CAR Wash Part I: Identifying problems in your data

Integrity checks for every data set

- Make sure you know how many records you should have and that you have them all. If someone in your newsroom tells you they have 65,536 records, be concerned. That is the exact limit of earlier versions of Excel, which many government agencies still use.
- Double-check totals or counts. Check for studies or summary reports.
- Check for duplicates. Make sure they are real duplicates. Is it possible that there are hidden duplicates?
- Consistency-check all fields. Are all city names spelled the same? Are company names spelled the same?
- What internal consistency checks need to be made? Is there more money going to sub-contractors than went to the prime contractor? Are there more teachers than students?
- How about other important fields? Check by running a GROUP BY and sorting alphabetically by every important field. Check it for spelling inconsistencies. For example, if you're analyzing a database of highway accidents, GROUP BY and sort ascending on the road name to check for inconsistencies.
- Other basic checks: make sure all states/cities/counties are included. Check the range of fields. (For example, check for DOBs that would make people too old or too young.)
- Check for missing data or blank fields. Are they real values, or did something happen with an import or append query?
- Do you know what every field in the database means? Are there special codes? (We had some data where missing was listed as 99 in the field, so that messed with totals.)

Beyond basic checks

- Keep a data notebook (or computer file) and write down everything you do.
- Know the source of the data.
- If you need to make sure it's solid, get similar data from another source.
- Create a back-up copy of the database.
- Check against reports.
- Make sure you're using the right tool. You may need to do more than counting and sorting.
- Check with experts from different sides of the issue.
- Find similar stories and study what they did. (IRE resource center is great for this.)
- Look at it. If you can actually physically go spot check records, do it.
- Don't forget the gut check. If something just doesn't seem right, it probably isn't.
- If you think you're in over your head, call on an expert to help. Do not run with scissors.

Find the right methodology

- Read research reports.
- Finding an existing data model - There are some accepted methodologies for dealing with certain types of data.
- Find an expert to bounce your methodology off during the process.
- Show findings to the targets of the story.
- Duplicate your work. To make sure you didn't mess something up along the way.

- Maintain a consistent universe of cases. If you have to filter or redefine your universe, be able to explain why you isolated certain records or cases.
- Give yourself enough time to follow through on collecting information for your database before you start writing. If you've built an organic database, where information may need to be updated or will change after additional reporting, set a cut-off date and don't make any more changes to the database unless the data is inaccurate or the new information will change the meaning of the story.

For more information

- *Numbers in the Newsroom: Using Math and Statistics in News* by Sarah Cohen for Investigative Reporters and Editors, Inc.
- *Precision Journalism* by Philip Meyer. Indiana University Press, Bloomington. 4th Edition. 2002.
- *News and Numbers* by Victor Cohn. Iowa State University Press, Ames. 1989.
- *How to Lie with Statistics* by Darrell Huff. W. W. Norton & Company, New York. 1954 (renewed 1984)
- *Innumeracy: Mathematical Illiteracy and Its Consequences* by John Allen Paulos. Vintage Books, New York. 1990.
- *A Mathematician Reads the Newspaper* by John Allen Paulos. Anchor Books, New York. 1995. (Also, check out the tape from Paulos keynote address at NICAR 2002 in Philadelphia)
- IRE Resource Center: www.ire.org

Danielle Cervantes contributed information for this tip sheet.

