White Dudes Giving Speeches

Thank you for inviting me here today to be here with you all here at MARAC. I'll admit that I'm more than a bit nervous to be up here. I normally apologize for being a software developer right about now. But I'm not going to do that today. . . although I guess I just did. I'm not paying software developers any compliments by using them as a scapegoat for my public presentation skills. And the truth is that I've seen plenty of software developers give moving and inspiring talks.

The reason why I'm a bit more nervous today than usual is because you are *archivists*. I don't need to #askanarchivist to know that you think differently about things, in subtle and profound ways. To paraphrase Orwell: You work in the present to shape what we know about the past, in order to help create our future. You are a bunch of time travelers. How do you decide what to hold on to, and what to let go of? How do you care for this material? How do you let people know about it? You do this for organizations, communities and collectively for entire cultures. I want to pause a moment to thank you for your work. You should applaud yourselves.

My Twitter profile used to say I was a "hacker for libraries". I changed it a few years ago to "pragmatist, archivist and humanist". But the reality is that these are aspirational..these are the things I *want* to be. I have major imposter syndrome about claiming to be an archivist. That's why I'm nervous.

Can you believe that I went through a Masters in Library & Information Science program without learning a lick about archival theory? Maybe I just picked the wrong classes, but this was truly a missed opportunity, both for me and the school. After graduating I spent some time working with metadata in libraries, then as a software developer at a startup, then at a publisher, and then in government. It was in this last role helping bits move around at the Library of Congress (yes, some of the bits did still move, kinda, sorta) that I realized how much I had missed about the theory of archives in school.

I found that the literature of archives and archival theory spoke directly to what I was doing as a software developer in the area of digital preservation. With guidance from friends and colleagues I read about archives from people like Hugh Taylor, Helen Samuels, "the Terrys" (Cook and Eastwood), Verne Harris, Ernst Posner, Heather MacNeil, Sue McKemmish, Randall Jimerson, Tom Nesmith and more. I started following some of you on Twitter to read the tea leaves of

the profession. I became a member of SAA. I put some of the ideas into practice in my work. I felt like I was getting on a well traveled but not widely known road. I guess it was more like a path among many paths. It definitely wasn't an information superhighway. I have a lot more still to learn.

So why am *I* up *here* talking to *you*? This isn't about *me* right? It's about *we*. So I would like to talk about this thing *we* do, namely create things like this:

Don't worry I'm not really going to be talking about the creation of finding aids. I think that they are something we all roughly understand. We use them to manage physical and intellectual access to our collections right? Finding aids are also used by researchers to discover what collections we have. Hey, it happens. Instead what I would like to focus on in this talk is the nature of this particular collection. What are the records being described here?

Yes, they are tweets that the Cuban Heritage Collection at the University of Miami collected after the announcement by President Obama on December 17, 2014 that the United States was going to begin normalizing relations with Cuba. You can see information about what format the data is in, when the data was collection, how it was collected, how much data there is, and the rights associated with the data.

Why would you want to do this? What can 25 million tweets tell us about the public reaction to Obama's announcement? What will they tell us in 10, 25 or 50 years? Nathalie is thinking they could tell us a lot, and I think she is right. What I like about Nathalie's work is that she has managed to fold this data collection work in with the traditional work of the archive. I know there were some technical hoops to jump through regarding data collection, but the social engineering required to get people working together as a team so that data collection leads leads to processing and then to product in a timely manner is what I thought was truly impressive. Nathalie got in touch with Bergis Jules and I to help with some of the technical pieces since she knew that we had done some similar work in this area before. I thought I would tell you about how that work came to be. But if you take nothing else from my talk today take this example of Nathalie's work.

About a year ago I was at SAA in Washington, DC on a panel that Hillel Arnold set up to talk about Agency, Ethics and Information. Here's a quote from the panel description:

> From the Internet activism of Aaron Swartz to Wikileaks' release of confidential U.S. diplomatic cables, numerous events in recent years have challenged the scope and implications of privacy, confidentiality, and access for archives and archivists. With them comes the opportunity, and perhaps the imperative, to interrogate the role of power, ethics, and regulation in information systems. How are we to engage with these questions as archivists and citizens, and what are their implications for user access?

