Metadata made simple

by Angela Murphy
Image Management and Rights Clearance

When I was asked to write an article called 'Metadata Made Simple', it seemed a contradiction in terms. After all, the art and sorcery of metadata in all its guises is hardly simple. In addition, who was my audience to be? In the world of image libraries and agencies there is an extraordinary range of knowledge about this subject. In large agencies, a small group of experts will know all about metadata to a very deep level – but many other staff in the same agency will scarcely have heard the word. On the other hand, sole traders and smaller agencies will probably be feeling overwhelmed by the amount of information available to them.

As I remembered my first forays into the complex world of metadata, I certainly felt that I was in the latter group - and that I would have liked my own personal guide through the maze of resources. In this article, I will try to provide that guidance – and, in this way, I hope to assist both the expert and the beginner. I cannot claim that this is an exhaustive survey (I have probably left out some critical resource) but I hope it will provide a broad overview and an insight into current developments in the field. The following is a broad overview, a descriptive list of the best of the resources that exist on the internet, and a brief glimpse into the future.

What is metadata?

Metadata is structured information describing different aspects of a resource, such as a document, an image or a group of images. Metadata will make that resource easier to find, use or manage.
There are four main types of metadata: descriptive, structural, administrative and rights. Descriptive metadata describes the subject of an image and includes identifying elements such as title, creator, keywords, etc. Structural metadata describes the way in which a resource is organised, so it might describe the way different manuscript illustrations fit together within a whole work, or connect the images from a particular shoot. Administrative metadata generally refers to technical, management and preservation information and includes elements such as the format, pixel size and colour space of a digital image and its date of creation. Finally, rights metadata tells users about the owner(s) of a resource's intellectual property rights, and of any restrictions to its use. Some of this information can be stored within the digital image itself – for example, the extensive technical information produced by digital cameras (so-called EXIF information) or the core descriptive elements. More detailed information about the content of an image is more likely to be held within a separate related image database so it can be actively linked to other collective resources, such as thesauri, authority files, or rights and client information.

As digital images are constantly shuttled between databases, individuals and organisations, the pressure is increasing for richer descriptive and rights information to travel within the resource itself - especially if it can be copied into the image at the point of export. The ability to hold metadata within digital resources (in this case, images) is thanks to a labelling technology called xmp (extensible metadata platform) –which has been utilised by different organisations to introduce common metadata standards. For example, IPTC has designed four standard data entry panels to assist the interchange of news data, including images. This set of defined fields, or elements, makes up the IPTC metadata schema.

What are metadata schema ?

Metadata schema are sets of information fields (such as location, date, etc) with agreed definitions that allow information to be transferred between different databases or other resources. Perhaps the most famous – and globally accepted – metadata schema is that produced in 1995 by a set of information professionals in Dublin, Ohio, USA. The so-called Dublin Core Metadata Element Set consists of 15 elements – Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage, and Rights. Each of these elements is optional and each of them may be repeated. Although developments in metadata schema have tended to leave Dublin Core behind, it has still influenced or formed the basis of many broader schema. [N.B. For a full review of existing schema refer to the live links listed below, especially the TASI and CHIN resources.]

Attaching good metadata (information) to an image will not only make it visible in a sea of other images, it will also add value. In addition, the use of a standard metadata schema will allow that information to travel between resources and users more effectively. For example, metadata can ensure that rights holders can be contacted – or that researchers can find an image. The better the information – and the more widely held the standards or metadata schema - the more discoverable the image. Indeed, choosing the right metadata schema or mixture of schema is a critical decision that may make the difference between a worry-free project and one that involves extensive manual intervention every time you transfer data.

Entering the same type of information into the same IPTC fields, enables news organisations to automate the transfer of important information – and this in turn helps organisations to introduce automated processes to interact with this information. Organisations that want to use image headers to carry more extensive information can create customised panels to hold other types of information. In theory, xmp is 'platform independent', which means that the metadata will remain with the image as it moves between operating systems (e.g. Windows and Mac).

More complex data can be transferred between databases using an XML (extensible markup language) standard, and these are now being developed for particular metadata schema (e.g. VRA Core 4.0 XML schema). Nevertheless, many individuals or smaller organisations with less sophisticated software may still need to export information manually into a csv, or tab-delimited, file before importing it into another database.

