Mining Metadata Gold with Semantics

Experts are turning to new models to better define metadata and to store it more flexibly and accurately

By Matt Turner, Chief Technology Officer, Media and Publishing, MarkLogic Corporation

Abstract: The value of metadata continues to grow in today's digital marketplaces and maximizing this value impacts nearly every part of entertainment and media organizations. Metadata is also becoming more complex, covering not just the asset information but now including product and title data, contributor and topics, and details on usage and consumption. With the rise of new technologies tailored to handle the intricacy of today's information, new approaches are available to manage metadata. Enterprise NoSQL databases and Semantic technology can make a wider set of data available to users and reduce the complexity and cost of delivering this valuable resource.

With the rise in the value of metadata and its direct impact on the new digital business, metadata is becoming a critical asset for product creation and distribution. This is driving the need for more accurate and flexible metadata. To address this need, organizations need to look beyond the traditional tools and technologies for managing metadata and embrace new approaches, including NoSQL databases and semantic technologies to create, manage and deliver more accurate and complete metadata.

Traditional metadata

Traditional metadata management processes define, up-front, the metadata required for the organization, using tools that range from Excel spreadsheets to taxonomy and ontology management tools. These tools all define a fixed set of attributes for the metadata, typically in the rows and columns of a relational database.

This process means creating a row for every possible type of metadata including dates, contributors, asset information and categories. This can become rapidly complex due to the need to define in advance all the possible permutations of the metadata. Problems may include defining a single model across multiple types of assets, selecting attributes that cover specific uses and adopting new information as data and purposes change.

In addition to the items of metadata, expressed as the columns in the model, the values of the columns also need to be managed. Many of the items of metadata are managed in lists called controlled vocabularies or look-up tables. Taxonomy tools are used to create and manage these lists and the selection of these items gives the metadata context for a given purpose. The typical taxonomy approach is to manage categories in a hierarchy with the distinct levels all having distinct sub-levels of choices.

Categories and taxonomies are also tailored to specific business purposes. Internal processes have different needs than external distribution processes, for example. The metadata may have either many similar categories or generic categories that do not have the specificity required for each distinct business process. This can cause additional data processing or clean up as different systems use the metadata.

These approaches also require that all the items of metadata including categories be explicitly associated with the metadata record. If an asset is associated with several categories, all of those categories must be associated with the asset directly in the metadata record. This can lead to metadata models that have hundreds of attributes to explicitly capture all of

the complexity of the information in the single metadata record.

To address these issues, metadata experts are turning to new models of data to both better define metadata and more flexibly and accurately store it.

NoSQL databases

Hierarchical, complex and sparse metadata are not well suited for relational databases and their strict rows and columns. Therefore many organizations are adopting enterprise NoSQL database technology that allows metadata to be defined and stored with a flexible schema without giving up data integrity or security. Instead of defining up-front every item that must go into the metadata record, the NoSQL approach allows for each record of metadata to store only (and all!) the attributes that make sense for that particular asset.

Schema flexibility addresses some of the issues found with the traditional approach:

- Varying types of assets can have the different types of metadata they need with out having to also carry or fit into a model not suited for that asset.
- A wider set of data can be stored including data suited for a specific purpose for one kind of asset and not another because the model can allow for the flexible storage of additional information without having to pre-define that information.
- New information can be added as data sources or metadata models change without having to rebuild the system.

With enterprise NoSQL systems like MarkLogic, this flexibility can be used without any loss in data integrity or reduction in security. Enterprise NoSQL allows for processes to ensure the data is accurate and has the core attributes needed for the system and that the information can only be accessed with the correct permissions and rights.

Semantics

To address the complexity of managing categories and taxonomies and to record more accurately the intricate relationships for an asset such as categories, genres and contributors, organizations are also embracing Semantic Web technology.

Semantic Web technology is a new approach to defining and managing relationship data. Using a simple format called a triple, Semantic Web technology indicates facts, concepts and object and how they are related.

Example of a Semantic Triple

A simple example about a contributor could be information about his location starting with a fact: "John lives in London". Combined with the information that "London is in England," we can now also state "John lives in England".

The data modeled in this fashion can be used with other data about England to derive new information about John – for instance that he lives in a monarchy.

The model can also be extended with more information about John. We can state that John is an actor and, if John is the lead character for a show, we can link John to the details of that show, all actors of that show, when that show is in production. We now have new levels of insight into his activity for instance when he was working in a given year or what other actors he has worked with.

The creation of this type of data requires data management structures called ontologies. Ontologies are similar to taxonomies that describe look-up lists except that the ontologies describe the collection of triples and are not only hierarchical. In our example, the ontology would describe the types of data we are recording (people and places) and the types of relationships they can have (lives in and is in).

