


IRON MOUNTAIN



Demystifying Taxonomies

Susan L. Cisco, Ph.D., CRM, FAI
Program Manager, Consulting Services
March 2006

©2006 Iron Mountain Incorporated. All rights reserved. Iron Mountain and design of the mountain are registered trademarks of Iron Mountain Incorporated.

Agenda/Contents

- Why Taxonomies Matter
- What Are Taxonomies
- Build Custom Taxonomies
- Buy Pre-Built Taxonomies
- Automatically Generate Taxonomies
- Best Practices

CONFIDENTIAL AND PROPRIETARY INFORMATION OF IRON MOUNTAIN. The information set forth herein represents the confidential and proprietary information of Iron Mountain. Such information shall only be used for the express purpose authorized by Iron Mountain and shall not be published, communicated, disclosed or divulged to any person, firm, corporation or legal entity, directly or indirectly, to any third person without the prior written consent of Iron Mountain.

IRON MOUNTAIN CONFIDENTIAL AND PROPRIETARY INFORMATION OF IRON MOUNTAIN 2

Everyone Uses Taxonomies

- We use taxonomies in everyday life to help navigate 'how to search' and 'where to find' what we need



IRON MOUNTAIN CONFIDENTIAL AND PROPRIETARY INFORMATION OF IRON MOUNTAIN 3

Taxonomy Example: Grocery Store



- While the average grocery store in the US stocks over 50,000 items, we navigate our way to the products we desire with ease
- It is the taxonomy that provides users with a reference system to help them find what they need

IRON MOUNTAIN
CONFIDENTIAL AND PROPRIETARY INFORMATION OF IRON MOUNTAIN
4


Terminology

- Taxonomy – Structured, often hierarchical, classification system of categories/topics
- Ontology – Study of entities and their relations, often used for web applications
- Tagging and Folksonomies – Free-form labeling (tagging) of documents with keywords

IRON MOUNTAIN
CONFIDENTIAL AND PROPRIETARY INFORMATION OF IRON MOUNTAIN
5

Why Taxonomies Matter

1. Speedup searching information
2. Enable classification for records retention compliance
3. Leverage trapped knowledge



IRON MOUNTAIN
CONFIDENTIAL AND PROPRIETARY INFORMATION OF IRON MOUNTAIN
6

Taxonomies Speedup Searching for Records

- Not everyone searches for information in the same way. Some people prefer the somewhat unstructured search provided by Google Inc. and others prefer the taxonomy provided by Yahoo! Inc.
- With an unprecedented amount of recorded information presented to the user, taxonomies will be employed to not only facilitate search but also to limit the search to a defined category of records

The image shows two side-by-side screenshots of search engines. On the left is a Yahoo! search result for 'Records Management Iron Mountain', showing a list of related terms like 'Records Management Services', 'Iron Mountain', and 'Records Management'. On the right is a Google search result for the same query, showing a snippet of text from a website and a 'Web Results' section.

Taxonomies Enable Classification for Records Retention

- A taxonomy and properly categorized documents enable implementation of retention rules
- At Iron Mountain, taxonomies will become increasingly relevant as the systems that enable users to locate records expand to a wider audience

The diagram consists of four circular nodes. At the top is 'IMConnect SKP' (green), connected by a double-headed arrow to 'Client System' (yellow) in the center. Below 'Client System' are two other nodes: 'Digital Archive' (blue) on the left and '3rd Party Data' (purple) on the right, both connected to the central 'Client System' node.

Taxonomies Leverage Trapped Knowledge

- Taxonomies allow access to information that is not possible through simple or complex keyword of the records (eDiscovery, forensics, online shopping, research)
- Taxonomies are critical to successful ECM systems because they provide structure for unstructured content
 - Systems can enforce classification of documents
 - Lessons learned from document imaging – companies still don't destroy paper originals from scanning operations because they lack the confidence in their ability to locate documents when needed for litigation or research

Taxonomies Originated in Life Sciences

A taxonomy for records is a structured, often hierarchical, classification system of topics or subject categories

