THE CASE FOR A SEMANTIC LAYER

Hippolyte Lefebvre, *Research Collaborator, MIT CISR*
Barbara H. Wixom, *Principal Research Scientist*
MIT Center for Information Systems Research (CISR)
Christine Legner, *Research Collaborator, MIT CISR*
Nick van der Meulen, *Research Scientist*
MIT Center for Information Systems Research (CISR)
Cynthia M. Beath, *Academic Research Fellow*, MIT CISR

Probably the biggest challenge we've had is that largely all the GenAI use cases are using unstructured data. . . . We're talking SharePoint data, or mortgage documents. Inventorying them is not an issue, but understanding the quality of the information in the document [is]: is it good enough for GenAI to make decisions? Today, understanding quality is human intensive, and it's not scalable. So, we need ways to augment that, to come up with outcomes that are similar to what we've got in the structured world.

AN MIT CISR DATA BOARD MEMBER[1]

Today, the democratization of data and the proliferation of GenAI tools and solutions[2] requires that data services, AI agents, and business users using natural language interfaces all have a comprehensive view of what data represents, where it came from, and how it can be used. GenAI and analytics tools and solutions pull data from applications and from downstream and external data sources. However, when data is decontextualized from the applications where it was created, it often loses the business meaning, relationships, and rules that were embedded in the original application. This information is critical for producing reliable, business-relevant output using AI.

A *semantic layer* is a system of technologies and techniques that organizes and maintains a consistent and unified representation of data from various sources that is interpretable by humans and machines.

Some organizations are investing in a robust semantic layer to address this problem. A **semantic layer** is a system of technologies and techniques that organizes and maintains a consistent and unified representation of data from various sources that is interpretable by humans and machines. Functioning conceptually between data and the user (human or machine), a semantic layer describes data's business meaning, its relationships with other data, and the rules the data must adhere to. MIT CISR research[3] indicates that leaders are feeling pressure to increase their investments in semantic technologies and techniques that can store metadata and instantaneously make data assets more useful to business users, AI models, and AI agents. Key enterprise practices that build a strong semantic layer—such as documenting metadata and sustaining taxonomies and ontologies that categorize data elements and catalog their relationships—remain underdeveloped: In our MIT CISR 2024 Data Monetization Survey,[4] just twenty-one percent of the executive respondents rated their organization's data curation practices as somewhat or very well developed. Yet organizations that had adopted these practices were more than three times as likely to be effective at implementing value-realizing data and AI initiatives, compared to organizations without such practices. They were also twice as likely to report that those data and AI initiatives gave them a meaningful competitive advantage.[5]

1 The MIT CISR Data Research Advisory Board (the "Data Board") is a community of data and analytics leaders from MIT CISR member organizations who participate in and inform MIT CISR research.

2 See an explanation distinguishing between GenAI tools and solutions in N. van der Meulen and B. H. Wixom, "Managing the Two Faces of Generative AI," MIT CISR Research Briefing, Vol. XXIV, No. 9, September 2024, https://cisr.mit.edu/publication/2024_0901_GenAI_VanderMeulenWixom.

3 The authors conducted interviews between Q1 and Q2 2025 with members of the MIT CISR Data Research Advisory Board from forty-one organizations; they hosted virtual Data Research Advisory Board discussions, two each in February and April 2026, on topics related to the semantic layer; and they performed a literature review of semantic layer articles.

4 2024 MIT CISR Data Monetization Survey (N=349); survey respondents were executives with an understanding of their organization's data investments and outcomes.

5 We measured data monetization effectiveness as the average of executives' ratings of their organization's effectiveness at improving (making work better, cheaper, or faster), wrapping (enhancing existing products with data-driven features), and selling (offering information solutions that customers pay for). Among respondents to our survey with somewhat or very well developed data curation practices—72 organizations—46% rated as effective or highly effective at data monetization. Among most other respondents—273 organizations—14% rated as such. The remaining 4 organizations did not respond to this question.

THE TARGET STATE FOR A SEMANTIC LAYER

Historically, to produce trusted applications and outputs, organizations relied on metadata stored in system-specific and enterprise data dictionaries and data models as well as knowledge in the heads of domain experts. However, as organizations have increasingly pursued AI initiatives and implemented enterprise data and analytics platforms, they have needed to develop a robust semantic layer. To capture and convey metadata in this layer, they have invested in semantic technologies and techniques such as taxonomies, ontologies, and business rules engines.

Today, GenAI sets an even higher bar for accessing data context in metadata at speed and scale; without context, GenAI tools cannot generate relevant and reliable responses for business users and are more prone to misinterpretations and hallucinations. Thus, GenAI tools and solutions require contemporary technologies and techniques, such as semantic search, which does not rely on keyword matches but instead uses metadata to understand the intent behind a query. They benefit from knowledge graphs,[6] which create a machine-readable representation of enterprise knowledge for use with applications such as recommendation engines, fraud detection, and customer chatbots.

In the age of GenAI, a semantic layer needs a bigger scope of contextual information. It must encode, for example, data usage constraints, permitted data combinations, rules for handling personal data, and data to support compliance audits.