Ontologies are also expressed as triples and, along with the types data and relationships, they can also define the values of that data. For instance we can restrict place to a list of known locations including London and England.

This approach is particularly valuable for entertainment as it allows for the accurate creation of models – or ontologies – that are tailored to the specific uses of the metadata within an organization. It can also record and manage the details of that organization's specific products. If our simple example explains where John can live, a complex real-world example for an entertainment organization might model all the possible characteristics a famous fictional character might possess, map those attributes over the many different times that character has appeared in films or shows and provide different metadata values tailored for production and distribution.

Semantic Web and NoSQL together

New technology capabilities are bringing the schema flexibility of a NoSQL database together with power of semantic technology. MarkLogic's Enterprise NoSQL has seamlessly combined these two data models into a single component, removing the final hurdle to fully leveraging the value of the semantic models with the active, operational metadata.

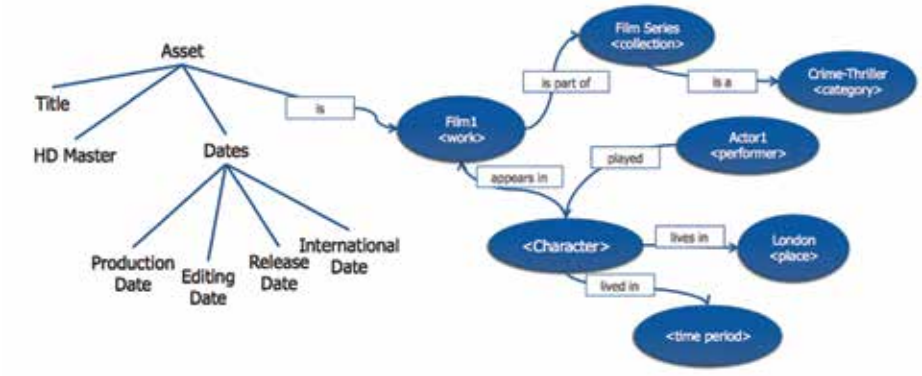
This combination addresses some of the major issues with taxonomy management:

- The relationships do not need to be only hierarchical but can express many different associations and be linked together to correctly model complex metadata attributes.
- Using several ontologies, metadata can now record relationships (instead of categories) for the many different purposes of the asset.
- Because the ontologies are independent of the individual metadata records, changes to the data and additions of new types of data can be done at any time without having to reprocess or remodel the metadata.

Using semantic triples, the metadata record does not need to store attributes for every single possible value related to the record. With only a few defined semantic relationships, an asset can be linked to a much broader set of data -- making metadata management much more efficient. For example, a

Matt Turner works with customers and prospects to create leading edge information and digital content applications with MarkLogic's Enterprise NoSQL database. Matt works closely with MarkLogic's customers McGraw-Hill, Warner Bros., Conde Nast and LexisNexis. Before joining MarkLogic, Matt was at Sony Music and PC World.

Melding Data Models



NoSQL Metadata record (l.) with semantic triples (r.) associated to an ontology.

metadata record using our fictional character, the asset only needs to be associated with the specific title where the character appears. All the other data about the character, for instance where and when they lived, what genres and categories they fit into, would all be automatically available to that record with that one item of metadata.

With this model, the metadata can now provide accurate and complete information to a much wider range of audiences, the processes for maintaining the metadata are much more efficient and the metadata will be a much more valuable asset to the organization.

Dynamic delivery

In preparation for the hosting of the London Olympics in 2012, BBC Sport adopted a Semantic Web approach to not only the management of the metadata of the assets it would be delivering on its web pages, but also the delivery of the experience to the end users.

The BBC Sport team developed sports ontologies that contained the categories of events and sports and the key players and organizations in each sport. Individual articles and information about the teams and athletes were tagged with just the links to the ontology, greatly reducing the amount of editorial work and streamlining the content creation process. The live video feeds, sport scores and schedules were also tagged to the ontology with automated feeds updating the information.

This data was used to create the public website (bbc.sport.co.uk) that, for the Olympics alone, featured over 10,000 pages. All of this content was dynamically generated from the ontologies and the articles, video feeds, scores and schedules and was constantly updated as results and scores were tallied. The resulting user experience broke many records in the UK with, at times, more people watching and interacting with the events via the site than over traditional broadcast.

Conclusion

As metadata becomes even more important to the delivery of digital products for media and entertainment organizations, leveraging new approaches to address the many challenges with traditional tools is becoming critically important.

Semantic Web technology combined with Enterprise NoSQL allows metadata to address more information, including data for multiple purposes with more flexibility in the management of the data while maintaining the data integrity and security. ■