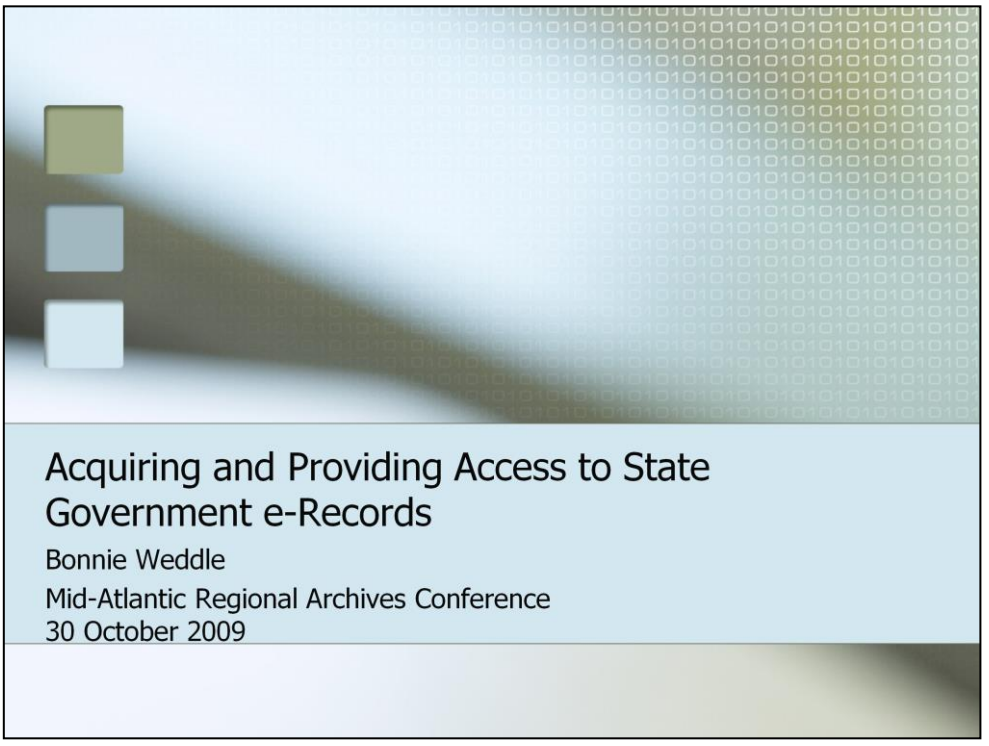
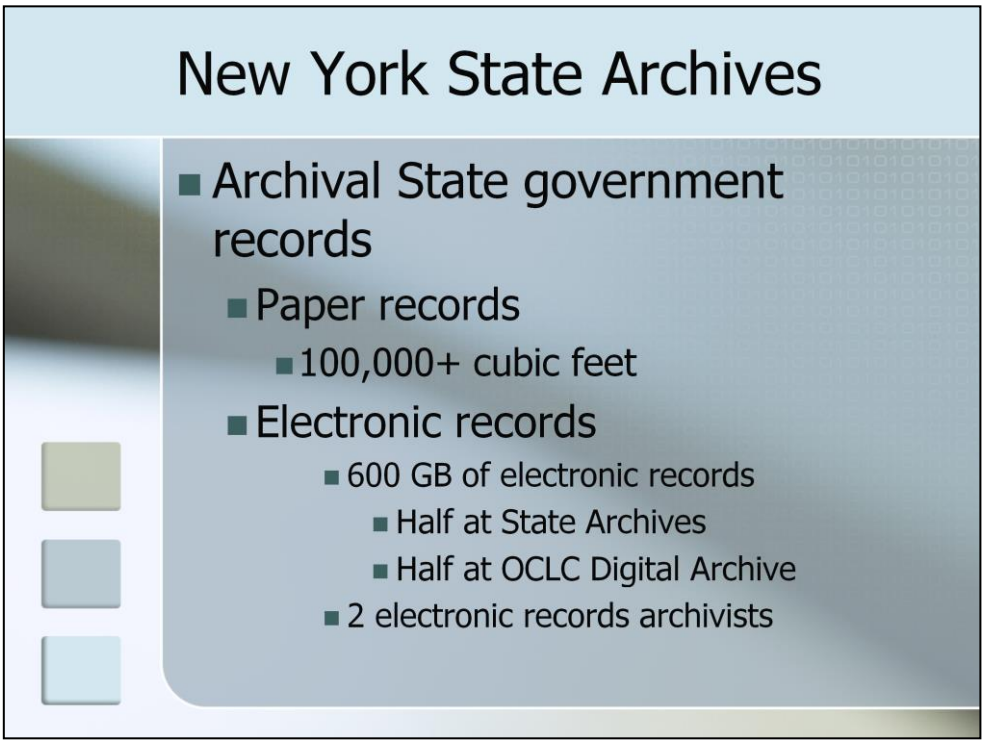


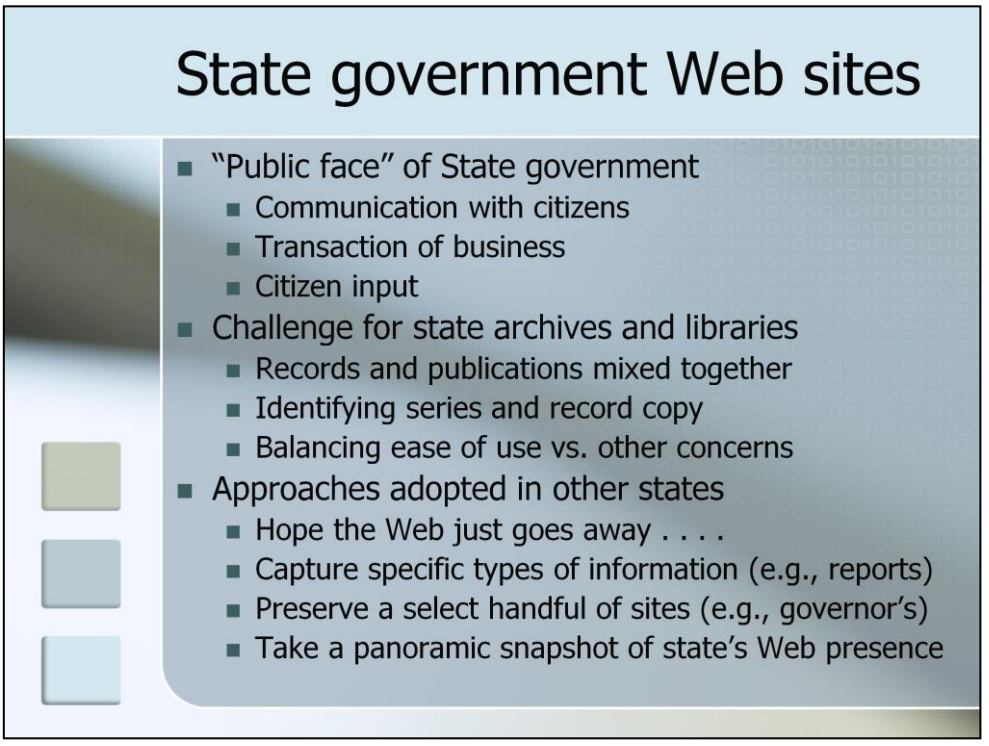
Good morning. For the next few minutes, I'm going to talk about how we've captured State government Web sites and how we've made electronic records containing legally restricted information accessible to a researcher.



