

Automating maintenance of public sector survey frames

Keith Finlay and Elizabeth Willhide

U.S. Census Bureau

September 23, 2019

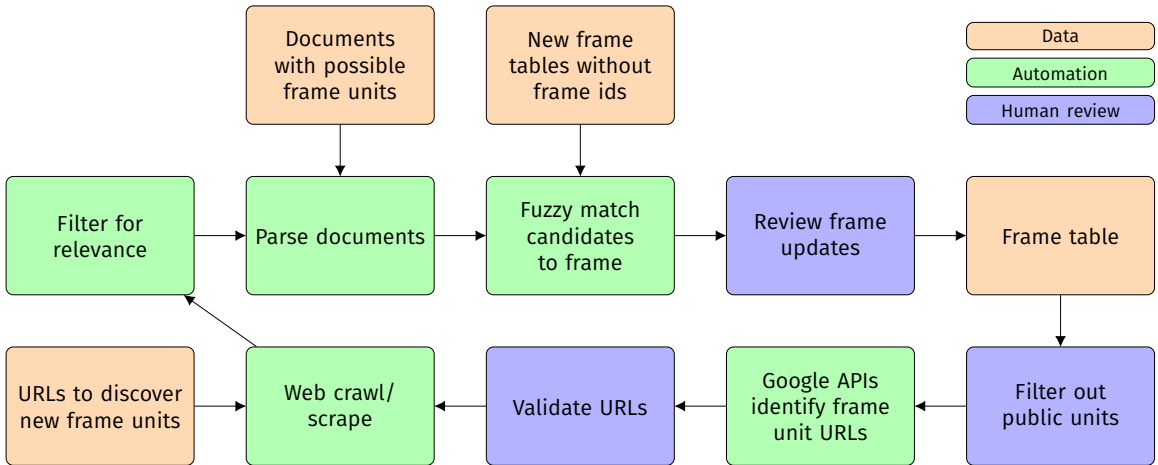
Disclaimer: This presentation is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the author and not necessarily those of the U.S. Census Bureau.

