. It does have the start of the J, but it doesn't affect it as much as the Canterbury Tales one, it's likely unrecognized. Now it will run through, again. This is a multi-page document. So, it will run through every page of that document, and it will do the same OCR recognition on every page, through the process. It will recognize captions for plates and things as well. So, you can recognize, look at captions and see figures and things. Once it's finished that process, which is a really good way of making sure that the students have access to the captions. And I'll just read this to run a little bit and then I'll stop it.

So just, yeah, just be aware that it does, even on my quite beefy machine, it still takes a while to run the OCR through the process, and it will run it on every page to recognize every single page. It also will detect, the pho-, the photos as well and the headers is in the document. I probably should have done a couple of pages. It's 16 pages, so I'll leave it to finish. So once that's finished now, you'll be able to select the text on the whole document. So, I can go through now and I can select that text. So, this one is not as bad as the Canterbury Tales where it will, if it does, it slightly understands a bit better, that the tell, with works. Once you get below that it will be out, it's fine. It'll

recognize that individual sections and you can go through, and it'll select every single page now.

So, this is now a converted OCR text file, all PDF. You can go through if you want, even on this document and correct, if there are certain words that you feel are not quite right. The software will also indicate if a word doesn't look quite right. So here, it's saying that it doesn't understand what this word is here. So, it's saying it's wrong, it's spelled wrong in the text up here. So, you can go through, and I can change this and say, this is per rather than the T I was pulling in and I click accept. So now when the students cut and paste this, or select that section of the text, it will be the correct word. So sometimes the software is only 97 to 98%, correct. If there's weird documents or weird things in the document itself, then it will recognize them. So here, uh, there's a dot, a slight misprint above the Albany U, so that's, the software's not recognizing that correctly. And I click accept. So, I can, you can go through this and just check through and make sure that these words are correct. Most documents you will have access to will be much better than this. This is a pretty old 1960s paper, I think.

So that's pretty much, that, that's the, the basics of the OCR software itself and just OCR and standard English language. But if you want to do, if you want to tell OCR something else, I have a few examples as well of all the languages. I'll just open up. I'll start with Chinese or Japanese first. So, here's a Japanese text, I my, I don't speak any Japanese. I have no understanding of this text at all. So, I apologize in advance of anything that's inside of it. This is just a damaged, a sample, like on pull, that I got from an OCR website. And then it's, again, it's the same process as you would do with the English language text. You, you go to scan and OCR. You then recognize text in this file. The only difference here is you want to make sure you change the language because the software is looking right now for English US.

You just want to change it to Japanese. It's on here. And then again, just click recognize text. And then it'll run through the text again. I don't do the same thing. So now I can select each one of these characters in the same process. And again, I think you can go through and correct words, or no, you can't do this one, but you can run through the, you can now use this as an OCR. And it will now search. Because I don't speak the language, I don't know how it searches or how the documents recognizes the texts. So, it might be searchable by the characters, it might not. I would need someone to check that through before I could confirm that. The, the final process is just making sure you save that copy of the PDF. It doesn't save automatically. So, you have to go through and just save as, and then just rename it as however you want. So, I can save this as an OCR test. I'll just save it to my desktop for now. So that's now saved as an OCR text document. You can now import that into Hypothesis, and that will work with the annotation tools within, within, inside of my Hypothesis. I'll just do...

Sorry again, can I ask you quickly? What if there are more than one language and one document? Is it going to work only on the, on the language that you chose?

It will do, yeah, it will work on the language you chose and initials. You, you, you can get around that by running the software twice. I can just go back. Let me jump out of this again. So, what you could do is you could now save this as the Japanese. So, I run it through once with the recognition text, just Japanese, save it, load it back into Adobe, and then run it through again as English US. So, what I'll do is it will first recognize the Japanese text, and then the second time you run it through it will recognize the English text. So, you'll probably have to run it through twice in that case because it's looking for the specific language when you're running the software in the

first stage. I think I don't need a lot. And then this one, but yeah, so for that type of thing, you would have to just run it through twice and it would then, hit would then work.

Just make sure you save it in between both versions. So, make sure once you've done Japanese version you save it and then just import it back in. I think I have another example of, other, other languages. So, I have a Chinese document as well to import. And again, the same process. You just click on the scan and OCR, recognize text in this file. And just changed the language to the language you're looking for. As far as I'm aware, this is simplified Chinese, and then click recognize, and then it'll go through, and it will recognize all the text and it will recognize the individual characters as well.

Well, I'll just jump back in. So again, just some considerations for how you use the software to make sure you understand what the output is. Have you looking for this type of OCR I've just done, is very quick. It's not really fine. So, it's going to be, it's fine for using with the students to put up on Hypothesis. If you're looking for archiving, this is probably not the best solution for archive. You want to go through, and you want to tweak the sentences and want to make sure that it's getting everything correct. You then want to go through and correct the mistakes if you're going to put it into an archive. Also, consider the functionality of the software as well. That, there's a lot of OCR packages out there. Adobe the one that the library uses and what we're recommending right now. It's quick and easy to use. You can just put the, press the OCR button and then press go.

There's a more advanced version of these types of software, but then that involves a lot of code or messing with the settings to make sure that it gives you the perfect OCR set, setup. But for the simplest easiest uses, Adobe is definitely the best for that type of thing. Also, to consider, things to consider are file formats. Adobe will take images. So, we do take a picture with your cell phone. You can import that directly into Adobe, and it will recognize the text within the image itself. A lot of the pages I imported just then, like the Canterbury Tales, one is from scan documents. So they will be, it will accept TIFF and JPEG and all the other image formats that you, you're aware, you get out of your cell phone or your scanner. And other things we consider, OCR is not a hundred percent perfect.

