Modernizing Census Bureau Economic Statistics through Web Scraping

Joint Statistical Meetings
Vancouver, Canada
August 1, 2018

Brian Dumbacher
Carma Hogue
U.S. Census Bureau

United States Census Bureau
U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

Disclaimer: Any views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

Outline

- Big Data Context
- Web Scraping Background
- Scraping Assisted by Learning (SABLE)
  - State Government Tax Revenue Collections
  - Public Pension Statistics
- Securities and Exchange Commission (SEC) Filing Metadata
- Building Permit Data
- Efforts to Improve Sampling Frames
- Next Steps with Web Scraping

Big Data Context

- U.S. Census Bureau’s Economic Directorate has been researching alternative data sources and Big Data methodologies
- Evaluation criteria include
  - Quality
  - Cost
  - Skillset
- Machine learning, “tableplots” for edit reduction, web scraping, and web crawling are beneficial methods

Web Scraping Background

- For many economic surveys, respondent data or equivalent-quality data are available online
  - Respondent websites
  - Public filings with the SEC
  - Application Programming Interfaces (APIs)
  - Publications on state and local government websites
- Current data collection efforts along these lines are manual
- Going directly to online sources and collecting data passively could reduce respondent and analyst burden

Web Scraping Background (cont.)

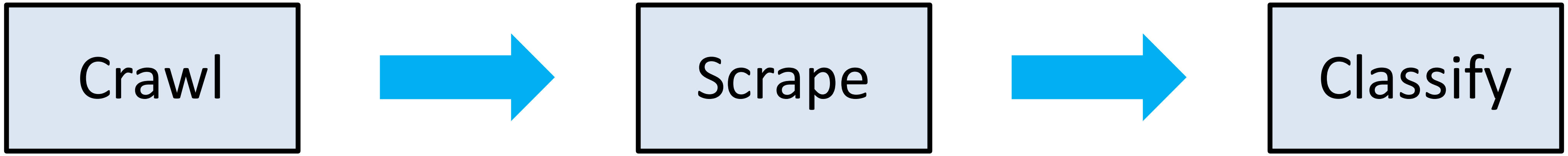
- Web scraping: automated process of collecting data from an online source
- Web crawling: automated process of systematically visiting and reading web pages
- Policy issues
  - Informed consent
  - Websites of private companies vs. government websites
  - Statistics Canada's "About us" page informs data users and respondents about web scraping

Source: Statistics Canada. (2018). About us. Accessed July 6, 2018. https://www.statcan.gc.ca/eng/about/about

SABLE

- Scraping Assisted by Learning
- Collection of tools for
  - Crawling websites
  - Scraping documents and data
  - Classifying text
- Models based on text analysis and machine learning
- Implemented using free, open-source software
  - Apache Nutch
  - Python

Three Main Tasks



Crawl → Scrape → Classify

Crawl

Given a website,

- Scan website
- Find documents and extract text
- Apply classification model to predict whether document contains useful data