GROUP NAME	ORGANISM				
	HUMAN	CHIMPANZEE	HOUSE CAT	LION	HOUSEFLY
KINGDOM	Animalia	Animalia	Animalia	Animalia	Animalia
PHYLUM	Chordata	Chordata	Chordata	Chordata	Arthropoda
CLASS	Mammal	Mammal	Mammal	Mammal	Insect
ORDER	Primates	Primates	Carnivora	Carnivora	Diptera
FAMILY	Hominidae	Pongidae	Felidae	Felidae	Muscidae
GENUS	Homo	Pan	Felis	Felis	Musca
SPECIES	sapiens	troglydtes	domestica	leo	domestica

Scientific Name *Homo sapiens* *Pan troglodytes* *Felis domestica* *Felis leo* *Musca domestica*

IRON MOUNTAIN CONFIDENTIAL AND PROPRIETARY INFORMATION OF IRON MOUNTAIN 10

Taxonomies in Records Management

Tax	Iron Mountain Records Management Record Classes By Business Function	Records Retention Schedule
Record Class Code	Record Class Description / Record Type	Retention
TAR100	Federal 1099 Reporting Records retained to 9/30/100 for reporting. Retention Event: The retention period begins when the record is created. Exempts Include: 1099 Reporting	5
TAR110	Federal Income Tax Records that represent actual Corporate tax returns, related work papers, audits, and appeals for U.S. Federal income taxes. Retention Event: The retention period begins when the record is created. Exempts Include: Federal income Tax Returns Federal income Tax Returns Federal Tax Agreements Federal Tax Appeals Federal Tax Audits	90
TAR120	Foreign Tax Returns Records that represent actual Corporate foreign income, sales, property, and other tax returns and work papers. Retention Event: The retention period begins when the record is created. Exempts Include: Tax Returns Tax Returns Tax Returns	90
TAR130	Local and Franchise Tax Records that represent actual tax returns and related work papers for local income and	10

IRON MOUNTAIN CONFIDENTIAL AND PROPRIETARY INFORMATION OF IRON MOUNTAIN 11

Faceted Classification

- **Faceted classification is a technique for structuring a taxonomy so that records are organized into categories based on the systematic combination of characteristics (facets) of the records**
- **Instead of building one huge tree of categories, a faceted classification uses multiple smaller trees (facets) that can be accessed and retrieved either alone or in any desired combination**

IRON MOUNTAIN CONFIDENTIAL AND PROPRIETARY INFORMATION OF IRON MOUNTAIN 12

Faceted Classification Example

Course
Appetizer, Beverage, Bread, Breakfast, Bunch, Casseroles/Candy, Sofa Plate, Main Course, Salad, Sandwiches, Sauces/Condiments, Side Dish, Snack, Soup

Cuisine
African, Chinese, Eastern European, French, German, Greek, Indian, Italian, Japanese, Jewish, Mexican, Middle Eastern, North American, Other Asian, Other European, Latin American/Caribbean, Spanish, Thai, UK/Irish, Vietnamese, Other

Main Ingredients
Alcohol, Bean, Beef, Bread, Cheese, Chicken, Chocolate, Corn, Dairy, Egg, Fish, Fruit, Game, Grain, Lamb, Leafy Greens, Nuts, Pasta/Noodle, Pork, Potato, Rice, Shellfish, Tomato, Turkey, Vegetable, Other

Cooking Method
Baking, Blending, Grilling/BBQ, Microwaving, Roasting, Sauteing, Smoking, Steaming, Stir-frying, Other

IRON MOUNTAIN
CONFIDENTIAL AND PROPRIETARY INFORMATION OF IRON MOUNTAIN
13

Faceted Classification Example

Cuisine
African, Chinese, Eastern European, French, German, Greek, Indian, Italian, Japanese, Jewish, Mexican, Middle Eastern, North American, Other Asian, Other European, Latin American/Caribbean, Spanish, Thai, UK/Irish, Vietnamese, Other

Recipe Browse > North American
Narrow View Results
By Course

Appetizer (158)	Beverage (117)	Bread (166)	Breakfast (104)
Bunch (65)	Dessert/Candy (1178)	Sofa Plate (29)	Main Course (1152)
Salad (284)	Sandwiches (32)	Sauces/Condiments (277)	Side Dish (284)
Snack (73)	Soup (193)		

IRON MOUNTAIN
CONFIDENTIAL AND PROPRIETARY INFORMATION OF IRON MOUNTAIN
14

Faceted Classification Example

Cuisine
African, Chinese, Eastern European, French, German, Greek, Indian, Italian, Japanese, Jewish, Mexican, Middle Eastern, North American, Other Asian, Other European, Latin American/Caribbean, Spanish, Thai, UK/Irish, Vietnamese, Other

