Revisiting Indexing and Abstracting in the Digital Era

Samantha King, Howard Boyedoe, Andrea Chacon, Molli Hall, Kathleen O’Bryant,
Madison Meagher, John Puga, Dr. Daniel Alemneh

February 12, 2018

University of North Texas
Revisiting Indexing and Abstracting in the Digital Era

Samantha King, Howard Boyedoe, Andrea Chacon, Molli Hall, Maria Krizansky, Madison Meagher, Kathleen O’Bryant, John Puga, Dr. Daniel Alemneh

University of North Texas

Abstract

The synergies of current trends are shaping creation, discovery, access, use and reuse of information resources. The digital landscape has exploded, providing greater access to a larger portion of the world’s population since the creation and domestic spread of the internet. This new frontier has created a surplus of information and a need for increased organization. The variety of platforms, media types, and metadata criteria available for organization require innovative approach and methodology development for efficient information organization. Automation practices are enhancing, but have not reached peak efficiency for indexing and abstracting processes. Social networking tools, such as social tagging in image databases, have shown increasing success with the creation, distribution and sharing of information. Web 2.0 is influencing how involved users can be in classification and indexing processes. The future of indexing and abstracting will require continuous facilitation of online content management in varying contexts due to the increase in available resources and digital storage capabilities.

This proposal discusses and provides an overview of some of the emerging trends and argues that indexing and abstracting services are still the path to rediscovering and effectively accessing knowledge in the digital era. Users’ needs, digital library developments, information access methods and indexing techniques will need to be routinely assessed as time passes to maintain effective information delivery.

Keywords
Introduction

The future of indexing and abstracting, based on the current literature, appears multi-faceted. With the shift towards a more socially accepting internet, Web 2.0 is influencing how involved users can now be with the classification and indexing processes. There are ongoing developments not just in computer capabilities, but also the continued spread of information via digital libraries, focus on user needs and experiences, and shifting of indexes from physical texts to digital texts and media. The shift towards digital media is extremely evident in the use of social tags for image databases. Image indexing has long been a source of contention for indexers, but with the inclusion of social tagging, this could change. What originally began on social networking sites like Flickr, has migrated to collections from art museums and libraries. Though this trend is gaining popularity, it is still important to understand both the pros and cons that come with it. This paper aims to present discussion on current trends regarding indexing and abstracting, while analyzing the literature and trends to predict shifts that may occur or factors that could significantly impact changes within the field. It also aims to analyze the concepts of image indexing and folksonomies, and how the collaboration of the two seeks to solve their individual problems.

User Interfaces: Google Scholar and Other Services

Abstracting and indexing (A&I) services have long been a primary means for locating scholarly information. With the advent of Google Scholar in 2004 there has been a transition from
subscription-based services, such as ProQuest and EBSCO, to general search engines, like Google Scholar, that provide similar services at no additional cost or training. Tucci (2010) states that, “The quest for a quick answer and instant gratification often circumvents the use of traditional fee based A&I services that take some knowledge and training to use successfully.” Tucci further goes on to explain that A&I services are expensive, but do not have interfaces that are easy for users to operate immediately based on instinctual or learned search techniques. This criticism is echoed by ProQuest’s President Tim Collins, who has said that relevancy is the most important criteria that needs meeting in order for users to be satisfied with their experiences. Library databases need to reflect that kind of sophistication to stay relevant and useful enough to users (Collins, 2013). Studies have also shown preference for Google Scholar over other scholarly databases among undergraduate students, as reflected on by Johnson & Simonsen (2015) about work done by “Salisbury et al. (2012).” This is said to be because searchers rely heavily on immediate access to electronic information and forgo the value librarian competence adds (Haglund & Olsson, 2008). To bypass this kind of critique and absence of a need for librarians, these professionals need to adapt to their users’ needs and see which techniques or other services they can employ that would better aid their users, instead of operating in ways that drive them to other tools without a second thought.

In another study, Meier and Conkling (2008) found that Google Scholar contained ninety percent of the engineering literature published after 1990 in Compendex, the premier engineering database. Meier and Conkling (2008) report that an average of 84% aeronautical, 56% civil, 71% computer, 68% electrical, 65% environmental, 48% industrial, 51% mechanical, 82% nuclear, and a 66% overall match of sampled Compendex records also appeared in Google Scholar (from the 1950s to the 2000s). These studies succeed in highlighting the importance of how “the easy-
to-use interface of Google Scholar also has appeal for individuals with more advanced information needs such as graduate students, faculty, and research staff” (Meier & Conkling, 2008, p. 201). Meier and Conkling (2008) bring up another good point when it comes to A&I services: their cost. We have to wonder how people outside of the academic and public library setting without access find academic information in A&I services, without having to pay. Google Scholar may be an answer. Meier and Conkling (2008) explain:

> Acquiring experience with Google Scholar may serve the students well after they graduate since not all engineering firms have the resources to supply their employees with access to commercial databases. For engineers in many companies, Google Scholar may offer the best way to review the publishing activity in their fields.

