From *Proceedings of the International Broadcasting Convention*, Amsterdam, Holland, September 2002.

DRINKING FROM THE FIRE HOSE:

HOW TO MANAGE ALL THE METADATA

Sharon Flank (DataStrategy Consulting) and

Serena Brinkman (Blue Order)

ABSTRACT

Media assets can be managed wisely through the careful use of metadata and smart search options, including natural language search. Archives should be crafted with long-term objectives in mind, including reuse and repurposing.

INTRODUCTION

Over time, media-rich organisations have come to realise the value of their media assets. When viewed as large units, such as movies and TV programmes, there is some reuse and repurposing potential: movies can be re-released and programmes can be rerun. When considered as smaller units, however, the reuse potential explodes. There are numerous possibilities to repurpose tropical sunrises, car crashes, and so on. When the cost and difficulty of obtaining the footage is considered, reuse becomes an extremely attractive option. If you already have a shot of a polar bear devouring a seal, there is no reason to outfit a team to risk their lives to obtain additional footage. And some footage cannot be re-shot at any cost: volcanoes erupting, political and historical events, and so on.

Once your organisation embraces the notion of media reuse, a new challenge becomes evident. You must maintain a media library, and someone (or preferably many people) must be able to catalogue its contents so that other people can find what they want, on a deadline. To complicate matters, sometimes the searchers do not know exactly what they want. Sometimes they know, but the characteristic of the existing footage that matters to them is not the characteristic that the cataloguer focused upon (e.g. the lava or the bloody seal rather than the spewing volcano or ravenous polar bear).

Until recently, media librarians relied on home-grown cataloguing schemes for tape and photo libraries. The advent of digital media brings forth a new set of possibilities, and highlights the utility of software for *digital media management*. Digital media management (DMM, also referred to as digital asset management or media asset management) incorporates media archiving, a workflow for media projects including review and approval, and collaboration tools for creative development.

Now the key issues in managing a media archive are transformed: how do you best foster reuse and repurposing, while minimising the cost of maintaining the archive? How do you help creative users find the most appropriate files, quickly and painlessly, improve them, and archive the improved version for others to access? How do you ensure that the latest versions and the latest logos are used for publicity and branding? How do you restrict access so that outsiders can use approved versions while insiders can collaborate to choose the next generation? Many of these matters were addressed in an ad hoc way before the advent of DMM; now they can be managed without relying on an individual's personal knowledge and processes.

The key element in cataloguing assets is the ability to find them again. A picture may be worth a thousand words, but it isn't worth a penny if you cannot find it when you need it. In this paper we consider how metadata, correctly used, preserves and enhances the monetary value of media assets. What is metadata? It is information about the asset: who created it, when, perhaps how, and what the pictures are about. Metadata may include literal information (e.g. the polar bear devouring the seal), and it may include more impressionistic information as well (e.g. carnivore, gore, violence, "nature red in tooth and claw").

1) Creating metadata is generally time-consuming and expensive. You should make every effort to do it right and do it once. We will explore what exactly this means later in the paper, but certainly the first step is to figure out what you want your metadata to do. First and foremost, the metadata must facilitate reuse by making it possible for people to find things. Second, the metadata should be as inexpensive as possible to create. Third, the metadata should be designed to last, to support eventual new users and new uses.

Metadata must:

Support search

Be inexpensive to create

Be timeless

Figure 1 – Metadata must facilitate inexpensive reuse, preferably over a very long period.

Metcalf's Law

-"The value of any asset increases by the factor of the number of people using it."

FILE HIERARCHIES AND CATEGORISATION ARE NOT ENOUGH

Organisations creating digital media archives often begin with hierarchies and categories. If only we create enough categories, they reason, we will be able to store and retrieve everything. As the task proceeds, they change their view: if only we can create the *right* categories, then everything will work perfectly. Unfortunately, the effort is doomed. First, consider the file hierarchy on your own PC. Do you always know exactly where everything is? Or do you resort to that little *Find* icon to help you track down where you have put your files? Now consider looking for files on someone else's computer – or on a computer owned

by a group of people. Either files will get lost, or the group will spend endless time in meetings trying to agree on where to put things. If you could simply search in plain English, you could avoid the problem entirely.

