

# PDF Liberation Hackathon

NYC, DC, Chicago, Dayton, SF & Buenos Aires- January 2014

Resource List and Links to Winning Entries at:

<http://pdfliberation.wordpress.com/>

# An Example of How PDF Liberation Can Generate News

- Working with Mortgage Resolution Partners, the City of Richmond has proposed to use its power of eminent domain to refinance mortgages for underwater homeowners
- In July, the media reported that 624 properties had been chosen
- I wanted to know which ones, so I filed a California Public Records Act request . . .

# The Request...(Make it Very Specific)

Dear Ms. Holmes,

Pursuant to my rights under the California Public Records Act (Government Code Section 6250 et seq.), I ask to obtain a copy of the following, which I understand to be held by your agency:

Attachments A, B and C to letters sent to mortgage servicers offering to purchase mortgage loans dated on or about July 31, 2013. The form letter is available on the internet at [http://www.contracostatimes.com/west-county-times/ci\\_23760190/document-city-richmond-letter-mortgage-lenders?source=pkg](http://www.contracostatimes.com/west-county-times/ci_23760190/document-city-richmond-letter-mortgage-lenders?source=pkg). I understand that 32 such letters have been sent, so this request involves as many as 96 unique documents.

The purpose of this request is to obtain a list of 624 mortgages which Richmond is offering to purchase containing the property addresses, mortgage amounts, appraised values, servicer names, and, if possible, the name of the Residential Mortgage Backed Securities (RMBS) deal holding each mortgage. If you can provide this listing in a more concise format, I will accept it in lieu of the attachments described in the previous paragraph.

I ask for a determination on this request within 10 days of your receipt of it, and an even prompter reply if you can make that determination without having to review the record[s] in question.

If you determine that some but not all of the information is exempt from disclosure and that you intend to withhold it, I ask that you redact it for the time being and make the rest available as requested.

In any event, please provide a signed notification citing the legal authorities on which you rely if you determine that any or all of the information is exempt and will not be disclosed.

If I can provide any clarification that will help expedite your attention to my request, please contact me by phone at 415-578-0558 or by email at [marc@publicsectorcredit.org](mailto:marc@publicsectorcredit.org). I ask that the requested documents be sent to be in electronic format via return email. If you must provide paper documents, I ask that you notify me of any duplication costs exceeding \$50 before you duplicate the records so that I may decide which records I want copied. I can visit your office to collect the documents once they have been duplicated.

Thank you for your time and attention to this matter.

Sincerely,

Marc D. Joffe  
1655 North California Blvd. Unit 162  
Walnut Creek, CA 94596

# The Response...

- Four PDFs

Servicer Exhibit A

| Current Servicer Parent | LoanId     | BloombergDealName | LewtanDealName   |
|-------------------------|------------|-------------------|--|
| Ally Financial          | 114733141  | DBALT 2007-OA3    | Deutsche Alt-A Securities Mortgage Loan Trust 2007-OA3 |
| Ally Financial          | 534040     | HVMLT 2006-10     | HarborView Mortgage Loan Trust 2006-10                 |
| Ally Financial          | 6014223    | HVMLT 2006-10     | HarborView Mortgage Loan Trust 2006-10                 |
| Ally Financial          | 1000104209 | HVMLT 2006-10     | HarborView Mortgage Loan Trust 2006-10                 |
| Ally Financial          | 1000105602 | HVMLT 2006-10     | HarborView Mortgage Loan Trust 2006-10                 |
| Ally Financial          | 49965257   | HVMLT 2007-3      | HarborView Mortgage Loan Trust 2007-3                  |
| Ally Financial          | 643544     | HVMLT 2007-4      | HarborView Mortgage Loan Trust 2007-4                  |

Trustee Exhibit A

| Trustee          | LoanId     | Parcel Number | House Number | Dir. | Street Name    | Street Suffix | Unit | Unit Value | City     | Zip   | Plus 4 |
|------------------|------------|---------------|--------------|------|----------------|---------------|------|------------|----------|-------|--------|
| Bank of New York | 1765493317 | 5192100179    | 544          |      | MCLAUGHLIN     | ST            |      |            | RICHMOND | 94805 | 1947   |
| Bank of New York | 1844561126 | 4321920110    | 5537         |      | CABRILLO NORTE |               |      |            | RICHMOND | 94803 | 3877   |
| Bank of New York | 1730035940 | 5192400058    | 5215         |      | SILVA          | AVE           |      |            | RICHMOND | 94805 | 2409   |
| Bank of New York | 1190465323 | 4334310036    | 208          |      | PIONEER        | CT            |      |            | RICHMOND | 94803 | 2648   |
| Bank of New York | 1846634720 | 5561520023    | 68           |      | IDAHO          | ST            |      |            | RICHMOND | 94801 | 4045   |
| Bank of New York | 58451350   | 5181120022    | 677          |      | 37TH           | ST            |      |            | RICHMOND | 94805 | 1776   |

# Processing

- Loaded the four PDFs into Able2Extract – a commercial PDF conversion tool that costs about \$100\*
- Converted the PDFs to Microsoft Excel
- I had now had multiple lists of properties with different fields
- I sorted the lists into the same order and then joined them together into one master spreadsheet
- I found that three properties had mortgage balances over \$800,000 and was able to connect the balances to the addresses
- This made it possible to map the properties and to see the houses themselves on Google Street View

\* ***Tabula***, an open source tool, is reaching the point at which it could perform the same function.

