John Perry

Center for Public Integrity Policy

Data Accuracy

To ensure data accuracy and the integrity of the process, the following mandatory steps will be required of all Center staff for all projects and stories that we publish. We believe that they are absolutely necessary and will be enforced.

1. Anyone working on a story or project involving a data set must consult the database editor, research editor, managing editor and managing director about methods and plans before any of the data is gathered or input.

2. The database editor should, in conjunction with the research editor and the project manager, design a plan to check the data according to agreed upon methods once the data is gathered.

3. Someone not involved with the story at any phase should have autonomy in fact-checking.

4. Fact-checking documents must include print-outs of the source documents (even when those documents are in a database) unless there will be more than 200 source documents. In such cases the database editor and the lead fact-checker are to be consulted about a suitable alternative to facilitate fact-checking.

5. All outlying statistics (those that are more than 50% higher or lower than the next nearest figure in a grouping) will be checked by hand against source documents by someone not on the team doing the story.

6. All rankings list entries will be compared against source documents. All ranked parties listed or mentioned must be contacted for quality assurance verification, though not necessarily for comment.

7. Databases will be checked by the database team staff or reviewed line by line by a staffer not involved in entry of or reporting on the data.

8. Databases entered by the Center staff or other nongovernmental entity are not to be treated as primary source documents. When using electronic data, there is no substitute for reviewing actual documents and comparing them against the electronic dataset.

9. The copy editor will verify with the database editor and research editor that steps 4-7 have been completed for each statistic used in a Center story and that they document their verification of data for inclusion in the fact-checking files.

10. Only the copy editor, upon instruction from the managing director, will tell the Web team when a Center story is ready to be published online. No story will be shared or promised to be delivered at a specific time before final edits from the managing director or executive director have been made it and the piece has been signed off on by the copy editor.

11. When the copy editor, fact checker or database editor is sick, on vacation or out of the office, their duties will be undertaken only by those who have been trained and briefed by them to do so.

12. Any variation from this protocol must be approved by the managing director in conjunction with the managing editor, database editor, research director and copy editor.

Official Policy of the Data Cave Regarding Data Collection and Analysis

The policies and procedures outlined in this document address the ways in which the Data Cave deals with three broadly defined sets of data. Procedures for entering, importation, cleaning, coding and updating data in a database are addressed in this document. To the extent that data is collected using scripts, data collection is addressed. But this document intentionally does not address data collection for the other types of data.

It is the responsibility of the Data Cave in conjunction with the project manager to determine the best method for collecting data on a project by project basis and a formal agreement must be reached and documented before the data collection begins.

Data Sources:

1. *Regularized Data Sets:*
   These are actual databases maintained by the government, in a formalized, documented structure that is available for inspection by the Center. Considered primary source material.
   EXAMPLE DATA SETS: Contracts, Census, FEC, FAADS, 527.

2. *Non-regularized Data Sets:*
   Let's call this the "data drop box." This form of disclosure is not collected and maintained in a consistent format. It also has no documented structure that is available for the Center to inspect. Instead, there are statutes, regulations or other forms of guidance that govern the disclosure and must be interpreted by the Center. While the data is available for the public to view and use, there typically is no regulator who vouches for the veracity of the data. It is merely a dumping ground for forms or electronic data. Also considered primary source material, but demands a higher level of scrutiny in the Center's data analysis and fact checking processes.
   EXAMPLE DATA SETS: Lobby, Trips, Personal Financial Disclosure, USAID spigots, Paper-based campaign finance disclosure.

3. *Unique Data Sets:*
   Any data creation and/or analysis done at the Center outside of the Data Cave and anything outside the purview of the first two categories. Data or statistics gleaned from non-governmental sources or governmental sources not made publicly available by the government in any way shape or form will never be considered primary source material and may only be cleared for publication when verified against the Center's strictest standards of scrutiny.

Data Entry:

Analyzing hard to obtain data is an important part of the Center's work, but must be approached with the highest amount of respect as the reputation of the Center as a whole rests on the veracity of its work.

