

Automatic Metadata Generation (AMG-UC)

Factual Metadata

Ian Watson, Project Manager (Knowledge Media) IRISS

ian.watson@iriss.org.uk

What is Factual Metadata?

Factual metadata is objective information about a resource, for example the duration of a moving image, the pixel dimensions of an image or the name of the user who downloaded a resource. Some consider title and author data to be factual¹. The table below is not exhaustive but lists the types of metadata that may be considered factual:

Date	Publication date, date of metadata contribution, last access date, last modified date etc
Time	Duration, time of creation or modification, time downloaded
User	Username
Software System	System on which resource is built or under which it runs
Unique identifier	URI
Price	More relevant in commercial environments
Rights	Copyright owner and licensing arrangements e.g. Creative Commons
Geographic information	Postcode, OS grid reference, latitude, longitude
Title	of document, film, track or album
Author	might also be artist in the case of performed work such as film or audio
EXIF data	digital photo metadata: includes static information such as the camera model and make, and information that varies with each image such as orientation, aperture, shutter speed, focal length, metering mode, and ISO speed
IPTC metadata	The IPTC standard contains fields for transmission dates and times; rights owner, restrictions on use, relationships to other objects etc.

Clearly some of these data types will be easier to generate automatically than

¹ Tonkin and Muller include Author and Title in the definition of factual metadata. "Keyword and metadata extraction from pre-prints"
http://elpub.scix.net/data/works/att/030_elpub2008.content.pdf

others. For example, date, times, URI and Exif data are usually system generated at the time of creation with no user input. Others, such as price and author usually require manual input, although research continues into ways of deducing such information from the content of the resource.

Why is Factual Metadata Important?

Metadata creates a link between the user's description of what he or she wants to find and resources that match that description. It is important and beneficial because it describes an object in terms that should be understood by a user's query, thus making the search more accurate and efficient, and therefore more satisfying for the user.

It is generally accepted that manually creating metadata is expensive because of the time and manual effort required to ensure quality and consistency. Where metadata can be generated or captured automatically the cost of creation decreases, potentially to zero. When digital photographs are imported from a digital camera to a digital archive, repository or album on a desktop computer, information such as date and time of creation of each image is automatically included with each image at no cost and with no effort on the part of the user.

This allows image handling programmes such as photo or Picassa automatically to arrange images by date and into albums. In commercial or institutional scale archives or repositories this kind factual data facilitates the management and organisation of images and can play a vital role in workflow management. On the other hand, adding descriptive information such as the names of the people in the photograph and the geographic name of the place where it was taken, or the occasion (Uncle Sam's birthday), requires time and effort and may be inconsistent because humans tend to be inconsistent, inaccurate or vague: whose Uncle is Uncle Sam and was it really his birthday?

When importing audio tracks from a music CD to an mp3 player such as iTunes, Gracenote² will automatically add artist and track name data to the imported files. With this information, the user can organise the audio files by album name or artist with virtually no effort. Before the existence of Gracenote, users had to transcribe this information manually, create and name folders and move the files to the appropriate folder.

Unlike subjective metadata such as classification terms or keywords, factual metadata ought to be unambiguous and objective if it is captured automatically from, say, a digital camera. More importantly, factual metadata creates much richer description of the resource, opening up the possibility for more granular searching and for creating links between resources that share the same 'facts': duration of an audio clip, date created, date last used, number of downloads etc. Indeed it has been suggested that the more fine-grained the level of description, the greater the scope for re-purposing learning and teaching resources and

² <http://www.gracenote.com/>

offering sophisticated end-user resource discovery experiences³. For example, MPEG-21 descriptors at any level of granularity from series or programme to the individual frame would make easier the task of finding shots, themes, key frames etc.

Factual metadata may be important in various circumstances. If a user wishes to find an object that was published in a specific year, month or day, then factual metadata will be helpful, especially if the date and time are recorded numerically. This would allow searching for date and times in relative terms, for example modified in the last month, day, or hour. One source of frustration when searching Google is that it is not possible to sort search results according to date, title, author etc. By contrast the commercial world makes extensive use of factual data: online retailers, for example, allow sorting results by price or model or both. Comparison sites such as comparethemarkets.com make extensive use of factual metadata to compare insurance products.

