Models of Digital Transformation

The role of context, governance, integration and industry models

SEPTEMBER 2019

A White Paper by
Dr. Barry Devlin, 9sight Consulting
barry@9sight.com

Digital transformation begins with data—collected, shared, and stored digitally—and ends with a business changed and charged by its use in automating and augmenting existing processes, with analytics and artificial intelligence. The destination and benefits are obvious: delighted customers, higher sales, streamlined operations, and more. The road to this digital nirvana is becoming clear.

As this white paper shows, IBM's DataOps offerings provide the major stepping-stones to success. Within that, IBM Industry Models offer the foundation for understanding the business in the context of industry norms and aligning to regulations, enabling enterprises to focus on achieving unique and specific business value from the information they hold.

We begin with data itself and its four dirty secrets. These particularly affect the externally sourced data on which digital transformation depends. Cleaning up the problem demands looking beyond the raw data to information and the context in which it is both created and used.

Turning to solutions, we start with IBM Industry Models, their history, their strengths, and their future directions. Industry Models play a key role in discovering and creating the business context of data, as seen in Folksam's experience. This leads to the importance of data governance, which IBM uniquely teams with integration as the basis for building absolute trust in data, a prerequisite for its widespread use.

Finally, we introduce IBM's Watson Knowledge Catalog, an important step in building a comprehensive, federated environment for the creation, management, and widespread availability of metadata describing the business context of all the information needed to navigate the digital transformation journey.

Contents

SPONSORED BY IBM CORPORATION
www.ibm.com

Digital transformation—challenges in context

Thriving at Digital Transformation is wholly dependent on the data or information available and how effectively it is used in decisions, actions, and daily operations. Getting that basic resource right is the greatest challenge.

Digital transformation is both a promise and a curse for pre-millennial businesses and for some of more recent vintage. The promise is of fending off marauding Internet-born giants, detect in real-time customers' most fleeting desires, and reinvent operations for instantaneous action and delivery, all at lower cost. The curse is that, no matter how appealing the destination, the journey is perilous, and the route bounded by steep learnings and dangerous pitfalls.

Understanding your destination—a digital business—is necessary but insufficient. Your path is determined not only by the destination but also by your starting point. Sadly, if you're like many mature businesses, you are likely already struggling to deliver on your current objectives, never mind dealing with the advanced analytical systems and externally sourced data demanded by digital transformation. What is a CDO or CIO to do?

The single most defining characteristic of a digital business is that it is—well—digital. Data and/or information are its lifeblood. We'll return to the distinction between data and information—it's ***really*** important—in a moment. For now, let's use *data*, because in common parlance, digital transformation means becoming data driven. All the opportunities and benefits accrue from basing decisions, actions and operations on data. By extension, all the challenges also begin with data. These fall into three distinct areas:

1. The new sources of data for digital business are largely external to the enterprise and are often poorly defined, variously structured and desperately dirty
2. The path from data to action is far more complex than is generally understood
3. Business—and the people of which it is constituted—disagrees continuously with itself on what most of the data it uses actually means

The trouble with data, especially when externally sourced

Consider two pieces of data: "Jane K. Smith" and "05021953". What do they represent? You can probably make a confident guess about the first item: a person. But could you be more specific—is this a customer, a prospect, or a supplier? The numerical item is more challenging. Is it a phone number, an identifier, or a date of birth? Prior knowledge might influence your judgement: it isn't a US phone number—there are too few digits.

Traditionally, business data was (mostly) well-labelled because it came from Systems of Record. However, labelling may be insufficient for Systems of Insight because the developers' assumptions were undocumented. If "05021953" is labelled "date of birth", is it 2nd May or 5th February? In a digital business—where much data comes from external sources—poor, untrustworthy, or non-existent labelling is a major problem.

In externally sourced data, labelling may be poor, content unreliable, and consistency with other sources dubious.

Even within well-labelled data, fields often contain incorrect values because of lax data entry control or poor data management. The field "first name" may contain titles, business names, or other items. External data has a much higher incidence of such problems. Furthermore, they cannot be fixed "at source." Lack of consistency is another problem.

1. ***Absence of context:*** Data without context is always useless and often dangerous in any type of decision making, especially in self-service analytics
2. ***Ill-fitting data:*** A lack of understanding of data sources, their content and provenance, as well as their relationships with other sources leads to poor data use in business
3. ***Poorly and variously structured data:*** Externally sourced social media and the Internet of Things data comes in all shapes and sizes—unpredictable and ever-changing
4. ***Data's got entropy:*** As data gets further from its original sources and is combined and used in various ways, its quality degrades and is trustworthiness diminishes
5. ***Data overload:*** Too much data, especially contextless, ill-fitting or recycled data, actually impedes business use in all areas of a digital business

Data fields with the same name from different sources may contain compatible data… or they may not, leading to double counting, incorrect joins, and other basic usage issues.

Data from the Internet of Things (IoT) is notoriously dirty with missing values and corrupted readings. Data structures change over time, without notice, and conflicting structures may exist simultaneously. Deliberate data tampering cannot be discounted.

Social media, another key source of external data, is especially challenging. Lacking any significant formal structuring and descriptions, and filled with colloquial usage, abbreviations, and implicit assumptions of context, it presents significant problems in interpreting the content and judging meaning.