According to Gistics, it takes, on average, 15 minutes to find a file in a file hierarchy if you do not know where it is. If you use a natural language search on the metadata, it takes less than three seconds. Assuming two searches per day, this capability alone could save more than 100,000EUR per year in a department of 20 people.

If a file hierarchy approach is inadequate, why not try assigning categories, and then filing the assets in the appropriate category? Unfortunately, categories end up with the same problem: they do not scale up. At about 1000 assets, it becomes clear that assigning categories consistently enough to facilitate search is equivalent to signing up for constant meetings and training sessions. Worse, it is inevitable that you will get to a stage where you need to recreate categories and reassign assets. Given the time and expense, you do not want to re-categorise what you have already catalogued. You need an approach that means you do not ever need to re-catalogue.

Keywords

If hierarchies and categories are inadequate, why not assign keywords to files? It makes sense to assign metadata to files, and then create an index. Keywords are a better approach than categories, because a file can then be indexed in multiple ways. But keywords can present another guessing game problem. How can you be sure that I have used the same term for cataloguing that you are now using for search? Do we need to have more of those coordination meetings? The terminology issue is especially problematic with international organisations. Consider *boot* and *trunk*, *rubbish* and *garbage*, and so on.

We have concluded that the idea of metadata is promising, and that hierarchies and categorisation do not work. Where does that leave us? What else can we use to search?

Natural Language Search

What is natural language search? It is more than simply allowing you to leave the funny parentheses and capitalised AND and OR out of your search. There are two key elements in NL search. First, a system must figure out what is important in the metadata. If you search for *tiger*, you want to find big furry felines, and not a *tiger butterfly*. To use George Miller's example, if you look for *Venetian blind* , you will not be happy to see a *blind Venetian*. Note that this capability goes beyond proximity search, and that restricting word order is not quite the right criterion, since *yellow spongy coral* should in fact retrieve *spongy yellow coral*. Second, search terms should be expanded to include synonyms and related

words, so that you do not have to play a guessing game as to what words were used in cataloguing. A search for *kids playing at the beach* should also find *children frolicking in the sand*.

Our studies on the PictureQuest photography portal (www.picturequest.com) show that it is six times cheaper for the cataloguer to write what they want without having to check what the company style guide is. If you can simply describe a file in plain English, it costs far less both to do the cataloguing and to train the cataloguers. Optimally, this fuzzy search capability goes beyond synonyms, to include subtypes, parts, and other relationships.

STANDARDS

We all know that standards-compliant software is better. Why? What do we want standards to accomplish for us? We want them to save us time and money. Optimally, good use of standards can make these things easier:

users learning a new system, since the new system "feels" standard

integrating new capabilities as "plug-ins"

data integration, as when AOL merges its media with Time-Warner, or Exxon merges with Mobil.

We need to step back and think about tradeoffs. It makes sense to develop a shortcut for a time-consuming, oft-repeated process, but it may not be useful to spend time on a shortcut for something that does not happen very often and is not particularly painful. Thus you might spend time writing a macro so you did not have to type "digital media management, also known as digital asset management or media asset management" over and over. But if you only used the phrase once every few months, it would not be worth your effort to create the shortcut. Similarly, we should focus our standards efforts on the activities that are both frequent and challenging. In general this means targeting the user's interactions with the system. Users may spend hours every day searching through metadata, and we owe it to them to make that process as painless as possible.

