John A. Walsh: Keep us moving along here, I'm going to shift gears away from film and back towards some of the other materials we were looking at yesterday. I was teaching in the morning so I missed the morning stuff but I was around for the afternoon sessions yesterday and really excited by what I saw. I'm a literary scholar working in the School of Informatics and Computing and digital humanities and digital libraries. I'm involved in a number of digital archive and digital edition projects from Petrarch the 14th Century Italian poet to Isaac Newton's Alchemical manuscripts to 19th Century British poetry, which is what I studied in graduate school and comic books and graphic novels, which I think share some formal similarities with film, and what ties these diverse subjects together for me is an interest in using digital technologies to represent complex document types with particular attention paid to graphic and visual aspects of these documents and informational issues of indexicality and addressability which relate to our ability to search, find, retrieve, and use and annotate these objects.

So most of my experience is with text and image based projects and I'll focus today on the paratextual documents around the film. We saw and heard about many different document types yesterday, scripts, filming scripts, letters, promotional materials, and so on. The document types in this slide are largely drawn from Barbara's [Tepa Lupack] handouts from her talk yesterday. In the display case downstairs at the coffee break, at the cinema, we saw movie posters, letters and so on. What I heard in the afternoon sessions, I heard very little about film itself and much more about the documentary record. These documents are clearly very important sources for your research, and I want to talk about some standard ways for digitizing those documents and enriching them with scholarly encoding and annotation.

I also heard a lot of interest in and sensitivity to the material artifact, maybe best exemplified by this interest in the edge to edge scanning of the film, and I'm interested in how you can use digital technologies best to represent not just the abstract content of the material object, the sequences of characters and words in a novel or poem or play, or the sequences of images in a moving picture, but also the physical material characteristics of the artifact, the binding, the size of the page, the watermarks on the paper, the type of film, size, manufacturer, and so on. As one example document, let's look at the example of one category of document.

Let's look at the example of letters. Yesterday at the coffee break I was looking at this letter in the display case downstairs. I took a picture of it with my iPhone, which we see here, and we can imagine a virtual collection of a few hundred or a few thousand such letters written by and to figures related to this period of African-American film. All the letters digitized, scanned, transcribed, and coded with detailed metadata about people, places, organizations, like film studios and theaters, and events, like screenings, performances, openings. One could have additional data files or databases with biographical details about people, metadata about films, movie houses, studios, and so on. In addition to transcribing the letters and making them searchable, you can add with some effort additional layers of encoding, identifying all the key people, places, works, organizations, and so on and linking references in the letters to these external data bases

with more information about these entities. When we've done that, and again, there is some effort involved in doing that detailed metadata and coding work, then we can do other interesting things. With a tool like Simile or Neatline, one could create temporal spatial visualizations of the places and events referenced in this growing collection of letters, so here is Simile's timeline from my [Algernon Charles] Swinburne project on the Victorian poet with the interactive timeline, the first band is his life and biographical and literary events. The second band is other non-Swinburne literary events. There's a third band below of other historical and cultural events going on during his lifetime, and there are links back to the collection, so his volume of poems and ballads published in 1866, you click on that and you move into the archive and you can read that volume of poetry.

Here is Simile's exhibit tool, which combines a timeline with Google Maps interface, so you could have this collection of letters and see, plot it out, all the events that are referenced in the letters, maybe the locations of theaters and studios that are referenced in the letters and so on, and that's in addition to browsing by the sender or the receiver of the letter or the date of the letter; you can browse by these other facets.

This is another similar tool; Neatline comes from the Scholar's Lab at the University of Virginia, also combining map and timeline, a different interface to do something similar. So here is our letter again, and I mentioned this encoding in these layers of metadata. What is that, and what does it look like? HTML is a hypertext markup language, the language of web pages and I assume many of you are familiar with that. If you haven't seen the HTML code and the pointing brackets and written those types of pages yourself, you're using them when you browse the web and so on. HTML has tags to mark paragraphs and hyperlinks and images embedded in the page, headings, tables, and so on. The Text Encoding Initiative or TEI is a standard analogous to HTML but it is the standard encoding language in the digital humanities and digital library communities, for digitization and representation of texts and text bearing objects.

From the introduction to the TEI guidelines that document all these codes and how to use them, they say these guidelines are addressed to anyone who works with any kind of textural resource in digital form. They make recommendations about suitable ways of representing those features of textural resources which need to be identified explicitly in order to facilitate processing by computer programs. In particular, they specify a set of markers or tags which may be inserted in the electronic representation of the text in order to mark the text structure and other features of interest, so in order for these tools like Neatline and Simile to process these documents and do interesting things with them, we need this tagging to formalize and normalize things like dates and places. So, how does Neatline or Simile know when it encounters Paris whether it is Paris, Texas or Paris, France? Or when it sees a date, if it is in the format of day/month/year or month/day/year, or it's written out in longhand in different formats, these codes can normalize that sort of thing.