Whether data is internally or externally sourced, the fundamental challenge is poor data governance. In digital transformation—compared to traditional business—the consequences are widespread, with potentially more severe impacts on daily operations. This is especially true when we factor in pervasive artificial intelligence (AI), trained with dirty data, that results in erroneous, biased or unethical actions.

We need to return to first principles, to *information architecture (IA)*. As has been discovered, there is no AI without IA. And there is no IA without an agreed language and definitions of the foundational business entities and their context of use. Digital transformation may be—simplistically—about being data driven, but the reality is more complex. We must, finally, address the distinction between data and information and, indeed, think about knowledge and meaning as well. Welcome to the world beyond data.

The winding road from data to action

So far, I've used the words *data* and *information* interchangeably. But they are fundamentally different, and confusing one with the other is at the heart of the second challenge above facing digital transformation. Despite philosophical debates, a simple definition suffices for business computing. Information is the basis for human comprehension and communication; data is a subset of information that has been optimized for and stored in computers[1]. Information is therefore fundamental, while data is derivative.

> Information is the basis for human communication; data is a subset of information, optimized for computers.

Converting information to data strips definitions and descriptors—*context-setting information*[2] (CSI) or metadata—from the information and stores them separately from the data values. IT systems continuously do this, a practice driven in part by the limitations of early computers, but that continues widely to this day. This split data/CSI structure provides for efficient processing and storage but often inhibits business understanding of and insights into the meaning of information.

Solving this problem demands we consider *context* whenever we use information. "Macintosh" is *naked data*. LastName: "Macintosh" allows us to distinguish it from (misspelt) raincoats and computer models, to decide what useful properties we should record, and how we should use this information in different business situations. Defining, discovering and storing context are at the heart of digital transformation and the foundation of catalogs and models, as discussed in the next section.

As we proceed with digital transformation, increasing volumes of data from external sources are embedded in the information fabric of the business, and used and reused across business functions and enterprise boundaries. Successful transformation depends on governing, managing and integrating both data and information with ease and elegance across the entire enterprise. Sadly, that seldom occurs even with internally sourced information, never mind poorly defined and dirty external data.

Understanding that information, rather than data, is the correct starting point suggests we should talk about being information informed instead of data driven. But even information doesn't tell the full story. Businesspeople with access to exactly the same information come to different conclusions. Their internalized information—*knowledge*—differs from person to person based on "gut feel"—personal experience and internal heuristics. Above that is *meaning*—the socially constructed stories we tell ourselves and others about our intentions, moral values and peer pressures that directly influence our decisions. These important considerations are beyond the scope of this paper but will inform future directions for digital transformation. Further details are provided in the appendix, "The modern meaning model."

Context is the golden thread

Woven through the first two digital transformation challenges is *context*. Only when data scientists truly understand context can they onboard and make full use of external data. Data wrangling defines or discovers context relating data to information. In a digital business, fast, insightful actions are what matter. However, most diagrams of digital transformation begin with data driven and end with automated action. Information, knowledge and meaning are consigned to some "magic happens here" cloud of artificial intelligence in the middle. There is no magic in AI. The only real magic is made of business context.

> There is no magic in AI; the only real magic is made of business context.

That leads us to the third challenge to data transformation mentioned above. How do we ensure that businesspeople throughout the enterprise, external providers or receivers of information, as well as IT all "speak the same language" about information used and goals asserted? In fact, they cannot. What they can do is to build a *lingua franca* of context bridging different languages, a golden thread that weaves the relationships between data, information and digital transformation, across the enterprise and beyond. It's time to move from challenges to solutions.

Industry models pin value to business context

Digital transformation begins with knowing your information assets. Industry level models are the foundation of that knowledge. Together with information architecture and governance, industry models show the path from design and development to governance and integration of business context.

From its origins[3] in the mid-1970s, data modelling[4] has been the method of choice for mapping business needs to information and its translation into well-behaved data structures and CSI. Entity-relationship modelling (ERM and its derivatives), as defined then, dominates current practice. Its co-development with relational database (RDB)

theory leaves ERM closely tied to physical RDB implementation, as seen in the often-direct translation from one to the other, and a poor fit to non-relational structures.

As digital transformation progresses, these and similar issues demand a broader perspective on data modelling. Thomas Frisendal[5] discusses many of these issues and provides a three-level, top-down methodology beginning with business concept maps, proceeding to specific business solutions using property graphs, and finally arriving at physical data models aligned with the target data stores: for example, an ERM for an RDBMS or a directed graph for a NoSQL store.

The three-level approach is not new. Since the 1980s, enterprise-wide projects such as data warehouses and master data management led to the creation of enterprise data models (EDM) spanning multiple business functions and applications. The cost of designing such models on a blank (and very large) sheet of paper and the recognition that most data—and processes—are very similar within any industry led to the development of *industry models*, of which IBM is a major provider[6]. These hierarchical—conceptual, logical and physical—data models align well with Frisendal's approach, offering different techniques at the various levels to bridge the gap between business and IT.

