Progress® MarkLogic®

U.S. GPO Deploys Progress® MarkLogic® to Modernize Document Publishing Capabilities

CASE STUDY

GPO

INDUSTRY
Publishing
Public Sector

PRODUCT
Progress® MarkLogic®

COUNTRY
United States

SUMMARY
Driven by its mission to publish trusted information of the U.S. Congress and congressional agencies for access by the American people, the U.S. Government Publishing Office releases millions of documents and publications annually. With a broad mandate to modernize and make content simultaneously available in various printed and digital formats, the agency chose Progress MarkLogic to help replace a legacy-based file management system.

"The MarkLogic platform gives us the tools we need to develop custom software for managing and publishing large volumes of rich XML content."

Martin Smith, Manager
XPub Department, GPO

Challenge

Established in 1860, the U.S. Government Publishing Office (GPO) is responsible for publishing all congressional bills, resolutions, amendments, and daily publications such as the Congressional Record and the Federal Register for public access.

In the 1970s and 80s the agency made a gradual transition from paper to electronic document production and distribution—and some of that technology remains in place today, therefore the present mandate to modernize technologies and work processes.

"We have a legacy file-based system where work flows from one shared directory to another to support the editorial process," explained Martin Smith, Manager, GPO, XPub Department. "The mission of the XPub department is to replace legacy tools and work processes with a system that enables the GPO to mark-up content in semantically rich XML and simultaneously publish and distribute that content in a variety of print and digital formats, including XML, PDF and responsive HTML."

In addition to modernizing work processes, there is a historical preservation aspect of the project, as well as the issue of scalability.

"We have employees working around the clock over three shifts," Smith explains. "If Congress is in session late, they might not adjourn until 8 o'clock at night. We're getting our last content then and it has to be proofread, typeset, corrected and turned into a book-length document that might be longer than 1,000 pages. We have to print and deliver it by 8 o'clock the next morning and put it online as well, so we've got to scale it."

That scalability comes with its own challenges. Smith explained that a typical XML editor could not provide the scalability the agency requires.

"For example, if you paste in a partially structured Word file directly into that editor, you're going to spend several days fighting with it because it's completely invalid against the schema," he said. "That's not going to cut it when we've got thousands of pages of content to mark up and publish—that kind of a workflow just simply doesn't scale."

Solution

To achieve that level of scalability, the GPO needed the ability to mark up all routine publications in semantically rich XML and then simultaneously publish and distribute this content in XML, PDF and responsive HTML formats. This capability enables intelligent search, retrieval and aggregation of content that the GPO makes available to the public.

Additionally, responsive HTML enables the public to conveniently access federal content on computers, tablets and mobile devices. PDF format enables the GPO to publish documents in printed form, as it has done since the founding of the agency. The ability to publish routine publications in XML and responsive HTML formats is a capability that the GPO does not have today, due mainly to the difficulty of marking up XML documents at scale.

These capabilities will be achieved with the new XPub system currently under development, enabled by the Progress® MarkLogic® platform's content management capabilities, which will enable the GPO to provide additional benefits to the public.

Results

The XPub team is currently focused on modernizing the publication of the Federal Register. This document is published daily, can run as large as a thousand pages and consists of content submitted by multiple federal agencies, typically in Microsoft Word format.

Each document will need to be marked up in XML, proofread, corrected and then assembled into the current edition of the Federal Register. Typically, the Word files are too large to be worked on by one individual, so GPO plans to use the MarkLogic platform to initiate workflow processes on section-level XML components.

GPO has developed software that converts Word files to XML and stores them in the MarkLogic data platform. The XPub client will provide supervisors and employees with views into the XML content managed by the platform. Supervisors will use the XPub client to monitor incoming work, then assign, review and approve the work. Employees will use the XPub client to receive, complete and return work. Finally, supervisors and employees will use the XPub client to assemble individual documents into an edition of the Federal Register and generate final XML, PDF and responsive HTML versions of the content for publishing and distribution.

Essentially, XPub implements document loading and check-in/check-out operations through a middle-tier REST service that integrates with MarkLogic's REST APIs. Other REST services convert Word files to XML, run XSLT transforms, manage workflows and generate PDF and responsive HTML output. GPO is also developing custom, enterprise software in-house to meet its unique business needs.

"The MarkLogic platform gives us the tools we need to develop custom software for managing and publishing large volumes of rich XML content," Smith explains. "It provides the enterprise-scale capabilities we need for capacity, performance, security and reliability."

Once XPub is in full production for the Federal Register, the GPO will have the opportunity to curate a large dataset of before and after XML files stored in the MarkLogic platform from which the agency intends to train AI models to provide user assistance tools for XML markup and proofreading content.

About GPO

The U.S. Government Publishing Office is responsible for publishing all congressional bills, resolutions, amendments and daily publications, such as the Congressional Record and the Federal Register for access by the American people.