Engaging with local and global search engines
Finally, the value of an image is enhanced not only by the quality of its attached metadata, but also by the way in which individual stock library search engines interact with that metadata, and use sophisticated search algorithms to refine their search functionality. An increased use of metadata standards will support cross-searching and the sharing of images and metadata records between different systems and collections – the so-called 'metasearch'.

Photographers and agencies can post many images onto the web – but without good metadata they will not be seen. At the same time, if you attach too much undifferentiated metadata to an image, you might well drive people away from your website when your images continually pop up in totally unrelated searches. Being clear about the way you define and use metadata will ensure that you strike the right balance between visibility and relevance.

Images, metadata and the 'Semantic Web'
The latest developments in the world wide web focus on the rapidly evolving concept of the 'Semantic Web'. This is based on the idea that existing methods of finding, sharing and combining information on the web can be extended by enabling computers to understand the meaning (semantics) within web resources. For example, Google's image search uses this approach, analysing the contextual information around images to asses their relevance to searches. This involves creating machine-readable languages (e.g. XML and RDF) and teaching machines to interact with that syntax. This originally relied on a 'bottom-up' approach that involved embedding metadata in web content and assumed that resources were available to create this 'machine-readable' metadata.

Metadata made simple – *continued*

by Angela Murphy
Image Management and Rights Clearance

More recently, the development of the Semantic Web has focused increasingly on the 'top-down' approach whereby software tools are designed to analyse existing information and create meaning from unstructured metadata. These tools are known as Semantic Apps and they try not just to search for particular information but to search for the meaning in web content, and then create connections between this content. This approach does not rely on metadata being applied in a controlled way, but tries to create its own connections. For example, such tools allow computers to understand the meaning of information by associating the meaning of different vocabularies. Descriptions of the Semantic Web can be seen at *www.youtube.com/watch?v=OGg8A2zfWKg*. This approach is less effective when applied to purely image-based resources as there is much less information to analyse. In image databases, controlled vocabulary is necessary to create accurate, filtered search results.

Organisations setting metadata standards

There are many online resources about metadata. Deciding which of these are the most useful is the difficult part. In the concluding section of this article I will outline the main standard-setting bodies and just a small selection of the resources that I have found most valuable.

The International Organization for Standardization (ISO) founded in 1947 is a non-governmental standard-setting consortium based in Geneva, Switzerland covering products as diverse as pipe threads, paper and food safety. It has strong links to governments and can therefore help accepted standards to become part of an individual country's legal framework. Members are appointed by national organisations. In recent years, a substantial part of ISO work has been through its Joint Technical Committee, setting standards for the communications and IT industries to ensure portability, interoperability and global harmonisation. Image-related standards ratified by ISO include the PNG, PDF, and JPEG2000 formats; XML Metadata Interchange (XMI); Hypertext Markup Language (HTML); and the Dublin Core Metadata element set.

As an international organisation drawing on the expertise of over 2000 specialists worldwide, it can be difficult for ISO to keep up with the fast pace of technological change. After a period of explosive growth (1994-2003) and a consultation period, ISO launched a new "ISO Strategic Plan 2005-2010" to address this. This can be found at www.iso.org/iso/isostrategies_2004-en.pdf. ISO members often have little time to digest and edit some of the standards submitted. For more views on this see the following links www.noooxml.org/ ; en.wikipedia.org/wiki/Office_Open_XML and the Open Letter to ISO at blogs.freecode.no/isene/2007/09/07/an-open-letter-to-iso/ articulating the fact that the ISO does not have a standard for creating standards.

ISO metadata standards include the useful two-letter (ISO 3166-1 alpha-2) and three-letter (ISO 3166-1 alpha-3) country codes. These and their use in various organisations are explained in a table prepared by Gwillim Law at *www.statoids.com/wab.html* and he has written an excellent article on their use in software at *www.statoids.com/w3166use.html*.

Probably the most influential of the national standards organisations is the National Information Standards Organisation (NISO) in the US whose mission is to both track and create standards. NISO issues guidelines on best working practices in the creation and distribution of digital records (*www.niso.org/publications/rp/*). Its metadata standards include the NISO/AIIM standard for describing technical metadata, Z39.87.

