Original Article

Enterprise metadata management: How consolidation simplifies control

Ron Roszkiewicz

career in graphic arts, publishing and technology spans 30 years. The first 13 years were spent in various editorial and production positions at Sterling Publishing, Charles Scribner's Sons, Macmillan Publishing and Time-Life Books. The next 17 years were spent developing technology for graphic arts, prepress and publishing industries at Scitex, Xyvision, Dainippon Screen and Pound Hill Software. Currently, Roszkiewicz consults with companies on strategies for transforming to digital workflows, digital asset management implementations, collaboration and product positioning with finitiv of San Mateo, California. He is a frequent contributor to The Seybold Report.

ABSTRACT Metadata has become a pesky fact of life for many businesses. In fact, one of the reasons many companies adopt digital asset management systems is to consolidate what are called silos of information into one controllable repository. The same is true of metadata. There exist relationships and dependencies between disparate systems that also call out for consolidation into a corporate taxonomy.

Journal of Digital Asset Management (2010) 6, 291–297. doi:10.1057/dam.2010.32

Keywords: metadata; metamodels; enterprise; consolidate

OVERVIEW

Metadata has become a pesky fact of life for many businesses. In the early days of establishing a web presence, it was attributed, and rightly so, to bringing more exposure to a site. That realization led to tactics for search engine optimization. As digital assets grew, the realization emerged that reducing the time searching for them would make a workgroup more productive, and that increased control would contribute to compliance requirements that arrived as part of tightened corporate governance rules. This realization led to adoption of taxonomies, controlled vocabularies, schema standards and the content management systems that would control them.

The latest realizations have to do with workflow and data input streams. Digital asset systems do not a workflow make. They may be an important component of a marketing supply chain, but are usually just a link in the larger manufacturing supply chain. Other systems, such as customer relationship management (CRM) and enterprise resource planning (ERP) are also

important and often share data through links to other chains. Each of these systems also has their own metadata as part of their database. In fact, one of the reasons that many companies adopt digital asset management systems is to consolidate what are called silos of information into one controllable repository. The same is true of metadata. There exist relationships and dependencies between disparate systems that also call out for consolidation into a corporate taxonomy.

Also important are new streams of data from social media such as wikis, blogs and instant messaging hosted and moderated by companies. The seemingly random flow of data they provide can be looked at as a dynamic form of intellectual property, stored as an asset and mined for information as needed. While it may seem to be an arbitrary stream of words, in a content management system it is a searchable .TXT or .XML file with metadata associated with it.

In human resources, valuable information about employees, benefits and policies are kept

Correspondence:
Ron Roszkiewicz
Finitiv Corp.
1840 Gateway Drive,
Second Floor, San Mateo,
CA 94402, USA
E-mail: ron.roszkiewicz@
finitiv.com

within a system as secure, controlled data. The same is true in accounting. Customer service maintains their own database as does production. For the most part, all of these data are structured data. It's linked to database records and accepts values that fit into programmed forms of labels. In metadata terms, each is managing a schema of properties and capturing values according to a predetermined set of value types. The data that are captured, when combined, form a valuable set of intellectual property that defines the company.

In recent years, unstructured content, or input data, has been introduced and has effectively broken down the control held by the structured data world. While often created and managed in an orderly fashion, this content was not bound together with either a schema of metadata or properties. Values, when added, were for the most part arbitrary and expedient. As more and more companies adopt databases to store and manage their content, the option of defining a schema and controlled vocabulary of company-specific terminology is emerging as an opportunity and a challenge.

The simple truth is that just as common query languages like Structured Query Language (SQL) emerged for linking and managing databases, so have languages such as Extensible Markup Language (XML) and protocols like web services and custom Application Programming Interfaces (APIs) built into applications. It is possible now to link or actually extract, transform and load (ETL) data from one system to the next. With this in mind, it's possible to think of logically related information being accessible and useful to different departments for different reasons.

The nature of the incoming data streams is also changing. Electronic forms are capturing arbitrary metadata online. Video chats and instant messages are being recorded and archived. Wikis, forums and blogs capture interactive feedback from customers and staff. Videos, recordings are all being tagged. The social media streams feed analytics; the recorded content tagging provides new revenue opportunities.

