Editor: Nevenka Dimitrova
Philips Research

Is It Time for a Moratorium on Metadata?

Dick C.A. Bulterman
CWI, Amsterdam

I'd like to begin this article with a short story about metadata:

Once upon a time, there was a small republic nestled in the heart of an inaccessible chain of mountains. The country was separated geographically into two regions by a range of towering peaks that ran from north to south. Originally, the citizens were mountain dwellers, but they had long since moved downhill, either to the eastern or western side of the country. Over time, two dialects of the national language had developed that, together with the terrain, provided nearly total regional autonomy.

The principal activity of the citizens in each region was producing and consuming bread: whole wheat, long grain, short grain —even rice bread. Hundreds of types of bread were sold in shops, and each town had its own specialty. Although many people were employed in the support tasks of growing wheat, making flour, and transporting raw and finished materials, it was the bakers who were at the top of the social ladder. The citizens were well fed and happy.

Life in the republic changed dramatically when experiments in the East with new, more powerful forms of yeast led to the development of the hot-air balloon. It was not long before the balloons were sent to study the western region. Adventurers returned with wondrous tales of life across the mountains. Although presentations on wildlife and natural history always drew polite applause, what the citizens in the East were really hungry for was information on the types of bread available across the national divide. When one insightful adventurer attempted to purchase a cookbook for western bread, a startling discovery was made: the entire population of the West was illiterate! Everything of interest and importance—such as recipes—was recorded in drawings and photographs, on 9-track cassette tape, or film. Where the East had extensive (and mandatory) vocabularies for describing every aspect of bread and bread production, the West had extensive film archives and audio descriptions, all of which were recorded in a proprietary format. All western bread was simply labeled: bread (where and when it was bought differentiated what kind it was).

The limited transport capacity of the balloons and the extended duration of the inter-regional journey made it impossible to import fresh western bread, but this did little to satisfy the appetites of the eastern population. The brightest young bakers were sent out to document every crumb of information available about the size, shape, color, texture, weight, and nutritional benefits of the neighboring breads so that every variety could be located on demand. Instead of learning how to bake bread, they were taught how to describe it. But alas, the ingredients and production methods used in the West were dissimilar enough to keep their bread indescribably delicious. Relations between regions became strained when the West's bakers and suppliers refused to adopt the East's naming schemes. Worst of all, given the comfort of their new air-conditioned offices—and the fact that they no longer needed to get up before dawn—those bright young eastern bakers had little interest in returning to the ovens once they got home. The baker's trade lost its popularity and status. Consolidation took place, and the variety of bread types in the East decreased dramatically.

All of this led to local unrest and disenchantment: the easterners felt they were missing out on an optimal experience because they were sure that, somewhere in the West, a better bread was being buttered. Wordsmiths were brought in to restore public confidence in eastern products and soon "E-Bread: The Crust You Can Trust!" billboards sprang up everywhere. To lure even more consumers back, generic bread

1070-986X/04/$20.00 © 2004 IEEE

Published by the IEEE Computer Society

produced at massive production facilities in the East was given trendy names and enriched with "information grains": nanocapsules that, when brought in contact with tooth enamel and saliva, caused a pico message to be sent out on the 2.4-GHz band. With each bite, pop-up information was broadcast about the (largely fictional) history of eastern bread—as well as marketing links for official baker's clothing and other commercial tie-ins. A whole "fact food" industry was established to define the most compelling bits per bite.

Initially, the public's appetite for information in their bread grew. Over time, however, the number of information grains exceeded the nutritional grains in each loaf. Eating even white bread became exhausting. The eastern population became overinformed but undernourished and weak. After a particularly harsh winter, everyone died.

This is not a pretty story. No frogs turned into princes: no happy endings, not even an IPO at the end of the rainbow. Still, it can serve as a useful backdrop for discussing the use of metadata at the beginning of the 21st century.

What metabuns, which meta-ovens?