My contribution to the panel was to talk about the Facebook Emotional Contagion study, and to try to get people thinking about Facebook as an archive (more about that in a bit). In the question and answer period someone (I wish I could remember his name) asked what archivists were doing to collect what was happening in social media and on the Web regarding the protests in Ferguson. The panel was on August 14th, just 5 days after Mike Brown was killed by police officer Darren Wilson in Ferguson, Missouri. It was starting to turn into a national story, but only after a large amount of protest, discussion and on the ground coverage happening also in Twitter. Someone helpfully pointed out that just a few hours earlier ArchiveIt (the Internet Archive's subscription service) had announced that it was seeking nominations of web pages to archive webpages related to the events in Ferguson. We can see today that close to 981 pages were collected. 236 of those were submitted using the the form that Internet Archive made available.

But what about the conversation that was happening in Twitter? That's what you've been watching a little bit of for the past few minutes up on the screen here. Right after the panel discussion a group of people made there way to the hotel bar to continue the discussion. At some point I remember talking to Bergis Jules who impressed on me the importance of trying to do what we could to collect the torrent of conversation about Ferguson going on in Twitter. I had done some work collecting data from the Twitter API before and offered to lend a hand. Little did I know what would happen.

When we stopped this initial round of data collection we had collected 13,480,000 tweets that mentioned the word "ferguson" between August 10, 2014 and August 27, 2014.

You can see from this graph of tweets per day, that there were definite cycles in the Twitter traffic. In fact the volume was so high at times, and we had started data collection 6 days late, that there are periods you can see there were periods where we weren't able to get the tweets. You might be wondering what this data collection looks like. Before looking closer at the data let me try to demystify it a little bit for you.

Here is a page from the online documentation for Twitter's API. If you haven't heard the term API before it stands for Application Programming Interface, and that's just a fancy name for a website that delivers up data (such as XML or JSON) instead of human readable web pages. If you have a Twitter app on your phone it most likely uses Twitter's API to access the tweets of people you follow. Twitter isn't the only place making APIs available: they are *everywhere* on the Web: Facebook, Google, YouTube, Wikipedia, OCLC, even the Library of Congress has APIs. In some ways if you make your EAD XML available on the Web it is a kind of API. I really hope I didn't just mansplain what APIs are, that's not what I was trying to do.

A single API can support multiple "calls" or questions that you can ask. Among the many calls Twitter's API has a call that allows you to do a search, and get

back 100 tweets that match your query plus a token to go and get the next 100. They let you ask this question 180 times every 15 minutes. If you do the math you can see that you can fetch 72,000 tweets an hour, or 1.7 million tweets per day. Unfortunately the API only lets you search the last 9 days of tweets, after which you can pay Twitter for data.

So what Bergis and I did was use a small Python program I had written previously called twarc to try to collect as much of the tweets as we could that had the word "ferguson" in them. twarc is just one tool for collecting data from the Twitter API.

Another tool you can use from the comfort of your Web browser (no command line fu required) is the popular Twitter Archiving Google Sheet TAGS. TAGS lets you collect data from the search API which it puts directly into a spreadsheet for analysis. This is super handy if you don't want to parse the original JSON data returned by the Twitter API. TAGS is great for individual use.

And another option is the Social Feed Manager (SFM) project. SFM is a project started by George Washington University with support from IMLS and the National Historical Publications and Records Commission. I think SFM is doubly important to bring up today since the theme for the conference is Ingenuity and Innovation in Archives. NHPRC's support for the SFM project has been instrumental in getting it to where it is today. SFM is an open source Web application that you download and set up at your institution and which users then log into using their Twitter credentials to to setup data collection jobs. GWU named it Social Feed Manager because they are in the process of adding other content sources such as Flickr and Tumblr. They are hoping that extending it in this way will provide an architecture that will allow interested people to add other social media sites, and contribute them back to the project. The other nice thing that both SFM and twarc do (but that TAGS does not) is collect the original JSON data from the Twitter API. In a world where original order matters I think this is an important factor keep in mind.

JSON is an acronym for JavaScript Object Notation. There are lots of other formats for sending data around on the Web, but JSON has emerged as the defacto standard for APIs. This has largely been the result of its versatility and that support for it is cooked into every Web browser that can run JavaScript.

So what's in the JSON data for a tweet? Twitter is famous for its 140 character message limit. But the text of a tweet only accounts for about 2% of the JSON data that is made available by the Twitter API. Some people might call this metadata, but I'm just going to call it data for now, since this is the original data that Twitter collected and pushed out to any clients that are listening for it.

Also included in the JSON data are things like: the time that the tweet was sent, any hashtags present, geo coordinates for the user (if they have geo-location turned on in their preferences), urls mentioned, places recognized, embedded media such as images or videos, retweet information, reply to information,

lots information about the user sending the message, handles for other users mentioned, the current follow count of the sender. And of course you can use the tweet ID to go back to the Twitter API to get all sorts of information such as who has retweeted or liked a tweet.

