The Principles of Data Management

Michael Atkin, Managing Director, EDM Council

August 2016

Data management is about managing meaning. And the management of meaning has been elusive because the harmonized capability didn't exist – until now. Data management is not about governance, although governance is certainly needed to get organizations to change the way they operate. Data management is not about lineage or legacy integration, although both tasks are needed. Let me give you some perspective on the management of meaning to help you understand why I am so optimistic about this moment.

My background stems from the beginnings of the information industry. By the information industry, I mean companies who manufacture, collect, aggregate and organize data. Let's call them publishers. These are the owners of intellectual property about legal rulings, technical studies, scientific research, medical innovation, economic activity and financial services. The one thing that all these information companies have in common is that they sell content. Their customers want to acquire "hard to obtain" knowledge and insights. This is the objective of an information company – the organization, presentation and sale of useful content. The more relevant the content, the more an information company can sell.

There are lots of important lessons that we can learn from the origins of the information industry and the discoveries of information companies. These lessons directly apply to the task that financial institutions face in managing data across the enterprise and (in particular) the task that regulators have in managing linked risk and unraveling interconnectivity.

Let's take a brief look at what we discovered about the "business of information" starting with the early (pre-computer) days when this valuable content was in the form of journals, academic studies and scientific papers. In those days, in order to get value from the content, one had to read the journals, understand the science and hope to find a nugget of meaningful content in support of their research objectives. In short – it was hard to find and the business of information was almost nascent.

Let's jump ahead from a historical perspective. The first technological revolution is happening and reliable business computing is emerging. COBOL is developed. Bell Labs develops UNIX. IBM develops the floppy disk. Researchers at Xerox park develop the mouse. Packet switching is released. PCs make their debut. VisiCalc emerges. Then comes MS-DOS. The National Science Foundation creates the precursor to the Internet. Microsoft releases Windows. Client Server technology makes computing business friendly. And a lot more developments in information technology are happening.

\

And while all this IT activity was underway, IBM (under the leadership of Charles Goldfarb) created an important standard known as "Generalized Markup Language" (GML) – to enable the sharing of machine-readable large-project documents in government, law and industry. GML became Standard Generalized Markup Language (SGML) and was propelled forward because of the need to search for information in electronic form to support military applications.

This is where we have to really start paying attention. Because this – in what became known at the American Publishers Association's Electronic Manuscript Project - is the beginning of the modern information industry. It is the development of this standard – SGML – that enabled search engines and information retrieval. It is important to understand that SGML is a standard for how to markup (or tag)

content in a document. It is metadata. It is based on the idea that documents have structural and semantic elements that can be described without reference to how they should be displayed. In essence, print documents can now be re-adapted for computer screen display.

SGML is important because it is based on the principles of descriptive markup to indicate the nature and content of the data rather than just conveying information about how it should be processed. With the advent of SGML, the information industry takes a big step forward. The information industry goes electronic. We can search for concepts in documents. Headings are identified as headings. Key words can be located. We can start structuring Boolean queries to narrow and refine our searches. Content can be organized and linked. Data in electronic form can be queried. SGML makes it possible to not just find data but to re-use and share it across different machines.

The information business is now an industry. Specialist information companies like Lockheed Dialog, Bibliographic Reference Service, Mead Data Central and the National Technical Information Service emerge. And big information originators like McGraw Hill, Dow Jones, The NY Times, Knight Ridder, The Washington Post, Reuters, Thomson Publishing, etc. have a way of extending their valuable content.

Let's jump ahead still further. Building off of packet networks, the US Department of Defense invents the Transmission Control Protocol (TCP) and the Internet Protocol (IP) – the standards that allow networks to communicate with each other. Before this – firms had to use their own network infrastructures or lease network access from commercial packet switching networks. So a big barrier to the rise of the information industry is eliminated.

But the real story centers on a British scientist - Tim Berners-Lee. In March of 1989, Berners-Lee submitted a proposal to his management at the European Organization for Nuclear Research (known as CERN) for a new kind of information management system. That proposal became the Hypertext Transfer Protocol (HTTP) – the underlying basis for the World Wide Web. HTTP allows for communication between computers using location identifiers - URIs and URLs

Tim Berners-Lee also created the HTML specification (just SGML with hyperlinks). But HTML is really important for the information industry. It's important because it allows for the separation of text, images and other content from the structure and format of the document on a computer screen. So in addition to creating the World Wide Web, the result of all this development is that we can now find critical information via keyword searching, retrieve it and display it on any device using open source standards. All of this because of a single data standard concept – that being SGML (and its extension to HTML) for tagging of data. So the concept of data tagging as the baseline for knowledge representation has been established.

This is where the financial information industry is standing today. We have lots of innovation from an IT perspective (database technology, business intelligence software, big data processing, cloud computing, query tools, machine learning, etc.) but we are still relying on the contribution of HTML for data tagging. And HTML has limitations because it is still just a markup language for documents. HTML allows us to make document level assertions – like this document has a title. What it doesn't do is create relationships that link bits of information together. TCP/IP, HTTP and XML let machines know how to read packets of bytes, but they do not tell machines what the information contained in the packets means – and that is critical.

And this is where the semantic web steps in to elevate data management to another level. So let me try to put this into context. The Semantic Web story starts in the year 2000 when the Defense Advanced Research Projects Agency (DARPA) started working on something known as agent-based markup language. The goal was to develop a new way to interpret data and decipher meaning to allow the military to react quickly and respond coherently to terrorist events and other forms of insurgency.

