Against the Grain

Volume 24 | Issue 6 Article 16

December 2012

Indexing and Indices: An Essential Component of Information Discovery

Donald T. Hawkins
dthawkins@verizon.net

Recommended Citation
Hawkins, Donald T. (2012) "Indexing and Indices: An Essential Component of Information Discovery," *Against the Grain*: Vol. 24: Iss. 6, Article 16.
DOI: https://doi.org/10.7771/2380-176X.6230

Biz of Acq Endnotes

1. **Spagnolo, Lisa**, **Buddy Pennington**, and **Kathy Carter**. "Serials Management Transitions in Turbulent Times." *Serials Review*. 36 (2010): 161-166.

2. **Anderson, Rick** and **Steven D. Zink**. "Implementing the Unthinkable : the Demise of Periodicals Check-in at the University of Nevada." *Library Collections, Acquisitions, and Technical Services*. 27 (1): 2003. 61-71.

3. Ibid.

4. **Carr, Patrick L**. "The Claim : Staking Out New Practices for Achieving the Objectives of Check-in." *The Serials Librarian*. 51 (1) : 2006. 74.

5. **Borchert, Carol Ann**. "To Check in or Not to Check In? That Is the Question." *Serials Review*. 33 (2007): 238-243.

6. **Council, Evelyn P.**, **Kittie Henderson**, **Daviess Menefee**, and **Mircea Stefancu**. "Is Checking In Checking Out?" *Serials Review*. 29 (2003): 228.

7. **Carr, Patrick L**. "The Claim : Staking Out New Practices for Achieving the Objectives of Check-in." *The Serials Librarian*. 51 (1) : 2006. 75.

8. **Anderson, Rick** and **Steven D. Zink**. "Implementing the Unthinkable : the Demise of Periodicals Check-in at the University of Nevada." *Library Collections, Acquisitions, and Technical Services*. 27 (1): 2003. 64.

9. **Peritore, Laura**. "Public Access to Serials Check-in Information and its Impact on Reference Services." *The Reference Librarian*. 27/28 (1990) : 17-38.

10. **Anderson, Rick.** "A Sacred Cow Bites the Dust." *Library Journal*. May 1, 2002. 56.

11. Ibid.

12. Ibid.

13. Ibid.

14. **Yue, Paoshan W.** and **Lisa Kurt**. "Nine Years after Implementing the Unthinkable: The Cessation of Periodical Check-in at the University of Nevada, Reno." *Serials Librarian*. 61 (2) : 2011. 248.

15. Ibid.. 232.

16. Ibid.. 240-244.

17. **Anderson, Rick** and **Steven D. Zink**. "Implementing the Unthinkable : the Demise of Periodicals Check-in at the University of Nevada." *Library Collections, Acquisitions, and Technical Services*. 27 (1): 2003. 61-71.

18. Ibid.

19. **Carr, Patrick L.** "The Claim : Staking Out New Practices for Achieving the Objectives of Check-in." *The Serials Librarian*. 51 (1) : 2006. 79.

20. **Tobia, Rajia C.** and **Susan C. Hunnicutt**. "Print Journals in the Electronic Library : What is Happening to Them?" *Journal of Electronic Resources in Medical Libraries*. 5 (2): 2008. 165-166.

21. Ibid. 161.

22. **Anderson, Rick** and **Steven D. Zink**. "Implementing the Unthinkable : the Demise of Periodicals Check-in at the University of Nevada." *Library Collections, Acquisitions, and Technical Services*. 27 (1): 2003. 65.

23. **Decker, Karen**, **Micheline Westfall**, and **Gracemary Smulewitz**. "To Claim or Not to Claim : Claiming Issues in the E-World." *Serials Librarian*. 56 (2009): 123-125.

24. **Yue, Paoshan W.** and **Lisa Kurt**. "Nine Years after Implementing the Unthinkable: The Cessation of Periodical Check-in at the University of Nevada, Reno." *Serials Librarian*. 61 (2011) : 232.

