

The Accidental KM Practice

MICHAEL IANTOSCA

SENIOR DIRECTOR OF KNOWLEDGE MANAGEMENT AND CONTENT SYSTEMS ENGINEERING

AVALARA INC.

Introduction

Five years ago, I retired from IBM after thirty-eight years leading enterprise product documentation strategy and technology. A few months later, I was invited to assist Avalara, an AI-driven tax computation company, in transforming its outdated content supply chain into a structured documentation model based on DITA, an open XML standard widely adopted for advanced documentation production and automation, at scale.

Having founded the team at IBM that created the DITA standard, I was hardly new to the field. My early career had been shaped by Dr. Charles Goldfarb, the inventor of SGML and the intellectual father of structured markup languages. Designing and implementing large-scale publishing systems had long been second nature. At IBM, I architected multiple generations of documentation systems, including one that managed sixty million pages of content across 3,500 products, 1,500 technical writers, and more than fifty languages.

The Transformation

At Avalara, I knew that semantic structure and metadata would be the foundation of success. These elements are essential not only for automation but for content discovery, reuse, faceted search, and intelligent content delivery. Building the publishing platform itself was the (relatively) easy part; what mattered most was embedding intelligence into the content itself.

Our team completed the transformation in under eighteen months, standing up a new structured component authoring and workflow system in record time - an omni-source/omni-channel content delivery platform, an AI-driven editorial quality system, and a national-language translation platform for both content and user interfaces. We also created a content operations analytics layer and migrated the entire content corpus into DITA's topic-oriented XML model - a formidable achievement on its own.

From Structure to Intelligence

Structured documentation is often called "intelligent content" because of its semantic richness and machine readability. It is modular, reusable, and predictable, enabling machine automation

and scalability – and even more so in the age of AI. Yet semantic markup alone is insufficient. True intelligence requires taxonomy-driven metadata, classification systems that add depth, context, and meaning.

We hired a professional taxonomist/Ontologist and soon realized that no centralized terminology base existed. Using the linguistic engine Acrolinx (now Markup AI), we built Avalara's first enterprise terminology foundation. It became the basis for constructing taxonomies and ontologies while also enabling real-time editorial governance, ensuring linguistic and conceptual consistency across content teams.

Building the Knowledge Foundation

Next came the challenge of taxonomy management. Anticipating our need for a knowledge graph to drive AI-based content creation, we selected PoolParty as our semantic platform and began developing universal taxonomies for both enterprise content and the complex sales tax domain. Integrating PoolParty with our component content management system required a custom connector, allowing it to serve as the authoritative source of truth for semantic tagging.

One of the earliest taxonomies we developed focused on product classification. To our surprise, no single authoritative list of the company's products existed, only scattered spreadsheets, none accurate or complete. Using PoolParty, we created a unified and dynamic taxonomy representing Avalara's complex product matrix. As word spread, other business units joined, from marketing and sales to finance and technical support. Our project quickly became the cornerstone of enterprise knowledge management.

The Accidental Knowledge Management Practice

What began as a documentation initiative evolved into an accidental knowledge management practice. Our product taxonomy matured into an ontology connecting multiple business functions, and our team grew into an enterprise center of competence. We established a unified knowledge foundation that supports content, data, and AI across the organization.

The Rise of AI and RAG

When we began our intelligent content journey in 2020, generative AI was still on the horizon. We planned to use structured content to train traditional AI/ML models, much like those behind IBM Watson.

But the sudden rise of generative AI changed everything. Within a week, we built our first retrieval-augmented generation (RAG) chatbot using open-source tools. Recognizing the limitations of purely predictive systems, we sought a deterministic approach built on ground truth.

The DOM Graph RAG Model

Thus was born the Document Object Model (DOM) Graph RAG project. Partnering with Semantic Web Company, we demonstrated that DITA XML - being structured and schema-based - could be transformed into a semantic knowledge graph with minimal effort. Using GraphDB and RDF/OWL, we converted DITA schemas and content into a scalable, trustworthy knowledge graph that linked structure, meaning, and automation.

This DOM Graph architecture now enables new capabilities: SPARQL-based content analytics, semantic retrieval, and AI-driven insights. Beyond RAG, it serves as the foundation for agentic AI: systems of cooperative agents that automate, assist, and make decisions grounded in structured knowledge rather than probabilistic prediction.

Beyond Content: Toward Agentic AI

RAG was merely the opening act. Our team is now designing and building agentic AI systems to automate technical content workflows through dozens of collaborating agents. Some agents operate autonomously, while others keep humans in the loop to ensure quality and compliance.

Yet unlike RAG, which can tolerate some predictive variance, agentic systems demand absolute certainty. Deterministic decision-making requires grounded truth that large language models and vector stores cannot provide on their own.

We are now exploring the integration of business process ontologies with our DOM Graph RAG model to enable deterministic reasoning, supported by data validation mechanisms such as SHACL. Though experimental, this fusion of structured content and structured knowledge promises to deliver a future where AI decisions are not merely intelligent but reliable,

explainable, and grounded in truth.

In the End

The convergence of structured content and structured knowledge marks a pivotal moment in enterprise AI evolution. What began as a documentation modernization project has grown into a model for intelligent, agentic knowledge systems. Sometimes, the most transformative innovations are born not by deliberate design, but by accident, and guided by those who recognize the deeper patterns hidden within the structure of knowledge itself.