HOW TO APPLY YOUR TAXONOMY TO YOUR CONTENT: USING AUTOCATEGORIZATION TO INDEX

SLA 2013 Annual Conference & INFO-EXPO, June 10, 2013

Paula McCoy
Managing Editor, Science & Taxonomy
ProQuest Editorial Operations
Louisville KY
Agenda

• What is autocategorization?
• Why use autocategorization?
• Implementing autocat
• The people side of autocat
• Managing your thesaurus with autocat
• The benefits of autocat
• Summary
What Is Autocategorization?

Automatic categorization ➔ Autocategorization ➔ Autocat

- Use of automation to apply categories, classifications, or controlled vocabulary terms to a set of content (journal or magazine articles, news items, miscellaneous documents, transcripts, etc.).

- Autocat technologies – statistical pattern matching, semantic text analysis, concept/entity extraction, knowledge or rules-base management.

- Autocat solution vendors – Data Harmony, Nstein (Open Text), Endeca, Smartlogic, Teragram (SAS), Temis . . . and on and on.

- Contexts for applying autocat – records management, document and knowledge management, website search, database indexing.
Why Use Autocategorization?

- You need to index, or tag, your content so it can be searched.
- You don’t have time to index everything fast enough by hand –
  - Daily newspaper articles
  - Newswire feeds
  - Reports, working papers, etc.
  - Magazines, journals
- For information aggregators like ProQuest, new content is added all the time.
- Human indexing resources are limited.
Basic Tenets of Autocategorization

1. Must have a controlled vocabulary in place.
2. Must have other controlled lists if you want to index entities, like companies, orgs, people & product names, locations.
3. Must have a way to manage your vocabularies—they will not remain static.
4. Must have a way to manage the results of the autocat—no automated indexing method is perfect; QA is required.

Autocat success rests upon existence of the right vocabulary, applied in combination with human management and control over the autocat software solution.
Implementing Autocat: Scenarios

From one extreme . . .

Have indexed for years using a controlled vocabulary. Want to start autocat on the same set of content. Have a long history of well-indexed content on which to draw in creating an autocat solution.

To the other . . .

Have a large set of content that has never been indexed, want to index all backfile content, but have no controlled vocabulary and no history of indexing upon which to base the autocat.

What you are starting with will determine what autocat solution you pick, what resources you need to implement it, and how long it will take.
Implementing Autocat: Key Questions

- What kind of content would you be applying to autocat?
- Has your content been indexed before?
- Do you have a controlled vocabulary or authority file?
- Does your existing CV match the content you want to autoindex?
- Do you have subject matter experts who can advise on the implementation and use of automated indexing?
- What level of indexing do you want to do?—what kinds of terms, granularity, # of terms expected per document, etc.
- What level of quality do you expect from autocat? Must set standards for accuracy, precision, and recall.
- Will the search engine support your indexing to enable retrieval of content? It is essential to understand how search will work with the indexing.
Define your objectives:

- What do you want to accomplish through autocat? What problem are you trying to solve?

- Do you want to continue what you’re doing now with indexing, or change something?

- Are you looking at an entirely new set of content, or content you already have?

- Is this a one-time, one-phase project, or will autocat be expanded or built upon in the future?
Implementing Autocat: First Steps

Review your content

- **Document type**: News stories, blogs, academic/scholarly articles, magazine articles, audio or video transcripts, reports, datasets, monographs, conference proceedings, white papers, etc.

Document/content type affects autocat results:

- **Article length**—
  - more words = more concepts to tag
  - can the autocat software handle the length?

- **Language**—formal, informal, use of jargon or slang, chatty, English or non-English.

- **Content**—some not suited for autocat, e.g., poetry, fiction, commentary/opinion, editorials, letters to editor, obituaries.

- **Content format**—content must arrive in format that can be read by the autocat software.
Review your vocabulary:

- Do you have a controlled vocabulary or authority file?
  - If not, can you obtain one?

- How sufficient is the vocabulary for the content you want to autocat?
  - Enough terms, the right terms, proper formatting, sufficient granularity

- What kind of shape is the vocabulary in? Do you need to update it before you apply it with autocat?
  - When was it last updated?
  - Do you need new terms for new content types?

- Should every term in the vocabulary be exposed to autocat?
  - Consider suppressing general, vague, or confusing terms, e.g., Management, Water, Security, Users, Diagrams, Shipping.
The People Side of Autocat

Manage the transition of people to automated indexing:

- Resourcing will change. Day-to-day jobs will change. New skills will have to be learned.
- Your current people are the subject matter experts.

Keys to a successful transition:

- Ensure everyone understands the goals and purpose of moving to autocategorization.
- Involve everyone in the transition process so they fully understand the technology and the outcomes and feel better about the changes.
- Train people on the autocat software so they can manage the results.
- Prepare them to accept changes in indexing practices.
Managing Your Thesaurus with Autocat

Manage the indexing by managing the vocabulary:

- Day-to-day review of automated indexing highlights correct and incorrect term usages, leading to greater discipline in vocabulary and indexing management. Those who manage the autocat may constantly question how terms are used, especially in comparison to how human indexers used them.
  E.g., University presidents, University students, Cartels, Child placement

- Humans are expected to index better than autocat software, but are humans always consistent? Do they find all possible terms? Autocat will apply more terms than humans—is that good or bad?

- Close attention to how autocat is using terms can lead to refinements in a thesaurus, such as more scope notes and improved relationships among terms.

- The need for precision in subject terms means terms must be exact and descriptive—autocat will not work with vague, ambiguous term or one-word terms with multiple meanings—”Apologies,” “Affect,” “Articulation” (unless the content is restricted to a very narrow discipline). Thus, new terms are better formulated and defined.

- Some terms may no longer work with autocat. Natural language terms work best.
  E.g., Olympic games-2012, Olympic games-2014, etc. → now Olympic games.
**The Benefits of Autocat**

**Benefits and impacts—planned and unplanned:**

**Expected benefits from autocat:**
- Ability to index more content and get it to the customer faster.
- Improved ability to index different types of content.

**Unexpected results and impacts:**
- Changes to indexing policies—highlighting the need for management.
- Better control over indexing through use of a knowledge base management tool to obtain ongoing improvement and consistency.
- Greater insight into the vocabulary by virtue of autocat’s usage of all terms, in contrast to human indexers who might not be aware of many terms.
- Coordination and collaboration among the people responsible for managing the indexing. Expansion of people’s areas of skill; exposure to new types of work and ways of working.
- Ability to create new indexed products faster and easier because the indexing method is already in place.
Summary

- Autocategorization is often a necessary solution for content providers.

- Properly managed, autocat can work very well and may bring improvements over human indexing.

- Autocat cannot be implemented and then ignored. It must be managed.

- When implementing autocat, know what you want, pick the best solution, and involve the right people.
THANK YOU!

QUESTIONS?

paula.mccoy@proquest.com
www.proquest.com