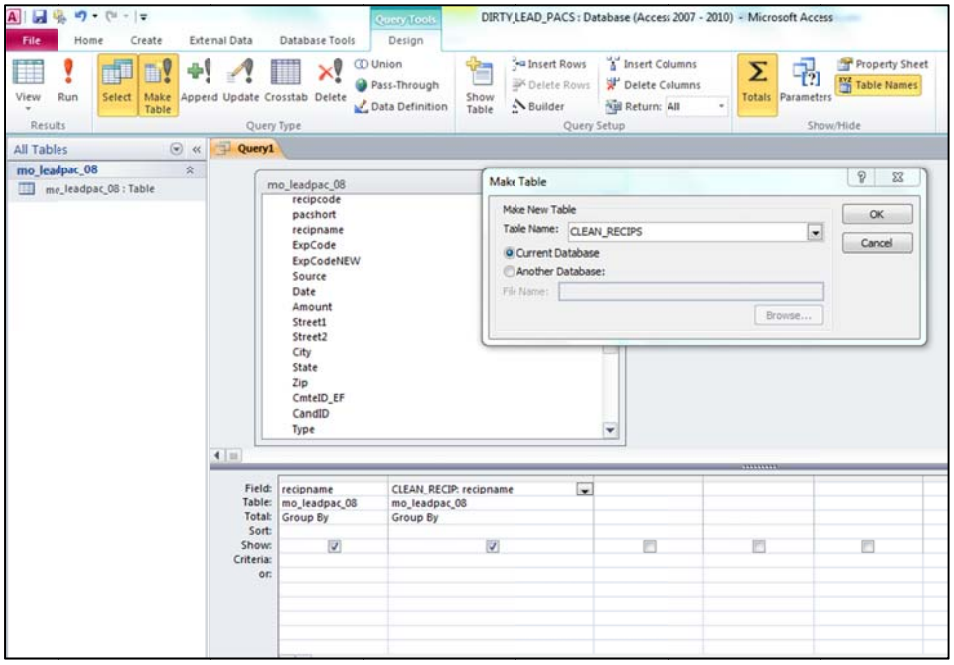
CAR Wash Part II: Cleaning your data

Problem: Inconsistent names/cities/counties/companies…
Solution: There are several approaches to fixing what is likely the most problem dirty data issue we run into. There are tools such as Google Refine (see separate class by Dan Nguyen and this tip sheet: http://www.propublica.org/nerds/item/using-google-refine-for-data-cleaning) Another new tool (courtesy David Donald) called Fuzzy Lookup, which is an add-on to Excel: http://www.microsoft.com/download/en/details.aspx?id=15011

We'll clean using our database manager.

Rule #1: Don't change existing data. You will regret it. Instead, create a new field called CLEAN_NAME where you will put your fixed up data.

In Access: Create a new table with the grouped results of the field you want to fix. Put the field in twice. Call the second field CLEAN_NAME.

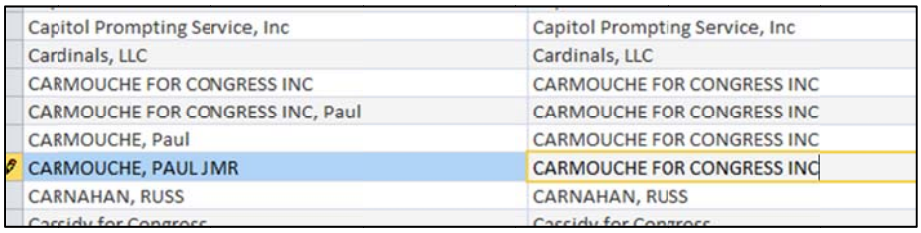


When you come to an inconsistency like this one:

Capitol Prompting Service, Inc	Capitol Prompting Service, Inc
Cardinals, LLC	Cardinals,LLC
CARMOUCHE FOR CONGRESS INC	CARMOUCHE FOR CONGRESS INC
CARMOUCHE FOR CONGRESS INC, Paul	CARMOUCHE FOR CONGRESS INC, Paul
CARMOUCHE, Paul	CARMOUCHE, Paul
CARMOUCHE, PAUL J MR	CARMOUCHE, PAUL J MR
CARNAHAN, RUSS	CARNAHAN, RUSS

Decide which version you want to use. In this case, let's use the first one: CARMOUCHE FOR CONGRESS INC

Correct only the second column, so that every version of CARMOUCHE in the field column has a consistent fix in the second column:



Capitol Prompting Service, Inc	Capitol Prompting Service, Inc
Cardinals, LLC	Cardinals, LLC
CARMOUCHE FOR CONGRESS INC	CARMOUCHE FOR CONGRESS INC
CARMOUCHE FOR CONGRESS INC, Paul	CARMOUCHE FOR CONGRESS INC
CARMOUCHE, Paul	CARMOUCHE FOR CONGRESS INC
CARMOUCHE, PAUL JMR	CARMOUCHE FOR CONGRESS INC
CARNAHAN, RUSS	CARNAHAN, RUSS
Cassidy for Congress	Cassidy for Congress

Use this table to join back to your original on the original name field and tell it to replace the contents of CLEAN_RECIP with the CLEAN_RECIP in the lookup.

Problem: Numbers don't match up
Solution: There's no easy fix for this other than going back to the agency and finding out what's wrong. There also may be another data source you can cross-check with that might give you more answers.

Problem: Duplicate records
Solution: Make sure they are real duplicates and not just that they appear to be. If they match on every field, just group by all fields to screen out filters, otherwise, you may have to take the "max" value of some fields. Don't just willy-nilly delete records. You will regret doing that. Instead, add a field called DUPE or something similar and mark Y to indicate it's a duplicate. Then when you run your final queries screen for that DUPE is null.

Problem: Missing data
Solution: This will often be the case. Get a sense of how serious the problem is. Go back to the agency. Find out if there is another source of the information that you could merge. For example, in the case of schools, we relied on the NCES schools data for filling in missing information and for double-checking the existing data.
We've also used Mechanical Turk to fill in missing information or to check records. (For more info see: http://www.propublica.org/article/propublicas-guide-to-mechanical-turk)

Problem: You don't know if there are problems
Solution: Pull a random sample of records to do spot checks on – sometimes even physical spot checks are necessary.

Problem: But there still may be problems
Solution: Include that caveat in your methodology – particularly for online databases and give folks a chance to correct the information. We did this with our national schools database. We then went back and verified with the district/school.