

Evaluating Taxonomies

Agenda

- Evaluation overview
- Editorial evaluation
- Collection analysis
- Market analysis
- Summary and questions

Framework for evaluating taxonomies: How will the taxonomies be used?

➤ Use cases

- Data management, Data warehouse, MDM, Big data
- Business intelligence, Text analytics
- eCommerce
- Search and browse, Web publishing

➤ Case studies

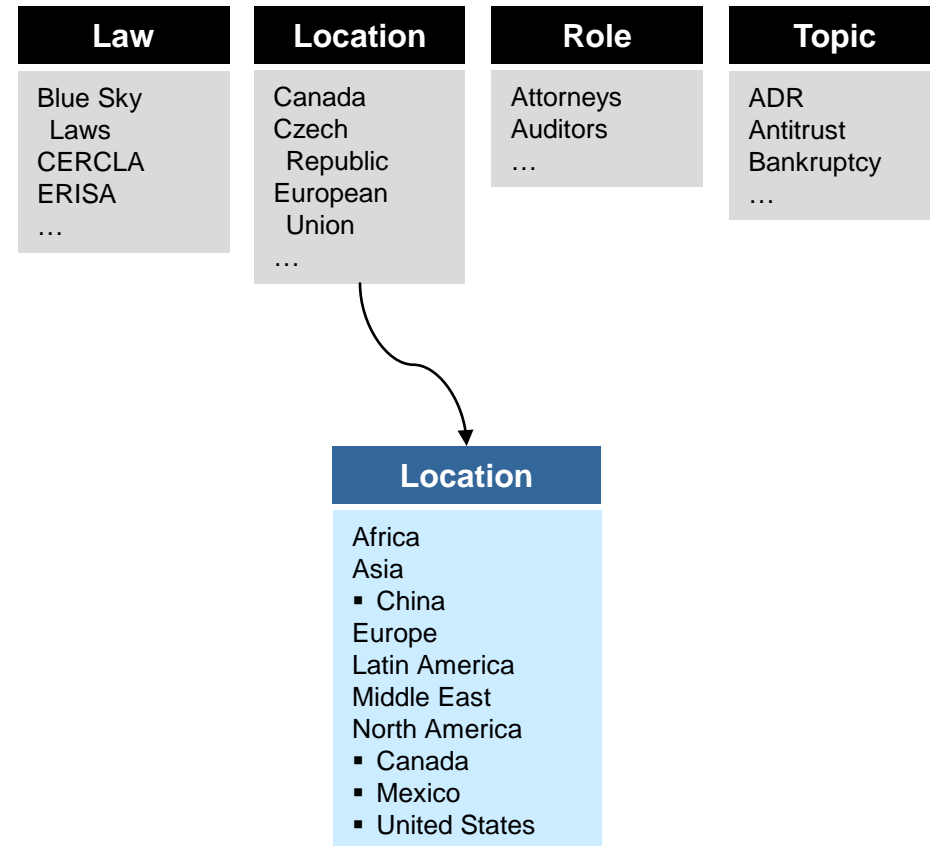
- Healthcare.gov – findable web content, transaction help, customer service
- Energy companies – technical training, operational documentation, EHS
- Retail and eCommerce – POS, labels, dynamic web content, ecommerce
- Financial services organizations – AML, SAR, trading, analysis

