Starting a Taxonomy Project: *Taxonomy Basics*

Alice Redmond-Neal
Chief Taxonomist
Access Innovations, Inc.

SLA Annual Conference, June 9, 2013

#slataxo

Access Innovations.

Let’s talk about taxonomies

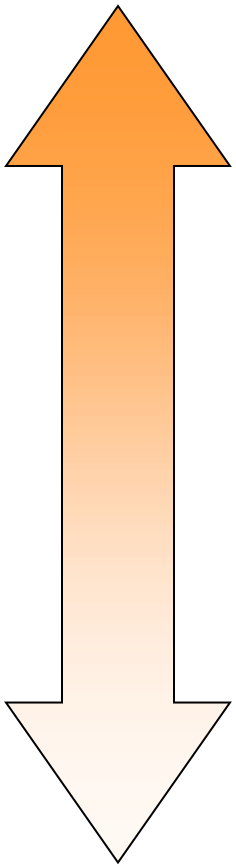
- What’s a taxonomy
- Why use a taxonomy
- Guidelines and standards
- Parts and pieces – term level
- Taxonomy organization – project level
- Review
- *Miraida’s taxonomy in practice*

What's a taxonomy?

- Words – controlled vocabulary
- Organized as a hierarchy
- Labels attached to documents – descriptive metadata to aid retrieval

Terms + Organization + Use

Info management starts with a knowledge organization system



Complex
High value

Simple
Low value

- Semantic network
- Ontology
- ***Thesaurus***
- ***Taxonomy***
- Controlled vocabulary
- Synonym set/ring
- Name authority file
- Uncontrolled list

Taxonomy vs. Thesaurus

- Controlled vocabulary
- Hierarchical format
- (Specific data at final nodes)

- Controlled vocabulary
- Hierarchical format
  - Other formats available
- (Less specific, more abstract tone)
- Network of relationships
- Scope notes, term history
- ***Richer and more informative***
- ***Taxonomy with extras***

Term detail progression

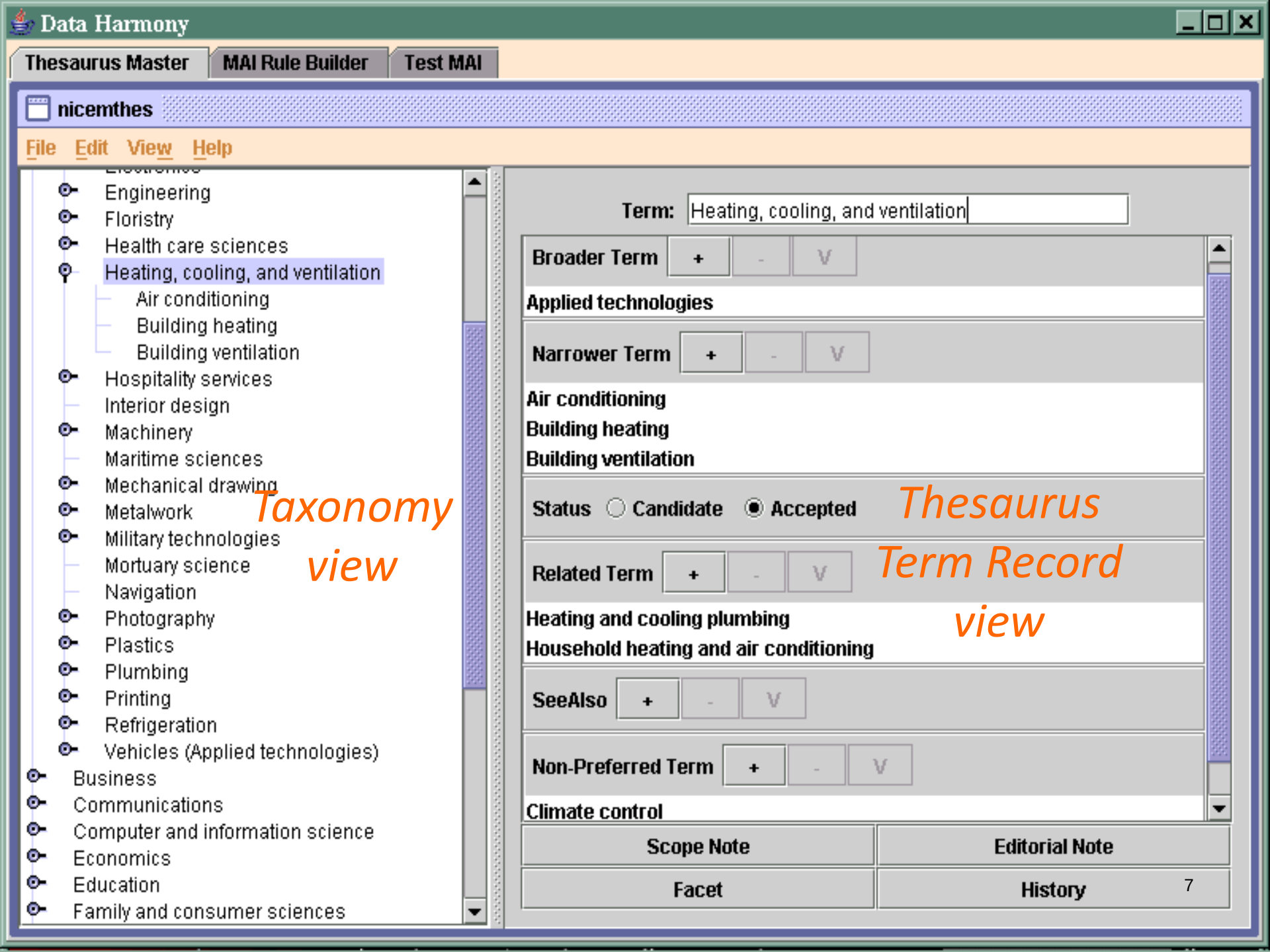
Main Term (MT)
Top Term (TT)
Broader Terms (BT)
Narrower Terms (NT)
Related Terms (RT)
See also (SA)
Non-Preferred Term (NP)
Used for (UF), See (S)
Synonyms
Scope Note (SN)
History (H)

