# Mining Domain-Specific Thesauri from Wikipedia: A case study

David Milne, Olena Medelyan and Ian H. Witten

*Department of Computer Science, University of Waikato*
*{dnk2, olena, ihw}@cs.waikato.ac.nz*

## Abstract

*Domain-specific thesauri are high-cost, high-maintenance, high-value knowledge structures. We show how the classic thesaurus structure of terms and links can be mined automatically from Wikipedia. In a comparison with a professional thesaurus for agriculture we find that Wikipedia contains a substantial proportion of its concepts and semantic relations; furthermore it has impressive coverage of contemporary documents in the domain. Thesauri derived using our techniques capitalize on existing public efforts and tend to reflect contemporary language usage better than their costly, painstakingly-constructed manual counterparts.*

## 1. Introduction

Lack of electronically encoded semantic knowledge is a major obstacle in natural language applications of computers. General lexical databases such as WordNet provide limited coverage of restricted domains; domain-specific thesauri are rarely available for a given field. It is hard to keep manually-maintained thesauri up to date in rapidly developing areas such as entertainment or technology.

Automatically constructed thesauri offer a potential solution. They are usually built by analyzing large document collections, employing statistical methods to identify concepts and semantic relations. However, the complexity of natural language and the primitive state of language technology means that such thesauri are inferior to manual ones in terms of accuracy and conciseness [3].

An alternative approach is to exploit collaborative folksonomies, a recent burgeoning web phenomenon. These provide a medium in which speakers of any language define, describe and discuss topics of contemporary relevance. The resulting information is freely available, electronically encoded and conveniently presented. Wikipedia is a classic example whose immense potential is just beginning to be explored scientifically. Previous work has used part of its structure as a general thesaurus [10]. The present paper extends this by using the entirety of Wikipedia, and shows how this can be intersected with document collections to provide comprehensive, detailed corpora-specific thesauri.

We present a case study that uses Agrovoc, a manually-created professional thesaurus in the domain of agriculture, as the gold standard. We compare Wikipedia articles and links to the terms and semantic relations encoded in Agrovoc. We also analyze its coverage of terms that occur in a sample document collection in the domain, and compare this with Agrovoc's coverage.

## 2. Thesauri

A thesaurus is a map of semantic relations between words and phrases. Terms represent concepts; relations between them encode the organization of knowledge. This property has been explored in information retrieval, where electronic thesauri serve as useful tools. They have been successfully exploited for content-based categorization of large document collections, yielding an improved ability to locate relevant parts and a more perspicuous representation of search results [2].

When retrieving information from a particular document corpus, an ideal thesaurus would be crafted to reflect its content. Manually constructing domain-specific thesauri is an arduous and demanding art that requires substantial investment of time by experts in the domain. Consequently thesauri used for practical information retrieval rarely match the domain of the documents. To make matters worse, collections evolve whereas thesauri remain static—they are as costly to maintain as they are to create. And because of the intellectual investment they represent, they are rarely made publicly available.

Deriving thesauri automatically from text is an interesting research challenge [3]. The resulting structures are far cheaper to produce and maintain than their hand-crafted counterparts and more closely matched to the document content. However they do not compare in accuracy and conciseness. Although useful for many information processing and retrieval tasks, they cannot yet compete with manually constructed thesauri.

How can you obtain a thesaurus to support a library of documents in a particular domain? Manual construction is prohibitively expensive; automatic generation is woefully inaccurate. General thesauri do not incorporate the specialist terminology that pervades our professions, nor can they keep pace with the deluge of new topics and concepts that arrive each day. Yet a contemporary resource that incorporates expertise in all fields of human endeavour already exists: the widely known Wikipedia.

## 3. Wikipedia

Wikipedia was launched in 2001 with the goal of building free encyclopedias in all languages. Today it outstrips all

other encyclopedias in size and coverage, and is one of the most visited sites on the web. Out of more than three million articles in 125 different languages, one-third are in English, yielding an encyclopedia almost ten times as big as the *Encyclopedia Britannica*, its closest rival. Wikipedia is also controversial; we return to this in Section 6.

Wikipedia's success is due to its editing policy. By using a collaborative *wiki* environment it turns the entire world into a panel of experts, authors and reviewers [6]. Anyone who wants to make knowledge available to the public can contribute an article. Anyone who encounters an article is able to correct errors, augment its scope, or compensate for bias.

There are many similarities between the structure of traditional thesauri and the ways in which Wikipedia organizes its content.

## 3.1 Wikipedia as a thesaurus

Our strategy is to use Wikipedia as a source of manually defined terms and relations; the building blocks of thesauri. Although never intended to be used in this way, it seems well suited to the task. Each article describes a single concept; its title is a succinct, well-formed phrase that resembles a term in a conventional thesaurus—and we treat it as such. Hyperlinks between articles capture many of the same semantic relations as defined in the international standard for thesauri (ISO 2788):

a) The **equivalence** relation connects one or more terms to a single preferred term (or descriptor), if they are synonymous. It is denoted by USE, with the inverse form with USE FOR.
b) The **hierarchical** relation occurs between more general and more specific terms, denoted by BT (broader term) and NT (narrower term).
c) The **associative** relation stands of any other kind of semantic relation and is denoted by RT (related term).

