Original Article

Search relevancy got you down? Learn how to use metadata to lift your results (and your mood)

Marko Hurst

is a consultant, author and speaker in the fields of Search, Measurement, Content Strategy and User Experience. He is the author of *Onsite Search Analytics: Conversations with Your Customers* and currently the Measurement Practice Lead and Account Manager for 'MISI Company' based in New York City, where he combines a hybrid quantitative and qualitative background into a single unique practice and offering. This combination provides not only a complete user-model, but a structured, repeatable and successful framework for success that has increased profitability for MISI Company and its clients. MISI Company Ltd., an NTT DATA Group company, provides Experience Design consulting services and Technical Staffing to a diverse client base on a global level. The Experience Design (XD) practice is the art and science of ensuring the success of every interaction between your business and your target audience. Regardless of how you measure success – increased sales, customer loyalty, greater efficiency, faster adoption, shorter development time – well-designed experiences move you closer to your goals. MISI provides objective, evidence-based guidance necessary to craft successful interactions. MISI XD comprises four areas of competency. Together, our Architecture, Creative, Measurement and Technology teams provide a comprehensive set of core skills that enable you to define, design and deliver successful experiences.

ABSTRACT When it comes to search an information retrieval system within Digital Asset Management (DAM) systems, the burden is often placed on the users who are expected to solve the problem of irrelevant search results. Why? The biggest contributor to this problem is simply an overreliance on algorithms alone to deliver relevant results, when it's a human linguistic issue. Machine-understandable information, semantic, algorithms will not be advanced enough for many years, if ever, to properly understand and interpret the nuances of 'natural language'. To deliver the most relevant results possible today you need more than algorithms, you a human understanding of language. You need someone who understands content modeling and structure, your audience's natural language, and your business vernacular in the form of metadata: taxonomies, ontologies and metadata schemas. To compound the issue of delivering relevant results further, there is an industry wide lack of knowledge and or practice of measuring the quality of search results, that is search analytics. Search analytics provides a quantifiable way to measure the relevancy of search results, as well as inform how well or poor the design and overall user experience are performing against your business objectives.

Journal of Digital Asset Management (2010) 6, 285–290. doi:10.1057/dam.2010.30

Keywords: metadata schemas; taxonomy; search analytics; user experience; key performance indicators; search

SEARCH AND THE USER EXPERIENCE

Everyone who has ever used a search function, within a Digital Asset Management (DAM) or any other type of information retrieval (IR) system, has experienced something similar to the following:

A user enters a query or query phrase into the search field and then selects a few pre-search

filters to narrow their results to their specific region and file types they are looking for. When they are satisfied, they click the search button and the page refreshes. They immediately begin to scan the Search Engine Results Page (SERP) seeking visual clues (called information scent) that will indicate that they have been successful in their search and will allow them to move on. As the user scrolls down the page he user becomes

Correspondence:

Marko Hurst
Search & Measurement,
MISI Company, 45 W.
36th Street, New York
10018, USA

frustrated with not only too many results, but finds them mostly irrelevant to his query.

Depending on the level of frustration, domain expertise and search expertise the user is experiencing, they will perform one or more of the following search behaviors, or search patterns¹:

1. *Pogosticking*: A user repeatedly bouncing between the SERP and the individual results.
2. *Thrashing*: A user sets the anchor with their initial query, and despite poor results insists on small variations of the flawed search phrase rather than trying a new approach.
3. *Refine*: A user will broaden or narrow criteria in an attempt to get more relevant results.
4. *Quit*: A user quits and leaves the application before achieving goal.

We have all experienced the above scenario in one form or another. In fact, it happens so often that having a negative search experience has become a common experience. While there are many factors that contribute to successful or failure of any IR system, in the author's experience he has observed and experienced one major contributing issue to irrelevant results more than any other – *an overreliance on technology alone, algorithms, and not leveraging metadata and metadata structures to deliver relevant results*.

METADATA FOR BEGINNERS TO ADVANCED

When working with metadata solutions, there are levels of complexity that must be considered when deciding what and how to implement or update a new approach, as well maintenance post launch. There are three main metadata initiative types that we will discuss in order from the cheapest and easiest to develop and maintain the most expensive and most complex to implement.

