Restoring Access to ERIC's PDFs

Erin Pollard
ERIC Project Officer
US Department of Education

Speaking Notes: Good afternoon,

Welcome to the ERIC webinar on "Restoring Access to ERIC's PDFs". I am Erin Pollard, the Project Officer for ERIC. I am going to be speaking to you today about the PDF restoration process, including why the decision was made and how we restored the documents.

Overview

- Policy decisions and background
- Process for clearing the documents
- Next steps to clear the remaining documents
- Questions

Speaking Notes: During the registration process, we asked users for their questions. We have tried to answer as many questions as possible in this presentation. If your question is not answered, please submit it through the text box on your screen. I am going to spend the first half of the webinar explaining the policy decisions that we made and why we made them. Then I will stop for a few questions and will continue to describe the process of how we actually cleared the documents. I will stop for questions again and then we will finish up by talking about next steps. We will then answer as many questions as possible at the end of the presentation.

Some of the questions that were submitted were not directly related to the PDF restoration process. I will be happy to answer them at the end if time allows.

Goals for Today

- How did we get into this situation?
- Why did ERIC take down the PDFs?
- Was there a risk?
- What process did ERIC use to restore the PDFs?
- Why did it take almost two years for ERIC to restore the documents?
- Was it a concern that the information was still available via microfiche?
- What are the next steps?

Speaking Notes: As background, starting August 3rd, 2012 the full-text of all ERIC PDFs was temporarily disabled due to the discovery of personally identifiable information in some documents. This was far from the ideal situation—this was the beginning of the semester and the start of ERIC's highest use period. However, it was the only option ERIC had and we hope to explain our reasoning and how we solved the problem.

Today we will be addressing the questions on the slide

What is PII?

- In OMB M-06-19 (July 12, 2006), "the term Personally Identifiable Information means any information about an individual maintained by an agency, including, but not limited to, education, financial transactions, medical history, and criminal or employment history and information which can be used to distinguish or trace an individual's identity, such as their name, **social security number, date and place of birth, mother's maiden name, biometric records, etc**., including any other personal information which is linked or linkable to an individual."

Speaking Notes: First, to begin, I want to clarify some of the terms that I will be using. The government defines the term "personally identifiable information", or PII, to information which can be used to distinguish or trace an individual's identity, such as their name, social security number, date and place of birth, mother's maiden name, biometric records, etc. It is the government's responsibility to preserve individual's privacy regarding PII. While an author's name or office phone number is fine to include on a government website, their social security number is problematic.

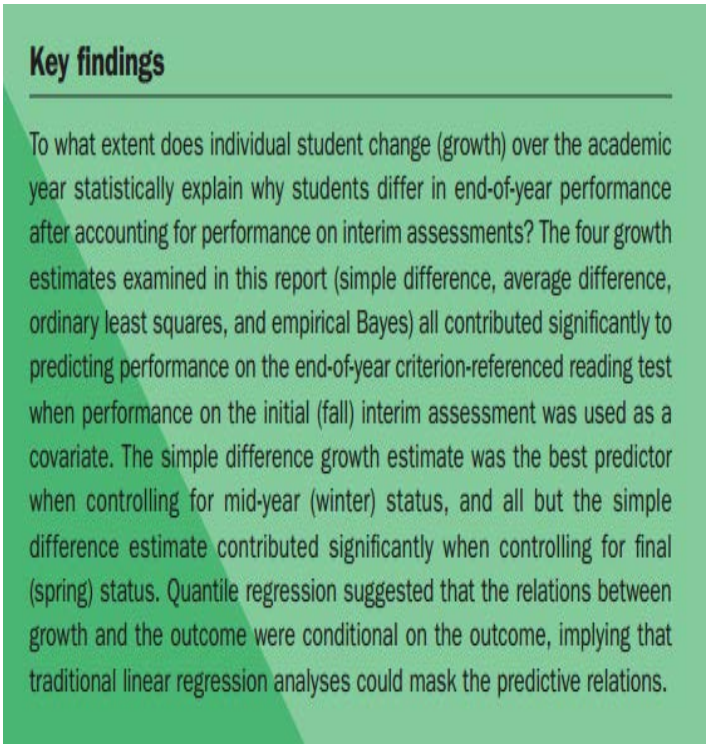
History

- 1960s: ERIC was a purely microfiche and paper system
- 2000s: ERIC converts microfiche to PDF
- 2010s: ERIC converts microfiche to readable PDFs

Speaking Notes: When ERIC was created almost 50 years ago, it was primarily a microfiche-based system. When ERIC went online, the microfiche was eventually converted to PDFs to make it easy for users to access documents. However, the technology for PDFs 10+ years ago was different than it is today and the PDFs were made as images. This means that they were not machine readable, which enables searching and is ideal for individuals with disabilities.

OCRing of a Crisp Document

PDF image (below)
Alternate text (right)

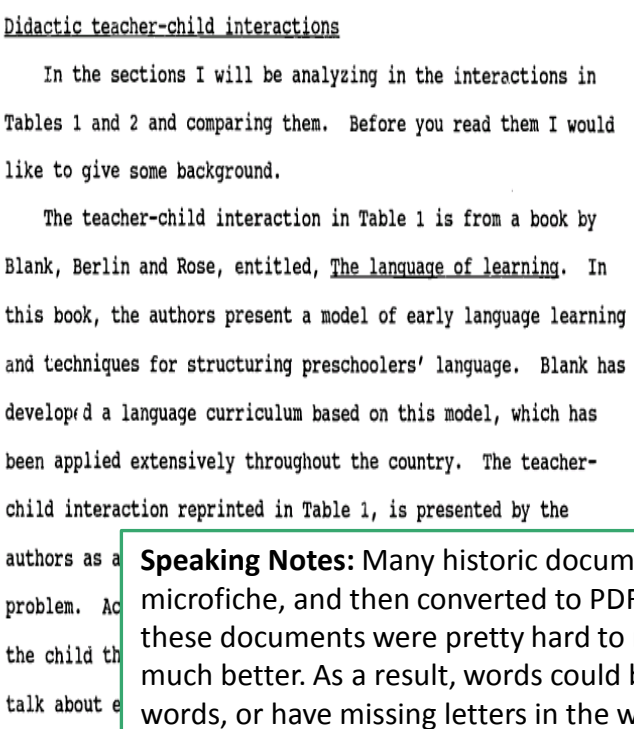


Key findings
To what extent does individual student change (growth) over the academic
year statistically explain why students differ in end-of-year performance
after accounting for performance on interim assessments? The four growth
estimates examined in this report (simple difference, average difference,
ordinary least squares, and empirical Bayes) all contributed significantly to
predicting performance on the end-of-year criterion-referenced reading test
when performance on the initial (fall) interim assessment was used as a
covariate. The simple difference growth estimate was the best predictor
when controlling for mid-year (winter) status, and all but the simple
difference estimate contributed significantly when controlling for final
(spring) status. Quantile regression suggested that the relations between
growth and the outcome were conditional on the outcome, implying that
traditional linear regression analyses could mask the predictive relations.