Scrape

Given a document classified as useful,

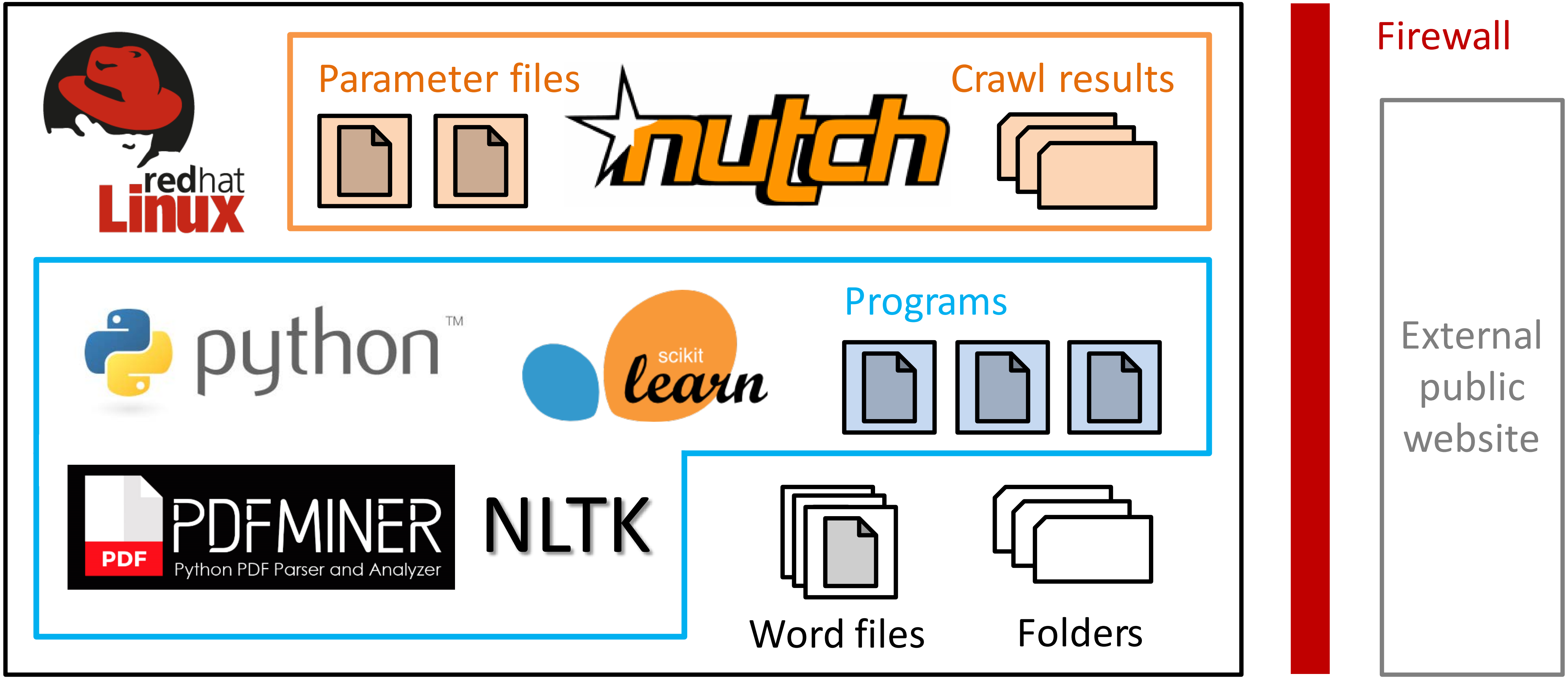
- Apply model to learn the location of useful data
- Extract numerical values and corresponding text

Classify

Given scraped data,

- Preprocess data
- Apply classification model to map text to Census Bureau definitions and classification codes

