MDM and Taxonomy

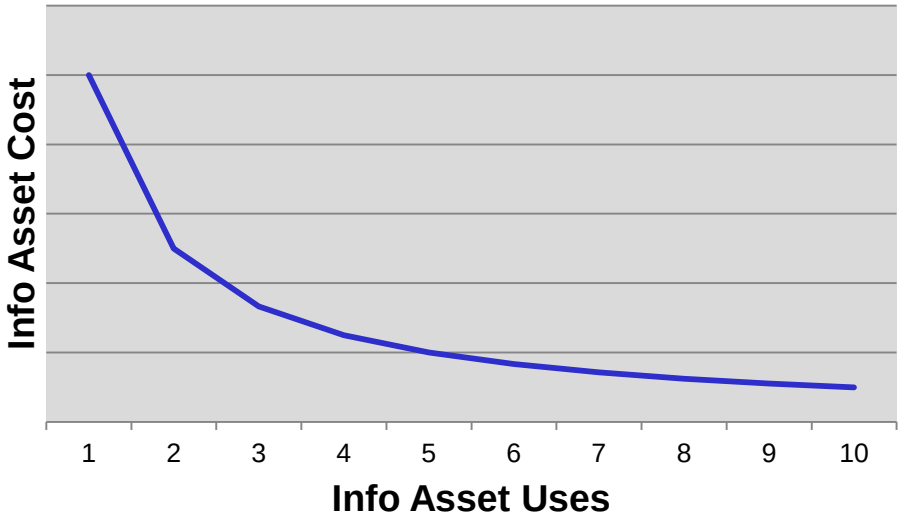
Mitre Technical Exchange Meeting

Interoperability

- **The ability of diverse systems and organizations to work together by exchanging information.**
- Semantic interoperability is the ability to automatically interpret the information exchanged meaningfully and accurately.

Interoperability ROI

- Information assets are expensive to create so it's critical that they can be found, so they can be used and re-used.
- Every re-use decreases the information asset creation cost and increases the information asset value.



Interoperability ROI (2)

- If information assets are so important, why can't they be found?
  - **They are named in different ways.**
  - There is no metadata, or the metadata is incomplete and inconsistent.
  - There is no searchable text (data, graphics, visualizations, etc.)
  - They exists in different applications, file shares and/or desktops.
  - They have been discarded or lost.
- When they are found why can't information assets be reused?
  - **There are no authoritative sources.**
  - When there are multiple versions, it's difficult to choose which one to use.
  - The source, accuracy and/or authority are unclear.
  - The usage rights may not be clear.

Interoperability ROI (3)

- Information assets are sourced from multiple applications and locations
  - Product lifecycle management (PLM) application
  - Product information management (PIM) application
  - Enterprise content management system application
  - Third party contractors' systems
  - Another department or agency

Interoperability vision

- I want to easily find any information assets in a particular format that can be used for a specific purpose regardless of where they are located.
- I want an authoritative source for key named entity* data such as “customer” or “product”.

---

* Named entities - people, organizations, locations, events, things, etc.

Agenda

- Problems with metadata
- Two types of vocabularies
- Business intelligence tools requirements

Problems with data and metadata

- Inconsistent category assignments
  - CA vs. California
  - RiM vs. Research in Motion
- Changes to classification systems over time
  - ICD-9 vs. ICD-10
  - SIC vs. NAICS
- Use of multiple overlapping or different categorization schemes
  - States vs. SMSA's
  - ICD-9 vs. CDC Diseases and Conditions
  - NASA Taxonomy vs. NASA Thesaurus

Case Study: Inconsistent categories (1)

Problem:

- Inaccurate reporting with incorrect product counts at global health and beauty products company.
- Some SKUs are sold as units, as well as a part of a kit, a set and/or a bill of materials.
- Lacked a consistent, standard language to enable data sharing including:
  - Rules for SKUs.
  - Business processes related to product data.
  - Product data definitions.
  - Single owner for data elements.
  - Roles and responsibilities related to product data.
  - Product data integration points and relationships.

Case Study: Inconsistent categories (2)

Solution:

- Faceted SKU taxonomy instead of a single, monolithic taxonomy tree
  - More flexible design.
  - Describe every item with a combination of facets.
  - Focus on ***universal facets*** applied to all products, or to all products within a large grouping such as a product line.
  - Provides the basis for MDM entity resolution.

Case Study: Inconsistent categories (3)
Universal facets/entities

Major grouping of products based on lines of business. A SKU can be in one or more product lines.

A single product or family of products with a distinct, copywrited, and sometimes trademarked label.

Broad, generic categories used to organize and group products for merchandising and/or business purposes.

A key, active ingredient that is part of the formulation that yields the desired effect in the product.

Indicates whether a product is composed of one or multiple SKUs. If the product is a kit, set or custom assembled BOM, then the component SKUs need to be identified.

Distinguishes products that are specifically intended for one or more age groups.

Distinguishes between products for women and products for men.

Regions and locales within regions that identify target markets or business regions..

Short description of the product.

Indicates type of measure such as number of items, or fluid ounces or milliliters.