Editorial evaluation

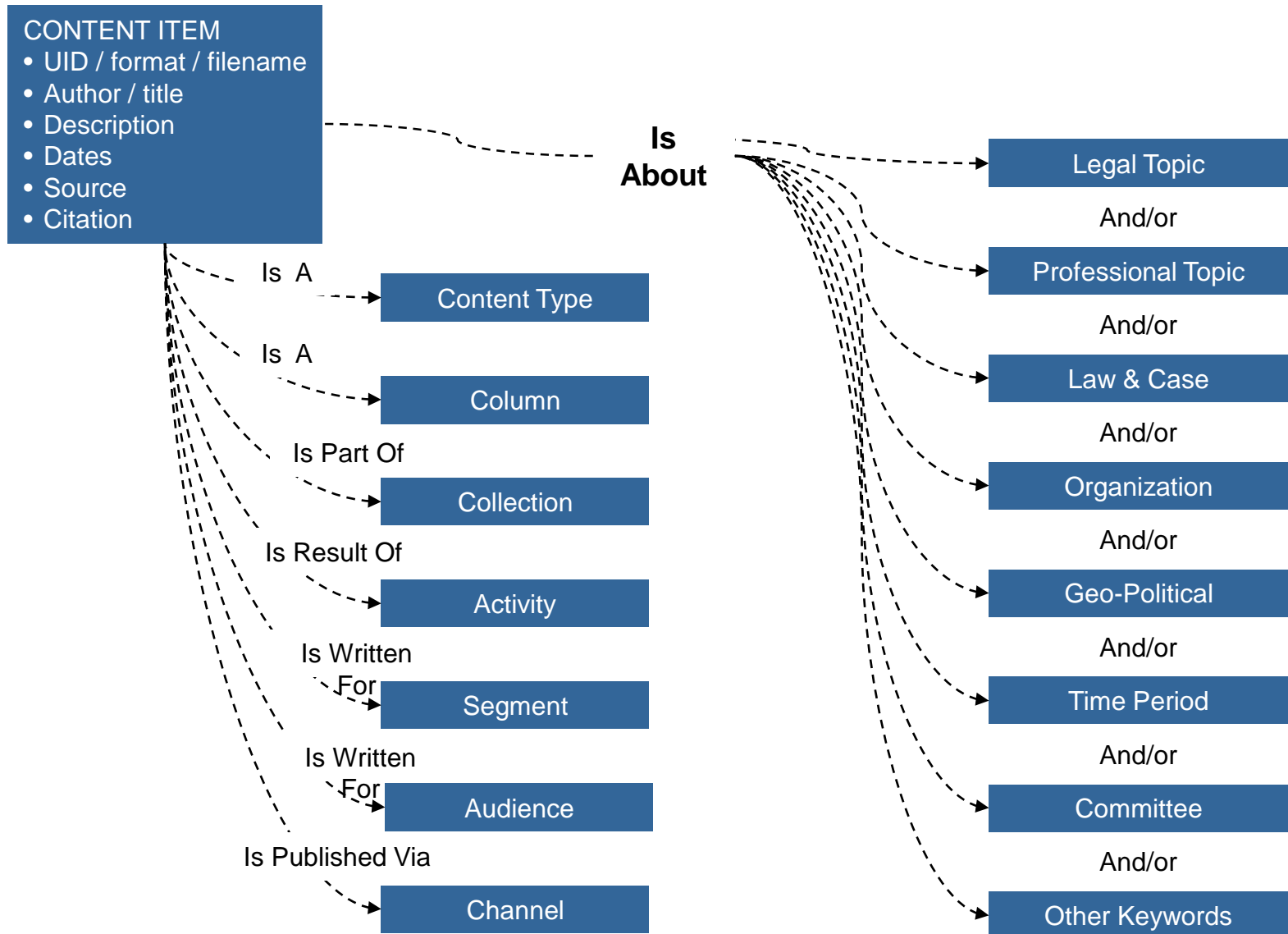
- Depth and breadth
- Comprehensiveness
- Currency
- Relationships
- Polyhierarchy (is it applied appropriately)
- Naming conventions.