A semantic layer holds the sum of the organization's understanding of data, regardless of where data resides or how it is technically structured at any point in time. In the age of GenAI, a semantic layer needs a bigger scope of contextual information. It must encode, for example, data usage constraints, permitted data combinations, rules for handling personal data, and data to support compliance audits. A semantic layer can also capture and expose codified expertise, such as judgments about the quality of data or permitted exceptions to business rules in technologies like data dictionaries and business rules engines. With such a comprehensive semantic layer, an organization can expand AI use without reinventing data governance for every use case. And AI agents as well as GenAI models using retrieval-augmented generation (RAG) will perform more accurately and with less risk.

Organizations like digital companies and information businesses have long needed sophisticated and trustworthy metadata to carry out their standard business operations. As a result, they have organically built increasingly advanced semantic layers over time. Consider the following case of information business Healthcare IQ.

HEALTHCARE IQ'S SEMANTIC LAYER

Healthcare IQ[7] is a privately held data and analytics company that has provided hospitals with data and analytics to inform purchasing, pricing, and reimbursement decisions for thirty-six years. By 2022, the company was curating two key data assets: a hospital supply chain data asset, containing detailed records of historical spending from many hospitals (Healthcare IQ's customers); and a product catalog data asset, containing a comprehensive catalog of medical products. The hospital supply chain data asset contained data from hospitals' invoices, purchase orders, pricing contracts, item and vendor master files, electronic health records, and payor reimbursements. The product catalog data asset included nearly six million unique healthcare products from more than 25,000 manufacturers, along with their attributes and relationships.

Healthcare IQ customers accessed the hospital supply chain and product catalog data assets using a suite of Healthcare IQ-developed reporting and analysis software tools. On average, Healthcare IQ expected a $100 million hospital to experience at least $8 million in price savings from such tool and data asset use.

Healthcare IQ stored its hospital supply chain and product catalog data assets in a cloud-based data lake with a microservices architecture using a set of Healthcare IQ semantic technologies and techniques. The source for the medical product data was mainly unstructured text on the internet that Healthcare IQ had scraped, then analyzed, categorized, and enhanced with new data attributes such as product type, sizing, sterility, and disposability. This process used a natural language processing (NLP) technique that required data dictionaries and collections of associated words. The engineer-

6 Knowledge graphs are flexible, graph-based data models that capture semantics explicitly in the form of concepts (e.g., customers, products, suppliers) and relationships (e.g., "purchased," "belongs to," "located in"). They are machine-readable and consumable by applications and AI. Knowledge graphs are often based on ontologies or semantically enriched by using ontologies.

7 B. H. Wixom, C. M. Beath, J. Duane, and I. A. Someh, "Healthcare IQ: Sensing and Responding to Change," MIT CISR Working Paper No. 458, February 2023, https://cisr.mit.edu/publication/MIT_CISRwp458_HealthcareIQDataAssets_WixomBeathDuaneSomeh.

ing team sourced words from existing medical dictionaries, and also built custom dictionaries of the vocabularies of their hospital customers, including their annotations, synonyms, and abbreviations. Healthcare IQ's team of content domain experts, which included former hospital procurement analysts and healthcare clinicians, helped to train the models.

Similar to the messy unstructured internet text used to build the product catalog data asset, the source data for the hospital supply chain data asset required curation: for example, Hospital A referred to a product as "Sterile Gauze Pad, 4x4," while Hospital B used the code "SG-44-STR." Because each hospital's information systems and data challenges were unique, Healthcare IQ used data pipelines to build a data curation engine the company called Semantic ETL (SETL).[8] This engine harmonized and mapped the source data. It also identified data quality problems[9] with hospital data during customer onboarding and ongoing data loading: Is this a catalog number? Is this a valid vendor? SETL performed entity recognition using ontologies and business rules; mapped products to standard product codes and descriptions; integrated records; and enhanced data records with additional information, such as Healthcare IQ descriptions, manufacturer names, benchmark pricing, and contract pricing.

Healthcare IQ used its semantic layer, including its data dictionaries, data models, taxonomies, ontologies, and access control databases, to source, clean, standardize, and enhance data assets cost efficiently and quickly. Its semantic technologies and techniques automated 80 percent of the effort required to onboard new hospitals. The semantic layer helped the company turn data from an ecosystem of manufacturer websites, public databases, and hospital systems into two meaningful, standardized, high-quality data assets. Healthcare IQ's semantic layer established shared terminology and definitions that made heterogeneous data understandable and combinable. Also, the metadata Healthcare IQ created helped the organization comply with HIPAA and other regulatory obligations, by powering deidentification techniques such as data blinding, data aggregation, and data access control management.

8 Healthcare IQ's SETL engine was based on semantic ETL, an industry-agnostic data extract, transform, and load framework that allows Healthcare IQ to link data in a manner defined by and inherently meaningful to the company's customers.

9 Issues with the data would prevent it from meeting Healthcare IQ's specifications, and the hospital would need to address any issues—such as by fixing data entry processes or improving supply chain and clinical processes—before it could resubmit the data.

