

Metadata as a Service (Are We There Yet?)

By Ron Roszkiewicz

Now that we've moved from workgroup silo to enterprise silo — thanks to DAM systems and metadata — are we ready for the pervasive Internet cloud and global search or have we just created a new set of problems? The recent Henry Stewart DAM conference provided some answers.

Content creators and publishers are struggling with choices for the best way to store, retrieve and protect data. Many organizations have made a first and second pass at archiving and retrieval. They've made progress; money is being saved every day on searches and reuse. As long as searches are carried out within the closed loop of the company schema all is well. Unfortunately, most companies need more. They need dynamic taxonomies, adaptable controlled vocabularies, and the ability to distribute assets with these taxonomies without fear of mischaracterization and inadvertent metadata changes. In addition, they need the ability to maintain some digital integrity over their metadata and treat it as the important intellectual property that it is. This concept recognizes that for many companies, metadata is more than a search enabler, it is the value added to a workflow that streamlines the product search for a customer and acts as the basis for a successful sales process. Recognizing this potential raises metadata above the hit-and-run copyright and authorship tags we use now. It raises it to a level where the assets it supports in turn support the global knowledgebase that the cloud makes possible. Metadata as a Service (MaaS) attempts to define this broader workflow relationship and strengthen the digital integrity of the metadata in the assets through synchronization and validation while providing tools to manage the taxonomy and controlled vocabularies as needed, when needed.

One of the best ways to gauge the health of an industry is by attending a user-driven conference. The November Henry Stewart DAM conference in Los Angeles was a mixture of user group meeting and vendor networking event. As always it provided a snapshot of what's happening in the industry. A similar snapshot taken a few years ago would show DAM/CMS systems as the answer to information silos and scalable, open interoperable solutions. Metadata was emerging as a flexible but powerful answer to data search on unstructured assets just as it had always been for structured data. Adobe's Extensible Metadata Platform (XMP) was also emerging as the standard for supporting metadata across applications and platforms. Developments seemed to be plodding along. With that in mind, it was interesting to see the preponderance of sessions still

devoted to metadata and the technology that supports it. Vendors continue to show applications that use metadata and focus on the value proposition for using it. Users describe workflows created to support their publishing environment. In the context of the vast Internet cloud, global searching, Software as a Service (SaaS) and other collaborative relationships that have made the notion of a firewall a virtual artifact, we seem to have moved from the workgroup as silo to the organization as silo without acknowledging the irony. This is not to say that all systems should adhere to one standard, one taxonomy or even one set of best practices.

Metadata is more than a search enabler. It is the value added to the workflow

In fact, the proposition is just the opposite. In companies where there exist myriad repositories it is impossible to control and validate metadata between systems without complicated connecting software or a new approach to metadata in the content lifecycle. The fact is that without controls over the use of metadata, some approaches to managing and validating it will never achieve in the open era of computing the importance it had during the long run of the relational database.

Déjà vu again

This workflow argument is reminiscent of the discussions of color management, trapping and PDF-based workflows that once dominated Seybold conferences. The problem then and now is that there are few applications or engines that support global metadata control. Metadata, for example, is still subsumed within DAM systems and limited mostly to the needs of that system. That is not to say that some systems don't provide tools to customize schema within the application. They do, but the market requirement exceeds what has been provided.

The problem is also a lack of consistent metadata support from application to application, and tools to manage metadata from end-to-end. As more and more users implement some form of metadata approach, they become increasingly aware of the difficulty maintaining metadata with files as they are distributed in the field. It's equally important to maintain metadata

upstream and downstream throughout the content creation workflow, into prepress and beyond. This holds true for web and print. There must be digital integrity for embedded metadata throughout the process. We faced similar challenges in the past with PostScript, color management and digital file formats.

