Taxonomies and metadata for digital asset management

Received (in revised form): 4th February, 2018

Heather Hedden
is a senior vocabulary editor at Gale, a Cengage Company, where she edits the subject thesaurus for Gale research database products and develops discipline taxonomies for the management of content for Cengage online educational products. Previously she was a full-time taxonomy consultant through her own business, Hedden Information Management. She is the author of 'The Accidental Taxonomist', 2nd edn (Information Today Inc., 2016) and currently provides limited taxonomy consulting services, offers onsite workshops and teaches an online course in taxonomy creation.

Hedden Information Management, 98 East Riding Drive, Carlisle, MA 01741, USA
Tel: +1 978 467 5195; E-mail: heather@hedden.net

Abstract Taxonomies, which can aid in the retrieval of digital assets, are of various kinds, and different kinds (hierarchical, faceted and thesauri) are suited for different situations. Taxonomies may be implemented in a digital asset or content management system either for tagging assets or for categorising assets or for both, and there are different circumstances that favour tagging versus categorising. Standards for taxonomies include both guidelines of standards bodies for best practices in developing controlled vocabulary, and specifically thesauri, and various industry recommendations for metadata schema so that controlled vocabularies can be shared. Taxonomies typically form part of a larger metadata architecture. Therefore, taxonomy design and metadata design should go together.

KEYWORDS: categories, controlled vocabularies, metadata, search filters, tagging, taxonomies, thesauri

INTRODUCTION

A taxonomy provides a combination of benefits: more accurate and complete retrieval of digital assets than can be retrieved with search alone, and the ability to browse topics that are structured in a way to guide the user to select the most appropriate topics. As such, taxonomies can be powerful tools for managing and retrieving content. Nevertheless, not all taxonomies are the same in their structure. Some are hierarchical, some are faceted, some support search and some provide a combination. With a better understanding of taxonomies, the right kind of taxonomy can be implemented. Furthermore, taxonomies usually form part of a larger metadata architecture. Therefore, taxonomy design and metadata design should go together.

Taxonomies can refer to various related schemes for organising topics and, by extension, content and information. A view of taxonomies that is too limited can result in overlooking their broader applications and benefits. A traditional and limited view of taxonomies is that of hierarchical arrangements of terms for browsing from the top down, from the broadest terms to the most specific, as in inverse tree structures. These types of taxonomies are useful in certain subject domains and content retrieval use cases but are not suitable in other cases. It is also important not to confuse a taxonomy with a navigation scheme (as menu labels

or a site map) as for a website, intranet or portal. Navigation schemes provide a guide to a site as it is structured, but they do not serve as an independent reference lookup tool as a taxonomy or metadata does.

TYPES OF TAXONOMIES AND USES

A broader view of taxonomies considers them as structured controlled vocabularies. A *controlled vocabulary*, in the context of information management, is defined by the standards body ISO as a 'prescribed list of terms, headings or codes, each representing a concept'.[1] A taxonomy may then refer to a kind of controlled vocabulary with some structure/relationships among the terms. ISO defines a taxonomy as a 'scheme of categories and subcategories that can be used to sort and otherwise organise items of knowledge or information', and in a note, explains that the simplest taxonomies may not have subcategories.[2] In a broad sense, a taxonomy may be presented in the following forms:

- one or more sets of hierarchies of terms, whereby all individual terms are related to each other in hierarchical relationships;
- a set of terms related to each other by either hierarchical or associative (related-term), or equivalence (synonym/variant) relationships, also known as a thesaurus; or
- a set of distinct types or facets of terms that are intended to be used in combination for search and retrieval of content. The structure is not necessarily the relationships between terms but the grouping of terms into facets.

Each of these examples of taxonomies is suited for a different purpose, both with respect to content types and content retrieval/management situations.

Hierarchical taxonomies may be suited for content and terms that naturally can be categorised and for a subject area with a defined scope. These could be a set of product types, industries, geographic places, academic disciplines, arts and crafts, occupations, organisational departments, news organised like newspaper sections, etc. Hierarchical taxonomies work well if the taxonomy is not too large, where the number of concepts is in the hundreds or less. The content retrieval situation for which hierarchical taxonomies are best suited is to provide guidance to nonexpert users who want to explore topics to discover what they are looking for. The taxonomy can even serve an instructional purpose to outline the subject domain for the users.

