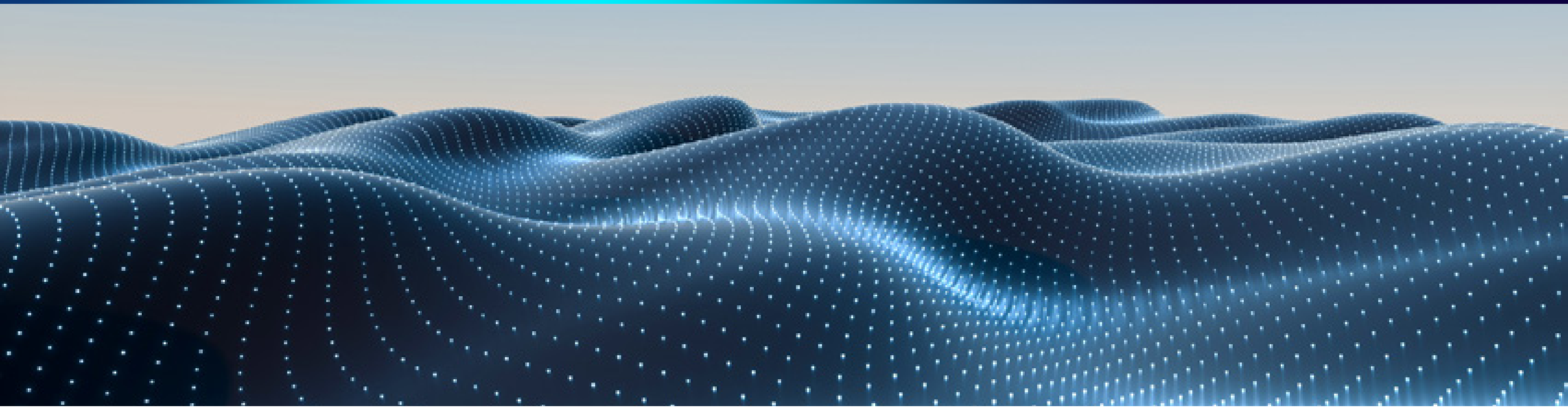
Technology Executive Update

CUTTER
AN ARTHUR D. LITTLE COMMUNITY



Vol. 2, No. 1, 2023

KNOWLEDGE GRAPH IMPLEMENTATION: COSTS & OBSTACLES

by Michael Atkin

Breaking through psychological barriers to entry is key to succeeding with any data management initiative. This is doubly true when seeking to adopt semantic standards to implement a knowledge graph within your organization — because change is risky. Application owners don't want to give up control. Most key stakeholders don't really understand the principles of data; they just want a near-term solution to an isolated use case. And the data dilemma is often viewed as too low-level for C-level executives to get their arms around. In this *Executive Update*, we explore how to fundamentally fix data so that it becomes a resource organizations can truly leverage.

THE DATA DILEMMA

Psychological barriers were among the top-line findings in my inquiry into the costs and obstacles associated with knowledge graph implementation. I began this inquiry because I was perplexed. The data dilemma (i.e., content incongruence and structural rigidity due to technology fragmentation) has been revealed as a significant liability to organizations. There is no question that it diverts resources from business goals, extends time to value, leads to business frustration, and inhibits an organization's ability to automate operational processes.

It is equally clear that we are not going to solve this dilemma by continuing to use yesterday's processing models to independently manage data in these fragmented silos. We've been on that path for well over a decade and have barely succeeded in achieving basic hygiene and putting core data governance in place. And while both hygiene and governance are important — critical in fact — they are not enough to transform data from a "problem to manage" into data as a "resource to exploit."

What is required is fundamentally fixing the data itself. We must unshackle it from the tables and joins that have become our conventional legacy. We must lock down granular meaning and embed it directly into the content itself. We must free our analysts from the business of being data janitors and change their "transform and revise" mindset that defines how most developers learned to operate.