Case Study: Multiple categorization schemes (1)

SAMHSA

Problem:

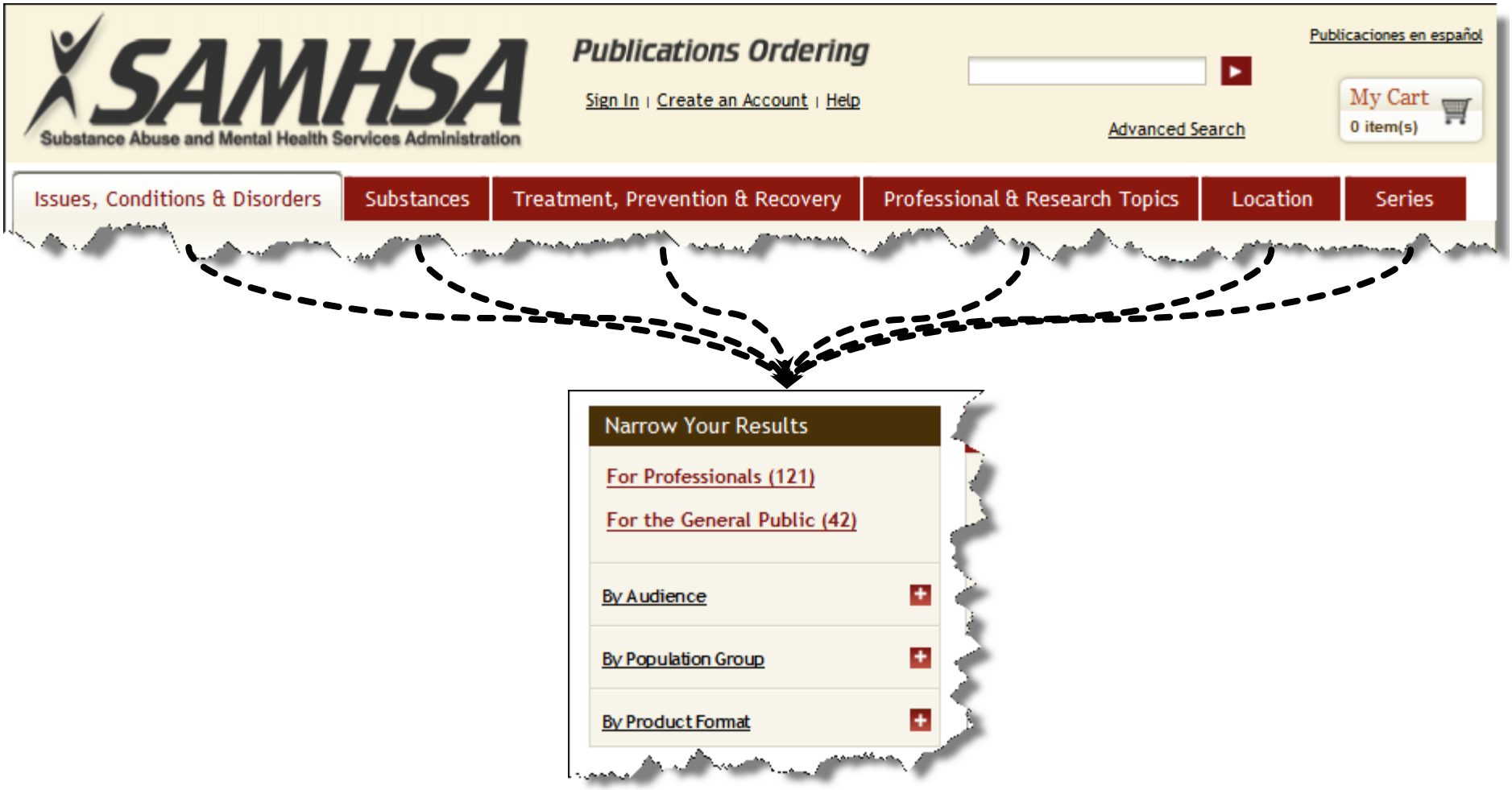
- Need to promote agency ***behavioral health*** program to heterogeneous audiences:
  - Human services professionals
  - Concerned family
  - Policy makers
- Merge heterogeneous information sources:
  - Alcohol and drug information
  - Mental health information
  - Other agency and inter-agency resources
    - Drug Abuse Warning Network (DAWN)
    - Treatment Episode Data Set (TEDS)
    - Uniform Reporting System (URS)

Case Study: Multiple categorization schemes (2)

Solution:

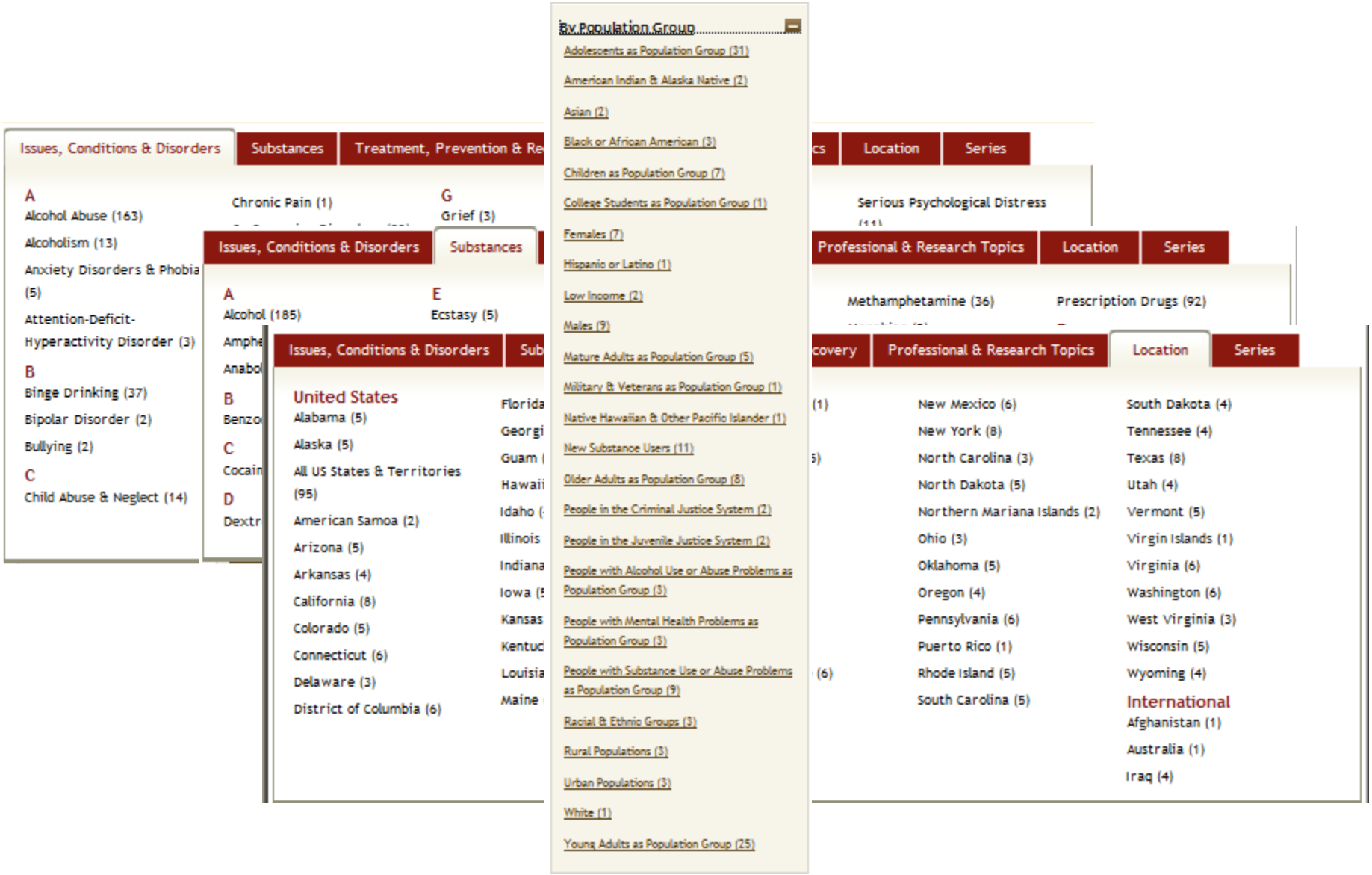
- Faceted taxonomy identifies and resolves key named entities
  - Powers the SAMHSA Store as illustrated in a YouTube video.
  - Provides framework for agency key performance indicators.
  - Increases the availability and visibility of SAMHSA information.
  - Offers tools for analysis, visualization and mash ups with other sources.

Case Study: Multiple categorization schemes (3)