During the past 10 years, three major metadata specifications have been released: The Dublin Core Metadata Initiative (DCMI) provided the first systematic attempt at defining an interoperable metadata standard that could be extended with specialized vocabularies; ISO's Motion Pictures Experts Group provided MPEG-7, a multilayered standard used to describe the structure and, to a lesser extent, the substance of a composite multimedia data stream encoding; and the World Wide Web Consortium (W3C) released various components of its megametadata initiative for creating and managing the Semantic Web. The DCMI was media agnostic, but still heavily text biased. MPEG-7 expanded the scope with the—in retrospect—relatively modest goal of describing a single audiovisual object. The Semantic Web—perhaps befitting its World Wide heritage—set its sights on characterizing all information, regardless of location or encoding.

Dublin Core

The origins of the Dublin (Ohio) Core were rooted in the 1994 World Wide Web Conference (not to be confused with the W3C), where attendees voiced concerns over the tractability of the then-500,000 documents stored across the Web.

Table 1. The Dublin Core.

Content	Content Instantiation	Intellectual Property
Title	Date	Creator
Subject	Format	Publisher
Description	Identifier	Contributor
Type	Language	Rights
Source		
Relation		
Coverage		

One conference led to another, and by 1998, the DCMI standard had been developed and encoded as an Internet standard. The DCMI, reflecting its strong library sciences roots, was geared to describing objects so that they could be easily located. To achieve this goal of then-modern metadata, the Dublin Core defined 15 elements to identify content and its access rights (see Table 1).

In most Web documents, DCMI information is encoded as a set of name-value pairs:

```
<meta name="Title"
    content="Is it Time for a
    Moratorium on Metadata?" />
<meta name="Creator"
    content="Dick Bulterman" />
```

Unfortunately, the simplicity of the DCMI had significant built-in limitations. It was one thing to agree on element classes for structuring descriptive content and quite another to agree on the format of the content. For the most basic forms of metadata, such as titles and authors, this wasn't a major issue, but providing uniform descriptions for object subjects and content proved much more difficult. Although the DCMI provided a simple unification model, the consistency of the model's content was left as an exercise for the writer. From a multimedia perspective, the Dublin Core also left much to be desired, because it didn't provide guidance for nontext objects beyond what had been used in a library card catalog.

MPEG-7

MPEG-7 was the product of a massive, five-year standardization effort to define metadata that could be used with video and—to a more limited extent—audio media objects. The principal (and most used) aspect of MPEG-7 is its location service, which defines a structure that allows content to be found on demand from a server.

Perhaps you're thinking: "Instead of trotting

Figure 1. An image from a scene of one story within a 22-minute newscast. (Photo courtesy of the author.)

around the globe for five years, couldn't they have found a spare afternoon to agree on the top 10 fields for labeling a piece of media content?" If so, you're missing the point. MPEG-7, unlike the DCMI, doesn't simply label content, it describes it. MPEG-7 provides a document description language (DDL) to encode a structured, schema-based model to describe media-specific properties of audio, video, and text data, as well as the individual content objects within each primitive media stream. It also contains something that broadcasters and equipment manufacturers really like: licensed technology. (The DDL alone is covered by 17 patents.) The reason that commercial parties tend to favor licensed components is that there's an identifiable controlling organization with a commercial incentive to develop and maintain—and protect—a standard's essential technological components. This provides, they feel, the promise of long-term interoperability and a level technological playing field. (This also raises the entry bar for organizations that can't leverage extensive patent portfolios.)

Many of the content-feature metadata supported by MPEG-7 (such as the camera angles used, the motion activity of dollies, or the sound-effect set used) are relatively esoteric, but it's metadata that is easy to gather and maintain during production. Data that's probably much more important includes characteristics of the media content itself, such as the shapes or geometric extents of content objects. Why? Because doing so opens a whole new world of content use. Consider the newscast in Figure 1. Suppose you wanted to purchase that snazzy sports coat or those hip eyeglasses while watching the news: An interested commercial party could associate a shopping link with a set of object regions and encode both using MPEG-7. Best of all, this information need not be stored directly in the media file but could exist in an independent metadata stream. This not only allows for localization based on a set of user profile preferences, but it also means that the original content owner might not need to get a slice of the revenue from the sales, since the owner's copyrighted media isn't being altered. Of course, MPEG-7 could also contain links to biographical data on actors or extra plot information, but this kind of data *costs* money to make; home shopping generates income, which in the real world of metadata production is an incentive that shouldn't be underestimated.

