DCPAPERS

The Microsoft Profiling Taxonomy

Amy Sweigert
Microsoft Corporation
Tel: (425) 706-2409
Fax: (425) 936-7329
amysw@microsoft.com

Abstract:

This paper provides an overview of the Microsoft profiling taxonomy, including its use of the Dublin Core metadata schema.

Keywords:

Dublin Core metadata schema, enterprise taxonomies, controlled vocabularies

1. Overview of the metadata landscape at Microsoft

The Microsoft profiling taxonomy represents an attempt to implement Dublin Core, with Microsoft-specific extensions, as well as controlled vocabularies to support specific elements. This effort began in 2001.

Historically, there have been two parallel taxonomy efforts at Microsoft:

- **Content taxonomy:** Used for tagging content with standard metadata to improve search (both on the intranet and on Microsoft.com).
- **Data taxonomy:** Used for collecting standardized data about customers and customer interactions (for example, tracking sales and marketing efforts).

Both efforts have required integration with other corporate data systems, such as product release and distribution information (SKUs, pricing, etc.). The need to integrate these two parallel systems into a single profiling taxonomy is becoming increasingly apparent as we develop plans to support targeted content delivery to customers (for example, delivering content based on the products they own), and to deliver training to Microsoft employees based on their job role (as tracked by human resource systems).

Data taxonomies within the company tend to be more standardized, both because the data is used for tracking and reporting revenue, and because we use a common set of tools for revenue tracking. However, the taxonomy is only updated annually and there is no automatic way for systems to consume it. As a result, compliance and data quality are inconsistent across groups.

On the other hand, content taxonomies have been *less* standardized, because Microsoft lacks a standard toolset or process for managing and publishing content. Being a software company, we only occasionally mandate use of a single toolset. More often, individual groups develop custom solutions for their specific problem areas. Because new content and customer search queries are added continually, the content taxonomy needs frequent updates and changes need to flow to consuming systems in a predictable and automated way. To use a common taxonomy requires consuming systems to make technical updates and coordinate with other teams. Individual toolsets represent an easier solution in the short term, but do not support data sharing in the long term.

In general, Microsoft has an entrepreneurial culture that is not friendly to mandated standards. As a result, the value of the taxonomy has to be constantly reinforced, both to executive management and to individual groups trying to implement standards in their systems.

2. Content taxonomy consumers

Portions of the content taxonomy are currently consumed by 14 content groups across the company, representing both internal and external content.

3. Metadata schema

The content metadata schema being driven across the company is based on the Dublin Core (DC)

https://doi.org/10.23106/dcmi.952108336

schema. DC elements used at Microsoft include: contributor, coverage (scoped to geographic coverage), creator, description, identifier, subject, title, audience, dateCreated, dateIssued, and dateModified.

In addition, the schema includes a set of extensions, some appropriate for internal Microsoft content and some for customer-facing content. Examples of extension elements that are required include: characterSet, contentID, cultureCode, lastModifiedBy, submitterEmailName, confidentiality, and contentState.

Some portions of the existing DC schema have proved problematic for Microsoft to implement.

- **Format:** File extension is a more common way to capture this than MIME type, and is more easily extracted from files.
- **Language:** The values for language need to be locale specific, we use the cultureCode element instead.
- **Relation/Source:** These overlap conceptually, source is really a specific type of relation.
- **Type:** This element name is too general to be useful, we have named it contentType in the Microsoft schema.

In addition to modifying some portions of the DC schema, we have extended portions of the schema by creating domain-specific sub-elements. For example, the audience element has the following extensions: organization, customer segment, and job role.

The users of the Microsoft content metadata schema are providing access to content with a diverse set of goals, audiences, and content domains, so even within Microsoft a single schema adopted by everyone is not possible.

Ideally, we want to manage the pool of elements centrally. Individual groups could customize the core schema to meet their needs by re-using the pool of elements and refining the supporting vocabularies (and in some cases the element properties). For example, one group might require a specific audience sub-element that is optional in the core schema. Or a group could specify the set of product names that would be available for tagging a certain set of content.

We also want to be able to support mapping of legacy or third-party schemas that cannot comply with the standard. For example, content built into shipped products cannot be updated, so the prior schema must be continually supported, but must comply with the standard in the enterprise search environment. Today we do not have a standard way to manage or implement schema mappings.

4. Controlled vocabularies

Currently, the Microsoft taxonomy contains approximately 50 vocabularies and approximately 60,000 terms (both authorized and equivalent). Terms can be re-used across multiple vocabularies.

Vocabularies are collections of terms within a given domain (for example, product names, industries, geographic place names, content types, etc.). We define scope, standard term form, structural guidelines, and business ownership for each vocabulary collection.

Each term within a vocabulary has a term string, a definition, a globally unique identifier (GUID), and may also have localized term strings.