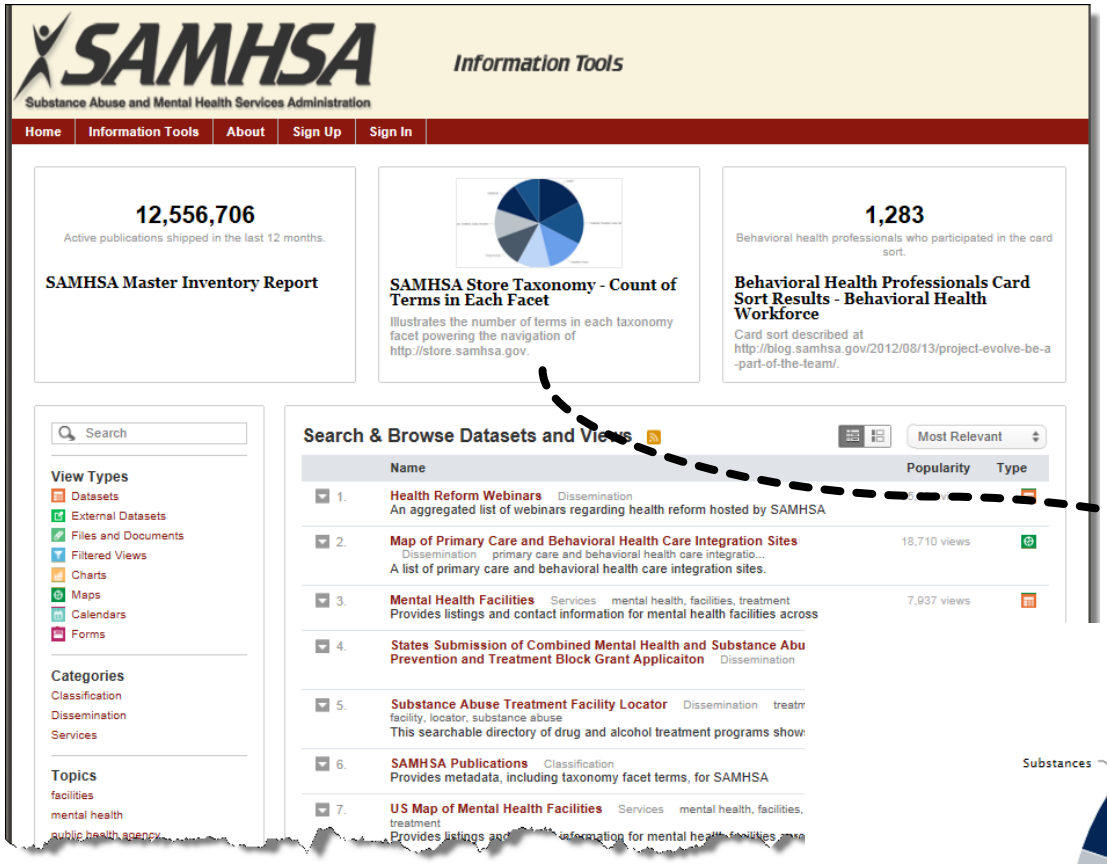


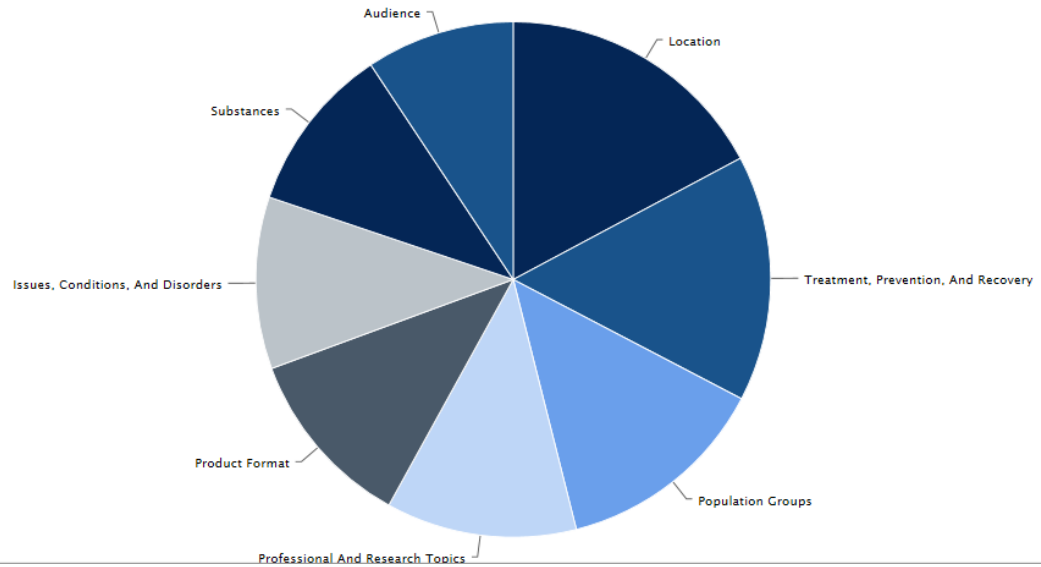
SAMHSA Store Taxonomy facets

Case Study: Multiple categorization schemes (4)



Case Study: Multiple categorization schemes (5)





SAMHSA Info Tools

MDM vs. Taxonomy

- Taxonomy aims to standardize metadata values and the relationships between them
  - Especially term strings.
- Taxonomy can act as a precursor to MDM in that it helps organizations understand what data to master and how to organize this data.
- MDM aims to normalize metadata schemas and valid values across heterogeneous data management systems.