The World Wide Web Consortium (W3C) is an international consortium of web industry organisations headed by Tim Berners-Lee, the inventor of the World Wide Web. Its purpose is to create and maintain web standards and guidelines. These standards include RDF, SOAP, HTML, OWL and XML. W3C has also introduced the concept of SKOS (Simple Knowledge Organisation System) which enables developers to share and link knowledge organisation systems (such as thesauri, and taxonomies) via the Semantic Web.

The International Press Telecommunications Council (IPTC) is a consortium of newspaper agencies and other news organizations that develop and publish industry standards to assist the exchange of data. The organization has a Photo Metadata Working Group that released a White Paper at its first conference just before the 2007 CEPIC Congress. Presentations are at *www.phmdc.org/*

IPTC have defined a number of fields that can hold essential information (such as caption, rights information, and keywords) about individual images in the image's own file header. This information is embedded in the file using Adobe's xmp (Extensible Metadata Platform) which is an image labelling technology using XML and RDF (both are digital languages used to 'mark-up' text and have been developed to provide common languages for exchanging data). This group are now working to extend the range of information that can be held in IPTC headers and utilising new technology to improve the way in which this information is held. The current IPTC specifications, technical documentation, and Photoshop CS panels can be downloaded from *www.iptc.org/IPTC4XMP/*. Further information about the background to IPTC headers can be found at *www.controlledvocabulary.com/imagedatabases/iptc_naa.html*

While the ISO, NISO, W3C and IPTC have focused on common technical standards and developing tools to enable interoperability and smooth data migration, the academic and cultural heritage world has primarily focused on the standards that apply to the descriptive content of resources such as the historical context in which the resource was created, its geographical context, or contextual information about its creator. In addition, these standards have evolved over a considerable period of time.

The complex information that is created to describe cultural objects has traditionally been held in detailed catalogues, which are now gradually being transferred to online databases. These link to related resources to help users standardise place names, dates, creator names and the relationships between them. The explosion of information resources on the web, including images, can no longer be accurately referenced by sophisticated search engines and automatic indexing tools. As a result, the sophisticated cataloguing standards developed by cultural organizations are becoming increasingly relevant to organizations dealing with more ephemeral resources.

A leading exponent of descriptive metadata standards is the Visual Resources Association (VRA) which started as an association of art slide librarians working in the educational and cultural heritage sectors who came together to further research and education in the field of image management. With the advent of the digital age, their activities appeal increasingly to a wider community and, as its membership has grown, its aims have also broadened to include commercial objectives. Today the VRA has over 700 members and its publications and meetings deal with all aspects of image management including copyright, technical digital image issues, data standards, cataloguing standards, and other emerging issues.

Having its roots in the cultural heritage sector has enabled the VRA to become expert in the cataloguing of visual information and it has worked hard to create standards to describe images. The VRA has

Metadata made simple – *continued*

by Angela Murphy
Image Management and Rights Clearance

produced a widely-adopted metadata schema (latest version is VRA Core 4 *www.vraweb.org/projects/vracore4/index.html*) and it has collaborated on the "Cataloging Cultural Objects Project" (CCO) funded by the Getty Institute and the Mellon Foundation. An important result of this was the publication of a guide to 'Cataloguing Cultural Objects' (with selected excerpts online) – now the leading work in this field. More about CCO can be found at *vraweb.org/ccoweb/cco/about.html*

The VRA website (*www.vraweb.org*) contains a variety of other useful resources including the Digital Image Rights Computator (DIRC) program – an online interactive resource intended to help non-commercial users to assess the copyright status of a specific image "documenting a work of art, a designed object, or a portion of the built environment".

The Getty Institute has been an important force in the development of metadata standards, This renowned cultural organisation has not only been central to CCO, but it has also achieved a consensus with the Visual Resources Association (VRA) so that information can be migrated easily between the VRA's new schema VRA Core4 and the Getty Institute's own metadata schema – CDWA and CDWA-Lite. The Institute is also the source of three significant vocabularies – the Art & Architecture Thesaurus (AAT); the Getty Thesaurus of Geographic Names (TGN) and the Union List of Artist Names (ULAN).