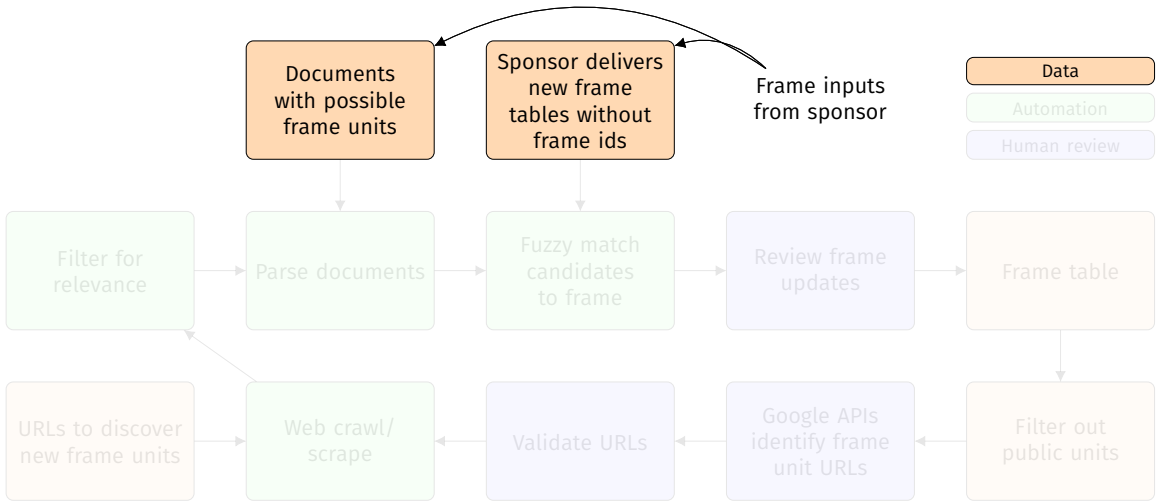
Frame maintenance is costly

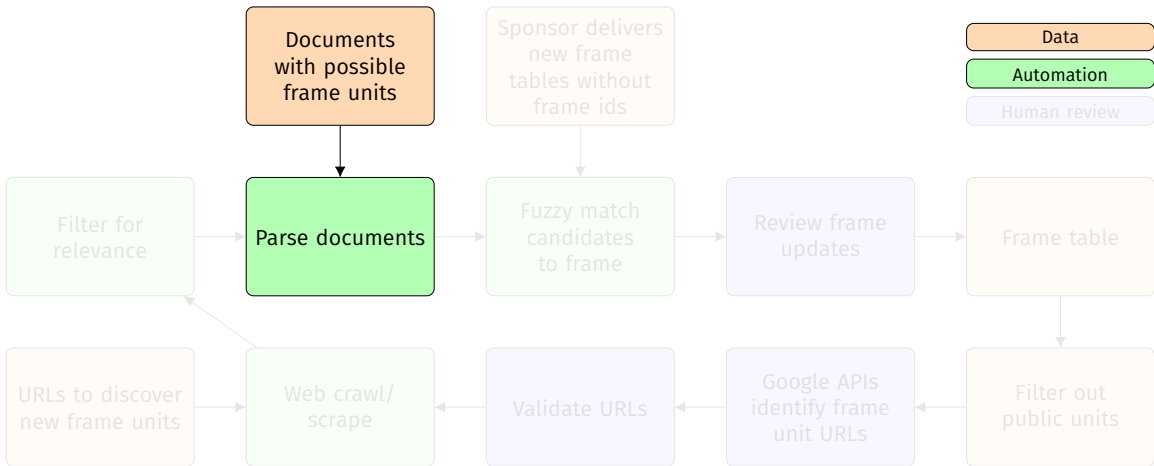
- ▶ Public-sector (criminal justice) collections in the Economic Reimbursable Surveys Division
 - ▶ Staff manually update frame information, hand match new tables to existing frames
- ▶ Much frame information is publicly available on the web
 - ▶ Use web scraping to harvest unstructured data
- ▶ How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

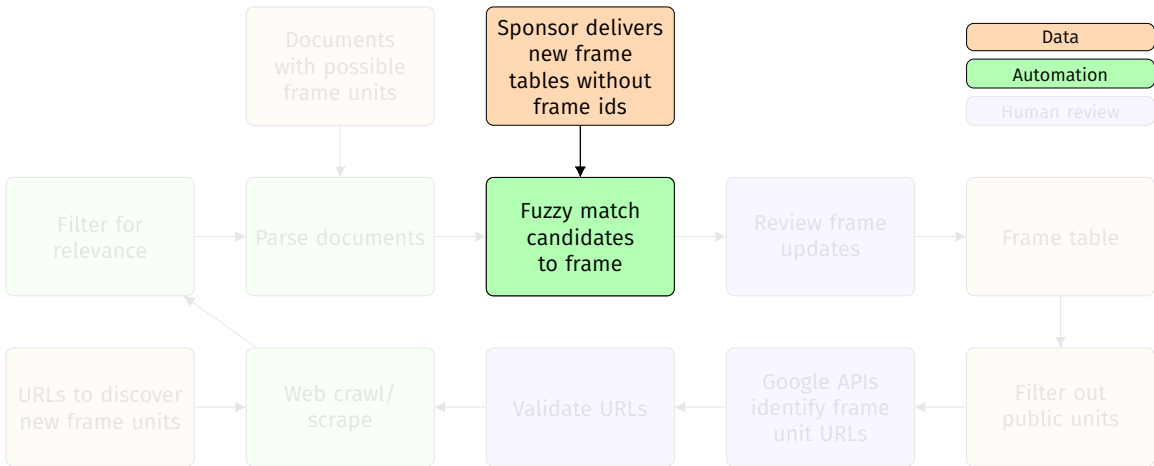
Civic Digital Fellows developed modular system

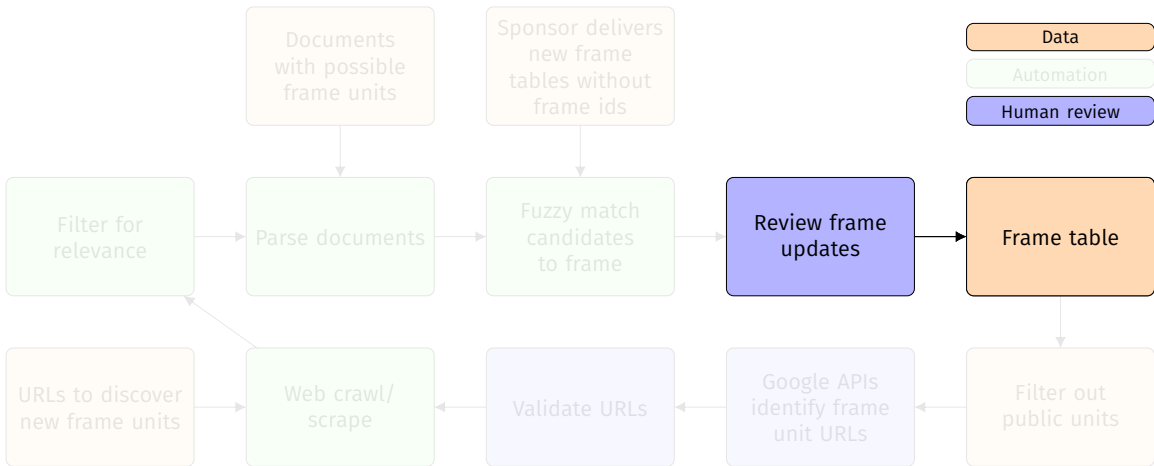
- ▶ Worked on this system the last two summers
- ▶ Carlos Ortega, a 2019 Civic Digital Fellow, helped us break the problem into smaller pieces
- ▶ Everything uses open source tools
- ▶ A number of the modules can be used for further process automation in our division

