

COMPUTATION + JOURNALISM IN THE PUBLIC INTEREST

Sarah Cohen, DeWitt Wallace Center

Public interest reporting defined

- Information of importance to the public that powerful institutions would prefer to be kept hidden or secret.
- A method of comparing what *ought* to happen with what *actually* happens
 - ▢ Breaking laws or rules
 - ▢ Taking advantage of the helpless
- Comforting the afflicted and afflicting the comfortable.

Common traits

Reporting shares some traits with other investigative or exploratory fields :

- Use of multiple sources with varying accuracy and collection methods
- Practitioners have little or no technical ability or interest
- Much of the data is in the form of unstructured text
- Start with broad, ill-defined questions, not discrete goals
- Combines field work and lab work (street reporting and data or document analysis)

Uncommon traits

- Rarely spend much time on a given data source or question
- Reporters look for tips and examples, not evidence or statistical patterns.
- Work from primary sources
- Analysis is not mission-critical
- Unique identifiers are censored.
- Little or no control over the form of the data: it's up to the source

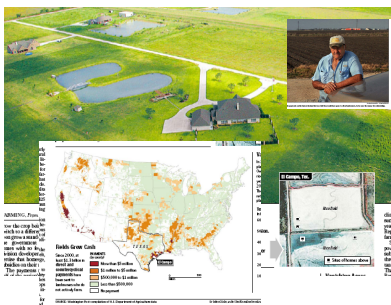
Typical data sources

- Interview notes, both on the record and for background only. Very little is taped, especially early in a story.
- Taped or videotaped hearings, sometimes transcribed already
- Calendars, correspondence, investigative and inspection reports
- Emails and press releases
- Excel spreadsheets created for printing, not analysis
- Structured databases
- Court documents: complaints, indictments, settlements, discovery.
- Online document collections, such as regulatory enforcement actions or audits.
- Handwritten forms, such as inspections or police incident reports, usually heavily redacted

Transparency myths

- Government records are easy to get in a usable form? Not so much.
- Examples, even when it is:
 - [Recovery Act](#)
 - [Campaign finance](#)
 - [Crime](#)
- Most government information is still hidden, secret or difficult to access
 - data.gov: Promise of 100,000 datasets. Currently have 728, half of which are annual state-level cuts of existing EPA data that has been available for years. No records, mainly just aggregated statistics. Almost nothing on it that wasn't already readily available.

Harvesting Cash

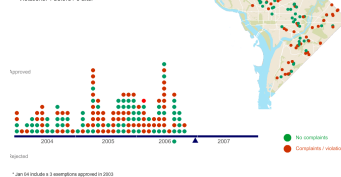


Sources:

- People
- database of 200 million farm subsidy payment
- crop estimates
- weather records
- emergency records
- property ownership records (local)
- letters from members of Congress

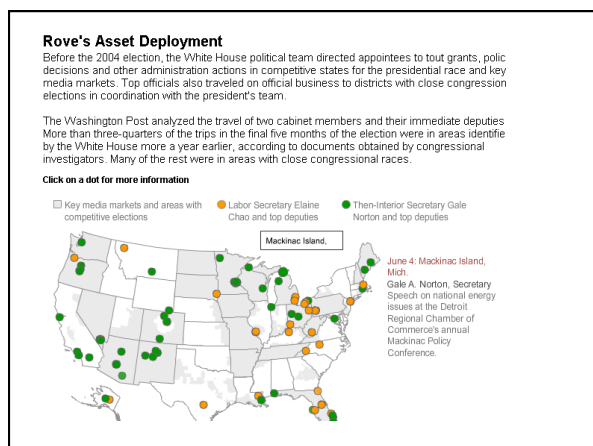
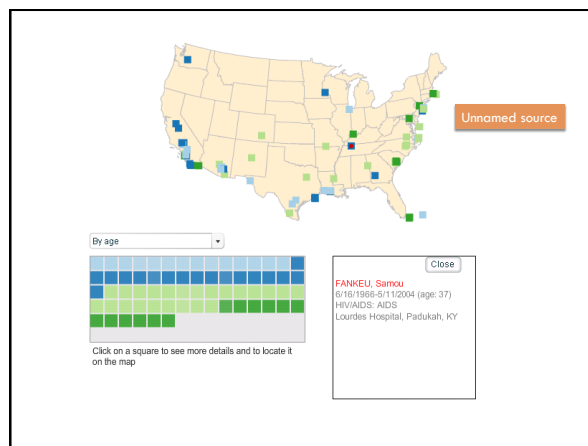
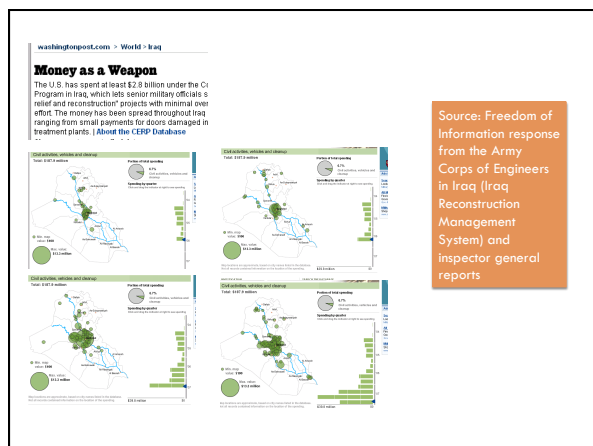
Forced Out