The vast Library of Congress Subject Headings were developed for libraries to classify bibliographic records. They are also useful as a basis for the classification of images as is the Thesaurus for Graphic Materials. Other useful thesauri are listed on David Rieck's Controlled Vocabulary site at *www.controlledvocabulary.com/examples.html*

Another useful source of information about metadata standards is UKOLN which is a centre of digital information management expertise in the UK. UKOLN advises the cultural and educational sector, raises awareness of standards and carries out research and development. Links to further information are at *www.ukoln.ac.uk/metadata/.*

Finally, an important set of standards relating to rights is fast emerging from the Picture Licensing Universal System (PLUS) initiative - this time generated from the commercial sector. The development of a universal set of clearly-defined licensing terms is a welcome addition to the standards field. PLUS aims to simplify and facilitate the communication and management of image rights through its use of machine-readable coding and standardized language. Its website (*www.useplus.com*) contains downloadable helper software, including the PLUS License Generator and the PLUS License Embedder & Reader lists of participating institutions,

Recent Developments

Professional photographers and stock agencies are playing an increasingly important role in the development of metadata standards. Taking the lead in the United States is the Stock Artists Alliance (SAA), the professional photographers' trade association that has just been granted funding as part of a major US initiative to promote awareness of standards. This initiative is funded by the Library of Congress through its National Digital Information Infrastructure and Preservation Program (NDIIPP). NDIIPP partners also include the ASMP (American Society of Media Photographers) and ARTstor, as well as organisations focusing on sound and motion picture archives. Details are at *www.loc.gov/today/pr/2007/07-156.html*. These organisations also contribute to the Universal Photographic Imaging Guidelines (UPDIG) which has just released version 3 of their guidelines for image submission

A recent European project is the MILE (Metadata

Image Library Exploitation) Project – a three-year European Union funded project that is being co-ordinated by the Bridgeman Art Library (*www.mileproject.eu/*). The project aims to promote the development and use of image metadata, especially where this can improve the use and trade of digital images throughout Europe. It is helping to raise awareness of existing metadata standards in the cultural heritage sector and elsewhere through seminars and via the website. Participants in the project and other interested parties hold discussions online at mile-forums.ssl.co.uk/phpBB2/.

Despite the growing amount of activity in this field, a lack of common standards still costs the stock industry an enormous amount in hidden costs and lost time. Many libraries still use non-standard keywords and few access standard thesauri when building their own vocabularies. The use of non-standard structured ontologies and taxonomies is widespread. Nevertheless, although the complex metadata schema developed by academic institutions are not usually suitable for commercial image libraries, they do provide a set of metadata tools that are capable of giving any of their users a head start in ensuring that their assets are discoverable on the web. Even Google are using the Library of Congress Subject Headings to enhance their Google Book Search.

It is vital that commercial image libraries explore and adapt the extensive work already done by cultural sector if only to ensure that when it comes to defining terms more precisely, they do not reinvent the wheel. Once IPTC's suggestions for the extension of standardised IPTC header fields are fully embedded in software upgrades, it should also be possible to import

Sources of Information and Training

Many of the organisations listed here provide very useful online and downloadable resources. However, there are a few whose online resources are outstanding. One of the most rich and varied set of web resources about image management and metadata standards is that produced by the UK's Technical Advisory Service for Images (TASI) at www.tasi.ac.uk . TASI is a Bristol-based higher-education funded training centre that runs expert training sessions in all aspects of image management. Of broader interest are the exceptional online resources published on their website. Articles relating to metadata are at *www.tasi.ac.uk/advice/delivering/delivering.html*. Another rich source of information and online training courses is the Canadian Heritage Information Network (CHIN) which also hosts globally-accessible, interactive courses on all aspects of digital image management. Although aimed at the cultural heritage sector they are of use to all. The SAA mentioned above also has a user-friendly set of tutorials for stock photographers at (*www.stockartistsalliance.org/tutorials/*). The website of the Periodical Publishers Association in the UK offers free downloads of two excellent guides – pic4press and pass4press – at *www.pass4press.com/*. Other organisations with valuable free downloadable information are UPDIG (*www.updig.org*) and PLUS (*www.useplus.org*).

Finally, there is the excellent resource created by photographer, David Riecks for *www.controlledvocabulary.com* with its subject hierarchies and plain spoken text.
Angela Murphy is a consultant in image management, digital workflow, and rights clearance and management, working in both the commercial and cultural heritage sector.

Angela Murphy, Consultant
Image Management and Rights Clearance
E: angela@angelamurphy.co.uk