Managing Taxonomy Tagging

Taxonomy Boot Camp

Washington, DC

November 4, 2019

Heather Hedden & Terry Casey

About Us

Heather Hedden

- Taxonomist
  – Independent consultant, Hedden Information Management
  – Previously employed and contract consultant and staff taxonomist
- Former indexer
  – Periodical article indexer at library vendor IAC (acquired by Gale)
  – Freelance back-of-the-book indexer
- Author of *The Accidental Taxonomist* (2010, 2016, Information Today, Inc.)

Terry Casey

- Taxonomist
  – Independent consultant, Casey Indexing and Information Service
  – Currently contract staff taxonomist
- Back-of-the book indexer
  – Textbook, scholarly, trade book and periodical indexer.
  – Embedded indexes for digital publications

Outline

- Introduction: Tagging, Indexing, Categorizing
- Taxonomy Design and Display for Indexing
- Indexing Policy, Documentation, and Training
- Automated Indexing Methods
- Manual Indexing
- Finding Indexers
- Examples of Indexing Projects

Introduction

Tagging vs. indexing vs. categorizing/classifying

Tagging – assigning metadata labels (“tags”)

- By identifying topics and names within a document or content item
- By content creators or editors (minimally trained in tagging), not as their primary job responsibility
- For metadata both with and without controlled vocabularies
- To support search
- Can also be semi-automated

Indexing – assigning index terms (subject metadata and related elements)

- By identifying topics and names within a document or content item
- By trained indexers, often as their primary job responsibility
- By selecting terms from a large controlled vocabulary/thesaurus/taxonomy
- To create a browsable index (and now also to support search)
- Can also be semi-automated

Introduction

Tagging vs. indexing vs. categorizing/classifying

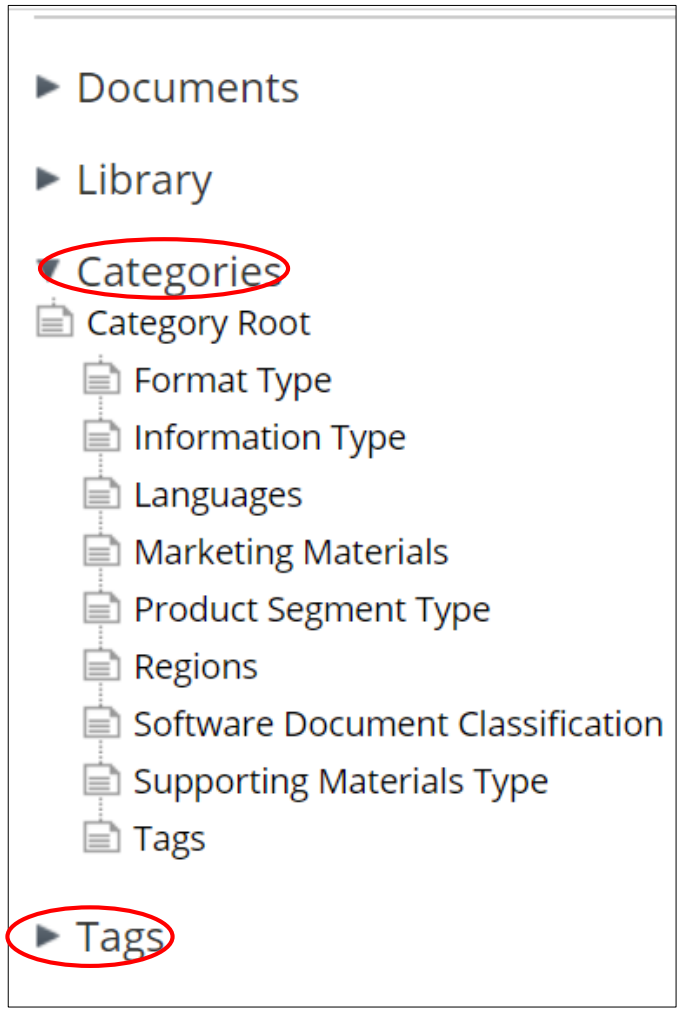
Categorizing/classifying – organizing & assigning content into named categories