Linked Data technologies rely on factual metadata as the 'facts' are the hooks for making links between data sets and between data sets and users. Wikipedia defines Linked Data as "a term used to describe a recommended best practice for exposing, sharing, and connecting pieces of data, information, and knowledge on the Semantic Web using URIs and RDF." In an interview with the BBC on 12 June 2009⁴ Tim Berners-Lee used the example of how a map of bicycle accidents was created within hours of the release of raw data by government. This is possible because the raw data contains the postcode for the location. He further extols the virtues and applications of Linked Data in a lecture available on TED⁵.

In summary, automatically generated factual metadata can increase both the recall and precision of searching. Recall is the proportion of relevant documents retrieved out of the total number of expected relevant documents whereas precision is the ratio of relevant documents retrieved to the number of documents retrieved⁶. The two measures are usually characterised as being in an inverse relationship, i.e. the greater the number of results the less likely they are to precisely match the query (more noise); and the more precise the query, the fewer the results, but with possibility that some relevant objects may have been missed. Typically search engines such as Google deliver high recall but often low precision.

Automatically generated factual metadata helps optimise both recall and precision. For example if the user requires objects created between 1st and 31st March 2009 and the date of creation has been generated automatically then the user will retrieve all objects (maximum recall) and only objects (maximum precision) created within that date range.

³ Metadata Generation for Resource Discovery, Polfreman, M et al p.20

<http://www.jisc.ac.uk/whatwedo/programmes/resourcediscovery/autometgen.aspx>
⁴ (http://www.bbc.co.uk/blogs/technology/2009/06/sir_tims_cry_raw_data_now.html)

⁵ (http://www.ted.com/talks/tim_berniers_lee_on_the_next_web.html).

⁶ Metadata Generation for Resource Discovery, Polfreman, M et al
<http://www.jisc.ac.uk/whatwedo/programmes/resourcediscovery/autometgen.aspx>

Scenario 1

Dora has created a learning object about the French Revolution. When reviewing the content as part of the rights clearance exercise she discovers that the royalty free licence she obtained for the image of a painting allows classroom projection but does not allow electronic distribution.

As she wishes to disseminate the learning object on the web she decides the simpler, and cheaper, option would be to find a substitute image that is licensed under creative commons. She searches Wikimedia Commons and finds a photograph of a painting depicting fighting in Paris during the revolution. As the picture carries a creative commons licence she may use it in the learning object.

Scenario 2

Danny is searching a digital repository, has found something useful and would like to find more objects like it. Further he would like to know about what other users think. He enjoys using online music services such as Last.fm and Spotify, which seem to be rather good at selecting music he likes. He also likes the way Apple iTunes's 'Genius' feature is able to suggest tunes based on what he is currently listening to. He thinks it would be good if the repository could also recommend in the same way and perhaps even find material from other repositories.

Although the repository allows users to allocate star ratings to an object's metadata as well as recording which users have downloaded the object, it does not correlate these two bits of metadata to tell Danny that the object he likes (object A) was rated five stars by user X who also rated objects B and C with five stars and object D with only one. Factoring in the date and time of download and the star rating would add a temporal dimension. This use of factual metadata (user and star ratings) could form the basis of a community around the repository.

Scenario 3

Oscar, a photographer, wishes to reduce the time spent on adding caption information to his digital photographs. As his digital camera has GPRS it captures the longitude and latitude of every image. The problem is that these co-ordinates are not very human understandable. Google can plot a location on a map if it is given a postcode. Oscar thinks it would be helpful if there were a tool that could a) plot the exact location on a map and b) suggest a geographic name for that location.

Scenario 4

Maria is picture researcher, searching for still images of a town centre as part of a project to document changes to the streetscape over time. She has found some digital images created by scanning hard copy prints and these contain approximate dates entered either by the photographer or by a cataloguer. Fortunately the repository contains other images that were 'born digital', and for these images the repository manager used a macro to extract the date of

creation from the Exif metadata in the original JPEG and add it automatically to the metadata template. Not only did this provide an exact date, it also provided the time at which the image was captured, thereby providing more exact information for the project and the possibility of linking the image to other events happening at that date and time.

Tools

Dbpedia (<http://en.wikipedia.org/wiki/Dbpedia>) is a community effort to extract structured information from [Wikipedia](#) and to make this information available on the Web.