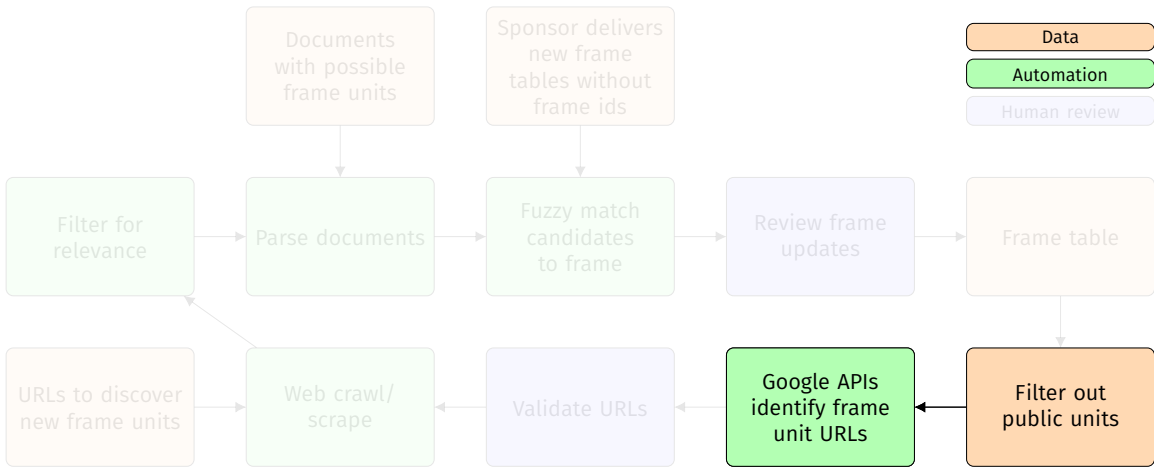


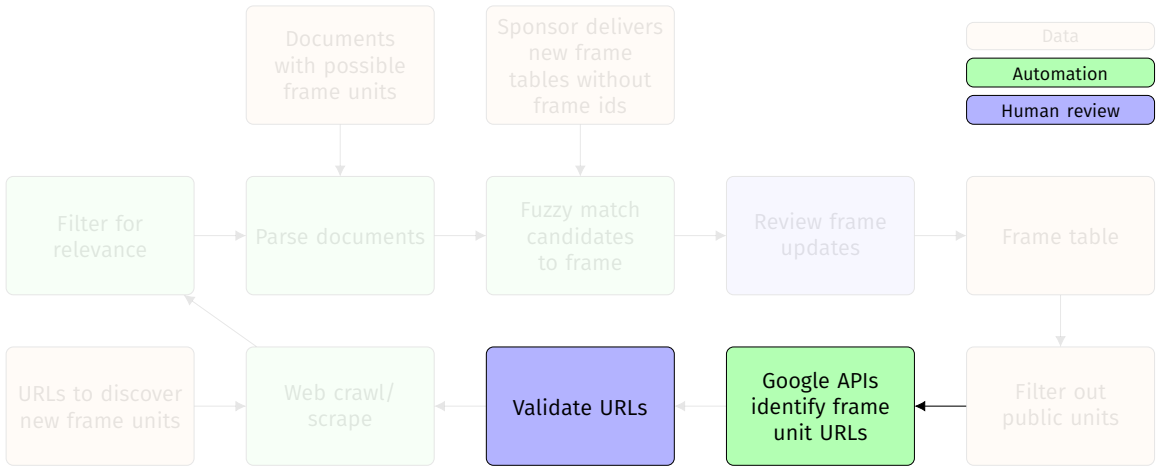


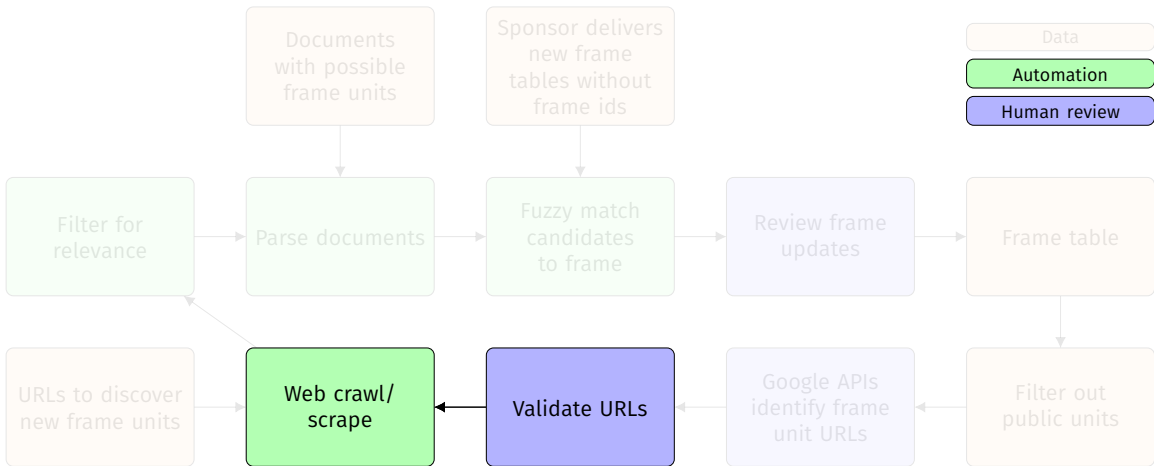


