ontotext

Information Literacy Series for Executives

KNOWLEDGE GRAPH IMPLEMENTATION: COSTS AND OBSTACLES TO CONSIDER

Michael Atkin, Managing Director, Content Strategies LLC

Breaking through the psychological barriers to entry is the key to making any data management initiative a success. This is doubly true when seeking to adopt semantic standards to implement a knowledge graph within your organization. Change can be scary. 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 a use case. And C-level executives often don't own the 'data dilemma, although they drive the business need.

These were among the top line findings in the inquiry into the costs and obstacles associated with knowledge graph implementation. Driving this research is that the data dilemma (i.e., content incongruence and structural rigidity due to technology fragmentation) has been demonstrated 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 independently manage data in these fragmented silos using yesterday's processing models. We've been on that path for well over a decade and still struggle with basic hygiene and putting core data governance in place. And while both are important (critical in fact) – they are not sufficient to turn data from a 'problem to manage' into data as a 'resource to exploit.'

What is required is to fundamentally fix 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 out of the mindset of 'transform and revise' that define how most developers were taught to operate.

What is even more puzzling is that the shift from the limitations of technology that was state-of-the-art two generations ago is absolutely achievable. 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 that is fit for the digital age is clearly necessary to address the complexity of today's business environment. So, what is the problem?

This research began 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 a knowledge graph. The focus is on companies where quality, traceability and flexibility of data are essential ingredients – because not every company is an initial candidate.

After interviewing a variety of experts and practitioners, findings are organized into three parts:

- organizational issues including positioning and dealing with bureaucratic roadblocks
- the costs of operational discovery and technology to deliver the initial use cases
- the importance of practitioner capability for the people needed to manage the data pipeline and engineer the content

ORGANIZATIONAL ALIGNMENT

We 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. It is both essential and meaningful. It's been seen many times how the clear (and visible) articulation by senior executives of 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.

With all that said, we 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 input into every aspect of operations, but often is only understood as something we process. Perhaps they view this area as too primal and technical. 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 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 seem to accept the separation of systems from databases as a fact of life and something that will always exist. The reality is that we haven't been overwhelmingly successful at getting organizations to understand and embrace the concept of value available from Linked Data where independent data sources have commonality that can be both shared and linked together to drive gains.

It may be the 'fear of missing out' that will finally facilitate broad adoption. We are in desperate need of clear and visible demonstrations of value by industry leaders. Once that is established, the rest of the industry will be more likely to follow. This is not about knowledge graph capabilities. The technology works as advertised. The problem is we are still stymied by enough clear evidence of knowledge graph working at scale to combat the organizational forces at work. It is clear that the goal of broad adoption will not be advanced with 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. Data visionaries are both rare and short-lived. The truth is that implementation of 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.

The most important people to advance this cooperation might be those that own the 'data mesh.' For some companies, there has been a change 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 is leading to an understanding of data as a 'product' that must fit into the supply chain ownership approach.

The good news is that this has shifted the orientation of data governance from being focused primarily on the perspective of the provider (with emphasis on systems of record, authorized domains, lineage traceability, syntax and operating models) to 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 owners of the data mesh 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 facilitate the prime goal of ensuring that data has defined meaning for trust and in a flexible format for intuitive use.

PSYCHOLOGY OF DATA MANAGEMENT

The biggest challenge to the adoption of semantic standards and knowledge graphs is not always convincing executive management. People in positions of leadership can understand the story – and it can be quite convincing – particularly when there has been visible failure using conventional technology. The problems are often more with middle management.

The first of these challenges is vested self-interest. Many systems and applications owners do not want to give up control and most think in terms of current objectives rather than organizational requirements. To many systems owners, the concept of sharing data, resources and approaches is an 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 systems of record. The architects that are in control of the existing (relational) environment have already made an investment in SQL and are 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 ETL, transformation and integration.

Some of this is about self-preservation and fear about the loss of autonomy in making model change decisions. Doing development for something that is not an 'application' is not something that most developers understand. The delivery of apps is the definition of 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 of these challenges are the multiple levels of bureaucracy that exist in many organizations. This is not restricted to knowledge graph – there is friction in getting 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 it is hard to convince the infrastructure group to run the procurement gauntlet for data experiments.