A thesaurus is the better option for content and terms that cannot neatly be categories into a limited number of hierarchies, such as business-related activities or current trends in popular culture. In addition, a thesaurus is also more suitable for multiple, overlapping subject areas or domains with diverse content, such as topics of research reports, and for very large and growing controlled vocabularies, in the thousands or tens of thousands of terms. Thesauri are best used for indexing and retrieval situations such as when detailed indexing is needed with highly specific terms, indexing is done by trained or professional human indexers or the users are subject-matter experts who will likely look for specific terms. Thesauri are useful to support searching, especially with search-support type-ahead/auto-complete features, and in user interfaces with full alphabetical browsing. There are also national and international standards for thesaurus creation, ISO 25964-1 and ANSI/NISO Z39.19-2005. Because thesauri are more complex than taxonomies, it takes a higher level of skill and expertise in library or information science to create them.

A taxonomy structured as a set of facets is best suited for managing and retrieving content that is of a somewhat unified or limited kind so that the content items share certain aspects which can be covered by shared taxonomy facets. These content items could be all the same type, such as product records, people records, customer documents, reports, marketing collateral, and content for digital publishing. The implementations for

which faceted taxonomies are best suited are more varied. Limiting/filtering/refining search results by facets is suitable both for novice and expert users. The use of taxonomy facets, as part of a larger set of metadata properties, is also suitable for the work of content managers or digital asset managers who have other workflow tasks, such as identifying records with certain rights or retention status, audience or market, and source or owner.

TAGGING VERSUS CATEGORISING WITH A TAXONOMY

There are different options for using a taxonomy in a digital asset management (DAM) system or content management system (CMS). Such a system typically provides a feature for tagging content with controlled vocabulary terms, which in this context might be called 'tags'. Tags may be any keywords of the digital asset manager's choice, or they may be restricted to terms from a controlled vocabulary/taxonomy. In some DAM system or CMS, there is also the option to categorise content items according to a limited number of predefined categories, often represented as virtual folders. If both features exist, a decision needs to be made about taxonomy implementation and use: whether the controlled vocabulary be implemented a controlled list of tags for tagging content, whether it be implemented as a taxonomy of categories for categorising content items, or whether both methods will be implemented with different controlled vocabularies for each. The choice depends on various factors.

Tags and tagging is preferred under the following circumstances:

- if the workflow involves content files 'travelling' downstream to other applications or systems (as commonly seen in work-in-progress DAM systems) so that the tags are always associated with the content;
- if the controlled vocabulary is very large, because a large set of folders may be cumbersome to browse through;
- if the controlled vocabulary includes synonyms, which tends to be supported in tag lookup, but not in categories; and
- if multiple topics are relevant for a content item, because tagging supports assigning multiple tags per asset, in contrast to categories, which may be limited to only one per asset.

Categories (such as virtual folders) and categorising is preferred under the following circumstance:

- if a single preferred means of categorising (eg content type, discipline, brand) is preferred by the users;
- if the same set of users usually work in the same category, so that team members can regularly access their 'go-to' folder;
- if the files always stay in this repository rather than 'travel' downstream to other applications;
- if the taxonomy of folders is relatively small (and there is no need for synonyms);
- if there is the desirability for a hierarchical taxonomy but none of the metadata properties support it;
- if there are problems with user compliance in tagging for such terms; and
- if users clearly prefer category folders (based on use cases).

Both tags and categories can be implemented in the same system for the same repository of content if serving different purposes. For example, categories could be broader topics than any of the topics in the tags, or categories can be for a different method of categorising than covered in tags, such as content types instead of topics.

METADATA AND TAXONOMIES

There is significant overlap between taxonomy and metadata. The term *metadata*, otherwise known as 'data about data', refers to all the recorded, structured information about a content item, document, digital asset or webpage. Taxonomies (or more

generally, controlled vocabularies) are often, but not always, used for metadata, and much, but not all, of metadata utilises controlled vocabularies or taxonomies. Metadata properties or fields get filled or 'populated' with specific controlled vocabulary terms as appropriate for each individual content item.

