

# Metadata in Context

BY RON ROSZKIEWICZ

**If content is king, metadata is the crown jewels.**

Long recognized as a key component in indexing information in the world of print media, metadata is emerging from the shadows as a key ingredient in initiatives to bring order to the Internet. If these initiatives are successful, the effects will be far-reaching and will dwarf our current ability to “google” our way to information.

Technologies have to be in place for us to achieve this global, metadata-enabled world. An open and scalable system architecture and metadata language standards have to be in place to satisfy evolving and expanding requirements. Applications to read and write metadata must exist to prepare the content that feeds the medium.

In recent years, activity defining the rules governing how the online medium is accessed and represented has focused on its structure. One cornerstone of this structure is search technology. Standards such as machine-readable USMARC used by libraries ([www.oclc.org](http://www.oclc.org)), the ANSI Z39.50 search standard ([www.loc.gov/z3950/agency/](http://www.loc.gov/z3950/agency/)), and other schemes being developed at universities and in some cases funded by the Department of Defense’s Advanced Research Projects Administration (ARPA) are being developed to determine how search engines will work. Regardless of which approach or compromise prevails, the final architectural scheme will be required to recognize metadata approaches currently being used.

The commercial incentive for these initiatives is to reach a day when metadata-based search engines allow us to access and filter authoritative information as a chargeable commodity where usage rights are expressed and persistent because they are embedded in the file, a commodity that will be known by scores of brand names in addition to the unassuming “smart document” and “rich media” it is known by today.

**Metadata defined: data about data.** Metadata is data about data. It can describe the author, creation date, file type and version of a file. The basic concept is simple: define a file with words or phrases, tag it by embedding or associating the definition with the file, and use software tools to read, edit and act upon this information. This simple definition is still true but a bit out of date. As collections of digital data become large and widespread, relevant, precise, contextual metadata

is even more critical. With all this focus on expanding the definition of metadata, it’s easy to understand the current effort to codify semantics and search technology standards.

**Metadata redefined.** Like most concepts and technologies that attempt to solve communications problems in a truly democratic way, the first step is often a disruptive technology or a bold initiative to establish an open standard or architecture. In the case of metadata, the bold initiative is from the W3C committee and its work on the Semantic Web.

In “The Semantic Web” (*Scientific American*, May 2001), authors Tim Berners-Lee, James Hendler and Ora Lassila write, “The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise and community boundaries.”

The standards upon which the semantic web are being built are the Resource Definition Format (RDF), Extensible Metadata Language (XML) and Uniform Resource Identifiers (URI). RDF is a formal specification of how to represent objects and concepts and the relationships to hold them together. XML, a subset of Standard Generalized Markup Language (SGML), is the syntax for exchanging data. URIs are short strings that look very much like URLs and identify resources on the Web. Resources can be images, downloadable files or schema, or collections of metadata properties and values.

RDF, XML and URIs are being stress-tested every day on the Web. As open and extensible standards, they seem up to the task. It’s no wonder that Adobe chose all three as the basis for its own open standard, the Extensible Metadata Platform (XMP) [www.adobe.com/products/xmp/main.html](http://www.adobe.com/products/xmp/main.html). Adobe applied these standards as a core technology to content creation applications. This approach recognizes the value of rich media and a distributed database. It also begins the system integration discussion between standards committees, application developers and digital asset management system developers.

Currently, most metadata is embedded in files on a system-by-system, application-to-application basis. For example, the data might be embedded in the file header and tagged so it can be retrieved by the host application or system. A proprietary or standard

## Metadata Moves from Analog to Digital

**P**inning down the first use of metadata of any sort is probably not possible. Libraries use metadata, book indices are a form of metadata, and most packaging includes some form of metadata.

Modern standards for defining metadata properties and values exist and in some cases have been tested in the context of real-time production. Two examples come quickly to mind. One is the International Press Telecommunications Council (IPTC) standard for content and metadata and the other is Dublin Core.

The first formal IPTC news exchange standard, IPTC 7901, had a line with various metadata elements for news management and a line for descriptive metadata, the equivalent of a keyword property in 1979. Top line in this case does not refer to a digital file header. It's the top line on the paper printout from a teletype machine that corresponded to the image file sent by wire. The next standard, the Information Interchange Model (IIM) of 1990, included an extended set of metadata that is the basis for the IPTC Headers introduced by Adobe for image files in Photoshop 4 and later. IPTC and Adobe recently announced a formal collaboration to bring the core IPTC metadata set to

Photoshop as an XMP-compliant schema.

