Summary of the FSRDC Capstone Project

Chenxu Li

***Abstract* -** This capstone project aims to create an inventory of research outputs based on utilizing the Federal Statistical Research Data Centers (FSRDC) in an open format. The inventory can be used for subsequent visualizations and data analysis. The final results are displayed via GitHub Pages.

I. Introduction

The FSRDCs represent a cooperation between federal statistical agencies and leading research institutions around the USA. They allow qualified researchers to use restricted-access data under secure environments, also protect respondent confidentiality. Currently, the FSRDC program only researches on the outputs produced by the subset of projects that use Census data.

During the research processes depending on FSRDCs, researchers obtain various scholarly outputs, which means the data system generates a significant return on academic research. By September 2024, more than 1700 known publications are relevant to Census data from an FSRDC project. This can lead to an inference that the number of research outputs produced in the whole program might be grossly underestimated.

As a result, an FSRDC program-wide inventory of research outputs would be of great significance to the program itself, its partners and the public. The inventory should collect scholarly output information related to not only Census data, but also data from all participating Federal agencies. The final result can demonstrate the substantial value of the FERDC program and show a measurable return to the stakeholders supporting the FSRDCs.

To reach this goal, a problem is distinguishing research outputs relevant to an FSRDC project. Because many researchers using FSRDC data are not required to report their outputs in a uniform manner. The wanted outputs should be found from some online sources like Google Scholar and Crossref API. Significant coding and documentation is expected to be completed during the project.

II. Methodology

A. Data sources

Crossref is a DOI registration institution. It provides metadata of officially published scholar papers, periodical papers, conference papers and research papers. It is worth noting that almost all papers from Crossref have a DOI. The website also provides clear information about a paper, including title, authors, published year, name of the container and DOI, in a ordered format.

Compared with Crossref, Google Scholar has a wider scope of inclusion It is a scholar search engine, containing other papers like some preprints, technical reports and chapters of books. Some papers included might not have DOIs, or not have been officially published. Google Scholar secures the diversity of data collection, while Crossref is more professional.

B. Brief explanation of codes

The first step is data fetching. Based on VS Code as the Python editor, Crossref API supports using “requests” library to fetch data, and Google Scholar supports using “scholarly” library.

Data collected from these two platforms are then merged for subsequent process. The merged data file contains both high-quality papers from Crossref and papers that are not officially published, securing the comprehensiveness of data. Duplicate papers from the two platforms are removed by checking the same DOIs, the same title and the

same combination of authors and published year.

To visualize the collected data, three figures are produced. The first one shows the trend of FSRDC research paper quantity change with time increasing. The second tries to find the most popular containers that publish FSRDC research output. The last one compares the proportion of papers having a DOI and having no DOIs.

III. Results

According to **Figure 1**, FSRDC program began to be used more in nearly 1960, and became more and more popular through the next 40 years. After 2000, the number of FSRDC research papers increased sharply to more than 40 for one year. Since this project is done in 2025, subsequent results are not complete.

According to **Figure 2**, the container publishing the most FSRDC research papers is named *Federal Grants & Contracts*, which has published more than 70 papers, far more than the second container, *ICPSR Data Holdings*. **Figure 2** shows the top ten popular containers, actually the published paper numbers of the third to the tenth are close to each other, which