Different types of metadata serve different purposes. The National Information Standards Organisation (NISO) defines three kinds of metadata: descriptive, structural and administrative.[3] *Descriptive metadata* includes information on what a resource is about, expressed in keywords or short descriptions; and it also includes other descriptive information that could be used to look up and retrieve the item, such as title, author and document type. *Administrative metadata* describes information needed to manage a resource, such as its creation date, size, access rights, intellectual property rights and archival preservation information. *Structural metadata* describes the relationships of parts to one another, such as the sequence of content items in a series. There are other methods besides NISO for classifying metadata types, but most methods distinguish between metadata for aiding in search or discovery and retrieval of content and metadata for managing content.

Taxonomies are associated with the descriptive type of metadata, for two reasons. First, taxonomists, by the nature of their work, are focused on the goal of descriptive metadata, which is to help users find content. Secondly, descriptive metadata tends to use taxonomies more than other types of metadata do. If administrative or structural metadata properties require controlled vocabularies, these tend to be short, flat lists of values and not taxonomies.

Regardless of the type of metadata, (descriptive, administrative or structural), a specific metadata element or property or field may either allow free text or require the user to select from a controlled list of options. A controlled vocabulary is, of course, a type of controlled list, but a controlled list may be simpler. For example, the controlled list for a metadata property may consist of just a pair of values, such as yes or no, male or female, or new or used, or it may consist of just three or four values, such as small, medium and large. These types of lists are sometimes not considered controlled *vocabularies*, because part of the definition of a controlled vocabulary is that a term is designated for a concept, and concept-naming decisions need to be made.

Controlled vocabularies of any size, including hierarchical taxonomies, may be used to support one or more descriptive metadata properties, especially a property that is called *Subject*, *Topic* or *Descriptor*. A taxonomist is not necessarily responsible for all metadata, so he or she needs to work in collaboration with a metadata architect, metadata librarian or content architect, especially in the blurred area of responsibility between short controlled vocabularies and long controlled lists. In addition to determining the metadata properties and their values, other decisions need to be made: whether assigning/tagging values from a specific metadata property is required or optional, whether a metadata property may hold only one value or can permit multiple values and whether the property will be displayed in the user interface for end-user search-and-retrieval purposes.

While the majority of taxonomies are implemented as metadata, if a taxonomy is implemented in a way that the terms, unlike other metadata, are not attached to a content item, then the taxonomy might not be utilised as metadata. An example would be navigational topics on a website, where the topics are hyperlinks to pages. Another example would be a taxonomy that is implemented to support dynamic auto-indexing or search, and executed 'on the fly', rather than being permanently attached to a record, and then it is not metadata.

FACETED TAXONOMIES AND METADATA

A faceted taxonomy comprises a set of facets, each facet containing an individual controlled vocabulary whose terms are generally not linked/related to terms in the other controlled vocabulary facets, but the combination of terms, each selected from a combination of facets, is used to tag the same set of content, and users limit the search/filter on terms in combination from various facets. Examples of facets may be Product/Service, Market Segment, Location, Content Type, Supplier, Channel, etc. The user interface may also include refinements/filters, which do not utilise taxonomy terms, such as author, price or date. Figures 1 and 2 provide examples of excerpts of faceted taxonomies.

A faceted taxonomy is a common type taxonomy, whether for as enterprise taxonomies, DAM taxonomies, or e-commerce or product review taxonomies. It is called a 'taxonomy' even though it differs from the classical hierarchical 'tree' type of taxonomy, because it involves controlled vocabulary and classification. The name for each facet plus the terms within the facet constitutes what is essentially a simple two-level hierarchy.

Each facet is also a metadata property/element. The taxonomist designing a faceted taxonomy is thus also designing metadata, or at least some of it. There are usually more metadata properties to describe the content beyond those which comprise the taxonomy facets, because metadata can serve additional purposes beyond helping users find content. Metadata may describe content for purposes of full identification, source citation or information on how the content can be used, including rights data. Designing the full set of metadata may be the responsibility of a metadata architect rather than a taxonomist.

