Web-based content organization and the transformation of traditional classification systems

Joseph Busch

Locating objects

❖ Traditional classification schemes physically locate objects in one dimension.

Shelf Location

https://upload.wikimedia.org/wikipedia/commons/f/fd/Shelves_of_Language_Books_in_Library.JPG

Table of Contents

Stores with a large assortment use well known departments and sub-categories

Supermarket Categories

❖ Bakery

❖ Beverages & Snacks

❖ Dairy

❖ Deli & Prepared Foods

❖ Grocery

❖ Household

❖ Meat & Seafood

❖ Produce

Department Store Categories

❖ Men's Clothing

❖ Women's Clothing

❖ Baby & Children's Clothing

❖ Home Furnishings

❖ Electronics

❖ Toys & Sports

❖ Food

Hierarchical classification problem

Fresh Vegetables

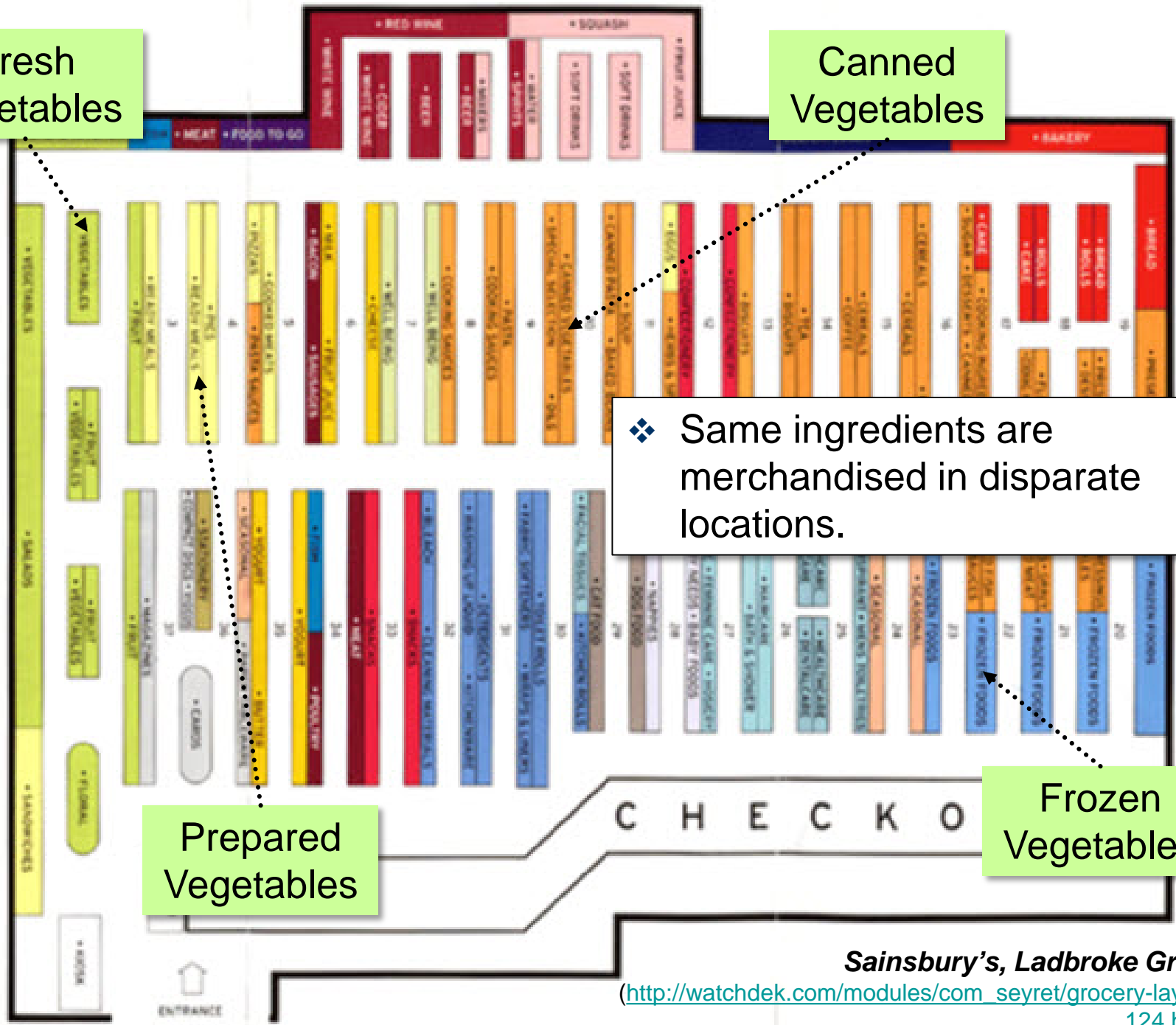Canned Vegetables

Prepared Vegetables

Frozen Vegetables

❖ Same ingredients are merchandised in disparate locations.

Sainsbury's, Ladbroke Grove
(http://watchdek.com/modules/com_seyret/grocery-layout-124.html)

Physics & Astronomy Classification Scheme (PACS 2010)

❖ Alphanumeric codes
❖ Strict (single-parent) hierarchy
❖ Incomplete and complex labels

❖ Some concepts are repeated many times.

PACS 2010 semiconductor-related categories are in disparate locations

00—General

01. Communication, education, history, and philosophy

02. Mathematical methods in physics

➤ 03. Quantum mechanics, field theories, and special relativity

04. General relativity and gravit...

05. Statistical physics, thermodyna...

06. Metrology, measurements, an...

07. Instruments, apparatus, and c...

➤ **60**—Condensed Matter: Structural, Mechanical and Thermal Properties

61. Structure of solids and liquids; crystallography

62. Mechanical and acoustical properties of condensed matter

63. Lattice dynamics

64. Equations of state, phase equilibria, and phase transitions

65. Thermal properties of condensed matter

66. Nonelectronic transport properties of condensed matter

67. Quantum fluids and solids

68. Surfaces and interfaces; thin films and nanosystems (structure and nonelectronic properties)

➤ **70**—Condensed Matter: Electronic Structure, Electrical, Magnetic, and Optical Properties