Depth and breadth

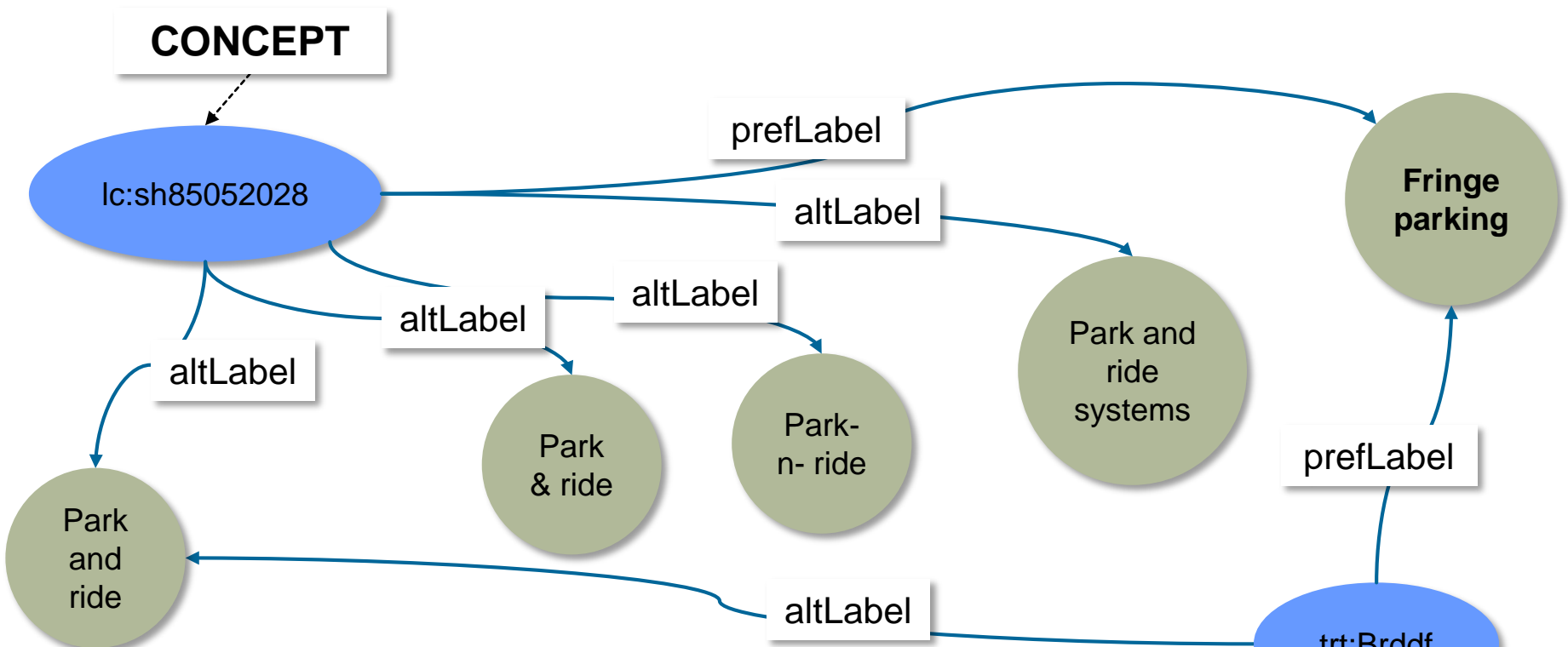
Category List	Facet
Alternative Dispute Resolution (ADR)	Topic
Antitrust	Topic
Attorneys	Role
Auditors	Role
Bankruptcy	Topic
Blue Sky Laws	Law
Canada	Location
Comprehensive Environmental Response, Compensation and Liability Act of 1980 (CERCLA)	Law
Czech Republic	Location
Employee Retirement Income Security Act of 1974 (ERISA)	Law
European Union	Location
...	



Taxonomy relationships



Taxonomy relationships



Subject	Predicate	Object
lc:sh85052028	skos:prefLabel	Fringe parking
lc:sh85052028	skos:altLabel	Park and ride systems
lc:sh85052028	skos:altLabel	Park and ride
lc:sh85052028	skos:altLabel	Park & ride
lc:sh85052028	skos:altLabel	Park-n-ride
trt:Brddf	skos:prefLabel	Fringe parking
trt:Brddf	skos:altLabel	Park and ride