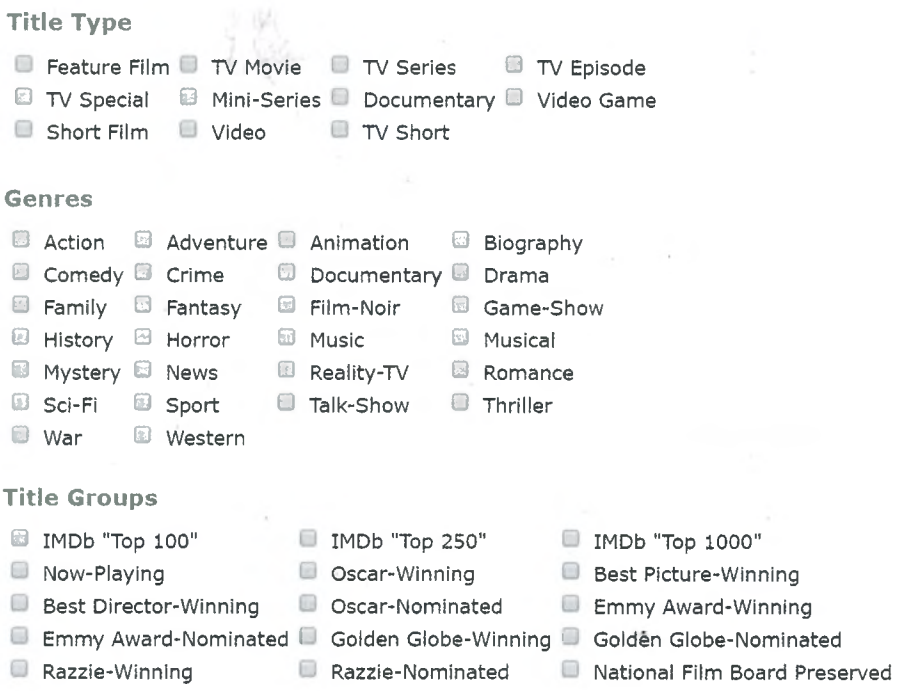

Figure 1: Select facets from the Internet Movie Database
Source: www.imdb.com/search/titles.

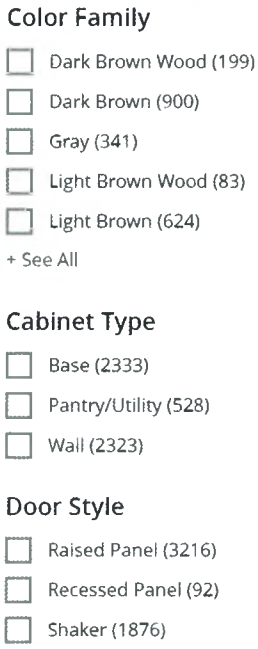

Figure 2: Facets from Home Depot
Source: www.homedepot.com.

Meanwhile, there may be additional metadata properties beyond the scope and definition of 'taxonomy' that are nevertheless made available to the end user to filter/refine results alongside the other, taxonomy facets. These could be for author/creator, date, title keyword, text keyword, file format, etc. Sometimes the distinction between taxonomy facet and other metadata in this case is not so clear, such as for Document/Content Type, Audience or Language, when these properties utilise controlled vocabularies.

DESIGNING FACETS

Designing facets overlaps with designing a metadata schema, but facets are displayed to the end users for their interaction. So, facet design needs to take the user interface and user experience into consideration. Following are some issues in designing usable facets.

For a faceted taxonomy to best serve the user who is trying to find/discover content based on what it is and what it is about, the number of facets should be limited, perhaps 5–10, keeping in mind that there may be additional, non-taxonomy refinements, such as date. Additionally, user interface space limitations may make even fewer facets preferable. Because the facets are few, they should be presented in some logical order, not in a default alphabetical order.

As for the number of terms within a facet, ideally these are also somewhat limited, so that they are not too many to be viewed easily in a short list or in a scroll box (without too much scrolling). Often the first three or four terms within a facet are displayed, and there is an option to click on 'more' to see the full list of terms within that facet. Thus, the number of terms may be 2–25, with only one or two exception facets that have far more terms. This way, the user can more easily keep track of selections of terms from multiple facets. Exceptions for facets with more terms include alphabetical

lists of known (anticipated) entities, such as states or countries, and a larger generic topic facet.

Other considerations in designing facets for the DAM or CMS include the following:

- ability to select multiple values from within the same facet at once (typically by means of check boxes);
- including other metadata (not 'taxonomy') in the same set of displayed facets (date, creator, price, etc);
- having all generic facets, the same in all contexts or also having some category-specific facets; and
- supporting a hierarchy of terms (no more than two levels recommended) within a single facet.

Designing facets is an integral task to designing and specifying all descriptive metadata, of which a faceted taxonomy is part. Due to this overlap and blurred distinction between taxonomy facets and displayed metadata for filtering, it is a good idea to design the taxonomy and metadata specification together as an integrated strategy.

