DATA: Now I’ve got it; what do I do with it?

Tom Johnson
Managing Director
Inst. for Analytic Journalism
Santa Fe, New Mexico USA

tom@jtjohnson.com
DATA:
Now I’ve Got It; what do I do with it?

Presentation at
IRE’s Albuquerque Watchdog Workshop
Feb. 12-13, 2011
Hosted by the University of New Mexico

• This PowerPoint deck posted at:

• Tipsheet handout posted at:
Important point

The document is **not** the data.
Important point

The document is *not* the data.
The data are *not* the story without analysis.
Important point

Nothing is as important, and valuable, as a good theory!
A good, important THEORY

Theory of Journalistic Process

Data In $\rightarrow$ Analysis $\rightarrow$ Info Out
DATA IN: Retrieve

Bookmark apps

• Objectives:
  • Access via browser – but not standard equipment
  • Create/manage sub-folders, categories & keywords, annotations
  • Private and/or public sharing
  • Save and Export to backup system(s)

• Examples:
  • Xmarks: www.xmarks.com/
  • Diigo: www.diigo.com/index
  • Freeware/shareware search for at www.tucows.com
OK, it’s downloaded. Where ya gonna save it?

- Multiple back-up sites: desktop and...
  - Safer in Cloud than otherwise
    - Passwords, but share capabilities
  - Easier with “Cloud-sync” apps
  - Free to low-cost
DATA IN: Store & Share in the Cloud

OK, it’s downloaded. Where you gonna save it?

- Avoid MS Windows Live, SkyDrive and Mesh – more trouble than they are worth
- Dropbox – www.dropbox.com

Your Hard Drive
OK, it's downloaded. Where ya gonna save it?

- Avoid MS Windows Live, SkyDrive and Mesh – more trouble than they are worth
- Dropbox - www.dropbox.com Viewed in your browser

DATA IN: Store & Share in the Cloud

- Folders, subfolders, sub-subfolders, etc.
- Nearly instant sync-ing with/from your desktop
DATA IN: Store & Share in the Cloud

OK, it’s downloaded. Where ya gonna save it?

• **Avoid** MS Windows Live, SkyDrive and Mesh – more trouble than they are worth
• Dropbox - [www.dropbox.com](http://www.dropbox.com)
• SugarSync - [www.sugarsync.com](http://www.sugarsync.com)
• Syncplicity - [www.syncplicity.com](http://www.syncplicity.com)
• Jungle Disk ($3p/m) - [www.jungledisk.com](http://www.jungledisk.com)
• Zumodrive (3p/m) - [www.zumodrive.com](http://www.zumodrive.com)
• AeroFS - [www.zumodrive.com](http://www.zumodrive.com)
• SpiderOak - [spaderoak.com](http://spaderoak.com)
• MiMedia, Wuala, Quanp,
Data In \(\rightarrow\) Analysis \(\rightarrow\) Info Out

- Notes
- Text
- Numeric
- Images
- Charts/Graphs
- Maps
- Audio
- Video
- Atoms \(\rightarrow\) Bits
- How? Who?
Data In: Objectives

• Move data from “out there” to analytic site/tools
• Seeking fine-grained data, NOT aggregations
  • Seek data in original form (i.e. NO PDFs)
  • Who collected the data? Why? How?
  • Who proofed/edited the data? Why? How?
  • If from data base, first ask for “record” or “code sheet” or “schema”
  • Definitions of variables or fields. Constant or ???
• Get data in lowest common denominator format: Comma-delimited files in ASCII or Text
Data In: Challenges

- New site in New Mexico: [www.sunshineportalnm.com](http://www.sunshineportalnm.com)
- “Beta,” but looks to be a cruel joke on taxpayers; torture for journos
Data In: “Typical” problems with SunshineportalNM

- Barriers data = barriers to analysis
  - **NO** site search capability; no site map
  - Completely abandoned open-standard HTML, going for the **closed-standard** Adobe Flash/Shockwave environment.
  - Page formats/layouts not standard; too many drill-downs instead of search-driven generators
  - Jiggly roll-overs; too much effort spent on bling
  - Impossible to download or scrape data for analysis
  - State makes information available only in Adobe PDF files; notoriously unfriendly to data analysis.
Data In: Challenges in SunshinePort