Speaking Notes: To make the documents searchable, ERIC used Optical Character Recognition software to scan the PDFs. This process takes the existing image and then assigns a letter to it to create a layer of readable text that a user can search, copy, and paste. This works really well for a crisp document, like one that you would type in Microsoft Word and then print on a laser printer, but many historic PDFs were not crisp.

OCRing of a Non-Crisp Document

PDF image (below)
Alternate text (right)



Didactic teacher-child interactions

In the sections I will be analyzing in the interactions in Tables 1 and 2 and comparing them. Before you read them I would like to give some background.

The teacher-child interaction in Table 1 is from a book by Blank, Berlin and Rose, entitled, The language of learning. In this book, the authors present a model of early language learning and techniques for structuring preschoolers' language. Blank has developed a language curriculum based on this model, which has been applied extensively throughout the country. The teacher-child interaction reprinted in Table 1, is presented by the authors as a

problem. Ac

the child th

talk about e

Didactic teacher-child interactions
In the sections I will be analyzing in the interactions inTables 1 and 2 and comparing them.Before you read them I wouldlike to give some background.
The teacher-child interaction in Table 1 is from a book byBlank, Berlin and Rose, entitled, The language of learning.Inthis book, the authors present a model of early language learningand techniques for structuring preschoolers' language.Blank hasdevelopEd a language curriculum based on this model, which hasbeen applied extensively throughout the country.The teacher-child interaction reprinted in Table 1, is presented by theauthors as an example

Speaking Notes: Many historic documents were often typed on a typewriter, converted to microfiche, and then converted to PDF. As many long-time ERIC users have seen, many of these documents were pretty hard to read with a human eye. Machines could not do much better. As a result, words could be misspelled, broken down into two or three words, or have missing letters in the word.

One of the reasons that we wanted these documents to have a readable layer is that it would enable users to search the full text. That way users would be more likely to find the information that they are looking for, even when it isn't in the abstract of the document. However, before ERIC could make the full text of our documents searchable, we needed to make sure that the background text of the document was correct.

To be clear, it wasn't that we OCRed the documents wrong, it is just very difficult to make a readable PDF from a document that was written typewriter, converted to microfiche, and then converted to an imaged-based PDF. We were in the process of addressing the OCR quality issue and finding a good way to make the documents searchable before the concerns about PII came to light.

Why did ERIC take down the PDFs?

- In July 2012 ERIC allowed commercial search engines to search ERIC's full text.
  - ERIC did not have the ability to make its PDFs searchable in house
  - Most of ERIC's users come from a commercial search engine, so this would improve their ability to find information
  - This would greatly improve the usability of ERIC

Speaking Notes: Around the same time that ERIC was having these conversations internally, ERIC was approached by commercial search engines who wanted to crawl ERIC's full text collection. Because they have different OCR technology, it was possible (and highly likely) that they would be able to OCR our collection far more accurately than we had been able to. And they would do this at no cost to us or taxpayers.

Because the vast majority of our users come from commercial search engines, being able to the search the full text for specific words and phrases would be incredibly beneficial for users. We thought that this would be an easy win for users and made the decision to let the crawling begin. We were excited to message this as a huge benefit for ERIC users.

Why did ERIC take down the PDFs?

- ERIC is highly ranked in search results
- Once commercial search engines indexed ERIC's full text, people could Google themselves. If their names appeared in ERIC, it would be one of the first results shown

Speaking Notes: Within a week of that decision, through a routine search, a user found that his highly sensitive personally identifiable information was showing up when he searched his name in Google. This was a problem. We absolutely did not want to have this type of information online. We made the call that the documents must be taken down immediately. The risk to individuals was too great—if it was your information or your child's information, I am sure you would feel the same way. However, as users, we also understand the frustration of not having the documents online. It was far from ideal timing.

One question you may be wondering is why we did not reverse the permission we gave to commercial search engines– the reason was that once we realized we had the problem, we needed to solve it. We wanted to make sure that commercial search engines did not have access to this information either.

Once the documents were taken down—we had to figure out how to get them back online. Because this was not a planned shut down, we needed to start from scratch to figure out how to solve the problem.

Was there really a risk?

- Types of places PII was found:
  - Graduate theses
  - Grant reports
  - Forms
  - Resumes
- Some PII were examples:
  - John Doe, 123-45-6789
  - Mary Smith, 111-11-1111
- Many (600+) were real

Speaking Notes: This raises the question, why would someone's social security number or other highly sensitive PII be in ERIC? One common example would be a graduate thesis. Students are often required to put their student ID number on the cover sheet of their thesis, normally directly below their name. Prior to the year 2000, student ID numbers were often their social security numbers. This would mean when a person Googled themselves (or someone else), their social security number would pop up right next to their name.

Similarly, the same type of PII was often included on resumes and invoices, and could even be a grant number. People simply thought about privacy differently in the past. Because ERIC has records from over 50 years ago, we needed to make sure that we protected individual's privacy in a responsible way.

Approach

- Clear any document classes that we felt posed no risk
- Manually clear as many documents as possible
- Hire a contractor to search as many documents as possible

Speaking Notes: We took a three step approach:
Clear any document classes that we felt posed no risk
Manually clear as many documents as possible
Hire a contractor to search as many documents as possible

Clearing Document Classes

- All peer-reviewed articles (September 2012)
- ERIC and Peace Corps Documents (October 2012)
- Any article published after 2005 (March 2012)
- Documents that have been scanned for PII either manually or through an automated process (July 2014)

Speaking Notes: In terms of clearing document classes—we had no idea what types of documents would be "safe" and which posed risk. We needed to go through several documents to get a sense of which may be okay to clear without searching. The first class of documents that we were able to clear were peer reviewed articles—we determined that it was highly unlikely for this type of information to be in these types of articles and that if it occurred, the information would have likely been removed in the peer review process. We then were able to clear the Peace Corps language learning documents based on the subject matter.

Finally, we were able to clear documents released after 2005. By that point, SSNs were no longer used as student ID numbers and were not on resumes or grant reports. The risk for these documents was severely reduced.

User Requests for Documents

- Over 10,000 emails in the first month
- Few popular requests– most documents were only requested by one or two people
- All requests were scanned in the order that they were received

Speaking Notes: During this period, we were working to clear as many documents as possible and we asked users to send us an email with the high priority documents that they wanted cleared. We got a huge response—over 10,000 emails in the first month alone. What was interesting about these requests is that they were mostly unique requests. We got very few documents requested by more than one individual.