First, however, a bit of background. The New York State Archives is responsible , among other things, for preserving and providing access to the archival records of New York State government. We hold over 100,000 cubic feet of paper records, and roughly 600 gigabytes of electronic records, half of which are housed at the State Archives and half of which are stored in OCLC's Digital Archive. We have two full-time electronic records archivists on staff, and other professional staff have some electronic records responsibilities as well.

A few months ago at SAA, Cal Lee asserted that there were three main strategies for dealing with electronic records created by individuals: getting the records on storage or transfer media, grabbing (or in other words, copying) the materials they place on the Web, and furnishing guidance to creators struggling to evaluate terms of service, preserve their records, and get their data out of the cloud. These strategies also work for records created by government agencies and other corporate bodies.

The first half of this presentation will focus on our grabbing of State government Web sites.

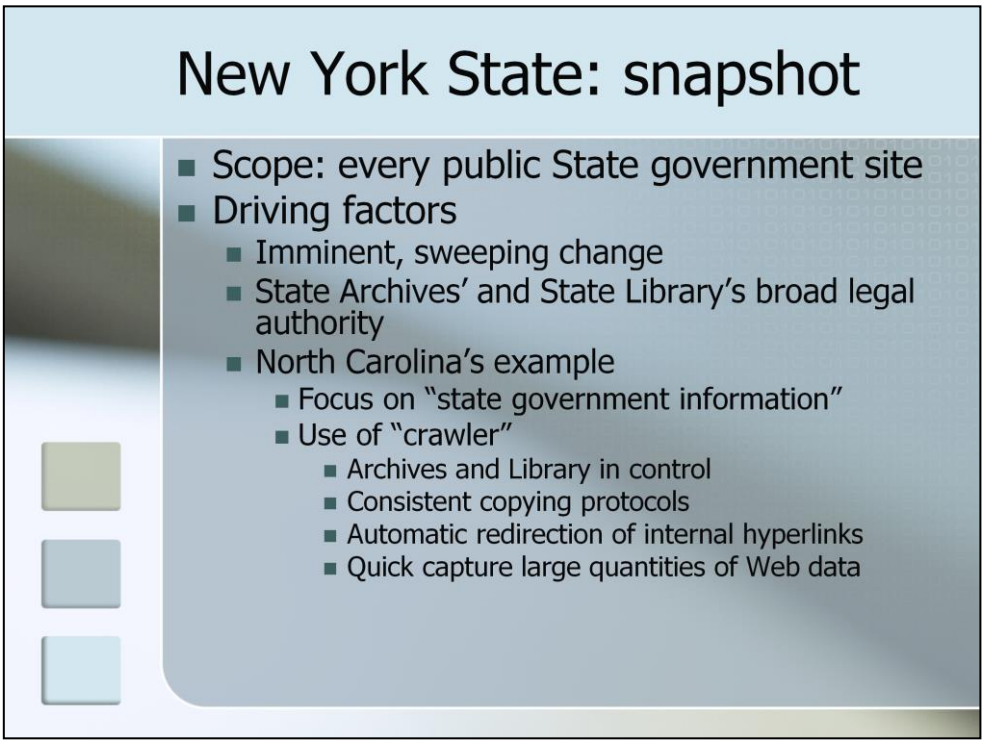


Web sites are increasingly the public face of government: they provide information about policies and programs an facilitate the transaction of business, and with the rise of Web 2.0, they're starting to serve as mechanisms for citizen input.

However, Web sites pose something of a challenge for archivists and librarians. Some Web content clearly meets the traditional definition of a record, some of it clearly meets the traditional definition of a government document, some of it resists easy classification, and all of it's mixed together. From an archival perspective, identifying series nestled within Web sites can be a real challenge. Web site files have traditionally been kept in file directories, and archivists could look at a site's directory structure and identify discrete series. However, a growing number of sites are database-driven, and it's getting harder to figure out where series begin and end.

Determining the record copy can also be a real challenge. Some agencies use the Web to disseminate convenience copies, but others use it as a digital filing cabinet; this is bad recordkeeping practice, but it persists nonetheless. However, even if all of the records on a given site are convenience copies, what if the convenience copies are a lot easier to access than the record copies? And what if the sum of the site is more than its parts? In other words, what if the site's structure and content convey important information about the creator that isn't fully reflected in other records or publications?

When Governor George Pataki announced in early 2006 that he would not seek a fourth term of office, the New York State Archives and New York State Library had no choice but to start thinking about how the State's Web presence would change as a result. We began examining how state archives and state libraries approached the Web. Were they hoping that the Web would just go away? Were they capturing specific types of information, such as publications or the texts of speeches? Were they preserving a select number of Web sites, such as those maintained by governors and other high-level elected officials? Or were they taking panoramic snapshots of their states' Web presence?