(p. 201)

Napp (2007) found that access to commercial engineering or science databases is surprisingly rare in most engineering firms. It is not unlikely to find similar results in many other fields. Michael Gorrell, chief information officer at EBSCO believes that “discovery services can provide a Google-like experience in terms of fast, single-search access to a library’s collection, but in academic research, the quality of results is the distinguishing factor”, according to Kelly (2012, p. 3). However, Kelly also discusses that A&I providers are concerned about discovery services and worry that they will threaten their services, thereby choosing not to implement their bibliographic databases to them (p. 9). To combat this, A&I providers introduced discovery, “which is modeled on the Google style approach of building and then searching a unified index of available resources, instead of searching each database individually” (Luther & Kelly, 2011, p. 1). Luther and Kelly (2011) explain the process:
In effect, discovery tools make good on the promise of those earlier search solutions by shifting some of the IT management responsibilities to the cloud, streamlining search, and improving the relevance ranking of results. And users get to enter a single query—à la Google—to search the rich content of the collection with the speed they have come to expect. Still brand new, and in action in only a handful of academic libraries, these tools are expected to transform search as we know it. (p. 1)

Luther and Kelly (2011) also state that, “the new unified-index discovery tools offer great potential for simplifying scholarly search and making it more effective” (p. 2). However, there are still many things to consider. They additionally say that, “Librarians have narrowed in on certain features and capabilities that are key to making decisions about these tools,” and, “naturally, different institutions weigh each factor differently based on local needs and objectives, collections, users, and staffing” (p. 2). As each institution adapts to their user needs, indexers and abstracters need to evaluate and continue to develop discovery tools that will continue to raise efficiency of researching and locating information. They need to simultaneously make sure that each tool they use fits within other criteria such as budget, expertise level, and operability.

Another issue that A&I services face is the need for academic institutions to teach scholarly communication. “Libraries spend a great deal of time teaching information literacy concepts to entering students” (Kesselman & Watstein, 2005, p. 384) and “much of this instruction includes an introduction to scholarly resources, database selection and use, and evaluating resources” (p. 384). Because of this need for instruction, there are many advantages of using Google Scholar
over traditional A&I services. Google Scholar, according to Kesselman & Watstein (2005), “is free and easy to use; results pages do not contain advertising (at least for now); it does a good job of delivering scholarly information; the "cited by" links are especially handy; there are many links to free full text; sometimes it is actually searching and retrieving from full text” (p. 384). For A&I services to compete with Google Scholar and provide better search functions for their users, they need to bridge the gap between subscription-based services and open access information. The need to keep information under lock and key may be the deciding factor in choosing one service over the other, especially in times of never ending budget cuts. If users are not members of academic communities, then Google Scholar is one of their only means of finding information. Google Scholar is free, it is accessible by every computer and mobile device capable of connecting to the Internet, and it requires very little training. In an era of same day shipping (Amazon), instant streaming services, and mobile apps for everything, the idea of having to go to a library to access content is becoming basically non-existent.

Certain abstracting and indexing services, particularly those involving traditional print sources, have not been affected by this in terms of indexing approach and methodology. Indexers of digital materials that require frequent retrieval, such as academic journals within database services, will need to be aware of the challenges and perceptions discussed above. They will also need to be aware of which locations materials will be most likely retrieved from by users. In order to improve these services, we need to increasingly evaluate the interfaces of each provided service and keep the user in mind. Users are shaping the future and causing many of the current shifts that are happening in the indexing and abstracting field.
Automatic Indexing and Abstracting

Due to the increasing volume of information that needs organization across all platforms, it is becoming necessary to automate processes such as indexing and abstracting. The information surplus that comes from content generation in Web 2.0 is too much for individual professional human indexers to process alone. This creates a need for automatic indexing. For example, Névéol, Shooshan, Humphrey, Mork, and Aronson (2009) estimate a "45% increase in the indexing load" (p. 814) for the National Library of Medicine's journal database from 2007 to 2015. Similarly, the German National Library increased its collection from an estimated 20,000 works annually to over 187,000 in the past decade (Junger, 2013). These libraries are investing in automatic indexing research to ensure their growing collections are properly organized for easy access by users.

Automatic Indexing Methods

For automatic indexing to be successful, multiple methods are required. Not only do different types of indexes require different programs, but a single index on a specific subject area will require a combination of different methods to achieve better results. The reason multiple methods are required is that each method has a strength, such as precision or recall, which can lead to an overall increase in both when combined in the right way (Névéol et al., 2009). However, finding the right combination of methods can be difficult and possibly still might not achieve a satisfactory result. The types of methods used in automatic indexing of text information include assembly, rule-based, and statistical (Névéol et al., 2009). Assembly methods, also called "'jigsaw puzzle’ method" (Névéol et al., 2009, p. 816), involve piecing
together an index from words that can be taken from either preexisting categories or a dictionary built from analyzing the information object. Various techniques can then be used to determine which words selected have more relevance. In rule-based methods, the index is created following rules that have been previously outlined. These include specific indexing rules that may be required by the index owner and natural-language processing. The statistical method uses data from previously indexed information to index similar objects consistently.