Unfortunately, the assets we need to manage are multiplying exponentially. We are collecting so much digital “stuff” that a large proportion of it becomes irrelevant the moment it is captured. There is no good way to find it again when it is needed. Editing the incoming data often helps. Editing and embedding metadata is even better. Best of all is treating metadata as *intellectual property* and controlling and managing it as such throughout the lifecycle of the asset it inhabits. This is the ultimate objective for any effective workflow and the key benefit of MaaS.

Can Chaos Have a Precedent?

At one time publishing was ruled by proprietary file formats. Interoperability was a chore, and the idea of assembling a hybrid workflow was daunting and often impossible. Out of that chaos emerged PostScript. Although it did some things well, like describing text and illustrations graphically without regard for output device, it did not provide many improvements over the predictable output provided by proprietary systems. It also had to go through many years of maturation before it was suitable for use as a commercial quality format capable of high-resolution color, illustrations and fonts. It was also too verbose, which degraded system performance. The answer to these issues, the Page Definition Format (PDF), did not immediately provide the final link in the workflow chain, but had to go through years of incremental improvements before it satisfied the requirements of prepress operators. By the time applications and operators were able to create readable and complete PDF files for prepress operators, there was only one major issue to resolve before PDF was accepted as an end-to-end solution. It had to certify in some way that the data used to form the file upstream was the same being read downstream. Data integrity was necessary to allow for that last bit of workflow streamlining, file checking and sign-off just before output and all of the proof sheets and film that entails. To achieve this final bit required a tweaking of the format into variations of PDF/X. Today there are successful end-to-end PDF workflows that combined with Job Definition Format (JDF) schema provide a glimpse into true automated print and web-to-print workflows.

Under the hood is no place for the user

The hours wasted tracking down and working around errors in PostScript is legendary. Some errors were caused by the applications, some by the user. None should have made it into the PostScript code. By the time PDF arrived, users were still too gun-shy from their experiences with PostScript to rejoice. In fact the

same problems were possible with PDF even though the format was cleaner and more compact. Garbage in equals garbage out. Getting files downstream intact and made up of predictable parts was solved by a combination of workflow best practices, industry supported standards (PDF/X) and a workflow process that recognizes host and client interdependencies. The same will be true for metadata and the XMP specification. It is no mystery that metadata can get trashed as it is making its way from desktop to desktop or desktop to server.

The problem is a lack of consistent metadata support from application to application, and tools to manage metadata from end-to-end.

Like PDF, users must manage a master template at the server end and validate the metadata in assets as they come and go. This can be done in cooperation with a DAM server or simply by identifying a straightforward target directory to serve as the repository. Once again this embedding, reading, writing and managing data is done by server-based host based on rules set up by the user, not manually by users.

XMP is an important technology but, like PostScript and PDF, it's better left under the hood. It's the application of this technology that is important. For that we turn to the Creative Suite and the File Info menu and templates first for applying pre-defined metadata properties and values to a file. Unfortunately, XMP is not completely finished. Adobe applications do not handle XMP metadata consistently from application to application. Some applications strip out metadata and some metadata actually conflicts internally with that which is contained in other templates. In this writer's opinion, this is wrong and the user should not have to deal with it.

Dealing with metadata that is somewhat unpredictable as it is stored by the application in the file presents a problem to the user who wants to build a workflow based on metadata. For Digital Asset Management system vendors the problem is moot. They define and ingest metadata for their own inter-DAM purposes with little attention paid to the outside workflow. For content creation workflows the problem is obvious. But the same issue is exponentially larger for consumer and Web based workflows. Chaos now rules and is leading to the issue of data irrelevance.

In point of fact, users — both consumer and professional — want much more metadata than they have access to now. They want metadata to bring order to video. They want metadata to complement image pattern matching for searching and sorting pictures of the kids or political figures. Professionals also want rights management terms to be integrated into the metadata package. System managers want metadata to trigger events along the workflow route with precision and predictability. How is any of this possible without

standards, centralized management and the concept of digital integrity applied to metadata?