The Semantic Web

The Semantic Web provides a layered set of components to define a scalable set of metadata definitions, allowing a generalization of the textual descriptions that form the basis of XML. Although somewhat undervalued, a primary component of the Semantic Web is the uniform resource indicator, or URI. (Without a URI, objects don't exist; without objects, you don't need metadata.)

At the next layer are RDF, RDF Schema, and OWL. RDF is the Resource Description Framework, a mechanism for encoding metadata based on statements containing subjects, predicates, and objects. (The terms resources, properties, and property values are also used for these building blocks of RDF.) RDF Schema provides a class hierarchy of descriptors for defining a structured vocabulary within and across statements. OWL is the Web Ontology Language, an extended set of property and class definitions that allow a formal definition of the terminology used within Semantic Web documents. (The fact that it's called OWL instead of WOL is a nice example of how you can't count on the names of objects to necessarily contain self-descriptive meta-information.)

Perhaps the principal difference between the Semantic Web and DCMI/MPEG-7 is the hope—or expectation or belief—that, in the future, you may be able to reason about the content using its statements instead of simply accessing the content described by them. This aspect, which has been the focus of intense (and intensely personal) debates—often, on both sides, with an exceptionally high "if you're not for us, you're against us" character—highlights that perhaps the greatest problems for the Semantic Web might not so much be technological as theological. There have been several intense online debates about the semantic potential of the Semantic Web. The W3C is currently bending over backward to show that the Semantic Web is not about AI and concerned only with "real world" applications. Semantic processing is now seen simply as the icing on the cake.

Metadata: The greatest thing since sliced bread?

The past decade has been good for metadata. New metadata standards were published before

the ink on the specifications of their predecessors (or competitors) had dried. Companies (large and small) issued testimonials to the potential values of new standards before the standards were deployed. Hundreds of academic papers were published describing metadata structuring, analysis, and (to a much lesser extent) creation tools. Standardization groups not only were born, they retired because their jobs were done. About the only thing that the past 10 years hasn't produced is much useful new metadata itself.

Perhaps one of the great metaparadoxes of our time is that although more information is being searched than ever—it's been said that more people use Google in a week than have used all of the world's libraries in a decade—the (relative) use of conventional metadata is probably at a 10-year low. The question, of course, is why.

Defining metadata

The first problem with metadata is that most definitions of it aren't very helpful. The US Geological Service defines it as information about data or other information. Various other sites define it as data about data. I'm sure someone has defined it as information about information, too. At the DCMI site ("Making it easier to find information"), a search for "metadata definition" yields nothing useful at all. For something purported to be essential to capturing all of human understanding, this is not a good start.

Where the original intent of the DCMI was to define metadata along the lines of the library's card catalog (using the kind of information that card catalogs typically contain), follow-on standardization efforts have consistently redefined and expanded the metadata problem instead of solving it. This is especially true for multimedia-related metadata, but it's also true for text.

To focus the discussions that follow, here is my personal back-to-basics definition of metadata: *Optional structured descriptions that are publicly available to explicitly assist in locating objects.*

The "optional" portion is key: If the descriptions aren't optional, then they're data, not metadata. The fact that they are structured is also key: Otherwise, you don't know how to apply the metadata to the search. Also, the metadata isn't there by accident: It's been explicitly added (manually or automatically) to help someone, somewhere, at some time find the associated information. To assist in locating objects, the metadata must be public; if it isn't, then it's simply part of a proprietary information management system. (Google's databases aren't filled with metadata, they're filled with Google's private data.) Where is it saved? Hopefully not within the data object itself, because then you can only see it once you've already found it!

In my view, metadata doesn't exist to describe or explain things. (That's called a definition or a summary or an abstract, all of which are valuable pieces of real information.) Annotation is not the process of creating metadata, in general: It's the process of enriching content to make it more useful or more understandable. Expanding the definition by calling everything "metadata" is about as useful as calling everything "data."

Manipulating metadata

With our location-centric definition of metadata in mind, the first question that pops up when reviewing the past 10 years of metadata research is: What did we know then, and how much more do we know about metadata now?

In 1994, we knew (more or less) that:

- Using a set of text-based keywords, term matching could be used to locate information.
- Using similar terms for similar concepts would lead to better search results, as long as those terms were unique.
- By applying fuzzy logic techniques, you could achieve broader matches than by using simple terms alone.
- The same techniques used to find information could be used to filter it (that is, additional terms could be used to exclude classes of undesired results).
- It took a thousand (key) words to describe a picture, and even this was probably a lower bound if you really wanted to describe it.