**Multimedia Indexing**

The methods previously mentioned have been in the context of textual documentation, but the automatic indexing of multimedia information objects is another growing area of research. Verbal audio and visual text is not too different from document indexing, as once the speech or text images are converted to text using the appropriate recognition software the above methods can be applied. Multimedia that does not lend itself to be directly translated to text is more challenging to index, as they usually only have limited textual data, such as title and artist (Hwang, Yang, & Ting, 2010). When analyzing multimedia content, it is important to note that recognition software can be used to retrieve information on the type of content as well as the content itself (Lichtenstein, Plank, & Neumann, 2014). Media such as videos, art, and music require their own standards and approaches for efficiently indexing them. The metadata for each item varies from not only one another, but also from text documents and other recording types.
Evaluating Automatic Indexing Efficiency

When judging the performance of automatic indexing it is important to note that there is no single correct way to index an information object. Measurements, such as recall and precision are important as mentioned previously, as well as the time it takes to run the program. Some factors can be considered more important than others, which results in the need for weights to be attached to these measurements. It is also important to note that an evaluation may be meaningless in a contained test environment that does not represent realistic conditions (Golub et al., 2015). Real indexers can also be used to evaluate automatic indexing results, but it is important to use experienced indexers because some results "might confuse junior indexers who may not have sufficient training to distinguish between almost-correct and correct recommendations" (Névéol et al., 2009, p. 823). Automatic indexing, while useful, should only be used as a tool to aid professional indexers process information faster, not as a substitute for human indexers. This is particularly true due to the difference in quality between human indexers and automatic indexing at this moment in time.

Challenges of Automatic Indexing

The results for automatic indexing efforts remain mixed. Junger (2013) shows that automated indexing tested for the German National Library could not achieve a high enough precision and "overall results achieved for automated indexing are considered not satisfactory" (p. 107). However, when automatic indexing is used with the purpose of aiding human indexes, a "performance of 32% precision...can be considered a positive result" (Névéol et al., 2009, p.
Another issue preventing the use of some of these systems is the amount of computation required. While computers can process a large amount of information, the amount of data contained in an information object, and the growing number of these information objects that need to be indexed can still overwhelm systems of today's capacity. The methods used in building an automatic index will also be affected by the language of the information object and the language the index is to be created in. These efforts in automatic indexing are limited to specific types of information and conditions, so a fully independent system that is on par with a human indexer is still far away. Another challenge to automatic indexing does not lie in the technical challenges but in user perception. Despite the need for automatic indexing and achievements made thus far in the field, there is large amount of information professionals that are skeptical of the technology. Surveys show that despite libraries recognizing the importance of indexing and the lack of current resources, over 70% of them do not use any automatic indexing (Keller, 2015, p. 900). Over 60% of responders say that they believe automated systems will never achieve the same quality as human indexers (Keller, 2015, p. 901). Going forward, we should try to combine automatic indexing with human indexers in order to provide a tool that speeds up the process, without sacrificing the quality of indexes.

Automatic Abstracting

Automatic abstracting has many similarities to automatic indexing. Because abstracts are geared towards textual documents, the research tends to focus on semantics and natural-language processing. Methods that extract sentences from a document to form an abstract already exist. The challenges come from choosing which sentences are the best to select and building an abstract from those sentences that is useful, as well as semantically correct. One issue related to
this is sentence simplification. Finegan-Dollak and Radev (2015) note that "an extracted sentence might contain both essential and extraneous information" (p. 2437), which can occur in longer sentences. Another challenge for the future of automatic abstracting is multimedia objects. Whether the term abstract will be used, or this type of information will be considered a different type of summary, it is equally as important for users who need to quickly determine the contents and purpose of information objects.

When looking to the future of indexing and abstracting, it is clear that automatic processes are necessary to keep up with the vast amount of information that is constantly growing. Coyle (2008) notes that use of these tools by human indexers is more efficient and time saving. Fully automatic indexing and abstracting is not currently acceptable in most applications and may never be fully realized. However, these programs need to be regarded as important tools for the human indexer or abstractor to have at their disposal.

**Multimedia Indexing**

As previously mentioned, multimedia indexing can provide a challenge for the field of indexing. It has different metadata standards, meets different user needs, and it is searched differently by various user groups. A big trend within the field currently is image indexing, which has its own approaches and methodologies. Innovations in this indexing subfield is especially important due to the daily increase of photos uploaded to the web that are being shared worldwide.
Image Indexing

Image indexing is a concept that is easy to understand in theory, but is very different from standard text indexing, due to the nature of images. Image indexing includes attaining, editing, storing, and retrieving visual images (Cleveland & Cleveland, 2013). Photos are now being digitized by every type of organization for different reasons ranging from historical to educational. All of these photos are being stored in databases that need to be searched by the public who want the information stored in them to build knowledge on subjects they are researching. Indexers and librarians have been trying to index and catalog these photos from the beginning, to keep order and organization in these databases. While these indexes are necessary, there are some issues and considerations of image indexing that affect how images are searched and found in the digital indexes. Some of those issues are lack of a specialized classification system, issues with users being able to understand how to get the right information they are looking for, and semantic gap.

Standard text-based systems, while predominantly used for information retrieval, are too time consuming to incorporate for visual media (Goodrum, 2000). Some of the approaches to image indexing include methods such as: atomic, term, structural feature, and semantic feature indexes, as well as more traditional methods (Cleveland & Cleveland, 2013) & (Nordbotten, 2008). Atomic indexes are geared towards more Relational Database (RDB) components that use context and structural metadata (Nordbotten, 2008). Term indexes are based on terms taken from the metadata that describe the image (Nordbotten, 2008). Feature indexes are focused towards aspects of images. Structural feature indexes use “low-level” features that are automatically extracted. These terms describe the basic features of the image such as; shape, color, and texture.
Semantic feature indexes are based on high-level features that include objects and other semantic features (Nordbotten, 2008). The more traditional methods include description-based indexing, concept-based indexing, and context-based indexing (Cleveland & Cleveland, 2013). Description based indexing assumes that a human indexer will be entering the descriptive information. The fields included start with the artist or creator and go on to include more descriptive information about the image (Cleveland & Cleveland, 2013). This is not an easy task since the indexer is attempting to give text-based descriptions about non-text-based content. Concept-based image indexing refers to retrieval from text-based indexing of images that may employ keywords, subject headings, captions, or natural language text. On the other hand, content-based image indexing means that the search analyzes the contents of the image rather than the metadata such as keywords, tags, or descriptions associated with the image (Castelli, 2002). Each indexing method has pros and cons, but they each evaluate different criteria, which may mean that we need to develop a more unified indexing methodology that encompasses them.