Term = subject term, heading, node, category, descriptor, class

TAXONOMY

ONTOLOGY

THESAURUS


Data Harmony
Thesaurus Master
MAI Rule Builder
Test MAI
nicemthes
File Edit View Help
Engineering
Floristry
Health care sciences
Heating, cooling, and ventilation
Air conditioning
Building heating
Building ventilation
Hospitality services
Interior design
Machinery
Maritime sciences
Mechanical drawing
Metalwork
Military technologies
Mortuary science
Navigation
Photography
Plastics
Plumbing
Printing
Refrigeration
Vehicles (Applied technologies)
Business
Communications
Computer and information science
Economics
Education
Family and consumer sciences
Taxonomy view
Term: Heating, cooling, and ventilation
Broader Term + - V
Applied technologies
Narrower Term + - V
Air conditioning
Building heating
Building ventilation
Status Candidate Accepted
Related Term + - V
Heating and cooling plumbing
Household heating and air conditioning
SeeAlso + - V
Non-Preferred Term + - V
Climate control
Scope Note
Editorial Note
Facet
History
Thesaurus Term Record view
7

Whatever you call it…

- It is a formalized labeling system
- Applies to
  - All content, Site navigation, Site headers, Site labeling
- Used for
  - Indexing/categorization, Search, Browsing/Navigation, Content management, Linking data, Filtering data, Web crawling
- Saves time, money through organization

Controlled vocabulary construction standards

- ANSI (American National Standards Institute)
- NISO (National Information Standards Organization)
  - ANSI/NISO Z39.19-2005
- ISO (International Organization for Standardization)
  - ISO 25964 parts 1 and 2
- BSI (British Standards Institute)
  - BS 8723

Standards and pragmatism

- Standards are your friends
  - Lead to richer, more informative product
  - Promote interoperability
    - Adopt or adapt other controlled vocabularies
  - Promote predictability
  - Allow repurposing within your organization and by other organizations
  - Get respect
- *Your taxonomy/thesaurus must meet your needs*

