When One Company Has the Data Our Researchers Need

Working with **CoreLogic®**
to Purchase US Tax, Deed,
and Non-Agency Backed Loan Data

Beyond the Numbers, Oct. 5, 2016

Adrienne Brennecke, Federal Reserve Bank of St. Louis

Kris Kasianovitz, Stanford University Libraries

STANFORD UNIVERSITY LIBRARIES

FEDERAL RESERVE BANK of ST. LOUIS
CENTRAL to AMERICA'S ECONOMY®

Disclaimers

The views expressed are my own and not the views of the Federal Reserve Bank of St. Louis or the Federal Reserve System

The views expressed are my own and not the views of the Stanford University

Agenda

- A Unique Data Resource

- St. Louis' Story

- Stanford's Story

- Lessons Learned

- Negotiation Strategies

- Resulting Research

A Unique Data Resource

- Public Record Data
  - County Assessor's Taxroll Data
  - County Clerk/Recorder Deed Data
  - AKA Parcel Data (attribute data not spatial data)

- Loan Level Market Analytics
  - Loan origination data from major US Banks

http://www.corelogic.com/solutions/loan-level-market-analytics-for-capital-markets.aspx

St. Louis' Story

- December 2010
- Accidental data librarian
- Focused on acquiring the resource, documenting and communicating terms of use
- Two different economists each wanted a separate custom dataset
- Started getting questions in October 2011
- Amended contract several times to get more variables and correct geographic coverage.

Making the initial request

- Research Question?

- 10 MSAs; 'all available fields'; 2003-2010

    Owner name and address, Sales Date, Sales Price, property characteristics (sq. feet, bedrooms, parking type, pool)

Amending the contract

- **Bulk Data I**: historical data 2003-2010, 10 MSAs, for the layout shown; wanted 'all available fields' from property records.
- Bulk Data II: updating separate dataset (1990) different geography.
- **Bulk Data III**: Updates Bulk Data I; then started to realize they didn't have what they needed.
- **Bulk Data IV**: Longer history back to 1990; More variables, including latitude and longitude and foreclosure-related variables, different geographies. Data from 2 different layouts – no way to connect them.
- **Amendment 7** added more variables, including unique identifier
- **Amendment 8** updated geography due to miscommunication

Geography problems

	STATE	COUNTY	Fips Code	Sale History
St. Louis	IL	Jefferson	17081	1999
Sacramento	CA	El Dorado	6017	1946
	CA	Placer	6061	1972
	CA	Sacramento	6067	1965

- Core Based Statistical Area (CBSA) versus Metropolitan Statistical Area (MSA)
- Postal Database versus Census

Foreclosure coverage

- No foreclosures in Denver?

- Lots of fields apply to foreclosures – maybe we just needed different ones.

- Requested layout for all available foreclosure-related fields

Amendment 8

St. Louis, MO-IL
Jefferson County, **MO**
Jersey County, IL
Lincoln County, MO
Sacramento-Arden-Arcade-Roseville, CA
El Dorado County, CA
Placer County, CA
Sacramento County, CA
Yolo County, CA

Epilogue

- Research Analyst left in May 2013

- 2014-2015, Federal Reserve System jointly acquired CoreLogic Real Estate data

- Housed in an internal Data Warehouse

- Staff of data experts worked to clean up the data and present it in an easy-to-use query tool

- This comprehensive collection supersedes our custom data set

Stanford's Story

"I am currently working on the reshaping of the American metropolis after **the subprime crisis**.

I am mainly interested in the evolution of the real estate dynamics, notable through the tenancy and the evolution of the housing price. My research focuses on the Northern San Joaquin Valley in California, but I am also interested in the 20 biggest MSA. Data Quick data would be extremely useful for my research, because **I would be able to access very accurate, homogeneous and reliable statistics on a long period of time**."