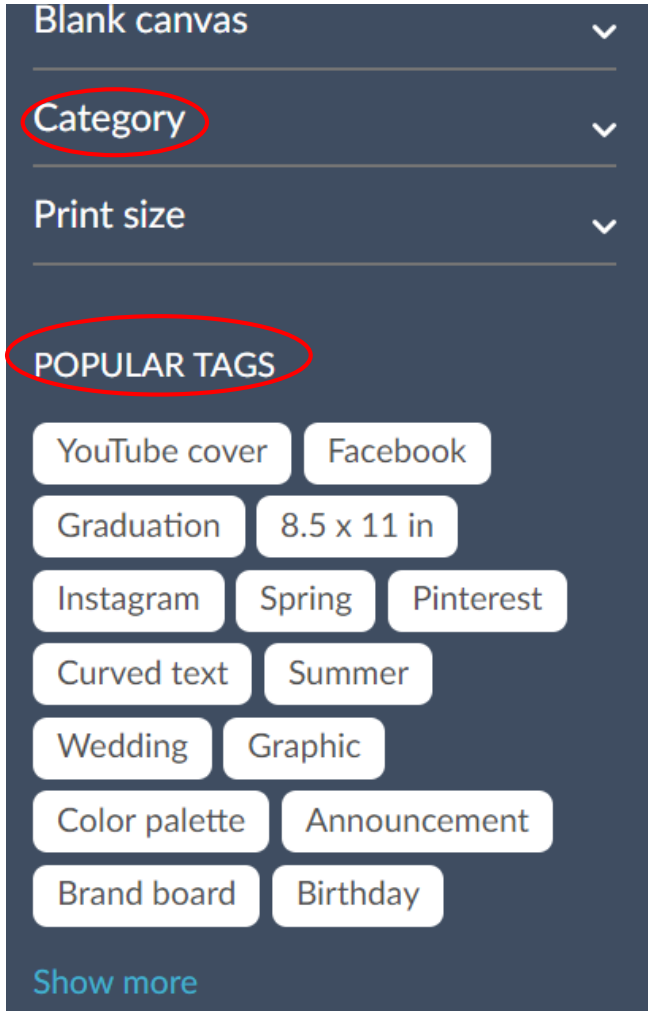
- By identifying which category a document or content item belongs within
- A feature of most content management systems, *in addition to* tagging.
- Often represented as virtual folders and subfolders.
- May be appropriate for Subjects or for Document Types.
- Content items can usually go into only one category, like classification.
- Categories are multi-level hierarchical.
- Category hierarchy is designed as a hierarchical taxonomy.
- Categories may or may not be metadata.
- Can also be automated or semi-automated.