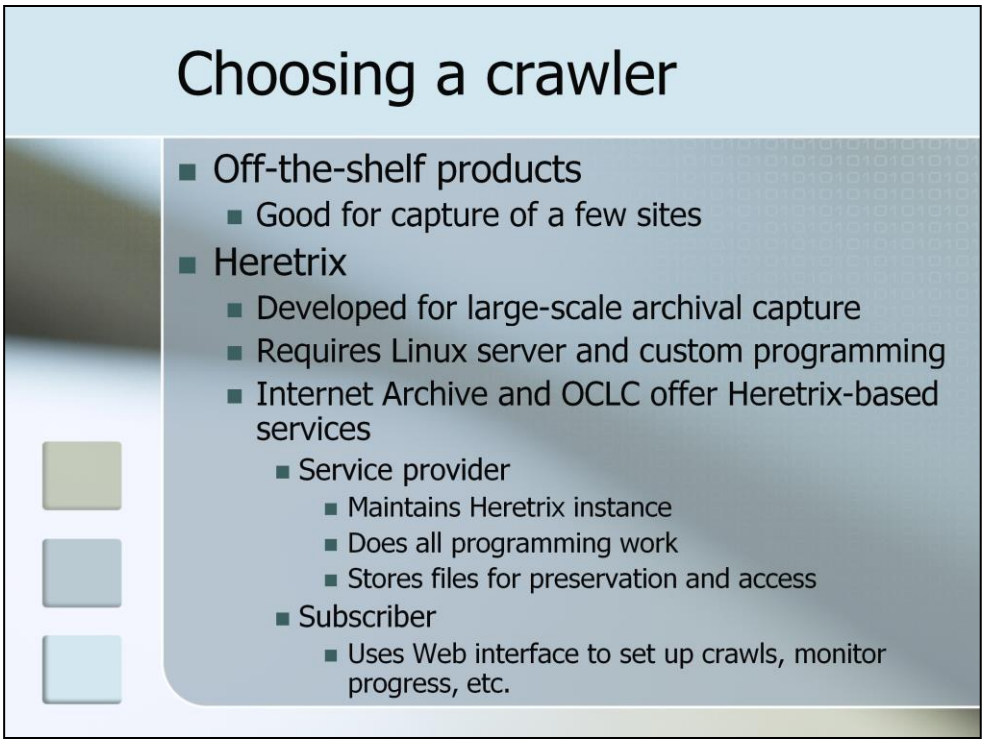
The New York State Archives and the New York State Library ultimately opted to take a panoramic snapshot of every publicly accessible State government Web site--executive, legislative, and judicial--as it existed in late 2006.

Why did we opt for this approach? First off, it was increasingly apparent that the State would experience a change of both officeholder and party. The State's Web presence would almost certainly change dramatically, and we simply didn't have the time needed to schedule every State government site.

Moreover, we were impressed with the example set by the State Archives and State Library in North Carolina, which ended a lengthy, unproductive debate about whether state government Web sites were records or publications by deciding that the sites contained important "state government information" that they would jointly capture and preserve.

We also liked North Carolina's decision to use a "crawler" to capture state government sites. Crawlers are software programs that move through a given Web site one page at a time, following links to other pages on the site until it has accessed all of them, much as a spider walks on a web. They're most commonly used to index content for search engines, but they have other uses.

Using a crawler put the North Carolina State Archives and State Library firmly in charge of capturing the information they wished to preserve and ensured that all of the copies were produced in the same way. The crawler also redirected each captured site's internal hyperlinks so that users can navigate through the archival copies much as they would the original sites. Finally, at present, crawling is really the only practical way to capture large quantities of Web data in a short period of time.



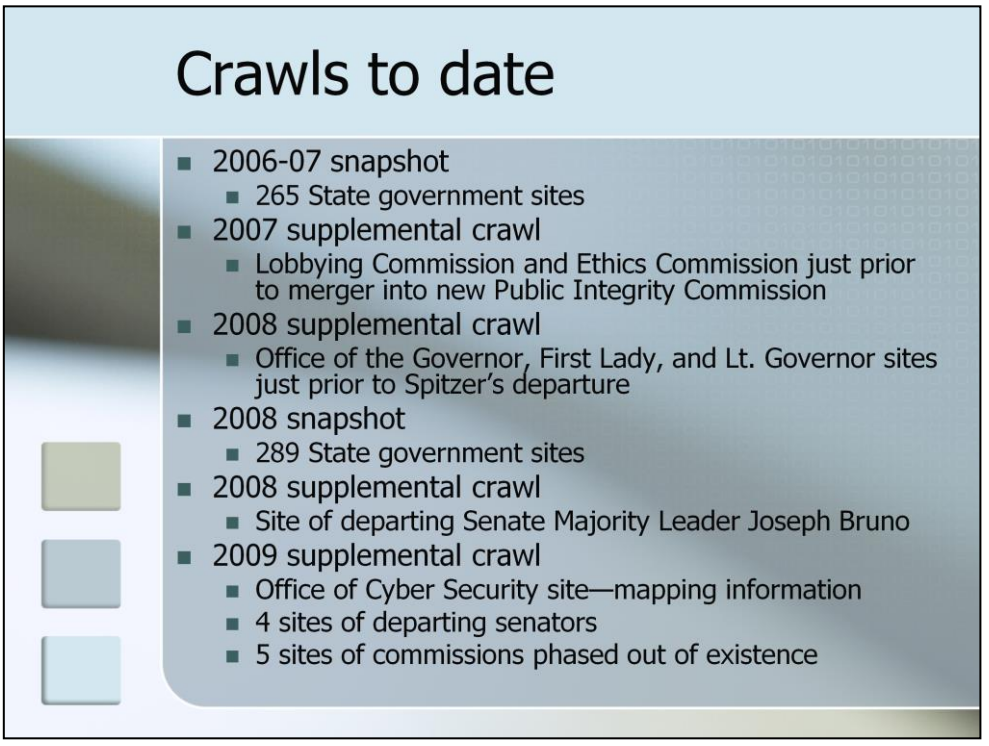
We then began exploring our crawling options.

There are a number of inexpensive commercial software packages that will copy Web sites and redirect internal hyperlinks so that the copies have both the look and feel of the originals. These packages are good options for people who need to preserve one site or a handful of sites, but New York State government maintains hundreds of Web sites.

At present, Heretrix, which was developed by the Internet Archive, is the only crawler suitable for archival capture of mass quantities of data. However, it requires a Linux server environment and a substantial amount of custom programming, and the New York State Archives and State Library have limited programming resources.

Fortunately, there are two vendors that, for a fee, offer Heretrix-based crawling and storage services. The Internet Archive and OCLC maintain Heretrix instances, do all the programming work, and store crawl results for preservation and access. They also provide Web-based interfaces that allow subscribers to specify which sites should be crawled, monitor the crawler's progress, and add descriptive metadata.

We talked with both Internet Archive and OCLC, and both of them, in my view, offer good services. However, we ultimately opted to go with OCLC.



