Congress.ai: Decentralizing the People's Branch

The problem: There is no single, comprehensive repository for research reports created by legislative branches of national governments around the world. Free resources are topically focused, and existing paid repositories typically host reports created by a single country. Some countries do not make their reports easily accessible. For example, before the US government's launch of crsreports.congress.gov, there was no official US Government website providing CRS reports to the general public. Still, the official website is missing older reports that are available from other sources. For a researcher, gathering comprehensive resources is a time intensive endeavor that prohibits the kind of comparative research that would promote good governance.

Proposed solution: Congress.ai is a project that will address this information gap by creating the first free resource of all nations' legislative research reports, preserved permanently and made available to all using IPFS. Congress.ai will include a website allowing journalists, researchers, activists, and others to easily search through the corpus to compare and contrast reports relevant to their work or interests. By uploading nation collections of reports to IPFS Congress.ai and anyone on the network can tenaciously access and host replications

Conceptually and practically speaking, parliamentary policy reports of the world are an ideal use case for IPFS because its decentralized state reflects the values and decisions of individuals as does the parliamentary branch.

Anticipated impact: Using IPFS to store and access these important documents will ensure that people can access all the parliamentary reports. This will enable any researcher to confidently focus on analysis and good governance recommendations rather than undertaking a complex resource gathering process. Government archivists will likely make use of IPFSs redundant hosting per LOCKSS principles which will increase the tenacity and speed at which others will be able to access the collection.

Timeline and budget:
We propose a 3 month timeline to build an initial prototype of Congress.ai. Our prototype will focus on 4 countries, and will involve two phases. In Phase One, we will ingest and label reports from official government websites, and use them to train a classifier to detect reports in a similar format. We will retool our existing codebase for our crsreports.com website to interface with IPFS. In Phase Two, we will begin applying our classifier to large datasets (such as the Internet Archive's corpus of PDFs) to identify and extract additional reports. Creating a unique and thorough collection requires dedication to resolve all of the technical, linguistic and challenges of scale that will arise, some solutions I will be capable of addressing on my own, others will require additional hardware, engineering and region specific archival expertise.
This project is independent, however it can also be seen as phase 2 of a larger effort to provide universal access to all law. Phase 1 CRSreports.com (detailed below) focused on US Congressional research reports.

Antoine@crsreports.com
Twitter @AGreenDCBike

Congress.ai will focus on global parliamentary reports. The tool and procedures developed for phase 2 can be applied in phase 3 beyond the focus of global parliamentary aggregation and onto a global judicial focus. Phase 3 would result in an all laws global repository project.

Expense estimate $30,000 USD

- OCR services
- H2O.ai license
- Cloud Provisioning
- Full Time engineer
- Supplemental Engineer
- Project Mgr Supplemental Expenses
- Crowdwork Budget (Mechanical Turk)
- International Federation of Library Associations and Institutions (IFLA) Membership
- Website
- ML Classifier based on 4 countries, 1 classifier set per country
- ML Libraries, training classifiers:
- Website build backed by database

Protocol Lab & Community Requested Resources

- Textileio guidance for website and database hosting
- Engineering guidance from IPFS: IPFS native website and database hosting, estimated 10hrs a week for the last 6 weeks
- IPFS storage, and country specific replications

A note on my commitment & experience

Post Internet Archive I designed a related workflow to Congress.ai for the US congressional research service which resulted in CRSReports.com This resource was profiled in the Washington Post and has become a cited resource by more than a dozen law libraries and internally by several US government agencies.

Antoine@crsreports.com
Twitter @AGreenDCBike

Contents

Antoine@crsreports.com
Twitter @AGreenDCBike

Aggregation plan

In order to aggregate and host research materials from "the people's branch" of as many nations as possible Global Government Researchers:

1. Accession as many lists of PDF urls as possible.
2. From this list we can prioritize the URLs that appear to host wanted content.
3. Download the URLs from publicly web archives
4. Identify reports from Congressional/Parliamentary research institutions
5. Improve prioritization and identification process
6. Host publically for human and programmatic consumption API
7. Decentralize for broad access

Download Prioritization

Initial prioritization will be based on regex (as was done for CRSreports.com), however a NLP ML method should be developed from the resulting positive matches.