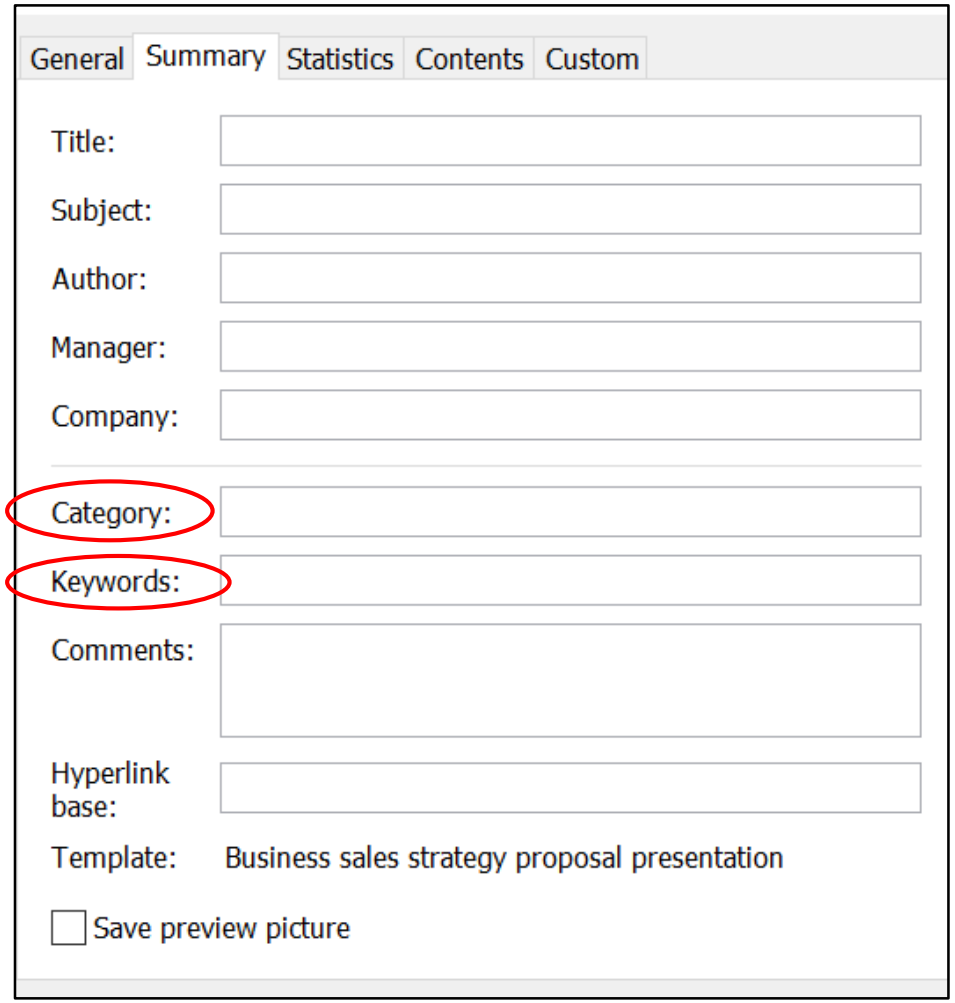
Introduction

Categories vs. Tags

Examples of both categories and tags within the same applications







Introduction

Categories vs. Tags vs. Index terms

Categories 

- What “buckets” the content goes into
- Like a table of contents
- Relatively broad
- Limited in number
- Mutually exclusive
- Sometimes hierarchical
- More controlled
- Pre-planned
- Supports browsing & filtering

Tags 

- What topics the content contains
- Like an index
- More specific
- More numerous
- Overlapping
- Unstructured
- Less controlled
- Ad hoc
- Supports searching & filtering

Index terms

- What topics the content contains
- For an index
- More specific
- More numerous
- Overlapping
- Structured
- More controlled
- Pre-planned
- Supports browsing, searching & filtering

Introduction

Indexing or tagging with a controlled vocabulary or not

Controlled vocabulary

- Using only pre-approved terms
- Used by indexers and content managers
- Ensures consistent indexing
- Slower to change and updates

Keywords

- Assigning any terms desired
- Used by authors and editors
- Tends to inconsistent terms and indexing
- Responsive to trends and dynamic
- May supplement a controlled vocabulary

Folksonomy

- Assigning any terms and reusing terms
- Used by authors, editors, content managers, users
- Tends to inconsistent terms and indexing
- Responsive to trends and dynamic
- May supplement a controlled vocabulary
- More collaborative as “social tagging”

Taxonomy Design and Display for Indexing

Taxonomy design for manual indexing

- Use of alternative labels/nonpreferred terms
  (considering also search or browse UI, from start of term)
- Use of associative (related term – RT) relationships in addition to hierarchical
- Scope notes, dedicated Indexer notes, occasional definitions of terms
- Grouped distinct term sets, hierarchies, or facets for comprehensive indexing
  (even if distinct term sets or facets are not supported in the end-user interface)

Taxonomy Design and Display for Indexing

Indexing user interface and experience (UI/UX) with taxonomy

Tagging interfaces of a commercial CMS are not user friendly.
For large volume manual tagging, develop your own.