We can finesse the problem of teaching users how to interact with unfamiliar systems. We can follow the Windows model of making everything feel approximately the same, or we can make the user interface so clear it can be learned quickly. There is a third choice as well, and that is making the UI tailorable: if you can rearrange the screen, rename the buttons, and change the work pattern to meet your own needs, then you can create your own intuitive and efficient work tool.

As for integrating new capabilities, it appears increasingly likely that XML will be the foundation of integrations for the foreseeable future. If your software can specify an XML interface, and mine can too, we can hook the two systems up quickly, and pass information

back and forth seamlessly.

Data integration is more challenging. Legacy systems offer a morass of competing standards and non-standards. Archivists made up whatever terms they wanted to, and stored them in fields with names they picked at random, and which may have changed over time. Often they began cataloguing before computers were invented, and assumed that their work would be processed by human intelligence, not a machine. Now what? Do we start over, recataloguing so that every metadata element follows some new standard? What if we guess wrong, and have to do this all again every decade or so?

There is a simpler solution: we can choose to allow variation inside the fields, and even in the field names. The process we are trying to simplify is data integration, i.e. merging libraries. This process occurs relatively rarely – perhaps once a year. When it occurs, we can map field names, e.g.

Old-field-*Summary*	maps to	New-field-*Description*
Old-field-*Place*	maps to	New-field-*Location*

Why would we subject ourselves to this effort, rather than insisting on a standard? Organisations will want to tweak the standard (e.g. *We need to include flower names, since this is a botanic film house!*). This semi-manual mapping process takes an hour or two, hardly enough time even for the first meeting of our standards body. And, until field name standards are widely adopted and flexible enough for everyone, it is quite adequate.

There is, however, a difference between standardising the metadata fields themselves and standardising the terms that go in them. Standardising fields is a laudable if difficult effort; standardising terms, as we shall see, is more controversial. The Dublin Core Metadata Initiative, discussed below, attempts to standardise fields, and even to abstract one level beyond that, to standardise how metadata fields are defined.

The Dublin Core Metadata Initiative develops interoperable online metadata standards. DCMI's activities include consensus-driven working groups, global workshops, conferences, standards liaison, and educational efforts to promote widespread acceptance of metadata standards and practices.

The Dublin Core is lightweight and extensible, and its goal, promoting interoperability, is certainly a good one. However, DCMI is still burdened by meetings in order to get its message out, and by controlled vocabularies. Here are some sample elements, just to show how it is used to define metadata fields in a flexible, general way.

Each Dublin Core element is defined using a set of ten attributes from the ISO/IEC 11179 standard for the description of data elements. These include:

1) **Name** - The label assigned to the data element
2) **Identifier** - The unique identifier assigned to the data element
3) **Version** - The version of the data element
4) **Registration Authority** - The entity authorised to register the data element
5) **Language** - The language in which the data element is specified
6) **Definition** - A statement that clearly represents the concept and essential nature of the data element
7) **Obligation** - Indicates if the data element is required to always or sometimes be present (contain a value)
8) **Datatype** - Indicates the type of data that can be represented in the value of the data element
9) **Maximum Occurrence** - Indicates any limit to the repeatability of the data element
10) **Comment** - A remark concerning the application of the data element

Fortunately, six of the above ten attributes are common to all the Dublin Core elements. These are, with their respective values:

Version: 1.1
Registration Authority: Dublin Core Metadata Initiative
Language: en
Obligation: Optional
Datatype: Character String
Maximum Occurrence: Unlimited

The elements themselves look like this:

Element: Title

```
Name:        Title
Identifier:  Title
Definition:  A name given to the resource.
Comment:     Typically, a Title will be a name by which the resource is
             formally known.
```

Element: Creator

```
Name:        Creator
Identifier:  Creator
Definition:  An entity primarily responsible for making the content of
             the resource.
Comment:     Examples of a Creator include a person, an organisation,
             or a service.
             Typically, the name of a Creator should be used to
             indicate the entity.
```