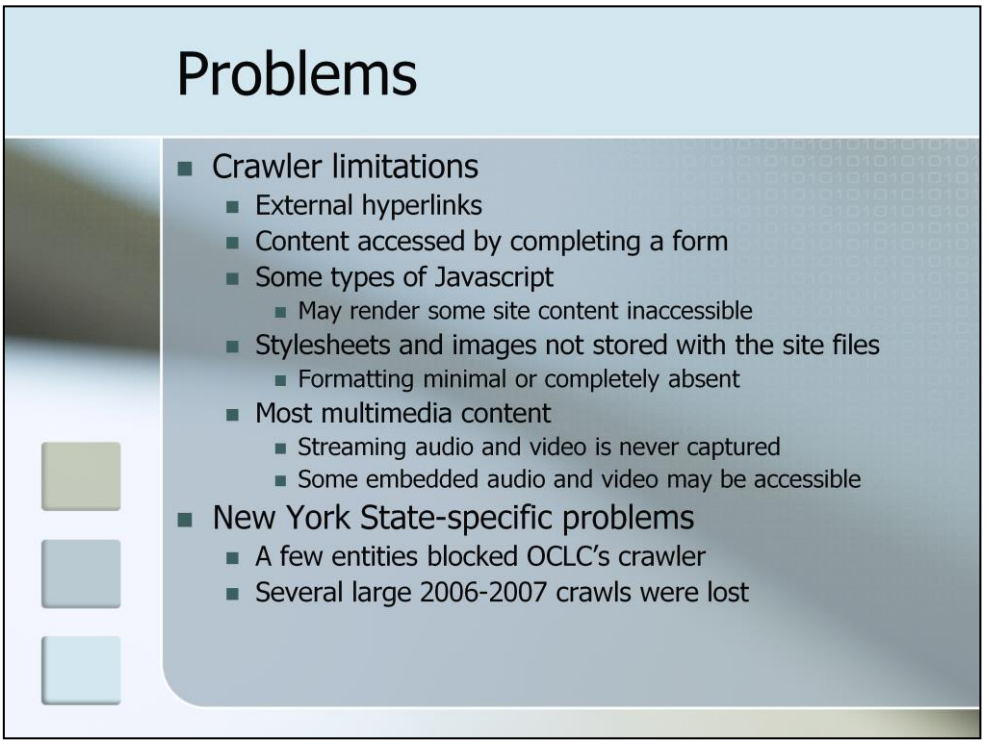
In December 2006, we gained access to OCLC's crawl utility and furiously began capturing State government sites. By the time we finished in early 2007, we had captured 265 sites comprising roughly 100 gigabytes of data.

Since that time, we've completed a number of supplemental crawls of sites of elected officials who left office and of commissions and other bodies that were merged with other bodies or phased out of existence.

In spring 2008, we completed a panoramic snapshot that we hoped would document the changes wrought by Governor Eliot Spitzer. It did so, albeit in a completely unexpected manner: we got the end of the Spitzer administration and the start of Governor David Paterson's administration.

We plan to take another panoramic snapshot after the 2010 election, and we'll continue doing supplemental crawls as circumstances dictate.

At present, we've captured about 295 gigabytes of Web site files, preservation copies of which are stored in OCLC's Digital Archive, and access copies of which are housed in CONTENTdm.



In theory, working with a service provider to capture Web sites is simple and tidy, and in many instances, it is. However, as with any large, technology-driven project, we did run into some problems.

Crawling is at this point the only practical way to capture large quantities of Web data, but it can't capture some types of Web content, including:

External hyperlinks. For example, if the Governor's Web site contains a link to a page on the State Health Department Web site, the crawler can't follow the link and capture that page.

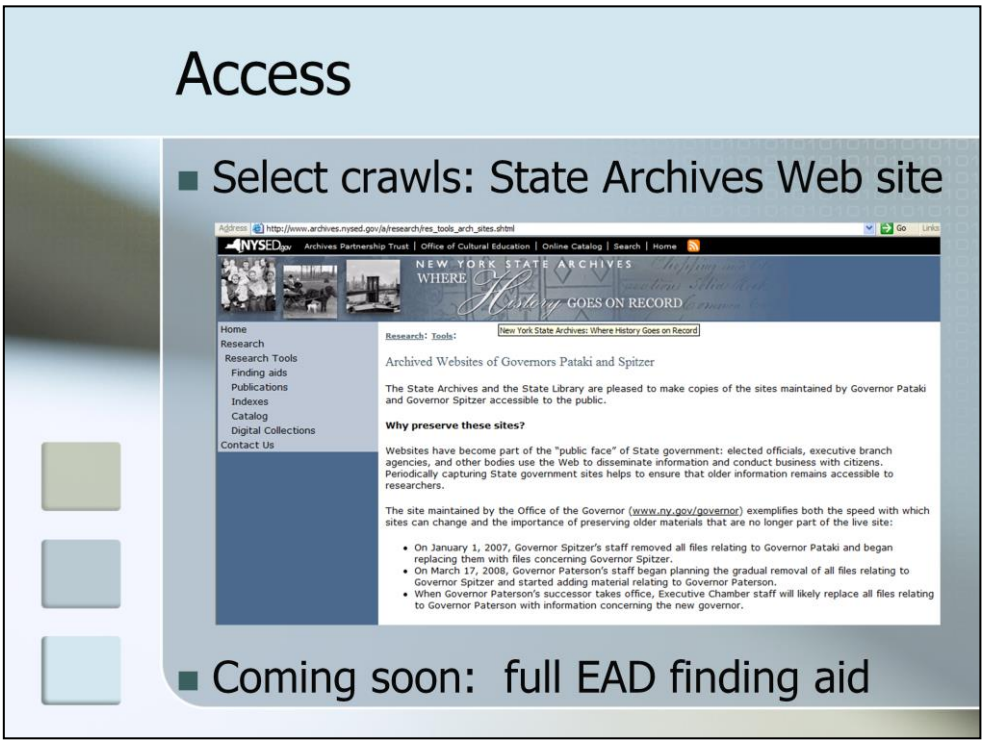
Information that is accessed by completing a form such as a login and password or a search screen.

External images and stylesheets. Crawlers can capture the site's text but not the accompanying images or the formatting supplied by the stylesheets.

Javascript. In some instances, Javascript keeps Heritrix and other crawlers from capturing some site files; in others, the files are captured but can't be displayed properly. Javascript display has gotten better since late 2006, but there's still a lot of room for improvement.

Multimedia files. We can capture some embedded audio and video files, but crawlers simply can't capture streaming audio and video.