Why Image Indexing Needs a Specialized Classification System

Usual indexing is text based and when these indexes are searched by the public, they will be using normal terms to search for the images. Librarians and indexers have tried applying general classification systems such as Dewey and LC subject headings with no success (Jörgensen, 2004). Every library or indexer has their own subject headings, which means that there is no cohesion when assigning headings to the images. When a person is searching for an image in a database or even a basic search engine, they will not have a list of possible subject headings to work from. Instead they will type in a few keywords to locate images that fit their parameters.
Librarians and indexers need to develop a common language or thesaurus not only to make the indexing process easier for those who are creating the database, but to make the database easier to search by the public.

Attempts to provide general systems for image indexing include the Getty's Art and Architecture Thesaurus (AAT), which consists of over 120,000 terms for the description of art, art history, architecture, and other cultural objects, and the Library of Congress Thesaurus of Graphic Materials (LCTGM). The AAT currently provides access to thirty-three hierarchical categories of image description using seven broad facets (Associated Concepts, Physical Attributes, Styles and Periods, Agents, Activities, Materials, and Objects). Assignment of terms to describe images is not solved entirely by the use of controlled vocabularies or classification scheme however (Goodrum, 2000, p.63). Term assignment and thesauri are valuable tools for increasing image indexing efficiency, but they are not the only considerations that indexers need to make. When it comes to indexing images that are pieces of art, photos of sculptures, and photos of architecture, factors other than color and images inside of the photo come into play. Descriptions of texture, color saturation, shape, and spatial relationships may be basic information for the indexer or historian analyzing the image to include as subject headings, but for the public using the database to find the images, those headings are not going to be the first keywords they are going to use.

When the public proposes a query in the search box of the database, it may not fit the exact thesaurus the indexer used while programming the database. According to Goodrum (2000), “The textual representation of images is problematic because images convey information relating to what is actually depicted in the image as well as what the image is about.” To fix this issue, a specific classification system with a dedicated vocabulary needs to be developed and used universally to make indexing and searching images easier.
This issue leads into a technical issue for the public. When they want to describe the possible texture of the image or the object in the image, the public is going to use basic terms such as rough, silky, or smooth. How are they supposed to know what a wavelet transform is? How are they supposed to know how to calculate image texture using co-occurrence matrices or laws texture energy measures? Both of these complex equations can be complicated for the novice indexer and librarian. The everyday public would not even think about analyzing an image using a complex equation to make a visual description of any image they are searching for. Sharpiro (2001) states that, “One negative aspect of the co-occurrence matrix is that the extracted features do not necessarily correspond to visual perception.” While indexers want to make sure that every image has as much possible descriptive information attached to it as possible, the end user will probably not know how to use the formulas to find the exact number for ripple or contrast and they may not even know what ripple is. This is not a natural way of thinking about which criteria to use as a descriptor, therefore it is not the best way to connect users to this particular type of information.

**Semantic Gap**

Semantic gap is a computational problem in image retrieval, making it another challenge for image indexing. When the metadata for the images is being analyzed, the end user is not included in this process (Zhao, 2002). If the image is of a classroom full of children, the metadata might include terms such as students, school, teacher, or classroom, but the end user might use the terms such as children, Ms. Greene, or even one of the students’ names depending on if it is a personal photo instead of a general stock image. User queries should be the driver but, “there has been a tendency for much image retrieval research to ignore the issue of user
queries and to concentrate on content-based techniques” (Hare, 2006). When dealing with content-based image indexing, the end user will not be using the usual metadata to search for the images, but rather the known objects in the images.

As technology continues to develop easier ways for people to access the internet and online databases, indexing images and other information will be searched for on laptops, desktops, or mobile devices. While indexers have their own vocabulary and classification system when creating indexes and databases, they need to remember that the end user is the one who will be accessing the information and the indexers need to take that into consideration when they are adding the metadata and terms to the thesaurus being used in the databases.

**Folksonomy**

Derived from the combination of the words folks and taxonomy, folksonomy is used to describe collaborative efforts in the creation and management of tags by users within a digital environment. The term folksonomy was coined in the early 2000s by Thomas Vander Wal in a discussion pertaining to information architecture (Mathes, 2004, p. 5). The environment of Web 2.0 fosters the idea of folksonomies because according to Kakali & Paptheodorou (2010), it “promotes user-centered technologies and infrastructures, which have spread to diverse information systems and communities, and which typically focus on the development of collaborative and interactive information services” (p. 192). User preferences and needs are unmistakable factors within folksonomies.
Elements of Folksonomies

There are two specific features accompanying a folksonomy. The first is the folksonomy’s ability to allow users to add tags to information. Once the tags are created the second feature appears. Navigational links are produced from the tags, thus somewhat allowing a way for the uploaded information to be organized (Cleveland & Cleveland, 2013, p. 80). Aside from the tags created by the user, folksonomies are unique with the elimination of relationships between terms. Without established relationships, terms exist in a flat space, lacking hierarchy. In a traditional setting, not within a folksonomy, “information resources are described, organized thematically, and classified either by experts, or by their creators” (Kakali & Paptheodorou, 2010, p. 192). Instead of traditional hierarchies, folksonomies evolve around the products of user-generated content (Cleveland & Cleveland, 2013). Additional terms relating to folksonomy include social indexing and collaborative tagging.