Here is an example of that letter we've been looking at, to [Richard E.] Norman,

encoded in TEI, and we see organizations like the Norman Film Manufacturing Company tagged with a metadata code that says it is a name of an organization. Here the Franklin Theater is tagged, org name, type equals theater. T equals Franklin which would allow us to go into that external data base, look up Franklin and get all the information about it; the street address, the years it was in operation, the owners, and so on. Here is Cooper-Hewitt, which I learned last night is someone who invented some sort of light bulb that is used in filming, so that can be interesting to film historians, references to these, that kind of equipment.

Allyson Nadia Field: So do you transcribe this letter yourself, or did you scan it?

John A. Walsh: I took my iPhone picture. I did OCR on it. OCR is Optical Character Recognition, so it will take a picture and it will turn it into transcribed text. I had to correct some things, and then I added these codes. It took…I worked on it for a couple hours, not just this part but then to make it look pretty in a web browser, so this is what it looks like in a web browser. Here is that letter and here is the letterhead and the transcription of the document, and all those codes are behind here, so there is a title code around this, which again would have a key where I could look up that information in the database. I've got links to the page images in those codes, so I can click on this, open up the full thing. I can expand it and then on a larger monitor I can put these side by side and compare. In my transcription I didn't do all this. The OCR didn't do a good job on that, so that would have to be done by hand, and then I can continue scrolling down to the second page.

Terri Francis: I'm so distracted by [Oscar] Micheaux's analysis that he's giving there, but I'm looking at too many things at one time.

John A. Walsh: By embedding, if I go back to the code, by embedding these codes in the document I filled the document with addresses, so the document is a neighborhood with thousands of houses or a building with thousands of rooms. You attach some intellectual content to that document, your analysis, your annotation about specific parts, you don't want to drop off that message at the entrance to the neighborhood or the building, you want it to be delivered directly to the specific house or room that you're interested in in this document, with these tags you've populated this neighborhood with lots of more finely grained addresses which you can then use tools like TILE, the Text Image Linking Environment or some system that is implementing open annotation framework or data model to attach annotations, not just to the gross level of the document but into this reference to Cooper-Hewitt or this reference to the Royal Theater in Philadelphia and that disambiguation, we have two Royal Theaters here, one in Philadelphia, one in Baltimore, and the tags can help us disambiguate between those two Royal Theaters.

So you do, I spent an hour or so on this last night, but we're imagining hundreds or thousands of these letters. You do all that work yourself and with students, maybe through a grant funded project, and then you want to do interesting things with them. One thing you could do is extract certain information about people and places and

events, like we've looked at and built those maps and time lines, but the most basic thing you probably want to do is publish these documents on the web for yourself and others to see and read. TEI is not HTML so if you just open this page in a web browser without doing other sorts of interventions, you're just going to see these tags. You won't see it nicely formatted like we did in this slide. The next part of my talk, we'll look at a tool that helps you publish these sorts of documents. Before I do that, I think I jumped ahead of myself a little bit, so again, imagining the scenario of the archive of letters related to this field of early black film, here is an analogous project, Vincent Van Gogh's letters and I'll jump to one. Here's his buddy, Paul Gauguin, and there is a specific one. This is a collection of TEI encoded letters and you have these tabs at the top of the two panes. You can view the original text in French, the original text with the same line endings as the original facsimile document, the facsimile, which we're already seeing on the other side, an English translation of the document, and you can do all these options on either tab and add multiple tabs if you have a wider monitor, but then you can see, there's metadata about all the artworks that are mentioned in this letter. So you can imagine the same thing in this collection of film correspondence where you have thumbnails of stills or links to embedded motion videos, movie posters, and so on, so this would be I think a really cool resource in the film world, similar to what we have here with Van Gogh, and it is again all enabled because of those layers of encoding over the documents.

And then there is another example. I also want to talk, before I get into this publishing tool, we were looking at letters, TEI also comes with tags for performance texts, used for encoding novels, poetry, drama, screen plays, and so on. These are some of the tags that are there for encoding like dramas and screenplays, so cast lists, cast item, within cast item you have role, the description of the role and the actor, for speeches. In the screen play you have the speech and then you identify the speaker, the label, who is speaking, stage directions, specific types of stage directions, like movement of actors or characters, camera angles, sound effects or technical instructions in the screen play and so on, so these are all tags that could be used to encode screen plays. This is the chapter in the TEI guidelines on performance text and then some examples of coding. So here's, I know most of you can't see that. Here's a speech with the speaker, telling the stage direction. She's speaking calmly, then the paragraphs of dialogue, another stage direction, a song imbedded in this drama or screen play with some poetic lines, so LG is for a group of verse lines, a stanza of poetry, L is individual verse lines. So lots of examples in the online guidelines about how one might use these things. Here is a section on camera angles and sound and so on, so new angle shot cut. I don't know what that means, but I assume the rest of you do. Just to illustrate some of the thinking that has gone into this TEI community about text that might be of interest to you.