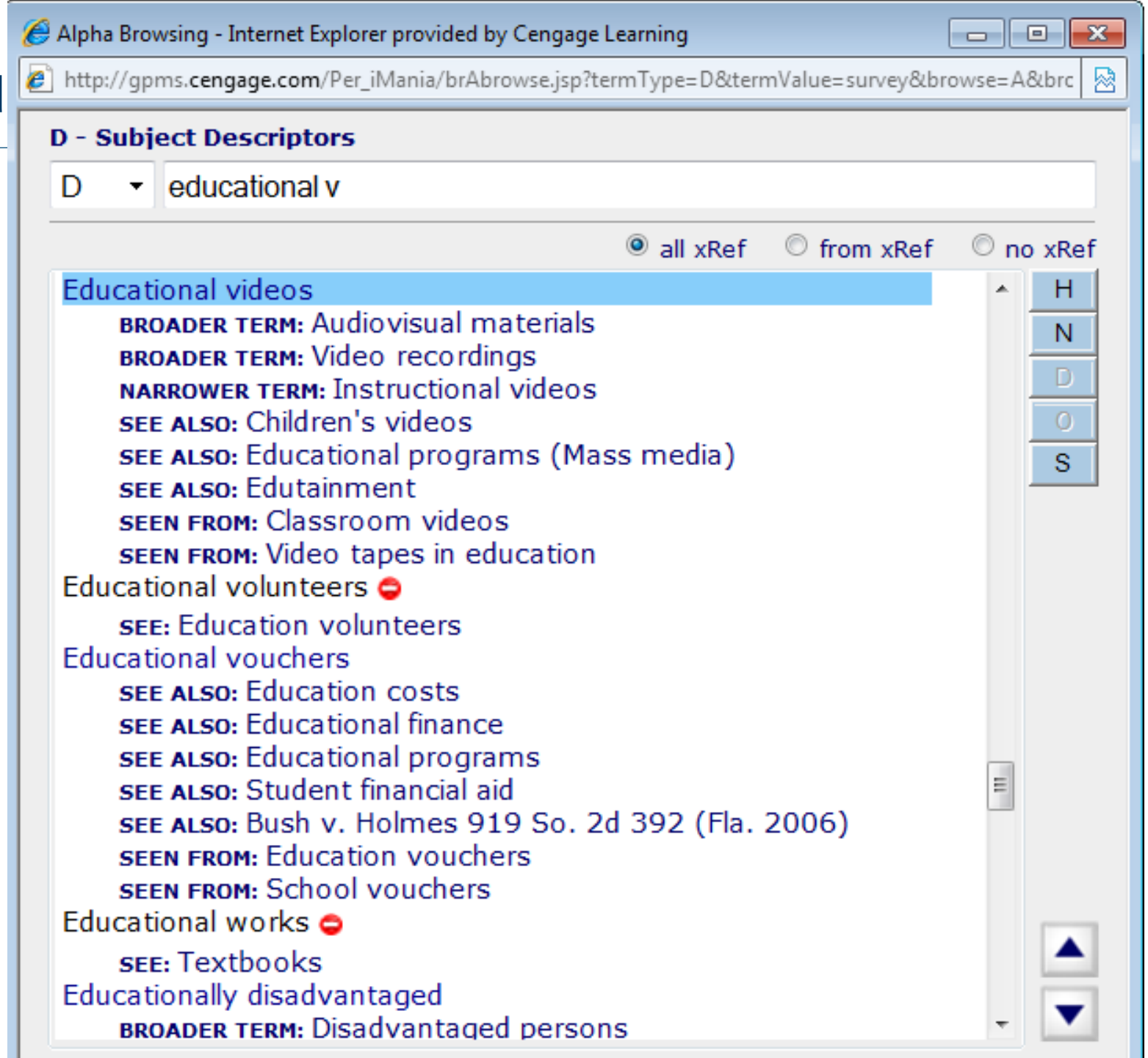
Desirable features

- Both alphabetical and hierarchical browse options
- Alphabetical browse with alternative labels/nonpreferred terms
- Various search options: Begins with, Word/phrase within, Exact, Smart
- Exact term matches are validated and don’t require searching/browsing
- Shortcuts (abbreviations) for commonly indexed terms
- Auto-conversion of selected alternative labels/nonpreferred terms to preferred
- Indexing steps with keyboard shortcuts, and not just mouse, for speed