From Wikipedia's structure, links corresponding to each relation can be identified as described below.

**3.1.1 Synonymy and polysemy.** Thesauri serve as controlled vocabularies that bridge the variety of idiolects and terminology present in a document collection. Each topic is named by a "preferred term" to which alternative expressions are linked via the USE relation. Likewise Wikipedia ensures that there is a single article for each concept by using "redirects" to link equivalent terms to a preferred one, namely the article's title. It copes with capitalization and spelling variations, abbreviations, synonyms, colloquialisms, and scientific terms. The top left of Figure 1 shows four redirects for *library*: the plural *libraries*, the common misspelling *libary*, the technical term *bibliotheca*, and a common variant *reading room.*

Scope notes specifying the meaning of each thesaurus term help users disambiguate terms that relate to multiple concepts. Wikipedia provides disambiguation pages that present various possible meanings from which users select the intended article. The term library yields several options, including *library*, a collection of books, and *library (computer science)*, a collection of subprograms used to develop software. The articles themselves serve as detailed scope notes—they fully describe the intended meaning of the term.

**3.1.2 Hierarchical relations.** The hierarchical organization of terms in a thesaurus is reflected in Wikipedia's categorization structure. Authors are encouraged to assign categories to their articles, and the categories themselves can be assigned to other more general categories. The right-hand side of Figure 1 shows a structure in Wikipedia that exemplifies these categorization principles. The article library has a corresponding category *libraries*, which contains several more specific subcategories and articles, such as *academic libraries* and *digital libraries*. Other categories, such as *libraries by country*, have no corresponding articles and serve only to organize the content. Both articles and categories can belong to more than one category. *Libraries* belongs to four: *buildings and structures*, *civil services*, *culture* and *library and information science*. Wikipedia's category structure does not form a simple tree-structured taxonomy but is a graph
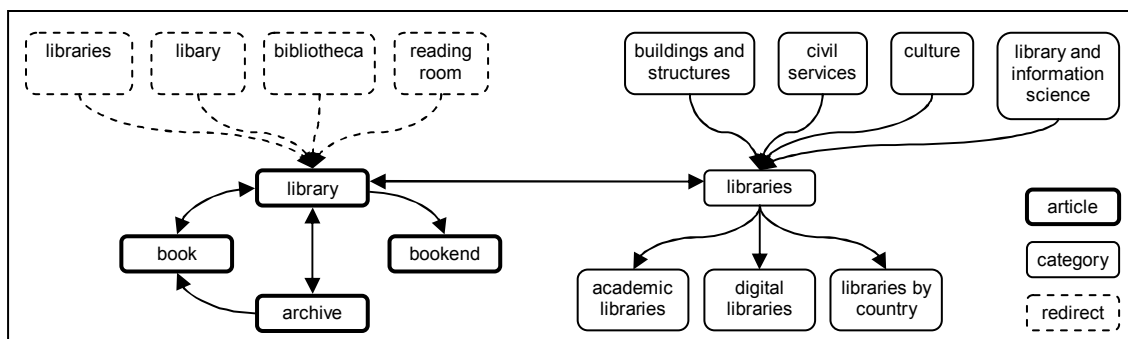


**Figure 1. Example structures from Wikipedia**

in which multiple organization schemes coexist.

**3.1.3 Associative relations.** Hyperlinks in Wikipedia express relatedness between articles. For example, the lower left of Figure 1 shows hyperlinks between the article library and those for book, archive, and bookend; some of these articles link back. Articles are peppered with such connections, which can be explored to mine the associative relations that are present in thesauri.

There are two problems: links often occur between articles that are only tenuously related, and there is no explicit typing of links. The first issue can be largely avoided by considering only mutual cross-links between articles—this discards the putative associative relation between *library* and *bookend* in Figure 1. As for the second, we must seek clues as to whether the relation is hierarchical or associative. If it already occurs within the category structure, it must be hierarchical. Statistical and lexical analysis can also be used (e.g. the *library* article has many more links and is therefore broader than *archive*).

## 3.2 Obtaining Wikipedia data

As an open source project, the entire content of Wikipedia is easily obtainable. It is available in the form of database dumps that are released sporadically, from several days to several weeks apart. The version used in this study was released on June 3, 2006. The full content and revision history at this point occupy 40 GB of compressed data. We consider only the link structure and basic statistics for articles, which consume 500 MB (compressed).