The shift from the limitations of technology that was state-of-the-art two generations ago is absolutely achievable, which makes the current circumstances even more puzzling. The value proposition based on semantic standards is overwhelming. The pathway to implementation is incremental, self-describing, reusable, and testable. And the importance of implementing a modern data infrastructure fit for the digital age is clearly necessary to address the complexity of today's business environment. So what is the problem?

DATA MANAGEMENT IS BOTH ESSENTIAL & MEANINGFUL

I began my research with the objective of defining the cost side of the equation — simply to make a reasonable business case to executive stakeholders — on the logic of adopting the knowledge graph. My focus is on companies where quality, traceability, and flexibility of data are essential ingredients — because not every company is an initial candidate for the adoption of semantic standards. After interviewing experts and practitioners across the spectrum, I have organized my findings into three parts:

1. The role of organizational issues, including positioning and dealing with bureaucratic roadblocks
2. The costs of operational discovery and technology to deliver the initial use cases
3. The importance of practitioner capability for the people needed to manage the data pipeline and engineer the content

ORGANIZATIONAL ALIGNMENT

I can't overestimate the importance of top-of-the-house buy-in to elevate the challenges of data management as a critical issue to address. Data management is both essential and meaningful. I have often witnessed how clear and visible articulation by senior executives regarding areas of importance really drives organizational priorities. One of the core problems, however, is that few at the top fully understand or are directly responsible for fixing the data dilemma.

I can't help but wonder why data remains the poor stepchild to people, process, and technology in the minds of executive management. It is an essential factor of input in every aspect of our operations, but is often only understood as something we process. Perhaps executive manager view this area as too primal and technical. As a result, they aren't paying attention to the paradigm shift that is underway. It doesn't help that we do a poor job of positioning this issue in either business or executive terms.

One would think that the big, fat failures of existing solutions (i.e., warehouses, data lakes, data marts, single masters) to fundamentally fix the data dilemma and reduce the price of technical debt would be enough to change the equation. Unfortunately, many stakeholders just seem to accept the separation of systems from databases as a fact of life and something that will always exist. In reality, we haven't been overwhelmingly successful at getting organizations to understand and embrace the concept of linked data, where independent data sources have commonality that can be both shared and linked together.

I shudder to think that it will only be the "fear of missing out" that ultimately will facilitate broad adoption. We are in desperate need of clear and visible demonstrations of value from an industry leader. Once that is established, the rest of the industry will be more likely to follow. We know this is not about the capability of the knowledge graph. The technology works as advertised. The problem is we are still stymied by little clear evidence of knowledge graphs working at scale to combat the organizational forces at work. It is clear to me that the goal of broad adoption will not be advanced by a bunch of isolated use cases, which characterizes the current state of maturity across much of the industry.

That's why it takes a visionary to own the pathway. And, of course, data visionaries are both rare and short-lived. Implementing a knowledge graph is a collaborative process that requires cooperation at scale across both operational and functional boundaries. And it is hard to get people to cooperate with the culture of competition that seems to exist in many companies.

Throughout my interviews, it appears that the most important people to advance this cooperation might be those who own the "data mesh." Some companies have undergone changes in senior technology and business views, from top-down organizational structures aligned by function to organization by product teams with responsibility for the full vertical supply chain. This leads to an understanding of data as a "product" that must fit into the supply chain ownership approach.

There is good news. The orientation of data governance has shifted from focusing primarily on the provider's perspective (with emphasis on systems of record, authorized domains, lineage traceability, syntax, and operating models) to examining the consumer point of view (with emphasis on integration, meaning, use cases, and harmonization). This confluence of circumstances is pushing entities to adopt some degree of data sharing capabilities — progress that is all too often derailed by the myopic focus on short-term deliverables.

And that is the downside of the equation. Many data advocates are finding it difficult to collaborate with the data mesh owners on semantics and data architecture. The creation of domain-related marketplaces to create local data products is (as usual) a technology approach to the problem. Just making data a product without fixing the underlying models is insufficient to reach the goal of ensuring that data is in a flexible format for intuitive use and has a defined meaning for trust.