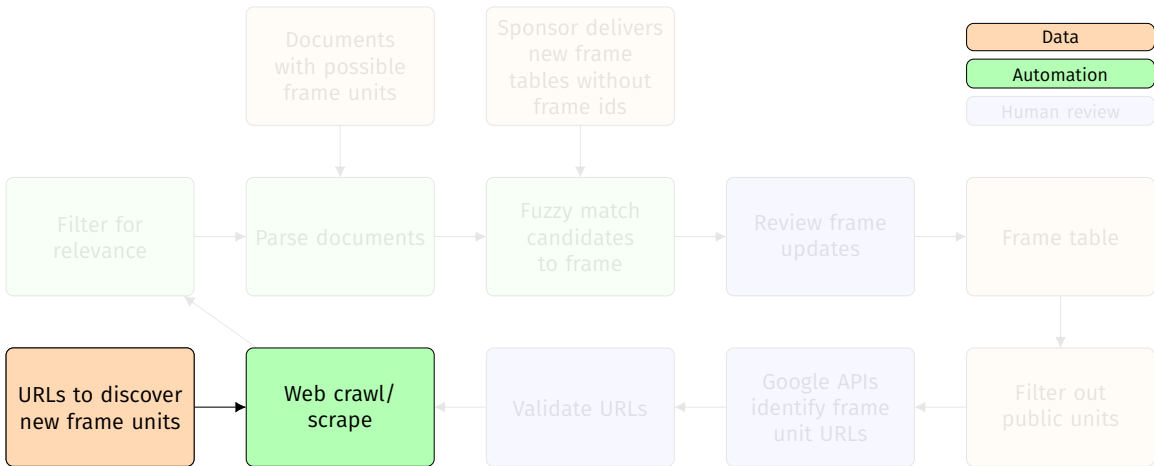


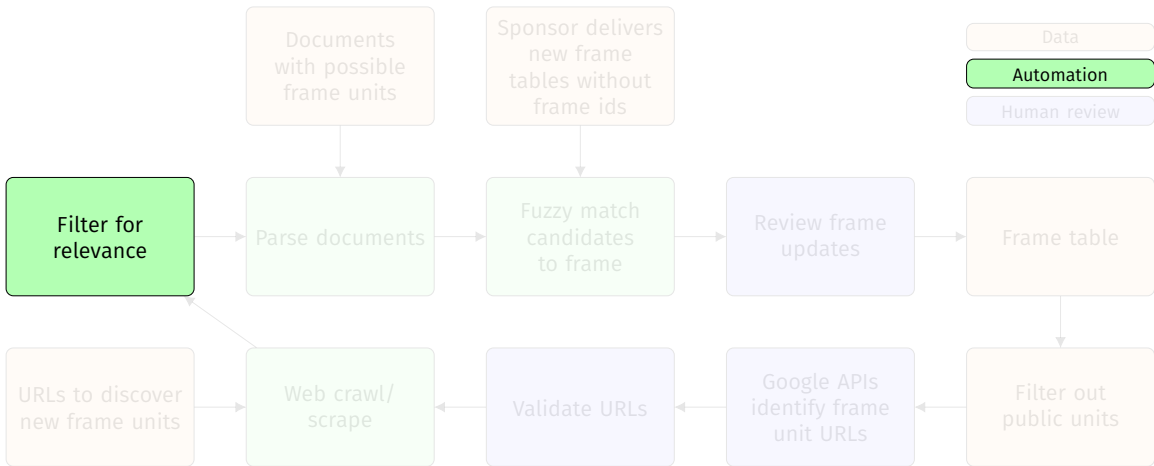


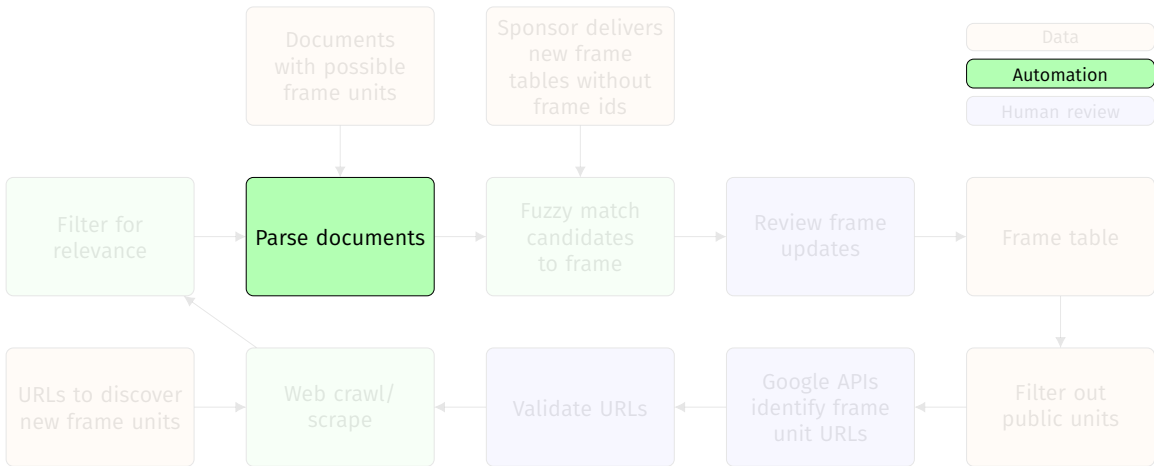