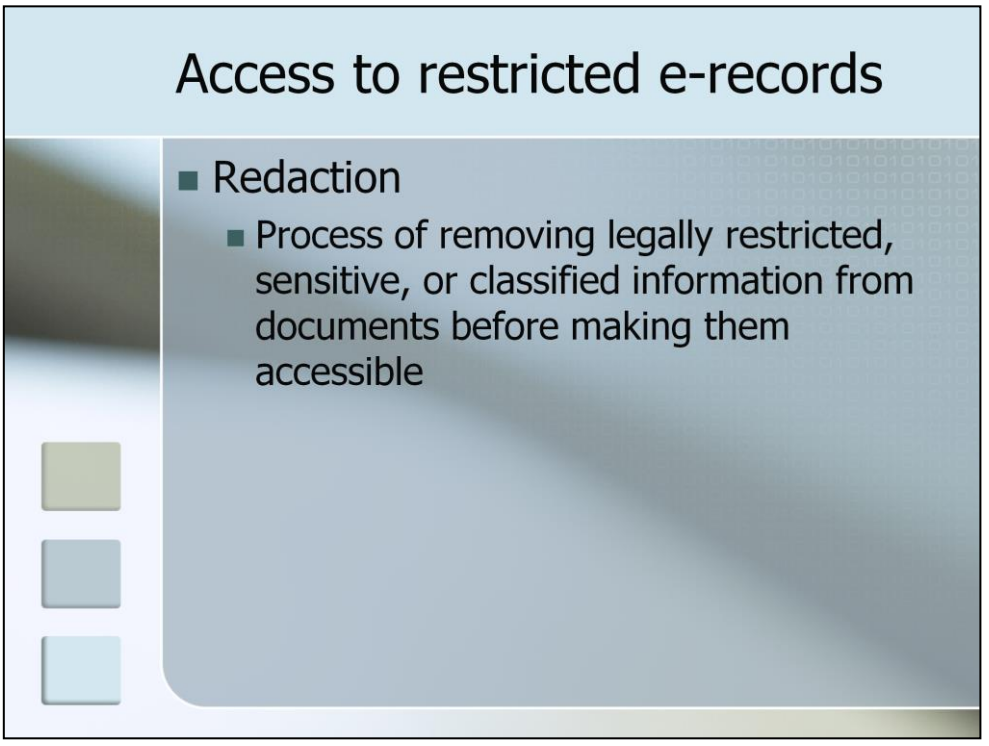
We also ran into some unexpected problems. Several entities have consistently blocked OCLC's crawler, and in 2006-07, we lost several crawls of very large sites because OCLC's crawl utility simply couldn't process them properly.



At present, we're focusing on finding the best way to make all of our crawls accessible. We need to make sure that researchers don't mistake our archived sites for the live versions, and we also need to explain how researchers can get around some of the problems associated with crawling's limitations; for example, Javascript issues sometimes disable navigational links on the left hand side of each page on a given site, but navigational links at the bottom of each page do work.

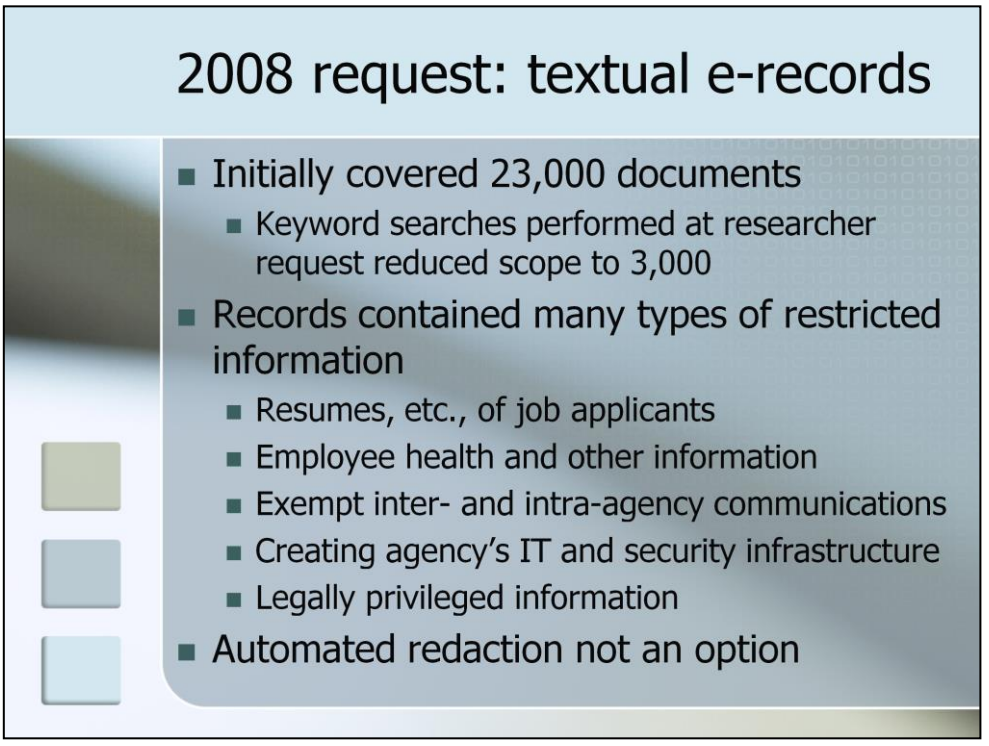
At the moment, we're using our Web site to publicize and provide access to our crawls of Governor Pataki's and Governor Spitzer's sites, and we've made other crawls accessible to researchers upon request. In a few weeks, we're going to expand the Web page to include our 2006 crawl of the Office of the Attorney General's Web site, which we captured a few days before Attorney General Spitzer became Governor Spitzer. The updated page will also include our crawls of sites of several defunct commissions whose sites have disappeared from the live Web.

As far as long-term plans are concerned, we're currently thinking of creating an EAD finding aid that will make all of our crawls readily accessible to the public.



Next, I want to talk about providing access to electronic records that contain legally restricted information. I'm going to focus on how we dealt with a recent researcher request, and I must emphasize up front that I am ethically obligated to be a bit vague about certain aspects of the request and our response to it. However, I think that the information I can share is interesting and might prove useful to other archivists.