> **MANY DATA ADVOCATES ARE FINDING IT DIFFICULT TO COLLABORATE WITH THE DATA MESH OWNERS**

THE PSYCHOLOGY OF DATA MANAGEMENT

The biggest challenge to adopting semantic standards and a knowledge graph is not always about convincing executive management. People in positions of leadership can understand the data dilemma story — and it can be quite convincing — particularly when there has been visible failure using conventional technology. We are still using 50-year-old relational technology to look at links, relationships, and perspectives, which will no longer work in today's complex and interrelated world. The problems are much more with middle management, which creates two challenges.

The first challenge is vested self-interest. Many systems and application owners do not want to give up control, and most think in terms of current objectives rather than organizational requirements. For many systems owners, the concept of sharing data, resources, and approaches is anathema to the way they operate. They have their own processes and their own data models and do not want the knowledge graph as their system of record. The architects who are in control of the existing (relational) environment have already made an investment in SQL and make themselves obstacles to adoption. They just want their standard reports, and they already know their relational databases. If they want something new, they will have to accept the burden of export, transform, load (ETL), transformation and integration.

The unfortunate reality is that people who thrive in corporate environments don't make waves. Some of it is self-preservation and some of it is fear of the loss of autonomy in making model change decisions. As it turns out, most developers don't understand the idea of doing development for an activity that is not an "application." The delivery of apps defines the value of computer science. This is the core challenge with master data management (i.e., the quest for the *single version of truth*) where everyone needs to see all things the same way. This is where old-school systems thinking clashes with the notion of shared concepts across distributed data sets.

The second challenge stems from the multiple levels of bureaucracy that exist in many organizations. This is not restricted to knowledge graphs; friction exists around bringing many new approaches and technologies into the organization. But it is a real obstacle. It is admittedly hard to get some people to change their orientation. Knowledge graphs and the adoption of semantic standards are not "organizational policy," and convincing the infrastructure group to run the procurement gauntlet for data experiments is hard.

People who run these data centers frequently look for reasons to say "no." Most entities are looking to reduce cost and complexity, not add another component into the mix. This makes it difficult to bring new approaches into an organization. Getting new platforms, databases, and tools up and running creates a lot of waiting. There is a plethora of permissions to secure before buying and installing new technologies. Investing in semantic standards takes effort and is viewed as risky. Think of this as the technology illustration of the "tragedy of the commons."

Perhaps this is why so many knowledge graph initiatives are relegated to "skunk works" and assigned to narrow use cases that carry the risk of being viewed as trite. This is particularly problematic because the real value of the knowledge graph happens when it is integrated across use cases — connecting things that weren't previously connected. Overcoming architectural inertia clearly is still the primary obstacle to progress. Direct quotes from my recent benchmarking research disclose specific examples of the predicament (see Table 1).

INTERVIEW REPONSES	
"Technocrats serving as roadblocks who require proof of success before implementation."	"Leadership highly risk-averse and wedded to legacy methods."
"Lack of understanding that adopting this technology does not require retooling."	"Entrenched data processing ecosystems culture."
"Business units at varying degrees of sophistication with regard to data literacy."	"Lack of technical expertise to move beyond proof-of-concept."
"Delivery managers who don't understand information architecture!"	"Reluctance to change, low sills, low accountability, zero jeopardy."
"Lack of a reference architecture. We are making it up."	"Organization is too large, too complex, too siloed. Weight of politics and posturing."
"The technology stack is not understood by IT in general."	"Management is lost and very cautious about any decision."
"The organization unable to grasp the semantic EKG way of thinking."	"Technologies and approaches that address such goals have been ignored for years."
"Lack of willingness among clients to invest in ontologies, taxonomies, and linked data."	"Inertia on current technology. Too many immediate crises."