How to build a taxonomy (thesaurus)

- Define subject scope
- Collect terms/concepts
  - Avoid confusing or ambiguous terms
  - Follow standards for formatting terms
- Group similar concepts together
  - Smallish number of general concepts
  - More specific concepts under general ones
- Add other term relationships and details
  - Synonyms
  - Related terms
  - Usage notes

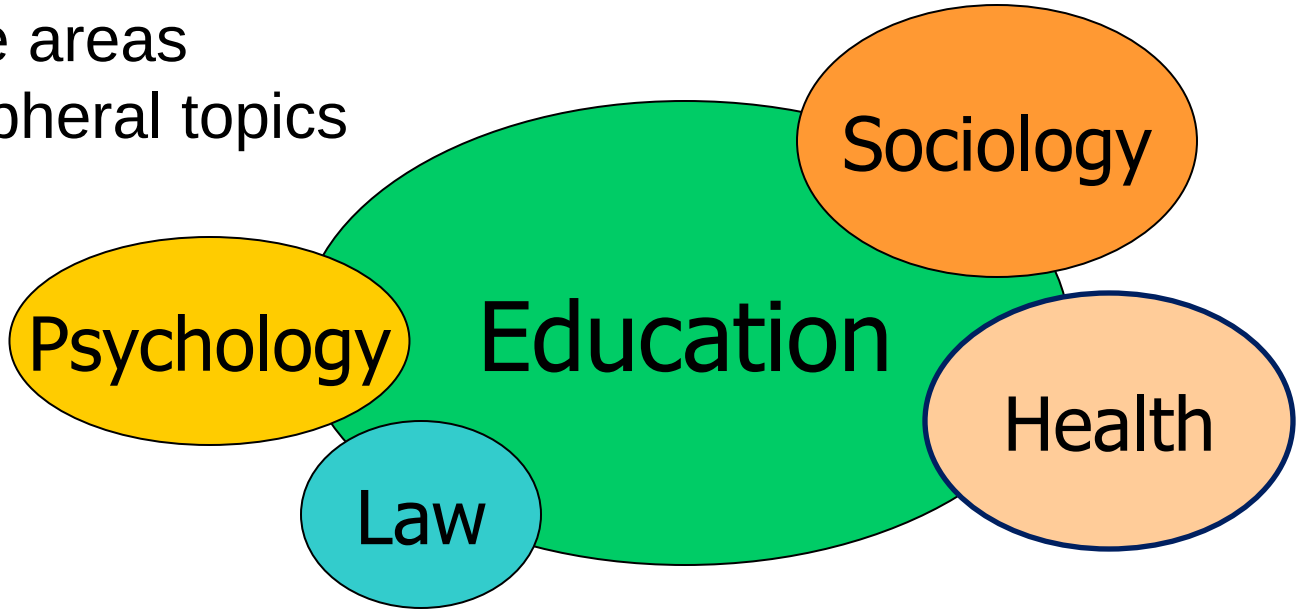
You're done!

DataHarmony

Access Innovations.

Scope – define subject field

- Review representative collection of content
- Determine:
  - Core areas
  - Peripheral topics



- Scope can be tweaked later

What should go in your taxonomy?

- Concepts important for subject area and end users
- Expressed in words they understand
  - Used in the literature, by the organization or community
  - Consider other controlled vocabularies on topic
  - Right degree of specificity or detail
  - Won't be confused with other words/concepts
- *Appreciate what you already have*

Sources for taxonomy terms

Top down approach

- Journals, textbooks, and other domain literature
- Tables of contents, indexes
- Dictionaries, glossaries, encyclopedias
- Lists of all kinds
- Web resources
- Users and experts

Bottom up approach

- Your documents or site content
- Search logs
- Social tags
- Brainstorming
- Custom terms created for concepts
- Collective terms as placeholders in hierarchy

Collect terms to express key concepts

- Concrete objects
  - Physical things and parts; Furniture; Materials
- Abstract concepts
  - Actions; Events; Disciplines; Properties; Measurements
  - Singing; Meetings; Education; Wellness; Value
- Proper nouns – “classes of one”
  - Consider separate authority files for long lists

