EPA Enterprise Records Management Division (ERMD) Content Ingestion Service APIs

Overview

The U.S. Environmental Protection Agency's (EPA) utilizes the Enterprise Content Management System (ECMS) to manage the retention of the Agency's electronic records. ECMS leverages Documentum as it's centralized records management platform. The Enterprise Records Management Division (ERMD) is developing Content Ingestion Services (CIS) to facilitate the assignment of record schedules to Agency records housed within ECMS. These series of web services may also be used by other EPA applications that have a need to consume and categorize records within their system.

The goal of ERMD is to provide the ability for application owners with minimal Documentum experience, to easily interface with ECMS and submit their records. CIS include the following:

- Document Transformation – Uses optical character recognition (OCR) technology to ensure scanned documents are fully searchable
- Text Extraction – Extracts and summarizes documents based on content within the document
- Categorization – Takes summarized text and suggest a records schedule/file plan based on machine learning
- Content Ingestion – Submits a record to ECMS and checks to see if the record exists and if there are changes. If changes are detected, a new version of the record is saved

The Challenge

ERMD's objective in developing CIS is to facilitate the process of submitting EPA's electronic records to ECMS by ensuring that all electronic records are appropriately assigned a records schedule. The CIS APIs also provide an end-point for application owners to be able to easily create interfaces between their applications and ECMS, to ensure compliance with EPA and federal records management requirements.

Alternatives Considered

ERMD worked with the ECMS user community to identify a set of short term improvement goals. This effort allowed them to break the opportunities into small, incremental pieces that could demonstrate value.

ERMD is committed to an open-source approach in order to avoid licensing costs. They experimented with and tested different technologies to determine how best to meet the document transformation, text extraction, categorization, and content consumption needs. As part of this analysis, different API Management solutions were considered including NREL's API Umbrella solution. However, ERMD ran into Federal Information Processing Standard (FIPS) compliance issues with API Umbrella and ultimately

selected apiman (JBoss' open source API Management system). The following is the full set of solutions supporting ERMD's Content Ingestion Services:

- apiman
- Postman (testing)
- Apache Tika (text extraction)
- Tesseract (OCR)
- spaCy (natural language processing)
- DeepDetect (Deep Learning platform)
- Python specific scripts developed to perform data cleansing and transformation operations and made publicly available on GitHub

Implementation Approach

ERMD is following an incremental approach for implementing CIS and other improvements to ECMS. Refer to the discussion of Alternatives for more information.

Best Practices and Lessons Learned

In describing their API strategy and development approach, ERMD identified the following best practices and lessons learned that may benefit other development teams:

- **Establish API Managers and Product Owners** – Establishing a central organizational role of API Manager is beneficial. The API Manager can ensure an organization sets consistent API management policies, standards, tools, and templates across the organization. The role of the API Product Owner is also critical as they ensure API development efforts reflect true customer business needs and priorities.
- **Advocate for "API First"** – API Development Teams must be advocates for their services and communicate the benefits of an "API First" philosophy. There needs to be coaching and mentoring opportunities for both technical and non-technical project managers and their teams to help them understand the benefits of APIs.
- **Provide Hands-On Training** – Providing developers with hands-on training is beneficial to show them best practices for developing and integrating APIs with commonly used technologies. A lot of different programming languages are used at EPA and other organizations and each have different ways of interacting with API requests. Recorded demonstrations could be made available via a collaboration site as well.
- **Encourage Collaboration** – Providing opportunities for collaboration and an API Integration Platform are beneficial. API Integration Platforms must provide the following:
    - Security – Tools to manage user authentication and API keys
    - Analytics – Tools to analyze which APIs are being used and who is using them
    - Cache – Support for call caching
    - Documentation – Tools to document APIs
    - Collaboration Forums – Tools or forums to support communications with developers and potential customers on new APIs, capabilities of existing APIs, discussions on best practices, etc.
    - Focus on both internal and external APIs – not just external.

Contacts

For more information on ERMD's API strategy, please contact the following:

- **Enterprise Content Management System (ECMS) Program Manager**: Andrew Yuen
  Email: yuen.andrew@epa.gov