Table 1. Inhibitors to adoption based on participant interviews

OPERATIONAL CONSIDERATIONS

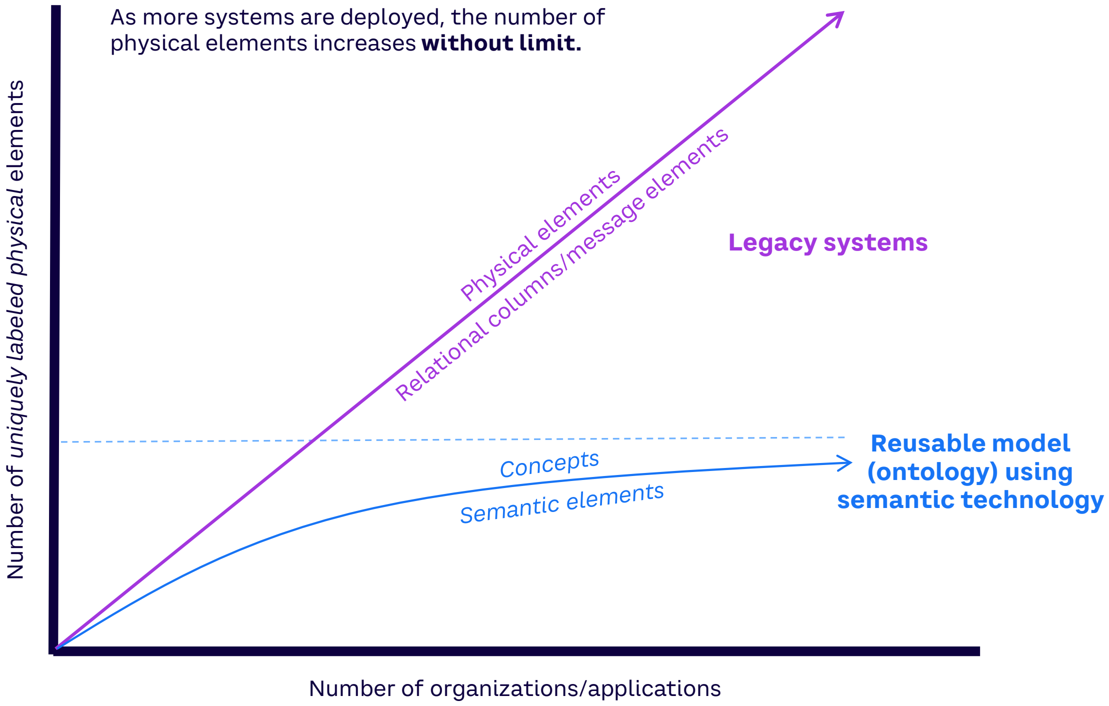
The cost of technical infrastructure for a knowledge graph is minimal and not viewed as an obstacle to adoption — particularly when considering the overwhelming cost of managing the cottage industry of silos and proprietary approaches many established organizations use. The direct cost, particularly for a proof of concept (POC), can be implemented within a sandbox environment, using trial software with a basic ontology constructed only from the data needed for the POC. In fact, as long as the interfaces exist for the data, there is minimal mandatory infrastructure.

> **EASY WINS ARE POSSIBLE ONCE BASIC COMPONENTS ARE CONSTRUCTED**

A single use case (POC) poses a challenge, as it might not be impressive enough to convince all involved stakeholders. There seems to be a big divide caused by the fact that conventional technology can almost always solve the immediate problem. The pathway to addressing the data dilemma's root cause starts with a "lighthouse" project as the first activity, which is designed to prove the point of the knowledge graph (i.e., reusable, testable, flexible, traceable, and contextual). The "digital twin," a virtual model of all systems, processes, databases, and applications, shows the most promise as the integration layer needed to get a holistic picture to better inform decisions about scalability, resilience, and lifecycle.

