XMP™
Adding Intelligence to Media

Adobe Extensible Metadata Platform in Newspapers and News Magazines

As recently as five years ago, not one in a hundred technologists and practically no one in the news business was aware of the term *metadata.* But that's changed now, and many in the Internet community view metadata and the related Semantic Web as important for the future.

Symptomatic of this shift, Adobe Systems has invested heavily in its metadata initiative, Adobe® Extensible Metadata Platform (XMP), and elevated it to the status of core Adobe technology, along with Adobe PDF and Adobe PostScript® software.

Because it carries the metadata for files created with the latest versions of Adobe applications, XMP already impacts the newspaper and news magazine community. But the full promise of XMP will be realized when businesses use the technology to reduce production costs and create new revenue-generating products.

Metadata is to

With structure to make human understanding precise

Nutrition Facts		
Serving Size 1/2 Cup (126g) Servings Per Container about 3.5		
Amount per serving		
Calories 260 **Calories from Fat** 80		
		% Daily Value
Total Fat 9g		7%
Cholesterol 6g		3%

As SOUP **is to**

And structure to make machine understanding precise

12345 67890

Metadata Analogies
Defining metadata through comparison with a label on a can.

Introduction: changes in the news business

The image of a cell phone serving as a newsreader reminds us that the Internet has dramatically altered the news business.

Internet News Delivery
Cell phone serving as a newsreader.

Adobe®

News cycles have collapsed. Cross-media publishing has fostered cross-media competition. While circulation of print news continues to trend downward, the consumption of news is high, due in large measure to the vibrant new delivery modalities.

RDF Site Summary (RSS) technology is a good example. As with the experience of the World Wide Web, this technology was developed to serve the clearly defined needs of a small user domain—the early bloggers in this case—and then entered widespread use.

WebNews.com   international edition

RSS Feed (Really Simple Syndication)

XML   **Click here for your RSS feed**

RSS Newsfeed
Link on a news website.

Specialized applications were critical to the popularity of RSS. Just as the browser was critical for the web, the invention of the RSS reader application increased the RSS business. The first half of 2004 was a watershed, during which most large newspapers and magazines began supporting RSS newsfeeds. There are over 100 RSS feeds from *The Washington Post* alone. By mid-2004, there were feeds from an estimated 4,000,000 bloggers, and now that the reader technology is being absorbed into the browser, people are forecasting RSS feeds in the hundreds of millions in three years.

Sign up for WebNews.com RSS feeds
Click content area(s) you're interested in subscribing to, and follow the instructions to add your news reader to your personal MyWeb.com page

Arts	**XML**	Multimedia	**XML**
Automobiles	**XML**	National	**XML**
Books	**XML**	Local Region	**XML**
Business	**XML**	Most E-Mailed Articles	**XML**
Campaign 2004	**XML**	Home Page	**XML**
Circuits	**XML**	Olympics 2004 NEW!	**XML**
Dining & Wine	**XML**	Real Estate	**XML**
Editorials/Op-Ed	**XML**	Science	**XML**
Education NEW!	**XML**	Sports	**XML**
Fashion & Style	**XML**	Technology	**XML**
Health	**XML**	Theatre	**XML**
Home & Garden NEW!	**XML**	Automotive	**XML**
International	**XML**	Travel	**XML**
Magazine	**XML**	Washington	**XML**
Media & Advertising	**XML**	Week in Review	**XML**

RSS Feed Sign-ups
Example of subscription choices.

But there is more to this than what some think of as merely a push technology for headlines. It represents a fundamental change in publishing. Where traditional publishing focuses on aggregating content into editions of pages of stories (and ads, of course), RSS and the related technologies represent a shift toward microcontent delivered to the consumer in response to specific requirements. As RSS technologist Nova Spivack puts it, "...information will be published in discrete, semantically defined 'postings' that can represent an entire site, a page, a part of a page, or an individual idea, picture, file, message, fact, opinion, note, data record, or comment."

The term *semantic* is critical in this context. It is a code word for the project called the Semantic Web. RSS depends on computers reading metadata that is coded according to a worldwide standard, so there can be meaning—semantics—in the sense that logical tests can be carried out without human intervention. This is at the core of the free but focused information transactions. From there, it is only a small step to remunerative transactions, whether for information or for other products.