Other elements of a folksonomy include the usage of uncontrolled keywords and common vocabularies. The words developed from uncontrolled vocabulary are used to create tags that provide enhanced subject description. Tags are defined as “any word that defines a relationship between the online resources and the concept of the user’s mind” (Kakali & Paptheodorou, 2010, p. 192). Collaborative tagging systems promote the usage of folksonomies. These systems allow users to create, upload, and share content using tags they choose. This set of tags forms the unorganized system known as a folksonomy (Movahedian, 2014). Examples of tagging systems are Flickr, Del.icio.us, and CiteULike. The rise of social bookmarking and photograph captioning have contributed to the development of the foundational concepts of folksonomies in Web 2.0. Folksonomies in Web 2.0 and into 3.0 demonstrate the movement away from older,
traditional indexing practices. The existence of folksonomies represents a shift away from terms and descriptors adopted by professionals to terms chosen directly by users that better reflect how they perceive these images (Mathes, 2004). As a result, users act have the agency to act as indexers and influence the overall retrieval of an object. Folksonomies do not replace the importance of information control, but they do play a role in highlighting work in academic work and aid in communication (Woolwine & Ferguson, 2011). Social tagging is in full effect through projects such as the Steve Collaboration, Penn Museum, and the Library Thing. All collections emphasize the user’s experience and call for a collaboration with users and information professionals.

Challenges of Folksonomies

Ultimately, folksonomy represents a double-edged sword for information professionals. The best and the worst aspects of information organization has been highlighted with the creation of folksonomies. The uncontrolled nature of folksonomies is fundamentally chaotic, suffers from problems of imprecision and ambiguity that well-developed controlled vocabularies and name authorities effectively ameliorate. Conversely, systems employing free-form tagging that are encouraging users to organize information in their own ways are supremely responsive to user needs and vocabularies, and involve the users of information actively in the organizational system (Mathes, 2004, p. 18). The idea of folksonomies thrives within the current digital age, because “the power of folksonomy is connected to the act of aggregating, the power is people here. The term–significance relationship emerges by means of an implicit contract between the users” (Kakali & Papiottheodorou, 2010, p. 192). While the existence of folksonomies is now inevitable in Web 2.0, there are several issues and consideration that still must be observed.
Folksonomies lack language control because they are open platforms. Per Porter (2011), “the synonyms, near-synonyms, homonyms, homographs, and lexical anomalies present in many user tags severely limit retrieval precision and hamper the location of similar or related resources” (p. 251). Inconsistent tagging and popular tagging influence high recall redundancy. Use of different, yet related, vocabulary from various sources retrieve related topics, but inapplicable results. According to Porter (2011) the issues that stem from misspellings, compound word groupings, inconsistencies with singular and plural forms, and the utilization of symbols negatively affect the “searchability” and value of folksonomies. The use of cross dialect of languages creates issues in folksonomy as well, because words that have similar meaning in different languages are not clarified by contributors, creating limited, irrelevant or incomprehensible search results. Basic level variation is an issue in folksonomy which occurs when multiple users describe the same item at a different specificity (Porter, 2011). Different users could describe a single object as a dog, hound and canine, but each description refers to the object at a level related to their experience with it.

Ambiguity is another issue that arises with the use of folksonomies. Guy (2006) states that, “the tagging terms used in those systems are imprecise. It is the users of a folksonomy system who add the tags, which means that the tags are often ambiguous, overly personalized and inexact” (Folksonomic Flaw section, para 1). Many folksonomy sites only provide single word metadata, which creates useless compound terms instead of single terms. Single terms are more effective for a search (Guy, 2006). Street jargon and even improper slander are used by friends on folksonomies, because there are no system administrators to regulate the folksonomies, leaving a disarray of tag terms. Not all of the terms are as effective for searching when compared to retrieval results attained when using a controlled vocabulary. Personalized tags in folksonomies
benefit the individual who initiated the tags, because it helps them allocate information, but they are not necessarily helpful to the general user who searches with controlled and academically correct vocabulary (Guy, 2006). Indexers and other information professionals, when employing a folksonomic system, need to consider the terms that users might select and decide which information would be best suited for that kind of labeling.

