

The Digital Asset Management Specialists www.thirdlight.com

Metadata: a simple guide to describing digital assets for search and retrieval

About the author

Michael Wells is the founder and Managing Director of Third Light, a digital asset management software company based in Cambridge, England. Founded in 2002, Third Light is a leading innovator in digital media library systems for business users, including search and metadata technology.

Abstract

This paper looks at the purpose of adding metadata to digital media, and the evolution of keywording into XML formats for metadata. Simple models for tagging assets and structuring metadata are discussed, with an emphasis on making effective use of the underlying technology and user interfaces.

| Introduction | 2 |
|--|----|
| What Is Metadata? | 2 |
| Which Standard To Use? | |
| The Mechanisms Behind Search | 4 |
| Basic Metadata: Creating Captions | |
| Basic Metadata: Creating Keywords | 6 |
| Example Files | 7 |
| Additional Metadata Fields | 8 |
| Custom Metadata Fields | g |
| Bulk Editing | 11 |
| Workflow And Engaging Your Staff | 11 |
| Browse Versus Search: Using Folders | |
| Smart Folders | 13 |
| Searching Non-textual Content: Hidden Gems | 13 |
| Summary | 14 |
| Acknowledgements | 14 |

Introduction

Building a searchable library of digital media is an everyday problem for businesses of all sizes. We can see that each extra file we store has some value, but we need to be able to connect that value to a future user by making the entire store of media searchable. The problem goes as follows:

- Digital media is prolific. Digital technology is accelerating the quantity of data in circulation and digital marketing has an unending appetite for new and engaging content.
- We are naturally keen to ensure that this digital media is searchable to encourage reuse and streamline our work, but we need textual descriptions to unlock this potential. This text simply does not exist by default.
- To succeed, the tagging process must be efficient and must be proportionate to the value of the files. In other words, it must provide a return on investment and not be unduly time-consuming or centered on one person.

Balancing the invested time and the return on this time is part of what I want to explore. This has far-reaching consequences for how we store the files in the first place, and how we minimise the manual effort and overheads that a large digital media library entails.

Perhaps you are faced with a back-catalogue of untagged content, and are looking for advice on where to start; or perhaps you are looking at metadata as an opportunity and wish to find a formula for success. The sort of questions I hope to address include:

- What are the common metadata fields and how should they be completed?
- How do you tag a large number of files?
- How can the labour involved be minimised, and re-work avoided in the future?
- Is there a way to use software to marshal the tagging of new files?
- What are the pitfalls, and how do I avoid them?

Firstly, let us look at the underlying system, metadata, which allows your digital media to be searched.

What Is Metadata?

Metadata means 'data about data'. For example, the ISBN database contains metadata about books. In terms of digital content, descriptions of images and videos (such as keywords, captions and copyright information) are metadata, and the value of supplying this information cannot be understated. Here is why.

Some assets (like web pages) contain text that can be easily consumed by search engines, but the the visual content of a picture cannot be represented as text automatically. Without accurate metadata, pictures are only ever browseable by the most basic attribute of all – the filename, itself now normally just a serial number from a camera – and are destined to be left untouched or manually browsed based on a very broad category such as the folder in which they are stored. This is the file server model or "shared network drive" model.

Metadata augments the value of this raw data dramatically. It is the cornerstone of the digital asset management industry and provides the solution to search, retrieval and management of visual content.

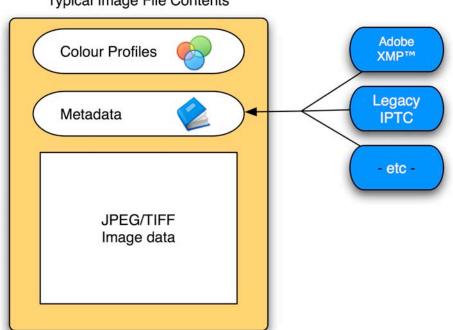
Which Standard To Use?

It is worth stepping aside for a moment to take a closer look at which metadata standards are available to you.

The most common approach today is to embed an XML (Extensible Markup Language) header into each asset, containing all of the metadata about the asset.

Embedding the metadata in this way is a good idea. It means that the information travels with the file itself when it is transferred between users, and this metadata will not be inadvertently separated from the asset. Compared to proprietary solutions which store metadata separately perhaps in a spreadsheet – embedded metadata is more dependable and does not create an unnecessary legacy or vendor lock-in. It is also elegant in terms of interoperability since tools supporting the same standard can read and write the same content.