Metadata in newspapers

Because the information systems in the news business cannot work without it, electronic metadata has been part of the news business for more than 40 years. It arises in a number of contexts and takes a number of different physical forms.

Metadata in database fields

Metadata arrived in the newsroom with the first word processing system. The first use was entries in databases that identified stories.

Metadata in early newspaper database structure

		Fields			
	Record number	Writer	Title	Date	Story
Records	1903045	Bernstein	It's Colson	8/21/72	0010010 1011001 1001100 0010010
	1903046	Woodward	It's Dean	8/21/72	0010010 1011001 1001100 0010010
		Metadata			Data

Early Newspaper Database Structure
Example of metadata in an early newspaper database structure.

In the database case, the metadata was logically joined to the data, but might have been physically located in a different part of the computer memory.

Metadata in the text file, as a header before the story

Once a means was found to transmit computer files over telephone lines, it was possible to identify the files by including metadata with the data in the same file, typically as a header before the data. To conserve space (and, at 1200 baud, time), the American Newspaper Publishers Association (ANPA) designed and promulgated a metadata format to be used by the Associated Press (AP). The metadata was located in at the beginning of the transmission in a header, providing values for each of these nine properties:

• Level of service

• Category of text information

• Identity number

• Text typesetting format

• Selector code identifie

• Keyword for specific and unique identity

• Priority of the message

• Month and day the story was filed

• Version and reference fields to assist in linking the text of a message to a previous or subsequent message

The precise format of the header was specified on a character-by-character basis and documented in an ANPA metadata header in sufficient detail for computer programmers to build receiver applications that could fill out the database metadata fields and store the wire stories in the editorial database.

SYN SYN SOG s 1378 TFI

"Start of message" codes

message number

"End of field" delimiter

"s" for sports, could be "a" for national news

ANPA Metadata Header
Example of the first few characters in an ANPA metadata header.

Though this format was totally idiosyncratic, inflexible, and inextensible, most of the other news agencies in the world adopted it exactly, regardless of their own requirements or perspectives. Though the transmission speeds are now up to a thousand times faster, it remains the standard at AP and other agencies. This system has even outlived the organization that developed it; ANPA has become The Newspaper Association of America.

Metadata in a special packet within an asset file

When image files arrived in the industry, the need for metadata became even more obvious. With a story, one can at least read the words and deduce the context. An image file, in contrast, is a very big array of numbers, meaningless without some information as a guide. Following the lead of the news stories, the International Press Telecommunications Council (IPTC) devised a set of metadata categories and a data structure standard. The IPTC structure is yet another serial format (delimiter, property value, delimiter, next property value). The list of metadata properties appends 10 properties to the original nine properties. It is, like the ANPA header, idiosyncratic and inextensible.

IPTC Metadata Packet
Metadata included within photo file boundary as distinct information packet.

While it was invented for identifying just wire photos from members of the IPTC, the evolution of this metadata initiative shows how an application's software vendor can contribute both functionality and applicability. For the release of Adobe Photoshop® version 4, Adobe developed a data mechanism for embedding the IPTC properties into an arbitrary image file and for reading them out, and the IPTC metadata properties emerged as the focus of the Get Info dialog box in Photoshop. Image database vendors rapidly added support for the Adobe IPTC structure, and it became the unofficial standard for metadata for digital photo files.

Digital camera technology brought yet another metadata concept to the newsroom. This time, the concept was determined by the Japan Electronics and Information Technology Industries Association (JEITA), more or less an association of digital camera vendors. JEITA's EXIF metadata format transmits values for up to 120 different metadata properties—primarily ones that deal with the physical characteristics of the photo—embedded into every digital image file. Though encyclopedic within its domain, EXIF is yet another idiosyncratic and inextensible metadata format.

Metadata in a companion file linked by reference to the asset file

Another notable example of newspaper metadata is the Adsend system administered by the AP. In the Adsend system, the advertiser fills out a web form that includes seven metadata properties, and then uploads the ad in PDF format. Then AP sends out the ad to the intended recipients, together with what is called a *sidecar* file that contains the metadata.