Taking a holistic, corporate view of metadata and managing it as core intellectual property is the logical next step for overall data consolidation,

and this article explores some of the issues faced in achieving this objective.

DEFINING METADATA

It's always good practice to define terms at the beginning of a document, and in the case of the word 'metadata' it couldn't be more necessary. As any beginner in Latin quickly learns there are certain words, like the Latin 'res', that means *thing* and can be used anywhere that a more descriptive word is not available. It seems that is what has become of the word metadata.

Classic metadata

For decades, metadata has been of critical importance to the systems that run our businesses. These are values that might be captured after they are entered into fields in a database, and used to define parameters in application building or used to trigger an event in a business process. This is the world of 'structured' data. In this case, structured data are a euphemism for a database. This is the classic example and has been in place for decades. The foundation of business is still the database. The ERP, CRM, data analytics and business process management for any supply chain that operates in a business is built on metadata models, or metamodels, and always has been. The tags and value types when dumped from such a system resemble a schema and should be considered as such. The daily feeding and nurturing of these systems is up to the IT department. Oracle or SAP are common brands in this environment.

A simplistic metadata definition

Today, the most commonly used definition favored by non-technical content managers for metadata is 'data about data'. This unfortunately identifies only one facet of what is becoming a more and more complex metadata matrix. Simplifying metadata's value is the result of its use in *unstructured data* systems. Unstructured data are, for lack of a better term, data not associated or stored within a database. This is not entirely true as everything digital is at least loosely part of a directory hierarchy with pointers to it, not unlike records in a database. *Unstructured data* can be moved around, from one system to another, without being tracked by any database. Unstructured assets take their metadata with them. Without some application

awareness of the asset and ability to track it, there can be no interoperability between applications.

THE METADATA CONUNDRUM

Unfortunately, the simplistic definition commonly held by content creators and data consumers, that the value of metadata for classifying and characterizing digital assets is limited to content life cycle upstream processes, is both short-sighted and harmful. The greater value is for improving the overall enterprise architecture so that business processes and the analytics they yield can be controlled and leveraged for compliance and increased overall productivity. It exposes a lack of understanding of metadata's role in structured data systems.

This is the situation we face today. In some companies, basic metadata management is taking place to support, for example, digital rights. At the same time, in the next building, some metadata modeling is taking place to improve human resources data access and new compliance rules. As both ends of the building become joined by a web browser and accessed by users within and without the organization, the differences between both metadata sets and models become increasingly blurred. By this I mean that a staff person, profiled in the HR database, has well-defined roles and permissions in his/her access to the company's intellectual property. This metadata might define their access to the digital rights database, which itself, in turn, is linked to the digital asset management system. This person may be the executive, or from marketing, sales or accounting. The point is that the rights, permissions, product and current state of the product are all communicated through metadata – often using different terminology or different nomenclature, misspellings, abbreviations and all kinds of legacy terms that often duplicate the tagged metadata from one system to another.

The above example is an interdepartmental condition. As data are captured by eForms, blog entries and instant message collaboration tools, and dumped into a repository without a metadata strategy to tag and categorize them, they just becomes a lot of noise and the value that can be extracted from them is lost.

Therefore, it is at this point of confusion that the need for a comprehensive, corporate metadata strategy becomes obvious. Obvious that is to an outsider looking down at the situation from 35000 feet. To those on the front lines it may be all about expedience. To the executive, however, the business drivers that compliance and security require overlap the same drivers that call out for metadata management. Although the need may be greatest in two or three departments, they are only the first to feel the pain of not having a metadata strategy, a pain that will soon be felt by any and all departments relying on an ERP or CRM system. In fact, a unified metadata strategy that encompasses structured and unstructured metadata, as well as classification of business process priorities, will simplify the increasing complexities of interdependent and connected systems. The bottom line is that metadata management must be considered a core component of information architecture.