Now, 10 years later, I'm not sure we've learned that much new about using metadata to locate generalized media except that metadata in the context of electronic processing is probably not nearly as useful as it was in conventional library catalogues.

Creating metadata for text has gone from tedious to insignificant

Here is an interesting experiment: Take a

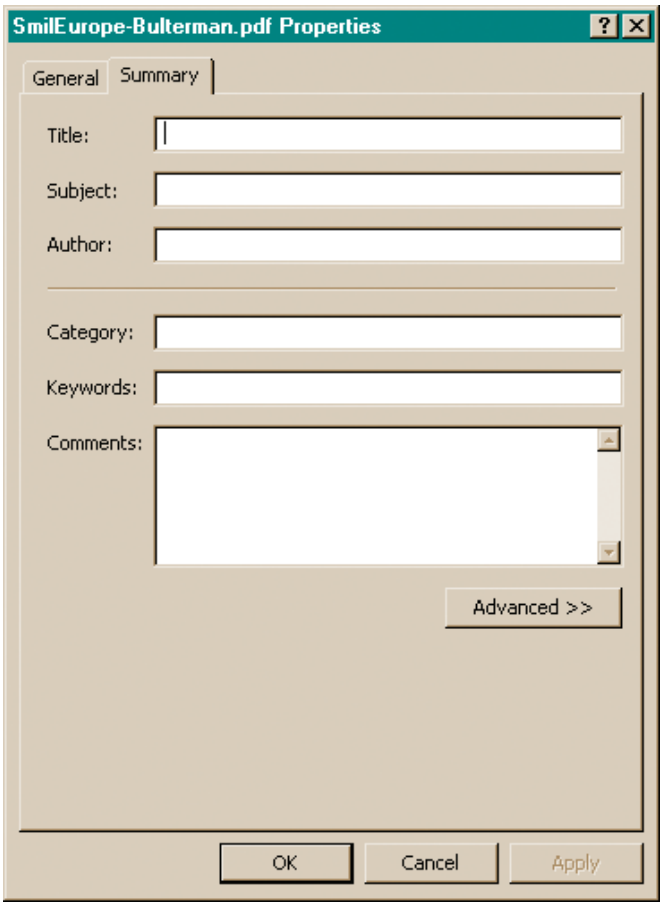

Figure 2. Metadata in practice.

random sample of PDF or PowerPoint presentations and look at the metadata attached to the files. For example, look at the properties box shown in Figure 2, which is from a PDF version of of paper that I wrote for a SMIL conference in Europe. Although most document container formats provide interesting metadata information fields, the utility of these facilities is limited. Either the fields will be empty (as in this example), or they will be filled with information that was attached to a template and is not relevant for the files to which it is attached. (This latter case seems to be especially true for PowerPoint files, in which information from an initial template is carried forth for generations of presentations.)

In a similar vein, here is the total amount of DCMI metadata contained in the W3C's RDF specification:

```
<meta name="rcsid"
    content="$Id: Overview.html,v1.9
    2004/02/10 15:29:30 sandro Exp $"/>
```

The point is this: People don't need to add metadata to text documents if documents are processed electronically. Experience has shown that the contents of text documents can be mined directly using a host of existing information retrieval technologies and that metadata descriptions are often superfluous.

For nontext data—such as video, images, audio, and so on—direct mining is difficult, but exactly at the point that metadata might be useful, manual creation simply doesn't get done because creating useful metadata descriptions (the proverbial thousands of words) is not in the critical path of content creation. Note that the problem isn't the richness or verbosity of MPEG-7 or the Semantic Web—these are cumbersome but could be managed using appropriate tool support. Instead, the real problem is that saying something nontrivial about audio, video, or image content requires too much effort and is probably not relevant because of the many contexts in which that data can be used. (More on this later.) At least MPEG-7 deserves some credit for coming up with a viable motivation for encoding metadata for commercial purposes, but I would argue strongly that the object recognition and link association information is not metadata: It's essential application data without which the home shopping process could not take place.