The Dublin Core standard is developed and maintained by the Dublin Core Metadata Initiative for the purpose of promoting the widespread adoption of interoperable metadata standards and vocabularies. Begun in 1995, Dublin Core (the original workshop for the initiative took place at Dublin, Ohio, hence the name) is closely linked to Web-based discovery systems, and because of this broad application is intended to apply to any industry. This means that the intent is to use the same methodology with appropriate domain-specific dictionaries. Dublin Core is used by most metadata applications as a de facto standard.

With IPTC, Dublin Core and others working to define their own schema, vocabularies and definitions, there seems to be a duplication of effort. But they are not competing to develop proprietary but openly available standards. They are maintaining the right to control domain-specific attributes. Domain-specific means that insurance companies have different terminology than healthcare or advertising. Since this terminology is reflected in schema property names, working groups are protective of their work.

— Ron Roszkiewicz

**TSR**

schema may be used and interoperability is not intended. This proprietary approach is difficult to maintain and support in an environment where other workflow variables are changing. It is generally not a problem in a static environment. Adopting standards such as those mentioned above has obvious advantages for developers and users.

### Metadata at the OS Level

Apple recently announced a core technology called the “metadata search engine” [www.apple.com/macosx/tiger/spotlighttech.html](http://www.apple.com/macosx/tiger/spotlighttech.html) as part of its operating system version 10.4, code-named Tiger, due out next year. Besides raising the level of visibility for metadata, it will add value to data enriched with metadata by providing another access point for the user. For the most part, Apple requires no effort on the part of the user to take advantage of this technology. The operating system will index file-related metadata and content in the background. This approach is useful for individual users but not much help in a production or automated environment. Using metadata in a professional or production environment requires active participation to make the metadata that

is embedded relevant to digital assets and workflow.

Since this is a core technology supported by a development kit, developers will be able to add filters to recognize metadata, including XMP. Having an XMP filter available will mean that photographers and graphic artists will be able to use this technology to conduct metadata-based searches using their hard disk as a freeform repository. This new functionality does not result in a digital asset management system (DAM) because there is no built-in functionality to manage and synchronize files, not to mention the ability to edit embedded metadata or view collections of images and documents.

Consumers face organizational problems similar to those faced by professional photographers and enterprises. Families amass hundreds and thousands of digital images of everyday life. Schemes are arising to help transmit and view images on a computer or television screen, but not to solve the problem of organizing and retrieving images based on metadata. Popular applications such as Apple's iPhoto, Adobe Album and Microsoft Picture-It do not provide a way to add metadata richness to images and they don't provide an application program interface (API) for third-party developers to add this functionality. Apple's metadata search engine and a similar development planned for Microsoft's Longhorn operating system could start the ball rolling.

### Metadata in Applications

Internet browsers represent the largest installed base of metadata users today, although few of them realize it.

iPhoto's simple keyword metadata field.



Search engines use HTML metadata to aggregate search results. Few users realize that peripherals such as digital cameras collect and store metadata with photos or that rudimentary copyright and file-related information is stored with every document created.

Even ubiquitous MP3 files include ID3 metadata tags. Although access to these tags might be available to developers, most of it is embedded in files in a non-standard or format-unique way. With standards such as XMP, reading metadata in a file is always the same, regardless of the file type or operating system. This approach has to be considered a benchmark for metadata that must be read by digital asset management systems, Internet or intranet search engines, and people using applications. This file-neutral approach also brings normalization to dissimilar schema. Dublin Core, IPTC, PRISM and other schema after being converted to XMP still retain their identity, properties and values.