Adsend System
A sidecar file becomes a newspaper ad.

These examples support the observation that metadata is ubiquitous in a news enterprise, and they also illustrate the spectrum of circumstances that the information systems departments have to deal with.

DOMAIN	WHO DETERMINES CATEGORIES?	WHO DETERMINES STRUCTURE?	WHO SETS VALUES OF CATEGORIES?	EXTENSIBLE?
Assets in content management system (CMS)	CMS vendor	CMS vendor	CMS user	With substantial difficulty and expense
Wire stories	ANPA	ANPA	AP, AFP, Reuters…	No
Wire photos	IPTC	IPTC	AP, AFP, Reuters…	No
Photoshop files through version 7.0	IPTC	Adobe	Photoshop users	No
Digital camera files	JEITA	JEITA	Digital camera users	No
PDF files of ads	AP	AP	AP	No

- There is a different metadata environment for each domain, each type of file, and each digital asset. Each one requires a different set of routines to extract and use the data.

- There can be a multiplicity of metadata structures for the same asset.

- Only the CMS structure is extensible, and that one only with substantial difficulty and expense.

- Though several of the metadata systems were set up by industry committees, none of them conform to a general standard developed for a broader domain. In particular, none of these systems relate easily to the emerging standard for digital publishing of microcontent.

Adobe XMP: an extensible metadata platform built on the standards of the Semantic Web

As many newspapers and magazines have already discovered, XMP is already making its presence felt in the enterprise.

DOMAIN	WHO DETERMINES CATEGORIES?	WHO DETERMINES STRUCTURE?	WHO SETS VALUES OF CATEGORIES?	EXTENSIBLE?
Files created with Adobe products: Adobe Creative Suite Adobe InDesign® Adobe InCopy® Adobe Photoshop Adobe Illustrator® Adobe GoLive® Adobe FrameMaker® Adobe Acrobat®	User selects from several standard category sets supported in base XMP and/or implements own category sets	Adobe, based on the Resource Definition Framework (RDF) standard of the World Wide Web Consortium (W3C)	User	Explicit procedure provided by Adobe together with tools and training materials

The domain of XMP is the files created by the different Adobe creative applications. In newspapers and magazines that use an Adobe-centric workflow, virtually all the content—all the files containing stories, pictures, graphics, ads, and pages—will include the XMP data packet.

The formal structure for the information is adopted from the RDF standard, which the W3C developed over more than a decade of work. XMP is implemented in XML to take advantage of the ease in parsing the metadata and the availability of open source code resources. XML brings XMP into the comfort region for legions of computer engineers with web skills.

In many respects, the essential difference is that RDF is designed for any web resource, while XMP is for Adobe application files. Because the structure is the same, XMP can easily draw on standard metadata category sets, such as the one developed by the Dublin Core Initiative to support what is often referred to as a "library card catalog for the web." The Dublin Core community defined the terms and wrote the documentation for the RDF implementation. With only minor modifications, Adobe has been able to include support for this metadata categorization—called a *schema* in both the RDF and XMP context—as part of the standard implementation of XMP.

One observer has termed XMP an example of "enlightened self-interest." Adobe had to adopt some sort of common metadata platform to make interapplication functionality work effectively within the scope of its creative applications. Placing a Photoshop file in an InDesign frame, for example, involved registering with the InDesign file database a number of separate metadata items that the layout application needed to do its job, including size, mode, and location.

Rather than develop yet another idiosyncratic metadata system, Adobe decided to follow the RDF lead. XMP inherited the benefit of thousands of work hours of RDF development that rested, in turn, on two decades of work in the artificial intelligence community. RDF was supported by knowledge engineers in a number of fields who were building metadata structures for their domains.

Though Adobe sacrificed the potential to sell access to a proprietary technology, it saved time and money in development and lowered the risks involved.

Adopting the RDF model, Adobe and its customers stood to gain as well from the eventual emergence of an application set of agent software that would manage information and commercial exchange automatically by reading the structured metadata and using the values to control routing and transactions.

