The Data Catalog Paradox: Why the Front Office Can't Find the Data It Already Has

In this article:

- **Why existing data catalogs fail front-office users**
- **The hidden cost of data discovery bottleneck**
- **What capital markets-native, business-first data discovery looks like in practice**

Paul Villena
Chief Technology Officer, RoZetta Technology

ROZETTA TECHNOLOGY

In a 2025 Bloomberg survey of over 150 quants, research analysts, and data scientists across capital markets, one bottleneck stood out: 72% could evaluate only three or fewer datasets at a time, and for 65%, evaluating even a single dataset takes a month or longer. In a market where alpha increasingly depends on who can ingest and act on new data fastest, most teams are stuck in a months-long queue before the real analysis even starts.

The mechanics of that queue are painfully familiar. When a quant needs to evaluate a new alternative dataset for a factor model, the journey typically starts with a message to the data engineering team, followed by days of back-and-forth to understand what the dataset contains, how it was sourced, whether the quality is reliable, and whether the firm is licensed to use it for that purpose. **By the time the data reaches the analyst's notebook, the trading signal may already be stale.** Worse, the institutional knowledge generated through that entire evaluation — what worked, what didn't, which fields were unreliable, which vendor was unresponsive — all of it lives in email threads and the heads of the people involved, never making it back into any system the next analyst can search.

The irony is that most of these firms have invested heavily in data infrastructure. They have technical data catalogs. They have cloud platforms. They have data lakes measured in petabytes. The problem is not the absence of data. **It is the absence of data discovery built for the people who actually need to use it.**

The Catalog Gap: Built for Pipelines, Not Portfolios

Technical data catalogs were designed primarily for data engineering teams. They excel at schema documentation, pipeline lineage tracking, and technical metadata. These are valuable capabilities. But they solve a different problem from the one facing the front office.

A portfolio manager evaluating whether to add satellite imagery data to a macro strategy does not need column-level metadata. They need to know:

- What does this dataset actually measure?
- How frequently is it updated?
- What is the data quality like in practice?
- Has anyone else at the firm used it, and for what?
- What are the licensing restrictions?
- Can I use it in a systematic strategy, or only for discretionary research?

Capital markets data carries context that generic business data catalogs were never designed to capture. A single dataset may have different licensing terms for proprietary trading versus client research versus index construction, embargo periods that vary by desk, and quality dimensions — **survivorship bias, point-in-time correctness, coverage universe** — that determine whether the data is usable before anyone looks at a single row.

The commercial impact is quantifiable, and it falls in two places. Data leaders running these environments know the numbers. Global financial data spend reached **$49.2 billion** in 2025 (Burton-Taylor). Yet meaningful, active utilisation of licensed datasets rarely exceeds **30%**. And for the front office, it compounds: longer evaluation cycles mean slower research, and slower research means the analysis arrives after the opportunity has moved. **Both pressures land on the same bottom line.**

The consequences compound across the organization:

- **Duplicated Effort.** Multiple teams independently evaluate the same datasets or build redundant pipelines.
- **Shadow Data Infrastructure.** Analysts bypass governance by creating their own copies, spreadsheets, and workarounds.
- **Delayed Research Velocity.** Every dataset needing an engineering ticket adds days or weeks to the research cycle.

The problem intensifies as firms adopt AI.Through 2026, organizations will abandon 60% of AI projects unsupported by AI-ready data. Data is the bottleneck, not algorithms or compute. When AI models are trained or prompted with data that analysts cannot independently verify for quality, provenance, and licensing compliance, the risk multiplies.

Source:
Bloomberg Survey, 2025 — dataset evaluation bottleneck
Burton-Taylor Global Data Management & Analytics Market Study, 2025
Gartner, 2024

What Business-First Data Discovery Looks Like

The alternative is a data catalog structured around how business users within the industry think. A portfolio manager can go from "I need credit spread data for European corporates" to evaluating a shortlist of licensed, quality-scored datasets in minutes, without filing a ticket or messaging the data engineering team.

What separates a capital markets business catalog from a generic one is the institutional knowledge it captures:

- **Plain-Language Descriptions:** Generated from vendor documentation using LLM-powered semantic enrichment, not cryptic column names
- **Quality Scores:** Displayed alongside every dataset so analysts can assess reliability before committing to evaluation
- **Usage Signals:** Which teams are actively using a dataset and for what purpose, providing built-in peer validation
- **Engineer's Notes:** The quirks, caveats, and known gaps that pipeline teams accumulate, surfaced at the point of need
- **Governed Access:** Licensing terms visible and enforced, so users can discover, evaluate, and use data without creating compliance exposure

From Technical Inventory to Research Enablement

The shift required is not incremental. It is a category shift: from technical data catalogs built for engineers to specialized business data catalogs built for the people whose decisions generate revenue. The firms that make this shift **convert data spend into research velocity.** Those that do not continue paying for data that sits unused.

How DataHex Data Library Delivers This

DataHex Data Library is a purpose-built business data catalog for capital markets. It organizes data around asset class, geography, provider, and use case, with multi-faceted discovery paths that let front-office users find what they need without knowing where it lives in the underlying infrastructure.

AI-powered semantic search and context enrichment, grounded in vendor specification documents, translates cryptic metadata into plain-language descriptions.

Engineer's notes, quality scores, usage analytics, and peer signals are surfaced at the point of evaluation. Interactive previews let analysts inspect sample data before committing to extraction, and one-click launch-to-tool integrations send data directly into an analytics platform or back-testing engine.

DataHex operates as a lightweight metadata layer over existing infrastructure, requiring no data migration. Deployment is measured in weeks, not years. **The front office moves faster. The data budget works harder.** Both, without replacing the infrastructure already in place.

See how front-office teams discover and evaluate datasets in minutes, not days

→ Request a DataHex Data Library walkthrough

RoZetta Technology believes in empowering every individual and every AI agent in the organization with institutional knowledge, so they can govern, discover, and maximize the value of their data.

We are a capital markets data infrastructure and intelligence company. Our systems power global firms to unlock the value of the data they already have. Our DataHex platform delivers purpose-built solutions for data vendors (DataHex Data Shop), data users (DataHex Data Library), and AI-powered intelligence.

inquiry@rozettatechnology.com
rozettatechnology.com