We made the decision to scan and release documents in order of request. This had the advantage of giving an undergraduate the same priority as his congressman and making it entirely fair. However, this wasn't popular because users were not getting the documents that they wanted quick enough to be useful for their project.

The reason for the delay is that document had to be manually searched and cleared by a government employee (not a contractor) due to security concerns. The federal ERIC staff consisted of only me. That meant I had to answer user request emails, clear the documents, as well as finding a more sustainable process on top of my other duties. I was trying to clear documents as quickly as possible, but the job was overwhelming. We were able to get some help from the National Library of Education librarians, but our ultimate goal was to hire a contractor with the appropriate clearances to do the work. We knew that this would ultimately be the best solution to release the documents quickly.

Federal Contracting

- It is government policy that unless the work is "inherently governmental", it should be contracted out.
- Federal contracts take 6-18 months to plan
- Funds expire on September 30th
- Contracts require the government to be very specific in what they want the contractor to do– which requires the government to know how to solve the problem

Speaking Notes: Hiring a contractor is typical for government work—almost all federal programs are contracted out (including ERIC). However, from a timing perspective, this was the worst possible time for this to occur. The end of the federal fiscal year is September 30th. This means that by August most of the government's budget has already been planned for. There was not money set aside for this project and we would not get additional funds until October.

The second wrinkle was timing—it normally takes 3-6 months to get a federal contract signed and you have to plan for new contracts at least a year in advance. We had less than 60 days until the end of the fiscal year. If we took the normal approach for hiring a firm to do this work, we would not have a firm hired until March or April of 2013. That was simply too late for our users—we wanted to get these online as soon as possible.

Finally, we had the challenge of figuring how to write up the scope of work and what we wanted the contractors to do. We knew that the OCRing of these documents wasn't perfect, so just writing a computer code to scan the documents would not work. We had to make sure that the OCRing was good enough that if we searched for PII within a document, and there was PII in the document, we would be able to find it.

However, we were not experts in doing this type of work and government contracting requires the government to put lots of information about exactly what we want in contracts. This ensures that we get the best value for the government– the best work for the lowest price. It also requires that we know what type of solution that we want from a contractor, which is something that takes time to consider.

Original solution

- Short term from October 2012-June 2012
- Scan 1000 documents a week for the quality of the OCR and for PII
- To continue with this approach, it would take over 7 years to clear all of the documents
- Revised plan: 70% lower cost and documents released in 2 years, but had a 5 month hiatus of releasing new documents

Speaking Notes: We were able to work around the clock and come up a creative solution for a short term contract to do this work. For anyone who has worked in contracting before, getting a federal contract in less than 2 months from inception to award is almost impossible. We pulled every solution at our disposal to get the documents back online.

We were underway by November of that year. We were able to scan about 1,000 documents a week. It is important to note that this was not clearing 1,000 documents a week—at least 10% of our documents did not have good enough OCRing to be able to check them. It also did not include the documents that had PII in them.

Scanning 1,000 documents a week was a lot of documents, but we had so many more requested. At the pace we were going, it would have taken about 6 years to get them cleared. We realized that 6 years was not going to be a sustainable plan and regrouped to find a way to get them cleared within 2 years of the original incident. This took additional resources and a hiatus of time to get a contract in place, but overall it allowed us to get the documents back online far quicker than before.

One question that was asked in advance of the webinar was about why we didn't notify people that we had fulfilled their request. There were two reasons for this– first, there was an element of risk. If you requested 3 documents at the same time, and 2 were put back online, that would tell you that the third might have PII in it. It could also mean that it was unreadable or locked, so that wasn't a huge risk.

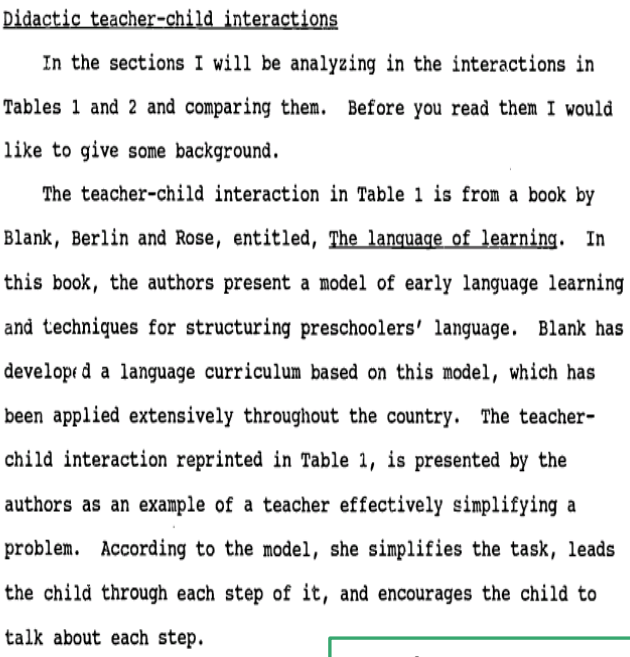
The bigger concern was the cost associated with doing this. To notify individuals would mean that we would have to pay someone to contact every person who requested a document when that document was released. We would have also had to set up a system where we kept people's names and email addresses, which requires approval that takes several months. Consequentially, we would have to significantly slow the PDF scanning and release process. We did not feel like that was a prudent release of resources.

Questions?

Speaking Notes: I want to stop for questions at this point. Are there any questions about what I have just talked about in terms of the policy setting/decision making aspect before I go into detail about how we cleared the documents? (None were asked)

Creating a Readable Layer

PDF image (below)
Alternate text (right)



Didactic teacher-child interactions

In the sections I will be analyzing in the interactions in Tables 1 and 2 and comparing them. Before you read them I would like to give some background.

The teacher-child interaction in Table 1 is from a book by Blank, Berlin and Rose, entitled, The language of learning. In this book, the authors present a model of early language learning and techniques for structuring preschoolers' language. Blank has developed a language curriculum based on this model, which has been applied extensively throughout the country. The teacher-child interaction reprinted in Table 1, is presented by the authors as an example of a teacher effectively simplifying a problem. According to the model, she simplifies the task, leads the child through each step of it, and encourages the child to talk about each step.

Didactic teacher-child interactions
In the sections I will be analyzing in the interactions inTables 1 and 2 and comparing them.Before you read them I wouldlike to give some background.
The teacher-child interaction in Table 1 is from a book byBlank, Berlin and Rose, entitled, The language of learning.Inthis book, the authors present a model of early language learningand techniques for structuring preschoolers' language.Blank hasdevelopEd a language curriculum based on this model, which hasbeen applied extensively throughout the country.The teacher-child interaction reprinted in Table 1, is presented by theauthors as an example of a teacher effectively simplifying aproblem.According to the model, she simplifies the task, leadsthe child through each step of it, and encourages the child totalk about each step.