71. Electronic structure of bulk materials

72. Electronic transport in condensed matter

73. Electronic structure and electrical properties of surfaces, interfaces, thin films, and low-dimension...

74. Superconductivity

75. Magnetic properties and materials

76. Magnetic resonances and relaxations in condensed matter, Mössbauer effect

77. Dielectrics, piezoelectrics, and ferroelectrics and their properties

78. Optical properties, condensed-matter spectroscopy and other interactions of radiation and partic...

79. Electron and ion emission by liquids and solids; impact phenomena

The *real* world of concepts is multidimensional

Type

Extent

Properties

Location

Chronology

Topic
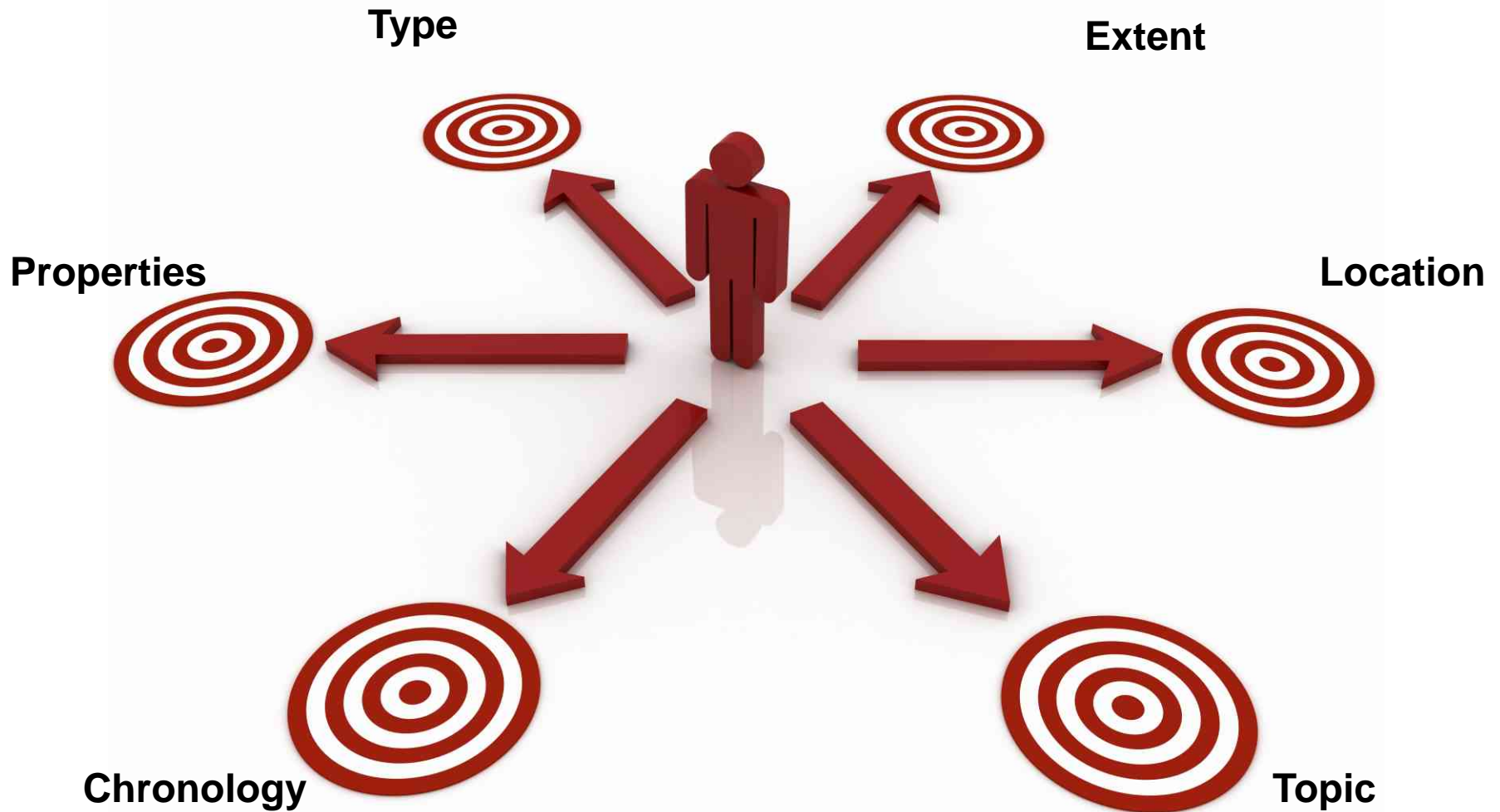
Image source: http://stratconcept.com/yahoo_site_admin/assets/images/Comp_ist2_6702437-many-targets-concept.12743816.jpg

Shopping for shoes on Zappos.com

- ❖ Size
- ❖ Width
- ❖ Styles
- ❖ Occasion
- ❖ Color
- ❖ Brand
- ❖ Price
- ❖ Heel Height
- ❖ Materials
- ❖ Pattern
- ❖ Accents

Filtering Search Results on rwjf.org

- ❖ By Topic
- ❖ By Content Type
- ❖ By Age
- ❖ By Gender
- ❖ By Race/Ethnicity
- ❖ By Location
- ❖ By States and Territories

Filtering search results (a.k.a. facetted navigation)

❖ Invites end users to *refine* their search results without typing in a new query, e.g.,

- Expose contextually relevant metadata.
- Show how many matching hits by attribute.
- Allow removal and addition of filters.

❖ Allows user to *explore* collection of search results

- Drill down or up, by applying or removing filters.

Online shopping vs. content collections

❖ Using facetted navigation in online shopping has become a commonly understood metaphor.

❖ But applying multi-dimensional classification and facetted navigation to content collections is not as common.

What are use cases?

❖ Understanding how an organization or individual might actually use or wish to use organized information.

❖ A "use case" explores various scenarios with multiple stakeholders.

   ▪ Formal and informal interviews.

   ▪ Analysis of quantitative data on content use and search behavior.

   ▪ Learning about organizational goals and expectations.

   ▪ Identifying potential activities or likely uses for a set of organized information.

❖ Use cases facilitate the development of a specialized taxonomy

   ▪ Describes activities and uses, or contexts important for particular applications in particular settings.