Any data converted from hard copy by employees of the Center (even an electronic hard copy such as a PDF file), or scraped from Web pages by Center staff for importation into a database must be reviewed line by line against the source material to verify the integrity of the data before analysis. There must be another round of verification done for fact checking before publication by staff not associated with the data creation and analysis. The data will not be considered ready for formal analysis until the initial review is completed.

Such data may be used for reporting purposes without the level of scrutiny described above, but ***only*** with the explicit understanding that any statistic coming from that data set will be considered inaccurate until every supporting piece of documentation that encompasses that stat is checked. If a reporter wishes to pursue a story based on information gleaned from such data, then he or she will bear the burden of verification and will consult a member of the data cave on how to document and verify the information. Wording matters here. If words such as "most" are used in any statistic referencing this kind of data, then every single record in the database must be checked. It is the duty of the Data Cave staff to clearly articulate this burden to the reporter and document in an email or another verifiable form that the reporter is aware of their responsibility.

If an outside vendor—most notably Secure Paper Solutions—provides the Center with data entry services, Data Cave employees must complete the following procedures before the data is considered ready for formal analysis. First, a Data Cave member must sign off on the data entry template.

Upon receipt of digitized data from the outside vendor, the importation and any cleaning should be memorialized in a MySql or Perl script to ensure that the data can be restored to its original state in case of catastrophic failure of some sort.

After the data is in the database, it should be checked for duplicate records and any systemic problems. Results of routine queries must be checked against the source documentation. For instance, a campaign finance filing should be summed up and checked against the original filing. If the summary total on the paper form does not match the sum of the itemized records in the database, a sufficient reason should be supplied as to why the totals do not match. The data in the database must be verified against the source documentation.

If the following steps have not been performed, then the data staff has not signed off and the data cave will not release it for formal analysis. Any member of the staff using the data for reporting purposes must vouch for the fact that it has been released by the data staff for formal analysis each and every time the data is used as it may have been updated since its last use.

Data Importation:

Regularized data sets collected and maintained by the Center require a formalized importation process, preferably performed by scripting methods, which are documented and archived in a binder and a secure location on the network. Data cleaning and stored procedure documentation will accompany this documentation.

Any other data set that can be imported from its raw format into a database should also be memorialized in a MySql or Perl script.

Data Cleaning:

It is the responsibility of the data staff to fully understand the data being used, therefore data staff must completely report out and document what the data shows and how it is compiled for use by Center employees. If other staff members are involved in this process, a member of the data staff should serve as the point person in compiling the data documentation.

All databases must have an accompanying data dictionary that describes what the data tables show, how they are linked and any limitations inherent to the data should be described here. This documentation will reside in a binder and must be securely archived on the network. Any stored procedure documentation created for any Web publication using a particular database must be filed in the binder for each database used, and must added to the network archive for each database used.

Data Coding:

Every database requires standardization of one form or anther. That standardization seeks to identify companies, individuals or other entities in a consistent format in the database. Often times, staff outside the Data Cave must be used to complete this often onerous process. Therefore it is the job of the Data Cave, in consultation with the project manager if applicable, to task a staff member as the lead coding czar.

The coding czar and only the coding czar will consult with the Data Cave on coding issues and will supervise the coding team to ensure consistency and accuracy. The coding czar will check the work of the coding team, and then turn over the complete, checked work to the Data Cave who will then check the coded data again to ensure there are no egregious errors.

After this process is complete then the Data Cave will release the data for initial rankings and other analysis relying on the coding. The finalized rankings or other analysis will be finalized ***only*** when it has been verified according to the Center's Data Accuracy and Integrity Policy.

Data Updates:

After an update is performed on a database, the data must be re-verified according to the same level of scrutiny applied to it. In the case of ***non-regularized*** and ***unique*** data sets only the new, non-verified records must be checked, but if there is no fool-proof way to check only the new records, then the entire data set must be re-checked.