Architecture Design



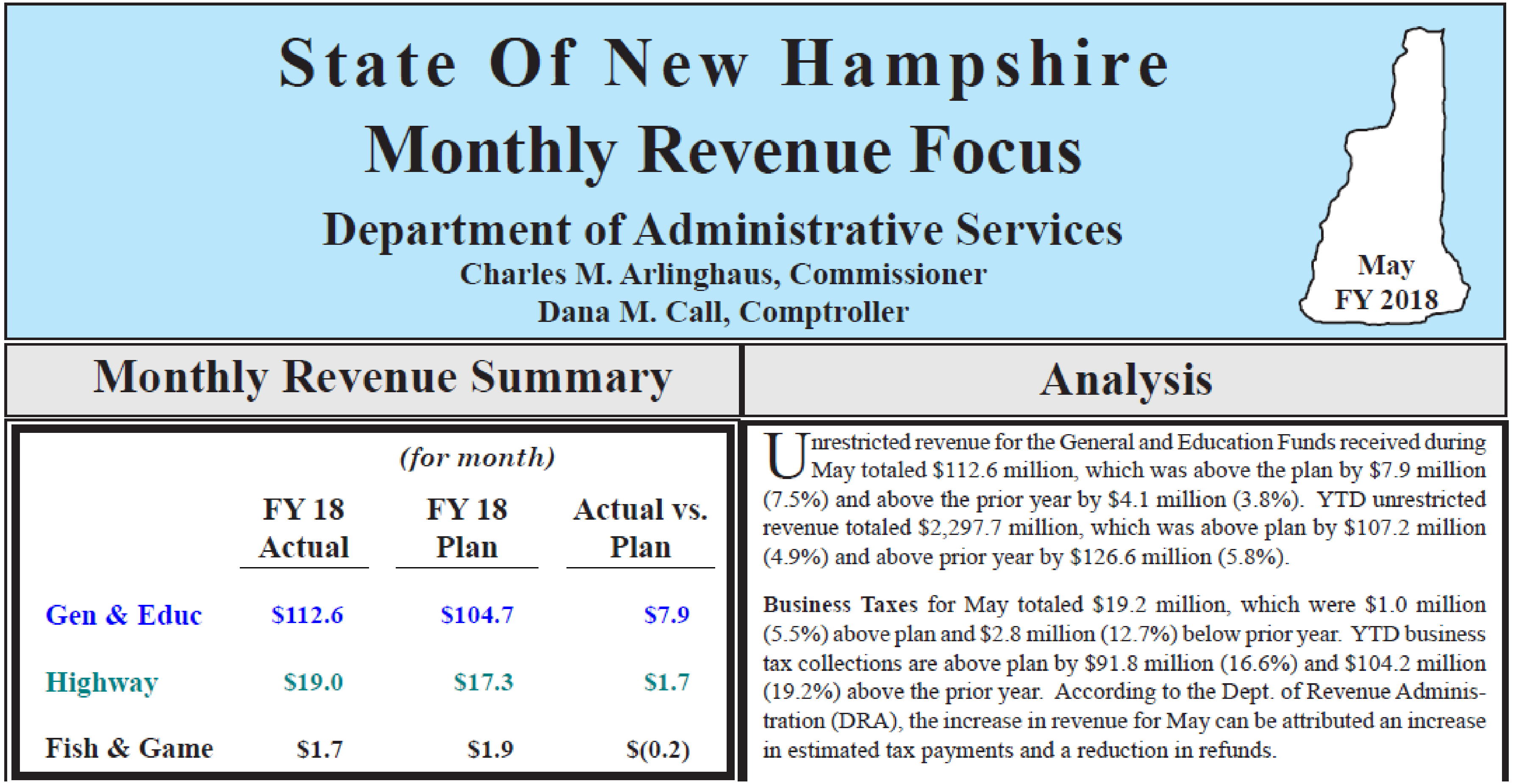
Moving to a Production Environment

- Authority to Operate
  - Risk profile and security assessment
  - Documentation and procedures
  - Audit trail system
  - Subversion for code management
- SABLE repository on the Census Bureau’s GitHub account
  - https://www.github.com/uscensusbureau/SABLE
  - Programs, supplementary files, examples, and documentation

State Government Tax Revenue Collections

- Data on state government tax revenue collections can be found online in Comprehensive Annual Financial Reports (CAFRs) and other publications
- Used SABLE to find additional online sources in Portable Document Format (PDF)
  - Crawled websites of state governments
  - Discovered approximately 60,000 PDFs
  - Manually classified a simple random sample of 6,000 PDFs as “Useful” or “Not Useful”
  - Applied machine learning to build text classification models based on occurrences of word sequences