Creating metadata descriptions is an error-prone task

The stated goal of the Semantic Web is to smoothly interconnect personal information management, enterprise application integration, and the global sharing of commercial, scientific, and cultural data. Figure 3 nicely sums up these goals, all in a single image. Here we see three people eating Danish pastry and discussing commercial, scientific, and cultural topics, with meta-information displayed by an application integrated for the enterprise. And, all of the metadata was created automatically for us. What could be tastier?

Along with its benefits, however, Figure 3 highlights a number of reliability problems with metadata, even when collected automatically.

- Much of the data collected is irrelevant: By adhering to the EXIF standard, the camera manufacturer has dutifully presented us with information that is either not filled in or not relevant. This is more than an annoyance: it could result in the image being rejected in a search.

- Some of the data is wrong: The date at the upper left is potentially useful metadata, were it not off by a full year. (The base date was set incorrectly when the camera was purchased.)

- Some of the data is useless. Although the time stamp in the metadata indicates that it's nearly midnight, there is clearly more daylight visible than one would expect in November. The problem here is not that the data is wrong in an absolute sense—it actually was about 23:45 at my home in Amsterdam when the photograph was taken—but that the metadata generated applied to the place at which the time was initialized rather than the place at which the photo was made.

- The metadata doesn't apply to this picture: The illustration you see is actually a composite of two photographs taken at a pastry shop in Berkeley, California, and then superimposed on a third image (taken a week later) of the Arizona desert. When writing out the final image, Photoshop 7 kept the original metadata (untouched) from one of the source images and stores it in the new composite.

Some would argue that these problems can be avoided or at least detected by automatic processing of the stored data. The camera could use satellite time from a GPS receiver; the individuals' identities could be recognized, their relationships could be analyzed, and their movements could be localized. Of course, if it could be done, we wouldn't need the metadata in the first place—we could simply analyze the image in the context of a particular query and be done with it.

A critical point is that this is not a "stump the stars" example. It illustrates the real problems that exist when one relies on automatically generated information. Coming up with a formal semantic framework is useless if you can't trust the information inside.

Creating metadata: Context-sensitive, culturally biased, and time-variant

Not long after MPEG-7 was announced, IBM made an Alphaworks application available for annotating video fragments. Figure 4 (next page) shows one of the standard screen shots of that application. Along with illustrating a scene-by-scene workflow that would cause even an outsourced army of ontologists to revolt, the application illustrates the limitations of "predictive" metadata—the use of predefined terms that assume potential access-time utility. Other than the obvious problem that, for example, there are more than three types of animals in the world (and that a duck is a subclass of bird, rather than its peer), there are dozens of contexts in which this video could potentially be used. Unfortunately, it's impossible to predict all of the contexts in advance.

Figure 3. Using automatically captured metadata. (Photo courtesy of the author.)

It is generally accepted that how we view information depends on the sum total of the cultural and experiential information that we possess at the time we make the observation. This is illustrated (on a small scale) by an experience I had at a recent talk by a theologian friend. Her work involved an analysis of the motivations of a famous Dutch poet who had contributed a significant number of texts to the standard Dutch hymnal. At the reception that followed, I asked: instead of developing a complex model for the philosophical motivations of this poet, why didn't you simply walk to the third row—where he was sitting—and ask him? She stared at me with that look of disbelief and disgust that only a humanities PhD can muster and replied: "Ninety percent of what people do is motivated by their subconscious. Even if he remembered why he wrote what he wrote, his recollection would be meaningless since he can't possibly really understand his deepest self!" So much for predictive metadata.