In the diagram below, a simple representation of how embedded metadata coexists with the image data is provided in the case of an image asset.



Typical Image File Contents

For completeness, the diagram shows that colour profiles are also stored with the image: in fact they are a form of metadata, too.

The idea of embedded, descriptive metadata is well-established, and was quickly adopted in the late 1980s by international news agencies. The original standard, still widely used, is referred to as IPTC, named after the organisation which devised the standard: the International Press and Telecommunications Council. Many vendors implemented this standard, making it almost ubiquitous. Other standards exist, notably Dublin Core which takes a more universal approach towards metadata in text, music and video.

In practical terms there is presently one standard that stands out. In 2001, Adobe Inc. announced a new open standard called XMP – Extensible Metadata Platform – and Adobe products started to transparently write documents with XMP embedded. As its name suggests, XMP is based on XML and can be extended to store customised information.

The existence of XMP has gradually allowed the industry to move in a single direction, which is of

great practical benefit. XMP is now the key standard for writing metadata and is supported by all major graphics software and digital asset management software, so it is the standard which I will single out.

The Mechanisms Behind Search

The apparent simplicity of typing in a search term tends to conceal some of the important details about how search engines work. Knowing some of this detail is informative. Here are some important points to consider.

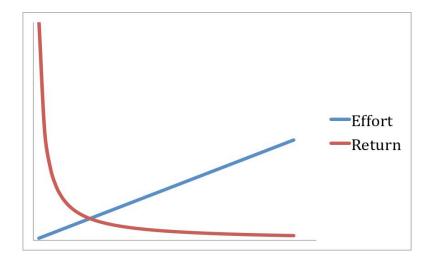
The database contains an index of all of your metadata, and can search this information based on patterns, words or other more innovative heuristics. A modern digital asset management system will usually:

- Ignore upper- and lower-case, instead treating both as equivalent;
- Offer partial matches (for example, searching for 'car' will usually return 'cars' but could also find 'caramel');
- Increasingly, offer some intelligence for matching like-sounding words, for instance 'colour' and 'color', or "did you mean" suggestions;
- Prioritise precise matches, but also provide a large number of possible matches (therefore ensuring that 'caramel' is scored lower than 'car' or 'cars' if you search for 'car');
- Offer a simple and advanced search facility, to balance convenience against precision according to the situation. For example, this may include specific field filtering or dropdown selections from a known vocabulary.

For comparison, let us also summarise how human beings usually work:

- We prefer typing short words, often just one keyword, into a simple box;
- We tend to search by broader concepts first (eg. "holiday", not "Madeira Island");
- We work by refining our searches in stages:
- We misspell or mistype words surprisingly often, but usually get the 'sound' right;
- We avoid advanced search forms if we possibly can, except simple checkboxes;
- We don't think about plurals, case or ambiguity very carefully.

Between the human being and the database, your influence as the author of the metadata is vital. You can pre-empt your user if you put yourself in their shoes, and then shape your metadata so that the database can be effective. This is your first opportunity to save time: don't enter too much metadata, just enough to catch the majority of the meaning. If each extra keyword adds just half as much meaning as the last keyword, for example, you will obtain a curve that falls steeply off with each additional keyword.



This shape of curve should make you think "diminishing returns". In practice, I've found that more than 10 keywords is usually unnecessary, for example. You can save a great deal of time by studying the behaviour of your users, and most powerful digital library solutions will allow you to browse a history of actual search terms used for this reason.

Basic Metadata: Creating Captions

Captions are short, descriptive sentences which give your users an accurate statement about the file's contents or meaning. Captions are factual and will normally be seen by the end user underneath or near to the file. A good caption contains carefully selected words, ideally remaining focused on the actual contents of the file to provide the highest quality of search results.

You may find the following approach gives you a sensible framework for photographs in particular.