Example Document



State Of New Hampshire

Monthly Revenue Focus

Department of Administrative Services

Charles M. Arlinghaus, Commissioner
Dana M. Call, Comptroller

May
FY 2018

Monthly Revenue Summary

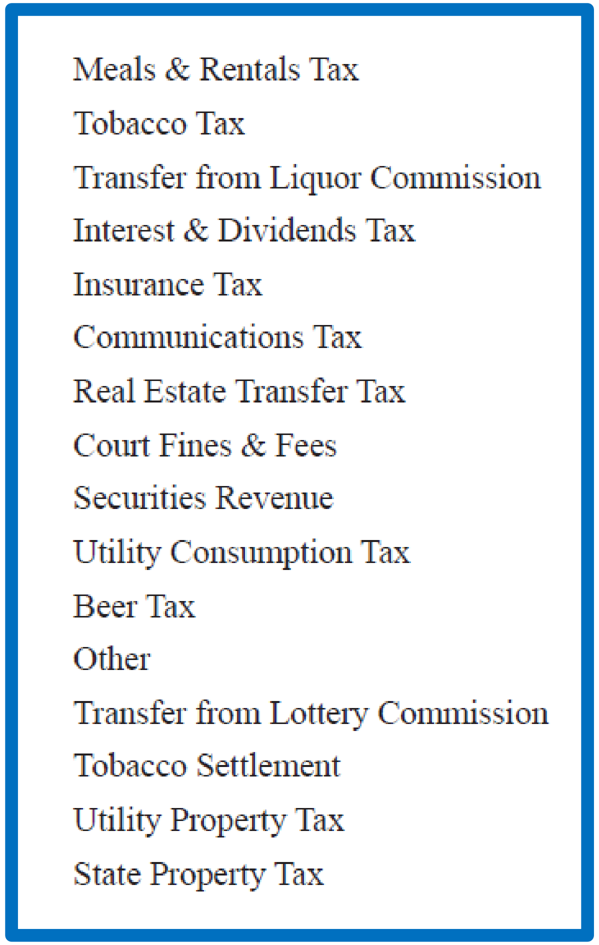
(for month)

	FY 18 Actual	FY 18 Plan	Actual vs. Plan
Gen & Educ	$112.6	$104.7	$7.9
Highway	$19.0	$17.3	$1.7
Fish & Game	$1.7	$1.9	$(0.2)

Analysis

Unrestricted revenue for the General and Education Funds received during May totaled $112.6 million, which was above the plan by $7.9 million (7.5%) and above the prior year by $4.1 million (3.8%). YTD unrestricted revenue totaled $2,297.7 million, which was above plan by $107.2 million (4.9%) and above prior year by $126.6 million (5.8%).

Business Taxes for May totaled $19.2 million, which were $1.0 million (5.5%) above plan and $2.8 million (12.7%) below prior year. YTD business tax collections are above plan by $91.8 million (16.6%) and $104.2 million (19.2%) above the prior year. According to the Dept. of Revenue Administration (DRA), the increase in revenue for May can be attributed an increase in estimated tax payments and a reduction in refunds.