MaaS Appeal

Metadata as a Service (MaaS) means metadata that is hosted, managed and validated at a central source and syndicated in the form of templates and panels to users to embed metadata into their assets. This is possible now because metadata technology is stable and mature enough. Many applications use this technology and more will in the future. We can identify best practices and build tools that apply technology to workflow thanks to the many installations of DAM in existence. Just as many users mis-identify DAM solutions as workflow, they also make the same mistake with XMP metadata by identifying it as an application. From the start, XMP was defined as a *platform* made up of technologies on which *applications* can be built. These applications use MaaS in the same client/server relationship as in a SaaS relationship.

Using MaaS overcomes the problem of many vendors using metadata subsets for their own purposes within their applications. As a part of a DAM application, there is limited functionality to customize or in any way dynamically change the nature of the captured metadata. Therefore, what is effectively happening is the open XMP metadata is subsumed in the application and treated like other proprietary metadata in the application. Clearly, this is not what was intended for the technology.

Mapping of metadata schema properties and values between interoperating applications is also taking place. This means when an occurrence of a metadata value is found in one user's DAM, it is mapped to a similar or equal value in the receiver's repository. This simplistic approach does not support the rich faceted nature of taxonomies or topics, synonyms and thesauri. It just facilitates the interchange.

In day-to-day database management, fields may be changed, deleted or added. Assets may or may not inherit these changes and the validity of the embedded metadata, if there is any, is typically not challenged. Not validating the metadata that is used to characterize the files is similar to not validating the PDF that carries the page elements. Prior to preflight software, files often slipped through the prepress process with RGB images, incorrect fonts and missing illustrations. Metadata that identifies the owner of the asset, the usage rights and allows internal and external users to successfully retrieve it is no less valuable. Without upstream or downstream validation, this metadata can become altered inadvertently and defeat the whole purpose for which it is used.

Users do not want to fill out metadata forms and for the most part, they shouldn't have to. Most information about a project is known and can be input by an administrator and embedded automatically as a server function. Batch processing project tags, digital rights and other workflow related metadata is better left to

a person assigned the task like a Cybrarian, familiar with the corporate taxonomy and controlled vocabulary. Most metadata embedding should be a server task and not a desktop task, even in small workgroups. XMP technology provides for this by supporting templates made up of taxonomy and values. All that is required is an application to execute the event or a script hand made for the task.

Finally, managing the digital integrity of metadata as intellectual property requires an application of its own. This application must complement the metadata handling functionality of DAM systems and repositories. It must allow for dynamic metadata maintenance from day-to-day as products change or subjects of the

Just as many users mis-identify DAM solutions as workflow, they also make the same mistake with XMP metadata by identifying it as an application.

daily news emerge. As a server-based technology it must provide for other search types to be layered on top of it. For example, it is possible to embed digital rights management URLs from the Plus Coalition into an XMP schema field. This means that two servers, the XMP metadata server and the Plus Coalition DRM server will be interacting interdependently and the status of the digital rights can be displayed in the standard XMP search view. Any changes to the URL can be identified during the validation and the metadata can be edited at that time.

New search types can be added to the applications such as face and pattern recognition software. This software can work in partnership with other XMP file characterization metadata to expand further the notion of multi-faceted searches. In an environment where an assortment of different DAM systems is in use, a centralized complementary application conducting searches based on a set of master criteria would be a powerful one indeed.

Conclusions

It's time that the subject of metadata shifts from technologies to applications. The killer application for metadata is not XMP but some form of search and retrieval application that uses XMP or a similar data characterization approach. Once the discussion is about the application of metadata, companies will begin identifying the functions, and features of search. The notion that a full-fledged metadata management tool will be built into every DAM system in the future is far fetched. As developments stand today, we are just at the beginning of metadata deployment. The excitement is over the fact that there is a technology that provides a way to stuff tags into files so they can be used to enable searching. On a planet where every business runs on metadata-based queries within a database, this is not

revolutionary. Providing every company with a custom schema development tool built into their database and have it automatically support schema built by others in their databases is also far fetched.