Advantages of Folksonomies

Folksonomies are simplifying the process of tagging for users to understand and implement with experience in indexing or classification. Vocabulary in folksonomies are a reflection of the user’s personal language and are accommodating each user’s needs, because the user can add or remove tags. Tags reflect both common content and personal content which enables users to browse or discover new content. They also represent the user’s conceptual model without cultural bias, in addition to providing a platform for creation of communities since users who apply the same tag may have a similar interest; folksonomies are versatile because users can assign any number or combination of tags to represent a concept (Revolvy, 2017). According to Porter (2011), “the effectiveness of folksonomies is reliant on the principle of collective intelligence, often referred to as the ‘hive mind’ or ‘The Wisdom of Crowds’, which posits that large, diverse groups of people are more capable of solving complex problems than individuals” (p. 249). Collective contributions of users allow folksonomies to cover larger grounds of vocabulary that a single editor of a controlled vocabulary would likely exclude or miss. The greater the number of users, the more dynamic the vocabulary content in a folksonomy. The contributions of multiple users with different interpretations of a resource allows users to retain new perspectives from other users (Porter, 2011). The variety of contributions from different
users in a folksonomy breaks the gap between search access points for users from different
cultures, genders or age. Folksonomies adapt to evolving cultural norms and societal climate
well because of the diversified contributions of different users (Porter, 2011). Social tagging has
a collective action component that makes it fascinating. Contributors enjoy the camaraderie they
feel with other users in their group, and seeing how others will come to a common consensus on
relevant terms. Shared cultural understanding also creates cohesive communities of users who
tag terms common to the cultural trends relatable to these communities of users (Boyd, 2005).
According to Boyd (2005), “The “in the know” groups using these services are very homogenous
and often have shared values and thus offers valuable related links. In tagging, quality is not just
about ‘accuracy’, but about what cultural assumptions dominate” (para 3). Users can become
part of the process of indexing, aiding in the tackling of surplus information processing, with
information sources that may or may not adhere to standard indexing practices. They also
provide new viewpoints than indexers, which may lead to more effective potential for future
retrieval and overall organization.

Social Tagging for Image Indexing and Retrieval

The advantages that folksonomies provide could be a potential solution to problems regarding
image indexing. Allowing users to tag images on social networking websites like Flickr, or even
on art museum websites, can benefit both users and professionals by including new descriptors
that were previously not considered. These methods are also gaining interest in the art museums
community as a means of better engaging the public (Trant, 2007). The use of social tagging for
image indexing has grown in popularity, and with this growth, new studies have been conducted
that analyze how the use of folksonomies can be effectively implemented in social image sites
and museum settings alike (Cairns, 2013; Chae & Kim, 2011; Choi, 2017; Jorgensen, Stvilia, & Wu, 2014; Rorissa, 2010; Trant, 2007). Rorissa (2010) studied the use of tags and index terms on the social image site Flickr. The focus of the study was a general-purpose collection of images uploaded and tagged during the first week of June 2007. Her findings reflected previous studies she had referenced, remarking that, though social tagging was useful, it should be done alongside traditional indexing. This is due to issues regarding how ambiguous and synonymous certain terms can be. Rorissa (2010) also explained the importance of context regarding the use of specific tags. These tags reflect the “context and the perspective of the person doing the tagging” (Rorissa, 2010, p. 2239). This results in more valuable content than index terms, which tend to lack context (Rorissa, 2010). Other studies have also shown the way folksonomies may prove to be beneficial for image indexing in the long run, particularly with museums.

One study into the practical use and implementation of social tagging in a museum was that published by Trant in 2007. Trant’s (2007) study followed Proof of Concept tests performed at the Metropolitan Museum of Art between fall 2004 and fall 2005. This test analyzed how well untrained users could develop image descriptions and access points. This study, unlike many others, served to understand the value of social tagging in establishing new or different terms for image descriptions than those that were previously used. These tests did not include general users, though. Most of the participants were library assistants or art history community members who had no prior cataloging training. Still, the subjects were able to provide more than 3,000 new terms, a large percentage of which could be considered useful as access points.

Cairns (2013) discussed the similar setting of museums, but brought up the concept of expertise. Her review followed the results of various museum studies incorporating the use of social tagging. The shift to Web 2.0 has created a new dynamic as to what expertise really means. What
was once regarded as knowledge of professional experts has shifted to the knowledge found on social networks. The inclusion of social tagging is changing the way formal institutions interact with and perceive “outsiders”. This is further emphasized by Choi (2017). Her study was conducted on the use of tags collected from the Networked Infrastructure for Nineteenth-Century Electronic Scholarship (NINES) of digital resources. While NINES does include some instructions regarding the format of tags, they are not always followed. The purpose of this study was not only to observe how images are tagged, but also how those tags relate to the metadata of the images already being used. Choi (2017) determined that social tags are important as they demonstrate what terms attract users. The results also show that, while the majority of terms are already present in the metadata, there is still a large percentage of new terms that can be used for better access for users.

Most of the literature surrounding this topic comes to the conclusion that social tagging has a myriad of benefits. This is especially the case regarding the semantic gap issue that was previously discussed (Chae & Kim, 2011; Trant, 2007). Trant (2007) pointed out that the implementation of social tagging exercises could create better “teaching moments” between staff and art museum patrons. This could create a better relationship between art museums and the public. Social tagging can also assist information professionals in other matters, as evidenced by the Lindstaedt, Morgenzer, Sorschag, Pammer, and Thallinger (2008) study which used behavior seen in social tagging as a means of creating better autoclassification and annotation systems.