To convert a metadata standard to a platform, Adobe first invented a packet structure for embedding the XML-tagged information in Photoshop, Illustrator, InDesign, InCopy, GoLive, and Acrobat PDF files. It developed a set of user tools for accessing these metadata packets and reading or setting the values of the metadata properties. It implemented a standard collection of different metadata property sets, including the Dublin Core set as well as a digital rights set.

Files in an InDesign Layout
Files from Photoshop, InCopy, Illustrator, and Acrobat in an InDesign layout.

In what is proving to be the defining step in this metadata initiative, Adobe adopted the full extensibility mechanism developed for RDF. Extensibility means that the scope for the XMP metadata is broad (or similar) in both depth and breadth. In principle, each newspaper and magazine can define its own metadata categories and implement them in XMP schemas. The categories from one newspaper can be borrowed by another or shared among the members of a regional association that has common interests. For example, a set of classified ad categories developed for Southern California newspapers might include surfboards, while those for Wisconsin might include milking parlor equipment. Unlike the wire story categories that have been virtually frozen for many years, extensibility means that special-purpose XMP metadata sets can change as the news or commercial environment changes.

Finally, Adobe created the support structure to facilitate the development work of sophisticated users and provide a foundation for independent consultants and software development companies to add XMP competence to their scope of work.

ROI and XMP

XMP presents a substantial return on investment (ROI) opportunity for news businesses. Initially, the benefit will come from productivity enhancements, followed by revenue opportunities for new information products and services.

Productivity enhancements with XMP

XMP metadata is already appearing in the newsroom. Beginning with the first Creative Suite editions of InDesign, Photoshop, and the other CS applications, the File Info functionality is based on XMP data storage. The dialog box returned when the user selects File>File Info is an interface for setting and displaying the XMP metadata properties.

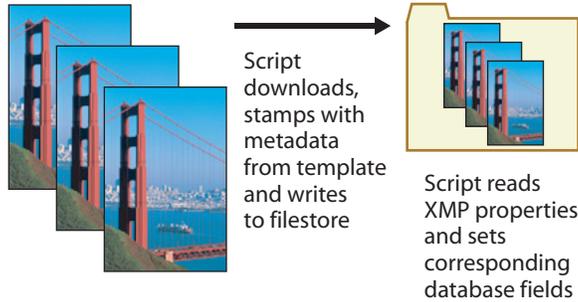**Adding Intelligence to Media**

The XMP Logo
The XMP logo appears at the bottom of the File Info dialog box in CS versions of Adobe print and web publishing applications.

Accordingly, using this built-in capability is a natural first step for a newspaper or magazine moving into the XMP age. Introducing the productivity proposition, this dialog box features a template procedure that writers, editors, designers, and photographers can use to personalize each file they create.

Asset archives

The next step is to use the information to enhance workflow productivity within the paper or magazine. One of the most frequent applications is specialized database structures for managing inventories of different kinds of digital assets. Increasingly, system vendors support XMP for this purpose. Additionally, special applications can be built by using either open source database applications or mass market products, such as Filemaker Pro or Microsoft Access, with a simple scripted interface.

A common example is a database of all the digital camera files that photographers bring back from an assignment, not just the one or two that are selected to run with the story. This is the digital equivalent of the files of negatives that are a feature of film-based systems. The pertinent metadata is little more than the name of the photographer, a description of the assignment, the date, and the assignment reference. The photographer sets the XMP metadata with a template as part of the upload from camera to enterprise file storage. When the folder watching the script recognizes a new file, it creates a corresponding database record and reads the metadata into its fields.

Script downloads, stamps with metadata from template and writes to filestore

Script reads XMP properties and sets corresponding database fields

Files Stamped with XMP Properties
Metadata is now stored in a database.

Another example is an archive for display ads created in house. In many publications, it is common to maintain a database of the work in progress in the ad layout department, and then move the ads to a filestore once they are published. This is another case where a simple database can be built and maintained inexpensively on an XMP foundation.

