Original Article

When metadata worlds collide: The hunter/gatherer dichotomy

Carol Owens
was one of the pioneers of media metadata when she led development and implementation of the BBC's Enterprise Data Model SMEF™ from 1998 to 2004. She was subsequently a managing consultant for Siemens, Head of Media Strategy for Serco, and is now Director, Strategic Engagements for Ascent Media.

ABSTRACT 'Metadata' is a term used by both information management (IM) and web professionals, yet they refer to different things, and the communities have different objectives. IM creates structured data to support focussed hunts for unambiguous answers, whereas the web gathers together multiple responses, links and associations. End users may need the help of both approaches at different times. Moving forward, there are increasing opportunities for collaboration between the two cultures, especially in the management and publication of media assets.
Journal of Digital Asset Management (2009) **5,** 181–184. doi:10.1057/dam.2009.14

Keywords: metadata; semantic web; search; controlled vocabularies; information management; archives

WHICH ARE YOU?

When you do a search through a web provider or an information management (IM) system, what are you hoping to find? If it is a specific target, and no other will do, you are a hunter. If you are interested in seeing what is available around a subject, then you are a gatherer. The web is ideally suited to content foraging and tracking knowledge across links and associations; that is why we use a 'browser'. If, however, the record or content returned must be unique, unambiguous and exactly what you need, you would typically expect to be accessing an information system (or ecosystem), holding rule-based data. As an individual, you may be a hunter or a gatherer at different times, depending on your objectives and role. Your aim may be retrieval of a singular answer, or discovery of multiple answers; it just depends what you need to do.

Simplistically speaking, support for hunting versus gathering has been split between the professional IM community, providing the spears, and the web publishing and aggregating community, providing the net. This is starting to change, however, with IM adopting web portals as data access mechanisms, and Web 3.0 exponents developing data structures to underpin the semantic web and 'disambiguate' results. There is exciting potential for the two cultures to work together more closely to enable the strengths of each to be fully exploited.

Correspondence:
Carol Owens
E-mail: carol.owens@ascentmedia.co.uk

WHAT ARE WE HUNTING?

As a starting point, it might be useful to establish some common terms and concepts for what is being hunted or gathered. Broadly speaking, in the broadcast media industry, it is either 'metadata' or 'content', where metadata is the information about the unstructured audiovisual data that identifies and describes it. Content may be composed of images, sounds and words that directly convey the editorial intent to the audience through playing or displaying them. The metadata functions as a proxy for the content in most searches and transactions – we make many of our decisions based on text. This is not only true within the professional production and management

domain, but also true for the audience checking out the Electronic Programme Guide (EPG) on-demand playlists and old-fashioned printed listings.

On the web, however, the term metadata applies more to the data tags that enable linking of content together to create information journeys, or aggregate disparate sources together to present new collections of answers. Content in this case may be purely text-based, or include audiovisual media. For example, the author of a web page about a TV programme may include hyperlinks to other pages or sites that relate to the core content, such as location information, artists' biographies, events or the music used. This is an editorial decision, where the author/producer can guide the audience through his or her original sources or lateral thinking, to broaden the conversation and enhance the overall experience.

Web technology also presents the opportunity to use automated linkage and aggregation tools such as spiders that crawl across published web content looking for shared terms and topics, and an indexing engine can tag the content and pull together possible associations for consideration by the author. The challenge for this technique is a tool's inability to identify whether topics or targets are really the same, or just have data elements in common, for example, a person's name. There are many 'John Smiths' in the world. For this reason, the web community has started to look for and adopt specific open sources of terms and definitions to serve as values in a controlled vocabulary. This is similar to the IM community's use of Reference Data sets, such as ISO country codes, designed to eliminate ambiguity in data entry, which are often presented as drop-down lists.

KNOWING YOUR TARGET

Wikipedia, the online encyclopedia, has been adopted as the source of a controlled vocabulary across a huge range of topics. Web authors can insert a Wikipedia subject HTML tag into their text as a unique identifier and link, and all readers will be able unambiguously to understand the term used and its meaning (and more). The 'dbpedia' project has gone further and gathered together a list of terms and unique identifiers for topics, people and locations, which can be universally referenced. The only note of caution is that as Wikipedia is dynamic and subject to change, the links may be superseded over time or the definitions may be updated. Managed data sources, such as Musicbrainz, are more reliable, and support the hunt for information about that song by that artist in that performance, and no other.

These open web-based controlled vocabularies could equally be used by the IM community within in-house business systems, so long as the data integrity issues can be understood and managed. This is one opportunity for crossover between the disciplines. A further area of useful common thinking may be around ontologies, or the high-level view of how key concepts relate to each other. An ontology identifies the things we know about, for example, a domestic telephone directory knows about people, addresses and telephone numbers, with the rule that a person is assumed to have a one-to-one relationship with an address, but there could be more than one phone number associated with that location.

The phone book ontology was built into it from the start, but web ontologies can be developed and imposed *post hoc* on existing data to support searches. For example, the Friend-of-a-Friend, or FOAF, ontology relates individuals, activities and relationships in an open protocol for connecting social networking sites. A user might create a FOAF profile among those who have posted entries to sites like Facebook, MySpace and so on to look for people interested in hang-gliding, who live in the London area. This will not be a comprehensive list, as it is constrained by people with the relevant entries, and the terms they have used; it is a gathering mechanism, not a definitive hunt.