Speaking Notes: So now I am going to switch gears and talk about how we went about clearing the process. This is really technical, but also really interesting to a non-technical audience. The first thing I am going to cover is how we determined whether or not a document was readable.

When you take a PDF and OCR it, the technology tries to figure out what word should be associated with the image. We call the underlying text the readable layer. If you open a PDF and copy a sentence or two and paste it into another document, you are copying from the readable layer. So if you have the word “information”, the goal is for the OCRing to have the readable layer read “information”. This raises the question--how do you check if this is accurate? Especially considering that ERIC’s full text collection is vast (estimated to be over 6 miles high if it was a printed stack)?

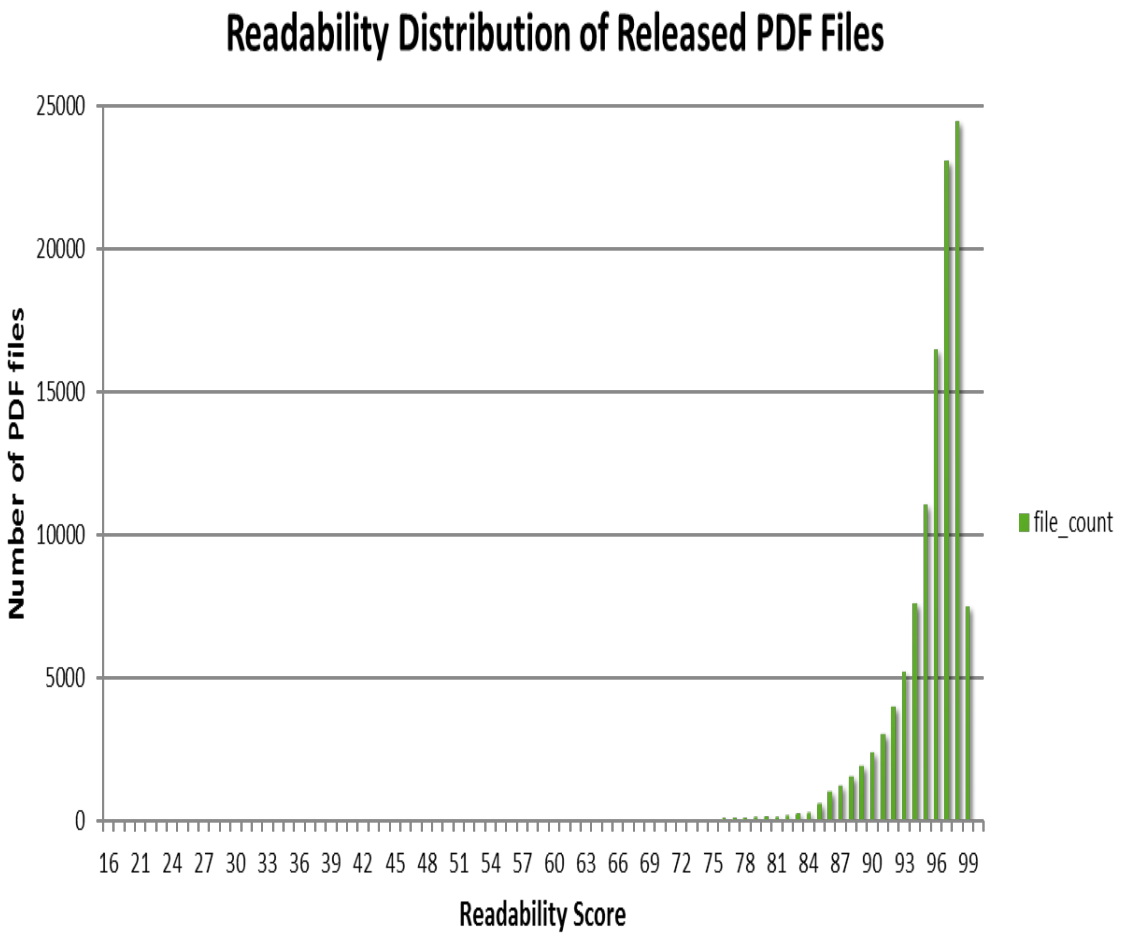
Quality Control for the Readable Layer

- Were 85% of the words spelled correctly?
- “Information” vs. “in form at ion”

Speaking Notes: As far as a timeline, I wish I had an exact timeline, but we are still working through the options and trying to prioritize resources. We want the documents all online as quickly as possible, but we also have other exciting things that we want to work on and we want to find the right resource balance so that we deliver the most benefit to users.

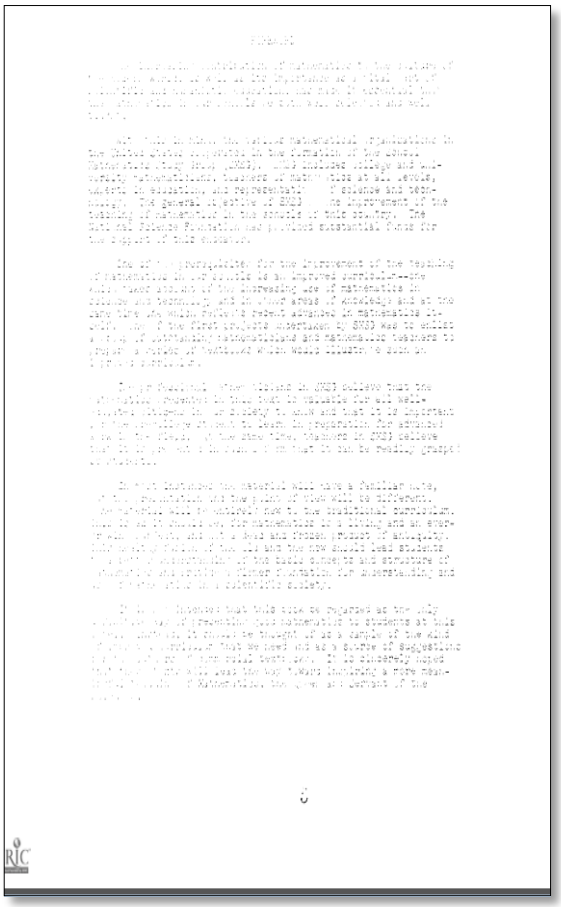
Because we pay by the page to get these restored, we are trying to balance our priorities. Is it better to prioritize shorter documents to get a more documents back online? Should we prioritize older and rarer documents that are harder to find? If you have ideas about which documents we should prioritize to restore or a fair way to do this, please let us know. We are open to your ideas!