25. Ibid.

Indexing and Indices: An Essential Component of Information Discovery

by **Donald T. Hawkins** <dthawkins@verizon.net>

NFAIS, the **National Federation of Advanced Information Services**, held another one-day workshop on November 30 in Philadelphia. Attendance was about 30 on-site and about 75 virtually, from as far afield as the U.S. West Coast and the U.K. The impetus for the workshop was the current environment of a world where content is being created at a tremendous rate, which is highly beneficial for many reasons, but has the downside that the desired information can be difficult and time-consuming to find.

Information discovery is vital to its users, creators, aggregators, and distributors. And now that an increasing amount of non-textual information is available, discovery has grown more complex, so high-quality indexing continues to be important. This workshop examined some new approaches to indexing techniques, as illustrated by some case studies.

Overview of Indexing Approaches

Joseph Busch, Founder and Principal of **Taxonomy Strategies**, reviewed some of the indexing approaches currently in use. He began with a discussion of four myths of indexing:

1. Taxonomies are monolithic hierarchies. Far from being rigid and unchanging, they are living and change frequently.

2. People retrieve content by subject. In fact, studies have found that retrieval is more often done by named entities.

3. Only librarians can index content. **Busch** said that today's systems have shown that many people have the ability to tag data.

4. All that a search engine can retrieve is a list. However, search today employs a panoply of technologies, allowing advanced retrieval of many different types of information.

Busch went on to say that only 21% of searches are successful. The debate about controlled vocabularies vs. natural language is a result of search failure. Most errors occur because users are unable to find the correct vocabulary terms for their search queries. There are also problems with metadata also. Indexers are inconsistent in assigning terms to categories; classification systems change over time; and different classification schemes may overlap.

These problems are not new. Solutions began as long ago as 1753 when **Linnaeus** published his *Systema Naturae* to bring order into the biological world. Here are some things that can be done today.

- ***Generate more consistent indexes.*** We must build relationships between terms the searcher enters in the query box and those that are used in the content, which will require the addition of semantic resources and processing of language categories.
- ***Correct user errors.*** Search engines can catch some errors, but not all of them. For example, adding synonyms to searches will allow them to be corrected or redirected.
- ***Map the language of users to the language of the target content.***
- ***Augment search results with linked data.*** Developing a faceted tagging and navigation taxonomy and using tools for analysis, visualization, and mashups to present search results will also help by producing predictable standardized structures and consistent semantics so that search engines can understand the content. Google has begun doing this for some sites, as can be seen in these search results for the National Museum of American Art.

continued on page 67

Notice how the search returns not only Websites, but also related links in the right panel, such as maps, photos, hours of operation, etc. The related links are produced using entity extraction and linked data techniques.

Managed vocabularies employ entity extraction for people, organizations, events, products, etc., to form a set of concepts and statements about the semantic relationships between those concepts. The goal is a unique identifier that will allow information to be extracted, and one way to do that is by using taxonomies. A taxonomy is a categorization framework developed by content owners, sometimes with the help of subject matter experts, which is used to tag and index content. Common taxonomy facets include type of content (genre), people, companies, and location.

Images present different challenges. They are "mute" because they do not have searchable text. Algorithms can be developed to retrieve images using metadata which has been attached to the image. Such algorithms are developed using collections of "training sets" of indexed content. Some training sets are available on the Web, such as the WordNet database (*http://wordnet.princeton.edu/*), which contains 117,000 English synonym sets, and the ImageNet database (*http://www.image-net.org/*), which has 14 million labeled images.

Finally, we must consider how to make Web pages indexable so that they will be more findable. Linked data can be used to combine information from more than one source and support mashups; it is now beginning to appear on a number of Websites (for example, see the linked terms on the *New York Times* site).

How do we know if an indexing system is effective, and how can it be monitored? Studies have shown that not only do levels of consistency vary, but high consistency is rare. Semantic tools and processes can help indexers to be more consistent.

