Optimizing Historical Knowledge Retrieval: Standardizing Metadata for Enhanced Research Access

The Challenge

A federal research and development center leverages outcomes from past projects and conducts experiments to improve present-day research on innovative scientific solutions. Most of these past reports are stored in a document repository to be made available to technical researchers and analysts. However, these researchers and analysts struggled to find the reports that were relevant to them because the document system lacked standardized metadata. A researcher would search for a report about a subject, but there was no guarantee that the document they sought would be tagged with that subject, or that other relevant documents would be returned in the search results. Meanwhile, the technical team working to upload both historic and new reports to the document management system had to manually add descriptive metadata such as Author, Subject, and Classification to each document. This meant that the document depositing process was extremely time-consuming and highly prone to errors. Furthermore, the Division's overall architecture suffered from information silos, with each system leveraging its own metadata model.

This lack of standardization for descriptive metadata fields made searching for documents, performing records management, and ensuring long-term information preservation difficult, These challenges resulted in loss of institutional knowledge, operational inefficiency, and potential poor decision-making due to inability to have reliable access to information.

The Solution

To help standardize descriptive metadata in the document repository systems, EK partnered with the Knowledge Management and Knowledge Apps team to tackle four key initiatives:
Integrate the Taxonomy & Ontology Management System (TOMS) with one of the primary document management systems at the organization to serve as the Semantic Layer representing contextual and descriptive knowledge models;

1. Extend the TOMS to integrate with existing gold sources (sources of truth) for models; and
2. Develop a solution to support automatic application of metadata to documents and streamline the document deposit process (auto-classification).

To begin working towards the first two goals, the research center and EK teams worked as one to develop use cases and prioritize the metadata fields for integration.

Optimizing Historical Knowledge Retrieval: Standardizing Metadata for Enhanced Research Access

The Solution

The integration use case focused on integrating the Subject and Author fields of technical research documents, two fields that were key for improving the searchability of content within the document management system. Additionally, the team also identified gold source models, which are existing models for key metadata fields such as People (authors) and Facilities at the organization.

After defining and documenting the use case, business requirements, and technical requirements, the EK team partnered with the organization's development team to build an API that would integrate the TOMS with the document management system and allow the TOMS to replicate existing gold source models. The resulting API's functionality is twofold:

1. It serves as an API abstraction to enable seamless integration, allowing data to flow smoothly between the TOMS and the document management system. This enables the document management system to consume standard metadata, while also enabling document depositors to submit new concepts to the TOMS model for approvers to review.
2. It supports ETL functionality to ensure that data remains synchronized between the internal systems and the TOMS. This "gold source replication" ensures that metadata is perpetually updated and connected across systems and is extensible to additional gold source models.

For the third initiative, streamlining the document deposit process and standardizing descriptive metadata in the document repository systems, EK architected and developed an auto-classification proof of concept (POC) which built on existing systems in the organization's environment. Built over 6 weeks, this POC took a subset of reports and leveraged an LLM and the TOMS to run auto-classification and produce a list of recommended subjects for each document. These recommended subjects provided the document depositors with a shortened list of subjects to choose from and ensured that the subjects adhered to a standardized model stored in the TOMS (e.g., the list of subjects was a controlled list/taxonomy). Finally, the EK team documented the auto-classification lifecycle and provided technical documentation for the API so the organization's development team could continue building on and enhancing auto-classification and integration capabilities.

The EK Difference

EK's extensive experience in semantic solutions, data engineering, semantic search, and content management enabled the team to deliver the API integration and gold source replication ahead of schedule and under budget. The EK team provided comprehensive code repository documentation to the organization's development team. Additionally, the EK team ensured there was continuous knowledge sharing between the EK and organization (Continued)

The EK Difference

development teams, enabling the team to continue expanding work after the engagement. The EK team also brought expertise in strategic planning and management and used this expertise to improve the team structure and project tracking during the engagement.

Finally, because the work was completed ahead of schedule, the EK team went above and beyond to guide the Knowledge Management Team in optimizing the solution through augmentation with a custom search front-end. The EK team facilitated several requirements-gathering sessions and developed initial wireframes to drive alignment and validation of use cases for search capabilities. These strategic documents can help drive business-buy in as the client team continues building on the metadata standardization work.

The Results

As a result of their engagement with EK, the research and development center has successfully integrated 3 models with their primary document repository, enabling 4 key metadata fields to pull from a standard set of values. This integration ensures that the 100,000+ documents in the document management system are using a consistent and correct set of metadata fields and values. The 4000+ users of the document management system can now better find and understand historical research documents, thereby increasing operational efficiency for new research efforts.

Furthermore, the organization will be able to continue to leverage the API to integrate the TOMS models to additional systems, enabling standardization of descriptive metadata across repositories and reducing manual effort and data anomalies. Additionally, the API enhancement allows document depositors to contribute to the accuracy and currency of the OMS models that support the metadata fields.

Finally, the auto-classification proof of concept and use definition for a search front-end empowers the organization's Knowledge Management team to advocate for future capabilities and initiatives within the Division. This comprehensive effort establishes the groundwork for the organization to improve the search and findability of documents and pursue additional data transformation capabilities.