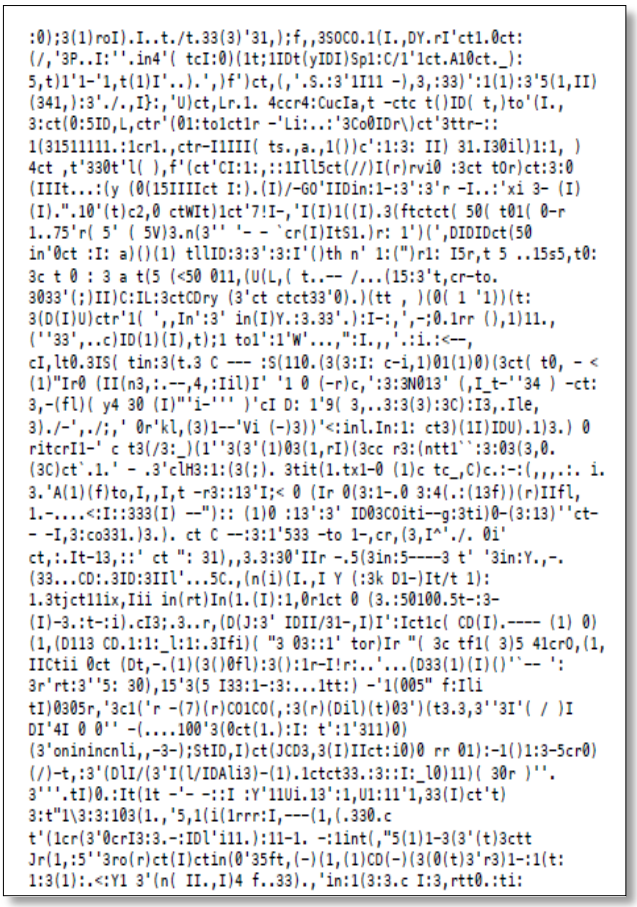
Quality of the Readable Layer



Speaking Notes: While the 85% number was initially determined in a non-scientific way, the number ultimately appeared to be quite a good threshold for quality. That was the tipping point for where a document was fairly good to where there were major problems in readability. This also allowed us to release as many documents as possible, while minimizing risk to users.

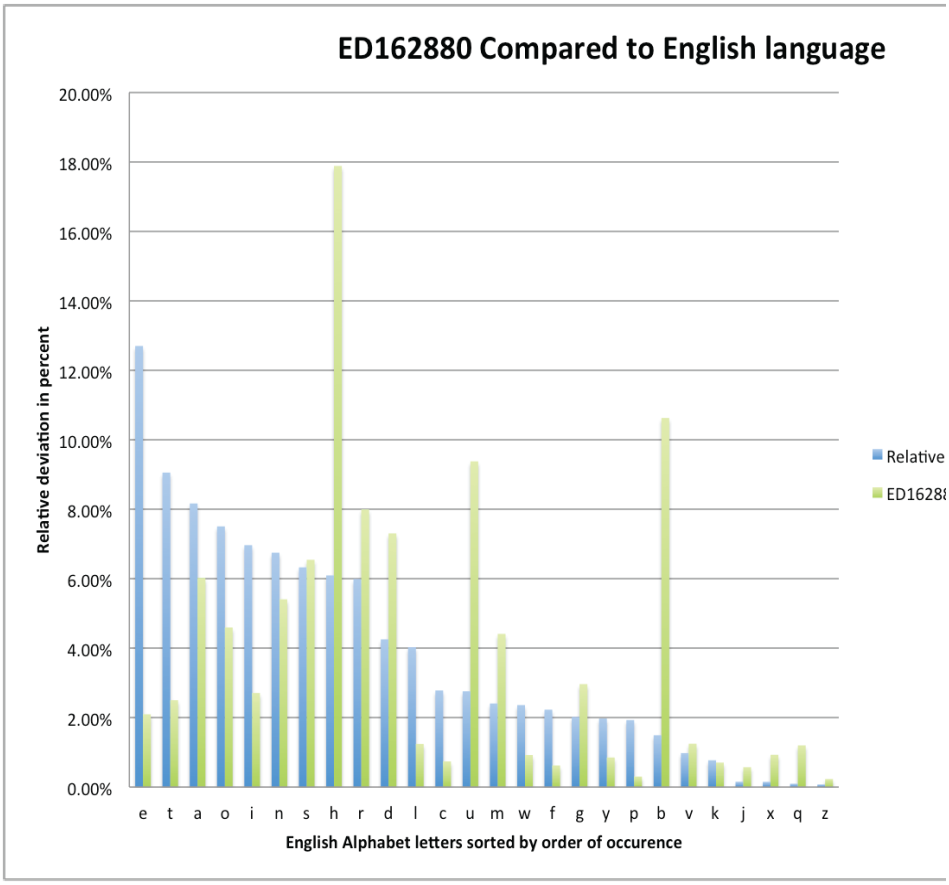
Problems with using spelling…





Speaking Notes: However, when writing the code for the software, we told the software to ignore mathematical equations. This is because they are in Greek and we only did the readability work on English language documents. We realized that some of the documents that had high readability scores had a readable layer that looked like a mathematical equation to our software, so it was ignored, when really it was just a very weak copy of text. We needed to find a way to weed out these documents.

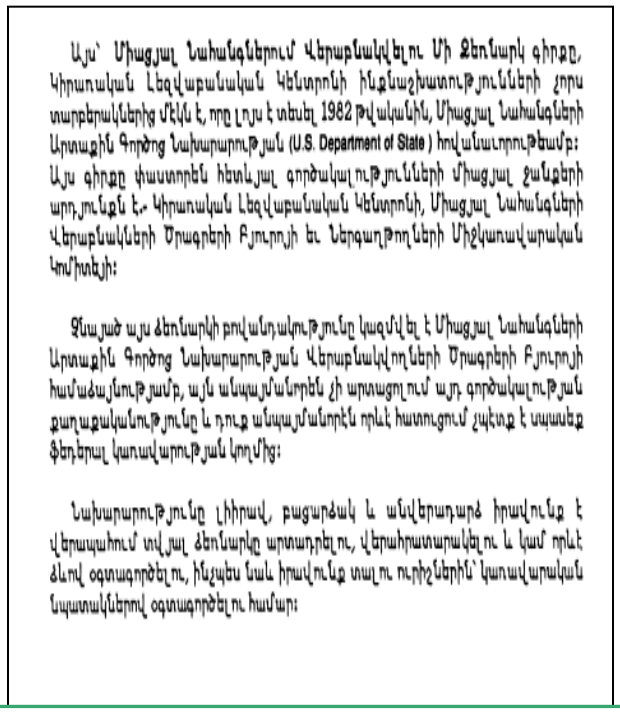
Distribution of Letters Approach



Speaking Notes: The way that we approached solving this problem was checking the distribution of letters in the document, compared to the distribution of letters in the English language. The thought was, in English there is a pretty known distribution of which letters are most common. Es, Ts, and As are all common, while Xs, Qs, and Zs are not common. While it would be possible for a paper to have a disproportionally high proportion of these letters, in general, over the course of a several page paper, the letters will generally fall into this pattern.