Recipe Browse > North American > Kid's Plate
Narrow View Results
By Main Ingredient

Bean (1)	Beef (4)	Bread (2)	Cheese (1)	Chicken (2)	Dairy (1)
Fish (1)	Fruit (2)	Game (2)	Nuts (1)	Pasta/Noodle (4)	Potato (1)
Rice (1)	Turkey (1)				

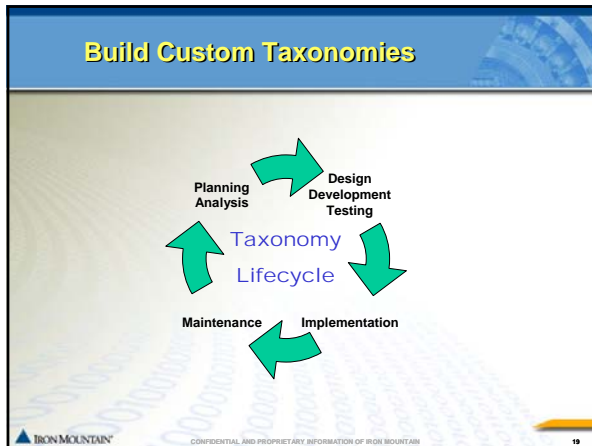
IRON MOUNTAIN
CONFIDENTIAL AND PROPRIETARY INFORMATION OF IRON MOUNTAIN
15

Faceted Classification Example

Metadata

Acquiring a Taxonomy

- BUILD Custom Taxonomies
- BUY Pre-Built Taxonomies
- Automatically GENERATE Taxonomies



- ### Phase 1: Planning and Analysis
- Assess resources and levels of expertise
 - Identify categories to be used
 - Identify records and retention requirements
 - Select development strategy
 - Develop cost models
 - Investigate outside consulting services
- IRON MOUNTAIN
CONFIDENTIAL AND PROPRIETARY INFORMATION OF IRON MOUNTAIN
20

- ### Phase 2: Design, Development, and Testing
- Design for short-term and long-term needs
 - Iterative process
 - Develop high level structure
 - Test with stakeholders
 - Modify structure
 - Repeat until consensus is reached
- IRON MOUNTAIN
CONFIDENTIAL AND PROPRIETARY INFORMATION OF IRON MOUNTAIN
21

Phase 3: Implementation

- **Change management process begins early in the taxonomy project**
- **Formalized with training**
 - Classifying new records
 - Searching and retrieving records

IRON MOUNTAIN CONFIDENTIAL AND PROPRIETARY INFORMATION OF IRON MOUNTAIN 22

Phase 4: Maintenance

- **Taxonomy will evolve over time**
- **Document decisions made throughout development and implementation**
- **Dedicate ongoing resources**

IRON MOUNTAIN CONFIDENTIAL AND PROPRIETARY INFORMATION OF IRON MOUNTAIN 23

Build Custom Taxonomies

- **Pros**
 - Custom taxonomy
- **Cons**
 - Requires investment in time and money
 - Can be political challenge

IRON MOUNTAIN CONFIDENTIAL AND PROPRIETARY INFORMATION OF IRON MOUNTAIN 24

Buy Pre-Built Taxonomies

- Taxonomy vendors
- Industry associations (POSC & PPDM for oil and gas industry; MeSH for medical profession)
- Public domain taxonomies (The Taxonomy Warehouse)
- Enterprise Content Management (ECM) companies – more of a driver than a provider

IRON MOUNTAIN
CONFIDENTIAL AND PROPRIETARY INFORMATION OF IRON MOUNTAIN
25

Buy Pre-Built Taxonomies

What You Get

- Hierarchical structure
 - List of categories/topics/nodes and tree structure of terms
- Rules
 - What are the variations (synonyms)? Broader terms? Narrow terms?
 - Retention rules - if "document type" is X, then retention period is X
 - Lexis/Nexis example