With a better understanding of both concepts, it is easier to see why there has been a trend in combining the two, but this must be done with some form of control on the part of professionals who operate and maintain these indexes. One such approach, suggested by Chae and Kim (2011), is creating a faceted structure to better organize user tags. This study found that creating
facets for users to categorize their tags was useful, but could also be seen as taking away users’ freedom of choosing terms and categories they found more appropriate (Chae & Kim, 2011). A similar approach to this method would be creating a controlled vocabulary for users to choose from. This was discussed by Jorgenses, Stvilia, and Wu (2014). They found that while users can sometimes generate new and useful terms, they still appreciate the involvement of controlled vocabularies as reference points. This is corroborated by the Lin, Trattner, Brusilovsky, and He (2014) article that asserts the usefulness of present image descriptions in aiding social tagging. Based on this study, users were more likely to include better tags when a description of the image was provided. Another issue, discussed by Chae, Park, Park, Shi, and Yeo (2015), was the lack of organization of all of these tags. Their solution included creating automatically-generated clusters to organize these tags. While the solution proved useful, they did remark that human intervention is still necessary for grouping tags, but human intervention is not always viable when these collections number in the thousands.

**E-Book Indexing**

**Current State of E-book Indexing**

Many people today go online for many reasons, the main of which is to read up on new information or new books that are being published. Some go to their devices to read a book, instead of a hard-back, physical one from a library or that they purchased directly within a store. The future of e-book indexing is still being developed and it often depends on whether publishers want their books to be published in print and e-book versions, or if they want just one way over another. A published article by Schoun (2011) states what 3 people who work in the indexing
services think about the importance and personal views about indexing with e-books. When it comes to the market, e-books make up 15 percent of the total market with their share growing at a rapid pace (Schoun, 2011). Unfortunately, there are minimal software programs that would lend an active and useable index for e-books, since the field of e-book indexing is still new. Indexers currently do not have much familiarization with indexing e-books, since the development of e-books is new, making indexing software not very capable at working with e-books presently. Since the e-book indexing specialty is still new, the jobs are varied since they do not provide all the same inputs or require the same outputs (Browne, 2016). As this trend of digital reading continues, e-book indexes will need to be created and perfected, so they are as effective as those found in the back of physical books.

E-book Indexing Tips

Some indexers have been dabbling into indexing e-books over the last several years, so they have some knowledge on what is required to make an index for an e-book. A woman named Nancy Humphreys decided to share her useful tips on how to create an index of an e-book (2012). An important thing to know is that e-book indexing is different from print book indexing. The first tip that she provides is that the index should only have one column due to the fact that devices today are small, such as the Nook or the Kindle. The second and third tips are to not continue headings, since they contain many subheadings. Indexers should keep the headings simple and subheadings short. The fourth tip is to place the following note, “This e-book contains an index. To go to the index…,” at the top of the contents page (Humphreys, 2012). The fifth and final tip is to make sure you use cross-references in your index. It is important to implement these tips passed on to future indexers who may come across having to index an e-book. Technology is on
the rise and so is the use of e-books among many users, so experience in developing e-indexes may also be on the rise in the near future, if not already.

EPUB

There is a company called EPUB, which is internationally based, known for their e-books. EPUB coincides with the Indexes Working Group (IWG), which has developed a draft standard for e-book indexes (Brown and Wright, 2013). There are multiple features that are listed in the proposed EPUB standard and these features will help make the index look polished and presentable for the little time it has been in development. Some features that are included are locator ranges, cross-references, locator targets, locator link text, indexing for handheld screens, identification of indexes within the package document, and navigation through indexes (Brown and Wright, 2013). Recently, there have been propositions on future developments for EPUB indexes that will have improve usability and necessity for e-book indexes. Even with a few software developments that have been created to start the progression of e-book indexing, it is not enough to get the whole task completed. A few propositions that have been presented to positively influence e-indexing include: text-to-speech synthesis, semantic markup (explaining what the subject is instead of showing it), dynamic lookups, and more structure (Brown and Wright, 2013). These are just some of the newer ideas that should be used to help improve the need for a form of software to index e-books.
Digital Libraries and Indexing

As materials have become more available on the internet and the internet has grown, libraries have begun to extend from physical to digital space. Although digital libraries have been present as a concept since the 1930-40s and libraries began achieving automation via records in the 1960s, they became more prominent in the 1990s due to the first Digital Library Initiative in 1991 (Calhoun, p. 1). Just as physical libraries have needed and supplied indexes, digital libraries also need them for organization and supply them to users to aid with navigation. Instead of simply cataloging materials, digital libraries need to make sure that items within their collections are retrievable. This requires indexing to be done on a digital level, similar to what is done in website indexing. Digital libraries need search engines and these engines need to be able to effectively locate and retrieve information. Indexing is necessary in digital libraries so that it can be retrieved for users. According to Nazim (2009), “automatic indexing techniques are used to index the content” (p. 32). Nazim discusses some of the models that may be used, such as vector-space and probabilistic information retrieval and classification or inquiry systems. Automatic indexing, as mentioned above, has limitations such as technical capabilities in terms of precision of content recall and negative perceptions from users and professionals. Digital libraries have also been using various access tools while research within the last decade has included user interfaces as they relate to digital collections (Nazim, p. 33). As digital libraries continue to develop and expand, technologies used to create and maintain them will need to be developed, improved, or continually maintained to provide a satisfactory product that fits user needs.
Challenges of Digital Library Indexing