Industry models—evolving with digital transformation

Industry models begin with the premise that much of what businesses in the same industry do is common and congruent across all players. In mature and highly regulated industries—such as Financial Services—business terminology, processes, and information needs are much the same for banks from Alabama to Alaska: they offer checking accounts and loans to individuals and companies, make profits by charging interest and fees, and so on. Even between America and Africa, business models are mostly the same, with regulations and even religious beliefs driving only limited divergence. Industry models offer a common business context upon which unique information and process differentiators can be easily built, without having to model the entire business from scratch.

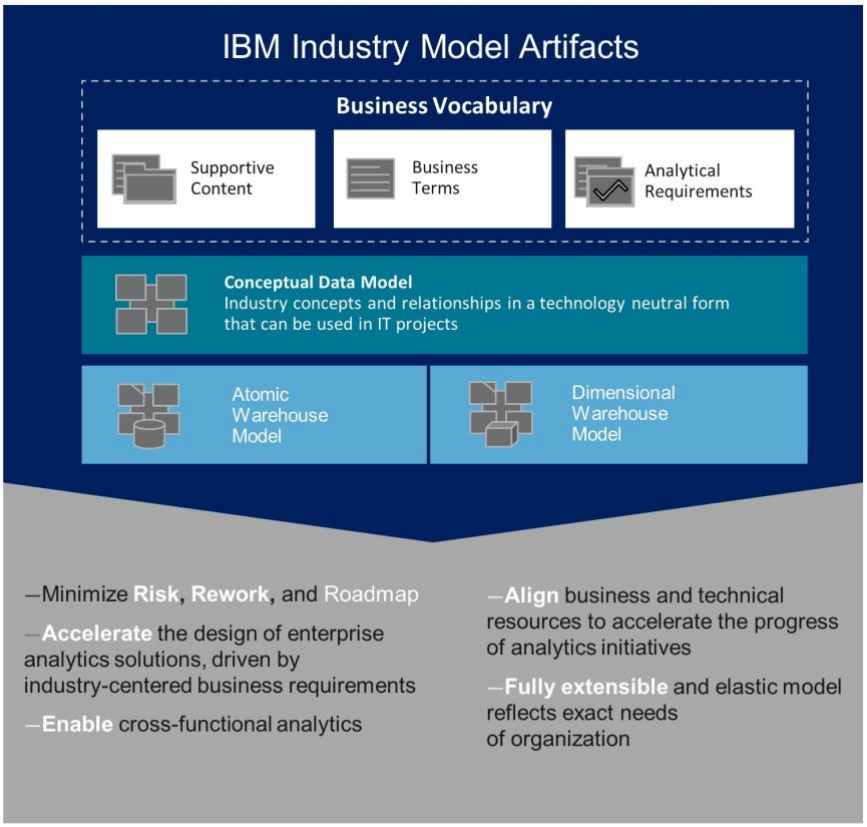
Industry models offer a common business context to build unique information and process differentiators, without having to model the entire business from scratch.

In emerging or loosely regulated industries, terminology, processes, and information needs may vary initially. Existing models from already established industries can be the seeds from which new models grow. People, for example—customers, prospects, employees and partners—have largely identical characteristics whether in Insurance or Online Gaming. Over time, best practices emerge and, as commonality grows across business behaviors, industry models can become established as the industry matures.

An overview of IBM Industry Models

Based on both these observations, IBM Industry Models embed more than two decades of experience and best practice in models for Banking and Financial Services, Insurance, Telecommunications, Retail, Energy and Utility, and Healthcare.

All these models start with a set of business-oriented assets comprising a vocabulary or glossary of business terminology, a set of analytical or reporting requirements, and supportive content that maps relevant governmental or industry standards and regulations to the models. The aim is to provide business users with a standardized and agreed approach to driving business goals and activities, first at a business level—with appropriate customization—and second, with IT as the basis for the required technical data structures.



IBM Industry Model Artifacts and Benefits

At the next level down, a common, enterprise-wide structure provides a logical, generic entity relationship model representing all the essential characteristics of the industry in question. As at the higher level, this model represents the best practices and knowledge of the industry and may be expected to match a significant percentage of the business requirements. It is designed, however, for easy modification and extension to meet specific client needs, allowing changes to be carried into other levels of the model. In addition, this level provides a strong foundation for integration projects that consolidate data from multiple diverse sources, a widespread need in many industries.

The final, lowest levels provide design models for physical implementation of systems and applications by the business. Examples include an enterprise data warehouse and lakes, data marts and analytics projects, and operational applications supporting specific business needs with broad governance requirements.

Traditionally, clients have leveraged IBM Industry Models in the design and development phase of major IT projects. The most common use has been in the development of Systems of Insight such as data warehouse / data mart environments. Such projects have a significant enterprise-wide aspect and require the development of a commonly agreed information asset where differences in definition and use of data across different departments of the business must be reconciled. Such usage has also led to the application of IBM Industry Models in efforts to re-architect IT design, development, and deployment toward more integrated and well-governed approaches.