- 1. **Name** any people in the photograph, and state their roles. For example, "*Prime Minister Tony Blair and Chancellor Gordon Brown*".
- If the photograph is a stock photograph, it can help your clients select appropriate images
 if you include the age group and gender of the subjects, For example, "Male student", or
 "Retired couple". Beware of storing (or guessing) sensitive information about ethnicity of
 age inadvertently, though.
- 3. A surprising amount of detail is uncovered if you describe the **subject and its attributes**, or layout. For instance, "Small yacht with sails against storm clouds in rough seas".
- 4. **Location** information. If you consistently include the location of photographs, a geographical context is added. For example, "Students graduating at the Senate House, Cambridge University".
- 5. **Colours and visual attributes** can make it easier for designers to find a match. "Large blue beach ball" or "Black cat with green eyes".
- 6. **Timing information.** Whether this is necessary depends on the subject, but you could include reference to the season, time of day or a specific instant. Examples include "Autumn forest", "The London Eye by night" or "September 11 2001". In general, all photos will have a date stamp from the camera, which will be searchable too, so only add timing data to the caption if it clarifies or improves the meaning of the caption.
- 7. The meaning or **purpose of the image**; an indication of its significance. For example, "The first Eurostar train leaving Waterloo Station, London, for Paris".

It is possible to produce a caption which still reads clearly, while being laden with information. This is the art of caption writing. However, it is not just a matter of following rules. The purpose of the caption is to assist a search but also to describe the picture in a coherent way.

If you find that you want to add more information but cannot make the caption flow correctly, or want to name concepts, you should move onto keywords.

Basic Metadata: Creating Keywords

Compared to writing captions, keywording is more logical and should be simple to do. Keywords are usually single words, or very short phrases, which are attributes of the image. Between **five** and ten keywords is a sensible aim.

When keywording images, you can choose to adopt a **controlled vocabulary**, which is a list of pre-defined keywords that you have consistently applied (in a sense, making keywords into categories). Digital asset management software will usually allow keywords to be hyperlinked, giving surprisingly helpful results that will surpass simple "folder browsing".

Another structure you may find helpful is a **taxonomy**, which is a tree-like structure of keywords which becomes more specific as you go down the tree. A taxonomy is often represented left to right in this form: "Europe -> Italy -> Rome -> Colosseum".

I have found that using either a controlled vocabulary or a taxonomy is all-but essential if you work in a team, since the likelihood of different approaches will grow with each additional person working with you. Using controlled vocabularies also lends itself well to using non-free text inputs (like dropdown options or checkboxes). These are usually quicker to complete and I will discuss them later.

Here is a successful keywording strategy:

- **Index by concept and by subject**, starting broad and becoming more specific. eg. "*sport, cricket, england, ashes*". Remember, users search iteratively.
- If you can, adopt a **controlled vocabulary or taxonomy**, but be prepared to extend this list very liberally at first, based on searches that your users are actually performing.
- Include synonyms and prefer to use plurals where possible, since this increases the chance of a match. Remember that many plurals will catch singular terms automatically, like 'table' and 'tables', although some plurals (eg. 'city' and 'cities') will need to be specified separately.
- **Do not use prepositions** or conjunctions like *at, from, to, with.* These do not add information to your picture and will normally be discarded by databases during searches. If you only include concepts, you will save yourself time.
- Do not add tenuous, long lists of keywords which are for concepts that the picture does not contain. It sounds obvious, but this has become quite common practice in certain stock libraries and has caused considerable tension (it is akin to polluting a search engine). Adding vaguely related keywords is not the same as adding closely related keywords. It degrades the quality of all search results and wastes time and money.

Lastly, a common mistake when keywording is to repeat the same information that is in the caption.

If you use a controlled vocabulary, it's certainly worth picking out the appropriate words to support hyperlinks, but the database will probably have already 'hit' an image if the information is in the caption, so you should aim to be brief. Instead, try using keywords to include a slightly broader list of concepts than you can achieve with the caption alone.

Example Files



Caption: Crew members of the Royal National Lifeboats Institute (RNLI), wearing helmets and floatation vests, conduct a training exercise at sea to lift an rigid inflatable boat into the water using an orange crane fitted to the Severn Class lifeboat "Katie Hannan" in August 2004.

Keywords: RNLI, portrush, lifeboats, sea, ocean, coastguard, cranes, maritime



Caption: Close up of Fluid RI-1393 preparation being injected onto a blue 10-sample ImmuGlow slide for microscopy analysis, using a Cambridge Life Sciences Ltd model C robotic tray handler.

Keywords: bioscience, model C tray handler, RI-1393, slide, needle

This image is an example of a very specific, technical caption which is designed for the benefit of knowledgeable members of staff.



Caption: Mighty Minis championship, Snetterton, Norfolk. Simon Wood driving his silver Mini Cooper around the "Esses" with three other competitors, Paul Morris, Gareth Neville and Steven Illingworth in pursuit.