DEFINING THE SCOPE OF METADATA

Each node in the content creation, capture and consumption life cycle and workflow has its own defined interest in metadata. One node may support the marketing supply chain, another the production supply chain. With all of the collaboration that takes place between departments and inside users and outside consumers, it's easy to overlook the dynamic content that is created through interactivity and social media. It is, however, captured metadata and stored for later retrieval and analysis. The overlap with production, marketing and sales can be obvious or discrete. It's there, of course, and because of this introduces a new area for metadata management. Metadata is pervasive.

STILL MORE SILOS

The focus in recent years has been to raze the silos, seed the clouds and herd packs of unruly cats. Such have been the metaphors for controlling digital data chaos and transforming it from data into information into knowledge. Taking control over the digital data through the use of some form of repository is happening across many industries, particularly in

departments where content is created. The advantages of this type of control are well documented in both hard and soft cost savings. The technology to achieve these advantages is mature, and standards are there for those who need to use them.

However, new metadata silos are emerging and creating a new set of challenges. For example, search engine optimization may be the result of a marketing initiative and contracted outside of IT. This exercise will require a new model with much of that resulting model mirroring models in other parts of the business – it has to. At the same time, there may be a digital rights initiative taking place. Once again this model is mirroring some or all of what is happening in marketing. The risk of these unconnected exercises is a loss of consistency and general transparency.

Trying to provide expedient fixes to metadata problems denies its emerging importance as intellectual property in its own right. By recognizing the corporate investment in all of the disassociated metadata modeling activities and resolving to invest in and develop a core, corporate taxonomy, the complexities that threaten to overwhelm systems and threaten the effectiveness of business processes are managed and risk is reduced.

FROM SOCIAL ENGINEERING TO BEST PRACTICES

Much of the activity that has been taking place in the latest phase of the data management transformation has been in changing corporate cultures to accept data management and workflow systems as reliable, transparent mechanisms – and not headcount reducers. This social engineering phase has passed and now it is time for a new wave of executive sponsorship to power a collaborative effort across IT and all business units to establish this metadata core. It can't happen spontaneously.

It also seems to follow that the best practices that achieved success for those who used them in the planning, implementation and deployment of other ERP or content management systems will work here. All of the same communications requirements, change management procedures and training will be needed to ensure success.

METADATA MANAGEMENT IS A HARD SELL

Selling metadata management is tough. The word metadata itself excites few beyond devout fellow travelers. Plus it takes time and a fair dose of outside professional help to get through the process of organizing it all, especially across departments. It's also rare for a corporation to have on staff a library scientist versed in a broad range of business use cases and digital technologies. The case for doing it must be not only sound business but produce enough heat to warm the heart of the executive who will champion the project against others in the company who cannot see the long term ROI value.

THE ADVANTAGES OF METADATA MANAGEMENT

What follows are a few key areas solved by introducing metadata management technology and practices to an organization:

Productivity – the case for increased productivity and reduction in hard and soft costs to a company is made each day for simply controlling digital assets. This argument can be extended to other departments and applications where well-organized data contribute to insightful analytical reports. Better metadata leads to better reports that in turn result in better decision making, which is not as easy a case to make as the one in which content workers spend 25 per cent of their time searching for digital assets. There are other benefits of course. Having reliable metadata in HR for workers may lead to reliable security for other systems based on the workers' rating or need to know. Data that are up-to-date and connected to from a master index informed by the HR database.

Compliance – the need to control digital assets and all intellectual property in order to retrieve it for custom publishing of documents for investors, digital rights management, Health Insurance Portability and Accountability Act (HIPAA) information control, Sarbanes–Oxley and Securities and Exchange Commission (SEC) regulations is critical to corporate oversight. As with any content management life cycle, all data are eventually archived. This includes SEC- and

IRS-mandated documents. Using metadata for retrieval as part of a corporate strategy is efficient and an appropriate use of the technology. Achieving this through a core taxonomy, with a recorded history of additions and changes, adds to the accessibility of the data.

Scalability – the typical definition of scalability is more users/bigger systems. The expansion of what content is and how quickly it's growing requires that this definition be expanded. As marketers battle store shelf and online wars for the eyeballs and dollars of prospective customers, they are also fine-tuning their message to target prospects more precisely through new channels. Metadata in the form of GPS coordinates, direct feedback from users about their preferences, search results, ballot choices and so on all feed the marketing machine with manipulatable data. Building a core metadata taxonomy for a corporation and using it as the master index for all forms of content simplify the process of filtering the data and bring increased clarity to the results. It enforces objectivity even though it may be multi-faceted and appended with folksonomies.