The taxonomy supports four classes of relationships: Hierarchical, Equivalence, Associative (Directional), and Reciprocal Associative (Bi-Directional). In addition, named types can be defined within each class. Examples are version, country/region, legal name, code, source, and instance-of. Hierarchical and Equivalence follow standard thesaurus management guidelines. Associative relationships are sometimes used to capture relationships that are traditionally considered hierarchical, such as instance-of. An example of this is that named products and generic products are maintained in two separate vocabularies. In order to create instance-of relationships between the two, associative named edges are used. Therefore, “Microsoft SQL Server” is an instance of “relational database management systems.” Associative relationships are also used extensively to identify subsets of vocabularies that are relevant to a particular user group. For example, individual terms are related to a consuming group name with a “used by” relationship. This allows core vocabularies to be maintained in a standard, central vocabulary, rather than having terms re-used in multiple small vocabularies customized for individual consumers.

Microsoft uses a tool called Taxonomy Manager (TaxMan) to create, manage, and store controlled vocabularies. The tool is a Web application built on a SQL back end. Taxonomy data is available to users via Web services.

5. Governance and change management

Currently, the content metadata schema is managed by a cross-company working group. The metadata schema is not currently consumed in an automated or validated way, so change management of the schema has not been formalized. Suggestions and refinements are evaluated by the working group and decisions are documented in a spreadsheet. Compliance with the schema is considered a voluntary best practice.

Change management for the vocabulary data is more formalized, because many downstream systems consume regular feeds of data. Certain types of changes (new terms added, term string edits, name

https://doi.org/10.23106/dcmi.952108336

changes, etc.) are not proactively communicated to users. Other types of changes (deletions, demotions, moves from one vocabulary to another) are communicated using a monthly notification process. Once notified, groups are given a week to respond before the changes are implemented in the system.

A major challenge has been the inability of many consuming systems to accommodate taxonomy changes, as well as the inability of the taxonomy management system to communicate changes in an automated way using Web services.

6. Integration with other data sources

In several cases, vocabulary domains that are needed for standard metadata creation are defined, owned, and stored in other corporate systems such as product release systems, human resource systems, and so on. Rather than duplicate research and data validation efforts, we try to consume vocabulary data from authorized sources and publish it out through our Web services to taxonomy consumers. For example, we have implemented a nightly feed of new product names and GUIDs from the product release system into the taxonomy management system. In this way, the same product GUIDs used to tag content can be used to query other corporate systems for part numbers and pricing information.

7. Next steps: Implementation challenges

One of the biggest challenges Microsoft faces is the lack of robust metadata tagging tools that can make use of the rich controlled vocabularies that are now available. Well-developed vocabularies require a significant investment in time and robust development tools. Until the taxonomy reached critical mass, there was little incentive to develop systems that take full advantage of the vocabularies and metadata. As a result, most authoring or tagging tools in use today only support flat lists of terms, usually without making definitions or variant term forms visible. In the future, we want to have tagging tools that support:

- A hierarchical browse of vocabularies.
- The ability to search against vocabularies, including equivalence terms.
- Search that is not limited to an element, but that populates the correct element based on term selection.

Another problematic issue is tagging with GUIDs versus human readable strings. Metadata records need to store both, and periodically refer back to the taxonomy store to validate the term strings.

Justifying improvements in Microsoft's taxonomy management and metadata creation tools has required showing successful, small-scale implementations. These have also been important in driving adoption out to the larger community.

Because we are only now developing tools that support rich metadata tagging, integrated systems with consistently high-quality data and metadata-driven features will likely follow in the next wave of investment, when metadata is applied to a critical mass of data.

It is important to emphasize the value of standard taxonomy beyond the information retrieval or content management space. Standard taxonomy is important for corporate data systems, and metadata schemas for other object types, such as customers or products, should be developed in a way that can be easily integrated with content metadata schemas.

8. General Lessons

Implementation of standard metadata in a corporate environment requires a long-term commitment. The development of a standard taxonomy includes several steps, which need to occur sequentially:

1. Acquire or develop taxonomy development and management tools.
2. Develop a standard schema and robust vocabularies to support the schema.
3. Acquire or develop metadata tagging tools and tagging guidance that make the schema and vocabularies easy to apply.
4. Demonstrate value and refine implementation with successful small-scale implementations.
5. Apply metadata to a critical mass of information.
6. Determine that consuming systems (such as search) effectively use metadata to solve business problems.

References

1. Dublin Core Metadata Element Set, Version 1.1: Reference Description. http://www.dublincore.org/documents/dces/
2. Aitchison, Jean, et al. Thesaurus Construction and Use: A Practical Manual. Chicago: Fitzroy Dearborn Publishers, 2000.
3. National Information Standards Organization. ANSI/NISO Z39.19-1993: Guidelines for the Construction, Format, and Management of Monolingual Thesauri. Bethesda: NISO Press, 1993.

https://doi.org/10.23106/dcmi.952108336