The most common classification use case is ordering a set of related content items in a search results.

APS Taxonomy use cases are not focused on search

- ❖ Organizing & facilitating editorial & publishing process.
  - Selection of taxonomy terms (indexing) for articles.
  - Authors assigning topics to their submissions.
  - Defining areas of responsibility and interest for editors.
  - Assigning articles to APS editors.
  - Referees describing their areas of expertise.
  - Selecting referees to review articles.
  - Assigning articles to journal sections.
  - Generating statistical reports and lists of articles by various subject criteria.

Grouping papers by topic and so that attendees can get from one session is another use case
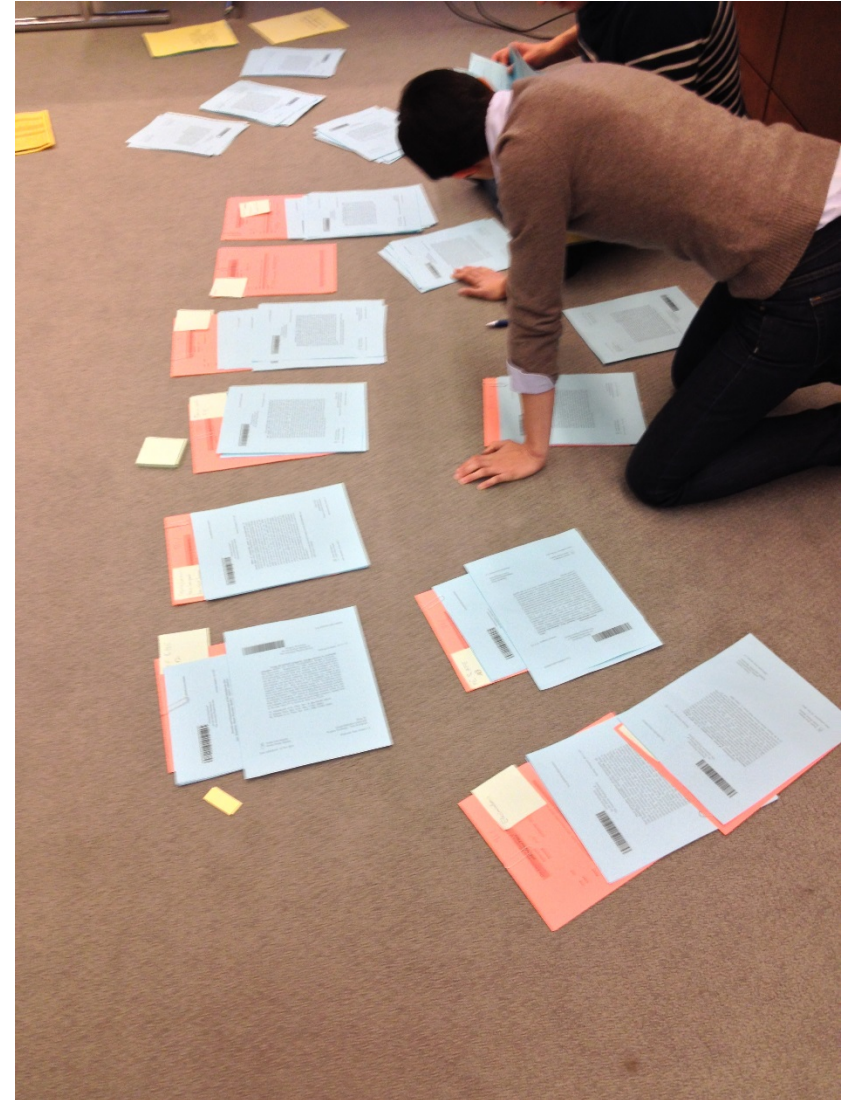
APS sorting meeting (Dec 5, 2013)

The "big use case": Overall strategy for locating products so consumers can buy them

❖ How to facilitate the identification and linking of a large and changing collection of content items with a large and changing assortment of related products.