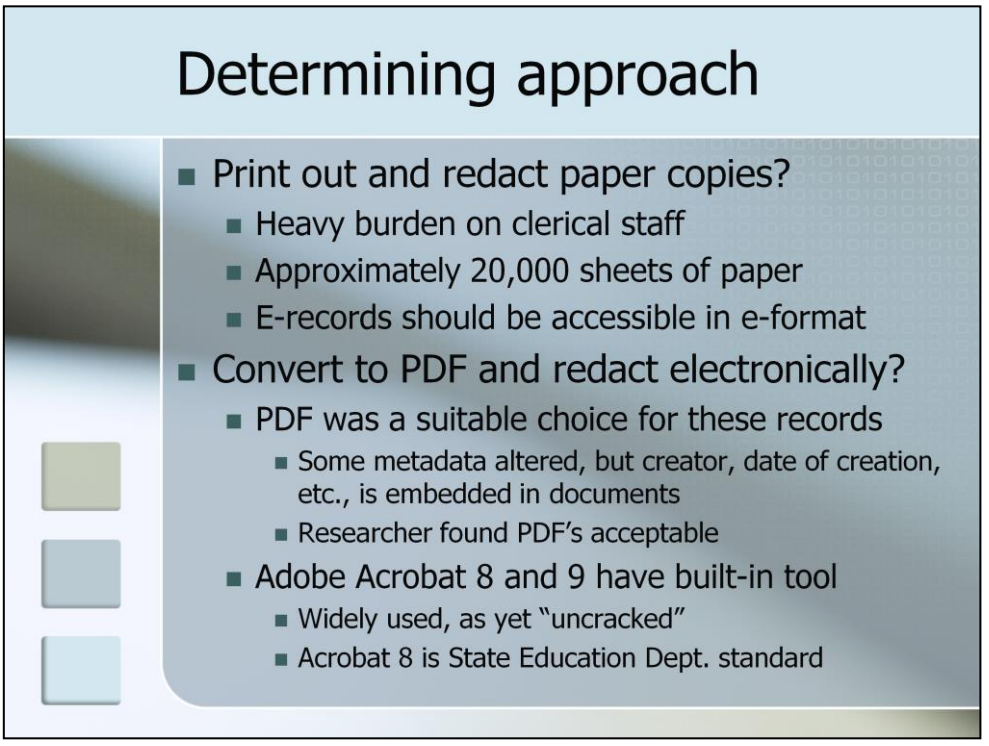
I also need to define up front a term that I'm going to use quite heavily during the next few minutes: redaction. Redaction is the process of removing legally restricted, sensitive, or classified information from documents before making them accessible to others. At the New York State Archives, our focus is solely on legally restricted information; terms such as "sensitive" or "confidential" have no legal meaning in New York State government, and we don't hold any classified records.



In 2008, a researcher requested access to approximately 23,000 electronic documents of widely varying length; we worked with the researcher and were able to narrow the scope of the request to roughly 3,000 documents – again, of widely varying length. These records had been transferred to us without advance notice a short time before, and our initial analysis revealed that legally restricted information was scattered throughout.

Types of restricted information include resumes of prospective employees and medical information, home addresses, and other information that would have compromised the personal privacy of State employees. We also found information about the creating agency's information technology infrastructure and the placement of cameras and other security mechanisms within its office spaces. Finally, some of the records contained the advice of agency counsel and other attorney work product.

Owing to the nature and diversity of the restricted information, we realized almost instantly that automated redaction was not possible. If, for example, we had been working with electronic forms that contained Social Security Numbers in a given field, we might have been able to use software that could hone in on that field and remove all of the data in it. However, we had no choice but to have professional staff read through and evaluate the content of each document.

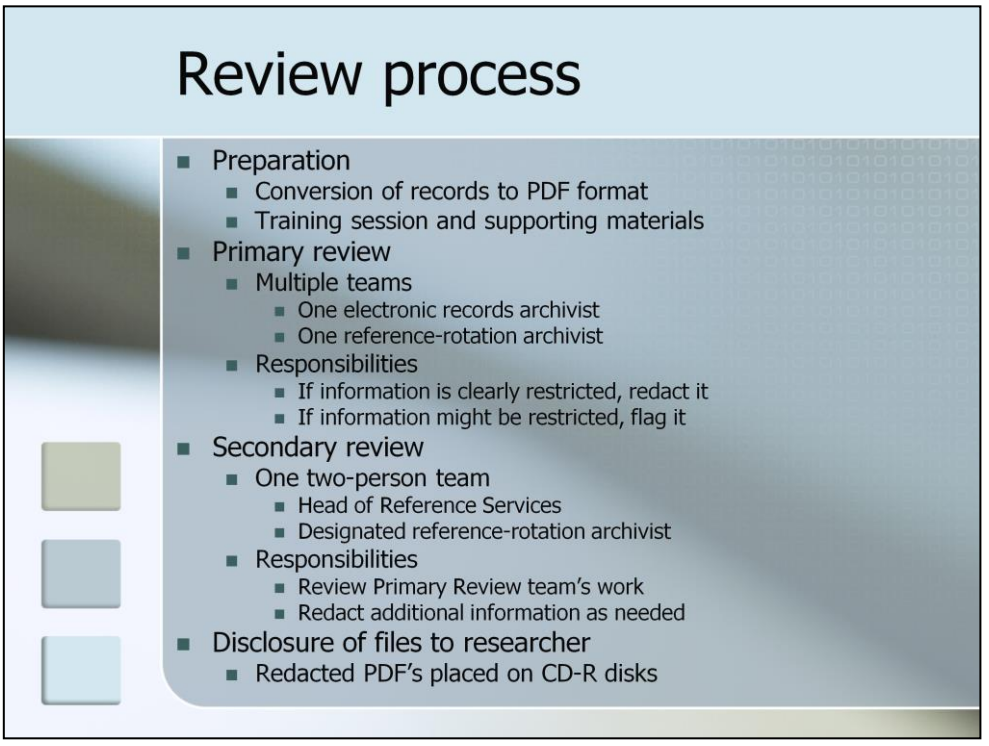


We first thought about printing out everything and using black markers to redact information as needed, but we estimated that we would need at least 10,000 sheets pf paper to print out the 3,000 or so documents the researcher wanted. If we opted for this approach, clerical staff would have to produce and folder printouts, professional staff would have to review and redact the printouts, and clerical staff would then have to photocopy the redacted printouts and folder the resulting copies for the researcher. We would thus need more than 20,000 sheets -- or 40 reams -- of paper. Moreover, New York State's Freedom of Information Law now states that whenever possible, copies of electronic records should be provided in electronic format.

We then began examining electronic redaction options. The records were transferred to us in a format that made redaction and, for that matter, access rather difficult, so we began researching various conversion options. We pretty quickly settled upon Portable Document Format, which seemed like a good choice for this particular set of records: although PDF copies would not have the same metadata as the original documents, all of the elements in which the researcher was interested -- authors, dates of creation -- appeared within the documents themselves. In addition, the researcher, who was familiar with the PDF format, indicated that PDF copies would be acceptable.