Naming conventions

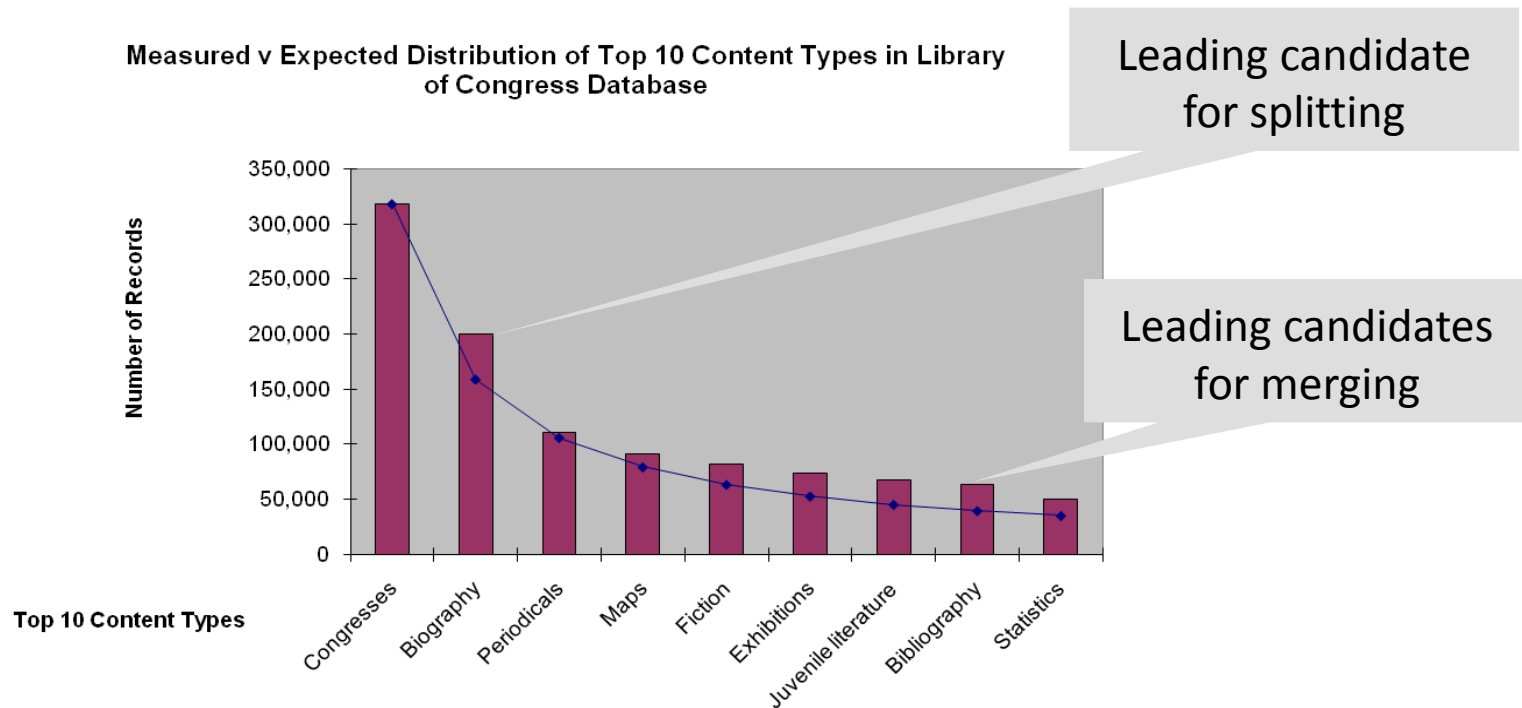
1. Label length
2. Nomenclature
3. Capitalization
4. Ampersands
5. Abbreviations & Acronyms
6. Languages
7. Special characters
8. Serial commas
9. Spaces
10. Synonyms
11. Term order
12. Term ordering
13. Compound term labels

Collection analysis

- Category usage analytics (is distribution of categories appropriate)
- Completeness and consistency
- Query log/content usage analysis

Category usage analytics: How evenly does it divide the content?

- Documents do not distribute uniformly across categories
- Zipf (long tail) distribution is expected behavior
- 80/20 Pareto rule in action



Category usage analysis: How does taxonomy “shape” match that of content?

Term Group	% Terms	% Docs
Administrators	7.8	15.8
Community Groups	2.8	1.8
Counselors	3.4	1.4
Federal Funds Recipients and Applicants	9.5	34.4
Librarians	2.8	1.1
News Media	0.6	3.1
Other	7.3	2.0
Parents and Families	2.8	6.0
Policymakers	4.5	11.5
Researchers	2.2	3.6
School Support Staff	2.2	0.2
Student Financial Aid Providers	1.7	0.7
Students	27.4	7.0
Teachers	25.1	11.4

- Background:
 - Hierarchical taxonomies allow comparison of “fit” between content and taxonomy areas.
- Methodology:
 - 25,380 resources tagged with taxonomy of 179 terms. (Avg. of 2 terms per resource)
 - Counts of terms and documents summed within taxonomy hierarchy.
- Results:
 - Roughly Zipf distributed (top 20 terms: 79%; top 30 terms: 87%)
 - Mismatches between term% and document% are flagged in red.

Source: Courtesy Keith Stubbs, US. Dept. of Ed.



Completeness and consistency: Indexer consistency

- Studies have consistently shown that levels of consistency vary, and that high levels of consistency are rare for:
 - Indexing
 - Choosing keywords
 - Prioritizing index terms
 - Choosing search terms
 - Assessing relevance
 - Choosing hypertext links

- Semantic tools and automated processes can help guide users to be more consistent.

30%

80%

Query log analysis: Description of analysis process

- Identify top query strings over annual period, average number of words per query and distribution of queries – Are there a few that make up the majority of the total number of queries?
- Review each query string to determine what the user is trying to find. Assign a concept/entity.
- Each concept/entity is a type of thing. Review each and identify the type or types of things.
- Identify the top concepts/entities.
- Perform analysis on internal and external queries as appropriate.