- Comprehensive Annual Financial Reports
  - Possible to machine download, but laborious to format for analysis
- Investment Holdings reports are far worse
  - They are poor-quality static image files, not machine-readable.
  - Tabular data roughly formatted; makes conversion for analysis an arduous, if not impossible task

CHAPTER 34

AN ACT

RELATING TO STATE GOVERNMENT; ENACTING THE SUNSHINE PORTAL TRANSPARENCY ACT; PROVIDING FOR THE DEVELOPMENT, OPERATION AND MAINTENANCE OF A WEB-BASED INFORMATION PORTAL THAT PROVIDES PUBLIC ACCESS TO STATE GOVERNMENT BUDGETS, EXPENDITURES, REVENUE AND OTHER INFORMATION.

BE IT ENACTED BY THE LEGISLATURE OF THE STATE OF NEW MEXICO:

Section 1. SHORT TITLE.--This act may be cited as the "Sunshine Portal Transparency Act".

Section 2. DEFINITIONS.--As used in the Sunshine Portal Transparency Act:

A. "department" means the department of information technology;

B. "exempt employee" means an employee of the state or a state agency who is in a policymaking or supervisory position and who serves at the discretion of the agency head or at the discretion of an appointee of the agency head;

C. "expenditure" means a disbursement of state or federal funds by a state agency, whether or not the funds have been appropriated by the legislature;

D. "federal funds" means money received by a state agency from the federal government or one of its branches, departments, agencies, offices, officers or
“A perfect example of creating the appearance of transparency without actually being transparent.”
Challenge for Watchdogs?

• Failure on the part of planners/bureaucrats to simply...

• Give The People THEIR Data...

• In The Most Basic, Original, Straightforward Form...

• And Let Them Figure Out What Should Be Done With It!
Positive example of gov’t data

Same data available in two formats!
NM HB 406

- Senate approved 39-0 on Feb. 9
  http://www.nmlegis.gov/Sessions/11%20Regular/bills/house/HB0406.html

- “An Act RELATING TO PUBLIC RECORDS; PROVIDING FOR THE INSPECTION OF ELECTRONIC RECORDS.”
But what if it wasn’t New Mexico state employees directly at fault?
Why is it sunshineportalNM.COM?

- Domain Name: SUNSHINEPORTALNM.COM
- Registrar: WILD WEST DOMAINS, INC
- Referral URL: http://www.wildwestdomains.com
- Name Server: ENESFOUR.SKS.COM
- Name Server: ENESONE.SKS.COM
- Name Server: ENESTHREE.SKS.COM
- Name Server: ENESTWO.SKS.COM
- Status: clientDeleteProhibited
- Status: clientRenewProhibited
- Status: clientTransferProhibited
- Status: clientUpdateProhibited
- Updated Date: 30-mar-2010
- Creation Date: 30-mar-2010
- Expiration Date: 30-mar-2011
Make money now with the HOTTEST Reseller plan going!
Our success depends on yours. That's why we keep our prices low - so you can sell more products and net more profits.

We set you up with everything you need:
- Pre-built website with shopping cart
- All credit card processing & billing
- We even provide the products

30-day money back guarantee

SIGN UP TODAY
SELL DOMAINS & RELATED PRODUCTS

Choose your plan

- $99.99/year
  - Basic Reseller
    - Turnkey eCommerce Website

- $179.99/year
  - Pro Reseller
    - Turnkey eCommerce Website
    - Best Buy Rates in the Industry
    - Pay-Per-Click Parked Domain Commissions

- $199.99/year
  - Super Reseller
    - All the features of the Pro Reseller
    - PLUS Sell Basic and Pro Plans
    - Create Multiple Revenue Streams
• Registrant:
  Wild West Domains, Inc.
  14455 N Hayden Rd Suite 219
  Scottsdale, Arizona 85260  United States