Data updates must be scheduled for *regularized data sets* on a recurring basis, conforming to the government's scheduled updates. Documentation describing the update process must accompany the Data Importation, Stored procedure and the Data cleaning documentation contained in both a binder and archived in a secure place on the network.

Methodology

For more than a year, researchers at the Center for Public Integrity developed plans to create a comprehensive federal lobbying project that coupled online resources with stories that examine under-investigated issues like compliance, the "revolving door" between government and the private sector and the vast influence lobbyists wield over the U.S. government. If you have any questions about the LobbyWatch project or process not addressed in this methodology, please contact Database Editor Daniel Lathrop.

How to Interpret the Figures on this Site

The Center has made every effort to ensure the accuracy of the numbers in this study. Despite this, there are certain to be occasional errors and inaccuracies in the figures we report on the LobbyWatch site. Among the reasons for this are: errors in the data as it was keypunched by the US Senate Office of Public Records, mistakes in the reports themselves and incongruity in the way organizations report their lobbying costs.

Of the data issues the Center has discovered, inconsistency in reporting semi-annual lobbying expenditures is the most common and consistent problem. Specifically, some organizations appear to pool the costs for in-house lobbying activities along with money they spend on outside firms as they are directed to do by the Senate's reporting guidelines. But in hundreds of other cases, organizations appear to be ignoring the guidelines by reporting just their internal lobbying costs.

Since there is no way to tell based on their filings whether organizations are reporting properly or not, the Center has developed a method to account for these apparent differences that results in a "best estimate" of the total each organization is spending on federal lobbying. Because of the conservative nature of the Center's approach, in many cases the figures significantly understate the actual amount of money certain organizations are spending. Because of this, the surest way to know what a company's full expenditures are on lobbying is to contact them directly.

Caution: This methodology has been revised since the initial release of this project in April 2005, so both the estimated aggregate total of lobbying as well as the totals for a number of individual organizations will have changed.

How the Center Arrived at These Totals

This method of calculating totals for each organization assumes that those groups that appear to be following the filing guidelines are doing so. Here's how it works:

The Center first tallies the totals each organization reports on forms it files with the Senate Office of Public Records for each six-month reporting period. That figure is compared to the aggregate total reported in the same period by outside lobby firms the organization hired.

If the organization's reported amount mathematically could include both its in-house and external lobbying costs, the Center assumes the report includes both and uses only that amount in its calculations. The money outside lobbyists report receiving from the company is ignored.

For those organizations that appear not to be following the guidelines—that is, an organization's self-reported total is lower than the total amount its lobbying firms report receiving that period—the Center counts only the outside money. The money the organization itself reported spending is ignored.

This ensures that there is no double-counting in the totals the Center reports. We'll continue to update these numbers and tweak the methodology over time.

How the Data was Collected

Researchers at the Center systematically downloaded every available lobbying disclosure record from the

Senate Office of Public Records, which, as of the beginning of this project included complete records covering the period between Jan.1, 1998 and June 30, 2004. Filings covering the second half of 2004·were due in February, but officials said they may not be completely available online until 90 days after the filing deadline. The totals for the second half of 2004 were added in late June 2005, after Senate officials told the Center data entry was complete for the forms received in February. From that point on, the Center has and will continue to update filings on a rolling basis and will when appropriate issue reports detailing overall activity. In general, it takes about three months for all filings given to the Senate in a timely manner to be posted.

In all, the Center downloaded 2.2 million records from the Senate's Web site which details information from just under 200,000 lobbying forms (and amendments) filed over the study period. These electronic records detailed amounts of money paid to lobbyists, the issues on which they lobbied, the agencies they lobbied and who hired them to do so. In addition, the primary place of business in the U.S. and the interests of foreign governments and foreign companies in those clients were disclosed in most cases.

Over a three-month period, Center researchers checked, cleaned and coded the data to make it searchable over time and consistent year to year. This, and the data analysis which followed, was supplemented with interviews with lobbyists, public interest groups and government regulators.