Break down of specific tactics by IA method

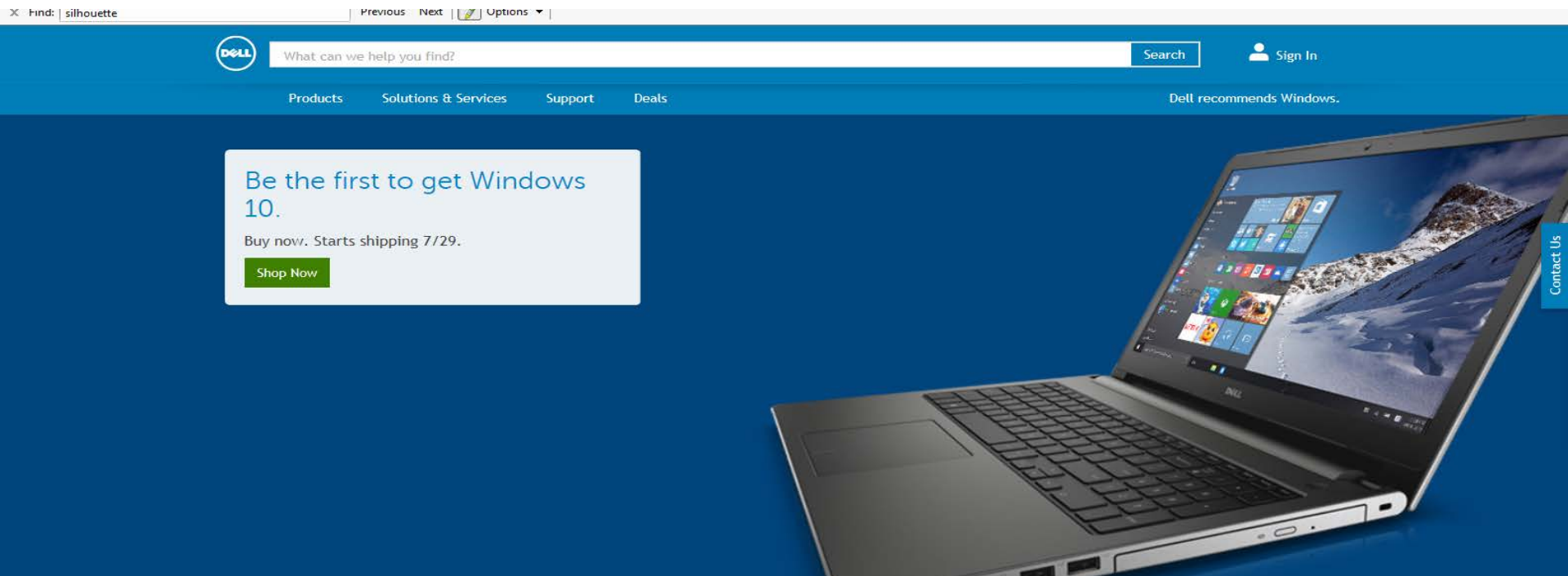| Use Case	Contextual Navigation	Site Architecture	Synonyms	Import Files
Improve Google search.			X	
Consistent experience across sites.		X		
Consistent terminology.	X		X	
Use technology to pivot between service & product.	X			
Associate learn content w/ specific products.	X			
Move from learn to product content.	X			
Provide context w/in industry solutions.	X			
Consistent solutions & best practices experiences.			X	X
Consolidate community content.	X	X		X
Unify support & community content.	X	X		
Integrate product support w/ product details.	X			
Surface software & peripherals info.				X
Surface parts & accessories w/ products.	X			X
De-segmentation.		X		
Integrate external content.	X			X

Importance of facets and relationships

❖ Facetted classifications deconstruct complex concepts into a grammar

   ▪ Statements of named entities modified by types and topics, e.g.