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 in new approaches into an organization. Investing in semantic standards does take effort and is viewed as risky.

Perhaps this is why so many knowledge graph initiatives are relegated to the 'skunk works' and focus on narrow use cases that carry the risk of being viewed as trite. This is particularly problematic because the real value of the knowledge graph is when it is integrated across use cases – to connect things that weren't previously connected. It is clear that overcoming architectural inertia is a significant obstacle to progress.

Below are direct quotes on inhibitors to adoption from the benchmarking research:

"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 eco-systems. Culture."
"Business units at varying degrees of sophistication with regards data literacy."	"Lack of technical expertise to move beyond a proof of concept."
"Delivery Managers who dont't understand Information Architecture!"	"Reluctance to change, low sills, low accountability, zero jeopardy."
"Lack or 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 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."

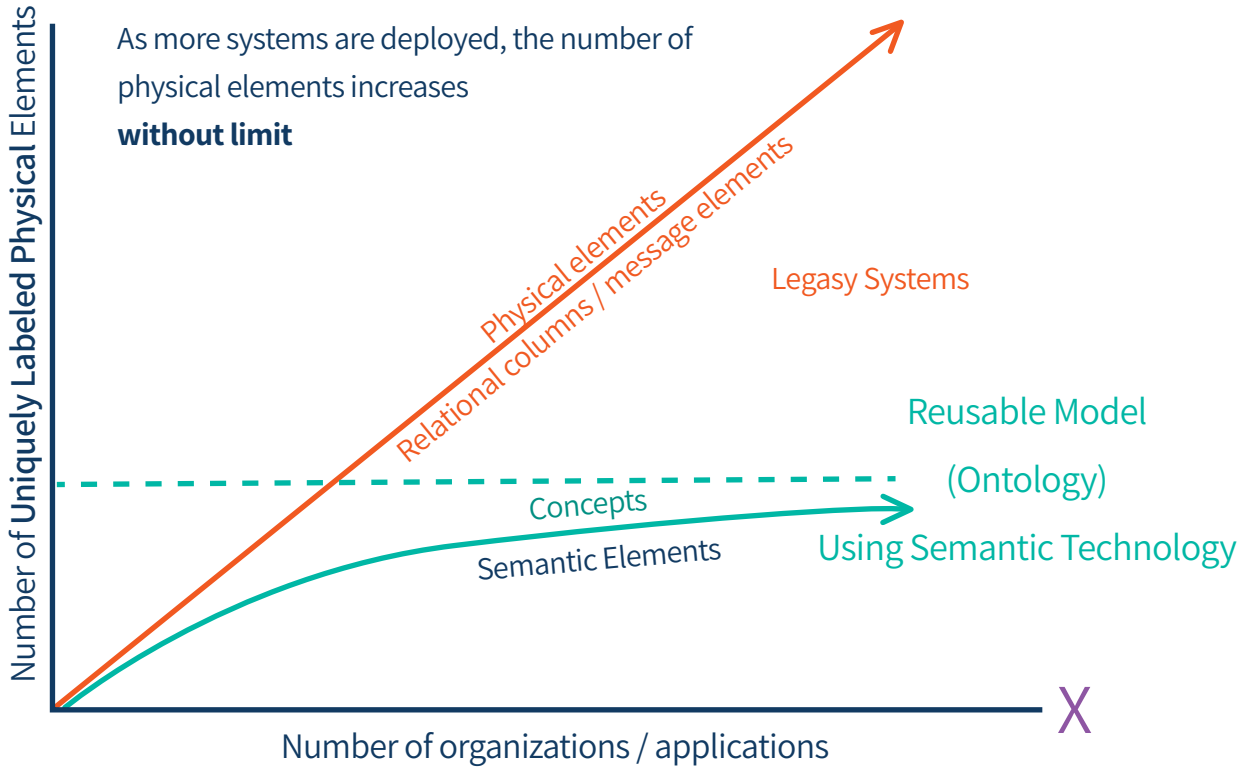
OPERATIONAL CONSIDERATIONS

The cost of the technical infrastructure for a knowledge graph is minimal and should not be viewed as an obstacle to adoption – particularly considering the overwhelming cost of managing the cottage industry of silos and proprietary approaches that characterize many established organizations. 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 not much in the way of mandatory infrastructure.

