CENSUS BUREAU STATISTICAL PRODUCT METADATA AND CITATION

25Sp DS 6015 Capstone

Capstone Students
University of Virginia
Angela Albarano
Chi Do
Will Milch
Becca Van Nostrand

Capstone Mentor
University of Virginia
Philip Waggoner

Sponsor
U.S. Census Bureau
Emily Molfino
Cass Dorius

INTRODUCTION

The U.S. Census Bureau engaged our team to define essential metadata fields and automate citation generation, with the goal of improving the discoverability of its statistical products.

STATISCIAL PRODUCTS ARE DEFINED AS "PRODUCTS THAT DESCRIBE, ESTIMATE, FORECAST, OR ANALZE THE PEOPLE, PLACES, AND THE ECONOMY, WITHOUT IDENTIFYING PERSONS, ORGANIZATION, OR INDIVIDUAL DATA." (MOLFINO & DORIUS)

OBJECTIVES

- Data products inventory
- Data products metadata
- Generate citations for data products
- Document site issues
- Investigate metadata use in other organizations

USER EXPERIENCE

User feedback revealed that navigating Census websites is often challenging, with users frequently struggling to locate the information they need and to discover existing data. While we encountered many of the same difficulties, we have documented specific instances below to support actionable remediation.

- The same data is presented across multiple URLs, making it difficult for users to determine which source to trust or how the pages relate to one another:
- https://www.census.gov/programs-surveys.html
- https://www.census.gov/programs-surveys/surveys-programs.html
- https://www.census.gov/programs-surveys/surveyhelp/list-of-surveys.html
- Data is distributed across multiple locations, with several interfaces available for users to search or query information. However, it is not intuitive for users to know which interface to use or where to find specific information.
- Same file may exist in multiple formats (pdf, txt, doc)
- Census Survey Explorer is an interface to all surveys and censuses. It provides an explicit set of dropdowns to assist a user in searching for survey and census data.
- Does not include all statistical products
- Topics are incomplete – additional topics at https://www.census.gov/topics.html
- Choices presented to the user are not comprehensive and limit user ability to adequately assess fitness for use
- No connection with internal systems to automatically feed and update the tool (dropdowns must be updated manually)
- Redundant terms in filter dropdowns (Parole/Parolees)
- Comparison of Survey Explorer filter compared to previous capstone work may be found on GitHub
- Census Data API Discovery Tool is dated 2014 leading to questions of currency
- Duplicate content—identical files are stored in multiple locations

INVENTORY

An analysis was conducted to better understand site content and structure using both manual and automated methods. The manual assessment provided insights into the user experience and helped contextualize existing end-user feedback. In parallel, automated scripts were employed to scrape and crawl Census websites and the datasets API endpoint was used to collect information on available data products across Census and related domains.

The resulting inventory consists of a CSV output listing the available data products and their corresponding paths. Additionally, random samples of Census PDF files were downloaded for ingestion and metadata extraction to aid in identifying key descriptive fields.

API

The API at `api.census.gov/data.json` provides access to Census data in a standardized JSON format. While the Census Bureau offers its APIs in multiple formats, we converted the JSON output to CSV to facilitate easier analysis. Each row in the resulting CSV captures metadata about the datasets exposed by the Census Bureau, including key descriptive metadata.

Web Scraping

Web scraping of `www2.census.gov` involves making multiple requests to web pages in order to programmatically collect information. This process respects the rules specified in the site's `robots.txt` file to ensure responsible crawling. Since web scraping is dependent on the structure of the HTML, any changes to the page layout can affect the effectiveness of the script. We used the BeautifulSoup (bs4) library to request and parse HTML content, enabling the identification and collection of file links from the site.

Web Crawling Process

1. Introduction to Web Crawling:

To complement manual exploration and targeted scraping, our team implemented a recursive web crawler to systematically explore Census.gov and identify downloadable data products and supporting documentation.

2. Tools and Setup:

We used a headless browser in combination with Python libraries such as `Selenium` and `BeautifulSoup`. The crawler was configured to respect `robots.txt` restrictions and simulate natural browser behavior. The web driver (e.g., ChromeDriver) was installed and used for page rendering and navigation.

3. Crawling Strategy:

The crawler was launched from the base URL `https://www.census.gov`, focusing on relevant directories such as `/programs-surveys`, `/library/publications`, and `/data.html`. Unlike a

typical sitemap crawler, our goal was to discover and log links to statistical products, particularly files with `.pdf`, `.doc`, `.zip`, `.xls`, and `.csv` extensions.

