Mining the Web for data and stories

Phoenix CAR 2010

Ted Mellnik, *The Charlotte Observer*
Jaimi Dowdell, IRE/NICAR

Searchable data

A lot of data is already available online, you just need to know where to look. Here is a list of places you can go, right now, to find searchable databases from government agencies:

- FedStats – Statistics from government agencies - www.fedstats.gov/
- Data.gov – "*The purpose of Data.gov is to increase public access to high value, machine readable datasets generated by the Executive Branch of the Federal Government.*" – www.data.gov
- USASpending.gov – Search and download federal contracts, grants and more - www.usaspending.gov/
- BRB Publications has a free directory of public records sites - www.brbpub.com/freeresources/pubrecsites.aspx?h=1
- OSHA – workplace safety data - www.osha.gov/oshstats/index.html
- Fatal Accident Reporting System data (FARS) - www-fars.nhtsa.dot.gov/QueryTool/QuerySection/SelectYear.aspx
- Mine Safety and Health Administration (MSHA) - www.msha.gov/drs/drshome.htm
- Federal Railroad Administration - safetydata.fra.dot.gov/OfficeofSafety/publicsite/query/query.aspx
- National Agricultural Statistics Service - www.nass.usda.gov/Data_and_Statistics/Quick_Stats/index.asp
- Census - American FactFinder quick profiles of states, counties, etc. - factfinder.census.gov/home/saff/main.html?_lang=en
- EPA Environment data searches - www.epa.gov/epahome/commsearch.htm
- FAA data (service difficulty reports, on-time data, aircraft registry, etc.) - www.faa.gov/data_research/
    - Service Difficulty Reports - av-info.faa.gov/sdrx/
    - On-time stats - www.transtats.bts.gov/OT_Delay/OT_DelayCause1.asp

- Landings (aircraft registry, pilots) - www.landings.com/evird.acgi$pass*129569481!mtd*7!map*_landings/images/landings-topbuttons.map?98,11

- Bureau of Transportation Statistics - www.bts.gov/data_and_statistics/

- National Sex Offender Registry - www.nsopw.gov/Core/OffenderSearchCriteria.aspx?Advanced=1

- Bureau of Labor Statistics - http://www.bls.gov/data/

Downloadable data

Besides searchable databases, there are plenty of places you can go to download information so you can analyze it on your own in a spreadsheet or database manager. Here are some places to get this kind of data (Note: Look for key words on government sites such as: Excel, .xls, download, database, etc.):

- Census data - query and download results - factfinder.census.gov/servlet/DatasetMainPageServlet?_lang=en&_ts=262968313843&_ds_name=ACS_2007_3YR_G00_&_program=

- Census download center - factfinder.census.gov/servlet/DownloadDatasetServlet?_lang=en

- Nursing Home Compare data download from Medicare.gov - www.medicare.gov/Download/DownloadDB.asp

- FRA data download - safetydata.fra.dot.gov/OfficeofSafety/publicsite/downloads/downloads.aspx

- MSHA accident and incident data download - www.msha.gov/STATS/PART50/p50y2k/p50y2k.HTM

- FBI Uniform Crime Reports download spreadsheets - www.fbi.gov/ucr/ucr.htm#cius

- Federal firearms licensees from the ATF - www.atf.gov/statistics/ffl-list.html

Advanced search strategies

You can also use advanced searches in Google to find data. Check out this guide to Google searching with the advanced form - www.googleguide.com/sharpening_queries.html:

- Search by file type: example – filetype:xls
- Search by domain: example – site:.gov
- Search by url: example – allinurl: ftp

Some folks have been asking me what the different file extensions mean. Here's a quick cheat sheet (Note: If you find an extension that you can't identify, just google "File extension .xxx" where "xxx" is the extension):

- Spreadsheet file
  - .xls – older versions of Excel
  - .xlsx – 2007 version of Excel
  - .ods – Open Office Calc
- Database file
  - .mdb – older versions of Access
  - .accdb – 2007 version of Access
  - .dbf - common database file
- Text files
  - .txt
  - .csv – comma separated value
  - .lst – usually a text file containing a list
  - .dat – data file
- PDF files - .pdf
- PowerPoint files - .ppt for older versions and .pptx for 2007 version.

Fun with URLs

Once you discover data online from a government agency, be sure to pay close attention to the URL. For example, the following link provides a download of an Excel spreadsheet with federal contracts relating to Haiti earthquake relief:

https://www.fpds.gov/downloads/top_requests/Haiti_Earthquake_Report.xls

Notice key parts of the URL: "downloads" and "top_requests"

If you paste this link into your browser and simply delete the URL a section or two, you might find more data from the FPDS. For example, if you back the link above one section you get this:

https://www.fpds.gov/downloads/top_requests/

It takes you to a page with the top requests from the Federal Procurement Data System. It includes hurricanes Katrina and Rita contracts, top 100 contractors by fiscal year, and more.