Certification date: Feb 23, 2008
1901 & 1903 Locust Street, N.E.
(No condo name given)
Complaints: 3 before / 0 after
Violations: 1 before / 0 after

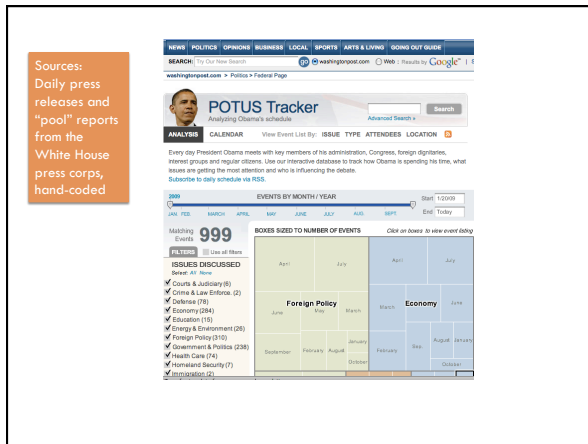
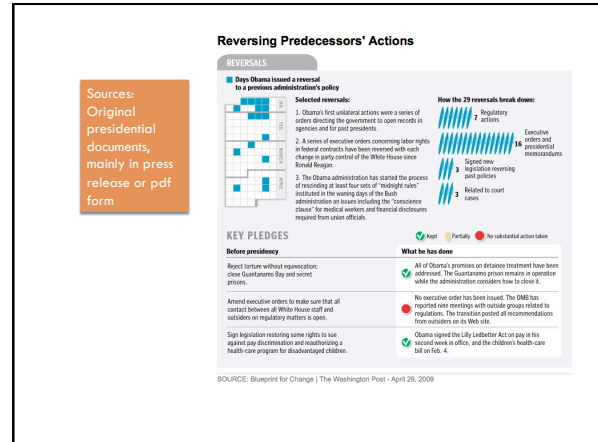


Sources:

- File cabinets (some missing)
- Spreadsheet where the color meant different things.
- 200,000 or so housing code complaints and violations
- Paper records from court
- People (residents, activists, government officials)
- Physical inspection of properties
- Landlord-tenant court disputes
- Property assessments and deeds



My prior Checking - Select Query								
Department	Edited	Past Checked	Link To City	Grant/Dates	Offical Name	Official Title	Location By Locality	Description
Interior	Yes	Yes	http://www.doi.gov/news/040802_pgs.htm	9/25/2004	Craig Manson	Assistant Secretary	Bayfield, WI	All Apostle Islands National Lakeshore, announced the completion of its wilderness study.
Labor	Yes	Yes	http://www.dailycoronadoalumni.com/articles/040619-1.html	8/29/2004	Dawn D. Lauriski	Assistant Secretary	Las Vegas, NV	With the National Mining Association, presented awards to firms operating for their safety records.
Summary								
			n	Amount	Notes	Program		
re,				\$0.00	Visited Apostle Islands National Park to announce the completion of the park's wilderness study, which was to be published in the Fed Register the next day. Fed govt wanted to protect about 80	Wilderness Act		
Mining entities operates onlands.				\$0.00	presented at the annual conference cooperative alliance to foster more safety and health in mining. Also presented awards for safety and health. DeRocco was also there. Can't tell if it is a conference or what.			



Theories of open data

- Governments prefer to lead you to the “right” answer, so they like colorful and “user-friendly” websites that convey their message.
- Transparency advocates look for government to first “wholesale” data, then work on their own website. They’ve had little luck.

New directions in newsgathering

- *What’s New?* Smart scraping and text mining – indictments, settlements and regulatory actions across agencies; lawmaker releases and grant announcements.
- *The Real Story*: Anti-aggregation of stories and news from blogs, news sites, RSS feeds, government agencies grouped by story, not by source.
- *Chronologies and social network tools* to organize notes and see new connections
- *Text mining of government documents: Jigsaw, Meandre, Document Cloud*
- *Audio and video analysis*

Opportunities for Computer Science students

- DukeEngage project (look on the DeWitt Wallace Center’s website for a link) to create newsgathering tools
- “Middle layer” of sense-making on unexplored local and national datasets
 - Stimulus
 - Contracts and grants
 - Dispersed records (calendars, etc.)
- Entity extraction, geocoding tools
- Visualizations that say something new