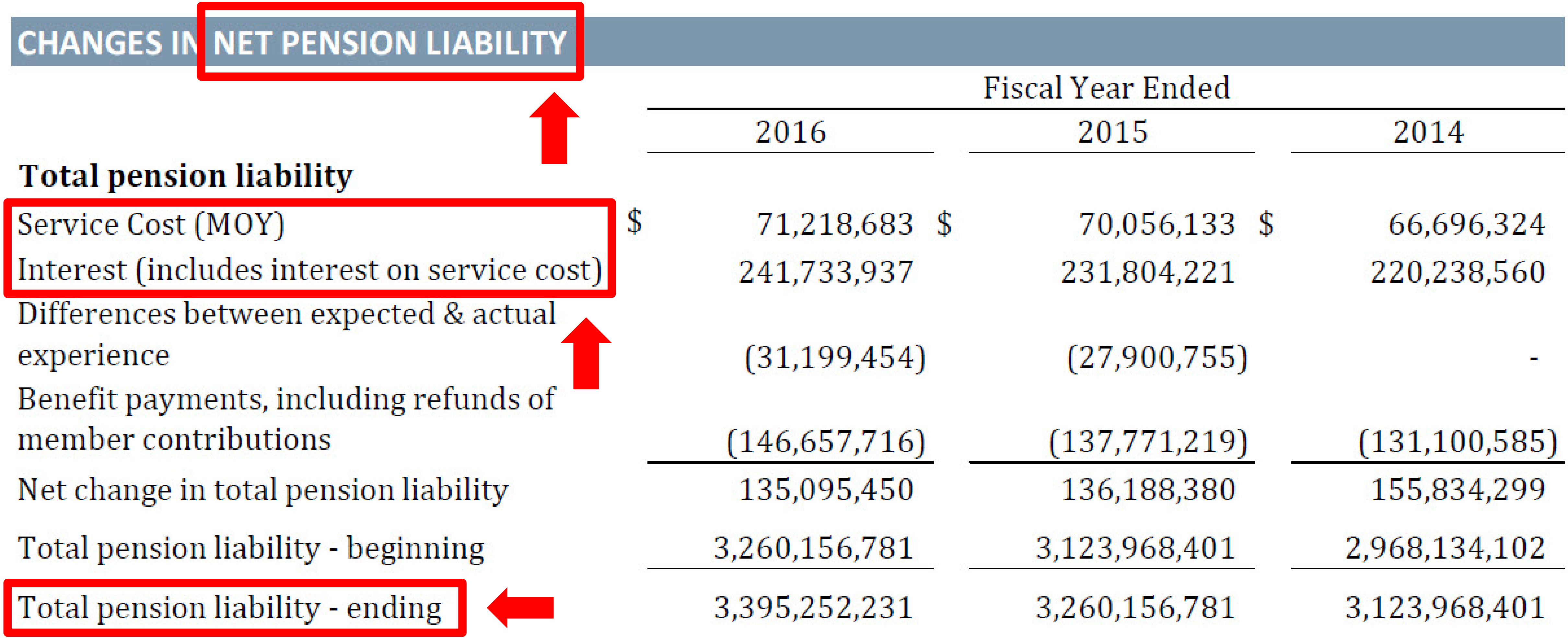
Meals & Rentals Tax
Tobacco Tax
Transfer from Liquor Commission
Interest & Dividends Tax
Insurance Tax
Communications Tax
Real Estate Transfer Tax
Court Fines & Fees
Securities Revenue
Utility Consumption Tax
Beer Tax
Other
Transfer from Lottery Commission
Tobacco Settlement
Utility Property Tax
State Property Tax

Source: New Hampshire Department of Administrative Services. Accessed July 6, 2018. https://das.nh.gov/accounting/FY%2018/Monthly_Rev_May.pdf

Pension Statistics

- Likewise, data on public pension funds can be found online and in CAFRs
- Examine feasibility of scraping service cost and interest statistics
- Create a data product based on the largest publicly administered pension plans
- Two-stage approach
  - – Identify tables using occurrences of word sequences
  - – Apply scraping algorithm based on table structure