What we did was compared the distribution of letters in the document to the known distribution of letters in the English language. If the deviation from the expected value was too great and the document was in English, we then declared it as “not readable”. This meant that it needed to be re-scanned and re-OCRed prior to being released on the ERIC website.

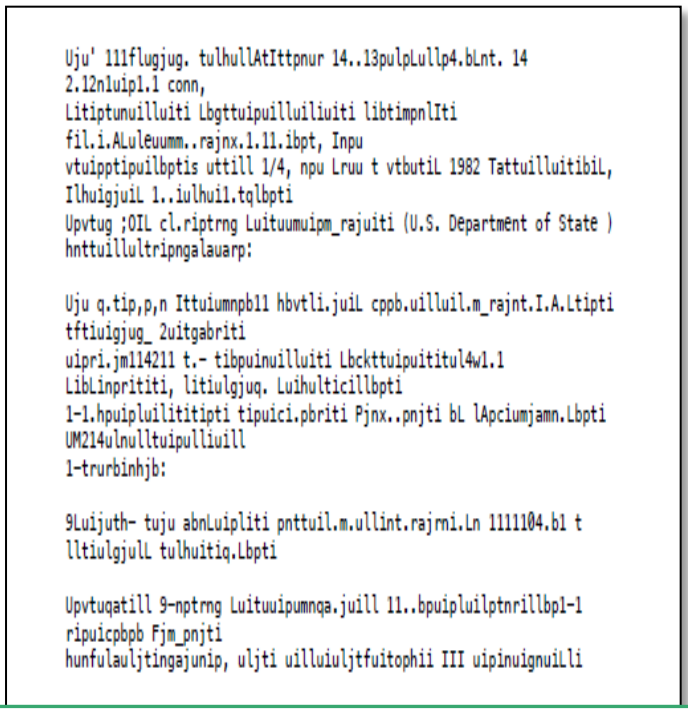
Foreign language documents?



Այս՝ Միացյալ Նահանգներում Վերաբնակվելու Մի Ձեռնարկ գիրքը, Կիրառական Լեզվաբանական Կենտրոնի ինքնաշխատությունների չորս տարբերակներից մէկն է, որը լոյս է տեսել 1982 թվականին, Միացյալ Նահանգների Արտաքին Գործոց Նախարարության (U.S. Department of State) հովանաւորութեամբ: Այս գիրքը փաստորեն հետևյալ գործակալությունների միացյալ ջանքերի արդյունքն է.- Կիրառական Լեզվաբանական Կենտրոնի, Միացյալ Նահանգների Վերաբնակների Ծրագրերի Բյուրոյի եւ Ներգաղթողների Միջկառավարական Կոմիտէի:

Ջնայած այս ձեռնարկի բովանդակությունը կազմվել է Միացյալ Նահանգների Արտաքին Գործոց Նախարարության Վերաբնակվողների Ծրագրերի Բյուրոյի համաձայնությամբ, այն անպայմանորեն չի արտացոլում այդ գործակալության քաղաքականությունը և դուք անպայմանորէն որևէ հատուցում չպէտք է սպասեք ֆեդերալ կառավարության կողմից:

Նախարարությունը լիիրավ, բացարձակ և անվերադարձ իրավունք է վերապահում տվյալ ձեռնարկը արտադրելու, վերաիրատարակելու և կամ որևէ ձևով օգտագործելու, ինչպես նաև իրավունք տալու ուրիշներին՝ կառավարական նպատակներով օգտագործելու համար:



Uju' 111flugjug. tulhullAtIttpnur 14..13pulpLullp4.bLnt. 14
2.12nluip1.1 conn,
Litiptunuilluiti Lbgttuipuilluiliuiti libtimpnlIti
fil.i.ALuleuumm..rajnx.1.11.ibpt, Inpu
vtuipptipuilbptis uttill 1/4, npu Lruu t vtbutiL 1982 Tattuilluitibil,
Ilhulgjuil 1..iulhui1.tqlbpti
Upvtug ;OIL cl.riptrng Luituumuipm_rajuiti (U.S. Department of State)
hnttuillultripngalauarp:

Uju q.tip,p,n Ittuiumnpb11 hbvtli.juiL cppb.uilluil.m_rajnt.I.A.Ltipti
tftiuigjug_ 2uitgabriti
uipri.jm114211 t.- tibpuinuilluiti Lbckttuipuititul4w1.1
LibLinprititi, litiulgjuq. Luihulticillbpti
1-1.hpuipluilititipti tipuici.pbriti Pjnx..pnjti bL lApciumjamn.Lbpti
UM214ulnulltuipulliuill
1-trurbinhjb:

9Luijuth- tuju abnLuipliti pnttuil.m.ullint.rajrni.Ln 1111104.b1 t
lltiulgjuiL tulhuitiq.Lbpti

Upvtuqatill 9-nptrng Luituuipumnqa.juill 11..bpuipluilptnrillbp1-1
ripuicpbpb Fjm_pnjti
hunfulauljtingajunip, uljti uilluiuljtfuitophii III uipinuignuiLli

Speaking Notes: While ERIC only collects documents in English today, some of our legacy documents are in foreign languages. If the document is written in a language that uses the Latin alphabet, this isn't a problem. The OCR technology will work correctly on documents in Spanish, French, and similar languages. But when the document is written on a language that isn't based on the Latin alphabet, such as Greek or Arabic, the OCR technology matches the Latin letter closest to what it sees, and it ends up with a readable layer that is not accurate.

If a document was written in a language other than English that used the Latin alphabet, we used a foreign language dictionary to do the spell check approach. This enabled us to make sure that the OCRing was correct.

This left us with the question, what do we do with documents not in the Latin alphabet? We need them to be OCRed to be put online due to federal IT regulations and doing OCRing on non-Latin letters is very expensive. For now, we are not going to be re-releasing these documents. If a specific request comes up in the future, we will consider if we have the resources to make that document accessible.

Checking for PII

- XXX-XX-XXXX
- XXXXXXXXX
- XXXXX-XXXX
- XXX-XXXXXX

Speaking Notes: Once we had determined that a document was readable, we could begin to check it for PII. This involved us creating a computer code to search the document for a 9 digit string of numbers, with or without hyphens. If a nine digit number was found, it would be flagged for human verification. Many of these were false positives, such as a dollar value or a zip code, but there were genuine positives and questionable results.