Agenda

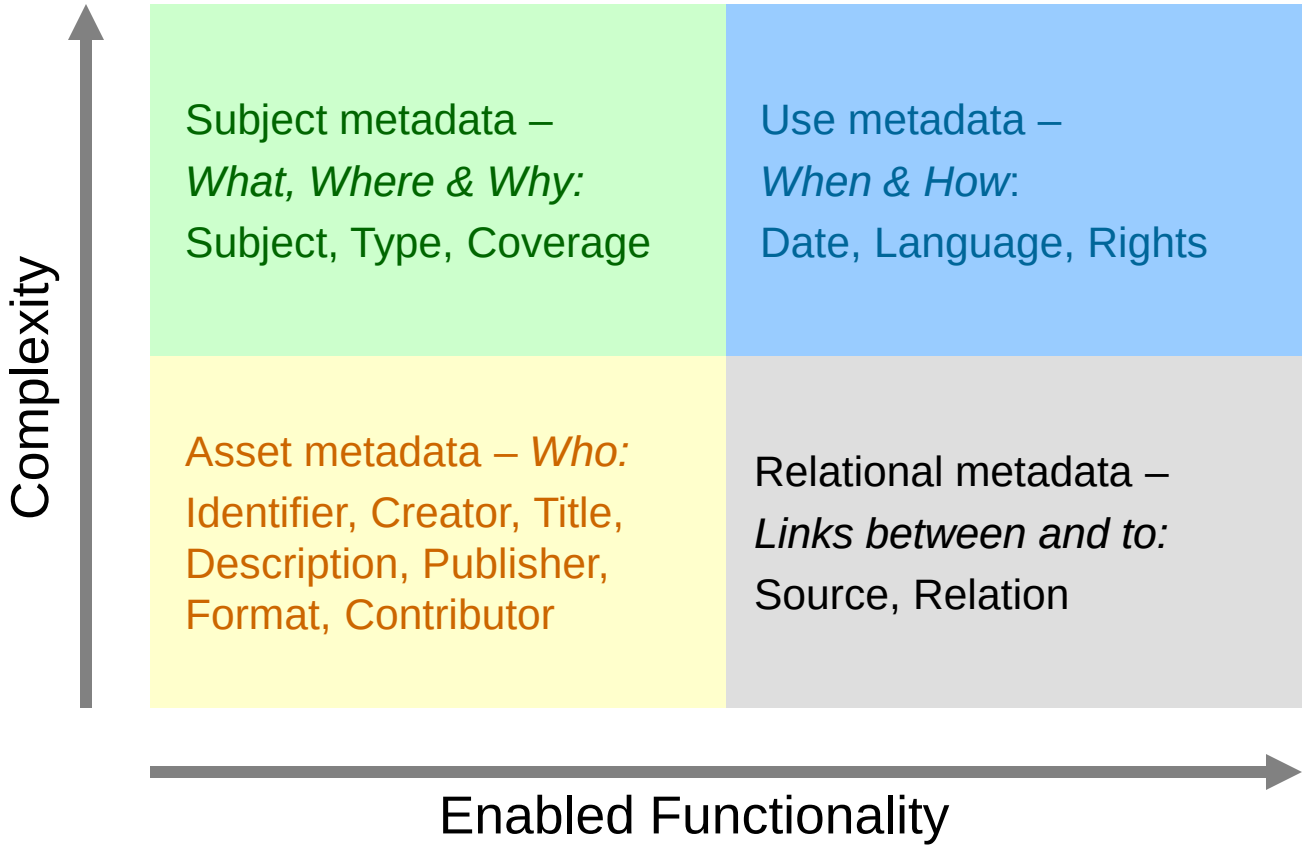
- Problems with metadata
- Two types of vocabularies
- Business intelligence tools requirements

MDM is concerned with two types of vocabularies

- Concept schemes – metadata schemes like Dublin Core, STEP (Standard for the Exchange of Product Model Data) and SEMI E36 (Semiconductor Equipment and Materials International)
- Semantic schemes – value vocabularies like taxonomies, thesauri, ontologies, etc.

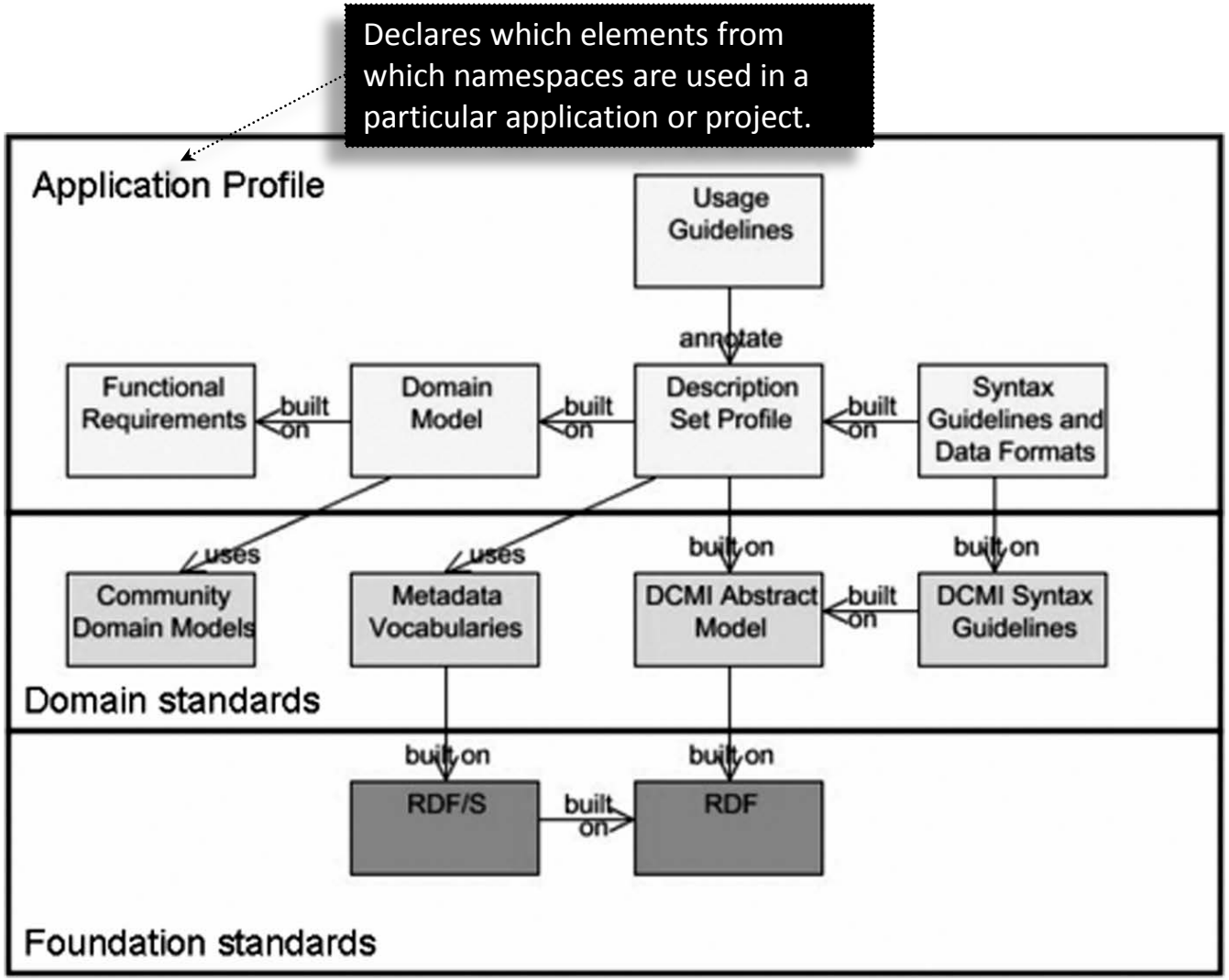
What is Dublin Core?