To conduct an industry-based analysis of lobbying trends, Center researchers assigned a trade-identifying code to each company or organization on whose behalf lobbying was conducted. The Center used a coding system originally developed by the Center for Responsive Politics, which has become something of an industry standard. This identification allowed the Center to analyze industry-wide spending trends and pinpoint the key lobbying spenders on specific issues. When company's had multiple business interests, these industry codes were based on a company's primary business.

To assess how much public money is spent within the federal lobbying system, the Center identified states, local governments, universities and other organizations operating as public entities that were registered as lobbying clients or as their own lobbyist between 1998 and mid-2004. Center researchers then analyzed the lobbying disclosure forms of these states, schools and local governments to total the lobbying spending on a national, state-by-state, and individual basis over the past six years.

In order to determine the number of former government employees now registered as federal lobbyists, the Center examined thousands of professional biographies posted on the Web sites of the top 250 lobbying firms. Although not all of the lobbying firms provided biographies, the majority did post detailed professional histories of their employees. This research was supplemented by a review of the federal lobbying disclosure forms filed with the Senate Office of Public Records. This proved less useful than biographies, however, because not all former government positions are required to be disclosed on the forms.

Researchers looked for any positions in both the executive and legislative branches, discounting positions as interns, consultants or members of advisory boards. To determine the number of former members of Congress now working as lobbyists, researchers additionally matched the names of former members of Congress born after 1890 with the names on federal lobbying registrations. Researchers then checked these matches against the law firms Web sites and lobbyist disclosure records to confirm the lobbyists' identities as former members.

Because the Lobbying Disclosure Act of 1995, which governs these reports, required companies to specify the amount they spent only when it exceeds $10,000 during a 6-month reporting period, many reports indicated only that "less than $10,000" had been spent.

The Center has treated those amounts a being the equivalent of $0 for the purposes of calculating money spent on lobbying.

This work was supplemented by additional research into public records and organizations' Web sites and public relations materials, regarding other lobbyists researchers believed may have been former officials.
Additional Notes

* In a very small number of cases, outside lobbyists will themselves subcontract with other firms to lobby on behalf of a particular client. Since the guidelines are unclear as to how this money should be reported, the Center does not include the subcontracting firm's total in the outside lobbying total for that client.

Power Trips
How We Did It
The anatomy of an investigation

By Daniel Lathrop

WASHINGTON, June 5, 2006 — Nine months of work, dozens of researchers, more than 26,000 documents and 7 million characters of data entry.

That's what it took to answer the question: Who is taking Congress for a ride?

Along the way, researchers from the Center for Public Integrity, American Public Media (producer of *Marketplace*) and Northwestern University's Medill News Service encountered all manners of misfiled, misreported and mystifying travel disclosures.

When a citizen, political action committee or lobbyist makes a contribution to an election fund, that information is reported to an independent federal agency, posted on the Internet and made available to reporters, researchers and the public.

But when the same people or groups pay for a "fact-finding mission," that information is put on paper forms, then filed in three-ring binders or input into a computer system, and made available only in the office buildings where the records are stored.

The House of Representatives' forms are kept in a sub-basement of the Cannon House Office Building, where the public copies were often hard to read, torn and misfiled. Researchers were told it was against House rules to digitally scan the documents — they had to make photocopies instead.

The Senate travel disclosure documents are stored in a computer system in the Hart Senate Office Building, and can be searched by the name of the traveler or the senator approving the travel. But those records are not available online. So researchers went to the building and printed them out.

Thus began an odyssey through the minutiae of congressional ethics rules, database software and company financial information.(See Methodology)

Along the way, we discovered trips for which no sponsor was listed; trips paid for by the federal government (not included in the totals of this report); trips for which no cost was listed — or with a reported $0 value, despite the fact that such trips do not require disclosure.

The overall numbers derived from the Center's study almost certainly are conservative. Always present was the concern that in having to photocopy and scan documents stored under less-than-ideal conditions, something had slipped through our fingers. Even without the possibility of our missing a document, the poor quality of the filings makes it hard to compare one member to another, one staffer to another, or one sponsor to another.