How useful is each term

- No value if term applies to everything in domain
  - “Education” in a taxonomy on education
  - “Sports” in *Sports Illustrated*
  - “Technology” in *Technology Review*
- How useful will the term be for indexing?
  - Apply to everything in the domain?
  - Distinguish important concepts?
  - If basic term is really needed, specify limited use conditions in Scope Note

One term – one concept

- "Terms … should represent simple or unitary concepts" *(ISO standard)*
- "Each descriptor … should represent a single concept (or unit of thought) … expressed by a single-word term but in many cases a multiword term is required."

(ANSI/NISO Z39.19-2005)

Terms must be unique, clear, mutually exclusive – avoid "and" in most cases

One term – clear meaning

Confusion comes from

- Synonyms – multiple terms with same meaning
- Homo*graphs* – same word/spelling with different meanings
- Insufficient information in a term
  - Term meaning should be inherently clear
  - Should not require hierarchy position to clarify
    - *"lake" "river" "ocean"* = terms for windsurfing

Vocabulary control

- Central to *controlled vocabularies*
- Needed to handle
  - Synonyms – *bushes* = *shrubs*

    De-duplicate, choose one as Preferred Term
  - Homographs – *bush* vs. *Bush*

    Disambiguate terms open to interpretation
  - Relationships among terms

    Organize hierarchical and associative links

Start organizing – Craft the major categories (Top Terms)

- Toughest job and most important step!
- Dictates further organization
- Determines how browsers/searchers perceive the taxonomy
  - Coverage
  - Formality
- How many?
  - Avoid top heavy structure
  - Avoid orphan terms

Organize terms – roughly

- Sort terms into several major categories, i.e. Top Terms – logical groups of similar concepts
  - Identify core areas and peripheral topics
  - 10 – 20 to start
  - Consider moving proper names to authority files
- Result: loose collection of terms under several main headings
  - Rough and tentative
  - Initial gap analysis
  - Add / modify / delete as needed

Connecting terms – How do terms relate?