According to most of the experts, easy wins are possible once basic components are constructed. Invest in the user interface. Build some "glamour" applications for business visibility. Don't talk about graph or ontology because no one really wants technology for the sake of technology. The clear message is to stop focusing on the solution before you understand the problem to solve. The knowledge graph is an elegant solution to the data dilemma and can be tied to many use cases.

The bottom line is that a clear ROI exists for adopting semantic standards and knowledge graphs. This is extremely difficult if your initial focus is only ROI. If the organization understands the principles of data and understands the nature of the problem, it becomes obvious that it will not be solved by conventional approaches. The pathway forward begins with a "self-fulfilling" journey. Start with the foundational components. Select the first project with a definable and valuable payoff (i.e., not only for a silo). Identify the related use cases because the onward applications can be accomplished for diminishing marginal costs. This is the opposite of conventional approaches, where every new system costs more because of the multiplicity of integration points (see Figure 1).



As more systems are deployed, the number of physical elements increases **without limit.**

But the number of business concepts and data elements **remains limited.**

Figure 1. The relationship between physical elements and business concepts (adapted from the US Department of Defense Business Mission Office)

LOWER BARRIERS TO ENTRY

The most pragmatic approach offered for collaboration at scale was to follow the linked data example of the Semantic Web. The building blocks of the Semantic Web are mature standards for ensuring unique identity (i.e., Internalized Resource Identifier [IRI]) and facilitating shared meaning (i.e., Resource Description Framework [RDF]), which guarantees that all data is structured and machine-readable. These linked data standards are a method of publishing to enable computers to share data and infer relationships. This enables data from different sources to be connected and queried. It works for the Internet, and it can work for any organization.

One of the big challenges many companies experience is the ability of technical stakeholders to understand the "point of divergence" between conventional and semantic approaches. Many organizations don't have an ontologist who understands RDF or analysts who know how to write queries in SPARQL. There is abundant confusion between people currently practicing taxonomy (linear classification) and people needed to create ontologies (models of concepts and relationships). The fundamental tenets of good ontology design are noted as a significant gap in the development chain. These are both mindset and skill set gaps that are hard to overcome. The consensus advice suggests not making this the prerequisite for initial success.

THE PATHWAY FORWARD STARTS WITH POLICY

The pathway forward starts with policy directing all independent lines of business to use Web standards for identity and expression for all applications for each local data set. No more construction in isolation. In essence, push the "data product" within the data mesh as a library of applications using standards. Most existing Web pages use linked data in the form of JSON-LD. This is a known process for most developers. At the local level, they know the data, the use cases, and the requirements.

Cultivating the owners of the data (i.e., subject matter experts) is the most important step for dealing with the data dilemma. Successful companies have found that allocating the task of ontology development and mapping to triples to a central data team of expert practitioners is a much more productive way to proceed.

BUILDING A CAPABILITY CENTER (EXTENSIBLE PLATFORM)

Once a firm has demonstrated the value proposition and progressed from a successful POC to an operational pilot, the pathway to progress centers mostly on investing in personnel. The team of experts who form the capability center, most likely between five and 15 people, will account for most of the cost of implementing an enterprise-level knowledge graph.

Expanding the identity of data owners who know the data's location and health is the first hurdle. Most of this is simply about organizational dynamics and understanding who the players are, who is trusted, who is feared, who elicits cooperation, and who is out to kill progress. This is a modeling exercise to identify principal and related use cases and coincides with developing the action plan, which includes capturing the inventory of the existing landscape.

Part of that exercise will focus on core operational information including:

- The scope of systems, processes, and components
- An understanding of how the above are connected
- The software dependencies
- The risks to consider
- A governance mechanism for developing policy and ensuring staff accountability

The practitioners I have been talking to suggest that an organization will need at least one experienced architect who fully understands the workings of the knowledge graph. This person will design the approach, build the use case tree, unravel dependencies, and lead the team. The organization will also need ontologists to design the content engineering framework, build the domain-specific ontologies, and manage the mapping of data.