The challenges of indexes in digital libraries are very similar to the challenges that are involved with search engines. These include database size and content types (Lewandowski, 2005). The challenges for digital libraries were previously outlined by the Library of Congress for the second Digital Library Initiative which was announced in 1998. Included in the challenges listed were questions regarding resource design, such as “Can automated tools permit the incorporation of factual knowledge…into descriptive information, indexing, or search and retrieval systems?” (Library of Congress, 1998). It also referred to the potential for protocols for unified searching to be similar to approaches towards distributed indexing in terms of interoperability. Another challenge includes name authority. Bennett and Williams (2006) discuss the difficulties that arise when author names are not consistent within databases. This can cause issues with information retrieval, since there is the potential for users to be unable to find the information that they are looking for should variations of author names not be properly indexed to refer to the other variations. Fetters (2013) discusses the challenges of website indexing, which arguably, digital libraries can fit into, since they are located online. She also discusses the process included in website indexing, which includes consideration of vocabulary and standard indexing entry procedure, but then adds in file names and links (p.102). While this may be similar to regular indexing, additional responsibilities are added since links need to be checked for accuracy and they need to be determined as relative or fixed. This refers to if the files will be in one place or in multiple folders (p. 102). Information locations are imperative in this indexing method because they will affect the ability for information to be retrieved and the potential longevity of the materials’ retrieval.
Professional indexers and indexing tool developers will need to take all of the above challenges into consideration to progress satisfactorily with the pace being set by digital libraries and the amount of information that is increasingly becoming available digitally. This is especially true when we consider that not all digital libraries present information in the same way. Therefore, each digital library may require a different type of index or may present their indexes differently to users, if at all. In some digital libraries, such as the Smithsonian Digital Library (2018), there can be multiple collections, which may or may not provide visible indexes for users to access and employ for navigation. Most collections are browsable by search using web indexes that are not visible to users; these indexes run in the background using metadata within the library system. Others provide a list of materials in the collection that serves as a visible index, such as the exhibition title index provided for the Library and archival exhibitions on the Web collection (Smithsonian Libraries, n.d.). This index features title alphabetization for the exhibitions and includes the institution that houses the exhibition, the location, and the subjects covered within the exhibition. The sections of the Modern African art: A basic reading list collection (Smithsonian Libraries, n.d.) are organized with indexes that are organized alphabetically by title, with an abstract or brief synopsis accompanying each. Some entries additionally include “see also” notes. While these may not be traditional indexes, such as those found in the back of published print texts, they reflect the way that libraries are also becoming more innovative and accessible.

Digital Library Indexing Approaches and Index Development

Over the last thirty years, many authors have suggested approaches to digital library collection management and digital indexing. These have included autonomous citation indexing (Lawrence,
Giles & Bollacker, 1999), keyphrase indexes (Gutwin, Paynter, Witten, NevillManning, & Frank, 1999), and web-based citation indexing (Liu & Cabrera, 2008). The type of approach used for developing indexes is evaluated based on the type of information being indexed. For example, citation indexing is used for scholarly materials, predominantly by academic libraries and research organizations. Citation indexing according to Gutwin et al. (1999) involves the use of citation tracking tools (p. 79). This type of indexing has been proven to be an efficient information seeking technique for some user queries. Citation indexes are still employed by university libraries and academic databases, despite Gutwin et al. presenting research on them back in 1999. They mention Google Scholar and EBSCO, which have both maintained longevity and wide use after at least a decade has passed since their inception and the study’s publication. The longevity of these digital libraries/databases and techniques maintain that indexing and abstracting move consistently with technological improvements, while keeping the general approach largely the same. Although there are more users employing Google Scholar and EBSCO, as well as faster computers for them to be used on, there is still a need to index information effectively. While the basic principles have remained, indexing has grown to include more formats, styles, and tool variety. The future of indexing and abstracting will continue requiring the management of content online in different contexts, due to an increase in available materials and the increase in digital storage capabilities. Software and digital techniques should continually be developed, evaluated, and improved so that user needs can still be met. User needs, digital library developments, and indexing techniques will need to be routinely assessed as time passes to maintain effective information delivery.
Conclusion

Although technology has changed dramatically over the last fifty years, the need for index and abstract creation and technique development has not. Techniques, models, approaches, and information access have been changed by the Internet, but information is in high demand and has become more widespread. It appears that even though most folksonomies and social tagging techniques can be unstructured and disorganized, a healthy medium can possibly be reached. Keeping the hierarchy done through traditional indexing is vital, but including new access points discovered through folksonomies can expand the original outreach and inclusion of these image collections. This also gives users a more interactive role in systems that were having issues with excluding the general public. Users are at the forefront of index and abstract creation, particularly when it comes to information organization and retrieval.
Reference


for Information Science and Technology, 67(10), 2437-2453. doi:10.1002/asi.23576


look at the problem of the semantic gap in image retrieval. In Multimedia Content
Analysis, Management, and Retrieval 2006 (Vol. 6073, p. 607309). International Society
for Optics and Photonics.


Hwang, S., Yang, W., & Ting, K. (2010). Automatic index construction for multimedia digital
doi:10.1016/j.ipm.2009.10.006

images across diverse domains. *Proceedings of the Association for Information Science
and Technology, 41*(1), 287-293.

Jorgensen, C., Stvilia, B., & Wu, S. (2014). Assessing the relationship among tag syntax,
semantics, and perceived usefulness. *Journal of the Association for Information Science

Junger, U. (2013). Can indexing be automated? The example of the Deutsche
doi:10.1080/01639374.2013.854127


Revisiting Indexing and Abstracting in the Digital Era