Query log analysis: Internal Queries

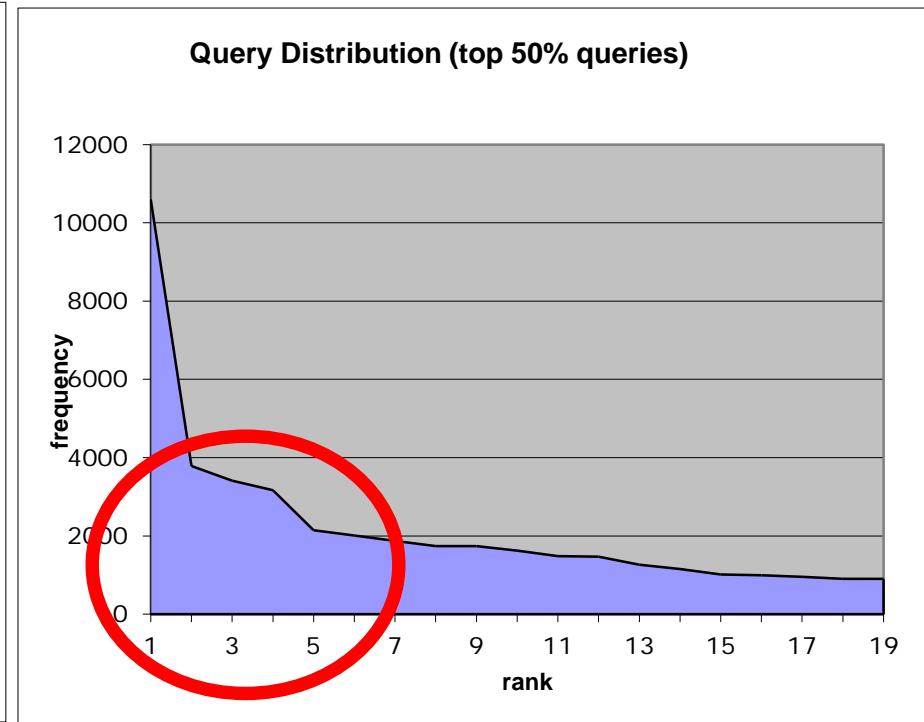
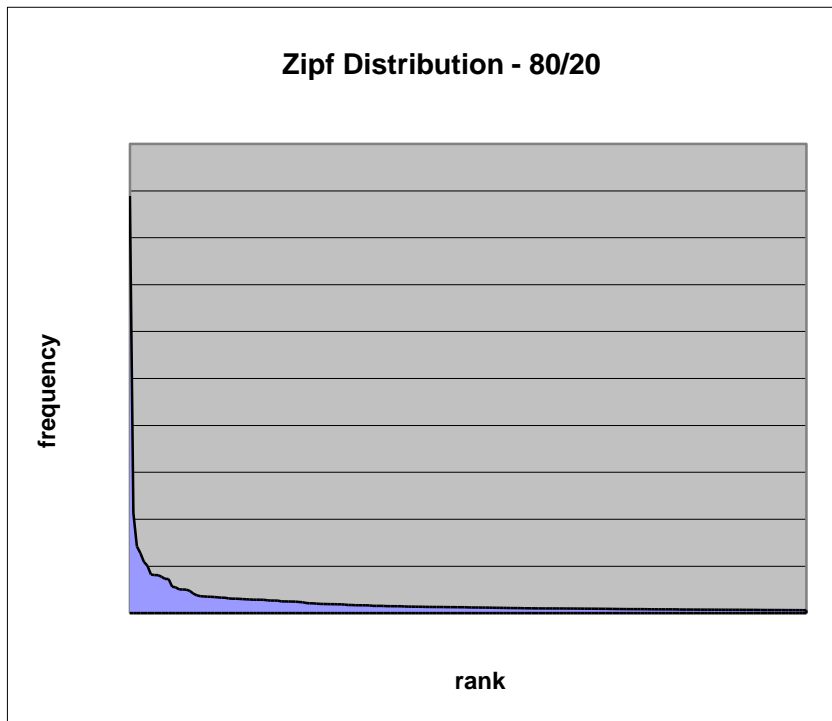
Words typed into search box on healthcare.gov Aug 2011-July 2012