Indexing from the Publisher's Perspective

Busch's overview was followed by sessions on indexing from the perspectives of publishers and librarians. Publishers were represented by speakers from the **American Theological Library Association (ATLA)** and the **National Library of Medicine (NLM)**, both of whom discussed the approaches they have taken to increase discoverability of their content. **ATLA** produces the Religion Database (RDB) and the Catholic Periodical and Literature Index (CPLI). The RDB indexes materials published in 103 languages — a significant challenge. Rhetorical language is tolerated in the database, making indexing difficult; thus, computer-aided indexing must be supplemented by human input. Humanities articles frequently have multiple comments, replies, and rejoinders associated with them, also presenting challenges to indexers. Authority files to guide RDB indexers have been created to help indexers resolve ambiguous personal names, oblique titles, and non-Roman scripts. For quality control, each indexer's work is checked by a colleague. Indexing remains critical to the discovery of complex records and will remain so into the future. **ATLA** has found that discovery is supported best by human computer-aided indexing.

NLM's MEDLINE database is created by a staff of 190, which includes specialists proficient in XML and OCR techniques, indexers, and a quality assurance division. It now contains about 20 million records. Medical Subject Heading (MeSH) terms are added to the database by indexers; the average time to scan and index an article is 15 minutes. A Medical Text Indexer (MTI, *http://il.nlm.nih.gov/mti.shtml*) program was developed to produce MeSH recommendations. Discoverability has been enhanced by supplementing the index with terms which do not exist in the MeSH vocabulary, thus providing additional access points. In common with **ATLA**, quality assurance is done by having senior indexers check the work of others.

Challenges to increasing discoverability include:

- The appearance of new content types, such as blogs, non-peer reviewed material, etc.,
- An increasing demand for large data sets with links to individual data elements,
- Emerging new areas of scientific research,
- Changing user needs and increased expectations,
- A continuing increase in the number of MEDLINE journals (100 were added last year, but no additional staff was added to cope with the increased workflow), and
- Leveraging the expertise of indexers by developing their role as curators.

And all these challenges must be met while maintaining the quality of MEDLINE!

Indexing from the Librarian's Perspective

At the **University of Florida (UF)**, search engine optimization (SEO) strategies have been found to significantly increase discoverability. For example, discovery of a collection of documents authored by a biologist was significantly enhanced by adding the author's biography to Wikipedia. Many of today's students do not know how to apply general online searching skills to the scholarly research environment, nor do they know how search engines work. They also tend to be naive and trusting of whatever results they get and may regard librarians merely as book curators or guides to the library building. Students' experiences can be improved by:

- Offering information literacy classes and using creative techniques such as gaming techniques to engage users,
- Creating better finding aids and improved metadata because OPAC records are not easily discoverable in general searches (federated systems can help solve this problem), and
- Using SEO.

Because they know their collections best, curators should work to strategically create SEO content in Wikipedia or blog entries, which will improve search results and provide links to highly-ranked sites.

Donors have given special collections that have unique metadata needs to the **Free Library of Philadelphia (FLP)**. Digitization of photo collections at **FLP** began in 2000, and as each item was scanned, a MARC record for it was created by catalogers. But because no central department controlled the cataloging, inconsistencies between digitization and cataloging developed over time. Some collections were curated; others were not; and separate rules governed the cataloging of each collection. To correct this disconnect, a Digital Collections Application (*http://libwww.freelibrary.org/diglib*) was created. The term "cataloging" was banned and replaced by "describing." Catalogers were thus free to use non-MARC terms and could be more flexible in their descriptions. A thesaurus subject engine containing over one million subject terms was created from a mashup of several different thesauri.

Case Studies

The afternoon featured case studies on automated indexing, managing vocabularies, and indexing non-traditional content.