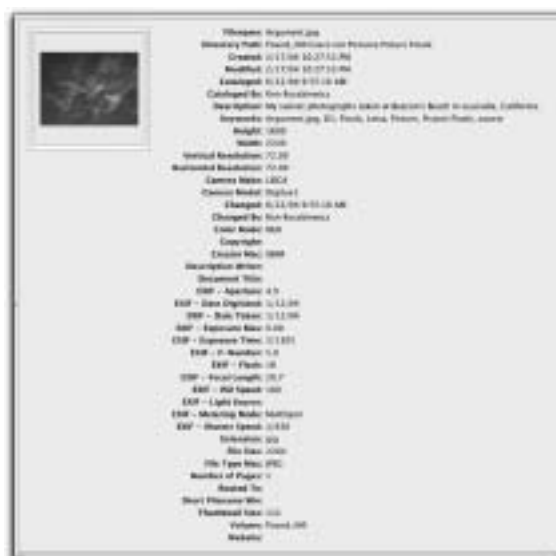
XMP is a platform and not a solution. It provides a consistent infrastructure from application to application. As a platform, it provides some control over the built-in graphical user interface and extensibility. Adobe supports two ways to use XMP. One is by developing custom panels using the [www.adobe.com/products/xmp/custompanel.html](http://www.adobe.com/products/xmp/custompanel.html). This approach does not provide complete control over user interface or platform extension. To exercise unlimited control over user interface and links to external resources, it is necessary to use another approach through Adobe's standard plug-in architecture in combination with the XMP Toolkit. The embedding of XMP metadata is still supported, but the plug-in architecture does not have any built-in limitations on user interface or workflow integration.

XMP is a platform and a methodology that can be applied to any application that allows developers to inject private data into a file. Since most applications, including Microsoft's, allow such private data, there is no reason that XMP can't become the de facto standard for creating rich media in any application.

Adobe already supports schema standards in its applications, including Dublin Core ([dublincore.org/](http://dublincore.org/)), Basic, PDF, Graphics, Media Management, Rights Management, Graphics, Photoshop and EXIF. Soon they will convert and support the IPTC [www.iptc.org](http://www.iptc.org) schema in Photoshop. Adobe also supports custom schema that plays by the RDF, XML and URI rules. Anyone can develop a custom schema. A custom schema can be used on its own or as part of a larger hybrid schema collection menu comprising properties from standard schema. To illustrate this point, one might build a schema using Dublin Core properties for file identification, Rights Management properties, IPTC properties for keywords, and custom properties for job ticket or automated workflow.

### Metadata Trends

**Graphic arts content creation.** It seems inevitable that all

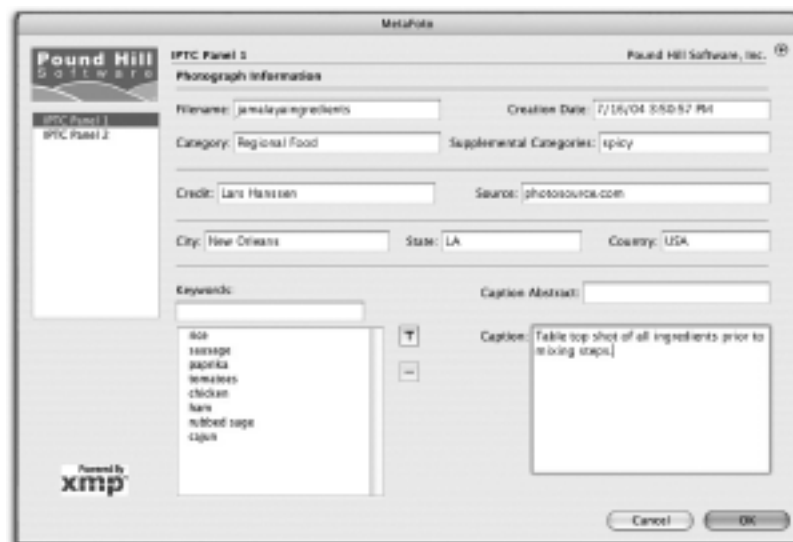


Extensis Portfolio 7 customizable metadata list view.

applications used to create content will include metadata capture and embedding functionality. The pressure to conform to the Internet standards for data sharing mentioned above will be too great. Developing a competing platform for applications that is not based on XML, RDF and URIs does not seem worthwhile. While there are weak spots in search and display support for metadata in Adobe applications, they do not detract much from the overall value.

**Office applications.** While many early adopters and implementers of standards such as XMP and metadata are in the graphic arts industry, the value in being able to retrieve intellectual property is not limited to layouts, illustrations and images. Microsoft Word, Powerpoint and Excel documents also need to be controlled and retrieved from digital filing systems. Microsoft provides rudimentary metadata support in its applications. There are no technical obstacles to implementing XMP technology in Microsoft applications. This includes support for standard and custom schema and custom user interfaces.