1 - Hierarchical relationships

-- *Parents and children*

2 - Equivalence relationships

-- *Alternate names*

3 - Associative relationships

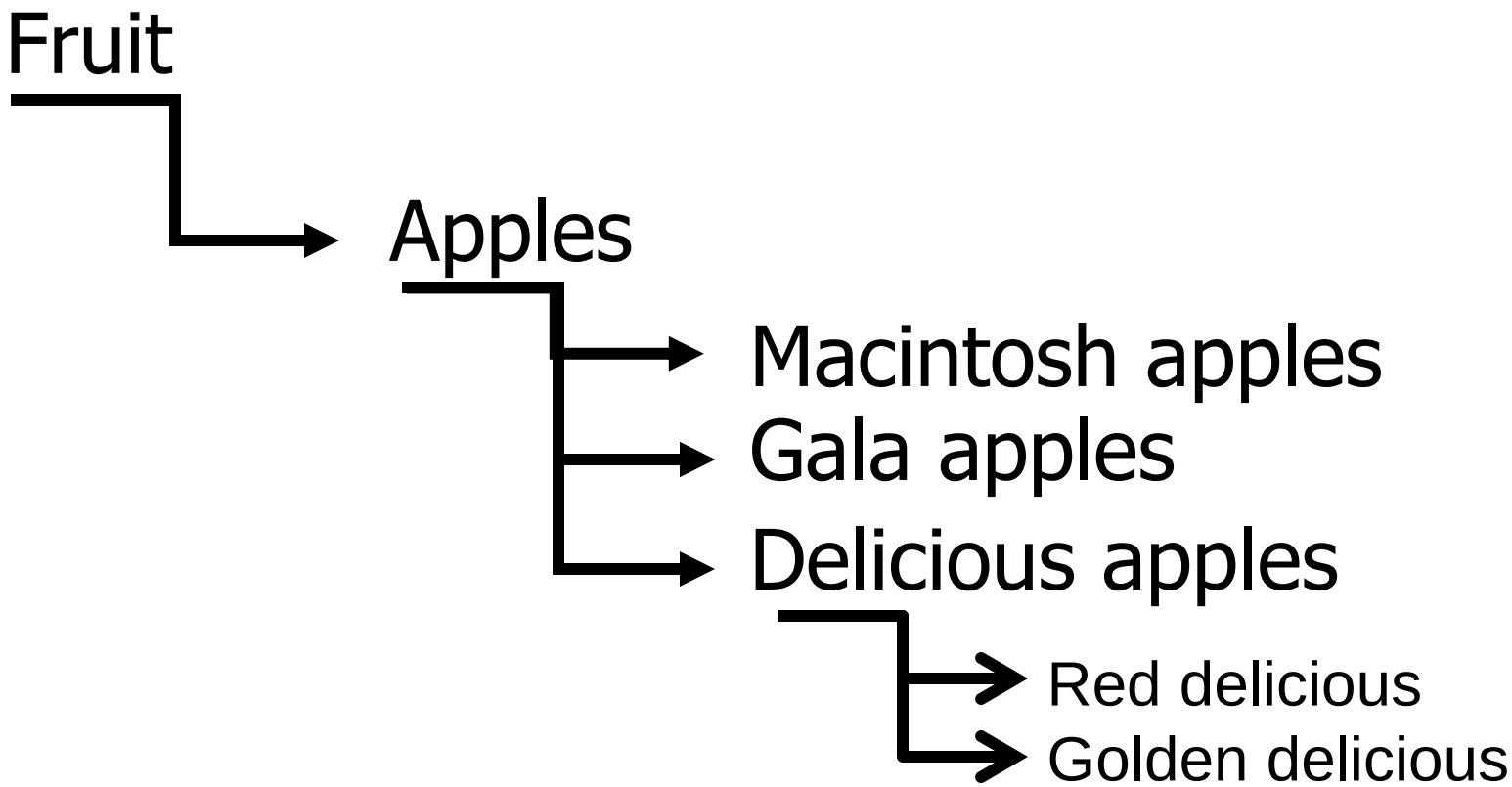
-- *Cousins*

4 - Custom associative relationships

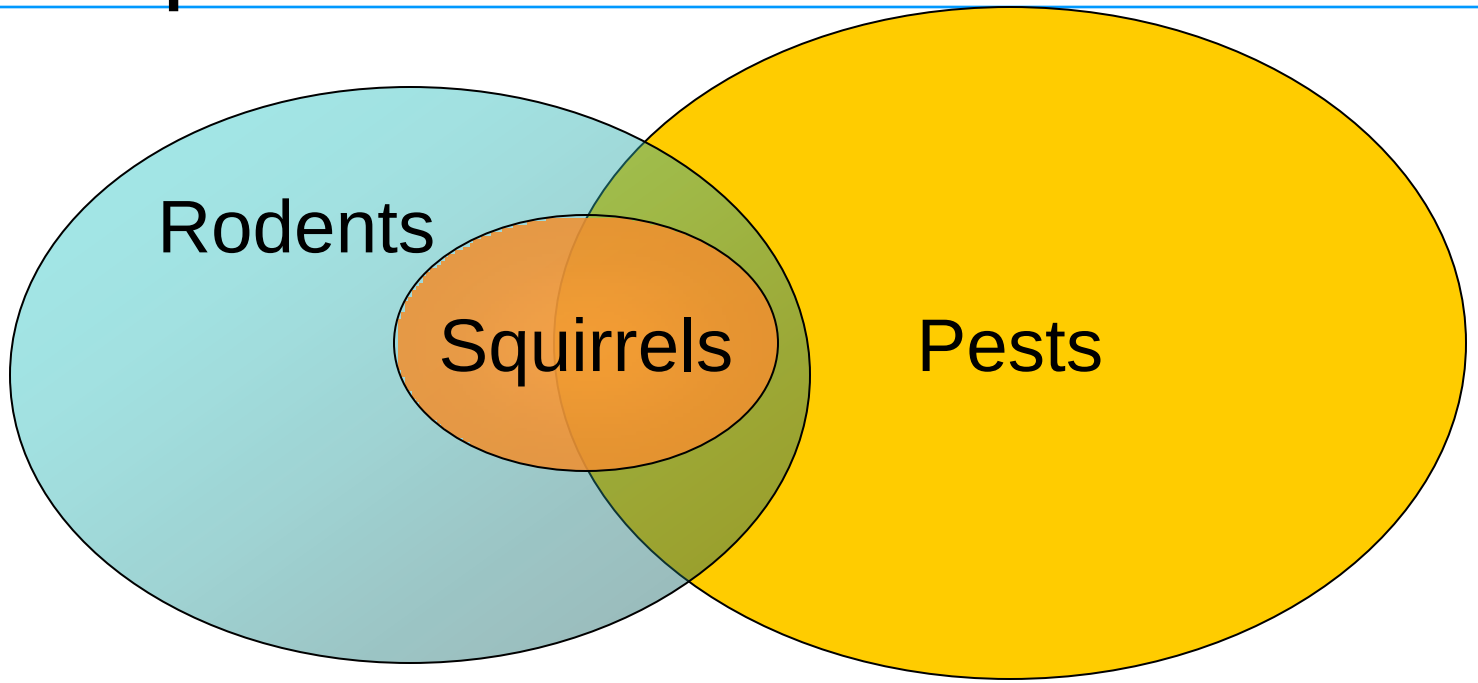
1 - Hierarchical relationship

- Broader Term represents the class, whole, or genus
  - Top Terms are the broadest of all
- Narrower Term is a type, part, or example
  - Generic relationship – NT is a **type** of the BT
  - Whole-part relationship – NT is **part** of the BT
  - Instance relationship – NT is an **example** of the BT

Broader to Narrower Terms



Is the NT a subtype, part, or example of the BT? – the “is a” test



- ✓ ALL squirrels are rodents
- x *NOT ALL* squirrels are pests - overlap
- x *NOT ALL* pests are rodents - overlap

Shifting relationships

- Term can be both BT and NT – depends on perspective

Fruit
Apples
Delicious apples
Red del. apples

Animals
Rodents
Squirrels
Flying squirrels

Inheritance

- What's true of the parent (BT) is true for all children (NTs)
- Can prevent child term's validity for indexing

Furniture

Tables

Tablecloths ?

Mental health

Stress

Durability ?

Polyhierarchical relationships

- Term can logically fit under more than one Broader Term –> Multiple Broader Terms

Spoons
Sporks

Forks
Sporks

Nurses
Nurse administrators

Health administrators
Nurse administrators

- Identical term, identical concept in both places
- Includes any+all children

Siblings – compatible or rivals?

- Siblings' sense and format should be consistent
- Subdivide large groups of siblings into subcategories
- Consider facets, e.g.
  - Price range, color, style, period, source
  - Document type, journal, publication date range

2 - Equivalence relationship

- Preferred Term
  - Thesaurus term, valid for indexing
- NonPreferred Term (synonym/equivalent)
  - Not valid for indexing
  - Entry point, redirects to use Preferred Term

San Diego
UF *Silicon Beach*

Silicon Beach
USE ***San Diego***

Equivalence – when to use

- Synonyms, slang, quasi-synonyms
- Scientific and trade names
  - *Ibuprofen* UF (Use For) *Motrin*™
- Lexical variants
  - *Fiber optics* UF *Fibre optics* *British spelling*
  - *Mouse* UF *Mice* *irregular plurals*
- Narrow concepts not specified in taxonomy
  - *Social class* UF *Elite, Middle class, Working class*

Get equivalent terms from search logs, brainstorming…
Include typos!