The ROI point is that the common metadata format across Adobe applications means that the same agent used in the first case (on the editorial side of the enterprise for Photoshop files) works identically in the second case (on the advertising side for InDesign files).
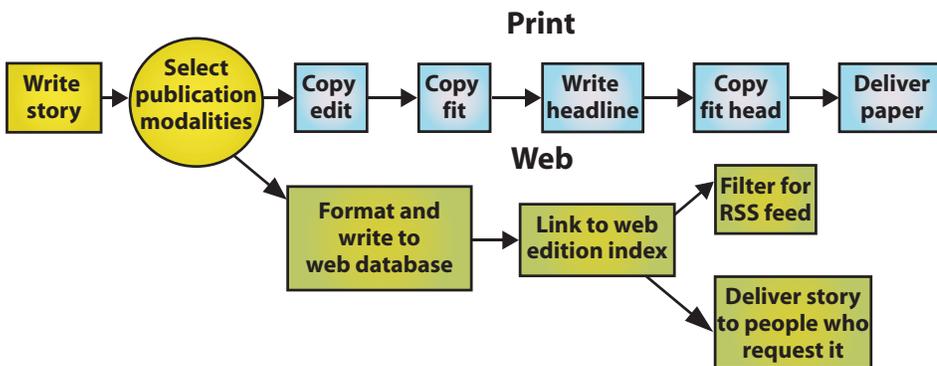
Workflow automation
The W3C standards at the foundation level of XMP were developed with a view toward machine comprehension, especially to serve in automatic process control environments. At its most simple level, this means having machines that test values of properties and then take actions depending on encoded rules. As a result, XMP metadata is particularly useful for workflow automation. In the newspaper setting, one familiar application is moving projects from queue to queue, as successive editorial steps.

Write story → **Copy edit** → **Copy fit** → **Write headline** → **Copy fit head** → **Deliver paper**

Newspaper Workflow
Moving projects from queue to queue.

In the environment of cross-media publishing, however, these workflow steps become much more complex, and the sort of machine processing as that shown below is an everyday feature.
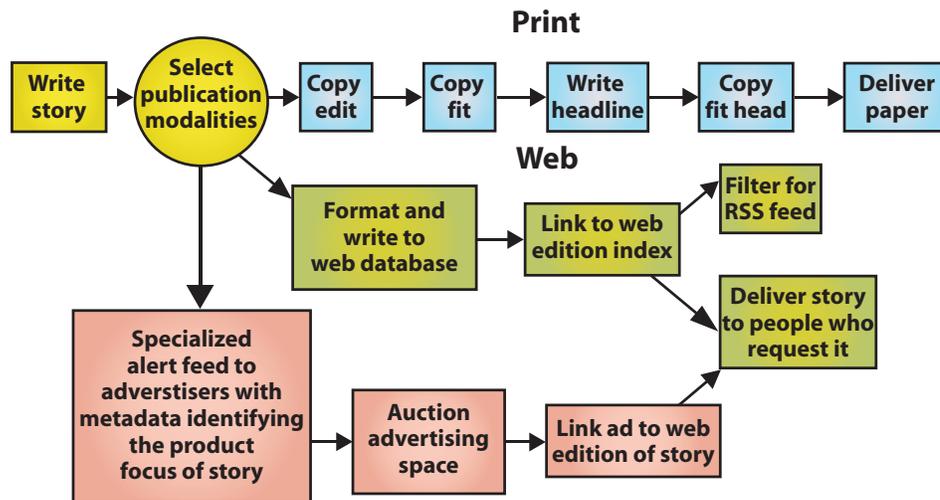
Print

Write story → **Select publication modalities** → **Copy edit** → **Copy fit** → **Write headline** → **Copy fit head** → **Deliver paper**

Web

Format and write to web database → **Link to web edition index** → **Filter for RSS feed**

Deliver story to people who request it

Cross-Media Publishing Workflow
Similar to newspaper workflow, but includes machine processing.

Revenue opportunities

The revenue opportunities for XMP occur from this stage in the process. The success of eBay matching buyers to sellers with a large central database, or the success of Google matching ad delivery to content from a astonishingly large database, shows the sort of economic success that stems from machine processing.