Adobe plug-in incorporating IPTC schema, embedded using XMP in Photoshop.



## A Metadata Metaphor

**E**mbedding files with metadata is the equivalent of creating a distributed database. Contextual information remains with the file and can be viewed, edited and indexed on a central database at any time.

Perhaps the best metaphor for the types of things stored in files and represented in metadata forms is a job ticket. In an analogous situation, a paper job ticket form is routed around with physical files. Each user logs onto the form and enters a time that they worked on the file, information about what they did and so on. Metadata can do the same thing digitally.

Using metadata, it is possible to select a user from a popup menu, automatically timestamp the file and add notes about the work done on the file. This rich information is ideal for searching and using as the basic data for reports. Capturing the information

upstream at the point of creation is also preferable to attaching it at the back end. The original creator of the work is the best one to identify the conditions under which the piece was created.

In the case of a hardcopy job ticket, a file is walked from station to station. The workflow is listed on the paper ticket. In a digital environment, we can choose the path of the data and use metadata to trigger scripted actions. It's possible to message someone about a file, update a layout with a new iteration of an element, convert an image to a different format or resolution for another use, or link to a production system using job definition format (JDF), all based on metadata selected and embedded in a file. As a result, a truly automated end-to-end publishing system is now possible. — RR **TSR**

**Consumer applications.** As we mentioned earlier, consumers are driving much of the activity related to digital photography. Consumers face the same issues as professional photographers when it comes to managing and retrieving images. Part of the problem is selecting and injecting keywords into the file that describes the image (or document) well enough to make sorting and retrieving useful. There is, unfortunately, a critical gap between what users want to do and what computer application developers are doing to solve their pain. A solution to this problem might result as an outgrowth of Apple's metadata search engine, the development that surrounds it, or repercussions reflected in future versions of Windows. Ironically, a solution to managing metadata for consumers could eventually result in solutions for the professional audience as well.

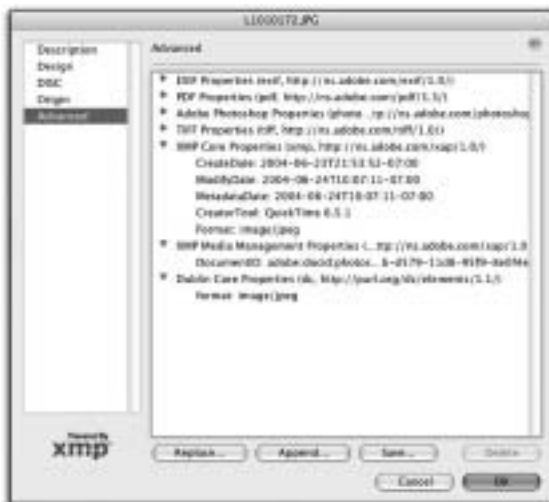
**Content and asset management systems.** Professional asset and content management systems use metadata every day to describe content and propel business processes. It is embedded in files, stored as separate associated files in databases and used ad hoc as an

event trigger in a workflow process. Many of these uses for metadata are discrete. Metadata that is tightly integrated into a solution can be very efficient and secure. Problems arise when a company wishes to build a best-of-breed workflow, is forced to have dissimilar coexisting systems, or is required to allow access to intellectual property from outside the organization. For these systems, the same modularity that standards such as SGML, SQL and ODBC bring to the database side, RDF, XML, URI and XMP can provide to the content creation side.

Without standards, merging hybrid systems is costly and disruptive. To date, nearly all content management vendors and digital asset management developers have recognized the potential and have expressed support for Adobe's XMP. While few have actually implemented it fully, users are beginning to apply pressure for direct support. A major reason for this pressure is the successful release of Adobe's Creative Suite products incorporating XMP.

Content and asset management systems do not inherently interoperate with production systems. One initiative printing and publishing vendors are using to bring about a standard for interoperability is CIP4's Job Definition Format. JDF is an extraordinary schema defining variables for most production processes. The schema begins at the job ticket stage and proceeds through all aspects of print production and delivery. Because it begins at the creation of the job ticket, it does not provide the solution to upstream design and content creation metadata needs. It also does not resolve, nor should it necessarily, unique workflow requirements for graphic arts or marketing communications environments. Other initiatives, such as IdeAlliance's Publishing requirements for Industry Standard Metadata (PRISM) [www.primstandard.org](http://www.primstandard.org) and Disc Image Submission Criteria (DISC) ([www.disc-info.org](http://www.disc-info.org)), IPTC's NewsML and related schema, and others are working to support these upstream metadata requirements.