One answer may be the way desktop applications solved the color management issue. Adobe, with its broad Creative Suite of applications, synchronized the color management in these applications through the Adobe Color Engine (ACE). This engine uses profiles (schema) for different output requirements (vertical applications) and runs in the background. Setup is easy. Once the ACE is invoked, all that is left to do is choose a profile based on the type of substrate (surface) or screen the image will be portrayed on. Images processed

in Photoshop have this profile embedded in the file and it is recognized by the other Creative Suite applications.

It's an accepted fact that every company needs custom metadata schema. Most can get along with a core set of a standard schema, but all will want to add to this core set and have the ability to make changes when needed. As in-house web site developers seek to automate their web-to-print mechanisms, they are looking for ways to leverage the power of metadata as a tool to streamline this processing. Once users are able to make the clear distinction between technology and application they will demand more workflow related functionality out of their asset management tools and out of new tools on the horizon. **TSR**

Scribus – *Continued from page 10*

active dialog box messages when the PDF is opened, or including page navigation and controls.

Summary

While the slow emergence of Scribus has largely gone unnoticed outside of open source communities, the product has quietly evolved into a viable, professional desktop publishing system. Its main value proposition is its broad cross-platform support, which is unmatched by other page layout software. Together with its support for 25 languages, impressive PDF authoring capabilities, built-in image editing support, Python scripting support and open source format makes the product attractive for several publishing environments and applications.

Scribus' main limitation is governed by product development. Built by a team of enthusiastic hobbyists, the product roadmap relies on developers donating their free time, outside of their day jobs, family lives and everyday responsibilities. Furthermore, Scribus lacks the luster of highly polished commercial software. Instead, its interface is raw and bears several marks of unfinished and deprecated features.

While the limitations of the OSS development model will continue to hinder Scribus from making any short-term market impact, there is future potential for growth. Following Wikipedia's funding model, it is conceivable for the Scribus team to secure donations and funding from users who share a vested interest in product development. In addition, it's conceivable that an organization could adopt an open source business model, similar to that of Red

Hat who sell subscriptions for the support, training, and integration services of their Linux-based open source operating system.

The products' free licensing model presents an attractive proposition to many in the midst of an economic crisis, where organizations are looking for quick ways to cut costs and software budgets are an inevitable victim. Software giants are already bearing the scars of a looming recession, including Adobe, who recently cited a weaker-than-expected demand for the newly released Creative Suite 4 and has taken steps to reduce its headcount by approximately 600 full-time positions.

Despite the attractive licensing model and its admirable features, Scribus may never be able to crack the mainstream publishing market. Existing commercial software products are heavily ingrained in publishing workflows. Convincing users to make the switch is no easy task. While Adobe has managed to convince many long-term QuarkXPress users to make the switch to InDesign, it hasn't been without considerable effort. The real cost of desktop publishing doesn't lie in the software alone or a robust feature set, but in the accumulated, associated cost of adopting and learning a new software tool, then changing and adapting workflows to suit it.

For many, Scribus may not signal an immediate mass-exodus in desktop publishing. However, as publishers struggle to survive and adapt, open source software will become an attractive model that no one can afford to ignore. **TSR**

Eliot Harper is Director of **Eliot & Company**, an information publishing firm based in Sydney, Australia.

Gilbane Boston – *Continued from page 3*

crisis would accelerate or decelerate the adoption of community applications and other Web 2.0 technologies, the panel and audience agreed that tight budgets cause companies to look for value and that because of their attractive pricing Web 2.0 applications will be easier to justify for cash strapped companies. Lundberg concluded that there will be continued pressure for

companies to innovate and grow despite less capital being available and that avoiding risk coupled with security, reliability, and accountability will drive many technology adoption decisions.

While the pervasive attitude could best be termed realistic, there were many productive exchanges of best practices and tactics for innovating and improving productivity in a difficult environment. **TSR**