4. Implementation Highlights:

- Implemented a recursive algorithm to crawl up to 10 levels deep
- Total pages crawled: 3,523.
- File types targeted: PDF, DOC, ZIP, XLS, CSV
- Output: A classified `.csv` file containing URLs, filenames, and paths

5. Observations:

- Duplicate files appeared under multiple URLs
- Some ZIP files exceeded 3GB
- Many documents lacked clear metadata in filenames
- Certain links led to outdated or deprecated pages
- FTP and mirror-style directories were less structured

6. Integration with Metadata Pipeline:

The discovered file paths were passed to the metadata extraction team for Named Entity Recognition (NER) processing, using both rule-based and large language model (LLM) techniques. Our crawling ensured comprehensive coverage of diverse file types and enhanced the breadth and quality of metadata extraction.

7. Limitations and Future Work:

- Some documents were inaccessible due to dynamic JavaScript rendering
- The crawler was limited by page timeouts and inconsistencies in site structure
- Future work could include scheduled crawling with automatic deduplication and integration into a broader data-ingestion pipeline

METADATA AND CITATION

We focused on developing a pipeline to extract structured metadata from U.S. Census Bureau PDF documents. Although the documents contain valuable information, their inconsistent formatting hinders discoverability. The primary goal was to generate machine-readable metadata fields—such as title, topic categories, keywords, data collection periods, and citation formats—and consolidate the results into CSV files to improve accessibility and integration into digital repositories.

The initial approach employed a rule-based natural language processing (NLP) pipeline built with the spaCy library (file and README can be found here). Custom heuristics and entity-matching patterns were designed to extract metadata from the PDFs. While this method showed promise on well-structured files, it proved fragile when faced with layout variations, abbreviations, and inconsistent fields across Census documents. Manual tuning was often required, and recall remained low for nuanced metadata such as topic domains and keywords. These challenges highlighted the need for a more flexible, context-aware solution.

The project team subsequently transitioned to a large language model (LLM) approach using GPT-4. Through carefully designed prompts, GPT-4 extracted metadata even from semi-structured and informally formatted PDFs. This is an example of a prompt we used:

You are a metadata extraction assistant trained to generate clean, structured metadata from U.S. Census Bureau PDF documents. Below is the extracted text from a government publication(s). Your job is to read the content and return structured metadata in JSON format.

Please extract the following fields:

1. Dataset Name – a clear title for the document
2. Topic Categories – 1 to 3 broad categories (e.g., Demographics, Housing, Economics)
3. Relevant Keywords – 5 to 10 specific terms found or implied in the text
4. Document Type – e.g., Report, Handbook, Factsheet, Technical Documentation, Dataset Guide from terms found or implied in the text
5. Data Collection Period – specific year(s) if mentioned
6. Suggested Citation – a complete citation in APA or government style
7. Path or Survey Code – e.g., ACS, SIPP, PUMS, P60, if referenced in the filename or content, retain source file name

Here is the document text:

<<<BEGIN DOCUMENT>>>
[Provide document or folder of documents here]
<<<END DOCUMENT>>>

Return the metadata in this JSON format:

```
{
  "Dataset Name": "",
  "Topic Categories": [],
  "Relevant Keywords": [],
  "Document Type": "",
  "Data Collection Period": "",
  "Suggested Citation": "",
  "Path or Survey Code": ""
  "Source file name": ""
}
```

GPT-4 consistently produced high-quality outputs, accurately inferring document titles, major themes, keywords, data periods, and citation formats. Its outputs aligned closely with the actual document content and required minimal post-processing. A CSV file was created to organize the extracted metadata, substantially improving the rule-based method.

The Gemini LLM was also tested on a subset of documents as a comparative exercise. While Gemini produced structured outputs for many files, it exhibited inconsistencies. It frequently omitted documents, introduced extraneous entries, and sometimes extrapolated metadata that lacked clear grounding in the document text. Additionally, while GPT-4 and Gemini have similar batch limits, GPT-4 could batch large amounts of files independently, while Gemini required feeding in documents 8K tokens at a time. After filtering for usable entries, Gemini's metadata was incorporated into a parallel CSV for evaluation. Ultimately, GPT-4 demonstrated superior performance in completeness, contextual accuracy, and document consistency.

Both GPT-4 and Gemini successfully extracted key metadata fields such as Dataset Name, Topic Categories, Relevant Keywords, Suggested Citation Format, Document Type, and Path or Survey Code, even when filenames were abbreviated or nonstandard. However, GPT-4 more consistently populated fields like Data Collection Period and Survey Code, whereas Gemini often defaulted to "N/A" when uncertainty was detected.