- Provides the basis for any user, tool, or program to find and use any information asset.



http://dublincore.org/

DCAM (Dublin Core Abstract Model) Singapore Framework



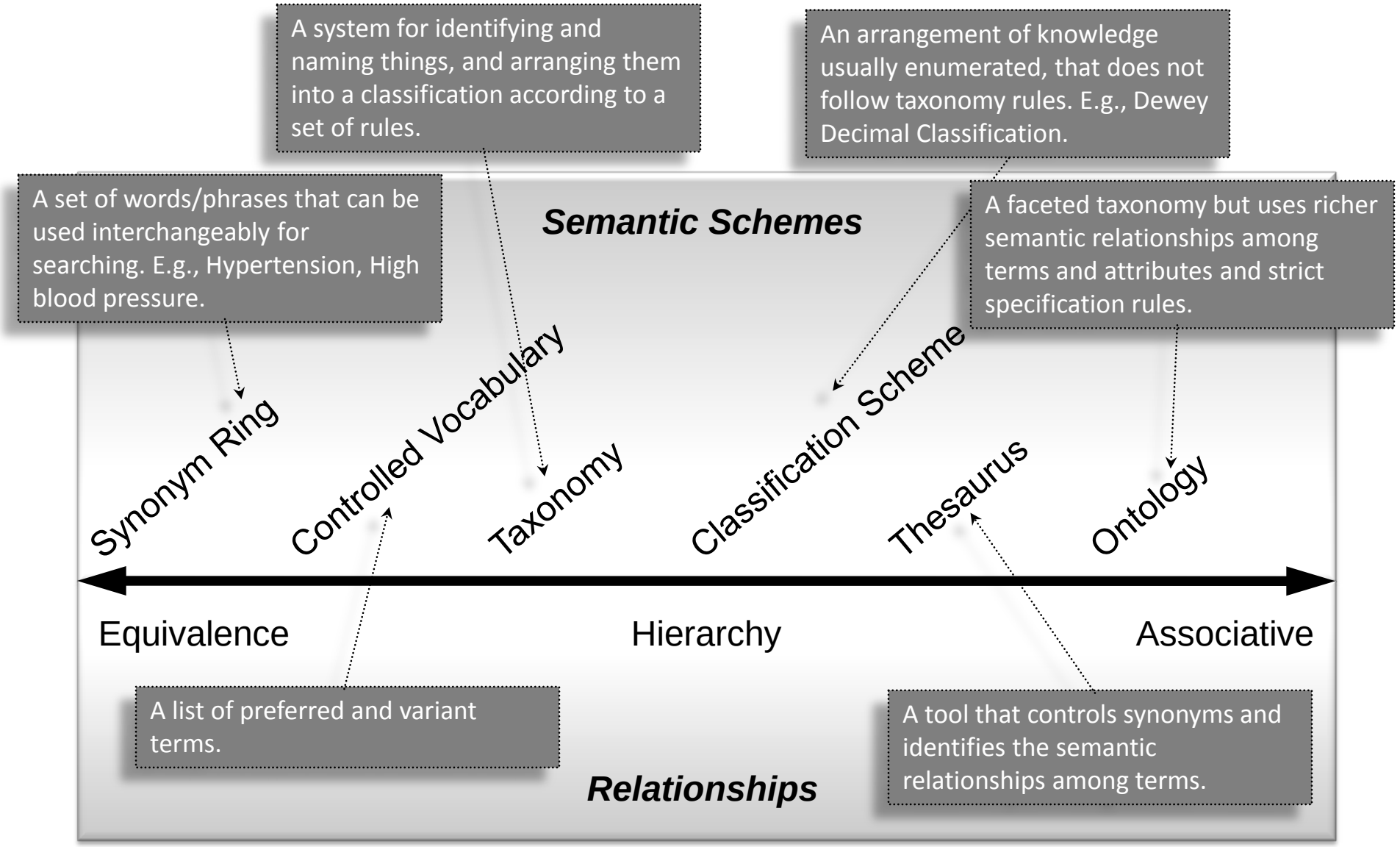
Why Dublin Core?
According to R. Todd Stephens*