It's not always going to get every word in every document. If there's a slight error in the print, or if there's, the page is scuffed when you're scanning it and, or any deformation of the page, the software might recognize, uh, I, as a T or like, there's a lot of weird requirements like that with, so just make sure that, let's make sure you understand that it might not do the whole document a hundred percent, perfectly. Some words might be incorrect. Again, consider the texts and consider how the texts looks. And as I showed in the example of the Canterbury Tales, it's, some texts don't always work perfectly, especially if you're using really old text, for any historians. It will require some tweaking and, to get the perfect scan if you are using that type of textbook. For most modern documents, it should work pretty flawlessly to at least 97, 98%, correct. Most cases a hundred percent, but there is that odd case sometimes where it won't recognize a word or a sentence. Also, if you are using it for something, consider correcting the errors in the text, if you're gonna use that document for the next couple of years. Yeah, just go through and change the words, correct the words, make, cause I know you're going to make them available to the students. Then it will save you a lot of time in the future when the students are misspelling a word because the OCR incorrectly spelled it.

And so that, that's pretty much OCR in a nutshell. Does anybody have any questions? Uh, if I may, um, you were just talking about corrections. So, would I be able to erase the wrong word and write it in?

For the English language you can definitely do that. So, you can go through, let me show you the, this one again. If you go through, let me show you the, this one again. If you go through, you click on the recognized text here. If you click on this correct recognized text, it will show you red, these little red sections. This is all of the parts of the document where the software doesn't understand what they are. So, there's obviously the three, three, and a half feet. It doesn't understand what this three and a half is. If I scroll down a little bit and find some words. So here, so kind, it doesn't recognize that word, even though it's pretty clear. So, at the top here, you can see, it says, this is what the image looks like with the software, what the software has picked out what the images are. It's recognized as K-I-N then less than L. So, it's not recognizing the D very well. If we zoom into the document itself, you can see it as a slight break in the D. All you do then is just go in here and type in K-I-N-D and then click accept. I mean, I'll go onto the next one then. So that, that's corrected that kind now. Okay.

Can you type in a different word even? Oh yeah. Yeah. I can type anything I want. I can type in...test. So now I can, I can say the at here in this location is the word test. So, it doesn't, it doesn't matter. So sometimes it'll be the correct word but sometimes it can be really, really wrong. And then if we go back to the Canterbury Tales one, you can see there's a lot of errors in this document. It doesn't understand what most of these words are. Maybe looking at the actual recognition at the top it's colon, colon, R, comma. So, it's just not recognizing the words, but you can get, you can go in and type in anything you want in that location. And just click accept. And then it will recognize that from that point on. And once you say that document, it will recognize this word assault rather than, or whatever you typed in.

But you cannot add a comment on the right side of the line or anything that, that would not stay? You can add comments, but not in this, not in this section of the software. Someday. The comment. So, you want to, you want to click on the different sections. You want to click on the comment here, click on that. Now I can start out and then comment into the document. Again.

Sorry. But that doesn't show it on the screen unless you point the cursor there. You can put a icon, so I could put an icon here and add the sticky, sticky notes but it won't, won't get it. It'll just show the sticky notes. It won't show that, that word has been changed to something else.

Yeah, it won't mean intermingle, it will remain separately. Yeah, it will remain separately. Yep. But it'll still be in the documents, as you can see here, it still says... Cause it's just, at this point it's just a standard PDF. So, you can do everything you can do with the PDFs. In theory, you can go in and add it to documents as well, but it probably won't work very well with this old document. This is what I can do about it. So, I could go in and change this. And change the documents itself. Any, any more questions?

This looks, uh, was JPEG as well? It'll work with JPEG as well. Yeah. You can just import, if you, I can just jump back in and show you. So, the Canterbury Tales is a JPEG. I can pull in another one, which is, or the other one. I'll pull out a really old, so this is it. This is also just another JPEG image, and it will just look like a normal document. And it goes to, to scan, recognize text. I don't think in this file. And it will recognize it from a JPEG.

Okay, great. Well, thanks so much, Andrew. No problem. But... This is Elizabeth. Could you show us how to straighten a crooked document? So, straighten is slightly different. You will have to edit the PDF. So, it's not part of the software. It's not the part of the OCR stuff, but you can use things like the, the crop, which is probably the better way of doing it. Unless it's really crooked, then you want to, want to edit the image first, before you bring it into the, into OCR software. With a crop, you can basically crop it out into a section. So, it's just the, the best possible solution. Okay. And then close it. You can use things like enhance and enhance the documents. And it sometimes will shift it a little bit to adjust text. But if there's any, there's a large amount of shifts, then you're probably better off doing it beforehand. One of the ways as well is there's, I think it's called Adobe, Adobe, no, it's Microsoft lens, and there's Adobe scan as well on your cell phone.

I'm just looking at my phone, sorry. Where if you're taking pictures of a document, you can use that and that will automatically shift it for you. That's what I've normally used. If you're taking pages of, we actually often have taken pictures of PowerPoint slides. So, I'll give a presentation, you can take pictures of the PowerPoint slide and it will automatically shift it, adjust the scale, and stuff. I would recommend probably doing it before you bring it in, making sure you, you adjust it before you bring it in. The OCR will adjust it a bit, but not if it's crazy. I definitely got some test stuff. I might have an example on here of a document. That's not... <inaudible>.

I'll try and recognize this and see what it does. Most of the time it will just, yeah, it didn't recognize that very well. So yeah, you're better off just doing it beforehand and then bringing it in once, once has been adjusted. Okay. Well, thank you very much. It seems pretty straightforward.

Yeah. It's pretty good, but especially with Adobe, where it's just one click, a couple of clicks and you're away, so. All right, well, thank you very much. Thank you. Any other questions? Then I will stop sharing my screen and stop the record.