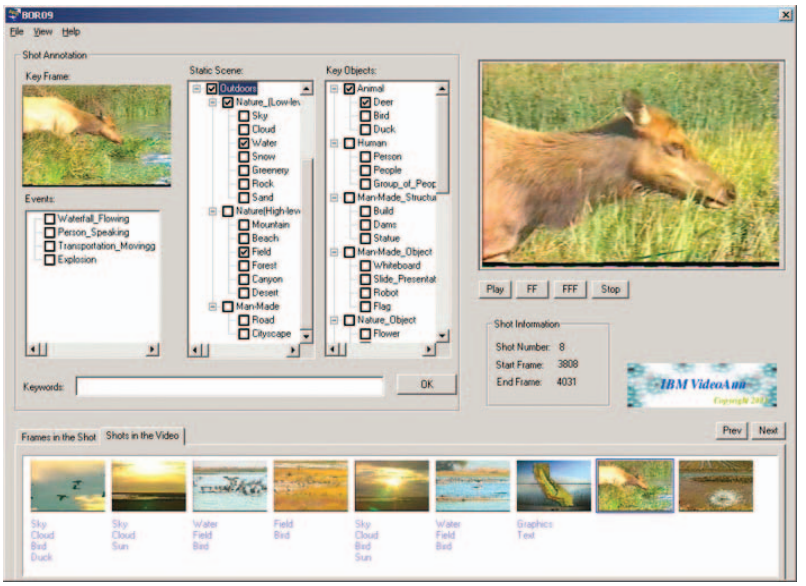

Figure 4. The VideoAnn video annotation system. (Illustration courtesy of IBM.)

Toward a moratorium on metadata

Locating information is a useful activity. It's so useful that it is a problem that has been studied for centuries. For most of that time, there was an implicit assumption that the description of the object being located would, by necessity, be separate from the object itself. Consequently, searches could at best be only indirect with respect to the source information encoding. This is the card catalogue model, in which data is separated from content because a card file is easier to search than a stack of books. During the past decade, it's become clear that for electronic assets, locating text is best done using the text itself rather than relying on metadata, because the context of the search is defined at query time rather than catalogue time. Also during the past decade, it has become clear that locating nontext assets remains an open problem. An open problem—isn't that an ungrateful assessment of all the sincere, hard work done in the past 10 years? I have to admit to feeling somewhat guilty of being so underwhelmed by these activities, especially when you consider that more than 40,000 person hours were dedicated to MPEG-7's morning coffee breaks alone.

I realize that it's good science to transform a problem that you don't know how to solve into one you do. But when the proposed solution for locating nontext objects requires me to create new text descriptions that are largely subject to the same limitations that already have made such metadata obsolete, my "Oh wow" factor is pretty low. The situation doesn't get better when I'm also required to use a baroque encoding structure and restrictive vocabularies that presuppose future use, even when we know that how I view my media today is a poor predictor for how someone else will use it tomorrow.

This is a growing problem. By recent industry estimates, there are now more digital cameras than conventional film cameras—and more telephones deployed with on-board cameras than conventional digital ones. (Nokia alone will make more than 200 million this year.) Finding all the images, audio captions, and video sequences created by these cameras will become more difficult than ever, especially because it is totally unrealistic to expect that the devices' users will spend any time thinking up descriptive filenames or adding extensive captions: They're too busy taking new pictures! Still, this is no time for nay-sayers and doom-and-gloom summaries. The growth of nontext digital media provides a great opportunity to rehabilitate metadata, but the process won't be trivial. (Important processes never are.) In fact, to save metadata, we first need to ignore it.

In this age of simple solutions for complex problems, here is my five-point plan:

1. Issue a joint proclamation that the DCMI, MPEG-7 and Semantic Web initiatives are all Official Successes and are Ready for Business. This will free up incredible amounts of creative resources in discussion groups and alleviate any need to post flame responses to this article.

2. Issue a second proclamation calling for a general moratorium on metadata. Because such metadata is rarely created anyway, this will not have a major impact except that designers of metadata analysis systems will no longer be able to simply assume that all of the required metadata already exists or will be created by somebody else.

3. We're now in a luxurious position: We've solved a problem (because of proclamation 1) that no longer exists (because of proclamation 2). During the moratorium period, we can get back to basics: concentrating on locating objects within a range of mixed-media assets based on context-sensitive queries. The challenge of this work is not only in finding the objects, but in maintaining the discipline to focus only on this core task.

4. Ask public-spirited citizens worldwide to contribute their favorite photos, audio fragments, or personal videos to create a culturally diverse corpus of 1 million nontext media assets. The only restriction is that it contains no predictive metadata, other than perhaps required citation information and a URI. Figure 5 can get the ball rolling.

5. Embark on a multimedia content differentiation competition that will allow a comprehensive but limited set of objects to be identified: people, places, objects, and life events (births, weddings, deaths, and so on). The catch: Any contributed techniques must apply to multiple encoding formats (pictures, video, audio), and it must include a user interface for managing media classification. Any desired technique or technology is permitted—such as template matching, geometric modeling, or direct object labeling—as long as it doesn't rely on preexisting metadata supplied by the author/creator (because these descriptions typically don't exist). Internally, any information structuring and labeling approach may be applied.