The challenge is that a single use case (POC) might not be impressive enough to be convincing to all the involved stakeholders. There seems to be a big divide between the fact that one can almost always solve the immediate problem using conventional technology. The pathway to addressing the root cause of the data dilemma would be to position the first activity as a 'lighthouse' project that is designed to prove the point of the knowledge graph (i.e., reusable, testable, flexible, traceable and contextual). The most promising of which is the 'digital twin' – a virtual model of multiple systems, processes, databases and applications – as the integration layer necessary to get a holistic picture of things for better-informed decisions about scalability, resilience and lifecycle.

According to most of the experts, easy wins are possible once the basic components are constructed. Invest in the user interface. Build some 'glamor' applications for business visibility. Don't talk about graphs 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. Knowledge graph is an elegant solution to the data dilemma and can be tied to many use cases.

The bottom line is that there is a clear ROI for adopting semantic standards and knowledge graphs. This is extremely difficult if the initial focus is only the 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 is to create 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 chart below).



But the number of business concepts and data elements remains limited

Adopted from the US Department of Defense Business Mission Office

CAPABILITY CENTER (Extensible Platform)

Once the company has demonstrated the value proposition and progressed from a successful POC to an operational pilot, the pathway to progress mostly centers on the investment in people. The team of experts (likely between 5 and 15 people) that form the Capability Center will account for most of the cost of implementing an enterprise-level knowledge graph.

The first hurdle is to expand the identity of the data owners who know the location and health of the data. Much 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 activities. This is a modeling exercise for the identification of principal and related use cases. This coincides with the development of the action plan including capturing the inventory of the existing landscape. As part of that exercise will be all core operational information including the scope of systems, processes and components … an understanding of how they are connected … the software dependencies … the risks to consider … and a governance mechanism for developing policy and ensuring staff accountability.

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

COSTS -
THINK OF AS VALUE OPPORTUNITIES

Let's put it all into perspective. The first POC is not expensive, somewhere between $50,000 and $100,000 depending on complexity. Converting the POC to an operational pilot adds some additional infrastructure cost as well as a team to manage the pipeline. Migration to an extensible platform shifts effort from building technical components to adding incremental use cases. The budget for these together is somewhere between $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 (year three) and continues to decrease as reusability advances. The long-term cost of a true enterprise knowledge graph is somewhere around $10 million - $20 million.

CONCLUSION

The essence of the data dilemma is clear. Due to our fragmented technology environments, we've 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 work. As a result, data has become incongruent. Meaning from one repository is not always the same as meaning from another – particularly as we try to connect business processes across independent lines of business.

Not only has data become misaligned, we suffer from the limitations of proprietary technology where data is organized into columns and stored into tables linked together using internal keys. Some companies are supporting many thousands of tables – many with conflicting column names - and all with relationships that must be explicitly structured. Because of this, companies spend countless time and money moving data from one place to another. They invest significant effort reconciling meaning. And changes are often difficult to implement because of the fear of disrupting critical processes. But it doesn't have to be this way. Data incongruence and structural rigidity are problems that can be solved.

The methodology for digital transformation using knowledge graphs is clear and definable. Adopt principles of data hygiene and implement the Semantic Web standards for identity and meaning. Don't overwhelm your stakeholders with semantic complexity. Develop your organization's own reference website of the concepts used to categorize and define information about your business. Focus on the user experience to answer business questions that can't be answered because of the limitations of the data. 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 move forward with building the data infrastructure for the digital world with business goals in mind.

Michael Atkin has been an analyst and advocate for data management since 1985. His experience spans from the foundations of the information industry to the adoption of semantic technology. He has served as an advisor to financial institutions, global regulators, publishers, consulting firms and technology companies.

The views and opinions expressed in this white paper belong solely to the author, and not necessarily identical to those of Ontotext.

ontotext
info@ontotext.com