Do’s and Don’ts of Data Migration

January 2018

Introduction

Jon Goldberg, Megaphone Technology Consulting LLC

- I’m a non-profit CRM consultant, specializing in data migration
- Ask questions! I’ll handle them between slides, also we’ll save time at the end.
- This slide deck is available at https://www.megaphonetech.com/resources.

Today’s Topics

Big Topics

- Collecting your data together
- Trimming your data
- The data dictionary
- Cleaning your data

Little topics

- The tech purchase process
- Some tools you can use
- Odds and ends

Collecting your data together

- Which data, exactly?
  - For most of you, much of this is in an existing system
  - But also your spreadsheets, address books, custom databases, mailing lists
- Is there a common identifier across your different sources?
  - Member ID, email. First/last name if it’s small.
  - It doesn’t have to be the same identifier for all sources – just be able to link them all together.
- Secondary data sources are often the biggest part of a migration budget
- Get all your data together before you even think of moving forward

Trimming your data

- Data has two dimensions – *depth* and *width*.
- It’s easier to import 100,000 records with just first/last/email than 50 records with a rich history.
- You can reduce budget by cutting the number of contacts you move – but you can cut far more by reducing how much data you bring over.

Trimming your data - width

- Drop contacts you don’t interact with!
  - It costs time/money to keep them. What conversion rate justifies that cost?
- Maintaining contacts costs time/money. Save that for higher-value contacts.
- While reducing “depth” is more important – we can reduce the number of exceptional cases. More on this later.

Trimming your data - depth

- Drop fields you don’t use.
  - OpenRefine can help you ID these.
- Delete/consolidate tags (“attributes”)
- What data is still relevant?
  - Fulfilled pledges, Short membership lapses
- How important are unusual transactions?
  - In-kind donations, tributes, stocks
- Refunds, unfulfilled pledges, partial/multiple payments
  - Can you simplify?
- Don’t keep addresses you know are bad!

Trimming your data – the human factor

- This can be a contentious process!
- Someone who's been with your organization for 30 years has different senses of what's important than someone who's been there for 3. Both have a contribution to make.
- Mollify your "data packrats".
  - Consider how to keep your old data WITHOUT migration. E.g. exports to spreadsheet.
- Many folks will hate a system because it's new.
  - Negative sentiment about the migration process can doom your project. More on this later.

The Data Dictionary

- Your data is collected and trimmed – now what?
- You need to map your existing data to the new system.
- Here is a Google doc of a sample data dictionary.
- My color code:
  - Green: I have no questions about how to migrate this data.
  - Red: I will not migrate this data.
  - Yellow: This needs further questions/clarification.
- Try very hard not to introduce new fields during the migration process!

Data dictionary – how to construct

- A business person will guide me through the old database, pointing out the most critical business functions performed. I'll screenshot the screens and take notes.
- I will review the internal data structure and match it to the fields I see in the screenshots.
- I will make a list of all the fields in the legacy tables.
- I will export that data and run it through OpenRefine (we'll cover later).
- I will note fields which have obvious/easy parallels in the new database; I'll add that to the dictionary.
- I'll put questions and notes into the dictionary for other fields.

Pop quiz!

Which one of these is a valid value in your email address field?

212-555-1234
jon@yahoo
jon@gmail.com (wife is jane@gmail.com)
ljksdljsdasjdasdjasl

- Is your name “Raiser’s Edge” or “Salsa”? If so, then you answered “all of them”!
- Sometimes you don’t need to clean data for migration. When the new system’s data validation is better/different, you do.