Keywords: racing, cars, motorsports, simon wood, 43, snetterton, mighty minis

Additional Metadata Fields

Having concentrated on two very popular fields (Caption and Keywords) we can turn our attention to other possible fields you may wish to complete, or fields you might want to create in order to finesse your metadata model.

When you are filling in metadata, depending on your software you will probably have the opportunity to supply some other information such as creation notes, copyright, special instructions, location details and so on. These are generally from IPTC. Other types of asset such as video may have different metadata models or may also offer IPTC fields. This depends on the vendor.

Here is what these additional fields are designed to achieve.

- Object Name the name of the subject item, eg "Ben Nevis"
- Category Three characters identifying a category notably, this is not intended to be a free text field but is part of the IPTC category standard
- By-Line / Author the author or photographer's name, eg. "Joe Cornish"
- By-Line Title / Author's position the author or photographer's position, eg. "Editor"
- Credit Credit information for the image not necessarily the Author, eg. "Royal Photographic Society"
- Source Identifies the original owner / creator who provided the source, eg. "National Trust Picture Library"
- Copyright Notice The copyright notice to be displayed with the image, eg. "Copyright 2005 Third Light Ltd"
- Date Created a date, normally in the form 'YYYYMMDD', eq. 20060123
- City eg. "Moscow"
- State eg. "New Jersey"
- Country / Primary Location Name eg. "South Africa"
- Headline Similar to the caption (description) field, used as the title for the article, eg.

"Man rescued from sea by RNLI crew".

- Special Instructions eg. "No reproduction other than internal magazines email publications team for exceptions".
- Transmission Reference a code identifying where the object was transmitted from, eg. "UK5". This field is highly specific to agencies.

Notice the distinct emphasis on press and news agencies. Although it is helpful that these fields are generally interoperable without configuration between different systems, the fields themselves may not be ideal for your purposes, in which case you should create custom metadata fields.

Custom Metadata Fields

A custom metadata field is simply an additional label, against which the values can be stored and independently searched. The input type depends on the sort of information being stored, and the use of a vocabulary allows this information to be restricted to specific, pre-determined values.

A customised metadata field generally has several parts, as follows:

- A label
- An input type (for example text, dropdown, checkbox, tree hierarchy, date)
- · Optionally, a vocabulary of possible values

Removing free text solves several problems. Firstly, personal language preferences or awareness of the subject matter can dramatically change the sort of words used in a free text field, so free text input is fairly weak if your hope is to obtain consistent tagging. Secondly, entering free text is relatively slow, while structured input is normally much quicker to complete as it provides a point-and-click tagging interface.

Example 1:

Field name: Aircraft Manufacturer

Input type: Dropdown

Vocabulary: Airbus, Boeing, Cessna, Dassault, Lockheed [etc]



In this case, the expected input type would allow a user to tag exactly one manufacturer for an aircraft. The strength of a dropdown input is that it only allows a single option to be chosen. In some cases this is too formal: if there were two different aircraft in the same image, for example, then a set of checkboxes would be required (see below).

A similar input type to a dropdown is the radio button, named after the channel selector in early radio sets, again requiring at least one choice but better suited to very small lists such as Gender (male, female).

Example 2:

Field name: Product Features

Input type: Checkboxes

Vocabulary: FM Tuner, DAB Receiver, CD Player, iPod Dock [etc]



By using checkboxes, it is possible to select more than one feature (but only those features from the vocabulary). This is often a good choice as it allows additional breadth, perhaps lending itself well to fairly small category lists (up to 50, say).

Example 3:

Field name: Office Location

Input type: Tree

Vocabulary (showing input selections):

- ▼ □ Europe
 - ▼ □ United Kingdom and Northern Ireland
 - ▼ < London</p>

 - . ☑ Portsmouth Manufacturing
 - □ Manchester
 - □ Leeds
 - □ Edinburgh
 - ▼ □ France
 - □ Paris
 - □ Nice
 - □ Toulouse
 - ▼ □ Germany
 - □ Munich
 - □ Frankfurt
 - Italy
 - □ Rome
 - □ Bologna
 - ▼ □ Spain
 - □ Barcelona
 - 🗆 Zaragoza
- □ Asia
- □ North America
- □ South America
- □ Oceania

Tree structures offer a very rich interface to structured data, which can be helpful when there are too many items for a flat list, and the precedence (hierarchy) of the data is fundamentally meaningful. This is sometimes also a way to combine fields, allowing us to reduce Continent, Country and City into a single field in this example.