Cyber-crime, threats to national security, opposing irregular groups or unraveling linked risk in any complex environment requires the ability to analyze large data sets. That means there is the need to share data across federated and interdependent systems. And that data can come from unreliable sources or be incomplete. The DOD realized that there was no way to meet these information sharing needs until they solved their federation problem. In the DoD terms is this across Army, Navy, Air Force and Marines – and across Logistics, HR, Intelligence, Finance, Command and Control.

This is analogous to any large financial institution needing to share data across capital markets, retail lending, commercial banking, wealth management, asset servicing – and across risk, finance, marketing and compliance – and across issuance, trading, clearance and settlement – and across liquidity and collateral – and in light of continually changing macro-economic, political, cultural and operational events.

Let me put it succinctly – there is absolutely no way to solve this problem using conventional data management technologies and approaches. We can't analyze this type of complexity by restructuring our relational columns and rows – particularly when our data is not harmonized across the enterprise – especially when we have to do it in near real time and under pressure associated with times of stress. In order to solve the data federation problem in the financial industry, four things are needed:

1. The first is a precise identification of the "things" in our industry. And for the sake of simplicity, let's characterize those things as "instruments" and "business entities." That's why the financial industry is spending so much time on Legal Entity Identification and Unique Instrument Identification. Knowledge management starts with identification.

2. The second is a precise description of the financial instruments and processes that underpin our industry. This is where content engineering comes into play. Let's start by understanding that data is very precise at the point of origination. And it's precise because it was the result of a legal process that carries with it contractual obligation. It is also important to understand that most everything we do in the financial industry (from trading to clearing and settling trades to modifying master files via corporate events) is an agreement that carries with it a legal obligation. And this contractual certainty gives us an anchor because it allows us to link the data to precisely what it means. So the second principle is to forget the nomenclature (the words) used to create databases and link everything to what it means based on the obligations of the contract.

3. The third is to identify where this data is located. Data doesn't have to sit in one place, we just have to be able to find it. And that is accomplished via name space management with URIs and URLs courtesy of Sir Berners-Lee and his colleagues at CERN.

4. The final piece is a method of unshackling our data from the limitations of relational databases. And this is where the final piece of the puzzle comes into play (another taxpayer gift from the DoD) in the form of triple store processing. Triple store processing is what we all mean when we talk about the Semantic Web or Web 2.0.

You don't have to be a geek to understand the value of triple store processing. All you have to understand is that data is organized into groups of three (hence triples) that contain subjects and objects that are linked together by predicates and verbs. And these concepts are all precisely defined based on the terms, conditions and obligations of the contract. And once you define these concepts, you can link them together like tinker toys. That is facilitated by an ontology which captures the meaning of information elements and their relationship in a form that can be automatically processed by computers. Remember our DoD federation problem and the importance of having a common reference point of what each term means (i.e. a "tank" in one domain is related to a "liquid" in another domain it is related to a "vehicle").

Semantic processing (which is now a W3C standard) was a huge breakthrough for content management. And because it is an open standard it has propelled lots of companies into the world of knowledge management. It is the backbone of the Semantic Web. It is the infrastructure for bio-medical engineering in areas such as cancer research and for the human genome project. And it is the basis for information companies like what Google is doing with knowledge graphs. DOD has led the charge and solved the Information Systems federation problem with TCP/IP. DARPA created the Defense Agent Markup Language program to facilitate information federation. W3C took the work funded by DARPA and created the Web Ontology Language specifications. Taken together they form the standards on which an Enterprise Information Web can be formed.

This same approach to finding, interpreting and linking data is now available to be used by financial institutions to harmonize data, capitalize on opportunity and unravel systemic risk. That's why the EDM Council is putting so much emphasis on the Financial Industry Business Ontology (FIBO). It solves the data harmonization problem. It solves the need for scenario-based analysis problem. It gets us out of the business of data wrangling and into the business of data application. And it does so in a way that is cost-efficient, non-intrusive, based on open source standards and governed by trusted processes.

Let me conclude by putting all of this history lesson into current context. I call this the "Principles of Data Management" which I have divided into these four core objectives:

1. First are the data objectives. These start with the implementation of the data content infrastructure. That's what this presentation has been about. I like to think of this as the "holy trinity" of data management – identify, describe and locate. Without this infrastructure – the objective of data management will remain elusive. This is an activity that is common to all and where we should all be working collectively. And for the first time - this is absolutely and entirely possible.

2. The second data objective is about data quality. This is about defining the requirements for fit-for-purpose data against all the dimensions of data quality. This is about reverse engineering business processes and unraveling data lineage. The result is definition of critical data attributes and an understanding of how the data is manufactured within your organization.

3. After that – the task is about managing people. That's why we need data governance. It is about getting stakeholders to behave in a new way. It's about the development of policies and the assignment of responsibility. It's about toll gates and authorizations. It's about systems of record and provisioning points. It's about funding. It's about changing culture in the midst of operational turmoil - all while maintaining the continuity of business.

4. And finally it is about integration. There is no getting around the scope of work required to perform cross-references and manage interdependencies. The scope of work is significant and requires coordination across your entire operational ecosystem.

There you have it. This is my view of data management based on my observations of this industry starting way back in 1985 where the biggest technological innovation I encountered was getting a dumb Wang terminal installed right on my desktop. But let me suggest that this is the most exciting time ever in the information industry. The technological infrastructure exists. The mechanism to overcome organizational inertia (in the form of regulatory requirement) exists. The need is real. The payoff is extraordinary. The management of meaning will enhance productivity and usher in a new era of business opportunity – as long as we get the content infrastructure in place. So as the First Emperor of China in the year 221 BC declared, "STANDARDS WILL UNIFY THE EMPIRE."