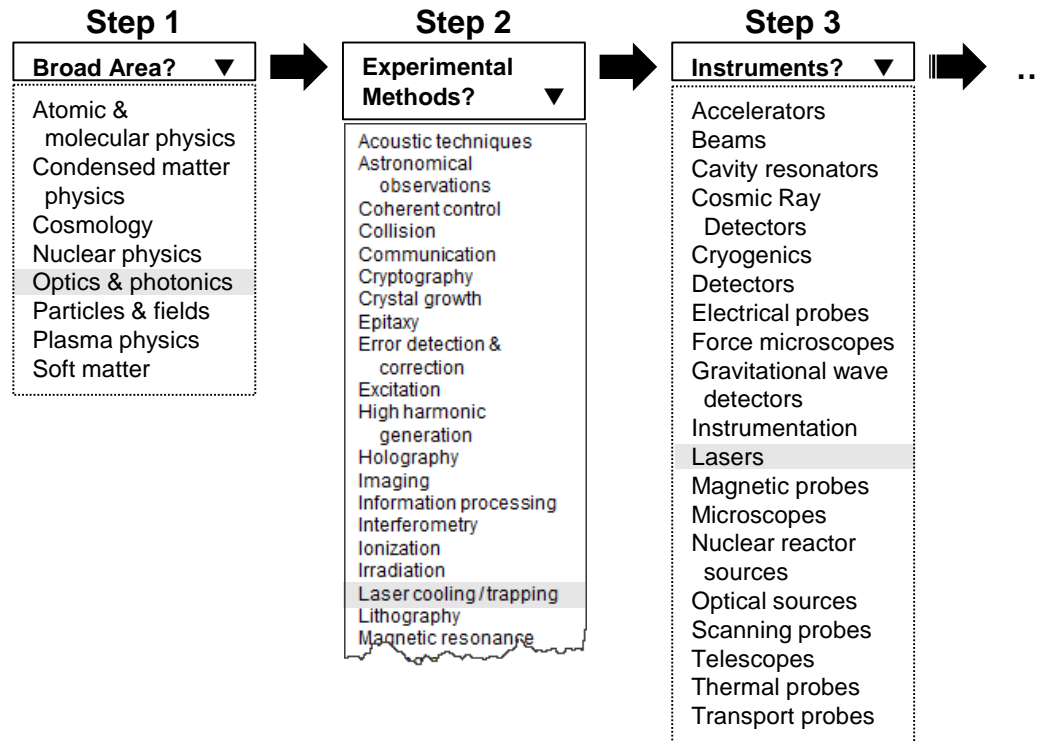| Named Entity	Type or Topic
PowerVault NX400	Network Attached Storage
Barak Obama	U.S. President
Fender Stratocaster	Electric Guitar

❖ Key semantic relationships commonly manifested in web classifications

   ▪ Equivalent (synonyms)
   ▪ Hierarchical (broader/narrower)
   ▪ Associative (related)

Facetted taxonomy example

❖ APS Taxonomy breaks up a complex categorization task into smaller chunks

- What you're studying → Broad area, materials & systems
- Why you're studying it → Phenomena & properties
- How you go about studying it → Apparatus, theory & techniques

Step 1	Step 2	Step 3	
Broad Area? ▼	**Experimental Methods?** ▼	**Instruments?** ▼	…

Step 1 — Broad Area?

Atomic & molecular physics
Condensed matter physics
Cosmology
Nuclear physics
Optics & photonics
Particles & fields
Plasma physics
Soft matter

Step 2 — Experimental Methods?

Acoustic techniques
Astronomical observations
Coherent control
Collision
Communication
Cryptography
Crystal growth
Epitaxy
Error detection & correction
Excitation
High harmonic generation
Holography
Imaging
Information processing
Interferometry
Ionization
Irradiation
Laser cooling / trapping
Lithography
Magnetic resonance

Step 3 — Instruments?

Accelerators
Beams
Cavity resonators
Cosmic Ray Detectors
Cryogenics
Detectors
Electrical probes
Force microscopes
Gravitational wave detectors
Instrumentation
Lasers
Magnetic probes
Microscopes
Nuclear reactor sources
Optical sources
Scanning probes
Telescopes
Thermal probes
Transport probes

9 Common taxonomy facets

Facet	Definition	Example Source
Content Type	Types of content created, managed and used to record or communicate information.	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.	Market segments, Educational stages/grade levels, etc.
People	Names of important people such as authors, politicians, leaders, actors, etc.	LC Name Authority File, NY Times Topics-People, etc.
Organization	Names of organizations, their aliases and the relationships between them.	LC Name Authority File, NY Times Topics-Organizations, etc.
Industry	Broad market categories such as industry sector codes.	North American Industry Classification System, ISIC, etc.
Location	Names of places of operations, activities, constituencies, etc.	ISO 3166, Geonames, NY Times Topics-Places, etc.
Function	Activities and processes performed to accomplish goals.	Federal Enterprise Architecture Business Reference Model, AAT Functions, etc.
Product	Names of products and services that are produced by an organization or people.	Household Products Database, UNSPSC, etc.
Topic	Topical subjects and themes that are not included in other facets.	LC Subject Headings, NY Times Topics-Subjects, etc.