- Dublin Core is a de-facto standard across many other systems and standards
  - RSS (1.0), OAI (Open Archives Initiative), SEMI E36, etc.
  - Inside organizations – ECMS, SharePoint, etc.
  - Federal public websites (to comply with OMB Circular A–130, http://www.howto.gov/web-content/manage/categorize/meta-data)
- Mapping to DC elements from most existing schemes is simple.
- Metadata already exists in enterprise applications
  - Windchill, OpenText, MarkLogic, SAP, Documentum, MS Office, SharePoint, Drupal, etc.

---

* Sr. Technical Architect (Collaboration and Online Services) at AT&T

Semantic Schemes: Simple to Complex



After: Amy Warner. *Metadata and Taxonomies for a More Flexible Information Architecture*

Q: How do you share a vocabulary across (and outside of) the enterprise?
A: With standards

- **ANSI/NISO Z39.19-2005** Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies
- **ISO 2788:1986** Guidelines for the Establishment and Development of Monolingual Thesauri
- **ISO 5964:1985** Guidelines for the Establishment and Development of Multilingual Thesauri
- **ISO 25964** (combines 2788 and 5964) Thesauri and Interoperability with other Vocabularies
- **Zthes** specifications for thesaurus representation, access and navigation
- **W3C** SKOS Simple Knowledge Organization System

Why SKOS?
According to Alistair Miles* (SKOS co-author)

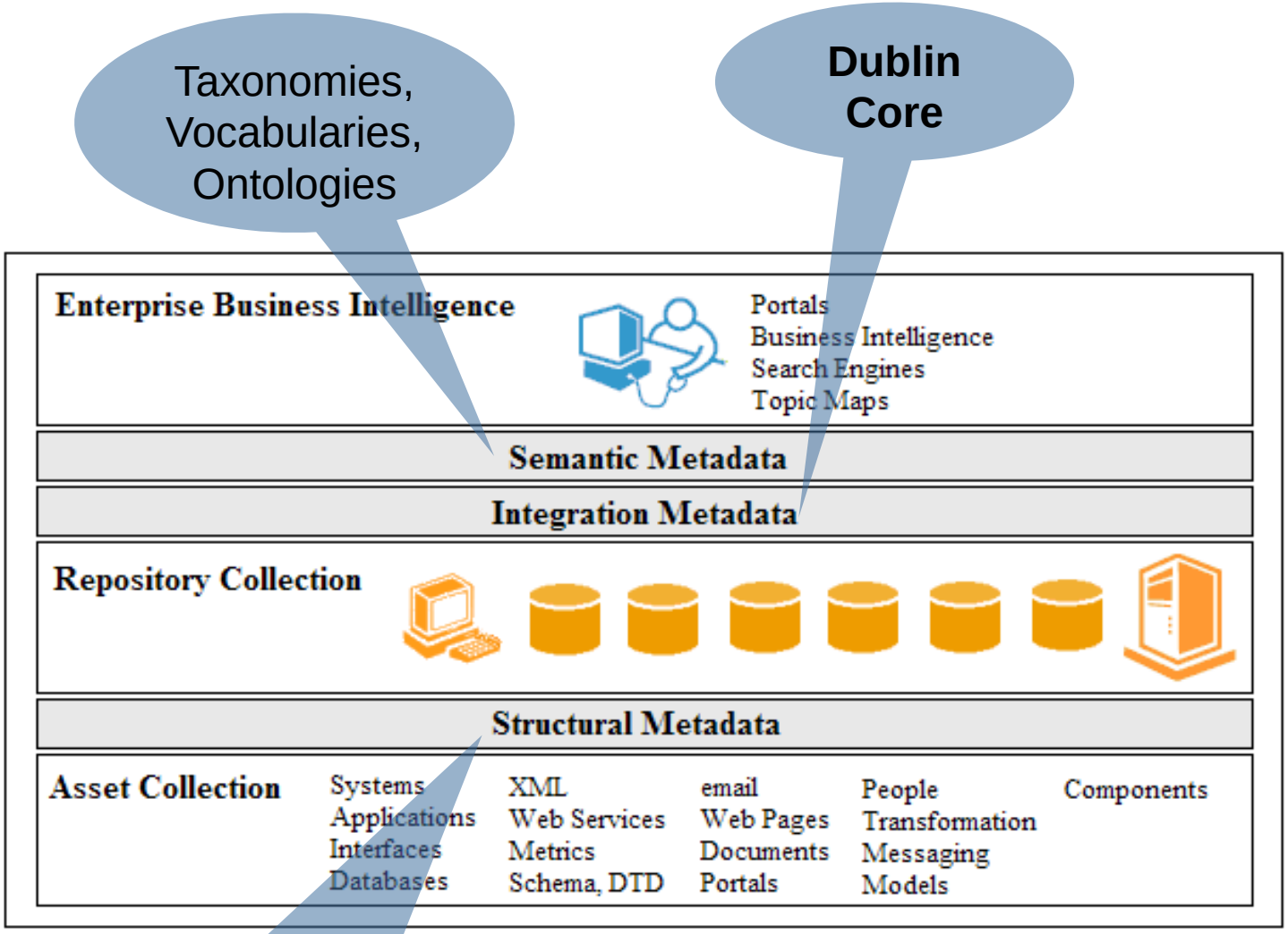
- **Ease of combination** with other standards
  - Vocabularies are used in great variety of contexts.
    - E.g., databases, faceted navigation, website browsing, linked open data, spellcheckers, etc.
  - Vocabularies are re-used in combination with other vocabularies.
    - E.g., ISO3166 country codes + USAID regions; USPS zip codes + US Congressional districts; USPS states + EPA regions, etc.
- **Flexibility and extensibility** to cope with variations in structure and style
  - Variations between types of vocabularies
    - E.g., list vs. classification scheme
  - Variations within types of vocabularies
    - E.g., Z39.19-2005 monolingual controlled vocabularies and the NASA Taxonomy

