The Civil Rights Litigation Clearinghouse: Access and Preservation

Seth Johnson
Michigan Publishing
University of Michigan Library

About

"The Civil Rights Litigation Clearinghouse, at the University of Michigan Law School, brings together and analyzes information and documents about important civil rights cases across the United States." http://clearinghouse.net/about.php

Consent Decrees hosted by the Clearinghouse often cannot be found anywhere else on the internet.

History

- Built by *Center for Empirical Research in the Law* at Washington University in St. Louis around 2005.
- Moved to University of Michigan Library when Prof. Margo Schlanger moved to Michigan Law around 2009
- ~ 40,000 PDF documents
- 1/4 FTE Law School school support

Overview and Demo

http://clearinghouse.net

http://chadmin.clearinghouse.net

PDFs come in from *everywhere*

- Scanned from individual lawyer's boxes
- Scans from local courts
- Random places on the internet
- ???

PDFs are renamed with the convention "Case Type"-"State"-"Case Number"-"Doc Number"

CJ-AL-0001-0001.pdf

(Criminal Justice, Alabama, Case #1, Doc #1)

Examples of where PDFs come from

```
% pdfinfo /n1/obj/c/ch/chDocs/not_public/CJ-AL-0001-0001.
pdf
  Title:          Jones v. Allen - Opinion
  Author:         thompson
  Creator:        PScript5.dll Version 5.2
  Producer:       Acrobat Distiller 7.0 (Windows)
  CreationDate:   Tue Apr 17 16:25:24 2007
  ModDate:        Wed Jan 25 00:00:00 2012
  Tagged:         no
  Pages:          42
  Encrypted:      no
  Page size:      612 x 792 pts (letter)
  File size:      222408 bytes
  Optimized:      yes
  PDF version:    1.4
```

Examples of where PDFs come from

The ~40,000 PDFs have 1130 unique "Creator" metadata entries.

This is the Top 10 "Creator" metadata with counts

7834	PScript5.dll Version 5.2
6371	*no entry*
2848	PScript5.dll Version 5.2.2
2799	PDF reDirect v2
2483	ABBYY FineReader
977	TIFF2PDF v1.14, 1999-12-14
575	ExperVision
453	POP90
395	c42pdf v. 0.12 args:
389	TIFF2PDF v1.14a,

Examples of where PDFs come from

"PDF Version" metadata with counts

12781	1.4
10252	1.3
7599	1.6
3383	1.2
3232	1.5
126	1.7
11	*no entry*
9	1.0

Full Text Search

- This process of uploading PDFs went on from 2006 until 2011.
- In 2011, development on a long asked for feature for “Full Text Search” started.
- Google Site Search was chosen due to simplicity and cost.
- GSS automatically indexes the site PDFs and displays results inline on the site.

PDFs Were A Mess

OCR Problems

OCR was “usually” there with varying quality as students performed OCR on PDFs prior to upload but used any software available to them to do this.

Title Metadata Problems

Google uses the Title metadata field from the indexed PDFs for search headings.

Consultation with the Digital Preservation Librarian

Is it ok to make changes to these documents?

- We don't currently have truly accurate information about this material (who scanned it? when? did they miss a page? how can we know?)
- We're adding value, not taking it away
- The 'spirit' of the Clearinghouse is access, not preservation

Quarterly OCR and Title Metadata Process (for new PDFs)

- Adds OCR if needed
- Adds new Title metadata using the Clearinghouse Case Name and Doc Name ("Jones v. Allen - Opinion")
- Saves as PDF/A
- Overwrites old PDFs in production
- Google indexes the results

The Clearinghouse isn't an Archive

The "spirit" of the Clearinghouse is access, not preservation (but we should try to get better about preservation where it makes sense).

- Source PDFs still come from all over
- Students no longer do OCR, it’s done centrally for new PDFs that need OCR
- Title metadata is now useful information
- New PDFs are converted to PDF/A