In both of these cases, the cost of technology is in the hundreds of millions, and no single news organization can afford it. However, a new generation of services will soon decentralize these sorts of transactions, based on the metadata standards used in XMP. Consider the example diagrammed below: An agent alerts the community of advertisers about the content of a coming story, and a specialized application auctions ad space in the web version a story.

Print

Write story → Select publication modalities → Copy edit → Copy fit → Write headline → Copy fit head → Deliver paper

Web

Format and write to web database → Link to web edition index → Filter for RSS feed

Link to web edition index → Deliver story to people who request it

Specialized alert feed to advertisers with metadata identifying the product focus of story → Auction advertising space → Link ad to web edition of story

Workflow with Advertising Opportunity
Using metadata to alert potential advertisers of a possible match of product and story.

Special software tools make the examples in these diagrams work. One tool is a script or application that reads the values of the XMP properties, applies logical tests to that information, and takes some sort of action. In the first diagram, the action is moving a file from queue to queue. In the second diagram, it is moving a file into a web publishing system, setting the fields of the web content management from the metadata values. In the third diagram, there is a collection of applications that interact with each other on the basis of metadata and manage the auction, including the financial transaction as well as the information transaction.

In all the cases, there is a dependence on the careful specification of the metadata and a standard structure for processing. XMP provides the framework for this specification, and Adobe has included support for several different types of specifications that are sufficient for the basic applications. The innovative and exotic applications will depend on new ways of categorizing new domains. To auction ad space, for example, you would need to have a set of metadata properties that allow a match between the content of the story and the scope of interest of the advertiser.

Increased acceptance

There is a rapidly growing community of industry associations, governmental bodies, and private companies focusing on the Semantic Web, and XMP developments are in the vanguard.

Extending XMP

XMP is extended by defining and documenting new ways of using the metadata platform in different domains. Increasingly, these new classifications—called *ontologies* in the Semantic Web community—are the outcome of collaborative work by experts from different companies and agencies, often with direct Adobe support and involvement.

For example, Adobe supported the initiative in the magazine and printing business to define and document a metadata ontology to be used in submission of digital photographs and illustrations as part of the Digital Image Submission Criteria (DISC). The members of the working group created the scheme, and then worked with Adobe to implement it through input and display panels for the different Creative Suite applications. Similarly, Adobe has been working with the IPTC to migrate that organization's existing metadata format to the XMP standard.

Adobe has been active in training the developer community as well, so that consultants and companies are ready to help any individual publisher or group that has its own special requirements.

Database connectivity

Vendors of editorial and content management systems are adding features that support capturing metadata from the XMP packet into the fields of the database records corresponding to Photoshop, InDesign, and other Creative Suite applications. XMP support is available in some client/server digital asset management systems. Specialists in editorial solutions for small newspapers and magazines also use XMP as the basis for metadata control.

Visions of the future

The agent technologies that will be built on XMP and other semantic initiatives are only at the beginning of their product cycles. They are moving out of universities and research labs and into applications pioneered by early adopters in both business and government.

Summary and conclusions

With cable news on the video side and bloggers on the Internet side, the newspaper and news magazine business is challenged to shorten news cycles and deliver microcontent efficiently. To flourish in this competitive setting, print news businesses have to increase the velocity of their information mobilization and shift into entrepreneurial mode with new news and ad products.

The effective use of metadata is emerging as a key technology resource to meet the new challenges because it offers both productivity advantages and the prospect of new machine-managed information services.

Metadata is nothing new in news organizations. For about 40 years, first stories, and then photos and more recently ads, have been arriving in the newsroom in digital format, identified by companion metadata. However, in each case, the structure was inflexible, idiosyncratic, and inextensible.

XMP offers an alternative. It is a core technology at Adobe Systems, strictly based on open standards that are both flexible and extensible.

While the objective of Adobe is a common metadata structure across all the files in the Creative Suite applications, XMP has attracted support from industry associations and from system vendors who look to the technology for much broader applications. Innovative solutions are already available in the area of workflow and content management, and early adopters are experimenting with automated information utilities.

In summary, an enterprise commitment to XMP may be critical for leadership in the next era in news and ad delivery to the consumer.