NEXT STEPS TO BUILD A SEMANTIC LAYER

For any data-driven organization, investing in a robust semantic layer is a commitment to creating data assets that are understandable for use cases that need trustworthy data quickly and at scale. With GenAI, it is critical that a semantic layer offer a machine-readable representation of enterprise knowledge, because that will allow AI tools and solutions to interpret data, reason, and generate more trustworthy outputs. This is no easy feat: both the number of GenAI solutions and tools and the volume of unstructured content feeding them are ever-increasing.

Here are next steps to ensure that your investments in semantic technologies and techniques pay off:

- **Incrementally build your semantic layer, starting with priority data assets.** Identify what data you need to fuel high-priority GenAI tools and solutions. Invest in technologies and techniques that help identify and remediate contextual gaps in data for those initiatives. For example, a taxonomy could reconcile varying naming conventions so that common fields can be identified and aggregated. Or a knowledge graph could identify common entities across varying databases and tools so that records can be integrated. Expect your semantic layer to grow as you repeatedly solve problems associated with data cleansing, cataloging, benchmarking, and permissioning.
- **Govern the semantic layer itself.** Because the semantic layer contains critical contextual information about data assets, the layer needs an owner who is accountable for its quality and results. Ideally this owner works closely with data platform owners and weighs in on data deprecation and lifecycle management decisions. The semantic layer owner will help data platform owners secure investments needed to exploit new semantic technologies and techniques.
- **Look for opportunities to use AI to generate metadata that describes and makes explicit the context of data assets.** Data curation practices benefit from using AI to enhance tasks such as metadata generation, data cleansing, access control, data classification, and data connection. AI will only become more important for ongoing semantic layer management as the AI era unfolds.

As GenAI tools and solutions proliferate, competitive advantage will depend more on how well organizations can make their own data readily accessible and understandable by humans and machines. The organizations that do this well will scale AI faster, at lower cost, and with greater confidence that their data is being used in acceptable ways.

MIT CENTER FOR INFORMATION SYSTEMS RESEARCH (CISR)

MIT CISR helps executives meet the challenge of leading increasingly digital and data-driven organizations. We provide insights on how organizations effectively realize value from approaches such as digital business transformation, data monetization, business ecosystems, and the digital workplace. Founded in 1974 and grounded in MIT's tradition of combining academic knowledge and practical purpose, we work directly with digital leaders, executives, and boards to develop our insights. Our research is funded by member organizations that support our work and participate in our consortium.

CISR RESEARCH PATRONS

AlixPartners
Avanade
Cognizant
Collibra
IFS
PwC

CISR SPONSORS

ABN Group (Australia)
Alcon Vision
ANZ Banking Group (Australia)
AustralianSuper
Banco Bradesco S.A. (Brazil)
Barclays (UK)
BNP Paribas (France)
Bupa (Australia)
CalSTRS
Caterpillar, Inc.
Cemex (Mexico)
Cencora
CIBC (Canada)
Commonwealth Superannuation Corp. (Australia)
Cuscal Limited (Australia)
Dawn Foods
DBS Bank Ltd. (Singapore)
Doosan Corporation (Korea)
Ericsson (Sweden)
Fidelity Investments
Fomento Economico Mexicano, S.A.B., de C.V.
Genentech
HCF (Australia)
Hunter Water (Australia)
International Motors
JERA Co., Inc.
JPMorgan Chase
Kaiser Permanente
Keurig Dr Pepper
Mallesons (Australia)
Mater Private Hospital (Ireland)
Nasdaq, Inc.
National Australia Bank, Ltd.
Nomura Holdings, Inc. (Japan)
Nomura Research Institute, Ltd. Systems Consulting Division (Japan)
Novo Nordisk A/S (Denmark)
OCP Group
Pentagon Federal Credit Union
Principal Life Insurance Company
Ralliant
Reserve Bank of Austrailia
RTX
Saint-Gobain
Scentre Group Limited (Australia)
Schneider Electric Industries SAS (France)
TabCorp Holdings (Australia)
Telstra Limited (Australia)
Terumo Corporation (Japan)
UniSuper Management Pty Ltd (Australia)
Uniting (Australia)
Vanguard
WestRock Company
Xenco Medical
Zoetis Services LLC

Information as of May 2026

MIT CISR is funded by our members, and we gratefully acknowledge their financial support and their many contributions to our work.

Membership and benefits: cisr.mit.edu/content/member-benefits

MIT CISR research publications: cisr.mit.edu/research-library

MIT Sloan School of Management
Center for Information Systems Research

One Main Street, E90-9th Floor
Cambridge, MA 02142
t 617-253-2348 | e cisr@mit.edu

cisr.mit.edu |

Team | Isobela Byerly-Chapman, Margherita Di Pinto, Christine G. Foglia *Associate Director*, Dorothea Gray-Papastathis, Cheryl A. Miller, Ina M. Sebastian, Alan Thorogood, Nick van der Meulen, Peter Weill *Chairman*, Barbara H. Wixom, Stephanie L. Woerner *Director*