Five top benefits of IBM Industry Models

Whether building new systems, implementing data governance, or driving information usage via self-service analytics:

1. ***Reduced time to value:*** Reusing existing industry knowledge speeds delivery
2. ***Lower cost of context discovery:*** A single source of common terminology, agreed usage, industry regulations, and more improves productivity

3. ***Increased project success rates:*** Reuse and adaptation of proven knowledge reduces failure rates
4. ***Reduced redundancy:*** Avoids re-inventing the same materials for different programs and shares expertise across the organization
5. ***Flexibility and scalability:*** Provides an agreed starting point for work and offer ease of expansion to other related business areas

Folksam treads the path from development to governance

An example of this is Folksam—a large, long-standing insurance company in Sweden operating in multiple areas of the insurance market, including life, property, and commercial. Having developed their own proprietary data models over a number of years, they had experienced the challenges of this method: it can be expensive and time consuming to develop a model and can lead to data quality issues and limited lineage traceability.

When the company decided to adopt a model-driven development approach for enterprise-wide systems across all areas of IT in 2012, it adopted IBM's Insurance Information Warehouse (IIW) as the foundation. This was chosen because it offered an extensive, common industry model that was more easily agreed and customized to specific business needs. Using this reference model addressed many of the problems of proprietary model development, although the level of business involvement was under-estimated.

IBM's IIW provides a template where business terms from different functions can be compared and consolidated into Folksam's Common Information Model.

With the advent of digital transformation and increasing compliance demands from Financial Regulators, Folksam saw that new initiatives in data/information governance were required. The business further recognized that IIW could also be applied to this activity. In fact, switching the focus from the more technical, lower levels of the model to the more business-oriented, upper levels has led to greater business involvement, more effective use of data, and improved return on investment.

A team reporting to the Chief Data Officer is building a glossary and taxonomy of agreed business terms for use in setting business goals, defining key performance indicators, and delivering requirements to IT for new systems' development and the upgrades. With ever more diverse sources of data and increasingly strict compliance requirements, tracing lineage from business context to IT artifacts is a key driver for data governance. IIW provides a foundational template against which business terms from different functions can be compared and consolidated into Folksam's Common Information Model.

In parallel, Folksam have embarked on the redevelopment of their data management environment in line with modern thinking, including a revamped data warehouse and data lake, with virtualized access where appropriate across all storage platforms. Within the warehouse, Folksam are aiming for "physical conformance"—full reconciliation—of data from all sources, while the aim for the data lake is "catalog conformance"—the use of a common vocabulary for all data. IIW addresses both needs with a self-consistent model at both business level for catalog-based data discovery and access, and at a detailed atomic design level for the data warehouse.

Starting with a generic industry model allows Folksam to benefit from intellectual capital from across the breadth of the insurance industry, while the extensibility of IIW enables the unique features of the company's business approach to be fully described and embedded in their own information model. These advantages are further enhanced by the availability of a linked detailed, atomic warehouse model that can be used to keep the physical data warehouse in full alignment with their Common Information Model.

Folksam's journey is not uncommon. The initial IT focus on using Industry Models in design and development of enterprise-scope systems—while relatively successful—has shifted toward a more business-oriented, IT-supported view of information governance.

With a generic industry model, Folksam benefitted from intellectual capital from across the entire insurance industry as input to their unique model.

Integrating good governance

Success in digital transformation is wholly dependent on the quality of the data/information available and how effectively it is used in decision making, action taking, and operations. Good governance of this data/information is therefore mandatory.

As we've already seen, a key difficulty in digital transformation is—in a soundbite—big, dirty data. The bad news is that this is a long-standing, cross-industry problem; and it's growing. The good news is that solutions exist, are getting better, and growing in scope. IBM gathers these solutions under the banner of "DataOps". It means: *getting the right data (and information) right*.

Governance matters, quality compels

According to the Data Governance Institute (DGI), data governance is *"a system of decision rights and accountabilities for information-related processes, executed according to agreed-upon models which describe who can take what actions with what information, and when, under what circumstances, using what methods."*[7] While there are various possible levels of complexity of such a system, the formality and importance of governance grow as the shared and interdependent use of information increases.

In essence, data governance is a structure and set of processes by which decisions about the management, use and integration of information are made. The DGI Data Governance Framework[8] defines ten components of a data/information governance program. They speak first to the people and their roles and responsibilities within the organization. Second, they define the processes—proactive, reactive and ongoing—necessary to ensure and manage the quality of the information used across the organization. Third, and of central importance to the success of governance processes, is a component named *Data Rules and Definitions*. The required rules and definitions apply both to information and data, as well as to the interrelationship between them.

Establishing these rules and definitions falls within the remit of data modelling in the broadest sense of the term. As digital businesses are established, these models take on an increasingly central role, and—with data coming from ever more diverse sources—hint at a necessary convergence of thinking around governance and integration.

IBM has combined governance with data integration because digital transformation redefines data ownership, stewardship, and the processes around them.