- *Beginner*: Metadata Schema
- *Intermediate*: Taxonomy
- *Advanced*: Ontology

Keep in mind that the metadata structures discussed within this article are NOT meant to replace technology for delivering results, but

instead are meant to be an additional layer of support that will help deliver more relevant results than algorithms alone. How? By adding a level of humanness that is still required to understand the nuances of the human language.

METADATA OVERVIEW

Whether you utilize a metadata schema, taxonomy, ontology, or any combination of the three, it all comes down to metadata.

What is metadata?

The most common definition for metadata is 'data about data', but something a little more useful is that it 'describes other data'. There are different types of metadata that include:

1. *Content Metadata*, which describes the form and contents of information of any type (paper-based or electronic media) and where it can be found.
2. *System metadata*, which is systemic data that describe a system or application. It is often helpful to treat an application as if it were a piece of content. The application can be targeted to a user in the same way a piece of content can.
3. *User metadata*, which describes the user and used within user profiles.

In this article, we will focus on *Content Metadata*.

Content metadata

Content Metadata is used in both digital and non-digital information storage systems to describe the form and contents of information of any type, paper-based or electronic media. Content metadata describes what a particular piece of content – for example, book, research article, MS Word document – is about and where it can be found. Examples of metadata items include: article name, author name, date of publication, abstract, ISBN number and so on.

Digital publishing – for example, DAM Systems, Content Management System (CMS), Search Engines and so on – relies heavily on metadata to match the right content to the right users. Both push and pull technologies rely on metadata.

In the electronic domain, metadata takes the form of tags associated with content. Attaching

metadata facilitates the efficient, effective management, location and retrieval of information, be it in the form of web pages, electronic documents or data in databases.

When applying metadata to content, the metadata attributes describe what kind of content this is. When applying metadata to users, attributes can define what the user sees as relevant. Defining metadata for content greatly improves the sophistication of searches, as search can now be carried out against a range of dimensions; for example, allowing searches by category, region, creation time, role relevance, interest and so on.

In addition, outside the context of a search, metadata becomes the mechanism by which content similarity can be assessed. Content items that share similar and related metadata attributes are likely to contain similar and related content. Definition of metadata attributes for users extends the potential for filtering even further. Examining the similarity between the content and user metadata can assess the relevance of a particular piece of content to a particular user. Obviously, for this comparison to be effective, both the user and content must possess similar metadata attributes. For example, if content is classified by region, the user must have metadata recording their interests in different regions.

CONTENT MODELING: BEGINNERS

Best practice dictates that metadata elements developed along an industry standard metadata schema. There are a number of these standards, a few include:

- IEEE²: Institute of Electrical Engineers
- PRISM³: The Publishing Requirements for Industry Standard Metadata
- MODS⁴: Metadata Object Description Schema (Library of Congress)

Perhaps the most famous and best suited for the web is the *Dublin Core Metadata Initiative* (DCMI). The DCMI utilizes a base of 15 elements that can be used to describe digital and physical objects.

The DCMI comprises the following 15 elements⁵:

1. *Contributor*: An entity responsible for making contributions to the resource.

2. *Coverage*: The spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant.
3. *Creator*: An entity primarily responsible for making the resource.
4. *Date*: A point or period of time associated with an event in the life cycle of the resource.
5. *Description*: An account of the resource.
6. *Format*: The file format, physical medium, or dimensions of the resource.
7. *Identifier*: An unambiguous reference to the resource within a given context.
8. *Language*: A language of the resource.
9. *Publisher*: An entity responsible for making the resource available.
10. *Relation*: A related resource.
11. *Rights*: Information about rights held in and over the resource.
12. *Source*: A related resource from which the described resource is derived.
13. *Subject*: The topic of the resource.
14. *Title*: A name given to the resource.
15. *Type*: The nature or genre of the resource.

These core elements can convey metadata information in three principle ways:

1. As an element themselves.
2. Qualified by an extension, for example DC.Subject.Keywords = 'information retrieval'.
3. With reference to a scheme, that is, controlled vocabulary.

The extensibility of Dublin Core metadata elements through the use of an extension, that is, adding the word 'Keywords' to the core element 'Subject', as in the above example, or a controlled vocabulary demonstrates the flexibility of the schema. This flexibility really means it is possible to develop a bespoke metadata framework for any situation or client.