BUSINESS DRIVERS AND MEASURING SUCCESS

Survey after survey taken online or in person at conferences reveals that the majority of users with content management systems are unhappy or find their system is under-utilized. Some realize why. It is often blamed on the initial planning and workflow development. Shortcuts were taken and money was sliced from these parts of the budget. This is code, I believe, for unsatisfactory metadata management planning and implementation. Of course, reasons also include poorly designed user interfaces and missing or unfinished functions. The result often is lost jobs for those who championed the solution and a general wariness about new systems in general.

It's rare that a company scraps the installed system and brings in a replacement. Instead, the strategy is to proceed incrementally so that problems can be identified sooner and before things have gone too far. Although incremental phasing in of new applications and systems is always wise, it doesn't replace a clear vision and strategy for workflow and metadata

management. It's always critical to plan how the upstream content creation and data streams will be handled from a metadata and a functional point of view. The change may be a tweak here and there, but all options should be considered, at least to determine whether the platform being adopted in the new system can support these options in the future.

Metadata management is so tightly linked to information architecture that they feed each other in determining what components are required to achieve the customer's desired objectives. Priorities are shifted depending on what the business plan calls for and the best ways to achieve it. There is no generic model, no universal taxonomy and no one-size-fits-all set of business rules. Not that some vendors don't attempt to sell on the notion that there is.

Not so long ago, most companies, as far as content was concerned, were print-centric. The supply chain from content creation to print distribution was straightforward. There were some technical issues that had to be resolved over time, but the nomenclature was pretty mature and static. Today, we've moved to a multichannel distribution method where print still exists, is still essentially static, but the dynamic web-centricity of the enterprise enforces its own needs. Adapting to this multichannel world, with staff and remote contributors, interdepartmental collaboration, behind and through the firewall interactivity and data capture and all of the additional media forms and approaches without a sound metadata strategy will make the systems that users are currently unhappy with into even less useful under the weight of the incoming data.

Surveys conducted by Forrester and others lead us to believe that companies are recognizing the need to develop a metadata strategy. It is still emanating out of the IT database side and includes data repositories, enterprise architecture and service-oriented architecture. This seems to recognize that the metadata landscape is broader than what was formerly accepted and that a corporate solution must match this broadness to be effective.

HAVING A PLAN

Planning for metadata management on a broader scale than at the repository level takes into account the broader implications of both the

source and the intended use of the data. We expect that there will be new sources and new uses in the future, and therefore planning for it means taking a more holistic approach. To do so, the following points must be taken into account:

Who are the people and what are the processes that will capture the data? How are they filtered now and how should they be filtered in the future. Do these data need to be transformed? How data are handled at the source can determine their effectiveness when they are needed later on.

Where do these data go? Do data repositories already exist to store these data? What form are they in and will there be a need to normalize them in some fashion. This is of course particularly critical in blending data from CRM and ERP-structured systems with a system storing unstructured data.

Map out the content life cycle for each of the metadata streams. Controlling the data implies understanding how they flow and where they may converge with other data streams.

With the various life cycles charted, what rules or policies are in place (or not in place) to govern the different data streams. There will undoubtedly be rules that are in place that are absolutely necessary to specific business processes and compliance and other streams in dire need of policies to be controlled. This is the time to identify them.

Once the data are identified, the stakeholders must evaluate and validate them in order for the process to go forward. This is the time to consider how metadata can cover gaps in any of the processes under examination. This evaluation is a good time to rationalize the current state and define requirements for going forward.

The main objective of identifying the data streams, recognizing how they support business processes, is to develop or re-use a revised or new metadata, models or schema and create a holistic information architecture.

The metadata discovery charted above will typically take place during a similar discovery into the current state of the workflow for each of the supply chains under investigation. There's a synergy between the two as one informs what is needed in the other and in the best of situations leads to a clearly defined approach to data governance and a solid foundation for

capturing and using data in the near and long term as data capture and utilization evolve.