IBM DataOps

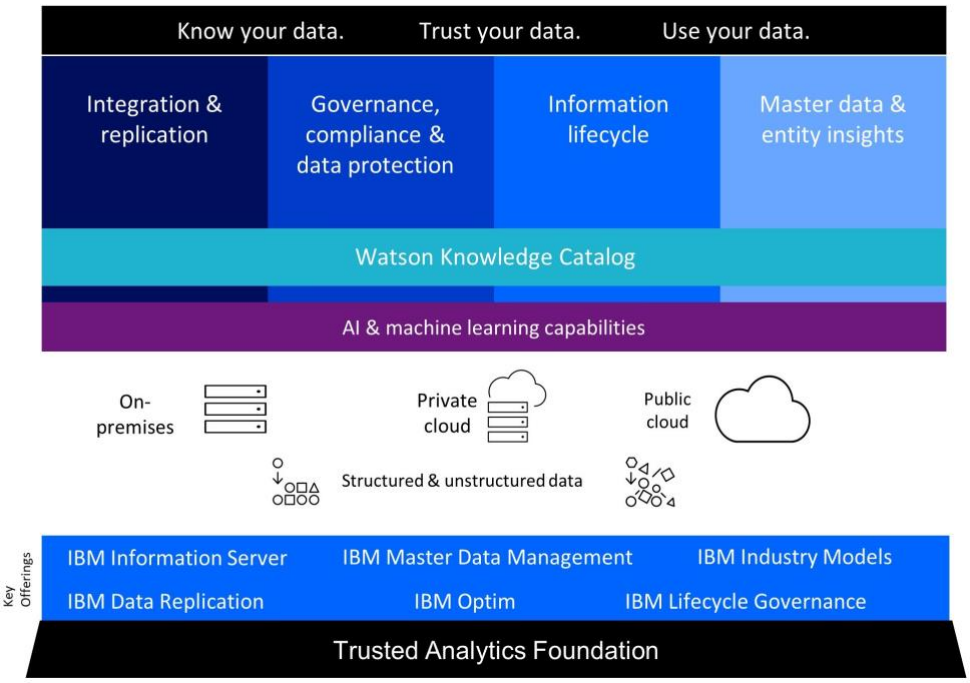
While most data governance thinking emphasizes the organizational and process aspects of the topic, IBM has taken the bold step of combining it with the more technological field of data integration. The reason is simple: digital transformation undermines the foundations of data ownership, stewardship, and the processes around them. In a traditional business, where most data has clear Systems of Record sources,

data governance could often follow functional ownership lines in the business. Systems of Insight, such as data warehouses, where data is integrated, followed from there.

Today, digital transformation disrupts both data sourcing and the importance of integration. Data comes from many more sources—external providers, Systems of Engagement, and so on—revamping traditional governance responsibilities. Systems where data coalesces—data lakes, AI systems, operational analytics, and more—have become central to the digital business. Data sources and targets are intermixed. Data lineage can start and end anywhere. Governance and integration now go hand in hand.

In order to ensure data is trustworthy for numerous, high-value decisions and actions, data lineage must be comprehensive. For speed of reaction, existing data must be easily discovered, its preparation streamlined, its meaning clear. New data must be quickly on-boarded, clearly defined, linked to existing information. All this data must be managed across multiple environments, duplication and conflict eliminated, personal information protected, government regulations respected. The list seems endless. The need to unify governance and integration is obvious. The motto is: "Know your data, trust your data, use your data."

Figure 2:
IBM DataOps framework



The IBM DataOps framework is shown in Figure 2. It consists of an open and extensible platform, designed for scalability and parallel processing and equally supporting all types of data from traditionally highly structured records to more loosely structured information, both on-premises and in all varieties of cloud. It includes components from a wide range to data management and governance areas, including data integration, preparation and cleansing, master data management, and lifecycle governance, as well as taking extensive advantage of emerging advanced AI technologies.

Two further components are of relevance in the context of this paper. First is the inclusion of Industry Models as a foundation for data governance and the drive to more easily create and manage the business context of information, already covered at length.

Second is the centrally positioned Watson Knowledge Catalog, which leads us neatly into the final section of this paper. We have so far emphasized the importance of business context, from the creation and collection of relevant and valuable information, through its storage and management as data, to its valid and vital use in decision making and action taking, whether in self-service analytics or AI-driven automation. But business context is—in practice—nothing more than information that must be stored, managed and used consistently and easily across the full range of the business and IT.

How is this to be achieved? The simple answer is the Watson Knowledge Catalog. However, as digital transformation progresses, we need a more comprehensive reply.

A firm foundation for business context

The creation and management of business context is a complex undertaking. Previous attempts have had limited success. IBM's current approach—built on its DataOps initiative, IBM Industry Models, open-source projects and tooling, and extensive AI—holds high promise.

Since the earliest data warehouses, experts have recognized the key role of metadata[9]—technical, operational and business—in building and maintaining such systems, and in empowering businesspeople to benefit from this newly available information. Many projects were undertaken, and tools built to collect, create, manage and deliver metadata. Most failed, through a combination of business and technical issues, and interest waned.

The emergence of data lakes in this decade—with their multiple dirty data sources, their distributed, disparate, and poorly described data stores, and their often-naïve self-service users—has reawakened interest but simultaneously presents an even more challenging set of technical and business issues. The more recent growth in advanced analytics, AI and machine learning—all of which depend intrinsically on well-understood and clean information—has further raised the stakes. Solving the challenge of metadata—or better, context-setting information—is a vital step in achieving digital transformation.

Solving the challenge of context-setting information is a vital step in achieving digital transformation.