Examples of Key Word Sequences



CHANGES IN NET PENSION LIABILITY

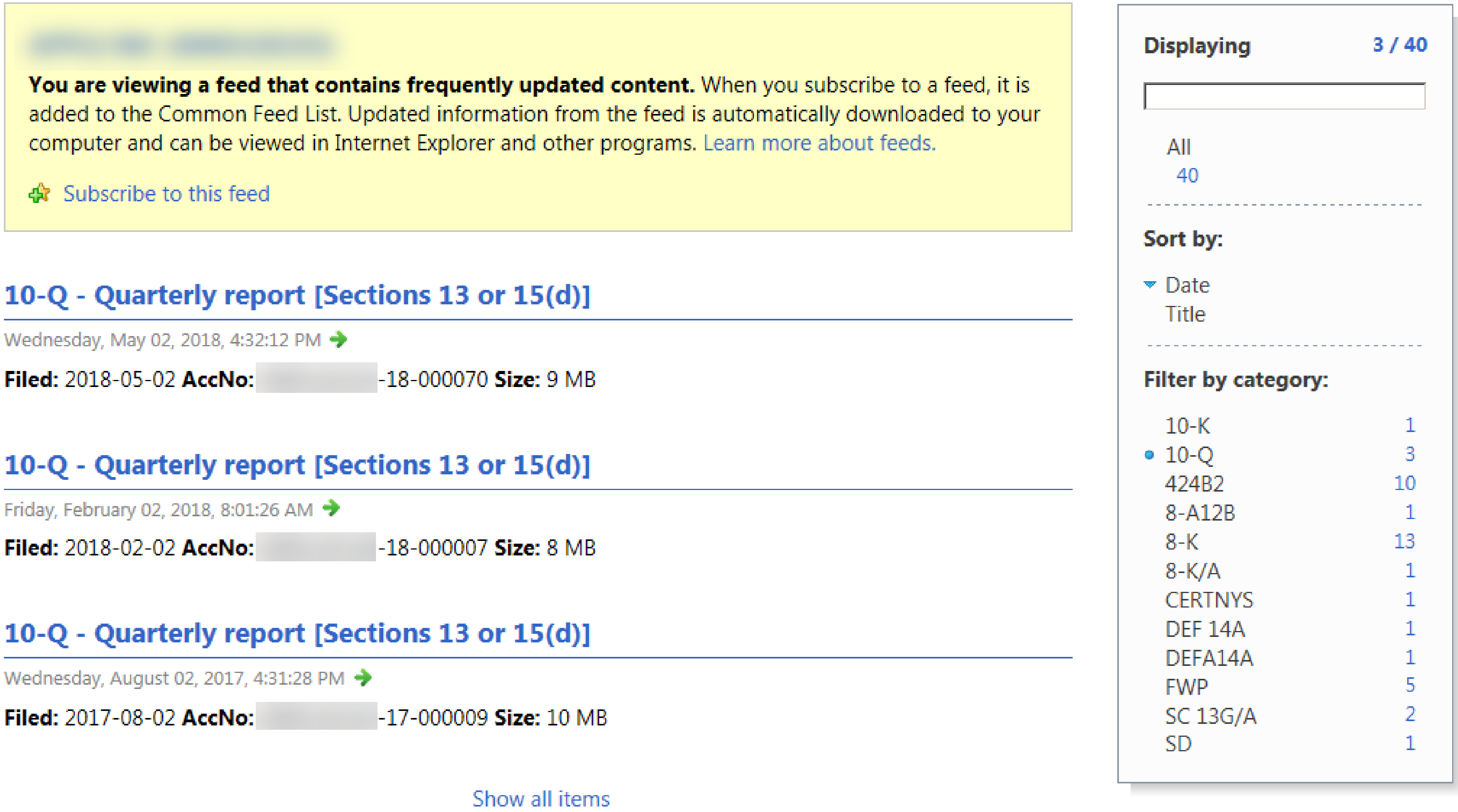
	Fiscal Year Ended		
	2016	2015	2014
Total pension liability			
Service Cost (MOY)	$ 71,218,683	$ 70,056,133	$ 66,696,324
Interest (includes interest on service cost)	241,733,937	231,804,221	220,238,560
Differences between expected & actual experience	(31,199,454)	(27,900,755)	-
Benefit payments, including refunds of member contributions	(146,657,716)	(137,771,219)	(131,100,585)
Net change in total pension liability	135,095,450	136,188,380	155,834,299
Total pension liability - beginning	3,260,156,781	3,123,968,401	2,968,134,102
Total pension liability - ending	3,395,252,231	3,260,156,781	3,123,968,401

Source: Comprehensive Annual Financial Report For Fiscal Years Ended June 30, 2016 and 2015; Santa Barbara County Employees' Retirement System; A Pension Trust Fund for the County of Santa Barbara, California. Accessed July 6, 2018. http://cosb.countyofsb.org/uploadedFiles/sbcers/benefits/SBCERS%206-30-2016%20CAFR%20With%20Letters.pdf

SEC Filing Metadata

- Online database of financial performance reports for publicly traded companies
- Really Simple Syndication (RSS) feed provides information about recent SEC filings such as filing dates
- Data obtainable in Extensible Markup Language (XML) format
- One can query this RSS feed by supplying
  - Filing type [e.g., 10-K (annual report) or 10-Q (quarterly report)]
  - Central Index Key, which the SEC uses to identify companies that have filed disclosures

RSS Feed



[redacted]

You are viewing a feed that contains frequently updated content. When you subscribe to a feed, it is added to the Common Feed List. Updated information from the feed is automatically downloaded to your computer and can be viewed in Internet Explorer and other programs. Learn more about feeds.