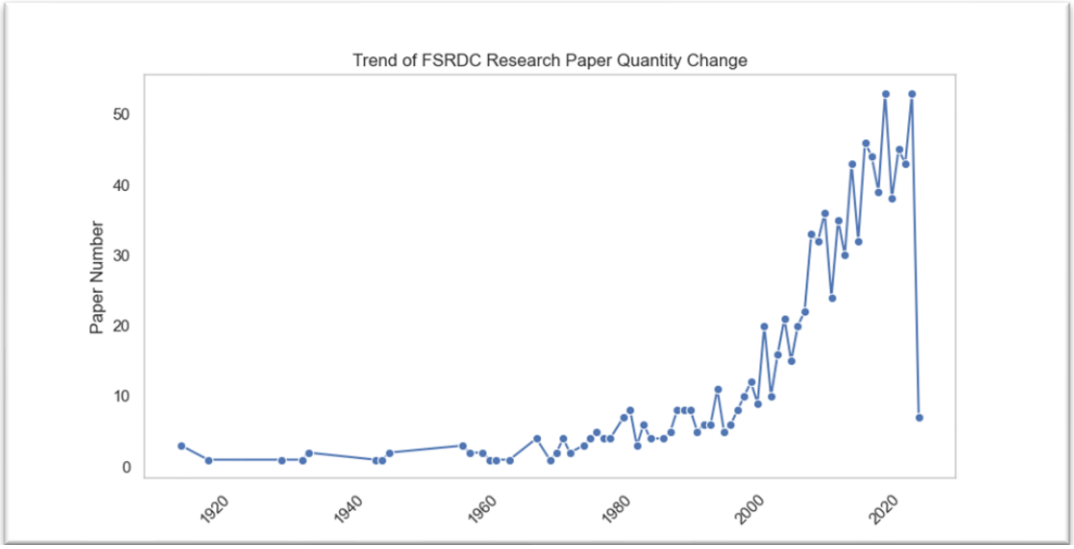

Figure 1

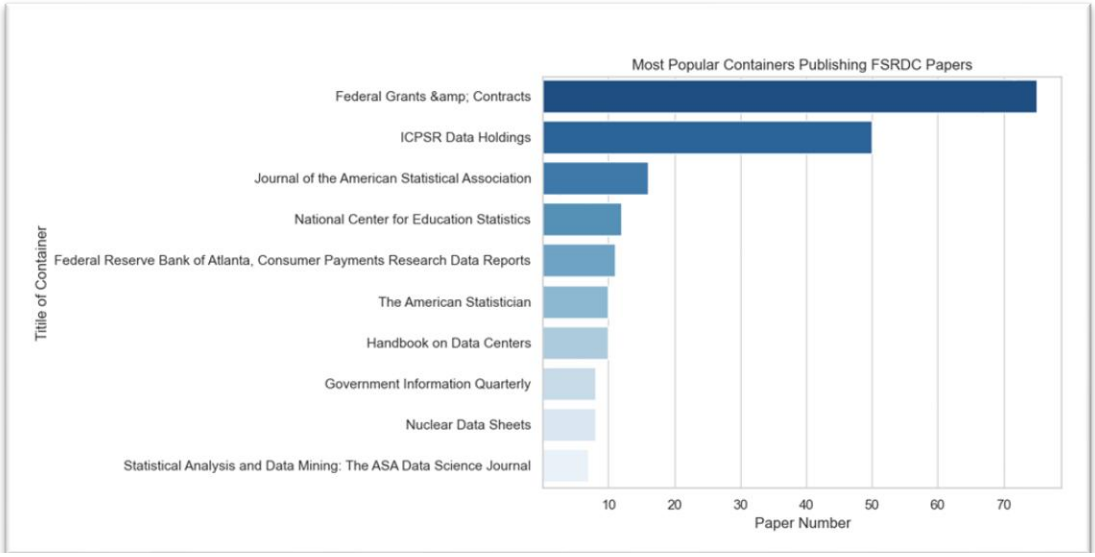

Figure 2

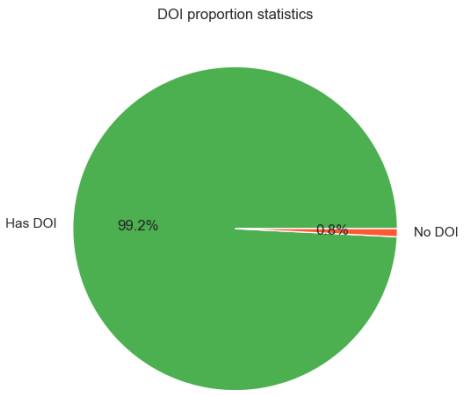

Figure 3

are fluctuated around 10 papers.

According to **Figure 3**, research papers collected without DOIs only takes 0.8 percent of the whole dataset. Having a DOI often means that a paper is published by an official periodic or scholar conference. **Figure 3** shows that research papers related to FSRDC program have high reliability and might play a significant role in future scholar researches.

All visualization results are then uploaded to GitHub Pages for references. The link is https://chenxuli2001.github.io/FSRDC-Inventory/.

IV. Future Improvements

The current capstone project still has many drawbacks to be developed. For example, the data is fetched from only two platforms. There are many other choices like arXiv that can provide different FSRDC research papers in various areas. More papers that have not been published also can be counted.

Another main problem is that the removal algorithm of this project is too simple. More advanced algorithm might be utilized to check if the fetched papers are duplicate, rather than simply compare via titles and so on.

Additionally, the run time of the script fetching data from Google Scholar is obviously longer than that fetching from Crossref. The reason is probably that Google Scholar has anti-scraping measures. Future methods might be explored to avoid this problem.

Appendix

When running the two data fetching scripts, the sampling outputs are shown as follows.

```
Success! Data is stored into data/crossref_data.csv
Fetch successfully!
                                                 Title  ...                                      DOI
0    How Does Data Access Shape Science? The Impact...  ...                 https://doi.org/10.3386/w31372
1        Restricted-Access Research Data Centers (NSF)  ...               https://doi.org/10.1002/fgc.31156
2    BFS: Get Data from the Swiss Federal Statistic...  ...       https://doi.org/10.32614/cran.package.bfs
3    Federal Automotive Statistical Tool: FY 2022 F...  ...                 https://doi.org/10.2172/2299525
4        Data Editing in Federal Statistical Agencies.  ...                 https://doi.org/10.2307/2290618
..                                                 ...  ...                                      ...
495                                          Editorial  ...               https://doi.org/10.1002/sam.11257
496  Comments on “visualizing statistical models”: ...  ...               https://doi.org/10.1002/sam.11272
497                      The Federal Rules of Evidence  ...               https://doi.org/10.1201/b13103-24
498                       Statistical Data Compression  ...  https://doi.org/10.1007/springerreference_57918
499                 Statistical interpretation of data  ...              https://doi.org/10.3403/bsiso16269

[500 rows x 5 columns]
```

```
Success! Data is stored into data/google_data.csv
Fetch successfully!
                                                 Title  ...                                      DOI
0    The research data centres of the Federal Stati...  ...  https://elibrary.duncker-humblot.com/article/7...
1    Federal Statistical Research Data Center Infos...  ...  https://scholarworks.iu.edu/dspace/bitstream/h...
2        United states data center energy usage report  ...  https://escholarship.org/content/qt84p772fc/qt...
3    Role of federal statistical research data cent...  ...  https://deepblue.lib.umich.edu/handle/2027.42/...
4    Statistical déjà vu: The National Data Center ...  ...  https://journalprivacyconfidentiality.org/inde...
..                                                 ...  ...                                      ...
495  When do losses count? Six fallacies of natural...  ...  https://journals.ametsoc.org/view/journals/bam...
496  A survey of software-defined networking: Past,...  ...  https://ieeexplore.ieee.org/abstract/document/...
497  Design and estimation for the national health ...  ...  https://books.google.com/books?hl=en&lr=&id=zH...
498                          Vanishing trial, the  ...  https://heinonline.org/hol-cgi-bin/get_pdf.cgi...
499  Trust In US Federal, State, And Local Public H...  ...  https://www.healthaffairs.org/doi/abs/10.1377/...

[500 rows x 5 columns]
```