Driven by the needs of some of its largest customers, IBM is at the start of a journey to create a comprehensive solution to the management and use of *Smart Metadata* (or better called Context Setting Information - CSI) throughout the entire data lake and related Systems of Insight and Engagement. The first step, arguably mundane but absolutely necessary is the integration of IBM's existing catalog offerings—IBM Information Governance Catalog, IBM Watson Knowledge Catalog, as well as the catalog within IBM Cloud Pak for Data into one offering, called Watson Knowledge Catalog, with a single strategy and roadmap.

More exciting are the architectural and technological directions that are already in progress. Smart Metadata includes an extended set of CSI categories—social, operational, business, technical, user activity logging, recommendations, and more. With such a range of structures, sources and uses, it would be madness to try to cram all this into a single repository. And IBM does not, rather envisaging a set of federated context stores operating according to a new Apache 2.0 open metadata and governance project called ODPi Egeria[10]. This development, managed by a group of software companies, including IBM, and business in banking and telcoms, aims to provide standards, APIs, services, frameworks, and a graph database linking different metadata stores to support automation, business value, and connection for metadata and governance.

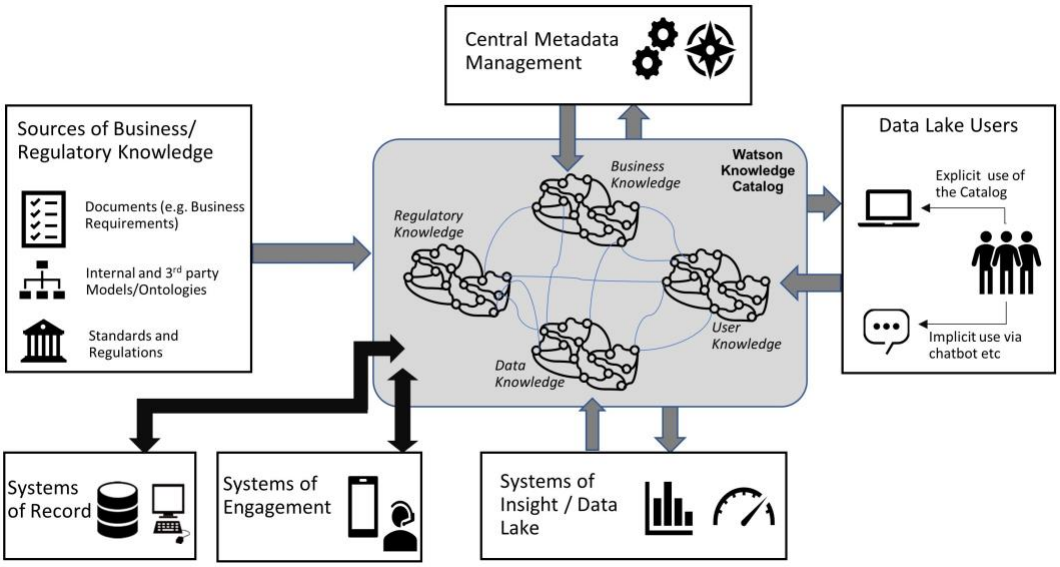


Figure 3: The Watson Knowledge Catalog Landscape

As shown in Figure 3, IBM's approach is to create the Watson Knowledge Catalog, a federated collection of technical, business and user knowledge stores, supported by a dictionary of terms and relationships generated via machine learning.

Of course, as all practitioners know, any metadata project stands or falls on the metadata content it holds. Sourcing of technical metadata from Systems of Record and Engagement is well-understood. Data warehouse developers have been doing it for decades, but AI-based tools are enabling faster and more accurate metadata generation. In a data lake environment, such mapping is particularly important. AI can automatically classify the raw data and then enable connecting the gathered technical metadata with related business metadata via AI-driven auto term assignment. This approach is especially vital in the case of the poorly and variously structured information gathered from external sources.

IBM Industry Models provide a unique source of business knowledge about information required and its data structure. They ease creation of customer-specific glossaries and models, as well as providing supportive information on industry, open and governmental standards, models and regulations. These models are also evolving to support modern needs, such as self-service, AI-enabled analytics by new classes of users from data scientists to citizen analysts, radically faster time to value, and the blurring of boundaries between traditional operational, informational and collaborative worlds.

Using AI techniques with the Industry Models enables the automated connection of business terms to one another, for instance, connecting taxonomies describing a particular regulation with the central canonical set of business terms for the enterprise, a type of mapping that is traditionally a highly manual, oft-repeated, and error prone task.

Collection of user knowledge, supported by collaborative and AI tools, is an area of growing interest across the industry, and IBM also aims to gather extensive context-setting information in this area. The contextual information created as people onboard and wrangle external data, explore and link it to internal data, and use it in self-service analytics contributes a wealth of detail and color to traditional technical metadata. It is also a vital source of business knowledge—of how the business actually operates, as opposed to the formal processes by which it believes it does.

This federated Watson Knowledge Catalog is the foundation for management and governance of the enterprise's information assets by IT. It tracks technical lineage, aids in discoverability and classification of enterprise data sources, and offers robust and

automated policy management. Knowing enterprise information to this extent allows business to trust it.