84,277	Total Queries in Sample
214	Total Unique Queries in Sample
393.82	Average # Times Unique Queries were Performed
153.00	Median # Times Unique Queries were Performed
1.86	Average # Terms/Unique Query
13.36	Average # Characters/Unique Query

Query log analysis: Query distribution

Comparing to Zipf – 80/20

- 80/42
- 80% of the query volume is made up of 42% of the unique queries
- 80% of the 84,277 queries is made up of the top 64 unique queries



Query log analysis: Top queries grouped into buckets

Buckets	% of Total Queries	Count
Medical Loss Ratio	19.07993877	16080
Conditions/Treatment/Equipment/Devices	11.39456791	9603
Federal & State Programs	10.28513117	8668
Pre-existing Conditions	7.264140869	6122
Healthcare Services	4.037875102	3403
Prevention	3.792256488	3196
Coverage Mandated/Coverage Exemption	3.146766022	2652
Grandfathered Health Plans	2.593827497	2186
Spanish/English "to seek"	2.513141189	2118
Essential Health Benefits	2.142933422	1806
Payments/Deductibles	1.89138199	1594
Health Insurance Exchange	1.724076557	1453
Patient's Bill of Rights	1.396585071	1177
Accountable Care Organization	1.160458963	978
Age/Gender/Class	0.950437249	801
Timeline	0.939758178	792

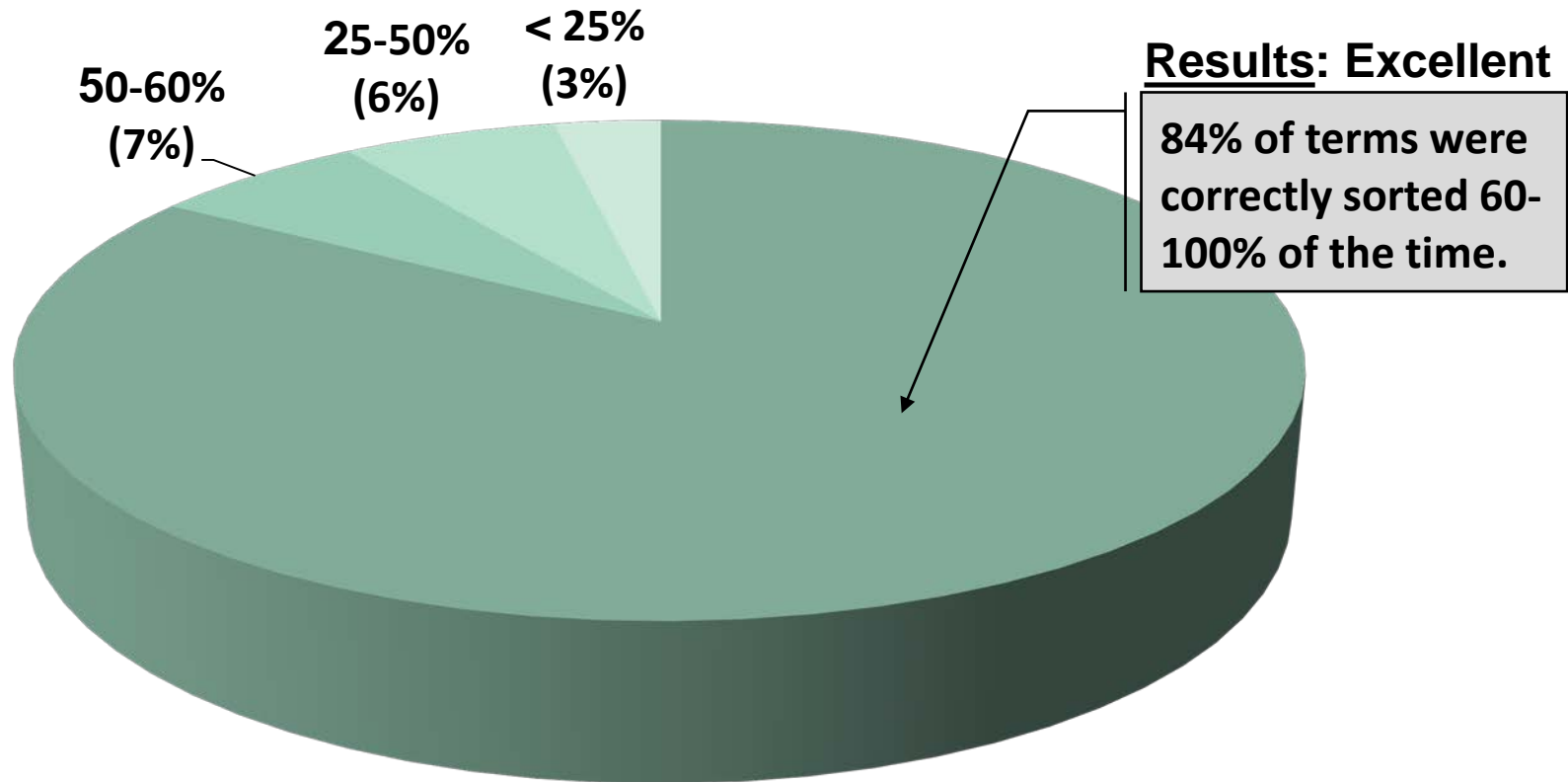
Market analysis: The best thing about standards is there are so many to choose from