At the end of the competition period, a "bake-off" will be held to separate the wheat from the chaff. Each entry will be required to provide a solution for locating a class of objects in the mixed media dataset. (A sample search request: Find all of the media objects containing references to strawberry Danish pastry, consumed in the wintry desert.) The key to any approach is its ability to support multiple media within one system. A solution for images isn't nearly as useful without a solution for audio, just as a solution for English-only results is only useful if the search request specified that as a requirement.

Figure 6 provides an example of a nontextual approach. Here we see a system for organizing digital photographs in which all of the instances of a particular person can be found based on face recognition rather than keyword matching. (The fact that image processing algorithm also included George Washington and the tail of a Cessna illustrates that even incorrect results can be interesting and potentially useful.) Of course, this is only a start. All media should be searchable, such as the visual baking instructions shown in Figure 5. But Figure 5 does demonstrate that, for nontext data, the future is not text based.

At the end of the process, the nontextual systems would be compared with a conventional metadata effort that would use predictive labeling of objects in the base data set. The evaluation of all systems (automatic and manual) will include not only the number of useful media objects located, but also the overhead required to define, maintain, and extend the information used for internal bookkeeping. The goal of the entire process is to determine which approach really provides a useful basis for locating content in mixed-media for context-sensitive queries. It could be that the on-demand processing of nontext information requires extensive text labeling, or it could be that like text itself, such labeling will play only a minor role. Hopefully, the process will shed more insight into whether predictive metadata really is the greatest thing since sliced bread, or if it is a stale concept whose "best before" date expired 10 years ago. **MM**

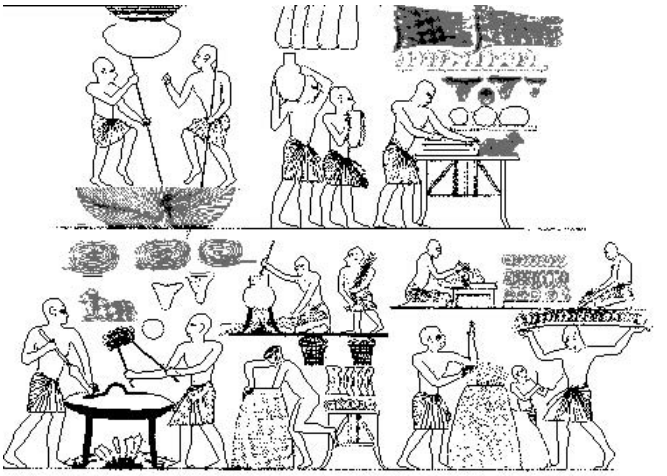

Figure 5. Illustrating a 4,600-year tradition. (Illustration courtesy of Artemis Verlag.)

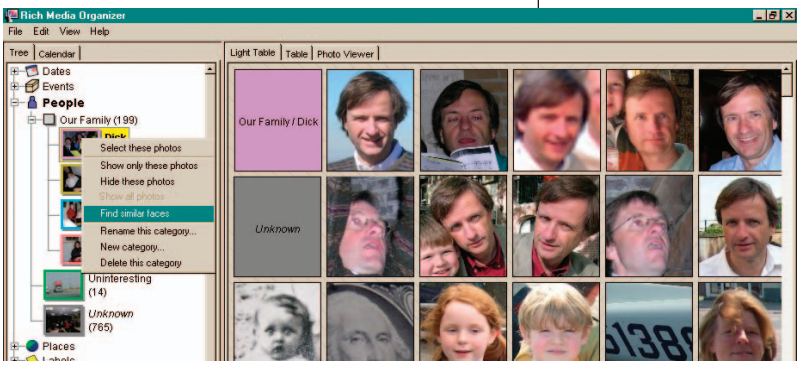

Figure 6. Locating by example using the Rich Media Organizer developed by FXPal. (Illustration courtesy of the author.)

Readers may contact Dick Bulterman at CWI, Kruislaan 413, 1098 SJ Amsterdam, The Netherlands. Dick.Bulterman@cwi.nl.