The extensive metadata / CSI in the Watson Knowledge Catalog is made available to businesspeople in the environment and tools they use. A business-friendly view of all information and its provenance enables self-service consumption, driven by intelligent discoverability and classification of enterprise, external and local data sources. Data science and analytics, as well as data preparation are supported by an operationalized governance program with active policy management. Trusting information is the basis for its productive use across the enterprise. But, what about the path to implementation?

IBM's CDO adopts the DataOps recipe

This journey is illustrated by the involvement of the Global Chief Data Office at IBM (IBM CDO) in the development, testing and early adoption of the above approaches. The mission of the IBM CDO is to deliver a trusted, enterprise-wide, data and AI backbone for transforming IBM into an AI Enterprise. Achieving this goal starts with developing a clear data strategy, and then creating and driving enterprise-wide governance and management systems for all the data used by IBM to run and manage its business. Success lies in becoming seen as the central data source and AI framework for IBM and being sought out for deep data and analytics partnerships by all segments of the enterprise.

Digital transformation starts with a clear data strategy and enterprise-wide governance and management systems for all the data used to run and manage the business.

Like many large enterprises, IBM is challenged by the time-consuming and costly maintenance of metadata, both business and technical, as it tries to make its enormous but fragile data lake more agile to meet rapidly changing business needs. One key strategy to address this challenge is to create a virtual metadata repository or catalog of all necessary enterprise business and technical knowledge. This will support natural language queries and self-service BI from all businesspeople through auto-discovery and classification of information, as well as the auto-generation of new data lake artefacts, such as data marts and ETL.

Work on this strategy is still in the early stages but has already benefited from its use of IBM Industry Models. This may, at first, seem surprising, given that no Industry Model covering IBM's business exists. However, the generality of the tooling and models has enabled the IBM CDO to validate and extend their initial ontologies using intellectual capital embedded in the Industry Models about common entities such customer, location and product. This ensures that IBM's internal taxonomy is consistent with cross-industry best practices, common regulatory frameworks, and external industry standards.

Conclusions

The path of digital transformation is today partly paved with aspirations. IBM is upgrading it to a free-flowing highway with business context, defined in Industry Models, a DataOps environment, and a new initiative to deliver the comprehensive Watson Knowledge Catalog.

Digital transformation is both the goal and the journey for businesses in every industry on every continent. The goal—to maximize business benefit from the tsunami of data being generated everywhere—is plausible and welcome. The journey—in its route and waypoints—does, however, enter uncharted territories and cross dangerous terrain. So, how can businesses arrive safely and comfortably at their destination?

First, they must recognize that, despite calls to become data driven, it is less about data and more about information and, in particular, about the context that defines the

difference between them. Context-setting information—from basic technical metadata through specific business taxonomies and generic ontologies to industry regulations—is vital on every step of the digital transformation path, from discovering or creating it when onboarding external data to ensuring the ethical use of AI models. Defining, aligning and using context is the *raison d'être* of IBM Industry Models. Maybe, you could do it without them, but why build a Porsche from its parts when you can simply rent one to get from New York to Washington?

A viable environment and comprehensive tooling for defining, discovering, managing, and benefiting from context are the next critical factors for successful digital transformation. Recognizing that the two are intimately linked, IBM has created a DataOps framework that puts the guardrails on either side of the road to a digital business. With an extensive set of both mature and emerging technology, IBM can offer tools from Information Server and integration, through Master Data Management and Lifecycle Governance, and a converging set of catalog tools.

Finally, we must recognize that we are still as an industry taking the first steps on what will be a long and arduous journey. Our prior approach has been more "garbage in, garbage out" than best practice. Recognizing that context management provides the correct starting point and the comprehensive roadmap to information governance and data integration, IBM has launched an initiative to define and create the federated IBM Watson Knowledge Catalog based on open source principles and technology.

IBM's three-part roadmap—context defined through IBM Industry Models; known, trusted and widely used information delivered via the IBM DataOps environment; and the comprehensive, federated, IBM Watson Knowledge Catalog to enable businesspeople and IT to truly collaborate on information-informed decisions and actions—will transform the currently rutted and rocky track to digital business into a high-quality, well-travelled superhighway.

IBM's three steps to digital transformation: (1) Industry Models for business context, (2) trusted information via DataOps, and (3) the comprehensive Watson Knowledge Catalog.

Appendix: The modern meaning model

The modern meaning model (m3), introduced in my 2013 book, *"Business unIntelligence,"* defines the difference between data and information, and introduces context-setting information (CSI) as key aspects of IA. *Information* is the physical representation of human communication and thinking, either created directly or gathered through by machines, varying in degree of structure, and today primarily digital in form. *Data* is a subset of information that has been structured specifically for computational needs, consisting of well-bounded numerical and textual elements. CSI (simplistically, metadata) is also a subset of information that describes the relationship between data and information.

It is important to note that information is conceptually a precursor of data. Understanding what we want to communicate—within the business, within IT, and between them—is the starting point for digital transformation. Startlingly, even after three decades of business intelligence, this low bar is seldom met. Only when CSI is treated as a first-class component of IA can we hope to progress. Information— data and CSI—is the physical locus of m3: it is the entire content of our data systems, from billions of smartphones to the server farms of the cloud, and all the on-premises systems in between.