Initial prioritization

1. Identify an online cache of these reports
2. Create a regex that matches the report URLs

We find congressional research service reports we find them hosted at the URLS:

www.house.gov/rules/RL30725.pdf
www.house.gov/rules/RL40608.pdf
www.house.gov/rules/RB60776.pdf
www.house.gov/congressional_rpts/CRS_95-408_20100413.pdf

A regular expression (regex) that would encompass all of these report URLs is:

R.{6,12}pdf$\|(crs\|congress\|congressional).*pdf'

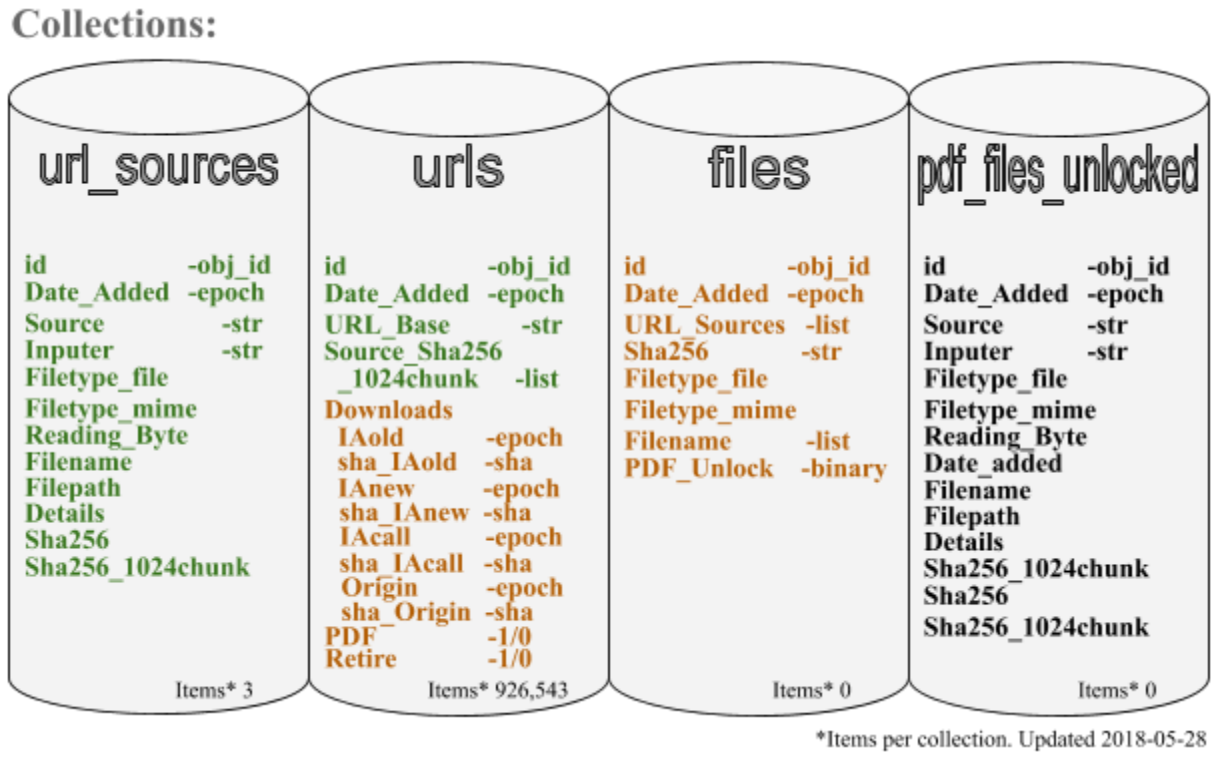
3. Discover and download all URLs matching the regex
4. Run identification of reports
5. Extract metadata from reports
6. Clean reports (as needed: unlock pdfs, remove 3rd party additions (unofficial cover pages, marginalia))
7. Create file conversions and metadata documentation
8. Upload to IPFS

Antoine@crsreports.com
Twitter @AGreenDCBike

Identification of reports

Initial report identification will be based on expected content within manually described bounding boxes (as was done for CRSreports.com), however a visually based ML method will be developed. Some early visually based ML tests on identifying CRS reports have shown promise, alternatively a more traditional but computationally heavy method of OCR'ing every PDF may be used.