Index of /downloads/top_requests

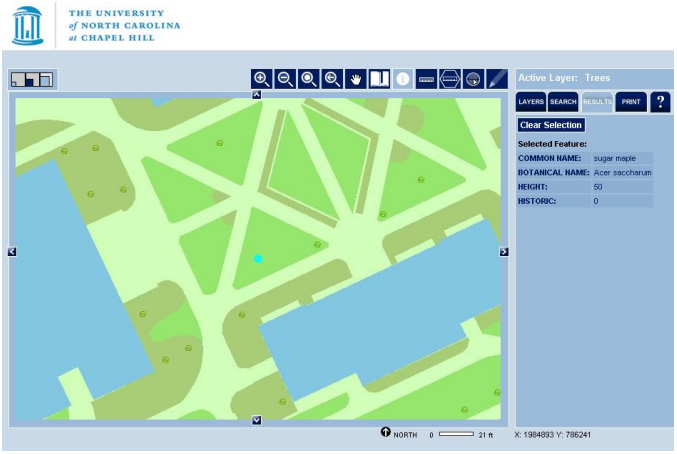
Icon	Name	Last modified	Size	De
[DIR]	Parent Directory		-	
[ ]	katrina_contracts.xls	09-Mar-2010 11:32	20M	
[ ]	rita_contracts.xls	09-Mar-2010 11:19	3.0M	
[ ]	American_Samoa_Earth..>	09-Mar-2010 10:43	109K	
[ ]	Haiti_Earthquake_Rep..>	09-Mar-2010 10:41	116K	
[ ]	TAS_Report.xls	09-Mar-2010 09:28	36M	
[ ]	TAS_Report_03_02_201..>	02-Mar-2010 08:53	35M	
[ ]	other_disaster_contr..>	22-Feb-2010 15:26	10M	
[DIR]	TAS_Report/	02-Feb-2010 10:44	-	
[ ]	FPDSNG_Contracting_O..>	14-Jan-2010 11:14	2.1M	

Mining map services

Some believe GIS data should and will be free.

- Web 2.0 Geoservers will make GIS knowledge widely available for mashups.
- Open standards, RESTful interfaces, and formats like kml, georss & json, will make the data, etc readily exchangeable.
-

We may not be there yet. But whenever you see an interactive map on the Web, there may be a map service that also is being published.



Take this example, at http://www.maps.unc.edu/uncmaps/default.aspx. The map is zoomed into the front of the J-School building, & a maple tree is selected. It's based on an ArcIMS map service

In ArcMap, you can add a GIS Server for the url domain. Then you'll see all the available map services, and their layers.

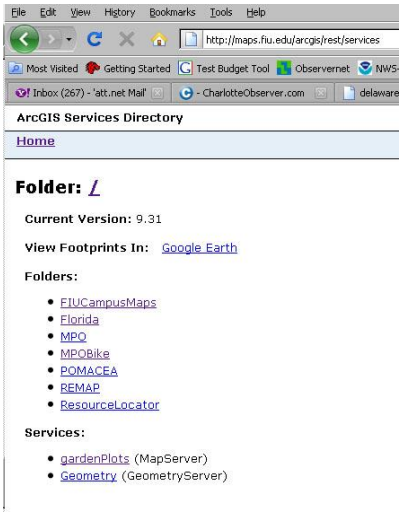
Selected Attributes of Trees

COMMONNAME	BOTANICALNAME	REMOVE	TYPE	DBH_IN	HEIGHT	CANOPY	CAMPUSZONE
		<Null>	D	2	12	2	SB
		<Null>	D	1.4	9	4	SB
red maple	Acer rubrum	<Null>	D	12.6	30	18	SB
flowering cherry	Prunus sp.	<Null>	D	10.7	20	16	SB
nelle stevens holly	Ilex x "Nellie Stevens"	<Null>	E	6	10	6	SB
american holly	Ilex opaca	<Null>	E	12.4	20	7	SB
flowering cherry	Prunus sp.	<Null>	D	14	20	19	SB
pecan	Carya illinoinensis	<Null>	D	27.6	70	31	SB
sugar maple	Acer saccharum	<Null>	D	19.5	50	25	SB
lusterleaf holly	Ilex latifolia	<Null>	E	20.799999	30	13	SB
southern magnolia	Magnolia grandiflora	<Null>	E	15.5	30	13	SB
sugar maple	Acer saccharum	<Null>	D	27.1	50	30	SB
willow oak	Quercus phellos	<Null>	D	3.5	20	9	SB
willow oak	Quercus phellos	<Null>	D	5.8	30	11	SB
american holly	Ilex opaca	<Null>	E	7	10	8	SB
american holly	Ilex opaca	<Null>	E	7	10	8	SB

Record: 1 Show: All Selected Records (16 out of 2000 Selected) Options

- The attribute table for the map service tree layer has columns not available through the Web map.
- You can use ArcMap tools to do searches not available on the Web version.
- You can export data.

Respect customary Web-scraping courtesies. Don't download too much or too fast.



Another kind of map service can be explored through a browser. You can see info on layers & attributes. Preview layers. Query layers, if allowed. Get query results in kml or json.

To find map services: Try a Web map domain. Use View | Source to look for URLs. Use Firebug to spot map services being queried.