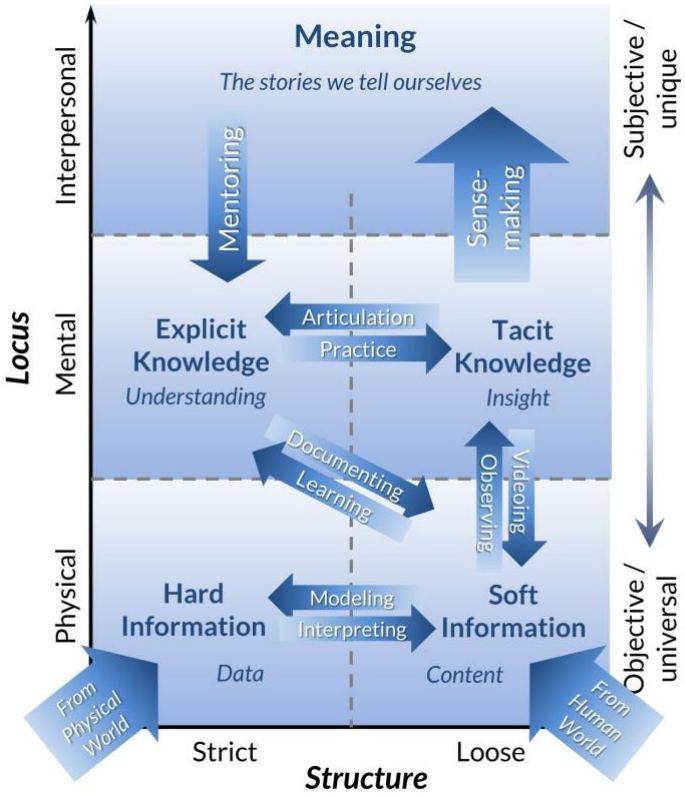


Figure 4: The modern meaning model (m3)

Knowledge, the mental locus in m3, is information transmuted into human knowledge in order to be truly useful. It takes two forms. Explicit knowledge relates to information directly available and more easily documented. Tacit knowledge is more internal and difficult to document; the usual example is how to ride a bike. In decision making, tacit knowledge equates to insight that goes beyond explicit, immediately available, physical information, often called *gut feel*. The insight may come from predigested information or internal mental heuristics prompted by the information at hand. Digital transformation demands agility in moving between explicit and tacit knowledge, especially in relating prior and often unconscious information and context to the current situation.

The interpersonal locus, *meaning*, reflects the reality that we humans are, at heart, social animals engaged in business as a social enterprise. Understanding what information means—and, therefore, what actions are appropriate—always occur in a social context. Meaning is, in the final reckoning, the stories we tell ourselves and others about the information we gather and the knowledge we hold. Decisions are seldom, if ever, fully rational. They are influenced by our emotional states, social conventions, and especially by our intentions, swayed by our own and others' expectations. It is in this locus that attempts to apply AI come face to face with ethical and societal context and the need to pay attention to personal and social implications of decisions and actions.

To learn more about IBM Industry Models, visit ibm.com/industry-models

Ready to assess your digital transformation maturity? Start the assessment today. Visit: ibm.com/analytics/digital-transformation-assessment/

Dr. Barry Devlin is among the foremost authorities on business insight and one of the founders of data warehousing, having published the first architectural paper on the topic in 1988. With over 30 years of IT experience, including 20 years with IBM as a Distinguished Engineer, he is a widely respected analyst, consultant, lecturer and author of the seminal book, "Data Warehouse—from Architecture to Implementation" and numerous White Papers. His book, ***"Business unIntelligence—Insight and Innovation Beyond Analytics and Big Data"*** *(http://bit.ly/BunI-TP2) was published in October 2013.*

Barry is founder and principal of 9sight Consulting. He specializes in the human, organizational and technological implications of deep business insight solutions combining all aspects of internally and externally sourced information, analytics, and artificial intelligence. A regular contributor to TDWI Upside, Twitter (@BarryDevlin), LinkedIn, and more, Barry is based in Bristol. UK and operates worldwide.

1 Devlin, Barry, *"Business unIntelligence"*, (2013), Technics Publications LLC, http://bit.ly/BunI_Book

2 I prefer the term *context-setting information* to the widely used *metadata* because it more precisely defines the intention and meaning of this subset of information.

3 Chen, P. P.-S., The Entity-Relationship Model—Toward a Unified View of Data. ACM TODS, Vol. 1(1), (1976)

4 As in normal parlance, I use the term *data model* to include *information model*, unless I need to specifically distinguish between them.

5 Frisendal, Thomas, *"Graph Data Modeling for NoSQL and SQL"*, Technics Publications, NJ, (2016)

6 https://www.ibm.com/analytics/industry-models

7 Even here, the data / information dichotomy is evident. The discipline is usually called *data governance*, but the definition speaks solely to information.

8 Thomas, Gwen, *"The DGI Data Governance Framework"*, http://www.datagovernance.com/the-dgi-framework/

9 Marco, David, *"Building and Managing the Meta Data Repository: A Full Lifecycle Guide"*, Wiley, NY, (2000)

10 https://egeria.odpi.org/

SPONSORED BY IBM CORPORATION

www.ibm.com