Registered through: WWDomains.com
•  Domain Name: WILDWESTDOMAINS.COM
  Created on: 22-Aug-00
  Expires on: 22-Jul-19
  Last Updated on: 08-Dec-09

Administrative Contact:
•  Wild West Domains, Inc.
  dns@wildwestdomains.com
•  Wild West Domains, Inc.
•  14455 N Hayden Rd Suite 219
•  Scottsdale, Arizona 85260  United States
•  +1.4805058800      Fax -- +1.4805058844
Post-data recovery: Analytic DNA

**Qualitative**
- Who
- What
- When
- Why
- Where
- How

**Quantitative**
- How many/much
- What categories
- What type data and levels
- What changes
- "What timeline"

**Geo-location**
- All stories have geography
- People are interested in how close is this to me?
Data In → **Analysis** → Info Out

- What are we looking for? How can we be surprised?
- Source
- Definition
- Context
- Estimating
- Counting
- Statistical
- Geostatistical
- Social Network Analysis
- Forensic accounting
The “Fundamental Five” Statistics

1. Calculating percent of change
   - \[(\text{New}-\text{Old}) \div \text{Old} \times 100\]
   or
   - \[((\text{new/old}) - 1) \times 100\]

2. Calculating proportion:
   - \[(\# \text{ of parts} \div \text{TOTAL \# of parts}) \times 100 = \% \text{ of whole}\]
The “Fundamental Five” Statistics

3. Calculating Rates:
   (incidents ÷ population) * 10,000 (or 100,000)

4. Calculating Ratios:
   • Take first of two numbers being compared and divide by second.
   • $600 \div 30 = 20$ [Ratio is 20-to-1; if fraction, round off]
The “Fundamental Five” Statistics

5. Calculating Inflation:

- (CPI Now ÷ CPI Then) * Item Price Then = Item then in today’s $$$
  [Tool: http://www.westegg.com/inflation/]

- Calculating INFLATION RATE
  CPI in 2000 is 3,500 CPI in 2001 is 4,500 What's the inflation rate?

  4500 - 3500 = 1000
  1000/3500 = .2857....
  .2857 * 100 = 28.57 is the INFLATION RATE
Data In ➔ **Analysis** ➔ Info Out

- **Online tools**
  - [Google Docs Spreadsheets](https://docs.google.com/spreadsheets)
  - Google Refine
  - Freebase
  - Google Fusion Tables

**Google Refine** is a power tool for working with messy data, cleaning it up, transforming it from one format into another, extending it with web services, and linking it to databases like Freebase.

**Freebase** is an open, [Creative Commons](https://creativecommons.org) licensed repository of **structured data** of almost 20 million **entities**. An **entity** is a single person, place, or thing. Freebase connects entities together as a **graph**.

**Fusion Tables** is a service for managing large collections of tabular data in the cloud. You can upload tables of up to 100MB and share them with collaborators, or make them public.
Data In → Analysis → **Info Out**

**Data In**
- Notes
- Text
- Numeric
- Images
- Charts/Graphs
- Maps
- Audio
- Video
- Atoms → Bits

**Analysis**
- What are we looking for? How can we be surprised?
- Source
- Definition
- Context
- Estimating
- Counting
- Statistical
- Geostatistical
- Social Network Analysis
- Forensic accounting

**Info Out**
- Broadcast
- Web
- Audio
- Video
- Text
- Data visualization
- Maps
- Dynamic databases
- Archives
“Analytic tools” also for story-telling

- Spreadsheets:
  - Tables, charts, infographics
- Data base programs:
  - Charts, graphs, data tables
- Stats programs (SPSS or SAS or R):
  - Generate graphics
- Social network analytic graphics
- GIS
TimeFlow Analytical Timeline Visual timelines for investigation
Tomorrow?

• Our job is to “monitor the centres of power.”

-- Amira Haassaid

The document is *not* the data
DATA:
Now I’ve got it; what do I do with it?

Gracias a todos

Tom Johnson
Managing Director
Inst. for Analytic Journalism
Santa Fe, New Mexico USA
tom@jtjohnson.com