# The Results ...

- Lead story in the business section of the Chronicle
- Wall Street Journal blog post
- Finding raised at City Council meeting
- In December, Mayor Gayle McLaughlin altered the program to exclude mortgages above the conforming loan limit (\$729,500) and to focus on blighted neighborhoods.

*By the way:*

The owner of the house on the right was apparently unaware that her home had been included in the program. So my initial theory that this had been a case of cronyism was not borne out.


PDF Liberation Hackathon x Pricey homes in Richmond x

www.sfgate.com/business/article/Pricey-homes-in-Richmond...  
Apps Suggested Sites PSC Historical Data PSC PSCF CA City Credit Scoring...

## Pricey homes in Richmond's eminent domain plan

Carolyn Said  
Updated 4:56 pm, Tuesday, August 20, 2013

VIEW: LARGER | HIDE 1 of 8 ◀ PREV NEXT ▶



Point Richmond: Sold for \$1.195 million. Loan balance: \$888,361. City's offer: \$510,727. Estimated value: \$666,461. Photo: Brant Ward, The Chronicle

◀ ▶

10 106 5  
Tweet f Share g+1  
44  
in Share

Richmond's controversial plan to seize underwater mortgages through eminent domain includes loans for at least two homes purchased for over \$1 million as well as other high-end properties - a revelation that appears to undermine the city's argument that the plan would combat blight.

arlington\_cnty\_fy201....pdf  
Cancelled

Show all downloads...

# Some of Our Challenges

- Government Financial Statements
- IRS Form 990s (Non-Profit Disclosures)
- House of Representative Financial Disclosures
- Compiling a History of Torture

# Government Financial Statements: Finding the Next Detroit

City of Detroit, Michigan  
General Fund Balance Sheet  
June 30, 2012 and 2011  
(in millions)

|                              | <u>2012</u>    | <u>2011</u>    |
|------------------------------|----------------|----------------|
| Assets                       | \$ 246.9       | \$ 290.2       |
| Liabilities                  | 516.4          | 438.3          |
| Fund Balance                 |                |                |
| Nonspendable                 | 20.9           | 20.7           |
| Restricted                   | 1.0            | 1.0            |
| Committed                    | 35.2           | 26.8           |
| Unassigned for General Fund  |                |                |
| Deficit                      | <u>(326.6)</u> | <u>(196.6)</u> |
| Total Fund Balance (Deficit) | <u>(269.5)</u> | <u>(148.1)</u> |



IRS Form  
990s:  
Finding  
members  
of the 1%  
who work  
at not-for-  
profits

**Part VII** Section A. Officers, Directors, Trustees, Key Employees, and Highest Compensated Employees (continued)

| (A)<br>Name and title  | (B)<br>Average<br>hours per<br>week (list any<br>hours for<br>related<br>organizations<br>below dotted<br>line) | (C)<br>Position<br>(do not check more than one<br>box, unless person is both an<br>officer and a director/trustee) |                       |         |              |                                 |        | (D)<br>Reportable<br>compensation<br>from<br>the<br>organization<br>(W-2/1099-MISC) | (E)<br>Reportable<br>compensation from<br>related<br>organizations<br>(W-2/1099-MISC) | (F)<br>Estimated<br>amount of<br>other<br>compensation<br>from the<br>organization<br>and related<br>organizations |
|--|---|--|-----------------------|---------|--------------|---------------------------------|--------|---|---|--|
|  |   | Individual trustee<br>or director  | Institutional trustee | Officer | Key employee | Highest compensated<br>employee | Former |   |   |  |
| 15) BARRY WILLIAMS<br>BOARD MEMBER                             | 7.00<br>0   | X  |                       |         |              |                                 |        | 27,500.   | 0   | 0  |
| 16) FLO DI BENEDETTO<br>SVP & GENERAL COUNSEL/ASST SEC         | 40.00<br>1.00   |  |                       | X       |              |                                 |        | 1,003,764.  | 0   | 649,941.   |
| 17) ED ERWIN<br>DIR REAL ESTATE SRVCS/ASST SEC                 | 40.00<br>0  |  |                       | X       |              |                                 |        | 188,768.  | 0   | 26,063.  |
| 18) ROBERT REED<br>SVP & CFO SUTTER HEALTH                     | 40.00<br>2.00   |  |                       | X       |              |                                 |        | 1,773,201.  | 0   | 1,240,260.   |
| 19) PETER ANDERSON<br>SVP STRATEGY & BUS. DVLPMT               | 40.00<br>12.00  |  |                       |         | X            |                                 |        | 1,020,427.  | 0   | 567,768.   |
| 20) DAVID BENN<br>REG. PRES., CENTRAL VALLEY                   | 0<br>40.00  |  |                       | X       |              |                                 |        | 1,037,565.  | 0   | 639,864.   |
| 21) ED BERDICK<br>SVP SHARED SERVICES                          | 40.00<br>5.00   |  |                       | X       |              |                                 |        | 1,529,687.  | 0   | 878,098.   |
| 22) MARTIN BROTMAN<br>REGIONAL PRESIDENT, WEST BAY             | 0<br>40.00  |  |                       | X       |              |                                 |        | 2,765,896.  | 0   | 546,676.   |
| 23) JEFF BURNICH MD<br>SVP, EXEC OFFICER MED NETWORK           | 40.00<br>0  |  |                       | X       |              |                                 |        | 964,357.  | 0   | 644,593.   |
| 24) MIKE COHILL<br>SVP SUTTER HEALTH                           | 40.00<br>0  |  |                       | X       |              |                                 |        | 1,769,123.  | 0   | 676,890.   |
| 25) JAMES CONFORTI<br>REGIONAL PRESIDENT, SSR                  | 0<br>40.00  |  |                       | X       |              |                                 |        | 812,769.  | 0   | 526,031.   |
| <b>1b Sub-total</b>  |   |  |                       |         |              |                                 |        | 4,811,827.  | 0   | 2,911,191.   |
| <b>c Total from continuation sheets to Part VII, Section A</b> |   |  |                       |         |              |                                 |        | 26,426,032.   | 0   | 13,788,067.  |
| <b>d Total (add lines 1b and 1c)</b>                           |   |  |                       |         |              |                                 |        | 31,237,859.   | 0   | 16,699,258.  |