Facetted taxonomy example for business content

❖ Describe the various attributes of a content item to scope its context:

▪ What type is it?

▪ Who was it created for?

▪ What business activity is it related to?

▪ What people, organizations and/or products is it about?

▪ Is it related to particular location?

▪ Is it related to a particular industry sector or market?

▪ etc.

Facetted classification is more like filling in the attributes of a product, than choosing the single most important aspect of a content item.

Describing named entities: Schema.org

- ❖ Identifies named entities in web published content.
- ❖ Rich description that enables Google Knowledge Graph.

People

Organizations

Semantic relationships that can be used to improve information retrieval in search engines

Relationship	Description
Equivalent Label	An alternative label for the Concept such as a synonym or quasi-synonym.
Hierarchical Concept	Hierarchical link between two Concepts where one Concept is more general or more specific than the other.
Related Concept	Link between two Concepts where the two are inherently "related", but that one is not in any way more general than the other.

The equivalence problem

❖ An "exact match" is a high bar, yet some applications such as medical diagnostics require an exact match.

❖ Matching codes, acronyms and abbreviations to descriptive labels can be very helpful in improving search results.

❖ Sometimes it can be effective to roll-up narrower concepts into a broader category

  ▪ E.g., when there is not a critical mass of content for narrower concepts.

The value of name authority files

Common types of hierarchy

Whole-part
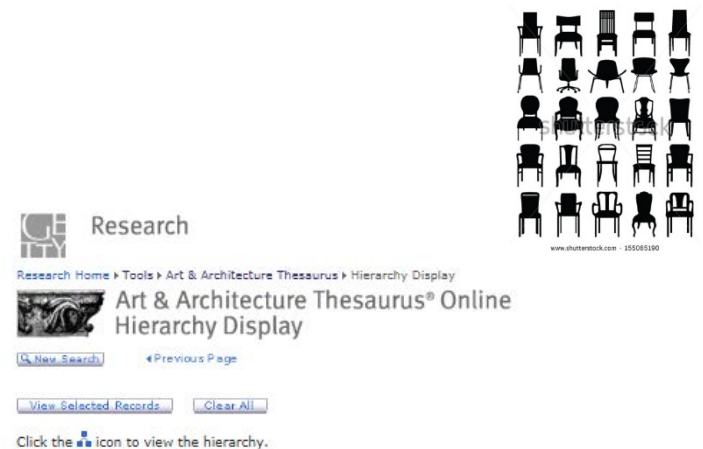
Research

Research Home ▶ Tools ▶ Thesaurus of Geographic Names ▶ Hierarchy Display

Getty Thesaurus of Geographic Names® Or
Hierarchy Display

[New Search]   ◀Previous Page

[View Selected Records]   [Clear All]

Click the ⊞ icon to view the hierarchy.
Check the boxes to view multiple records at once.

- ☐ ⊞ Top of the TGN hierarchy (hierarchy root)
- ☐ ⊞ .... World (facet)
- ☐ ⊞ .... North and Central America (continent)
- ☐ ➡ .... Canada (nation)
-     ⊞ .... [view physical features]
- ☐ ⊞ .... Alberta (province)
- ☐ ⊞ .... British Columbia (province)
- ☐ ⊞ .... <lost & found/Canada> (miscellaneous)
- ☐ ⊞ .... Manitoba (province)
- ☐ ⊞ .... New Brunswick (province)
- ☐ ⊞ .... Newfoundland and Labrador (province)
- ☐ ⊞ .... Northwest Territories (territory)
- ☐ ⊞ .... Nova Scotia (province)
- ☐ ⊞ .... Nunavut (territory)
- ☐ ⊞ .... Ontario (province)
- ☐   .... Prairies (general region)
- ☐   .... Prairies, Lake of the (reservoir)
- ☐ ⊞ .... Prince Edward Island (province)
- ☐ ⊞ .... Quebec (province)
- ☐ ⊞ .... Saskatchewan (province)
- ☐   .... Trans-Canada Highway (highway)
- ☐   .... Ungava district (general region)
- ☐ ⊞ .... Yukon Territory (territory)