Making the initial request

- Driving forces:
  - Several faculty, graduate students, and Post-Doc's needing parcel level data (Tax and Deed data), initially for 3 counties in California
  - County Assessor and Clerk Recorder Offices manage these records in all types of ways, which impact ease of access, especially to longitudinal data
  - Started with DataQuick then moved to CoreLogic (a merger and need for loan data)
  - Word got out! Research team in Management Science and Engineering needing parcel level data AND non-agency backed loan data

Making the initial request

- What
  - Public Record data = Tax Assessor and County Clerk Recorder
  - All counties in the US*
  - Historical up to 2014Q4
  - No ongoing updates*
  - Campuswide access
- Who
  - MSE, Law, Economics, Environmental Studies, Business School Faculty and graduate students
  - Subject Librarians, Library License Coordinator, Library Directors
  - Single Contact at CoreLogic (Government Solutions)
- How
  - Various funding sources
  - Year-long negotiation process

Negotiation Process and Data Issues

- We did negotiate the price
- We did negotiate access for co-authors – do not have to be at Stanford
- Reliant upon researchers for content analysis/determination
- Codebooks before signing the license
- We have a point of contact for questions, and got support written into our Statement of Work; Turnaround time for questions
- Comingling issue and EULA
- Serving the data

A Few More Data Issues

- Researchers determined the data we purchased was not complete once we took delivery
  - Missing files – resulted in having several new files to be merged
  - Lat/Lon file ended up being at the block not parcel level (fixed)
  - FIPS code issues
  - APN issues
- Not for the faint of heart data user

Lessons Learned

- **Don't assume:**
  - the sales rep knows the particulars of the data. Find someone who does before signing the contract.
  - the researcher fully understands the data, where it comes from, and what is available from the vendor.
  - the layout in your contract is accurate.
- Examine the data when it is delivered.
- Add a clause into your license to ensure data quality and a level of support
- Ask researchers what they hope to learn from the data; have a better sense of their projects and goals. How long are they going to need the data?
- Name custom data sets something meaningful in the contract.

Lessons Learned

In addition to FRB Lessons Learned…

- Be prepared for a lot of back and forth, especially if you are in an academic setting
- Try to find someone who has licensed the data or worked with the data provider to get insight
- Strategically involve the researchers in the negotiation process
- There will be a lot of issues to sort through once you get the data
- Keep a "paper trail" and create read.me files for the data
- Don't be afraid to send the rep questions about the data
- Provides a great opportunity to push data access, delivery, archiving of large data sets at your library

Negotiation strategies

- Prestige Factor – other researchers will see that we're using this data
- Plead your case - We don't have as much money as you think we do
- Mention competitors pricing; Zillow licenses content from CoreLogic
- Multi-year contract; research takes time
- Big bundle; multiple datasets

St. Louis Publications

Semi-Parametric Interpolations of Residential Location Values: Using Housing Price Data to Generate Balanced Panels

by Jeffrey P. Cohen, Cletus C. Coughlin, and John M. Clapp

https://research.stlouisfed.org/wp/more/2014-050

Local Polynomial Regressions versus OLS for Generating Location Value Estimates: Which is More Efficient in Out-of-Sample Forecasts?

by Jeffrey P. Cohen, Cletus C. Coughlin, and John M. Clapp

https://research.stlouisfed.org/wp/more/2015-014

Stanford Uses of the Data

- Subprime Mortgage Crisis
  - Foreclosures, delinquency
  - Stochastic modeling of mortgage default
- Regulatory issues with coastal properties
- Relationship between city density and energy consumption
- Sales prices from the last twenty years to study the effect of floods and hurricanes on real estate markets
- Mathematical modeling projects
- Courses needing a large data set to practice data cleaning, extraction, analyses
- Several Dissertations Forthcoming!

Stanford Uses of the Data

- Ohlrogge, Michael and Giesecke, Kay, Securitization and the Growth of Subprime Mortgage Lending (May 17, 2016).

  Available at SSRN: http://dx.doi.org/10.2139/ssrn.2781060

- Giesecke, Kay; Sirignano, Justin; and Sadhwani,Apaar. Deep Learning for Mortgage Risk. (July 7, 2016). Working paper.

  Available at: https://stanford.app.box.com/s/wubluayywt2d41tgg9bwuxmic7nsuvwx

Thank you & Questions

Adrienne Brennecke, adrienne.j.brennecke@stls.frb.org

Kris Kasianovitz, krisk11@stanford.edu