NORMALIZING ENTERPRISE METADATA

It's perfectly normal to find a variety of database systems in a company, regardless of size. Even Fortune 50 companies will have FileMaker databases, as will start-ups. The reason is expedience as much as cost. Solving a data management problem with a database built over the weekend makes a lot of sense. FileMaker databases and more recently MySQL and PHP web-centric databases seem to grow organically in many companies. What, in effect, has happened is that a number of smaller meta-models have been developed that are often not related to other meta-models in the organization. Very often, there will be the same or similar tags or properties in two or more databases (meta-models). The idea is to recognize these, normalize them and construct a meta-meta-model from the result.

Once the decision is made to consolidate records and digital assets into one repository, expedience gives way to control, productivity and the notion of building a system for future growth. The following project management steps are helpful in building this system:

Determine all of the systems that have metadata. This may include ERP, CRM, installed digital asset repositories, human resources, social media streams, accounting and so on. Assemble a team of individuals from each of the relevant departments to act as stakeholders for decisions on the core corporate taxonomy.

Understand the objectives of the core taxonomy. For this phase, it may be necessary to enlist the help of an outside professional with an objective perspective of the best ways to achieve the company's goals and how to manage the metadata to achieve these goals.

Consider as part of the process whether any existing standards are available to act as the core for the new corporate taxonomy. Certain schema have been developed just for this purpose, for example, IPTC for wire service photographs, Dublin Core for libraries, ONIX for book publishers, PRISM for magazine publishers and so on. If none exist specifically for your domain, pick one that comes close as

the model. If one becomes available in the future, the data will be mapped from one schema to the new one as part of the migration process.

Gather together with whatever means possible the printout of the properties and values or value types for each of the affected databases. In this case, a property is a label on a field (as in Architecture) and the value may be pulled from a controlled vocabulary (Early American, for example) or data input from some source.

Reconcile the industry standard with the company's legacy metadata wherever possible to create the core set of root set of taxonomy nodes and the controlled vocabulary of company metadata.

Apply this newly developed taxonomy set to assets, records, reports and whatever other interactions that take place between linked systems. This step may require using a new tool for managing the corporate metadata set or some tweaks to the databases involved to bring them consistent with the new schema.

The value of this type of consolidated and rationalized master metadata set is simplicity. With data coming and going in the form of digital and printed output, either as reports, marketing collateral or catalog cop, and having what amounts to a standard corporate nomenclature helps to maintain the company story and reinforces the brand identity. When digital data are under control and the corporate metadata are organized and managed, the process of building on the foundation laid makes it easier to take advantage of new revenue opportunities in a nimble manner while ensuring the identity of the brand is maintained.

MEANS TO AN END

Every company is made up of a number of intersecting workflows. Establishing a taxonomy that brings uniformity to corporate metadata

simplifies data interchange, rules development and control. Although it may seem like a good idea to consolidate into one taxonomy and a single schema, it's probably impossible to achieve. What is possible is to normalize metadata that overlaps systems. What this may mean is team member information from HR normalized with the permissions for access to digital assets in the content management system. Manufacturing and production also have links to the status of projects in the workflow-tracking system. Obviously, having project numbers, output type and materials required is valuable information to both systems.

One initiative that applies to this concept is the Job Definition Format (JDF) that exists for graphic arts and prepress. This initiative began in 1999 as a collaborative effort of Adobe, Man Roland, Heidelberg and AGFA. Today, it is managed by CIP4 and has many industry members. This XML-based schema includes to the best of the workgroup's ability all of the materials and processes that may be specified for a print job. The idea is that the information for each node of the workflow is added by the application according to the JDF and is then readable by the next node in the process. This *persistent* metadata is embedded in jobs or associated with them.

But taking advantage of a consolidated corporate metadata taxonomy does not require a JDF type of initiative. It can be built on an existing standard schema recognized by a particular domain and appended with properties and values relevant to the company creating it. It's the equivalent of having a corporate stylebook for written communications. A stylebook simplifies communications and maintains a consistent story and writing quality standard, the result of which is increased brand support and a valuable addition to the corporate intellectual property.