Here's what the JSON looks like for a tweet. I'm not going to go into this in detail, but I thought I would at least show it to you. I suspect catalogers or EAD finding aid creators out there might not find this too scary to look at. JSON is much more expressive than the rows and columns of a spreadsheet because you can have lists of things, key/value pairs and hierarchical structures that don't fit comfortably into a spreadsheet.

Ok, I can imagine some eyes glazing over at these mundane details so lets get back to the Ferguson tweets. Do you remember that form that ArchiveIt put together to let people submit URLs to archive? You may remember that 236 URLs were submitted. Bergis and I were curious so we extracted all the URLs mentioned in the 13 million tweets, unshortened them, and then ranked them by the number of times they were mentioned. You can see a list of the [top 100 shared links] in that time period. Notice at the time we checked to see if Internet Archive had archived the page.

We then took a look just within the first day of tweets that we had to eyeball what the most tweeted URLs tweeted initially looked like. Look at number #2 there, Racial Profiling Data/2013. It's a government document from the Missouri Attorney General's Office with statistics from the Ferguson Police Department. Let's take a moment to digest those stats along with the 1,538 Twitter users who did that day.

Now what's interesting is that the URL that was tweeted so many times then is already broken. And look, Internet Archive has it, but it was collected for the first time on August 12, 2014. Just as the conversation was erupting on Twitter. Perhaps this URL was submitted by an archivist to the form ArchiveIt put together. Or perhaps someone recognized the importance of archiving it and submitted it directly to the Internet Archive using their Save Now form.

The thing I didn't mention earlier is that we found 417,972 unique, unshortened URLs. Among them were 21,457 YouTube videos. Here's the fourth most shared YouTube video, that ended up being seen over half a million times.

As Bergis said in July of this year as he prepared for a [class] about archiving social media at Archival Education and Research Initiative (AERI):

Every time I hear we shouldn't build social media archives like #Ferguson, I think abt events in black history for which we have no records.

— Bergis Jules (@BergisJules) July 18, 2015

Bergis was thinking specifically about events like the Watts Riots in Los Angeles where 34 people were killed and 3,438 arrested.

Of course the story does not end there. As I mentioned I work at the Maryland Institute for Technology in the Humanities at the University of Maryland. We aren't an archive or a library, we're a digital humanities lab that is closely affiliated with the University library. Neil Fraistat, the director of MITH, immediately recognized the value of doing this work. He not only supported me in spending time on it with Bergis, but also talked about the work with his olleagues at the University.

When there was a Town Hall meeting on December 3, 2014 we were invited to speak along with other faculty, students and the University Police Commissioner. The slides you saw earlier of popular tweets during that period was originally created for the Town Hall. I spoke very briefly about the data we collected and invited students and faculty who were interested in working with the data to please get in touch. The meeting was attended by hundreds of students, and ended up lasting some 4 hours, with most of the time being taken up by students sharing stories from their experience on campus of harassment by police, concern about military grade weapons being deployed in the community, insight into the forms of institutionalized racism that we all still live with today. It was an incredibly moving experience, and our images from the "archive" were playing the whole time as a backdrop.

After the Town Hall meeting Neil and a group of faculty on campus organized the BlackLivesMatter at UMD group a set of teach-ins at UMD where the regularly scheduled syllabus was set aside to discuss the events in Ferguson and Black Lives Matter more generally. Porter Olsen (who taught the BitCurator workshop yesterday) helped organize a series of sessions we call digital incubators to build a community of practice around digital methods in the humanities. These sessions focused on tools for data collection, data analysis and advanced data analysis and rights and ethical issues. We had Laura Wrubel visit from George Washington University to talk about Social Feed Manager. Trevor Munoz, Katie Shilton and Ricky Punzalan spoke addressed the rights issues associated with working with the data. Josh Westgaard from the library spoke about working with JSON data. And Cody Buntain , Nick Diakopoulos and Ben Scheiderman helped us thing about using tools like Python Notebooks and NodeXL for data analysis.