DBpedia allows users to ask expressive queries against Wikipedia and interlink other datasets on the Web with DBpedia data. The BBC is experimenting with DBpedia and Linked Data technologies to create a better user experience (<http://derivadow.files.wordpress.com/2009/06/eswc2009-bbc-dbpedi-2.pdf>)

Leuven's Automatic Metageneration System is reported to extract information from both content and context and experiments have been tried on extracting LOM data in Dublin Core format⁷.

Standards

The Exif specification is important one for digital photos (http://en.wikipedia.org/wiki/Exchangeable_image_file_format). Exif contains standard tags for location information. Currently, few cameras support this but the iPhone, for example, has this capability and records the latitude and longitude of the device when the image was created.

The IPTC Standard (<http://www.iptc.org/>) is widely used in the media to manage text and image files contains a vast amount of factual metadata such as time of transmission, embargo date and rights.

The ID3 tagging standard (<http://www.id3.org/>) is relevant to sound files imported into mp3 players.

MPEG-21(<http://en.wikipedia.org/wiki/MPEG-21>) has relevance for tagging multimedia files as well as recording rights information.

The LOM standard (<http://ltsc.ieee.org/wg12/20020612-Final-LOM-Draft.html>) has much to offer in terms of metadata for multimedia learning objects.

Key Issues

Scenario 1 requires the existence of sites such as Flickr and Wikimedia commons on which materials are available that have licence data in their metadata. Local tools are required that allow users to attach licence information which can understood by Flickr or and other sites.

⁷ Metadata Generation for Resource Discovery, Polfreman, M et al p.73
http://www.jisc.ac.uk/media/documents/programmes/resourcediscovery/metgenreport_final_v5.doc

The key here is interoperability. An IntraLibrary repository, for example, has the facility to append Creative Commons, or any other, licence information to an object and that information will be included in an IMS package of the downloaded object. Services such as Wikimedia Commons and Flickr do not support IMS and the licence information would have to be applied again were the image uploaded to be uploaded to one of these sites, as would be allowed under CC.

Scenario 2 requires work on the use of linked data techniques to establish connections between objects in different repositories thereby breaking down silos. Some work has also been done on rule-based approaches (http://www.w3.org/2005/rules/wg/wiki/UCR/Candidate_Use_Cases_for_2nd_Draft/PublicationAlternative) which attempt to infer characteristics of a resource from certain factual metadata. For example if a film described in IMDB had a budget of less than a certain amount it would be characterised as 'low budget'. Other rules might invoke to deduce, for example, genre.

Scenario 3 requires exploration of the creative use of latitude and longitude to generate a geographic place name or a postcode. For example, the ability to discover that 56 13 mins N / 4 18 min W is a location in the Trossachs in the postcode area FK17 8HP would have a number of potential uses. We could go further, as advocated by the linked data movement, and uses dbpedia to link to other information about this location: nearest town, public transport services, local facilities etc.

Scenario 4 requires that image capturing device, in the case a digital camera, has been set to the correct date and time. Exact date and time of image creation add greatly to the granularity of cataloguing and help researchers place images in context. This factual data can also provide hooks for 'linked data' so that the images may be linked to other sources of information about, say, what events were taking place in that town or street at that time or on that date.

Recommendations

Factual metadata may be seen as the key to providing links in Linked Data technologies. When applied to non-textual resources, such as moving images and audio, it can provide the key to delivering richer and more detailed indexing, possibly down to individual frames, which will ultimately provide better access to specific information and help resource discovery.

1. Explore how to fully exploit factual metadata that is already being automatically created, such as latitude and longitude stored in digital images. This might include looking at how Linked Data technology as promoted by Tim Berners-Lee and others may be exploited. Essentially this approach uses factual metadata, potentially at highly granular level, to make links between resources in order to improve the relevance of search results.
2. Explore whether the methods being developed at Leuven have potential for reducing the costs of cataloguing objects using the LOM standard. If

this cannot be achieved automatically it is possible that LOM will cease to be relevant as the costs of manually entering the metadata are high.

3. Investigate technologies that automatically extract metadata from audio and moving images to enable granular indexing.
4. Linked Data initiatives appear to hold much promise for linking objects based on factual metadata and would merit further examination.

© 2009 Institute for Research and Innovation in Social Services. Licensed under Creative Commons CC-BY <http://creativecommons.org/licenses/by/2.5/scotland/>