As a result, figures reported in this series were calculated with an eye to finding numbers that show floor totals rather than ceilings of what was spent taking Congress for a ride.

Power Trips

Methodology

WASHINGTON, June 5, 2006 — In a nine-month analysis of privately financed travel data based on sometimes incomplete congressional disclosure forms, the Center for Public Integrity and its partners took the following steps to ensure that the resulting data would be as true as possible to the forms filed by members of Congress and their staffers.

What we did:

- Drew up detailed templates for data entry teams at Secure Paper Solutions of Fredericksburg, Va.
- Scanned batches of trip forms by member (and, where possible, by year) into PDF files, which were transmitted electronically to servers at SPS.
- Once data entry was complete, downloaded data files from SPS and imported them into the Center's internal database server.
- Reviewed 30,000 pages of raw documents — including forms and attachments — and compared them with the data entered. Each apparent error or inconsistency was reviewed by a database editor.
- Analyzed traveler names to standardize them to 715 members and 5,945 staffers.
- Compared start dates and end dates of trips taken by people with the same or similar names to identify amendments and duplicate filings.
- Built a calendar of all trips taken by people with the same or similar names to identify travel that overlapped to identify amendments and duplicate filings.
- Analyzed start, end and signature dates and examined cases in which they were non-sequential or inconsistent, identifying additional amendments and data entry errors.
- Identified the sole sponsor reported on 23,380 forms out of 26,577 received.
- Attributed the trips to 3,208 organizations and identified an additional 265 forms listing no identifiable sponsor.
- Identified the remaining trips as having been co-sponsored by multiple organizations or sponsored by organizations that could not be identified as having financed multiple trips.
- Removed from analysis 82 forms reported to be sponsored by lawmakers' offices, Cabinet agencies or other arms of the federal government.
- Excluded from analysis all trips that did not begin between Jan.1, 2000, and June 30, 2005 (one trip which began within that period ended outside of it, on July 1, 2005).
- Adjusted totals to account for forms on which travelers incorrectly added subtotals in the space reserved for "Other" expenses. Leaving them in could have led to double counting.
- Set to zero all totals reported in foreign currency.
- Compared trips with identical costs to identify duplicates and amendments with name variations not previously identified.
- Consulted five years of congressional directories to resolve inconsistencies in the entry of staffers' names and identify whether chiefs of staff, administrative assistants and others whose signatures appeared approving forms were in fact supervised by the member under whose name the forms were filed.
- Compared signature data to the name of the member under which congressional officials filed the trip. This corrected 524 forms filed under the name of someone other than the member responsible for approving them.

- Checked incomplete dates and assigned them to a year. By default, they were assigned to the year in which they were filed (in the House) and then that date was adjusted based on the signature and date stamp dates on the forms themselves.
- Reviewed trips reported as having no cost to identify "advanced authorization forms" that reported approval of a trip rather than that a trip had been taken and paid for by an outside sponsor.
- Called congressional offices about hundreds of trips that appeared to have violated ethics rules. Changes were only made to these trips when specific data entry errors were identified in the process.
- Researched reported destinations to identify whether the destination was foreign or domestic.

What we did not do:

- Attempt to determine whether committee staff members' forms were signed bv the proper supervisor.
- Attempt to harmonize sponsorships (except in a few limited cases in which it was necessitated by additional reporting) when different staffers listed conflicting sponsors for what appeared to be the same trip.
- Attempt to determine the precise job titles and roles of travelers other than members.
- Attempt to modify the data based on interviews; the goal was to ensure that the data accurately reflected what was disclosed on the forms, not what legislators and staffers later said they had intended to disclose.
- Attempt to attribute sponsorship proportionally to the various sponsors of co-sponsored trips.
- Attempt to determine whether ethics rules required the filing of specific forms, such as those without substantial costs or with no travel cost listed.

— The Center for Public Integrity Data Team