Optimizing Historical Knowledge Retrieval: Extracting Knowledge by Making Connections

The Challenge

A Federally Funded Research and Development Center (FFRDC) faced significant challenges with low-quality or incomplete metadata for managing and cataloging scientific reports, hindering researchers' ability to parse repositories and efficiently discover relevant content. As detailed in a previous case study, *Optimizing Historical Knowledge Retrieval: Standardizing Metadata for Enhanced Research Access*, Enterprise Knowledge (EK) partnered with the organization to improve the findability of these valuable research materials. EK employed complementary approaches to clean and enrich the content metadata and developed a proof-of-concept (POC) application to automatically classify documents with "about-ness" concept tags (auto-tags) from established taxonomies. The delivered POC is capable of processing batches of local files and returning a set of auto-tags for each document.

Following the successful demonstration of the POC, EK was engaged to advance the solution to a production release. The application needed to scale to enterprise-level workloads, processing hundreds of thousands of documents and publishing enriched metadata back to the source Document Management System (DMS). After unlocking meaning through semantic enrichment and metadata standardization, EK would extend the architecture to implement a document-centric knowledge graph with a user-friendly interface. This graph representation of the DMS would create a network of documents and entities derived from metadata tags, enabling additional discovery through their connections and added context.

The Solution

Effective auto-classification depends on a well-structured semantic foundation and content that is clean and standardized. To achieve this, EK needed to address inconsistent metadata spanning decades of research outputs and validate the relevance or gaps of available taxonomies. EK processed these components in parallel, using iterative tuning and assessment between taxonomy design and classifier performance to achieve meaningful, high-precision tagging.

Taxonomy & Semantic Modeling

Production classification targeted seven distinct domains. EK's taxonomy experts fulfilled this scope by helping the organization identify and integrate priority taxonomies in the client Taxonomy Management System (TMS), including established taxonomies, client institutional taxonomies and new custom taxonomies developed or refined by EK for this application. EK collaborated with the organization's subject matter experts to align labels and hierarchy with language that researchers actually use to describe the content. By calibrating the comprehensive semantic model to the specific use case, the team increased the breadth and depth of auto-tagging scope while reducing noise from vague or less informative tags.

To support continued evolution of the semantic models, EK enabled automated extraction of new concepts within select domains during the auto-classification process. A repeatable human-in-the-loop workflow for candidate concept review was integrated with the TMS. Approved concepts are added to the taxonomy, while denied or modified concepts are edited appropriately in the metadata store. Addition of this capability increases flexibility and supports a corpus-driven expansion of current taxonomies while safeguarding semantic accuracy and governance.

AI Readiness

To prepare content for classification, EK analyzed existing metadata to identify the fields most relevant to each target domain and removed low-quality or redundant metadata. The metadata structure was standardized to align with requirements for a new Document Management System (DMS) release, ensuring consistency and interoperability across repositories and future use cases.

To address the sometimes noisy text outputs from the OCR tool, EK deployed a client-side Large Language Model (LLM) to generate summaries and alternative titles to supplement existing metadata. For more on this use case, see the related case study: _Optimizing Historical Knowledge Retrieval: Leveraging an LLM for Content Cleanup_.

Enterprise AI

With clean, standardized content and a rigorously vetted semantic model This application uses a Natural Language Processing (NLP) auto-classification engine to apply concept tags at scale. The semantic NLP classification service performs concept extraction and classification using linguistic analysis and taxonomic rules. These rules can be configured at multiple levels of granularity, including per domain, per concept scheme, or per input field. Through iterative experimentation and review, EK optimized classifier configurations. This tuning ensured that the classification service prioritized the most significant signals when scoring concepts for each domain, delivering high-precision, context-aware results. Auto-tagging based on a semantic model yields tags paired with the persistent identifier for the tagged concept in the TMS. Metadata fields that once held free-text values (for example, author or publisher) are now standardized to a preferred label, with reference to the defined concept in the source taxonomy.

Enabling Connected Discovery

With standardized, semantically enriched metadata replacing years of inconsistent records, EK's next step was to help the organization move beyond document management toward exploring the relationships between research assets. Concepts bound to taxonomy definitions elevates tagged metadata values into resolvable entities - no longer simply attributes of documents, but integral components of a broader knowledge network. By making these connections visible and easily queryable with a user-friendly front-end application, the organization can uncover patterns, trace topic evolution, surface expertise, and identify collaboration opportunities across domains.

Knowledge Graph & Data Modeling

For this solution, EK designed a document centered ontology that models how documents and key metadata domains are related. Metadata that were previously static within the DMS become entities in the graph with their own connections. A knowledge graph brings together data from previously separate data sources and connects it in an informative, navigable way. This approach supports advanced use cases such as cross-domain analysis, expertise identification, collaboration analysis, and contextual exploration of research relationships that are otherwise hidden within the flat structure of the DMS.

Enterprise Search

To provide researchers with access to these graph insights, EK implemented a companion front-end application powered by a purpose-built API to query the graph database. The interface offers functionality similar to that of academic search engines such as Google Scholar. Users can execute natural language queries, explore connection chains, and filter based on taxonomy concepts to discover related content. This comprehensive solution not only enhances discoverability for researchers, but also transforms a static document repository into a dynamic, interconnected knowledge ecosystem.

The EK Difference

A notable highlight of this project was the strong partnership EK established with the client. Acting as an intermediary among cloud platform consultants, the client's developers, and their semantic team, EK ensured that all components of the project were fully integrated and aligned throughout the development process. Through ongoing knowledge-sharing sessions and demonstrations across the organization and to client leadership, EK showcased the solution's current and future capabilities to help build interest and support within the organization. Fostering an open, collaborative environment not only enhanced EK's understanding of the client's immediate needs but also allowed EK to help position the client for continued success as their data landscape evolves. This close working relationship enabled a smooth handoff of the solution and a client team ready to scale to additional repositories and domains.

- **Semantic Modeling & Implementation:** Taxonomy design and refinement, Taxonomy Management System (TMS) integration, and semantic governance planning
- **Content AI Readiness:** Content quality assessment, metadata standardization, and readiness strategy
- **Enterprise AI:** AI-augmented content cleanup and natural-language-processing(NLP)-driven auto-classification at scale
- **Knowledge Graph & Data Modeling:** Ontology design, data mapping, and graph-database integration
- **Enterprise Search & Discovery:** Search and discovery user experience design, custom APIs, and front-end user interface implementation

The Results

Previously, client research content was sparse and often inaccurately cataloged, resulting in poor content discovery and significant gaps in information access. Now, standardized and enriched metadata combined with the added context and query power of a knowledge graph, the full scope of the archival repository has become far more accessible. Researchers can better find target content and explore connections between entities through the new user interface, uncovering insights about experts, affiliations, and subject areas. With EK's expertise, the client has expanded their data strategy beyond data management and into true knowledge discovery, preserving the organization's institutional knowledge for future innovation.

Organizations facing similar challenges in managing and accessing complex or archival content can benefit from Enterprise Knowledge's expertise in optimizing content discovery through metadata enrichment, AI-driven classification, and knowledge graph solutions.