Developing a metadata schema

Developing a metadata schema is relatively straightforward, using the Dublin Core element set. The skillful part lies in only choosing those elements that make best sense for the business, to make sure the metadata schema is scalable and to keep the elements to a minimum.

SURFACING CONTENT: INTERMEDIATE

If and when it becomes fully realized, the future of surfacing content lies in perhaps what is known as the ‘Semantic Web’, or semantic search. Semantic search has been a hotly debated topic for some time now, but if it becomes a reality it will be able to deliver on the promise of understanding the linguistic relationships between words and phrases. Until and if that day ever comes to fruition, there are currently two main paradigms of surfacing content at your disposal: Algorithmic and Taxonomic.

While both paradigms have the same goal of delivering relevant results, each paradigm approach is unique in how it goes about achieving that goal. Algorithmic search obviously utilizes an algorithm and is entirely technology-based. Taxonomic search leverages linguistics in the form of metadata. Each approach inherently comes with its own set challenges, advantages and disadvantages that are discussed below.

Algorithmic search

Algorithmic Search, as the name conveys, is an algorithm that surfaces content through performing a search, making it a *pull*-only event. In order to surface content in this manner, the algorithm itself relies on textual matching and the parameters (weighting, stemming, ranking and so on), defined to be able to serve relevant search results.

How algorithmic search works?

- Two-step process: search and contents.
- Search is performed against the content of a database repository.
- It uses only the keyword(s) entered into the search engine.
- Where there are an exact match, a result is returned.
- This gives results which are less expansive, less controlled and often, less relevant.

Taxonomic search

Taxonomic Search, on the other hand, surfaces content by interrogating the taxonomy (hierarchical relationship) and the ontology (associative relationships), if it exists. Unlike

algorithmic search that is restricted to search only, taxonomic search is unique not only in its ability to search, but also provide customization and personalization, allowing it to leverage both *push* and *pull* within the system to surface relevant content.

How taxonomic search works?

- Three-step process: search, metadata and contents.
- Searches are performed against the metadata associated with content stored in a database repository.
- The metadata, not the search itself, is mapped against the contents of a database repository.
- This gives results that are more expansive and relevant.

Benefits of taxonomic search

- A primary benefit of the taxonomy is the simplification of search. Rather than force the user to refine or filter their search in the user interface (UI) such as with advanced searches, the filtering can be performed by the taxonomy all behind the scenes. This can help remove roadblocks and burden from the UI and the user.
- Aids with the precision of searches by matching documents tagged with taxonomy terms, defined synonyms and related terms, search results include information relevant only to the users’ entered search terms, for example ‘DAM’ maps to ‘Digital Asset Management’.
- Can match the user’s search word entered into the search engine with related terms mapped in the taxonomy (for example ‘Oil Prices’ might be mapped to ‘Automotive Industry’; ‘Steel Prices’ to ‘Shipping Industry’).
- Search results can be prioritized and categorized by filtering for pages and/or documents associated with specific search terms, for example ‘best bets’.
- Can provide query suggestions, that is auto-complete: Query keyword and phrase options are provided based on what is being entered into the search box. Where there is ambiguity, can ask users to refine their searches by providing ‘Did you mean ... ?’ feedback.

How taxonomy aids searching?

Taxonomy aids searching by:

- Restricting searches within a finite category or set of categories, for example a search for 'Ancient Rome' will be restricted to the category 'Ancient Rome'.
- Searches can be performed against a higher parent, lower child or equivalent sibling nodes: for example, a search for 'Rome' would search across 'Ancient Rome', 'Modern Rome', 'Roman Technology', 'Roman Emperors' and so on.
- Provide 'Did you mean?' feedback to users to refine searches, for example a search for 'Rome' might return: 'Did you mean Ancient Rome, Modern Rome, Roman Technology, Roman Emperors ... ?'
- Search against synonyms, that is alternate terms, for example a query against the acronym Search Systems' would map to, and search within the 'IR' category.
- Search against related terms, for example 'Julius Caesar' is a term that exists in both 'Ancient Rome' and 'Roman Emperors'. As this relationship is known (and mapped) in the taxonomy, searches on one usage will also return 'results' from the other.