Element: Subject

```
Name:        Subject and Keywords
Identifier:  Subject
Definition:  The topic of the content of the resource.
Comment:     Typically, a Subject will be expressed as keywords,
             key phrases or classification codes that describe a topic
             of the resource.
             Recommended best practice is to select a value from a
             controlled vocabulary or formal classification scheme.
```

Element: Description

```
Name:        Description
Identifier:  Description
Definition:  An account of the content of the resource.
Comment:     Description may include but is not limited to: an abstract,
             table of contents, reference to a graphical representation
             of content or a free-text account of the content.
```

There are other metadata standards efforts as well. The BBC has defined a Standard Media Exchange Framework (SMEF™) for media management. The SMEF Data Model (SMEF-DM) consists of a set of metadata definitions for the information required in production, distribution and management of media assets, currently expressed as a data dictionary and set of Entity Relationship Diagrams.

SMEF is the intellectual property of the BBC, designed for their internal purposes. They have decided to share it with the community, presumably both in order to improve general standards and also to make it easier for vendors to conform to BBC requirements.

PRISM (Publishing Requirements for Industry Standard Metadata) is an extensible XML metadata standard for syndicating, aggregating, post-processing and multi-purposing content from magazines, news, catalogs, books and mainstream journals. It builds on the Dublin Core, includes extensive subject description capabilities, and adds rights information. It helps companies move away from proprietary vocabularies into a common standard, but it is still a controlled vocabulary standard requiring cross-company agreement, updates, extensions… and meetings.

MPEG-7, the "Multimedia Content Description Interface," is another emerging standard, incorporating Description Tools and Description Schemes. Here, too, coordination is a major challenge, and the standard becomes more a framework for a framework.

Standardisation, then, should not be viewed uncritically. Does your approach address problems that are common and expensive? Does it do so economically? Does it put the business goals of your organisation first (rather than "let's all hold hands until we agree")? If you do standardise, you may choose to stop short of standardising on a controlled vocabulary, lest you end up back in a guessing game about which terms work for searching. Keep in mind that your systems must respect both the human memory and the human need for a comfortable user interface.

Finally, make sure your standardisation approach keeps the long-term view in mind. If you build an archive that is designed to last for 100 years, who will end up as its owner? The corporation that inherits your archive should not have to re-do the metadata. With a minimum of effort, they should be able to map it to a schema that is useful in their business.

SOURCES OF METADATA

Where do you get your metadata? What if you already have legacy metadata, and it

consists of keywords? Natural language retrieval can still improve your access to your data, since you eliminate some of the guessing game. Did we use *car* or *automobile*? Was it *boot* or *trunk*? Natural language search makes retrieval smarter on those keywords, boosting recall by 10% or more.

There is no reason you should have to create all of your own metadata from scratch. There are numerous existing sources of material describing content. These sources include programme notes, speech recognition, closed captions, and production notes. The U.S. has recently begun to require that a portion of television programming be described for the visually impaired, and these descriptions are an excellent source of thorough metadata.

What about futureproofing? If you choose natural language search over categories and controlled vocabulary, how do you know it will stay around? When you select natural language as a technology, you are not buying into the proprietary ideas of a single company. Natural language processing as a science is 40 years old, and all of the work done in it can be fed back into retrieving the metadata that you are now creating. All the technology for voice recognition, machine translation, automated question-answering, and speech translation, use the same techniques that will improve search retrieval for metadata. If you create your metadata so that people can read it, you can expect natural language processing to keep improving so that it can understand language more as people do.

Creating your metadata in plain human language (now English, soon other languages as well) is an investment in a technology that will improve as time goes on. Real English is unlikely to disappear. You are betting not on a single technology but a set of technologies. Choosing natural language as the basis for your search is not like deciding to record on Betamax. There was nothing in the real world making people choose Betamax, but there is the weight of hundreds of millions of people who already know English. Therefore the risky choice is to choose anything other than plain language for your metadata.