Subscribe to this feed

10-Q - Quarterly report [Sections 13 or 15(d)]

Wednesday, May 02, 2018, 4:32:12 PM →

Filed: 2018-05-02 **AccNo:** [redacted]-18-000070 **Size:** 9 MB

10-Q - Quarterly report [Sections 13 or 15(d)]

Friday, February 02, 2018, 8:01:26 AM →

Filed: 2018-02-02 **AccNo:** [redacted]-18-000007 **Size:** 8 MB

10-Q - Quarterly report [Sections 13 or 15(d)]

Wednesday, August 02, 2017, 4:31:28 PM →

Filed: 2017-08-02 **AccNo:** [redacted]-17-000009 **Size:** 10 MB

Show all items

Displaying 3 / 40

All
40

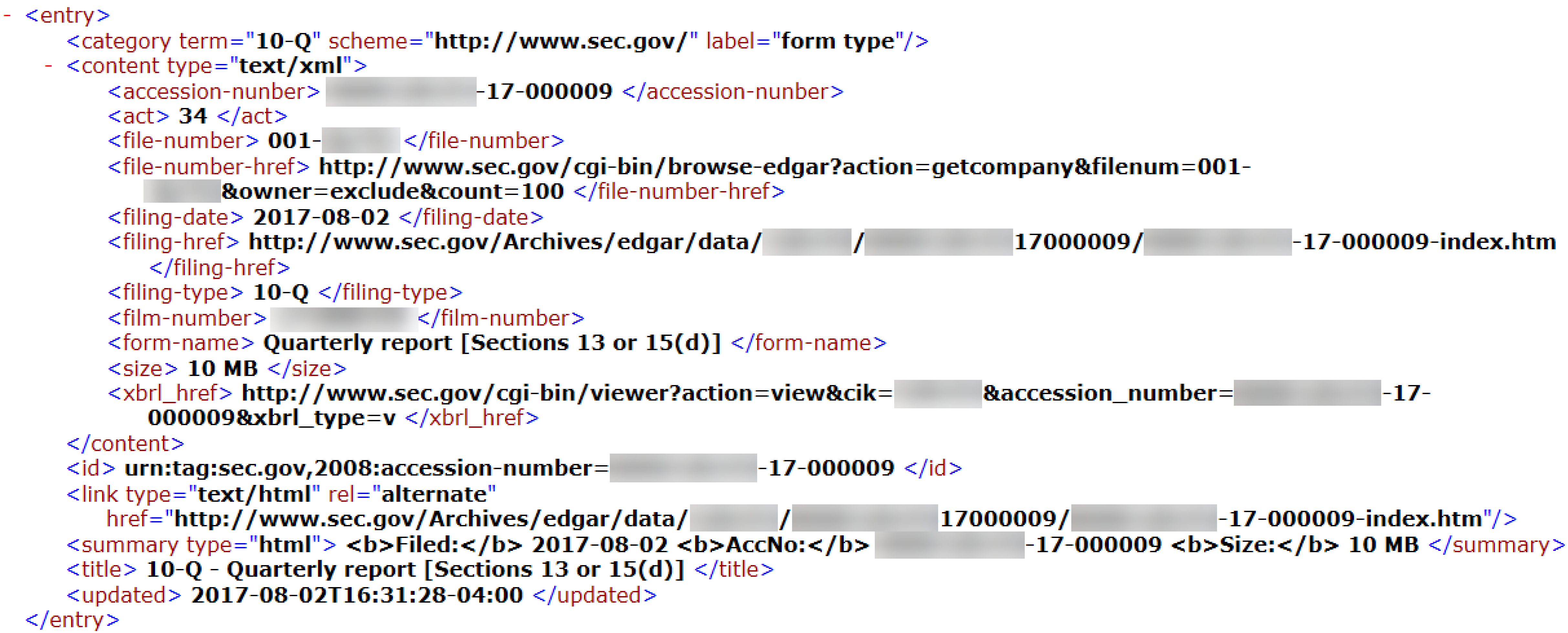
Sort by:

▾ Date
Title

Filter by category:

10-K	1
● 10-Q	3
424B2	10
8-A12B	1
8-K	13
8-K/A	1
CERTNYS	1
DEF 14A	1
DEFA14A	1
FWP	5
SC 13G/A	2
SD	1

Data from RSS Feed in XML Format



```
- <entry>
    <category term="10-Q" scheme="http://www.sec.gov/" label="form type"/>
  - <content type="text/xml">
      <accession-nunber> [redacted]-17-000009 </accession-nunber>
      <act> 34 </act>
      <file-number> 001-[redacted] </file-number>
      <file-number-href> http://www.sec.gov/cgi-bin/browse-edgar?action=getcompany&filenum=001-
        [redacted]&owner=exclude&count=100 </file-number-href>
      <filing-date> 2017-08-02 </filing-date>
      <filing-href> http://www.sec.gov/Archives/edgar/data/[redacted]/[redacted]17000009/[redacted]-17-000009-index.htm
        </filing-href>
      <filing-type> 10-Q </filing-type>
      <film-number> [redacted]</film-number>
      <form-name> Quarterly report [Sections 13 or 15(d)] </form-name>
      <size> 10 MB </size>
      <xbrl_href> http://www.sec.gov/cgi-bin/viewer?action=view&cik=[redacted]&accession_number=[redacted]-17-
        000009&xbrl_type=v </xbrl_href>
    </content>
    <id> urn:tag:sec.gov,2008:accession-number=[redacted]-17-000009 </id>
    <link type="text/html" rel="alternate"
      href="http://www.sec.gov/Archives/edgar/data/[redacted]/[redacted]17000009/[redacted]-17-000009-index.htm"/>
    <summary type="html"> <b>Filed:</b> 2017-08-02 <b>AccNo:</b> [redacted]-17-000009 <b>Size:</b> 10 MB </summary>
    <title> 10-Q - Quarterly report [Sections 13 or 15(d)] </title>
    <updated> 2017-08-02T16:31:28-04:00 </updated>
  </entry>
```

SEC Current Work

- Work with various survey teams to see how they can best use this information
- Incorporate web scraping and a filing notification system into production cycles
- Research how best to scrape actual financial information
  - Extensible Business Reporting Language (XBRL)
  - Arelle software
  - lxml XML parser

Building Permit Data

- Data on new construction
  - Used to measure and evaluate size, composition, and change in the construction sector
  - Building Permits Survey (BPS)
  - Survey of Construction (SOC)
  - Nonresidential Coverage Evaluation (NCE)
- Information on new, privately owned construction is often available from building permit jurisdictions
- Investigate feasibility of using publicly available building permit data to supplement new construction surveys

Research and Findings

- Chicago, IL and Seattle, WA building permit jurisdictions
  - Data available through APIs
  - Initial research indicated that these sources provide timely and valid data with respect to BPS
  - Definitional differences and insufficient detail to aid estimation
- Seven additional jurisdictions across the country
  - Data come in other formats
  - More standardized classification data items
  - Lack of information regarding housing units

Challenges and Future Work

- Challenges of using online building permit data
  - Representativeness
  - Consistency of data formats and terminology
- Future work
  - Ongoing validation of data compared to survey data from BPS, SOC, and NCE
  - Use of third-party data sources Zillow and Construction Monitor

Efforts to Improve Sampling Frames

- Scrape location and contact information for
  - Juvenile facilities
  - Franchisees and franchisors
  - Tax collectors
- Work done by Economic Directorate, Civic Digital Fellows, and Center for Economic Studies

Next Steps with Web Scraping

- Use SABLE in production
- Release a data product based in part on scraped data
- Scrape data from SEC's online database
- Look for guidance from a newly formed Census Bureau-wide working group to address policy issues regarding web scraping and web crawling

Contact Information

- Brian.Dumbacher@census.gov
- Carma.Ray.Hogue@census.gov