What can taxonomies be used for?

In the web space, taxonomies classify domains of knowledge and show the hierarchical relationships between categories, sub-categories and values within categories. Taxonomies are often created and used for a variety of purposes including:

- Developing navigation systems.
- Understanding content and the relationships between content items.
- Providing an organizational principle for the Content Management System.
- Populating related links determining ontological relationships.
- Developing faceted search capabilities.
- Approved auto-populate certain Metadata Schema elements, for example DC.Subject. Keyword = 'information retrieval'.
- Driving cross-sell.
- Personalization – targeting of specific content to specific users, according to user profiles and business rules.

- Security – access to site areas, content, portal pages and so on.
- Collaborative filtering – targeting of specific content to users, according to the behavior of users with similar profiles.
- One consistent taxonomy reinforces 'one voice'.
- Saves resources, as products are classified consistently, once, everywhere.
- Reduces or eliminates duplicative efforts.
- Makes internal and external website searching easier.
- Enables website profiling, filtering and pushing information.
- Supports category codes and descriptions in multiple languages.

Developing a taxonomy

Taxonomies can be developed by one of the following methods:

- Manually
- Automated
- Combination of both

RELATING CONTENT: ADVANCED ONTOLOGY

According to Wikipedia's definition,⁶ 'The core meaning within computer science is a model for describing the world that consists of a set of types, properties, and relationship types'. That is, an ontology is an associative relationship between more than one entity.

An example for 'Julius Caesar' would be relationships with the First Triumvirate, Gnaeus Pompeius Magnus, Pompey the Great, Gaul, Marcus Junius Brutus, Political Assassination, Ides of March, Republic, Military Commanders, Dictators, Shakespeare and so on.

Ontologies generally describe:

- *Individuals*: the basic or 'ground level' objects.
- *Classes*: sets, collections or types of objects.
- *Attributes*: properties, features, characteristics or parameters that objects can have and share.
- *Relations*: ways that objects can be related to one another.
- *Events*: the changing of attributes or relations.

In the context of DAM systems, ontologies are about associating content in usage. Ontologies are

especially useful in defining associative relationships between content items. Sometimes these relationships can be between concepts and other like concepts, for example MPEG4 and MOV; sometimes between concepts and whole categories, for example MPEG4 and video; and sometimes between whole categories themselves, for example audio and video. Often, these relationships are obvious and straightforward, for example, 'Leave of Absence' and 'Holiday Pay' and might even logically form part of the same node or category within a taxonomy, but they may equally well be surprising, even incongruous.

For an automated approach, there is analysis software that can data-mine DAM system data to identify and analyze such seemingly incongruous patterns to gauge their appropriateness for marketing exploitation. Ontologies, then, can be defined from the top-down, identify and document known, common or anticipated relationships, or from the bottom-up providing the unusual, unpredictable or even serendipitous patterns through data-mining. These latter, so-called 'emergent ontologies' come to the forefront in collaborative filtering/recommender systems and are an extremely useful for personalization.

Benefits of ontology

- Populate related content portlets, widgets and so on.
- Target related content to consumers of particular content without knowing anything about the user or having to know anything about them, but simply by monitoring usage behavior.
- Fuel recommender systems.
- Facilitate personalization.
- Drive cross-sell.

NOTES

- 1 Search Patterns – Peter Morville & Jeffery Callender. O'Reilly books ISBN: 978-0-596-80227-1, <http://oreilly.com/catalog/9780596802288>.
- 2 IEEE, <http://www.ieee.org/index.html>.
- 3 PRISM, http://www.idealliance.org/industry_resources/intelligent_content_informed_workflow/prism.
- 4 MODS, <http://www.loc.gov/standards/mods/>.
- 5 Dublin Core Metadata Initiative, <http://www.dublincore.org/documents/dces/>.
- 6 Wikipedia, [http://en.wikipedia.org/wiki/Ontology_\(information_science\)](http://en.wikipedia.org/wiki/Ontology_(information_science)).

Disclaimer This article mentions several organizations. Mention does not imply endorsement of any kind by the author or by his employer, MISI Company or any of its partners or affiliates.