Shola Lynch: How much do these communities talk to one another? So for instance, in the Schomberg archive, the previous curator would start doing a lot of databasing, and then that particular software or that particular platform would go out of style or out of fashion, so there are all these fits and starts, so is there a way for these things to talk to one another?

Doug Reside: There is a lot of kind of discussion, especially now of trying to solve exactly that problem, with these kind of isolated pods of data.

Shola Lynch: It's a lot of work.

Doug Reside: Yes, and my criticism of TEI, it's exactly that. That its very, that TEI has kind of an isolated community, I would say. You can disagree.

John A. Walsh: Yes, it's isolated, but then if you want to do a project like this where you're transcribing text and making them searchable and that and you're looking for NEH funding, if you were not using the TEI, they would want an explanation for why not. So, it's a fairly small community, but for certain types of projects, it's expected that you use this.

Shola Lynch: So for a scholarly focused project that has a finite amount of time an archive, we're trying to think about doing this kind of stuff, but to accumulate it over years and decades that you can continue...

John A. Walsh: Right, and if your goal is to get a push out of massive amounts of content, then this is not an efficient way to go. It's big data versus close reading and close attention. I'm imagining all the correspondence related to this period of early black film is probably not that massive a corpus to deal with and that it could be done in a reasonable amount of time, and people, scholars have spent decades or generations doing critical additions of one famous author. People spend a lot of time doing this, they always have, and that's not really different. If your goal is just to push out digital content and put it up on the web, there are more efficient ways to do it, but if you want to do the deep kind of analysis, this approach is often useful.

Shola Lynch: Actually what I would like to be able to do is to have a platform in the cataloging that can be accessible to this kind, so it's not duplicating the effort. So if I digitize all the letters or all the films, how do I forward my collection for public use, for scholarly use? Is there a way for these platforms to talk together?

Doug Reside: There is an increasing move towards linked open data that uses a different kind of standard called RDF that TEI information could be expressed within and then typed up, so not everyone would have to use or rather the TEI tags could be related to a database at Schomburg, a database at the Library of Performing Arts, etc. But ...

Shola Lynch: But that's future talk.

Doug Reside: It's future talk.

Shola Lynch: I'm asking because I don't know. I'm not asking to be critical.

John A. Walsh: One of the points of this, you mentioned the software going out of date;

this is an XML based format. One of the points of that is it is plain text that can be read by any computer platform, any computer program, and it's designed to be future proof, so the tool I was going to show is a simple piece of software that interprets these codes and publishes them on the web, sort of as if it were HTML, and TEI has been around since 1987, before HTML and before the web, so it has a longevity to it.

Allyson Nadia Field: Can you just explain again the relationship between TEI and HTML? Because that looks, reads like HTML. Is that different?

John A. Walsh: It shares some things in common, like the paragraph, but HTML doesn't have a speaker tag or speech tag or an epigraph tag or tags...

Allyson Nadia Field: So it's more specific?

John A. Walsh: It's more specific; it's more semantically rich. HTML doesn't have a specific tag for camera angles, or stage direction. It's got tags that are designed by scholars that match the things they are interested in.

Allyson Nadia Field: Could WordPress, or whatever you're using to publish this on the web read that?

John A. Walsh: That's the tool that I was, wherever I went with that. This tool where it looks like a web page to you, HTML. It's really a hybrid. I've taken the TEI document and put it inside a HTML document, so all those TEI codes are still there, present in the web page, and if you're familiar with CSS, which is used to style HTML documents, you can use CSS on any arbitrary XML and that's what I'm doing here. So it looks like HTML but it has those more specific semantically rich tags in it. I think I've used up my 20 minutes and the rest was technical overview of that system and I don't know how useful that is to go through, but there are two approaches to publishing TEI content on the web. The typical approach is you use a programming language called XSLT to just convert your TEI to HTML. But then you've lost all that semantic richness by the time it reaches the user. They may still be taking advantage of that in the searching and browsing, but the page that is sitting in their web browser has been converted to the less expressive, less semantically rich HTML. This tool uses a hybrid approach where it uses that transformation language to create like an empty shell of a HTML document and put the TEI document inside there, giving you kind of the power of both at once. The rest is some more technical stuff about the programming and how that works, and I can talk at a break to anyone else about that. So here's again is that letter. This is interesting. So here is this Petrarch project I'm working on, this is a diplomatic transcription of his manuscript where we're recording his archaic, Petrarch's older style of punctuation and abbreviations and so on, and then I can use this drop down switch from the diplomatic transcription to the edited text and I have different punctuation, abbreviations are expanded, and so on, so I have the edition and the original, it's all in one document and the codes are in there, saying what's from the original, what has been modified or changed by the editor and then this drop down just says okay, show me the original

stuff, or flip it and show me the edited version, and you don't have two separate documents and have to try and keep them in synch when you discover a mistake or whatever, you've got one master document that contains multiple views or versions of the document within itself. So, I'll stop there.

Brian Graney: Thanks.