3 - Associative relationship

- Related Terms (RTs) – cousins
- Terms related conceptually but not as Broader/Narrower terms and not as synonyms
- Both are valid terms for indexing
- Expands user's awareness
- Standards describe some specific types
  - Doesn't qualify as NT but feels connected → RT
  - Ontologies get creative with associations

Term format details

- Grammatical issues
  - Nouns, noun phrases (adj+noun), gerunds
- Singular and plural forms
  - Plural for count nouns, singular for mass nouns
  - Exceptions: body parts, unique terms, community warrant
- Spelling
  - Depends on audience
  - Capture variations as synonyms

Term format details (continued)

- Abbreviations and acronyms
  - Use the more common expression – always changing
  - Use alternative as nonpreferred term
- Capitalization
  - All lower case is standards compliant
  - Much variation in practice – you choose
- Parentheses – generally ☹ but limited OK
- Hyphens – generally ☹, interferes with search
- *Consistency*

Compound terms (multiword / bound terms)

- KISS – Keep it short and simple
  - 1-2-3 words
- Try to stick with a single concept
  - Avoid *"and"* in most cases
- Many are noun+adj phrases – nouns as "head"
- Keep compound term or factor out elements?
  - Can mix 'n' match elements

    California highway construction →

    California + Highways + Construction
  - Pros and cons

Scope notes

- Show the meaning of the term in the scope of the taxonomy
- Show any restriction in use
- Offer direction for indexers and searchers
- *May* suggest alternative term that users may confuse with the term
- Not necessary for every term
- Use concise and consistent format

So far you've got

- Terms to express key concepts
- Hierarchy
- Complete term records
  - Broader and Narrower Terms
    - Polyhierarchies when needed
  - Preferred/NonPreferred Terms (equivalence rel'ships)
  - Related Terms (associative relationships)
  - Scope Notes
  - Compound terms when needed
  - Correct term format
- Organizational backbone for content management

Review, *edit,* test, *edit,* use, *edit,* and maintain, i.e. *edit*

- Review
  - Users
  - Expert reviewers
- Test
  - Index 500+ documents (more with variable writing style; fewer with strict style)
  - Monitor search log

Consider machine automated / assisted indexing software

- Edit and maintain
  - Add terms
  - Change existing terms
  - Change term status
  - Delete terms
  - Add term relationships
  - Delete term relationships
  - Add/modify Scope Notes
  - Change overall structure

Now what?

- Taxonomy is used for
  - Indexing/categorization
  - Search
  - Browsing/Navigation
  - Content management
  - Filtering data
  - Web crawling
- Taxonomy applies to
  - All content
  - Site navigation
  - Site headers
  - Site “meta name keyword”
- Taxonomy enables
  - Consistent and complete data delivery
  - Redirection to correct terminology
  - Connections to related content

Closing thoughts on taxonomies

- Taxonomy – basis for content organization
- Plan how you'll apply and integrate in workflow
- Plan for growth
- Avoid software or system limitations
- Plan to invest intellectual capital, time, money
- If not building your own, borrow/license an existing taxonomy and customize it
- Start somewhere, start small, but start!

Check these resources

- ANSI/NISO Standard Z39.19-2005
  - http://www.niso.org/apps/group_public/project/details.php?project_id=46
    drill down to PDF
- Data Harmony, Inc. – Best Practices
  - http://taxodiary.com/2010/12/best-practices-for-creating-and-maintaining-a-thesaurus-with-thesaurus-master-or-maistro/
- SLA Taxonomy Division
  - http://wiki.sla.org/display/SLATAX/SLA+Taxonomy+Division
- University of Western Ontario
  - http://publish.uwo.ca/~craven/677/thesaur/main00.htm
- TaxoDiary – articles by M.M.K. Hlava
  - www.taxodiary.com

- Aitchison, Gilchrist, and Bawden
  - Thesaurus construction and use: a practical manual, 4th edition (2000) London: Aslib IMI
- Hedden, Heather
  - The Accidental Taxonomist, (2010): Information Today Inc.
- Lambe, Patrick
  - Organising Knowledge: Taxonomies, Knowledge and Organisational Effectiveness (2007) Oxford: Chandos Publishing
- Stewart, Darin L.
  - Building Enterprise Taxonomies (2008): Mokita Press

Questions?
Comments? #slataxo

Alice Redmond-Neal

ared@accessinn.com

Access Innovations, Inc. / Data Harmony

Ask about a *free trial* of Data Harmony software
for taxonomy and classification

Now see a taxonomy in action!

DataHarmony

Access Innovations.