Built-in schema in a Photoshop CS file.



There are no technological reasons that JDF, Dublin Core and a custom schema cannot coexist in the same file. Good planning is all that is necessary.

## Managing Metadata

**Taxonomies, controlled vocabularies, keywords.** One of the most mature metadata disciplines is the development of dictionaries and thesauri. Government agencies, scientific institutions, photo stock agencies, multimedia companies, libraries and document processing companies, among others, have all put enormous effort into developing controlled vocabularies for their domains. The ability to precisely locate a digital asset is a business advantage for whoever develops it. The retrieved information might be used for analysis, licensing or piped to another system for additional manipulation. Few controlled vocabularies today are part of the public domain or will ever become commercially available. One good example of a mature vocabulary that is freely available is the [www.iptc.org/metadata/](http://www.iptc.org/metadata/) topic vocabulary. More discussion and links to other available controlled vocabularies is at the [www.controlledvocabulary.com](http://www.controlledvocabulary.com) website.

Controlled vocabularies are often represented as keyword properties in schema. Most photographic stock agencies develop their own controlled vocabularies to aid clients in finding images on their sites or on disc-based catalogs. This type of metadata saves research time and provides control over the browsing process. The effectiveness of the search tool is directly related to the quality of the controlled vocabulary. The effectiveness of the controlled vocabulary will be reflected in the success of a site as a sales and marketing tool.

**Creating a controlled vocabulary.** Creating a keyword set or a controlled vocabulary is a daunting task. It's possible to build a vocabulary organically, adding words as the need arises. The best approach is probably to begin with a pre-made set and build on it with unique keywords. For example, while the standard IPTC collection is an excellent starting point for a controlled vocabulary, every newsroom will want to customize it with local and regional categories and keywords. A murder trial, natural disaster and political campaign will dictate a set of locations, people and milestones and will grow organically. Today there is a format known as the extensible faceted metadata language ([www.xfml.org/spec/1.0.html](http://www.xfml.org/spec/1.0.html)) for managing keyword sets. It is scalable and supports branching and keyword annotation.

**When to use standard and custom schema.** In practice, the only meaningful criteria that determines when a standard schema must be used is if interoperation with others at the metadata level is required. If you receive digital feeds with IPTC metadata embedded in it, you must incorporate the IPTC schema into your workflow.



Metadata properties window in Microsoft Word.

Many organizations are initially satisfied with the properties defined in schemas such as IPTC and Dublin Core. But when metadata is viewed as a distributed database meant to reflect each unique workflow environment, it's obvious that some modification or customization is required. Does this mean that everyone must build a schema? Do we have to make a choice between using a standard such as IPTC or a custom one? The answer is a resounding no. There's no technical reason why a core schema like IPTC or Dublin Core can't co-exist with properties from a custom schema. There's also no reason why each individual or organization cannot have its own individual custom schema. It's simply a matter of having the tools to do so, following the rules of the standard or platform, and taking care not to duplicate properties that are part of other standard or custom schema.

Adobe's File Info... function is a good example of different schema coexisting in the same application and file. Any metadata property can be mapped to a field in a database. Telling a field in a database what to expect as far as a schema and property is usually all that is required once the Adobe's XMP Toolkit is supported by the database.

**Mixing and matching standard and custom schema properties.** The following example illustrates a common request for metadata not covered by IPTC and Dublin Core and other widely used standard schema. A magazine might include sections such as Health, Food and Lifestyle for an upcoming issue. It has an edition and volume number. Stylists and photographers on staff do the assignments. The names of the magazine sections, stylists and photographers, and edition titles can be represented through custom schema property values. This metadata information is workflow related. Dublin Core properties can be used to represent standard



Managing IPTC Topicsets with MetaKey.

metadata properties such as creator, creation date, originating application, copyright and so on. The custom schema values can co-exist with Dublin Core values in the file. That is the nature of extensible standards.

### Tying It All Together

The obvious question is whether the time is right to begin using metadata. Is it still bleeding-edge technology? Are there tools available to allow integration into an existing workflow? Is this a transitional technology or one likely to be around in some form into the future?