Moreover, there are a number of options for redacting PDF files, including a tool that Adobe has built into Acrobat versions 8 and 9. Adobe Acrobat 8 is part of which is part of our parent agency's standard array of software. If its redaction tool worked, we wouldn't need to purchase any specialized software.

We began combing the legal, digital forensics, and other literature for information about the robustness of Adobe's redaction tool and discovered that to date, no one has been able to recover information redacted with it -- or, if they have, they're not telling anyone. We also prepared a small test file, used Acrobat 8's tool to redact some of the information within it, and sent it to two groups of State government digital forensics experts. Neither group could recover any of the information we had redacted.



We then developed a redaction workflow based upon that we've established for redacting paper documents. Our approach works best for larger repositories dealing with large requests. If your repository is small or you need to redact a handful of records, your workflow will be quite different.

During the preparatory stage, staff from the Electronic Records unit converted the files to PDF format and prepared a guide to using the redaction tool, and the head of our Reference Services unit developed a detailed overview of the different types of restricted information that might be present in the records and the statutes governing access to each type of information. Once the records were converted and the training materials were finished, we held a training session for staff.

We then created several two-person Primary Review teams that consisted of one electronic records archivist and one other archivist drawn from our reference rotation; we have only two electronic records archivists on staff, and the two of us were pretty heavily involved in the process. The Primary Review teams were were responsible for reading through a given set of electronic documents and using Acrobat 8's redaction tool to remove information that was clearly legally restricted, and for flagging information that might be restricted.

Once a Primary Review team finished reviewing a set of records, the Secondary Review team, which consisted of the Head of Reference Services unit and one reference-rotation archivist, went to work. This team spot-checked the work of the Primary Review teams, examined flagged information and determined whether it should be redacted, and, in a few particularly thorny cases, consulted with our parent agency's Office of Counsel.

After the Secondary Review team finished working on a given set of records, the Electronic Records unit burned copies of the redacted files onto CD-R discs and the head of Reference Services conveyed them to the researcher.

At every stage of the process, Electronic Records staff was responsible for backing up all work files.

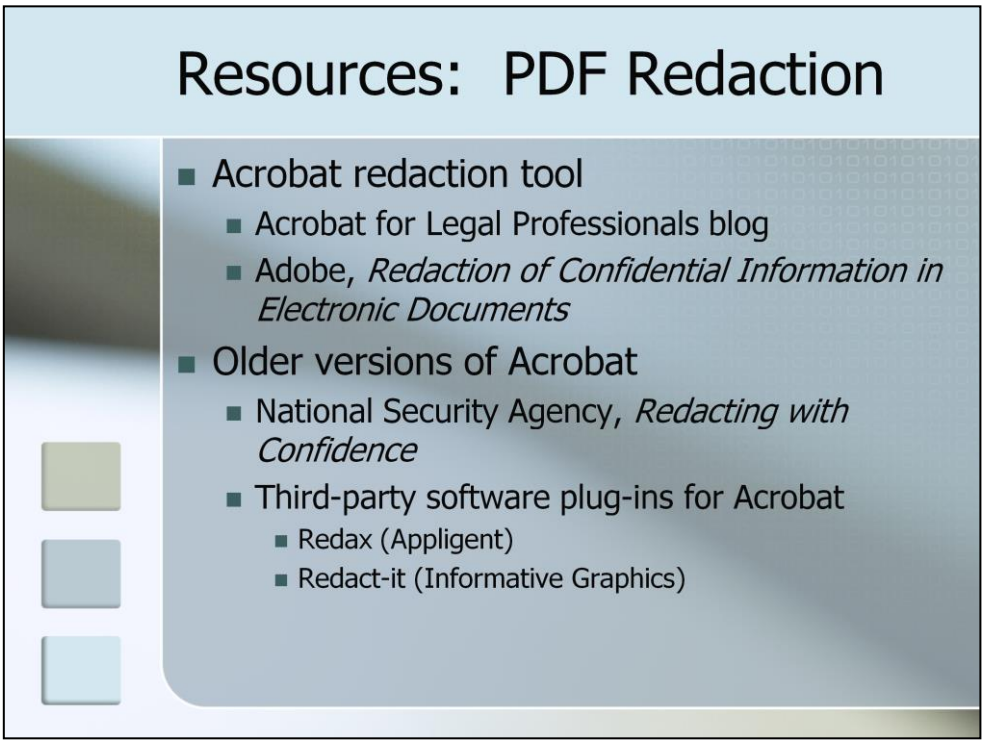
I have to say at that in response to circumstances I really can't discuss, we were forced to modify this workflow midway through our review. However, we did quite a bit of work before we were forced to

switch gears, and we plan to handle future requests the same way we initially handled this one.

What did this experience teach us? First of all, redaction of electronic records is about as labor-intensive as redaction of paper records. Although the burden on clerical staff, who do a lot of photocopying when we redact paper records, is reduced, the burden on professional staff actually increases a bit. It might be possible to delegate to clerical staff routine redactions such as removal of Social Security Numbers, but identifying attorney work product, proprietary information, and other less cut-and-dried forms of restricted information must be done by professional staff. Moreover, professional staff had to devote a lot of time and effort to research and testing, and the Electronic Records unit is now responsible for monitoring the legal, digital forensics, and publicly accessible national security literature relating to electronic redaction.

Second, electronic redaction produces electronic copies that will require at least some care. In addition to the original records, we now have full PDF and redacted PDF copies. We might not keep these copies permanently, but we want and need to retain them for an indefinite amount of time, which means that, at least in the immediate future, the copies require the same level of care as the original files.

Finally, we need better tools. Responding to this request consumed A LOT of professional staff time, and as the volume of electronic records in our holdings increases, demands on staff time are likely to overwhelm our current redaction workflow. We need tools that can do proximity and other types of searches that will reduce, if not eliminate, the need for detailed review of every single record.

