Automating maintenance of public sector survey frames

Keith Finlay and Elizabeth Willhide

U.S. Census Bureau

September 23, 2019

Disclaimer: This presentation is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the author and not necessarily those of the U.S. Census Bureau.
Frame maintenance is costly

- Public-sector (criminal justice) collections in the Economic Reimbursable Surveys Division
  - Staff manually update frame information, hand match new tables to existing frames
- Much frame information is publicly available on the web
  - Use web scraping to harvest unstructured data
- How can we reduce the labor required to integrate new frame tables/data?
  - Document parsing, fuzzy matching tools
Civic Digital Fellows developed modular system

- Worked on this system the last two summers
- Carlos Ortega, a 2019 Civic Digital Fellow, helped us break the problem into smaller pieces
- Everything uses open source tools
- A number of the modules can be used for further process automation in our division
Documents with possible frame units

New frame tables without frame ids

Filter for relevance

Parse documents

Fuzzy match candidates to frame

Review frame updates

Frame table

Filter out public units

Filter for relevance

Parse documents

Fuzzy match candidates to frame

Review frame updates

Frame table

Filter out public units

Web crawl/scrape

Validate URLs

Google APIs identify frame unit URLs

Data

Automation

Human review

URLs to discover new frame units

Web crawl/scrape

Validate URLs

Google APIs identify frame unit URLs

Filter out public units

CENSUS.GOV
Documents with possible frame units

Sponsor delivers new frame tables without frame ids

Frame inputs from sponsor

Filter for relevance

Parse documents

Fuzzy match candidates to frame

Review frame updates

Frame table

Filter out public units

URLs to discover new frame units

Web crawl/scrape

Validate URLs

Google APIs identify frame unit URLs

Data

Automation

Human review
Documents with possible frame units

Sponsor delivers new frame tables without frame ids

Parse documents

Fuzzy match candidates to frame

Review frame updates

Frame table

Filter for relevance

Web crawl/scrape

Validate URLs

Google APIs identify frame unit URLs

Filter out public units

URLs to discover new frame units

Data

Automation

Human review

URLs to discover new frame units
Documents with possible frame units

Sponsor delivers new frame tables without frame ids

Filter for relevance

Parse documents

Fuzzy match candidates to frame

Review frame updates

Frame table

URLs to discover new frame units

Web crawl/scrape

Validate URLs

Google APIs identify frame unit URLs

Filter out public units

Data

Automation

Human review
Documents with possible frame units → Parse documents → Fuzzy match candidates to frame → Review frame updates → Frame table

Sponsor delivers new frame tables without frame ids

Filter for relevance → Web crawl/scrape → Validate URLs → Google APIs identify frame unit URLs → Filter out public units

Data → Automation → Human review

Filter for relevance

URLs to discover new frame units

Web crawl/scrape

Validate URLs

Google APIs identify frame unit URLs

Data

Automation

Human review

CENSUS.GOV
Documents with possible frame units

Sponsor delivers new frame tables without frame ids

Filter for relevance

Parse documents

Fuzzy match candidates to frame

Review frame updates

Frame table

Filter for relevance

Parse documents

Fuzzy match candidates to frame

Review frame updates

Frame table

Google APIs identify frame unit URLs

Filter out public units

URLs to discover new frame units

Web crawl/scrape

Validate URLs

Google APIs identify frame unit URLs

Filter out public units
Documents with possible frame units

Sponsor delivers new frame tables without frame ids

Parse documents

Fuzzy match candidates to frame

Review frame updates

Validate URLs

Google APIs identify frame unit URLs

Filter out public units

Web crawl/scrape

Filter for relevance

URLs to discover new frame units
Documents with possible frame units

Sponsor delivers new frame tables without frame ids

Filter for relevance

Parse documents

Fuzzy match candidates to frame

Review frame updates

Frame table

Filter out public units

Data

Automation

Human review

URLs to discover new frame units

Web crawl/scrape

Validate URLs

Google APIs identify frame unit URLs

Web crawl/scrape

Parse documents

Fuzzy match candidates to frame

Review frame updates

Frame table
Documents with possible frame units

Sponsor delivers new frame tables without frame ids

Parse documents

Fuzzy match candidates to frame

Review frame updates

Data

Automation

Human review

Filter for relevance

Parse documents

Fuzzy match candidates to frame

Review frame updates

Frame table

Filter out public units

URLs to discover new frame units

Web crawl/scrape

Validate URLs

Google APIs identify frame unit URLs

Web crawl/scrape

Validate URLs

Google APIs identify frame unit URLs
Sponsor delivers new frame tables without frame ids

Documents with possible frame units

Parse documents

Fuzzy match candidates to frame

Review frame updates

Frame table

Filter for relevance

Validate URLs

Google APIs identify frame unit URLs

Filter out public units

URLs to discover new frame units

Web crawl/scrape

Data

Automation

Human review

Parse documents

Fuzzy match candidates to frame

Review frame updates

Frame table

Filter out public units

Parse documents
Documents with possible frame units → Parse documents → Fuzzy match candidates to frame → Review frame updates → Frame table

Sponsor delivers new frame tables without frame ids → Fuzzy match candidates to frame → Filter for relevance

Filter for relevance → Parse documents → Validate URLs

URLs to discover new frame units → Web crawl/scrape → Validate URLs

Validate URLs → Google APIs identify frame unit URLs → Filter out public units

Data

Automation

Human review

Google APIs identify frame unit URLs → Filter out public units
Next steps

- Improve accuracy of web scraper, expand scope
- Improve speed and effectiveness of frame matcher, move to server
- Move system into production
  - Has potential to reduce 4 weeks of 3 FTE down to days
  - Staff can be reallocated to response, quality improvements
Thank you

Keith Finlay
keith.ferguson.finlay@census.gov