Relationships within an ontology can be made explicit through the structural rules of web authoring languages, and also the data structures in traditional IM data modeling. For example, the Resource Description Framework (RDF) uses a three-part syntax of Subject/Predicate/Object, where, for example, the Subject might be a person, 'John Smith'; the Predicate could be 'e-mail address'; and the Object would be 'john.smith@email.com'. Logical IM data modeling tends to work at a higher level of abstraction, so that there could be an Entity or Class of 'PERSON', having Attributes or Properties, such as 'PERSON_NAME', for which the Value might be 'John Smith'. A further Attribute of the

Person might be 'PERSON_EMAIL_ADDRESS', with its corresponding Value. It may look complicated, but the rules are built into applications and profiles, and are made invisible to ordinary users, so that all they need to worry about is either entering or finding information that's important to them.

THE MEDIA TERRAIN

Why is this relevant to the media industry, and who are the hunters and gatherers? The terrain in which they operate divides between the professional in-house media production and management domain, and the external audience domain of push delivery (scheduled broadcast transmission) and on-demand consumption through the Web or VoD. The quarry is content, and metadata about the content, and the harvest is context and wider knowledge and enjoyment. To date, these two domains have often been thought of and managed separately, but the argument and funding for such a divide and potential duplication of effort is rapidly weakening. People need to hunt and gather in both domains at different times, and those who are responsible for providing the source data to enable these activities need to be aware of and support both aims.

It is obvious that when publishing information about content, or the content itself, on the web, web techniques and protocols must be employed and exploited. This has worked well while web content has been specifically authored for the platform, but is presenting problems as on-demand provision of broadcast content through channels, such as the BBC iPlayer, becomes more popular and the demand for content and descriptive metadata (as well as linking metadata) increases. The sources of information and content will probably lie within the broadcast media domain in business systems that store the data in formalized and managed structures.

A classic example would be an archive cataloging system, which uses a specialist taxonomy to describe a programme's content. This is vital to the archivist or producer, who is hunting down a specific programme or sequence, and requires a 100 per cent success rate on the search. However, it is not understandable or useful to the audience. It will be necessary to find other ways of creating simplified metadata for browsing purposes that is consistent with the archive data, and may open a window on to it if people are really interested. It should be possible to retrospectively derive a simplified ontology for the catalogue that makes connections and retrieval easier, and is preferably consistent with the conceptual models used in production and publication. If the raw catalogue data are made accessible to the web, could search tools then understand the relationships and perform the aggregations along the lines of FOAF? In addition, what would spider technology make of an archive catalogue? Could spiders be programmed with specific editorial principles to automatically create new collections around a theme, such as an anniversary or event? This line of research requires the web and archive specialists to collaborate, and bridge two very different cultures.

CREATING THE METADATA

In terms of data creation, a media archive catalogue may be derived in part from information provided by the original production team; the archivists do not have time to view or listen to all the content passing through a large company. The onus is then on the production teams to create good descriptive metadata in the first place, and also to design the links and associations to enhance and extend the audience experience. It may be possible for them to take a Wikipedia approach to content description, giving each editorial unit (for example, Series, Episode, Item, Version) its own unique resource identifier, and maintaining a consistent location for the data. Other information created through the course of production such as scripts, cast lists, costume designs or whatever could be associated with that unique ID and accessed through a portal. This also provides the rich content for a future audience version of the program information resource to accompany the playable media itself. Producers today understand the power of Wikipedia, and may find this an attractive and motivating opportunity.

It may be feasible to use the 'Wikimedia' page as the Primary Key/Unique ID for all associated information sources, to be added over time and linked using the identifier. Archivists could extend the original production description with catalogue terms, which could be derived from professional taxonomies published as open

sources of controlled vocabularies. Wherever possible, open controlled vocabularies should be referenced to make the published data accessible to searching and aggregating tools. Clear indication should be given as to which data are 'approved' by the content owner, to distinguish them from any user-generated contributions.

It seems possible that the same principles could be applied to traditional information systems, or within an intra- or extranet. Data from the Scheduling or Rights systems could be provided in a portal window while still being managed by the master systems. The system owners would be responsible for ensuring accuracy and timeliness in support of business requirements, for example, in terms of editorial compliance and parental control. It is vital that the precise version of a program can be identified and retrieved to avoid the broadcaster being lambasted by the Regulator for putting out the post-watershed version in the afternoon. Open, controlled vocabularies may replace in-house reference data sets where appropriate, provided that they are fit for purpose.

OPENING UP THE SOURCE

The decisions about which metadata to expose to the public audience are editorial and commercial. Viewers or users will need to hunt or gather as the mood takes them – to browse happily through 'comedy' or to track down the documentary about their hometown they remember seeing ten years ago. Any media portal and potentially EPG must support both sets of objectives, by supplying links and associations to the gatherers, and accurate, relevant metadata to the hunters.

Search analytics can provide useful insights into how people search, and perhaps reveal the implicit ontologies in their minds. This could be helpful in designing data capture and presentation, and may feed back into thinking about the professional domain and cataloging structures. Links can be designed-in during content production, or superimposed through open search ontologies. The important thing is to deliver value to the audience by answering their questions, while remaining true to the original content and its wider context – all at minimal or no extra cost.

This challenge can only get more interesting.