Each result was checked by two humans. If there was disagreement, I would look at the document to make the determination if it was an actual risk.

We allowed clear examples to stay, but rooted out anything which appeared to be an actual instance of PII. We used the Social Security Administration's convention system to verify if the number could be valid, as well as the region codes. There were over 650 cases of actual PII and at least 100 were close calls.

Questions?

Speaking Notes: Do we have any questions about the process and how we scanned documents?
(None were asked)

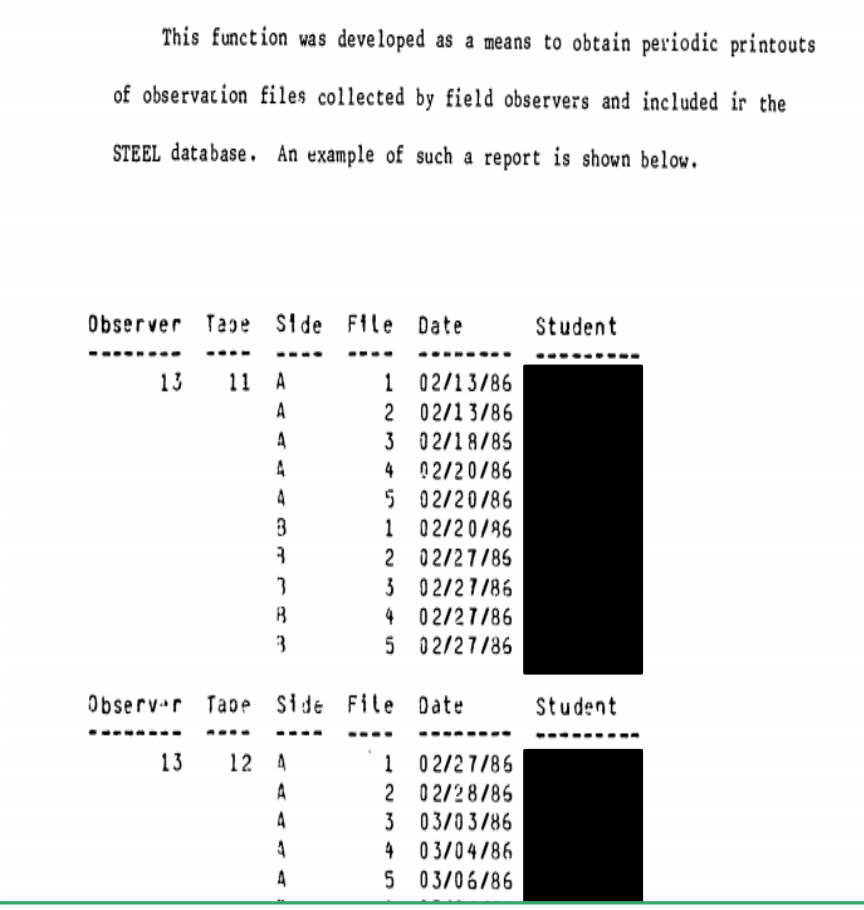
What has been cleared?

- Cleared:
  - All peer-reviewed documents
  - All documents published since 2005
  - All readable PDFs without PII
- Not cleared:
  - Documents with PII
  - Documents that are not readable

Speaking Notes: As far as a timeline, I wish I had an exact timeline, but we are still working through the options and trying to prioritize resources. We want the documents all online as quickly as possible, but we also have other exciting things that we want to work on and we want to find the right resource balance so that we deliver the most benefit to users.

Because we pay by the page to get these restored, we are trying to balance our priorities. Is it better to prioritize shorter documents to get a more documents back online? Should we prioritize older and rarer documents that are harder to find? If you have ideas about which documents we should prioritize to restore or a fair way to do this, please let us know. We are open to your ideas!

PDF Restoration Process



This function was developed as a means to obtain periodic printouts of observation files collected by field observers and included ir the STEEL database. An example of such a report is shown below.

Observer	Tape	Side	File	Date	Student
13	11	A	1	02/13/86	[redacted]
		A	2	02/13/86	
		A	3	02/18/86	
		A	4	02/20/86	
		A	5	02/20/86	
		B	1	02/20/86	
		B	2	02/27/86	
		B	3	02/27/86	
		B	4	02/27/86	
		B	5	02/27/86	

Observer	Tape	Side	File	Date	Student
13	12	A	1	02/27/86	[redacted]
		A	2	02/28/86	
		A	3	03/03/86	
		A	4	03/04/86	
		A	5	03/06/86	

Speaking Notes: One other question I have been asked is what are we going to do with the 650 documents with PII, which is about 0.2% of the collection. We have to figure out how to deal with them in a way that does not involve risk.

To try to answer a question that I know will be asked, because the microfiche is still out there we cannot simply black line the documents. If a user found a document that is black lined, like the one on the screen, they could go to the microfiche, to that page, and get that document. They could then use that information for identity fraud. Therefore, this isn't the best solution.

We are still exploring other options, but for now the PDFs are not available online and they probably won't be online anytime soon. We will be removing the "PDF pending" identifier for them, but their citations will remain in the collection.

Is there a risk of having the PII in the microfiche?

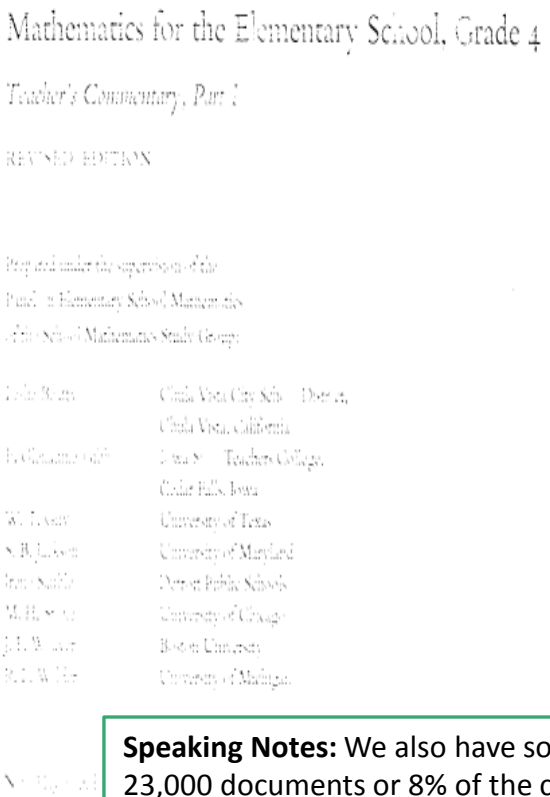
- The types of documents with PII are not the prime ERIC documents
- Microfiche must be searched by hand, lowering the risk
- Much of the PII is very hard for the human eye to find

Speaking Notes: One question that we got a lot during this period was "what about the microfiche? Is there a risk there?" Our answer to this is complicated. We made the decision to recommend institutions to keep their microfiche. We did revise the weeding "keep" list, or the list of microfiche you should retain because the PDFs are not available online. The list now reflects the current status of the online collection. We added the documents to the keep list which we could not scan because the OCRing was too poor and the documents with PII. We will update the lists annually as the documents change.