IRON MOUNTAIN
CONFIDENTIAL AND PROPRIETARY INFORMATION OF IRON MOUNTAIN
26

Rule Set – TAX FRAUD
Example LexisNexis® SmartIndexing Technology™

<p>Document Segments</p> <ul style="list-style-type: none"> Headline Lead Paragraph Body Summary/Abstract Publisher Indexing 	<p>Strong Words/Phrases</p> <ul style="list-style-type: none"> false W-2 filed fake returns IRS scam tax dodge VAT and duty fraud
<p>Strongest Words/Phrases</p> <ul style="list-style-type: none"> cheat on his taxes cheated on their taxes cheats on her taxes defraud the IRS dodging income tax evade paying taxes false claims for tax illegal tax shelter 	<p>Related Words/Phrases</p> <ul style="list-style-type: none"> delinquent tax false deduction hid income tax police
	<p>Weak Words/Phrases</p> <ul style="list-style-type: none"> deductions IRS tax taxman taxpayer allegations cheated crack down guilty suspicion

IRON MOUNTAIN
CONFIDENTIAL AND PROPRIETARY INFORMATION OF IRON MOUNTAIN
27

Pre-Built Taxonomies for Architecture

- At least 9 taxonomies for sale/license
- Long established resource is the Art and Architecture Thesaurus from The Getty (3 of the 9 taxonomies)
- Costs vary

IRON MOUNTAIN CONFIDENTIAL AND PROPRIETARY INFORMATION OF IRON MOUNTAIN 28

Pre-Built Taxonomies for Taxation

- At least 14 taxonomies for sale/license
- 9 taxonomies are available from The Taxonomy Warehouse
- Costs vary
- Test case details are in proceedings paper

IRON MOUNTAIN CONFIDENTIAL AND PROPRIETARY INFORMATION OF IRON MOUNTAIN 29

Buy Pre-Built Taxonomies

- **Pros**
 - Generally less expensive than custom development
 - Immediate results
- **Cons**
 - May not map to your business
 - Will still need IT/RM resources to integrate into existing eRecords systems and develop custom rules

IRON MOUNTAIN CONFIDENTIAL AND PROPRIETARY INFORMATION OF IRON MOUNTAIN 30

Automatically Generate Taxonomies

Automatic categorization provides the means to automatically describe or file records based upon their content, structure or metadata, thereby speeding up the filing process and improving retrieval and correlation. Automatic categorization software uses a wide variety of techniques that find similarities in the documents.*

**Auto-categorization methods include statistical (Bayesian) analysis, word similarity clustering, word frequency graphing, linguistic inference, the use of pre-existing sets of categories, and seeding categories with keywords.*

IRON MOUNTAIN
CONFIDENTIAL AND PROPRIETARY INFORMATION OF IRON MOUNTAIN 31

Automatic Categorization

A **cluster** is a group of documents that are more similar to each other than to the members of any other group of documents in the collection


A centroid is the center point of a cluster. Each subject-area cluster has an associated centroid. All documents whose feature-vector endpoints are near a given subject-area's centroid are related to that subject area

IRON MOUNTAIN
CONFIDENTIAL AND PROPRIETARY INFORMATION OF IRON MOUNTAIN 32

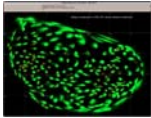
Techniques for Text Visualization

Different ways to locate groups of documents that are more similar to each other than to the members of any other group of documents in the collection


Topographical maps



Cluster maps



Networks of linked nodes



IRON MOUNTAIN
CONFIDENTIAL AND PROPRIETARY INFORMATION OF IRON MOUNTAIN 33

Automatically Generate Taxonomies

- **Pros**
 - May be less expensive than building a taxonomy
 - Enables high-level categorization of a large collection of documents without time-consuming interviews and secondary research
 - Shorter development time
- **Cons**
 - May generate more bad data than good
 - Requires subject matter expertise most organizations do not have

IRON MOUNTAIN CONFIDENTIAL AND PROPRIETARY INFORMATION OF IRON MOUNTAIN 34

Taxonomy Best Practices

- **Incorporate existing taxonomies and metadata whenever possible**
- **Make sure topics/categories are well-defined and distinct**
- **Iterative development**
- **Keep the taxonomy as simple as possible**
- **Provide for adequate resources to maintain the taxonomy**

IRON MOUNTAIN CONFIDENTIAL AND PROPRIETARY INFORMATION OF IRON MOUNTAIN 35

Conclusion

- **Taxonomies and the associated metadata can complement text searching, provide multiple access points to information, and incorporate retention requirements**

IRON MOUNTAIN CONFIDENTIAL AND PROPRIETARY INFORMATION OF IRON MOUNTAIN 36