Classified Lists

Research

Research Home ▶ Tools ▶ Art & Architecture Thesaurus ▶ Hierarchy Display

Art & Architecture Thesaurus® Online
Hierarchy Display

[New Search]   ◀Previous Page

[View Selected Records]   [Clear All]

Click the ⊞ icon to view the hierarchy.

Check the boxes to view multiple records at once.

- ☐ ⊞ Top of the AAT hierarchies
- ☐ ⊞ .... Objects Facet
- ☐ ➡ .... Furnishings and Equipment (Hierarchy Name)
- ☐ ⊞ .... Furnishings (Hierarchy Name)
- ☐ ⊞ .... furnishings (artifacts)
- ☐ ⊞ .... Costume (Hierarchy Name)
- ☐ ⊞ .... costume (mode of fashion)
- ☐ ⊞ .... Tools and Equipment (Hierarchy Name)
- ☐   .... apparatuses
- ☐ ⊞ .... equipment
- ☐ ⊞ .... tools [N]
- ☐ ⊞ .... Weapons and Ammunition (Hierarchy Name)
- ☐ ⊞ .... weapons
- ☐ ⊞ .... ammunition
- ☐ ⊞ .... Measuring Devices (Hierarchy Name)
- ☐ ⊞ .... measuring devices (instruments)
- ☐ ⊞ .... Containers (Hierarchy Name)
- ☐ ⊞ .... containers (receptacles)
- ☐ ⊞ .... Sound Devices (Hierarchy Name)
- ☐ ⊞ .... sound devices (equipment)
- ☐ ⊞ .... Recreational Artifacts (Hierarchy Name)
- ☐ ⊞ .... recreational artifacts (equipment)
- ☐ ⊞ .... Transportation Vehicles (Hierarchy Name)
- ☐ ⊞ .... vehicles (transportation)

http://www.getty.edu/research/tools/vocabularies/

Related term guidelines are complicated

In Same Hierarchy

- ❖ Overlapping sibling terms.
- ❖ Mutually exclusive sibling terms.
- ❖ Derivational relationships.

In Different Hierarchies

- ❖ Disciplines or fields of study and objects or phenomenon studied or practitioners.
- ❖ Operations or processes and their agents or instruments.

- ❖ Objects or processes and their counteragents.
- ❖ Actions and their products.
- ❖ Actions and their targets.
- ❖ Objects or substances and their unique properties.
- ❖ Concepts linked by causal dependence.
- ❖ Concepts and their units or mechanisms of measurement.
- ❖ Phrases in which the noun is not a true broader term.

NISO Guidelines for the Construction, Format, and Management of Monolingual Thesauri. ANSI/NISO Z39.19 – 2003. "Section 5. The Associative Relationship", pp. 19-20

Curated hyperlinks: Google knowledge graph

❖ Images

❖ Wikipedia snippet

❖ Birth & death dates & locations

❖ Full name

❖ Nationality

❖ Periods

❖ Works

❖ Related searches

Information retrieval has been the key enabling technology for the Web

❖ We are surrounded by named entities and clues about the relationships between and among them.

❖ Underpinning web search is the analysis and annotation of the relationships between entities.

❖ One goal of schema.org is to uniquely identify named entities URI's (Universal Resource Identifiers such as URL's).

❖ URI's are similar to traditional classifications whose codes aim to provide unique locations for physical objects.

❖ Perhaps we have come full circle from universal classifications systems to systems of universal unique named identifiers.

WIKIPEDIA
The Free Encyclopedia

DBpedia

Crowd-sourced classification

LOD

Joseph Busch, jbusch@taxonomystrategies.com, @joebusch

QUESTIONS?