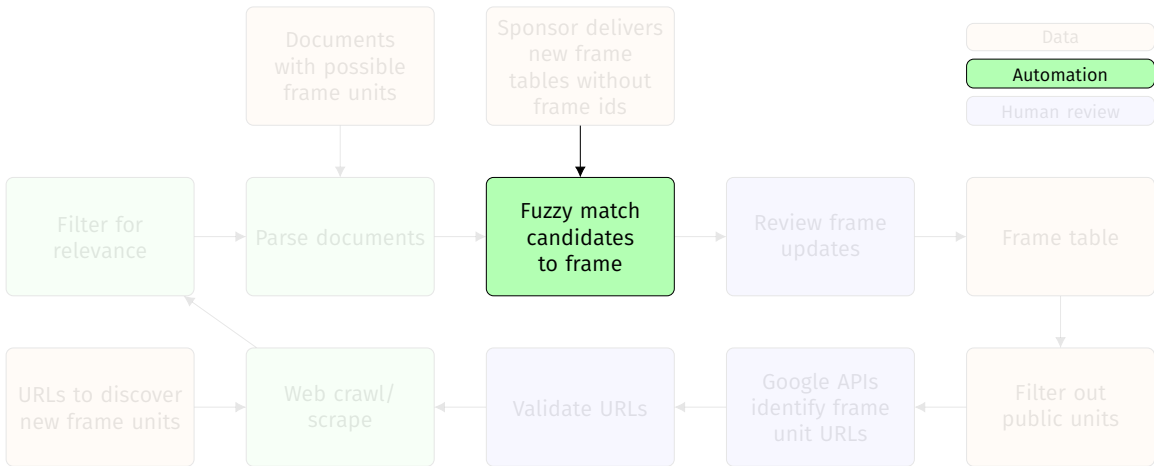












Next steps

- ▶ 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