* Senior Computing Officer at Oxford University

Why SKOS? (2)

- **Publish managed vocabularies** so they can readily be consumed by applications
  - Identify the concepts
    - What are the named entities?
  - Describe the relationships
    - Labels, definitions and other properties
  - Publish the data
    - Convert data structure to standard format
    - Put files on an http server (or load statements into an RDF server)
- **Ease of integration** with external applications
  - Use web services to use or link to a published concept, or to one or more entire vocabularies.
    - E.g., Google maps API, NY Times article search API, Linked open data
- **A W3C standard** like HTML, CSS, XML… and RDF, RDFS, and OWL

MDM model that integrates taxonomy and metadata



Source: Todd Stephens, BellSouth

Agenda

- Problems with metadata
- Two types of vocabularies
- Business intelligence tools requirements

Business intelligence tools requirements

- Requirements for integrating taxonomy with business intelligence metadata tools.

Tools

- Taxonomy editing
  - Data Harmony, Mondeca, MultiTes, PoolParty, protégé, SmartLogic, Synaptica, Top Braid Composer
- Metadata tagging (automated categorization)
  - CIS, ConceptSearching, Data Harmony, MetaTagger, nStein, Smartlogic, temis
- Enterprise content management
  - Alfresco, EMC Documentum, Drupal, IBM FileNet, Joomla!, OpenText, Oracle Content Management, SharePoint
- Business intelligence tools
  - Actuate, Business Objects (SAP), Cognos (IBM), Hyperion (Oracle), Informatica, MicroStrategy, SAS

Taxonomy tool functions (1)

Functional area	Functions
Taxonomy Development	Create a taxonomy User roles and permissions
Taxonomy Maintenance	Add, edit, move, delete items Assign or modify privileges to one or a group of items Activity logging
Taxonomy Governance	Approval workflow for additions and changes
Metadata Controlled Vocabulary	Assign attributes to a category Associate controlled vocabulary with metadata field Thesaurus capabilities
User Interface	Search and browse Drag and drop Multiple windows
Reporting	Alphabetical, hierarchical and other views Visualizations Importing and exporting taxonomies
Application Integration	APIs (WSDL, Scripts, Java, etc.) Application integration (CMS, DMS, search engine, etc.)

Taxonomy tool functions (2)

Functional area	Functions
Database Definition	How is the database created? Where is it stored? Is it Z39.19 and ISO 2788 compliant? Database license requirement?
Importing/Exporting Data	How are data imported? What file formats are supported? Can data files be in batches?
Add, Edit, Delete Categories	How easily are categories added, edited, or deleted? Can categories be added, edited, or deleted in batches?
Relationship Types	How are relationship types defined? What types are supported? How is polyhierarchy handled?
Add, Edit, Delete Relationships	How easily are relationships added, edited, or deleted? Can relationships be added, edited, or deleted in batches? Does a change propagate to all instances?
Reporting	How does the TMS report: new, edited, deleted taxonomies and categories; new, edited, deleted relationship types and relationships; mapped taxonomies and categories? How are the reports presented? What audit logs are available? Can changes be traced to users who suggested them? Is an "approval" step for changes available for administrators?
User Access	Can the TMS integrate user accounts with existing authentication systems, e.g. Active Directory, etc.? Is there support for role-based access or defined group membership with configurable access? Is there a workflow to approve changes? What functionality is available or restricted based on a user's security privileges?

Taxonomy tools and business intelligence

- No taxonomy tool vendors have connectors, custom APIs or other direct integrations with leading business intelligence tools.
- SAS acquired Teragram in 2010.
  - Teragram is primarily an OEM business, not integrated with SAS business intelligence products.
- Business Objects acquired Inxight in 2007, which was acquired by SAP in 2008.
  - Inxight is not evident in SAP business intelligence products.

Questions

Joseph A Busch

jbusch@taxonomystrategies.com

mobile 415-377-7912