Automated Indexing

The **American Association for Cancer Research (AACR)** has over 30,000 members and publishes eight scholarly journals. Why would a journal publisher decide to create an indexing system? And when members showed interest in an item, how could they be alerted to related items? Communication with the Association's members about activities of interest and new content had become difficult, so a proposal was made to classify all of the **Association's** content using a standardized vocabulary, which would allow a member viewing a journal article to be informed about a conference on the subject, other journal articles, relevant job postings, podcasts of interviews with other researchers studying the same topic, etc.

Access Innovations, Inc. (AI) developed a taxonomic structure and indexing rules for **AACR**. The **Association** provided MEDLINE records for articles in its journals, conference abstracts, and other data, and **AI** created the indexing rules, which were then reviewed and confirmed by subject experts (members and journal editors). An indexing process was added to the journal production workflow. A recommendation feature similar to **Amazon's** "people who bought this also bought..." is now undergoing beta testing and will soon be available for **AACR's** journal sites.

The **JSTOR** service, with 60 million pages of journals and 14,000 newly added books, has a significantly different problem from **AACR** because its content spans many topics. **JSTOR's** problem is the variation in subjects and topics; thus, semantic indexing was considered as a solution.

Semantic indexing is what the content is about, not its appearance or layout. (XML is used to describe both, but semantic tagging is "like XML on steroids," which has implications for discovery.) Semantic tagging can ensure that all synonyms are treated identically and point to the same place. The tagging is traditionally done by human subject matter experts.

Brute force searching has its limitations: it is costly and does not scale. Humans work at

continued on page 68

Indexing and Indices ...
from page 67

the controlled vocabulary level, but software works at the tagging level, so the human eye is always needed. Machine-aided indexing is a good middle ground and provides the best of both worlds. Users are increasingly demanding more value from content providers than brute force search can deliver, and content providers are in the best position to satisfy them because they know their disciplines and their terminology.

JSTOR did a pilot indexing project with three disciplines, built rule bases using their thesauri, and automatically indexed them. The pilot was successful, so **JSTOR** is now looking for controlled vocabularies to license and use to create a "**JSTOR** Thesaurus," which can be maintained by a staff of librarians without the need for subject matter experts.

Managing Vocabularies

Controlled vocabularies must be managed. They enable consistency in indexing and ways for searchers to find content, and they must be readily available to indexers and users. Because names and relationships among terms constantly change, vocabularies cannot be simply created then left alone. Without formal control, there is no system for adding new terms, either procedurally or technically, and indexers do not know to whom they should submit suggestions for new terms. And when terms are added, the formats are not consistent.

As the amount of content increases, vocabulary management and control become a full-time job for one or more persons, which was the case at **ProQuest**. The **ProQuest** vocabulary management department makes both procedural and technical decisions about the maintenance of the vocabulary. The procedural decisions include ensuring that the vocabulary adheres to industry standards such as ANSI/NISO Z39.9, establishes procedures for indexing names of entities which may change frequently, decides how users will suggest new terms, and what will be the criteria for accepting them. Technical decisions include selecting vocabulary management software tools, ensuring that they are available to the indexers, and developing methods for rapid updating.

Discovery of Non-Traditional Content

The **Department of Energy (DOE)** is increasingly being called upon to handle emerging forms of scientific information, such as videos of presentations, guest lectures, and recorded experiments, and many **DOE** facilities are beginning to make these available on sites such as YouTube, Vimeo, and SciVee. This type of information presents challenges because no transcripts exist, so there is no "full text" to search; metadata, if available, is often minimal; the vocabulary is highly specialized; and the videos are often lengthy, lasting up to an hour or more. Because of these challenges, searching this content is problematic.

DOE's Office of Scientific and Technical Information (OSTI) entered into collaboration with **Microsoft** using its **Microsoft** Audio Video Indexing Service (MAVIS), which uses state-of-the-art speech recognition technology to digitize and enable searching of spoken content. (**Microsoft** had not worked previously with an STM vocabulary and was anxious to experiment with one.) The MAVIS technology handled different voices and accents very well, which led to the launch of **DOE's** ScienceCinema (*http://www.osti.gov/sciencecinema*) in February 2011. Content from meetings, conference calls, voice mail, presentations, and call centers can be searched. The system is especially useful for videos because they can be searched for the occurrence of a word without the necessity of watching the entire video. The user experience is like searching for words in a document, and the search results are highly accurate. ScienceCinema now contains over 2,600 videos from **DOE** sites and **CERN** (the **European Organization for Nuclear Research**).