- Industry standards/leaders
- User surveys
- Card sorting
- Task based usability.

9 Common taxonomy facets

Facet	Definition	Example Source
Content Type	The various genres of content being created, managed and/or used.	AGLS Document Type, AAT Information Forms , Records management policy, etc.
Audience	Subset of constituents to whom a content item is directed or intended to be used.	GEM, ERIC Thesaurus, IEEE LOM, etc.
People	Names of important people such as authors, politicians, leaders, actors, etc.	LC NAF, NYTimes Topics-People
Organization	Names of organizations, their aliases and the relationships between them.	FIPS 95-2, D&B, Ticker Symbols, LC NAF, NYTimes Topics-Organizations, etc.
Industry	Broad market categories such as industry sector codes.	FIPS 66, SIC, NAICS, etc.
Location	Names of places of operations, activities, constituencies, etc.	ISO 3166, FIPS 5-2, FIPS 55-3, USPS, NYTimes Topics-Places etc.
Function	Activities and processes performed to accomplish goals.	FEA Business Reference Model, AAT Functions, etc.
Product	Names of products and services that are produced by an organization or people.	Household Products Database, etc.
Topic	Topical subjects and themes that are not included in other facets.	LCSH, NYTimes Topics-Subjects, etc.

Completeness and consistency: Blind sorting of popular search terms

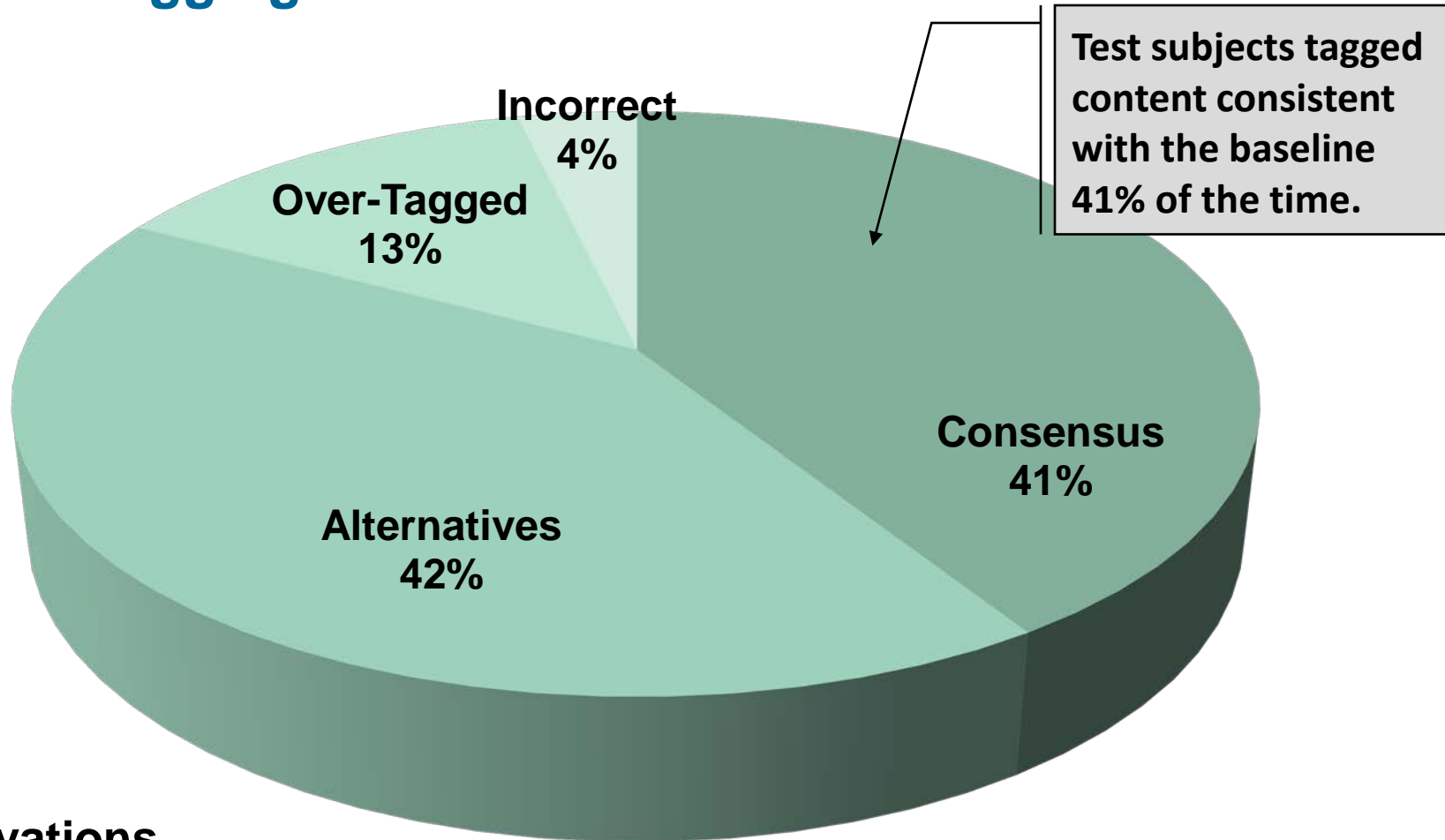


Difficulties

- For Methadone, confusion when, in this case, a substance is a treatment.
- For general terms such as Smoking, Substance Abuse and Suicide, confusion about whether these are Conditions or Research topics.

Completeness and consistency: Content tagging consensus

Results: Good



Observations

- Many other tags were reasonable alternatives.
- **Correct + Alternative tags accounted for 83% of tags.**
- Over tagging is a minor problem.



What are your primary goals when visiting Nike.com?

- Shop
- Research
- Sports information
- Training advice
- Other _____

Observation on top level of navigation:

- What do you expect to find under Product?
- What do you expect to find under Sport?
- What do you expect to find under Train?
- What do you expect to find under Athlete?
- What do you expect to find under Innovate?