A couple knowledge graph engineers will be needed to coordinate with the database administrators (DBAs) on the data's meaning and the content models. This is about extraction, transformation, validation, curation, testing, and other tasks associated with managing the data pipeline. ETL in a semantic environment is somewhat different from conventional ETL because the data sets in each pipeline are generic and designed to be reused for onward use cases. The organization also needs a project manager to advocate for the team and the development process.

Let's put it all into perspective. The first POC is not expensive (barely a rounding error for most organizations). Converting the POC to an operational pilot adds some additional infrastructure cost and also requires a team to manage the pipeline. The migration to an extensible platform shifts the effort from building the technical components to adding incremental use cases. The combined budget for these is somewhere between US $1 million and $3 million. This is where the reusability benefit kicks in; plan for 30% of the original cost, but three times faster. Self-sufficiency starts to arrive after the first few domains, during year three, and continues to decrease as reusability advances. The long-term cost of a true enterprise knowledge graph is somewhere around $10-$20 million.

THE LONG-TERM COST OF A TRUE ENTERPRISE KNOWLEDGE GRAPH IS SOMEWHERE AROUND US $10-$20 MILLION

CONCLUSION

The essence of the data dilemma is clear. Our fragmented technology environments have allowed data to become isolated into hundreds of independent silos. We have modified, transformed, and renamed the content many times to make the software that propels our business processes. As a result, data has become incongruent. The meaning from one repository is not always the same as the meaning from another, particularly as we try to connect an organization's processes across independent lines of business.

Not only has data become misaligned, people also suffer from the limitations of relational (and proprietary) technology, where data is organized into columns and stored in tables linked together using internal keys. Some firms support several thousand tables — many with conflicting column names — and all have relationships that must be explicitly structured. As a consequence, firms spend countless amounts of time and money moving data from one place to another. They invest significant effort in reconciling meaning. And fear of disrupting critical processes often makes it difficult to implement changes.

But it doesn't have to be this way. Data incongruence and structural rigidity are problems with solutions. And the methodology is clear and definable: adopt the principles of data hygiene and implement Web standards for identity and meaning. Don't overwhelm your stakeholders with semantic complexity. Develop your organization's own reference website of concepts used to categorize and define information about your business. Focus on user experience to answer business questions that can't otherwise be answered because of data limitations. Make it operational. Let the analysts use it and ask for more. Expose the work and let it speak for itself. It is time to get on with building the data infrastructure for the digital world.

About the Author

Michael Atkin is an expert in content management, reference data strategy, governance, and content engineering. He has been providing strategic advice to financial industry participants on the requirements associated with managing data as a business asset since 1985. Mr. Atkin is Managing Director of Content Strategies LLC. He has served as strategic advisor to the EDM Council and is a former member of the SEC Market Data Advisory Committee, the CFTC Technical Advisory Committee, various ISO Working Groups, and the Financial Stability Board's Advisory Group for LEI. Mr. Atkin was a two-term chair of the Data and Technology Subcommittee for the US Treasury's Financial Research Advisory Committee. Recently, he has been on the faculty of Columbia University, teaching master's degree candidates about the principles, practices, and operational realities of data management. He can be reached at atkin@content-strategies.com.

CUTTER

AN ARTHUR D. LITTLE
COMMUNITY

Cutter Consortium, an Arthur D. Little community, is dedicated to helping organizations leverage emerging technologies and the latest business management thinking to achieve competitive advantage and mission success.

Cutter helps clients address the spectrum of challenges disruption brings, from implementing new business models to creating a culture of innovation, and helps organizations adopt cutting-edge leadership practices, respond to the social and commercial requirements for sustainability, and create the sought-after workplaces that a new order demands.

Since 1986, Cutter has pushed the thinking in the field it addresses by fostering debate and collaboration among its global community of thought leaders. Coupled with its famously objective “no ties to vendors” policy, Cutter’s *Access to the Experts* approach delivers cutting-edge, objective information and innovative solutions to its community worldwide.

For more information, visit www.cutter.com
or call us at +1 781 648 8700.