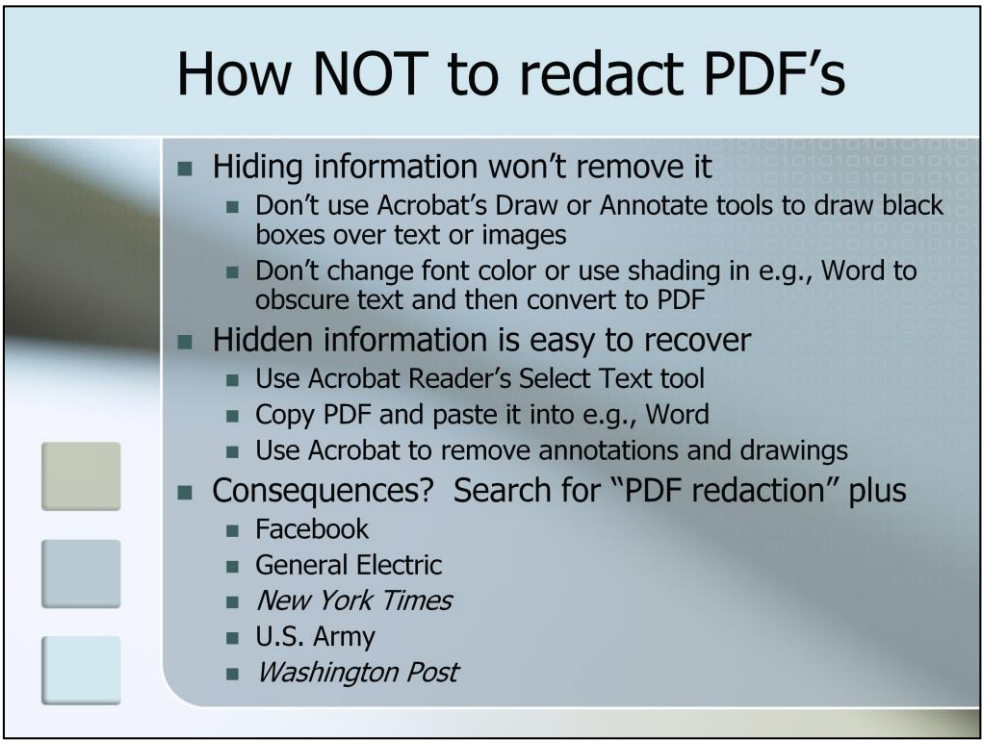


If you need to redact PDF documents, the following resources, all of which are readily available online, will be of interest to you.

Adobe's Web site contains a ton of information about the Acrobat redacting tool, and its Acrobat for Legal Professionals blog provides step-by-step instructions for using it. We found this blog immensely helpful.

If you're using a an older version of Acrobat and don't have the resources or the desire to upgrade, check out the National Security Agency's "Redacting with Confidence" document, which outlines a laborious but effective effective redaction method. You'll find this document on Web sites of the NSA and several other federal agencies.

You may also want to explore third-party software plug-ins. Redax, which is produced by Appligent, and Redact-It, an Informative Graphics product, both have dedicated followings in the legal community and in various federal agencies.



Lastly, if you take nothing else away from this presentation, please remember that there are some BAD PDF redaction techniques that should you should never, ever use.

Do not use Acrobat's Draw or Annotate tools to place black or white boxes over information they wish to redact.

Do not attempt to redact word processing documents by changing the font color to white or using a shading or highlighting feature to obscure the text and then converting the documents to PDF format.

Hiding information in an electronic file is not the same as removing it. In many instances, simply using Acrobat Reader's Select Text tool will reveal its presence. Copying the PDF in its entirety and pasting it into Word or some other application may also reveal hidden information. It's also really easy to pull a PDF back into Adobe Acrobat and get rid of drawings and annotations.

If you want to see what can happen when redaction is done badly, just Google or Bing the words "PDF redaction" plus Facebook, or General Electric, or New York Times, or U.S. Army, or Washington Post. You'll find lots of information about how these organizations and various Fortune 500 firms and federal agencies have learned hard lessons about how not to redact PDF files. This is one club you really don't want to join!

Notes:

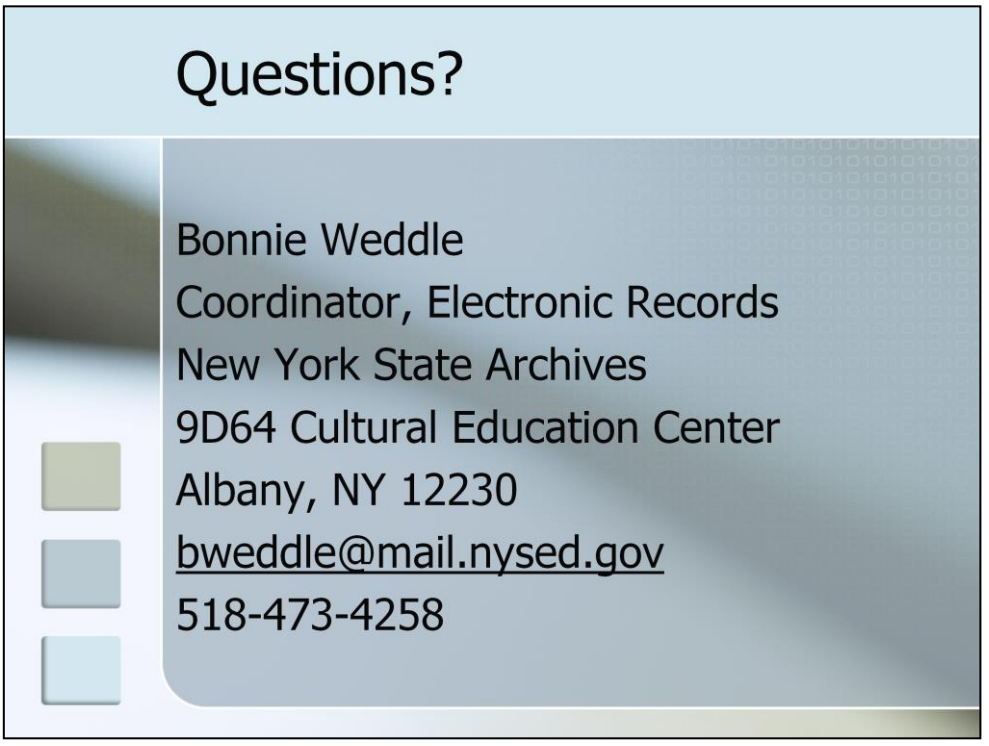
Facebook: in 2009, revealed terms of confidential settlement of a 2009 lawsuit and company's internal valuation.

GE: in 2008, opposing counsel in class-action sex discrimination case released materials documenting inner workings of GE's corporate culture.

NYT: in 2000, publication of a CIA document and ended up revealing identities of several agents.

U.S. Army: revealed classified info about 2005 US killing of Italian intelligence agent in Iraq.

WashPo: 2002 letter written by Washington sniper revealed bank account and other information



If you have any questions about Web crawling or redaction of PDF files, please feel free to ask me at the end of this session, track me down while we're here at MARAC, or get in touch with me. Thank you!