And of course, we didn't know it at the time, but Ferguson was just the beginning. Or rather it was the beginning of a growing awareness of police injustice towards African Americans and people of color in the United States that began to be known as the BlackLivesMatter movement. BlackLivesMatter was actually started by Alicia Garza, Patrisse Cullors, and Opal Tometi after the acquittal of George Zimmerman in the Florida shooting death of Trayvon Martin two years earlier. But the protests on the ground in Ferguson, elsewhere in the US, and in social media brought international attention to the issue. Names like Aiyana Jones, Rekia Boyd, Jordan Davis, Renisha McBride, Dontre Hamilton, Eric Garner, John Crawford, led up to Michael Brown, and were followed by Tamir Rice, Antonio Martin, Walter Scott, Freddie Gray, Sandra Bland and Samuel Dubose.

Bergis and I did our best to collect what we could from these sad, terrifying and enraging events. The protests in Baltimore were of particular interest to us at the University of Maryland since it was right in our backyard. Our data collection efforts got the attention of Rashawn Ray who is a professor in sociology at the University of Maryland. He and his student Melissa Brown were interested in studying how the discussion of Ferguson changed in four datasets we had collected: the initial killing of Michael Brown, the non-indictment of Darren Wilson, the Justice Department Report and then the one year anniversary. They have been exploring what the hashtags, images and text tell us about the shaping of narratives, sub-narratives and counter-narratives around the Black experience in the United States.

And we haven't even accessioned any of the data. It's sitting in MITH's Amazon cloud storage. This really isn't anybody's fault but my own. I haven't made it a priority to figure out how to get it into the University's Fedora repository. In theory it should be doable. This is why I'm such a fan of Nathalie's work at the University of Miami that I mentioned at the beginning. Not only did she get the data into the archive, but she described it with a finding aid that is now on the Web, waiting to be discovered by a researcher like Rashawn.

So what value do you think social media has as a tool for guiding appraisal in Web archives? Let me read you the first paragraph of a grant proposal Bergis wrote recently:

> The dramatic rise in the public's use of social media services to document events of historical significance presents archivists and others who build primary source research collections with a unique opportunity to transform appraisal, collecting, preservation and discovery of this new type of research data. The personal nature of documenting participation in historical events also presents researchers with new opportunities to engage with the data generated by individual users of services such as Twitter, which has emerged as one of the most important tools used in social activism to build support, share information and remain engaged. Twitter users document activities or events through the text, images, videos and audio embedded in or linked from their tweets. This means vast amounts of digital content is being shared and re-shared using Twitter as a platform for other social media applications like YouTube, Instagram, Flickr and the Web at large. While such digital content adds a new layer of documentary evidence that is immensely valuable to those interested in researching and understanding contemporary events, it also presents significant data management, rights management, access and visualization challenges.

As with all good ideas, we're not alone in seeing the usefulness of social media in archival work. Ed Fox and his team just down the road at Virginia Tech

have been working solidly on this problem for a few years and recently received an NSF grant to further develop their Integrated Digital Event Archiving and Library IDEAL. Here's a paragraph from their grant proposal:

> The Integrated Digital Event Archive and Library (IDEAL) system addresses the need for combining the best of digital library and archive technologies in support of stakeholders who are remembering and/or studying important events. It extends the work at Virginia Tech on the Crisis, Tragedy, and Recovery network (see http://www.ctrnet.net) to handle government and community events, in addition to a range of significant natural or manmade disasters. It addresses needs of those interested in emergency preparedness/response, digital government, and the social sciences. It proves the effectiveness of the 5S (Societies, Scenarios, Spaces, Structures, Streams) approach to intelligent information systems by crawling and archiving events of broad interest. It leverages and extends the capabilities of the Internet Archive to develop spontaneous event collections that can be permanently archived as well as searched and accessed, and of the LucidWorks Big Data software that supports scalable indexing, analyzing, and accessing of very large collections.

Maybe you should have another Ed up here speaking! Or an archivist like Bergis.

Seriously though, this has been fun. Before I leave you here are a few places you could go to get involved in and learn about this work.

1. If you are an SAA member please join the conversation at the SAA Web Archiving discussion list. One of the cool things that happend on this discussion list last year was drafting a letter to Facebook that was sent
2. If you're not an SAA member there's a new discussion list called Web Archives. It's just getting started, so it's a perfect time to join.
3. Bergis, Christie Peterson, Bert Lyons, Ryan Baumann and I have been writing short little stories about this kind of work on [Medium] in the On Archivy publication. If you have ideas for little stories, thought experiments, actual work, ideas or commentary please write it on Medium and send us request to include it.

And as Hillel Arnold pointed out recently:

Your semi-regular reminder that direct, local action is what makes change happen, not white dudes giving speeches. #saa15 #s610

— Hillel Arnold (@helrond) August 22, 2015

Let's get to work.