STANDARDS AND POLICIES FOR TAXONOMIES AND METADATA

Standards serve various purposes. Two leading purposes for standards are

- to ensure consistency and ease of use across different products or systems used by different users; and
- to ensure interoperability, the sharing or exchange of products/services/information.

Published standards for taxonomies and other controlled vocabularies are typically for the first purpose, of enabling consistency and east of use, and this is by means of best practices guidelines for term format/style and relationship types between terms. The leading standards are ANSI/NISO Z39.19 (2005, renewed 2010) 'Guidelines for Construction, Format, and Management of Monolingual Controlled Vocabularies' and ISO 25964-1 (2011) 'Thesauri and Interoperability with Other Vocabularies, Part 1: Thesauri for Information Retrieval'. There is a lot of overlap between the two. Taxonomies should be designed to follow such guidelines to the extent practical. The guidelines are especially relevant to correctly structuring hierarchical relationships. Such standards, however, are still not sufficient for any taxonomy implementation. Additional customised policies for the taxonomy maintenance and indexing use should be created as part of an overall taxonomy governance plan.

Metadata in general do not require such standards for ease of use. Rather, there are 'standards' of the second type, for interoperability for metadata, and these are known a metadata models or schema. There are a number of different published standards for different kinds of content, so each may be considered as just a suggestion. These include Dublin Core Metadata Elements for digital content, MARC (Machine-Readable Cataloging) for library materials and IPTC (International Press Telecommunications Council) for photographs, just to name a few. If an organisation does not have the need to exchange its fully tagged content externally, then there is no need to follow an established metadata schema. Rather, an organisation should develop its own internal, customised metadata schema. Like internal taxonomy policy, a metadata schema defines the policy for maintaining the metadata and how it should be applied.

Specifically, a metadata schema lists exactly what each of the metadata properties are, provides definitions for those properties

and spells out rules for use of those properties. Rules for metadata properties may include

- whether or not the property field is populated with terms from a controlled vocabulary;
- what the source entering terms is (automatically generated or human-created and by whom);
- whether applying a term from the property is required for each content item;
- whether only one term, a limited number or any number of terms can be applied with in the same field; and
- whether the application of term in one property are dependent on terms applied in another property;

These last three issues (required, number and dependency) are also relevant to facets in a faceted taxonomy.

DESIGNING TAXONOMIES ALONG WITH METADATA

As taxonomies and metadata are integrated, there may be uncertainty whether to start with creating the overall metadata strategy and schema and then build taxonomies as part of it as needed, or to start with creating a taxonomy and then, in the process, identify the various descriptive metadata. Ideally, the two are developed for implementation combination, as part of an integrated strategy. An expert in taxonomy development (a taxonomist) and an expert in metadata design (a metadata architect), however, are usually not the same person.

A metadata architect (one who develops, implements and manages metadata strategy, architecture and policies) can acquire taxonomy-creation skills, and a taxonomist can acquire metadata architecture skills, or the two individual experts can work together on the same project. A smaller organisation, however, might not have both types of experts on staff. Whether such an organisation has a metadata architect or a taxonomist depends on the nature of the organisation's content and content organisation needs.

Organisations that start with the metadata expertise and approach to information management tend to be those with significant needs in DAM (with image or other media collections), records management (in highly regulated industries), publishing or cultural preservation (museums or libraries). Organisations that start with the taxonomy expertise and approach include product or service providers, distributors and retailers (especially in e-commerce), and organisations focused on providing information resources.

In conclusion, well-designed taxonomies and metadata will facilitate the management and retrieval of digital assets or other content. As taxonomies and metadata are integrated, their design and development should also be an integrated process, whether undertaken by an individual or an interdisciplinary team. A good understanding of taxonomies and metadata is needed to choose the best type of taxonomy and select or design the appropriate metadata model.

References

1. International Organization for Standardization. (2013) 'ISO 25964-2:2013 Information and documentation — Thesauri and interoperability with other vocabularies — Part 2: Interoperability with other vocabularies', International Organization for Standardization, Geneva, p. 4.
2. *Ibid.*, p. 14.
3. Riley, J. (2017) 'Understanding Metadata: What is Metadata and What is it For?', National Information Standards Organization (NISO), Baltimore, MD, p. 6.