To use a tree structure well requires more planning than other types of metadata input, but will guide a new user more effectively if they are asked to tag an asset.

Bulk Editing

Avoiding re-entry of the same details is an important opportunity to save time and improve consistency.

There are many situations which lend themselves to bulk editing. Firstly, many sources of media are inherently from the same photo shoot, brand package, event and so on. Secondly, content which is being migrated from CD-ROMs or file servers may have been categorised into folders within which your files share many attributes.

Depending on your digital asset management tool, one or more time-saving tools should be available:

- 1. Use 'copy from previous' tools. This approach lends itself well to sequential files in the same subject area with minor differences. Normally, it is used to obtain fine detail while also saving time and retaining consistency.
- 2. Use batch editing tools. Typically using a system of checkboxes or selecting multiple files with the mouse, you can trigger an editing tool that will give you the ability to tag all of the selected files in one go. This is a particularly useful technique since you can judge by eye which files to group and tag in each batch. It is a sensible compromise between fine and coarse tagging.
- 3. Set up a folder as a metadata template. This means that any file placed in the folder in your digital asset management software will be automatically tagged with a standard set of information for that folder. One benefit of this approach is that it would be a suitable starting point for all of the files in the same folder, and will be applied consistently even as different people become involved.

Workflow And Engaging Your Staff

In digital asset management, the phrase 'peer sourcing' is particularly relevant. When a new file or collection of files is uploaded, you will want to capture the specialist information and awareness of the content. This is most probably known to the person doing the uploading.

Let's start with the basic problem of a shared file server model:



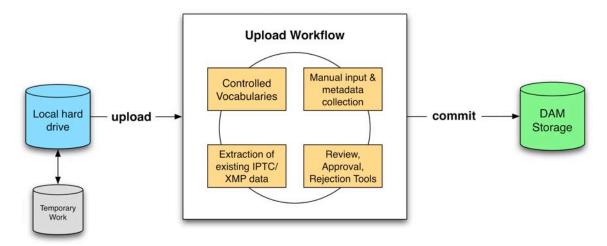
Version 1 - Upload without workflow ('shared network drive')

In Figure 1, when a user copies a file from their workstation to the storage, only the file – and perhaps some technical data or local metadata from the user's own graphics package – will be uploaded. This is not a good strategy for digital asset management, since it means that files are accumulating in folders as they always have in the past.

If only the files are stored, there must be a digital librarian who separately opens each new file at a later date and reviews what it is, how to categorise it and tag it with metadata. Otherwise, the file cannot be searched and is probably better deleted or returned to the user to keep on their workstation.

In reality, it may make sense in smaller systems to allow materials to be committed in this very simple way as it will tend to encourage submissions. However, for more sophisticated setups, the upload workflow tools in your digital asset management tool can provide an automatic intervention to capture metadata from the user performing the upload.

Workflow means that steps exist that lead to other steps; conditions triggering processes, just like a manufacturing process but applied to digital files. The nature of these processes will depend on the features that are configured to suit your organisation. In a generic sense, in digital asset management the upload workflow is about tagging activities and peer review.



Version 2 - Upload with workflow

Metadata tagging can be completed with the aid of controlled vocabularies, or by assigning defaults at the point of upload in templates. This can alleviate the pressure of providing a lot of metadata and encourage occasional users to use the software effectively.

Peer review (approval and rejection) usually means that the finished, tagged content is passed to an administrator or another group of users for their review. Until the requisite steps are completed the uploaded files stay in an unapproved state, rather than appearing in the digital asset management system to other users.

One thing that must be achieved is *adoption*. Users should not be left behind by accident or by designing metadata requirements that are too onerous to be practical.

Participation in metadata is the right way to activate the highly attractive power of the search engine and does need some extra commitment. So, if you can, it is a good idea to require only the minimum of metadata, especially at the earliest stages.

It is also worth establishing what does and does not belong in the digital asset management system. One good check is whether the file is a finished, reusable asset, for example. Keep a temporary work area available, too, for other works-in-progress outside the scope of the digital asset management system.

Browse Versus Search: Using Folders

When you open your library, you might choose to begin by searching or you might decide to browse. The bigger the library, the more powerful the idea of search becomes, but browsing is often an aide to creative research, and scanning through recent uploads is usually productive.