To ensure metadata reliability, the project followed standards consistent with emerging Office of Management and Budget (OMB) guidance for federal metadata. This guidance emphasizes the need for structured, machine-readable, and consistently applied metadata elements to improve data discoverability and reusability. Special care was taken to include full citation formats for each document under best practices for government publications, allowing future users to attribute source materials correctly.

GPT-4 emerged as the most effective tool for metadata extraction in this project. It demonstrated high contextual accuracy, strong recall across metadata fields, and minimal fabrication of information, even when working with semi-structured and inconsistently formatted PDFs. The resulting metadata provides a detailed, machine-readable foundation that improves access to Census Bureau documents and aligns with federal metadata standards. Although Gemini showed some potential, its outputs were less consistent and would require additional validation before being suitable for production use.

Future teams could build on this work by assessing the quality and consistency of the extracted metadata through targeted validation and manual review. Based on those findings, a logical next step would be to automate the tagging process using a streamlined pipeline that integrates GPT-4 or similar LLMs for batch metadata generation. Additionally, a more scalable long-term solution could involve developing a deep learning-based NER model trained on the outputs of this project, enabling high-accuracy, local metadata tagging without reliance on external APIs.

PEER ORGANIZATIONS METADATA USE

As part of our review of metadata across peer institutions, we conducted a multi-stage exploration focused on understanding the gaps in the current metadata structure improving metadata usability and value. We also brainstormed additional fields for key user personas.

First, we compiled a list of metadata fields currently published across publicly accessible datasets across both Census.gov and Data.Census.gov. This inventory highlighted variability in metadata depth and structure between different product lines, with limited standardization in metadata taxonomy. We categorized metadata fields as either *global* (shared across datasets) or *product-specific*, identifying core attributes such as dataset title, geographic coverage, and release date, but noting that interpretability, provenance, and methodological context were inconsistently provided.

Current Census Bureau Fields by Global Presence	
Global Fields	**Product-Specific Fields (available for some data products)**
Title, Description, Geographic Coverage, Release Date, Source, Contact Information, Update Frequency, Data Format, Data Access URL, Confidentiality Policy, License, Language, Tags & Keywords, API Availability	Variable Definition, Units of Measurement, Collection Method, Time Period, Revision History, Data Suppression Rules, Related Datasets, Citations, Dataset Size, Record Count, Data Dictionary

We then conducted a comparative analysis with peer institutions, including both domestic peers such as OMB or BEA as well as international bodies such as Eurostat, Statistics Canada, and OECD. We catalogued the presence of enriched metadata fields across these agencies—such as quality indicators, intended use cases, data lineage, licensing terms, and embedded methodological notes. The Census Bureau's current offering ranked strongly on coverage and update frequency but fell short in user-oriented fields such as interpretive guidance, interactive metadata access, and API discoverability. Peers often include metadata fields that directly guide interpretation, detail limitations, and enhance machine-readability. Many also support interactive metadata portals and user-

centric schema documentation. Fields that are offered by other institutions besides the bureau are citation information (addressed in other parts of our project), downloadable metadata files, full multilingual metadata support, related datasets, and searchable keyword indicators.

To guide future improvements, we created needs profiles for five high-priority user personas provided to us at the beginning of the report:

Persona	General Needs	Potential Additional Metadata Fields
Professor Penelope	Academic rigor, Comparability	Documentation for Academic-Use, Versioning History with Data Revision logs
Developer Maya	Integration-ready, Schema clarity	Format, Size, Rate Limit expectations, Codebooks & Definition sheets
Data Journalist Kate	Fast insights, Plain Language, Visualization support	High-level insights, Direct download links, Suggestions for visualizations
Community Leader Sarah	Usability, Local Relevance	Examples of Community Action for reports, basic usage notes, map-ready exports
Analyst David	Accuracy, Joinability, Granularity	Links to Related Data products, Quality flag explanations, inflation adjustments

Each persona analysis revealed metadata fields that would significantly improve usability— and provide ease of use for the key users of the bureau.

References

Albarano, A., Do, C., Milch, W., & Van Nostrand, B. census_capstone. Retrieved from GitHub: https://github.com/aalbaran/census_capstone

Molfino, E., & Dorius, C. Proposal 10: US Census Product Metadata Tool. U.S. Census Bureau.

Erickson T., Pippin C., Rico B., Toutsis V., & Turner T. Retrieved from Box: https://virginia.box.com/s/2rh0kdc7urzbcy9j0567all17b1tf0r8