Database congressai



Antoine@crsreports.com
Twitter @AGreenDCBike

Accessioning Scripts:

url_sources__add_source.py

-Accessions lists of URLs and documents list source to db collection url_sources
-Generates item for each URL in the db collection urls

download_urls.py

-If not retired nor PDF, script attempts cache downloads leaving epoch as trace
-Creates file record for each download attempt and determines file type
-Updates URL item

metadata_extraction.py

-Creates priority rank for downloading URLs

prioritize_downloads.py

-Creates priority rank for downloading URLs

tools.py **#Tools**
dbcontrols.py **#DB Controls**

Progress notes

URL Lists for Import

"InternetArchive"

"From Ubuntu Backups"

X /mnt/8TB/backups/backups/Backups_2015_Fall/2015 Jan Backup/Dropbox/Merged lists/MASTER_preface.txt

X /mnt/8TB/backups/backups/Backups_2015_Fall/2015 Jan Backup/Dropbox/Merged lists/1,508,392_no-dups_preface.txt

X /mnt/8TB/backups/backups/Backups_2015_Fall/2015 Jan Backup/Dropbox/Merged lists/Individuals/001,476,537_no-dups_preface.txt

/home/comp/Desktop/Master_URL_1of3.txt

/home/comp/Desktop/Master_URL_2of3.txt

/home/comp/Desktop/Master_URL_3of3.txt

/media/comp/320GB/urls/wbm-pdf-urls-20131106004342.txt

Redundant URL lists to discard once complete

Antoine@crsreports.com
Twitter @AGreenDCBike

441M /mnt/8TB/backups/backups/Backups_2015_Fall/2015 Jan Backup/Dropbox/Merged lists/MASTER_preface.txt

395M /mnt/8TB/backups/backups/Backups_2015_Fall/2015 Jan Backup/Dropbox/Merged lists/Individuals/MASTER.txt

325M /mnt/8TB/backups/backups/Backups_2015_Fall/2015 Jan Backup/Dropbox/Merged lists/Individuals/MASTER_nopreface.txt

149M /mnt/8TB/backups/backups/Backups_2015_Fall/2015 Jan Backup/Dropbox/Merged lists/Individuals/1799890_Lines_memo.txt

138M /mnt/8TB/backups/backups/Backups_2015_Fall/2015 Jan Backup/Dropbox/Merged lists/1,508,392_no-dups_preface.txt

135M /mnt/8TB/backups/backups/Backups_2015_Fall/2015 Jan Backup/Dropbox/Merged lists/Individuals/001,476,537_no-dups_preface.txt

128M /mnt/8TB/backups/backups/Backups_2015_Fall/2015 Jan Backup/001,476,537_no-dups_preface.txt

107M /mnt/8TB/backups/backups/Backups_2015_Fall/2015 Jan Backup/Dropbox/Merged lists/Individuals/1,585,900_condensed_nopreface.txt

/home/comp/Desktop/Master_URL_1of3.txt

/home/comp/Desktop/Master_URL_2of3.txt

/home/comp/Desktop/Master_URL_3of3.txt

Directory structure

Some (not all) notable directories and files:

- scripts/ - Scripts tasks and configuration for building and populating the database
- staging/ - Temporarily storing files before bulk transfer to external hard drive

API access suggestions or support

Tips: https://babolabs.blogspot.com/2011/08/using-restful-apis.html

Replication - Tenacious access:

- IPFS
- Internet Archive, NY Gov Labs, Stanford Gov Docs

Notifying researchers & academics

- Noted opportunity: Research effective influence by nation alliances & supranational memberships

To be examined

Institutional partnership suggestions

- IPFS
- Internet Archive, NY Gov Labs, Stanford Gov Docs

Antoine@crsreports.com
Twitter @AGreenDCBike