Taxonomy Design and

Indexing UI display

Screenshot example
(Gale/Cengage internal)



Indexing Policy, Documentation and Training

Indexing policy, rules, documentation, should cover:

- Criteria for determining topic or name relevancy for indexing
- Depth, level of detail
- Comprehensiveness of aspects (what, who, where, when, how, why, etc.)
- Required term types/facets (and any dependencies)
- Number of terms (of each type)
- Indexing of certain terms in combination
  e.g.: a parent/broader term in addition to its narrower/child term
- Other required metadata to enter

➢ Recommendations/guidelines *and* rules/requirements

Indexing Policy, Documentation and Training

Indexer training

- Instructing the indexing policy/guidelines as a live or web presentation
- Training with examples on indexing that captures the “aboutness” of a document rather than matching words in the text to taxonomy terms.
- Reviewing sample indexing and providing feedback.

Indexing Policy, Documentation and Training

Feedback from indexing to improve the taxonomy

Often based on statistics on term usage in indexing

- Underused terms may need added alternative labels or relationships.
- Overused terms may need to be split into more specific terms.
- Misused terms may need rewording, scope note, and/or alternative labels.
- Correctly used low-use terms can be dropped.

Also based on indexers' individual requests and queries

Indexing Policy, Documentation and Training

Indexer-taxonomist communication for new terms

- Taxonomist informs indexers of new and changed terms, and indexing tips (combinations of terms) for indexing new or recurring topics
- Indexers request taxonomist to clarify terms or create new terms

Methods:

- email distribution lists
- Intranet bulletin posts
- collaboration workspace posts
- indexing software feature for new term nomination

Automated Indexing Methods

Automated indexing/Auto-categorization/Auto-classification

2 primary methods: machine-learning and rules-based

Machine-learning based

Automatically categorizes/tags based on previous examples.

- System has complex mathematical algorithms.
- Content managers must provide multiple (10's or more) representative sample documents for each taxonomy term to "train" the system. Results are reviewed and training sets are "tuned."
- Matches are to terms and alternative labels, which can be individually weighted.
- System may also "suggest" additional terms to add to taxonomy.
- Best if large body of pre-indexed records already exists (such as migrating from manual to automated indexing)

Automated Indexing Methods

Rules-based auto-indexing

Rules are created for each taxonomy term.

- Rules are based on synonyms with more conditions.
- Some systems feature weighting of synonyms.
- Some systems feature more sophisticated rule-writing, like advanced Boolean searching (in reverse) and proximity operators or regular expressions.
- Some systems feature auto-generated suggested rules for each term/synonym which can be manually edited in addition to writing rules from scratch.

Automated Indexing Methods

Manual tasks for automated indexing

Continual update work is needed for each new term created.

- New training documents need to be added and taxonomy terms tuned.
- New rules need to created or edited.

- Identifying and tuning training documents is more appropriate for subject matter experts, editors, indexers.
- Writing rules is more appropriate for information professionals, taxonomists, knowledge engineers.

Manual Indexing

Benefits of manual indexing

- Can audit and check indexers' work immediately and make corrections or give instructions
- Can respond to indexers request for new tags quickly
- Can make and/or use compound headings
- Can handle terms that could go under multiple headings and make educated and nuanced decision of where to index the information correctly
- Human interpretation of complex subjects is hard to automate with rules. Autoindexing can lead to inconsistent and uncertain results for complex subjects
- Higher levels of precision and recall: indexers' inconsistencies are minor, compared with potential automated indexing errors.