The formal presentations concluded with one from **RSI Content Solutions** (formerly **Really Strategies, Inc.**) looking at some of the basic challenges currently being faced by publishers, which mainly revolve around the question, "How can we affordably create and deliver primary content to multiple channels while also developing new digital products?" A few publishers have already faced the issues of internal content challenges, efficient processes, and content markup. They are ahead of the pack and are aggressively experimenting with metadata development and management.

Acquiring metadata gives publishers a strategic advantage — metadata *is* content! It should be stored in a separate repository outside of the content management system, which will produce enhanced control and flexibility and easier updating processes. In such a configuration, it will not matter whether the content is textual, binary, or a database.

Here are three examples of forward-looking publishers that are using these principles:

1. Oxford University Press (OUP), wanted to create an index to its publications (the *Oxford Index*). It had the vision and the resources (knowledge of content, customers, and processes plus the technical insight, infrastructure, and the will) to complete the project. Such a project takes time, is expensive, and requires special expertise. These types of activities are fundamental to publishing. **OUP** is still defining some of the processes, investing in automated systems, and identifying links.

2. Audible.com, a supplier of audio books, invested as much effort in obtaining metadata about their products as in getting the audio itself. They received ONIX metadata from publishers, normalized it, and published it with their own products.

3. Meredith, publisher of *Better Homes & Gardens* and other magazines, developed a "standard recipe markup language," combined it with user-generated content, and organized it via metadata. Its service was enabled using an XML standard, and it can capture ratings and reviews by users.

These examples show that several publishers are experimenting with metadata, some of which is embedded in media files and some is based on related or extracted text. Currently there is little manual metadata assignment of non-textual content except where it has immediate business value.

Publishers are advised to push discoverability into their operations, build costs into standard operational budgets, invest in internal expertise, plan for ongoing evolution, and accept failures as learning experiences. Development timelines are long, so they must start now or risk losing their business.

Speaker slides are up on the **NFAIS** Website at *http://www.nfais.org/page/378-indexing-and-indices-nov-2012*. A follow-up workshop entitled "The Future Role of Abstracting and Indexing Services" will take place on March 15, 2013.

Donald T. Hawkins *is a freelance writer for* ***Information Today*** *and other publications. He blogs the* ***Computers in Libraries*** *and* ***Internet Librarian*** *conferences for* ***Information Today, Inc. (ITI)*** *and maintains the* ***Conference Calendar*** *on the* ***ITI*** *Website (****http://www.infotoday.com/calendar.asp****).*

Booklover — Poland

Column Editor: **Donna Jacobs** (Research Specialist, Transgenic Mouse Core Facility, MUSC, Charleston, SC 29425) <jacobsdf@musc.edu>

Poland is a country that has always fascinated me, probably because I grew up during the cold war and propaganda was my only source of information. It was hard for me to believe that people would be so different from what I knew to be true from my own surroundings. Working at a university I have had the good fortune to meet several people from this country. Many have become lifelong friends. These relationships afforded me the opportunity to travel to Poland in the early 1990s and discover that my theory about "normality" was true. Over the years I have been given numerous gifts of Polish origin, including three books written by Nobel Laureates. Two of the works are in English and one is in Polish. I have made attempts to wrap my tongue around the Slavic sounds and numerous consonants of this language. I am forever away from mastering this. However, I enjoy hearing the language, embracing the culture, and now and then chasing pickled herring with a shot of cold vodka.

The country of Poland has produced five Nobel Laureates in Literature: **Henryk Sien-**

continued on page 69