Data Governance Series: Part 2 - The Importance of Data Taxonomy

Modern

Data governance can be a powerful agent in scaling the use and distribution of trusted data throughout the company. However, more often than not, it conjures up the idea of a central authority strictly guarding against such access. In this 3-part series, we'll cover the critical and often misunderstood components of data governance and offer perspective on how to implement data governance strategies that deliver trusted data at the speed of business. If you missed it, make sure to catch up on Part 1 – Data Timeliness.

What Is Data Taxonomy?

A taxonomy, very broadly, is a system of organized information that allows the user to classify and show relationships between things. A common example of a taxonomy is the Dewey Decimal System of library classification, in which numbers form a code that correlate to topics, subtopics, and sub-subtopics. Wikipedia illustrates the way this hierarchy is set up:

500 Natural sciences and mathematics

510 Mathematics

516 Geometry

516.3 Analytic geometries

516.37 Metric differential geometries

516.375 Finsler geometry

In the Dewey classification system, each number is associated unambiguously with a single entry in the hierarchy. A number such as 516.375 above identifies a book or other resource specifically as dealing with Finsler Geometry. That number also shows how that book relates to others above and below it in the hierarchy.

A data taxonomy uses a system of unambiguous metadata terms (such as a filename or tags attached to a file) that allow an enterprise to classify a file or dataset into important business categories. Categories can be configured in any way that meets the needs of the organization, but some common ones include the date of creation, date last modified, account name of the creator/modifier, required access privileges, personal identifying information (PII), the department that owns the dataset, and the primary business use of the dataset.

Properly designed and developed, a data taxonomy improves discoverability, observability, and security for your data. Data that is properly classified, catalogued, and tagged is usually well-governed data.

How Does Taxonomy Aid Data Governance?

A proper data taxonomy addresses many problems in your data and metadata, including:

- **Ambiguity** - The same term can have different meanings according to different business users, leading to confusing query results. Taxonomy provides a hierarchy that helps remove ambiguity. It includes mechanisms for understanding context and making meaning precise.
- **Consistency** - Sometimes the terms used in legacy applications differ from those used in newer systems. Terminology can also vary between business units. For various reasons, the data sometimes can't be re-tagged to provide consistent metadata. A taxonomy lets you create a thesaurus to map disparate terms to a single label, mitigating the impact of natural inconsistency.
- **Connections** - Taxonomies can also represent related concepts (technically also part of a thesaurus) that can be used to connect processes, business logic, or dynamic/related content to support specific tasks.
- **Incompleteness** - Data may be missing important attributes. Consistent terminology, through a taxonomy, makes it easier to identify missing fields or metadata. The missing attributes may then be added from other sources or the incomplete data may be deleted or set aside and not used.
- **Utilization** - A data taxonomy can emphasize the suitability of a dataset for a particular purpose (for example, by tags that identify department ownership and business purposes). This can increase discoverability and thereby promote the use of that dataset for the intended purpose.

How to Build a Data Catalog with Taxonomy

The first and most important step to data discoverability is a data catalog. The first essential step in building a catalog is tagging data with business vocabulary so users can easily find the data they need. A data taxonomy makes cataloging much more powerful, improving data quality and discoverability. DataOS® can automate tagging and indexing to add incoming data to your catalog immediately.

How to Build a Taxonomy for Your Data

The two keys to building a usable data taxonomy from scratch are focused changes and using the language of your users as much as possible.

Focused Changes

Focus your taxonomy on one business area at a time. Balance your choice of area by beginning with high-priority targets, while keeping your scope manageable. For example, don't begin with something like compliance with HIPAA or GDPR. Those are too large and too sweeping to start with. Save those to address after you build the taxonomies for a few smaller areas, such as marketing, sales, or security. Not only will this give you more practice with the methods of taxonomy, but much of what you build there will be needed for something like GDPR, so you're whittling the scope of that project down as you go.

Use your narrow focus to plan and keep milestones as your taxonomy progresses from one target to the next.

User Vocabulary

More than many other data projects, a data taxonomy is a team effort. Your IT team or data steward can't do it on their own. A data taxonomy needs to use the language of your business users, which means a polling process and meetings with users to learn how they think of their data.

You may add a hierarchy to your taxonomy to address the variety of terms that users may have for the same thing. If users have terms like "POS revenues," "sales," and "revenues," then you can set up the taxonomy so all of those searches point back to "sales," which is the tag that appears in your metadata. This is one of the primary ways in which a taxonomy enforces consistency and aids discoverability.

The focus of your taxonomy efforts can also help users see the value of the taxonomy to their particular projects, increasing enthusiasm and interest in developing the vocabulary for their area.

A Data Taxonomy Improves Data Value

Most modern businesses spend a lot of money on collecting their data. The ROI on that effort depends on deriving business insights from the data. A data taxonomy makes data easier to find and easier to use while improving data governance and data quality. It makes your data more valuable to your business.

BY P. SCOTT

TheModern DataCompany

The Modern Data Company
306 Cambridge Ave
Palo Alto, CA 94306
TheModernDataCompany.com
info@TMDC.io