The PII is certainly in the microfiche. However, the kind of documents where the PII is found in are not exactly highly used research material. Few ERIC users would be looking for graduate theses from the 1970s or grant reports from the 1980s. These are not the prime ERIC documents that we see high usage from.

The second aspect is that it is really hard for the human eye to find PII in these documents. Some of the documents are over 10,000 pages long—the odds of someone finding the PII on page 8735 in size 4 font is highly unlikely. Even as I was examining documents when I knew what page it was on, it was often very hard to find.

PDF Restoration Process



Mathematics for the Elementary School, Grade 4

Teacher's Commentary, Part 1

REVISED EDITION

Prepared under the supervision of the
Panel on Elementary School Mathematics
of the School Mathematics Study Group

Chula Vista City Schools, Chula Vista, California
Teachers College, Cedar Falls, Iowa
University of Texas
University of Maryland
New York Public Schools
University of Chicago
Boston University
University of Michigan

Speaking Notes: We also have some PDFs (approximately 23,000 documents or 8% of the collection) that are just plain not readable. This is actually one of our more readable copy of the unreadable list. It is a horrible copy from microfiche and we are fairly positive we can get a better copy. What we are doing with these is not releasing them because they are not useful or readable– to either humans or machines.

We do not believe that there is PII in these documents, but you can't read them and they are not useful to our users. We have marked them as "pending restoration" and then will be going through them one by one to re-scan them and make them into fully searchable PDFs. This will be a long and expensive process, but it will ultimately be better for our users.

Timeline?

- It is hard to determine because this is really expensive to do
- We want your ideas--what are the best ways to prioritize the restoration?

Speaking Notes: As far as a timeline, I wish I had an exact timeline, but we are still working through the options and trying to prioritize resources. We want the documents all online as quickly as possible, but we also have other exciting things that we want to work on and we want to find the right resource balance so that we deliver the most benefit to users.

Because we pay by the page to get these restored, we are trying to balance our priorities. Is it better to prioritize shorter documents to get a more documents back online? Should we prioritize older and rarer documents that are harder to find? If you have ideas about which documents we should prioritize to restore or a fair way to do this, please let us know. We are open to your ideas!

Questions?

- Q: I never heard about the ERIC "keep" list. Where can I find it?
- A: The keep list can be found in the FAQs: http://eric.ed.gov/?faq

I am looking to weed my microfiche collection. Which ERIC records should I keep?

The documents which are currently available on microfiche, but not online, can be found here. This list will be updated annually to reflect any new documents made available online through the PDF restoration process.

Questions?

- Q: When did the PDF restoration project end and how many of the previously restricted documents are available now?
- A: The PDF restoration process is ongoing, but Phase I ended in July 2014. There are currently over 339,000 full text documents back online. We currently have about 650 removed from the collection due to privacy concerns and 80,000 removed due to readability concerns. We are also adding new full text documents every week.

Questions?

- Q: How many ERIC documents have been taken offline because they are unreadable?
- A: Approximately 80,000. This is a lot of documents, but it is important to remember that the documents were not readable prior to the privacy concerns. We are working to go back to the original microfiche to make a much better copy that will be far more useful to the ERIC community.

Questions?

- Q: How will aggregators handle the unreadable documents?
- The unreadable documents will no longer be available on the ERIC site and the metadata will be updated to reflect that the documents are not available. Each year we will update our metadata files for historic records. When the "PDF Pending Restoration" files have been restored, we will immediately add them to the site and then update the historical metadata and weeding list on an annual basis.

Questions?

- Q: How soon are full-text documents provided to aggregators/database vendors?
- A: ERIC does not provide the full text document to anyone– but we do provide our our metadata. This metadata is available for anyone to download at http://eric.ed.gov/?download. We typically update the site once a week and then release new metadata in the middle of the following month. So the public http://eric.ed.gov will always have more recent content than the downloads page.

  One thing to note is that we have literally just transitioned our servers! This is huge news because it means ERIC can stay online, even if the government is shut down again. However, the person responsible for uploading the new content to the site is the one in charge of transitioning lots of IES data to the new servers. We haven't updated the site or the downloads in a few weeks, but the data files are coming soon!

Questions?

- Q: Will ERIC need new copies of the documents to replace the unreadable files?
- A: No, we have lots and lots and lots of microfiche to work with. We are fairly positive that we have clear copies that can be used to make a better copy and will go about prioritizing resources to make sure we only convert good copies. ERIC has not retained paper copies of records.

Questions?

- Q: Approximately how many full-text files contained PII?
- A: About 650 actual cases and about 150 of example PII– like Jane Doe, 123-45-6789.

Questions?

- Q: Is there a way to get e-alerts from ERIC as new content comes online?
- A: Not currently, but that is something we will explore in the future. We are looking at a whole bunch of new ideas to make ERIC better going forward, but want to prioritize our resources in a way which gets the most value for users. Right now, the biggest priority is getting new, good content into ERIC and highlighting the online submissions system.

Questions?

- Q: Has everything been resolved with the issue EBSCO was having linking to the full-text PDFs in ERIC?
- A: Yes. When we developed the new website, we set up a new URL structure and then also set up a way to redirect the old URLs to the new format. Over the weekend we transitioned to new servers, all of the new links worked, but the old links weren't set up yet. We were able to fix the re-direct links quickly after the problem was identified. If you ever see something not working, please let us know at http://eric.ed.gov/?contact or email me directly.

Questions?

- Q: Will ERIC be considering the option of bringing back an Advanced Search feature on the ERIC website?
- A: We are very willing to bring back the Advanced Search feature, but we want to make sure we build an advance search that works with the current website and is useful to our audience. I recommend you checking out our advanced search tips : http://eric.ed.gov/?advanced and then watch the video on the ERIC search which should be out link late September at https://www.youtube.com/user/SearchEduResources

  If you have specific examples of searches that you used to be able to do in ERIC, but no longer can with the new site, please let us know (http://eric.ed.gov/?contact)! We will work to find a way to make that search possible in the new site.

Thank you

Erin Pollard
ERIC Project Officer
US Department of Education
Erin.Pollard@ed.gov