Cleaning your data

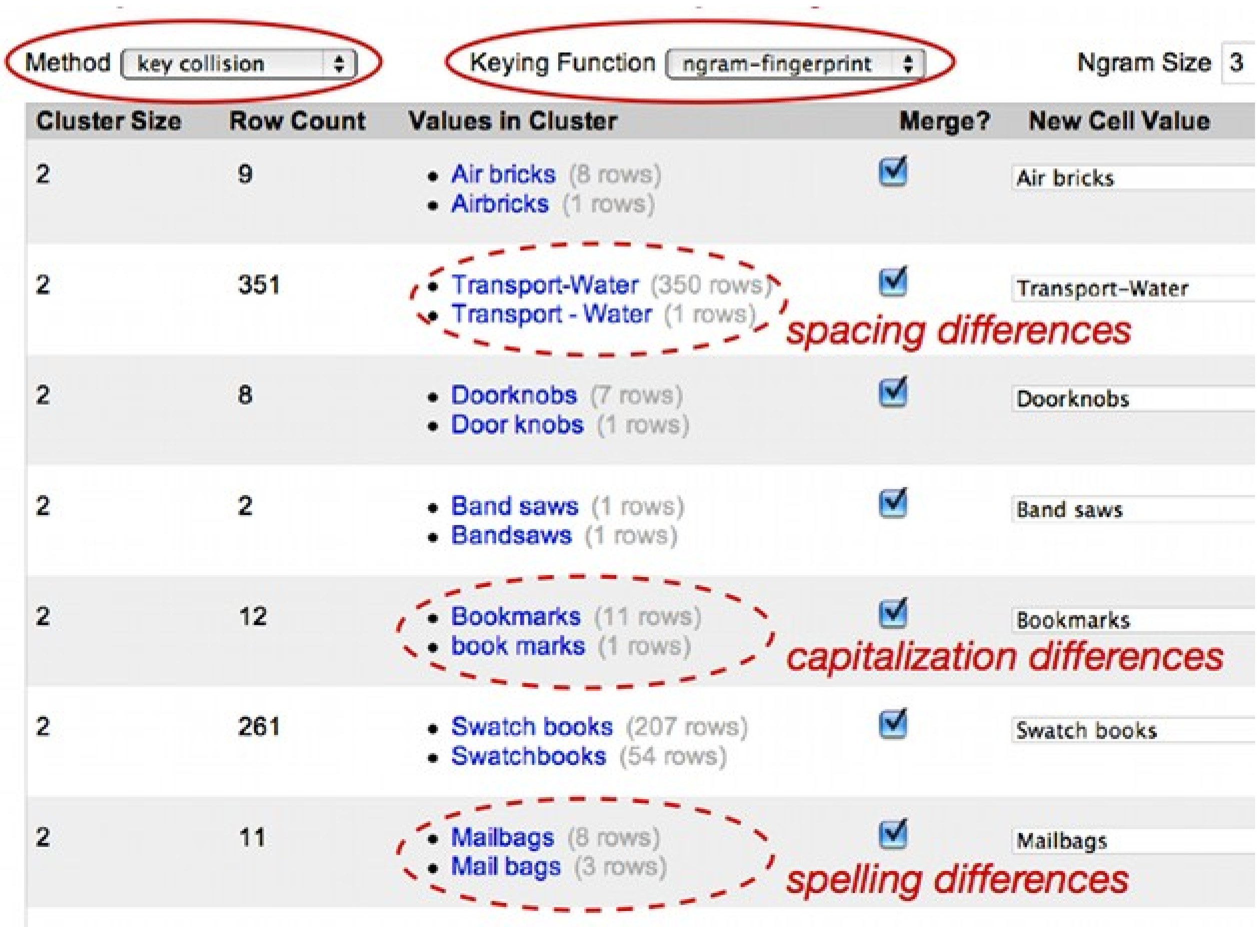
- IF YOU DON'T HAVE CLEAN DATA, PEOPLE WON'T TRUST THE NEW SYSTEM. People will blame the system, not the data.
- Some kinds of cleaning require knowledge of your data. Others not.
- International addresses – I’ve never seen folks get this right.
  - I’m not saying there’s no “Calgary, Alabama”, or “Tokyo, Japan, United States” - but there isn’t.
- If your new system sends email/SMS, cleaning notes and invalid values from email/phone is very important.
- If data can’t be cleaned, consider a read-only custom field for import.
- This data has been messy for a long time – do you actually use/need it?
- Often data is in the wrong field because of a messy previous import.

The tech purchase process/RFPs

- Good data conversion matters.
  - Automated tools help, but understanding how the data is used is important.
  - This is critical data; treat it as such.
- Get free help from Aspiration Technology.
- Get someone who understands data, but also people/internal politics.
- Cultural competency matters.
  - It only takes one slip-up by a “well-meaning” person to undermine staff’s trust in a new and complicated tool.

Some technical tools - OpenRefine

- Find values with spacing/capitalization differences.
- Non-numeric values in numeric fields
- Values that fall out of a range – e.g. that check received in “2099” probably arrived in 1999.



Method: key collision Keying Function: ngram-fingerprint Ngram Size: 3

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
2	9	• Air bricks (8 rows) • Airbricks (1 rows)	☑	Air bricks
2	351	• Transport-Water (350 rows) • Transport - Water (1 rows)	☑	Transport-Water
2	8	• Doorknobs (7 rows) • Door knobs (1 rows)	☑	Doorknobs
2	2	• Band saws (1 rows) • Bandsaws (1 rows)	☑	Band saws
2	12	• Bookmarks (11 rows) • book marks (1 rows)	☑	Bookmarks
2	261	• Swatch books (207 rows) • Swatchbooks (54 rows)	☑	Swatch books
2	11	• Mailbags (8 rows) • Mail bags (3 rows)	☑	Mailbags

spacing differences

capitalization differences

spelling differences

Some technical tools - ETL

- “Extract” (from old data source), “Transform” (to preferred format), “Load” (into new data source).
- In “naive” migrations, someone exports your data, munges it in Excel, and imports it.
- In the meantime, you’ve added new records, updated old ones, fixed mistakes, and told the migrator they did parts wrong. Start over?
- With ETL, you *script* the migration process, so you can run it over the original data source multiple times.
  - See how it looks on your latest data set quickly.

Continuous Migration

- ETL enables you to migrate without abandoning your old database.
- A script that will just pull in new/changed data into the new system hourly/daily.

Odds and ends.

- Recurring donations are awful!
  - Most recurring donation tools (PayPal, most versions of Authorize.Net, etc.) work by “IPN”. These just don't carry over.
  - Depending on how many you have, you may want to manually import every month, ask folks to give you a new card, or something else.
  - There's a new generation of payment processors ("token-based") that will hopefully eliminate this problem as folks transition to them.
- Consider your special use cases.
  - E.g. if your organization spans multiple timezones, make sure you're migrating with the correct set of timezone assumptions in place. Same for internationalization concerns.
- File attachments are the bane of conversions! Many CRMs don't offer an export tool for attached images, PDFs, etc. It's not ALWAYS impossible, but it's often unpleasant.

Thank you!

Links to resources I mentioned:

- https://www.megaphonetech.com/resources/
- http://openrefine.org/
- http://www.pentaho.com/product/data-integration - my ETL of choice
- https://github.com/MegaphoneJon/civicrm_kettle_transforms - links to sample ETL scripts for Constant Contact, DonorPerfect, Nationbuilder, Raiser's Edge, Salsa, and Wild Apricot.

More questions?

- jon@megaphonetech.com
- https://www.megaphonetech.com