Browsing is a visual process and relies on keeping related files in proximity to each other. Operating systems have been grouping content in folders for many years, so the analogy is firmly established and translates easily to a digital asset management system, although with some limitations.

My most practical suggestion is that the layout of your folders in your digital asset management tool should try to match as closely as possible the way that your organisation naturally prefers to categorise content.

Some options to consider would be:

- Try using a nested folder structure of subjects, departments, divisions, products and so
 on, depending on the most obvious fit or conventions that your organisation uses. It is
 best to be pragmatic and not too formal, in most cases.
- Try to place each asset uniquely in one folder without unnecessary duplication and without storing multiple variants in different formats – as this would fragment your metadata and record keeping efforts. Placing a file in exactly one folder can be quite challenging, particularly if your existing store of data has been divided into projects and has caused related materials to become dispersed.
- Allow your digital asset management tool to eliminate duplicates, and upload just one high-quality version from which other versions can be derived on demand.
- Avoid any scheme based on dates (which can be handled more easily by search)
- Try to keep the depth of your folder trees to a reasonable amount, perhaps less than 5 layers, to improve navigability.

Smart Folders

What if you could have a hybrid solution, where the same file can appear in multiple places at once? This is achieved using metadata as a grouping tool.

As you add metadata to your assets, you can use the search engine to group similar files together. For example, you can create a collection of files all sharing the same keyword, "London", and perhaps other conditions on various metadata fields. This search will gather results from any folder and bring the media together in a single place.

If this search is then saved, you have what is usually described as a Smart Folder or Search Folder. A Smart Folder behaves like a folder but has been constructed from a search, and is kept up-to-date when metadata changes in your library.

Smart Folders will give you a flexible library where you can have the same file appear in multiple locations, since there are various ways that a search can find the same file (so called "many-to-many" placement of the same file in different places).

Using search instead of manual grouping of the files is a superb opportunity to capitalise on your efforts invested in metadata, especially for generating ad hoc collections of files without having to physically relocate content. This is one of the more notable benefits of owning and operating a digital asset management system.

Searching Non-textual Content: Hidden Gems

An often-overlooked benefit of a metadata search engine is its ability to find missing data, or to filter by details which exist but were obtained automatically.

Starting with missing data, you could try searching for files that have no copyright data ('show me files where the copyright field is blank'). This can lead to a quick way to remedy rights issues in a bulk editing form. Alternatively, you can devolve responsibility for this crucial field by filtering such a search by the person who uploaded the file (or when it was created). This is another good example where a Smart Folder can transform metadata into a valuable working tool.

Another benefit of the digital era is the hidden, often voluminous technical data embedded in files. For example, digital cameras and video systems embed dates and times, or may have GPS receivers – especially common now in smart phones.

This technical data allows a search between a date range (say, May to September) through which you should obtain files that are taken in summer. This could be combined with geo-referencing if GPS information has been captured in the file. In this way you would identify seasonal images of a particular location. This is an emerging area of technology that is effective and "free" in terms of effort – and again, it works very well if combined with the idea of Smart Folders.

Other emerging technologies are likely to improve the working practices of digital librarians. These include face recognition and automated image recognition. Both are possible now, but require extensive training on any particular collection of files to succeed.

Summary

It is wise to begin adding metadata with a structured approach, and I hope I have shown that using suitable tools and storing just the right amount of information will also make your life easier. Digital asset management is not only about storing data, it is also about enriching data with metadata.

Techniques such as a controlled vocabulary or adding your own customised fields can make tagging more productive and durable when your library grows. More importantly, your forethought should improve the results your users receive and make your media collection valuable, or recover value that is hidden by the sheer quantity of data you store. Engaging users in the tagging process incrementally is certainly a way to tackle growing volumes of digital media.

I have advocated that using a future-proof technology like XMP is essential, since this means that your metadata will not be lost as you move from one application to another. Although adding metadata to your media will take time, it should be time well-spent, and by using standards you can avoid ever re-entering this data in the future.

If you need to justify the time and money required to start a formal library in your organisation, it is reasonable to propose that adding metadata to a library is an investment - leading to a significant saving in the time and effort made by everyone who searches the library - and that benefit persists in perpetuity.

I'd be delighted to hear from you if you would like to continue this discussion. You can email me at michael.wells@thirdlight.com.

Acknowledgements

Adobe and XMP are trademarks of Adobe Inc. Photographs featured are © Third Light Ltd. and may not be reproduced without prior consent. Captions and keywords used are devised for illustration purposes only.

