DATA
COALITION

Data Coalition
1003 K Street NW
Suite 200
Washington, D.C. 20001

☎ 312.493.7533
@DataCoalition
✉ info@datacoalition.org
datacoalition.org

October 13, 2017

The Honorable Gregg Harper, Chairman
The Honorable Robert Brady, Ranking Member
Committee on House Administration
United States House of Representatives
1309 Longworth House Office Building
Washington, DC 20515

Re: Transforming GPO for the 21st Century and Beyond: Part 3 – Federal Depository Library Program

Dear Chairman Harper and Ranking Member Brady:

On behalf of the Data Coalition, I am pleased to submit these comments for the record in response to the committee's September 26th, 2017, hearing entitled, "Transforming GPO for the 21st Century and Beyond: Part 3 – Federal Depository Library Program."

The Data Coalition was founded in 2012. We represent forty-two technology and consulting companies, employing over two hundred thousand Americans. Fourteen of our members are startups founded within the last decade and ten are public companies. All of our member companies support the publication of government information as machine-readable, open data.

The publication of government information as standardized, machine-readable data, is not only vital to democratic accountability – it also presents a tremendous opportunity for technology innovation driven by companies as represented by the Data Coalition's membership.

Title 44 of the U.S. Code ensures that government information is accessible to the public, but it does so using print-era terminology and standards. As Rep. Zoe Lofgren (CA-19) and Stanford Law School Library Director Beth Williams discussed in the September 26th hearing, these outdated standards can inhibit technological innovations that could unlock enormous value from the information.

One example that highlights the existing opportunities that lie in datasets consisting of government policy documents comes from FiscalNote, a Washington, D.C.-based policy tech company that connects people and organizations to their government. FiscalNote dramatically improves the way organizations build and manage their relationships with all levels of government, and empowers them to have maximum impact on legislation and regulation. The company pulls from the GPO's bulk data, isolates and extracts key text and metadata fields. FiscalNote's software utilizes cutting-edge data science techniques to help organizations identify all policy issues that could have a material impact on their issues and provides such analytical insights as legislative language trends and the likelihood of a bill passing.

As a result, organizations and companies are in a better position to keep abreast of policy risks posed to their lines of business, drive down regulatory compliance costs, and more nimbly adjust business strategies to account for the changing policy landscape.

Bulk Data

Key to these capabilities is access to timely, comprehensive, and well-structured data from the GPO's bulk bill data repositories. Currently, this data is split into two principal repositories. One contains bill texts, while the other contains bill metadata such as sponsors, bill actions, and committee referrals. Each repository contains a simple list of the individual bill files in eXtensible Markup Language (XML) format, with clearly-indicated timestamps that denote the most recent update for each individual file.

XML is a great method to deliver data in a machine-readable format, as it imposes a uniform and semantic structure to the data. This structure mitigates uncertainties and speeds development of any application or tool that desires to utilize this data. Due to its consistency, XML files can be read and utilized by a wide variety of companies, organizations, and citizens-- including hobbyist and academic researchers.

Bulk data is also simpler to access and use than an Application Program Interface (API), as it doesn't require the user to make thousands or millions of individual API calls for each piece of data. The entire dataset is delivered at once, which is faster to develop, more efficient to implement, and easier to analyze.

While fairly comprehensive, the GPO's bill text and bill metadata bulk data dumps could still be improved. For example, bill metadata files should include referential links to its matching bill texts within the bill text bulk data dump; right now, there are no explicit links between a bill metadata file and its associated bill text files. Alternatively, the GPO could combine the two bulk data dumps into a single dump which contains both bill texts and bill metadata. Unifying the two dumps into a single dump would simplify the acquisition of this data.

In addition, the GPO should add certain pieces of missing metadata into its bill metadata bulk data files. For example, bill voting rolls are currently not included within the metadata files. Nor are amendment texts or Congressional Record references.

APIs

However, well-structured, thorough, and well-documented APIs are still an acceptable machine-readable delivery for data.

For these APIs to be maximally useful, the APIs should:

- Make as many fields available via the API as possible. An API that only makes available a handful of fields is not very useful, as the developer would then have to assemble data from multiple sources.
- Offer a flexible search endpoint that enables the widest variety of filtering, sorting, pagination, and field selection options, or offer an endpoint that returns every row of data.
- Explicitly incorporate the version number in the API URL. This minimizes confusion about versioning.
- Explain errors via HTTP-standard status codes, such as 400 (Bad Request), 401 (Unauthorized), 403 (Forbidden), 404 (Resource Not Found), and 429 (Too Many Requests).
- Encode any preferred API rate limits in the HTTP response headers.
- Offer a variety of input and output formats, such as JSON, XML, and YAML.
- Be accompanied by well-written, thorough, and updated documentation, with version notes and changelogs. In addition to a list of fields and parameters, this documentation should also include expected characteristics and constraints for each field and parameter.

Equally Open

No matter the choice of delivery mechanism (e.g., APIs or bulk data dumps), another crucial factor in the purveyance of this data is the idea that the GPO should aim to make its data open and usable from the very outset. In other words, it is much easier to design a system in which data is open and usable from the very beginning, rather than retrofitting or altering an existing Closed Data system to make it public.

Designing a system that is equally open from day one prevents the formation of certain privileged client-provider relationships, and enables developers the maximum flexibility to create their own unique products atop the underlying data. For example, imagine the GPO desired to publish Congressional hearing testimonies. In this scenario, the GPO partners with Congress.gov to publish these testimonies. In doing so, it has created a privileged relationship with Congress.gov; no other users can access these testimonies except through Congress.gov.

Instead, the best approach would be for GPO to publish these testimonies as bulk data dumps on its own website in addition to any other partnerships. By publishing this data in an open format, it is now equally available for all users, including Congress.gov's users, to create their own applications atop this data. Rather than privilege a certain outlet, every user has an equal claim to this data.

Additional Policy Documents to Consider Including in GPO Bulk Data:

- Amendments
- Votes
- Hearing documentation:
    - Public testimony
    - Supporting hearing documentation
    - Transcripts to hearings
    - Witness "Truth-in-Testimony" statements
- Floor transcripts
- Dear Colleague letters

Summary

The potential for innovative companies like FiscalNote to serve the public includes a future, in which all government policy documents are available in a machine-readable format so that textual analytics can deliver new insights and organizations can better understand their government.

To unleash the economic potential of this vast public resource, the Data Coalition supports expanding free and open access to the information disseminated by the GPO to federal depository libraries and public. Reforms should guarantee free public access, ensure that agencies provide all applicable digital information to GPO, and direct GPO to make that information available to the public in open and bulk formats with supporting metadata to the greatest extent possible.

The transformation of the Federal Depository Library Program (FDLP) will require a rethinking of how information is collected, organized, published and distributed. This should include a review of open data standards, and how these standards can be incorporated in a 21st Century model for the FDLP. The GPO should be directed to initiate this process, in cooperation with FDLP libraries and members of the public including the tech community.

The Data Coalition supports the FDLP's mission to ensure the public's access to government information. We believe that modernizing the FDLP and Title 44 can both strengthen this core mission and spur innovation that drives economic growth.

Thank you for this opportunity for the Data Coalition to submit comments for the record in support of modernizing government information.

Respectfully,
Christian A. Hoehner

Director of Policy
Data Coalition