Manual Indexing

Benefits of manual indexing:

Process advantages

- Handles complex documents that require human interpretation to analyze terms
- Can figure out indexing parameters as you go. Everything does not need to be decided ahead of time-very responsive to change
- Make major structural changes right away because decision maker is right there
- Clean out old, out-of-date, unusable tags while tagging
- Create new tags immediately when needed. Able to adapt and change taxonomy while it is evolving with new information instead of later coming back to find information to tag

Manual versus Automated Indexing

Considerations in choosing an indexing method

Manual methods

- Manageable number of documents
- Higher accuracy in indexing
- May include non-text files
- Investing in people
- Low-tech: can build your own indexing tool/user interface
- Internal control

Automated methods

- Very large number of documents
- Greater speed in indexing
- Text files only
- Investing in technology
- High-tech: must purchase auto-indexing/classification software
- Software vendor relationship

Finding Indexers

Who are indexers and where to find them

- Full-time staff (for ongoing indexing)
  - Editorial or subject specialization background + thorough indexing training
- Freelance or contractors (for temporary or part-time projects)
  - Look for those with periodical article/database indexing experience
  - Many have back-of-the-book indexing experience only (so would require some additional training)
- Subject matter experts or not
  - Scholarly or highly technical documents require subject expertise
  - General enterprise or public content does not require subject expertise
- Rates can be hourly or per indexed record/document
- American Society for Indexing -- for finding freelance/contract indexers

AMERICAN SOCIETY FOR INDEXING

AMERICAN SOCIETY FOR INDEXING

Home | Join/Renew Now | Find an Indexer | Online Store | A – Z Index

- Home
- Find an Indexer
  - Indexer Locator
  - Jobs Hotline
  - Advice for Editors & Authors
  - Sample Indexing Agreement
- About Indexing
- About ASI
- Education & Training
- Webinars
- Reference Shelf
- Conferences
- Publications
- Join/Renew ASI
- Chapters and SIGs

Categories

Newsletter
Webinars
News
Online Learning
Conference

The Voice of Excellence in Indexing

educate, and provide the central resource for indexing.

VIEW CART

Member Center

UserID:

Password:

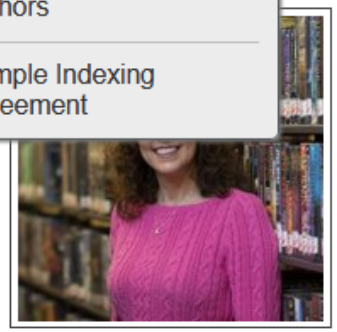
Remember me LOGIN

• Recover Password
• Help

Members: Log in for full web access!

www.asindexing.org

ber 2019



Autumn has officially arrived, and you have two new opportunities to enhance your indexing knowledge. Jan Wright brings us insight on working with clients who need embedded indexing, and Jim Fuhr gives us an intro to James Lamb's CUP/XML WordEmbed. In this issue of See Also, read about the Pacific Northwest Chapter's fall event, discover [...]

Webinars

ASI Webinar Replay: An Introduction to James Lamb's CUP/XML WordEmbed

October 16, 2019 Based on a text of Jim Fuhr's, the presentation will cover the basics of preparing a Word document for conversion into book form by Cambridge University Press or its affiliates. During this one-hour webinar, you will learn: Reasons for embedding; CUP/XML's relation to other embedding programs; Its basic operations: ease and difficulties; [...]

Of Note...

New Indexing Books

Indexing Tactics & Tidbits: An A to Z Guide by Janet Perlman and *Ten Characteristics of Quality Indexes: Confessions of an Award-Winning Indexer* by Margie Towery are now available for

Examples of Indexing Projects

- Educational institution: unique scholarly article for public research access
  - Professional indexer for first phase; will train others on subsequent phases
  - Taxonomy consultant remained available throughout first phase indexing
- International organization: SharePoint intranet taxonomy
  - Request for not just guidelines but also training for tagging
- Fortune 500 firm: enterprise taxonomy for tagging articles

Questions/Contact

Heather Hedden

Taxonomy Consultant
Hedden Information Management
Carlisle, MA
978-467-5195

heather@hedden.net
www.hedden-information.com

accidental-taxonomist.blogspot.com

www.linkedin.com/in/hedden

Twitter: @hhedden

Terry Casey

Taxonomy Consultant
Casey Indexing and Information Service

Saint Paul, MN

651-278-2023

terry@caseyindex.com

www.caseyindex.com

www.linkedin.com/in/terry-casey