Scenario 1: what would you click on to find out more about men's clothing?

- On a scale of 1-5 (1 = very difficult, 5 = very easy) did you find it easy to generally locate the object through the diagram navigation path?

1 2 3 4 5

Scenario 2: what would you click on to find out how to improve your performance?

- On a scale of 1-5 (1 = very difficult, 5 = very easy) did you find it easy to generally locate the object through the diagram navigation path?

1 2 3 4 5



Hybrid method: “Fashion-forward” product recommendations

Index Attribute	Value Type	Source
Boldness	1-5	Merchandising
Newness	Logarithmic	Derived from release date
Brand Fashion	1-5	Derived from brand
Lifestyle	1-5	Marketing
Product Review	1-5	Fashion Forward Customers

- Indexes are derived from multiple attributes and sources
 - Initial weighting can be heuristic and adapted based on user behavior
- Index attributes enable analytics and personalization to bootstrap from and leverage Macy’s merchandising expertise
- Likert scales (1-5) are sufficient for manually set index attributes
- For automated scoring, use more granular, relative scales.

Joseph A Busch, Principal
jbusch@taxonomystrategies.com
twitter.com/joebusch
415-377-7912

Vivian B Bliss, Associate
vbliss@taxonomystrategies.com
425-417-7628

QUESTIONS?

Evaluating Taxonomies

- ❖ Taxonomies are developed in communities and evolve over time. From the outset there is a need to evaluate existing schemes for organizing content and questions about whether to build or buy them. Once built out and implemented, taxonomies require ongoing revisions and periodic evaluation to keep them current and structurally consistent. Taxonomy evaluation includes the following dimensions which will be discussed in this webinar.
 - Editorial evaluation – including depth and breadth, comprehensiveness, currency, relationships, polyhierarchy (is it applied appropriately), and naming conventions.
 - Collection analysis - category usage analytics (is distribution of categories appropriate), completeness and consistency, and query log/content usage analysis.
 - Market analysis – including industry standards/leaders, user surveys, card sorting, and task based usability.
- ❖ Examples will be provided from public, non-profit and commercial client projects.