**Is this still bleeding edge technology?** Adding metadata to files is not new. New technologies such as XMP have moved metadata capture upstream into the creative process and have brought an adoptable, standard approach to embedding metadata in files. Users can now take complete control over what is embedded in

*The problem arriving at this data access nirvana has more to do with human than machine issues.*

files. There is little risk since even if standards or schema change in the future, there will always be a software approach to converting to a new scheme or schema. This is currently the case with IPTC. Any Photoshop file with IPTC metadata embedded in the header is automatically converted to XMP when it is saved in Photoshop CS.

**Are tools available to allow integration into an existing workflow?** Tools from Adobe can read XMP metadata in Adobe application files and apply it to data repositories or convert it for other purposes. The XMP Toolkit is available through an open license agreement without charge. Commercial and free tools [www.poundhill.com/index.php?pageid=2](http://www.poundhill.com/index.php?pageid=2) are available to start metadata implementation and exploration.

**Is this a transitional technology?** Perhaps. However, the stated objective of the W3C committee in developing its

new standards is to be inclusive of existing schemas. This means that the investment made in developing Dublin Core and IPTC schemas will not be lost. Because they conform to standards and methodologies such as XML, RDF and URI, they will continue to be supported in the future. Custom schema that conform to these standards should also be safe. We have already seen numerous variants of XML available under the Simple Object Access Protocol (SOAP). Microsoft supports SOAP and provides a migration path to .NET. This ensures that work done to support metadata in their environment is not lost. Adobe's commitment to XMP is real. Any change to another platform will likely recognize previous metadata standards and provide a similar migration path to developers supplying the tools.

### Conclusion

Metadata is perfect for computerization. A computer can read through a file, find a tag that matches a category (property), pick out keywords (values) that match search criteria and print the results on the screen. Perform the search on a fast processor or network and it seems to happen instantaneously.

The problem arriving at this data access nirvana has more to do with human than machine issues. The major obstacles to implementing metadata in content creation environments include a lack of tools to build custom and composite schema and manage controlled vocabularies, strategies to overcome (perceived) resistance to entering metadata during content creation, security, and support for open standards by data repository developers.

There are few options when it comes to tools to build schema or manage vocabularies. More will arrive to complement content and asset management tools because of the implicit synergy between the two. It's a chicken and egg dilemma, when customers demand support for XMP from system developers, they will develop or acquire the tools to make the process work.

Overcoming the perceived resistance to entering metadata at content creation can be solved by a combination of sound usability practices in software design and a dash of social engineering. We so rarely get to use well-designed software that we are pre-conditioned to the drudgery of filling out online forms. Fortunately, a combination of predefined metadata templates, auto-fill fields, default and saved values can turn a chore into a handful of mouse clicks. Accept nothing less.

Security is an unresolved metadata issue. Encrypting metadata in a file is not currently possible. There is no security value type to assign to a form field. If security is a major concern, it is possible to watermark a file and embed standard and Read Only metadata in it. The better solution is to control the movement of the data at the repository. Controlling access to the data combined with watermarking and data embedding will deter most unauthorized uses.

Rich media protected with copyright information,

enriched with searchable keywords, and annotated with information about the project and creator is valuable on its own. Using this data to access and distribute information locally or globally adds additional value. Following recognized standards is important in both instances. What's missing is a seamless metadata interchange between content creation applications and content and digital asset management systems.

The way it should work is metadata is captured on the creation end, checked into a data repository where the data is consumed and stored in fields in the repository's records. This data is used to generate reports or for searches. If data needs to be changed because of usage changes or repurposing, it is done directly to the file, in the repository.

The alternative to this approach is to inject metadata in file batches as they are being added to the repository. Unfortunately, you can't do this today, though you will be able to soon. Most system developers have expressed the desire to support XMP, and

Adobe has made integrating XMP a straightforward operation for most systems. System developers from the high end to the low end are waiting for customer

**A combination of predefined metadata templates, auto-fill fields, default and saved values can turn a chore into a handful of mouse clicks.**

demand before gluing it all together. When considering a personal or enterprise asset management system, be sure to inquire into the developer's "metadata support strategy." TSR

#### About the Author

Ron Roszkiewicz (ron@poundhill.com) is president and co-founder of Pound Hill Software, a specialist developer of metadata applications and metadata-based workflow aids based in Encinitas, Calif.