Tom Hyry
Remarks for panel on "Reconciling Modern Archival Practices and Ethics with Large Scale Digitization"
The Legal and Ethical Implications of Large-Scale Digitization of Manuscript Collections Symposium
Southern Historical Collections, University of North Carolina, Chapel Hill
12 February 2009
****

My short piece today looks at the next frontier of access to archival holdings in digital form and I will begin by imagining an access system coming to an archives near you in the not-so-distant future. Currently, our systems are comprised chiefly of metadata, created by archivists, that users search. That metadata then leads either to physical holdings or, increasingly, digitized holdings. As we've discussed today, those digitized holdings are far more discoverable now, and in an unmediated fashion, than they were when researchers needed to travel to our repositories to consult our holdings and this raises a host of ethical concerns.

What I would like to address is what we could call the third leg of the stool of emerging archival access systems: full text records, including born digital materials and transcribed oral histories and digitized collections.

I hesitate to bring born digital materials into this discussion, as I would like to avoid the quicksand of a conversation about electronic records. So a quick disclaimer: while fascinating and important in their own right, I do not want to devote any time or attention to the myriad issues related to the acquisition and preservation of born digital records, in order to stick as close as possible to today's topic of ethics, digitization, and contemporary archival practice. I would argue, however, that issues related to born-digital and other full text sources are relevant to our panel because I believe we will one day provide access to them in the same systems we use for our digitized holdings. We increasingly live in a world in which archival holdings come in hybrid form.

So, for the sake of the topic of this panel, I hope you will forgive me with brazenly going with three assumptions:

1. Repositories that document wide areas such as the American South need to include the born digital in their holdings in order to fulfill their core mission.
2. We will succeed in overcoming many of the myriad issues related to the acquisition and preservation of these materials
3. In addition to dealing with all of the hassles and difficulties of born digital holdings, we will also be compelled to utilize the chief advantages of materials in these formats.

Speaking to this third assumption: along with a whole host of difficulties, born digital records, at least text files such as email and word processing files, have at least three distinct advantages over analog counterparts: 1) they are portable and can be quickly sent or served over networks; 2) they usually have their own embedded metadata; and 3) in the right system, they can be searched in full text.

Moving into ethical issues, I would like you to imagine that I have a small archivist standing on each of my shoulders, whispering into my ear. Not an angel and a devil, just two solid and passionate professionals with differing points of view. The pro access archivist on my left shoulder, says: "This full text stuff is great. Users can log into an archival access system, perform complex searches on a vast array of metadata and free text data, return relevant hits, and have the digitized and born digital records from their search results served back to them over the network. Access to digital holdings has great potential for scholarship, as digital tools can not only foster discovery, but also provide the capacity for remixing, data mining, and other potential creative uses being developed in the emerging field of digital humanities. Having holdings in full text digital form will help us fill our missions of connecting users to our holdings better than we could have previously imagined. "

The archivist on the other shoulder says, "But wait a minute. Don't you think that email developed as a communication tool in a very informal way? Meaning that people have been far franker and more revealing? Making discovery of and access to these kinds of records easier makes for an even trickier situation than with their analog or digitized counterparts; it's kind of like third party privacy issues on steroids. And wouldn't a system that allows you to avoid the traditional contextual framework of the finding aid by drilling down into collections to search the records themselves run a great risk of separating content from context, making misinterpretation a significant possibility? And finally, the ease of creating digital records means that the volume of these collections are even greater than the mountain of paper we also face. Surely we couldn't effectively screen these records for private and confidential information during processing."

End of dialogue. So the competing ethics of providing open, broad and easy access and protecting private and confidential information becomes even more complicated.

In case you are thinking that these issues are less than immediate and can be put off until the future, I will note that most of these same issues apply to digitized holdings for which we can easily employ OCR to obtain full text records. Presumably this could be done to typescript holdings and arguably should be done for oral history transcripts. The Documenting the American South project already has made many transcripts and recordings of oral histories available and searchable in an impressive web site. Most of the same ethical issues arise for this material, as for born digital holdings.

All of this said, full text records also allow us an opportunity to screen for privacy and confidentiality not available for analog or digitized holdings. We could develop algorithms to perform automated searches for records that meet criteria we set up for privacy and confidentiality. Records that turn up as part of these searches could then be reviewed, segregated or automatically redacted. This technology already exists: recently at Yale, the university implemented a program that searched all computers on our network to identify Social Security Numbers and credit card numbers on users' hard drives, provided a list of those files to users, and asked them to delete or otherwise restrict them. The irony here, of course, is

that we would be back to defining, in a very explicit way, what we thought should be private and confidential. The more things change, the more they stay the same.

I will close by reiterating that the ethical issues surrounding access to archival holdings promise to become more complicated in the future, not less. Archivists do have a responsibility to be as transparent as possible to their creators and users about our purposes and methods. Respecting privacy and confidentiality fit into this ethical calculus, but we have to remember that access to archival holdings is also an ethical consideration, as we head off into the future.