... And  
finding the 1%  
in Congress by  
dissecting  
House  
Financial  
Disclosures

**Schedule III - Total Assets and "Unearned" Income**  
**Year: 2012**  
**Honorable Darrell Issa**

| Investment                               | Value of Asset |                |                     |                      |                       |                        |                        |                          |                            |                             |                              |                      |
|--|----------------|----------------|---------------------|----------------------|-----------------------|------------------------|------------------------|--------------------------|----------------------------|-----------------------------|------------------------------|----------------------|
|  | A. None        | B. \$1-\$1,000 | C. \$1,001-\$15,000 | D. \$15,001-\$50,000 | E. \$50,001-\$100,000 | F. \$100,001-\$250,000 | G. \$250,001-\$500,000 | H. \$500,001-\$1,000,000 | I. \$1,000,001-\$5,000,000 | J. \$5,000,001-\$25,000,000 | K. \$25,000,001-\$50,000,000 | L. Over \$50,000,000 |
| Alliance Bernstein High Income Fund CI A |                |                |                     |                      |                       |                        |                        |                          |                            |                             |                              | X                    |
| Allianz AGIC High Yield Bond Fd CI A     |                |                |                     |                      |                       |                        |                        |                          |                            |                             |                              | X                    |

This project was taken on by our second place prize winner. Their best results came from using [Captricity.com](http://Captricity.com).

# Documenting a History of Torture: Parsing Amnesty International Annual Reports

In November 1978, an Amnesty International research mission spent two weeks in Iran interviewing released prisoners, relatives of prisoners and lawyers. The information obtained confirmed allegations spanning the past 15 years that the torture of political prisoners had been practised systematically throughout the country and that although it appeared to have decreased since early in 1977, when the Shah announced that the use of torture had ceased, it had not stopped altogether. In a press release on 11 December 1978 Amnesty International published details of some recent cases of torture reported to its delegates. Methods of torture described included whipping with cables, the beating of the soles of the feet, kicking, punching, burning of parts of the body with cigarettes, prolonged sleep deprivation combined with forced standing, the application of nettles to sensitive parts of the body and long periods of solitary confinement.

**This project was taken on by our first place prize winner.**

# Three Inter-Related Problems ...

- Extracting data from PDFs that contain embedded text
- Using Optical Character Recognition (OCR) to generate text from PDFs of scans or photographs
- Transforming unstructured text and numbers into a form that can be readily analyzed. A related IT term is ETL (Extract-Transform-Load)

## ... and some Open Source Solutions

- Extracting data from PDFs that contain embedded text

PDFBox, Poppler

- Using Optical Character Recognition (OCR) to generate text from PDFs of scans or photographs

Tesseract

- Transforming unstructured text and numbers into a form that can be readily analyzed. A related IT term is ETL (Extract-Transform-Load)

Tabula (for table identification), OpenRefine

## ... or Licensed Solutions

- Extracting data from PDFs that contain embedded text

**PDFLib Text Extraction Tool**

- Using Optical Character Recognition (OCR) to generate text from PDFs of scans or photographs

**ABBYY (FineReader of Cloud SDK)**

- Transforming unstructured text and numbers into a form that can be readily analyzed. A related IT term is ETL (Extract-Transform-Load)

**SIMX Text Converter**