Table 1 breaks down the data. We identified over two million distinct terms (articles and redirections) that constitute the vocabulary of thesauri. These were organized into 120,000 categories with an average of two subcategories and 26 articles each. The articles themselves are highly inter-linked; each links to an average of 26 others.

## 4. Comparison of Wikipedia and Agrovoc

We aim to investigate the suitability of Wikipedia as a source of terms and relations from which thesauri can be

### Table 1. Content of Wikipedia

| | |
|---|---|
| **terms in Wikipedia** | **2,250,000** |
| articles | 1,110,000 |
| redirected terms | 1,020,000 |
| categories | 120,000 |
| **relations in Wikipedia** | **33, 060,000** |
| redirect to article | 1,020,000 |
| category to subcategory | 240,000 |
| category to article | 3,050,000 |
| article to article | 28,750,000 |

constructed. This section compares it with a manually created domain-specific thesaurus. We chose Agrovoc,[1] created and maintained by the UN Food and Agriculture Organization (FAO) to organize and provide efficient access to its document repository.[2] Table 2 shows pertinent statistics. Agrovoc is a substantial thesaurus, with approximately 28,000 terms describing topics relevant to the FAO and 54,000 relations between terms. The following subsections gives details of our analysis and presents results that summarize how well Wikipedia covers Agrovoc's terms and relations.

## 4.1 Comparison strategy

For effective comparison of terms, superficial differences—case, punctuation, plurality, stop words and word order—must be removed in order that equivalent terms match each other. For example, process recommendations, recommended processes and processing recommendations are superficially different phrases that all relate to the same key concept. To counter this, terms are case-folded, stripped of punctuation, and stemmed using the Porter stemmer [7]. Stopwords are removed and word order within each phrase is normalized alphabetically.

When comparing relations, differences in the terminology chosen to express the concepts should be ignored. Wikipedia and Agrovoc use different terms as descriptors. This is especially frequent for concepts that can be described either with a scientific term or an everyday expression: Wikipedia tends towards the latter. Figure 2 illustrates this by comparing the way in which the concepts harvesting and cultivation are related. While in Agrovoc these terms serve as descriptors, Wikipedia connects the articles on harvest and tillage to express the same relations. Through all possible permutations of redirects and USE relations we are able to overcome such differences and consider relations equivalent if they relate the same two concepts, regardless of the terms they use.

## 4.2 Coverage of terminology

Direct comparison of terminology, shown in Figure 3, reveals that Wikipedia covers approximately 50% of Agrovoc. The vast majority of terms found in the former but not the latter lie outside the domain of interest,

### Table 2. Content of Agrovoc

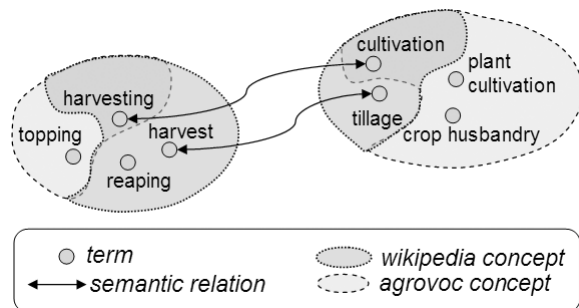| | |
|---|---|
| **terms in Agrovoc** | **28,000** |
| descriptors | 17,000 |
| non descriptors | 11,000 |
| **relations in Agrovoc** | **54,000** |
| USE to USE FOR | 11,000 |
| BT to NT | 16,000 |
| RT to RT | 27,000 |

**Figure 2. Comparing relations**

namely agriculture. More interesting are Agrovoc terms that are not covered by Wikipedia. Cursory examination indicates that these are generally scientific terms or highly specific multi-word phrases such as *margossa*, *bursaphelenchus* and *flow cytometry cells*. This is illustrated in Figure 4, in which terms in Agrovoc are stratified into groups according to whether they occur at general or specific levels of the thesaurus hierarchy. Wikipedia's coverage of Agrovoc degrades noticeably as concepts become more specific.

One third of the terms found in both structures are ambiguous according to Wikipedia; they match multiple articles. For example, the Agrovoc term *viruses* relates to separate articles for *biological viruses* and *computer viruses*. Agrovoc, being domain specific, does not consider multiple senses for terms.

## 4.3 Coverage and accuracy of relations

Next we examine Wikipedia's coverage of Agrovoc's relations, and evaluate our scheme for mapping Wikipedia's structural elements to particular semantic relations. First, for every pair of concepts related by Agrovoc that exist in both sources, we check whether a relation is present in Wikipedia. This was the case for 66% of Agrovoc relations. Some of the rest are encoded implicitly in Wikipedia. For example, Agrovoc's associative relation *gene transfer* → *gene fusion* is present because both terms are siblings under the Wikipedia category *genetics*. We did not consider these implicit

relations in this initial comparison.

Conversely, 94% of relations in Wikipedia are not present in Agrovoc. However, many of these are implicitly present through siblings in the BT/NT hierarchy or through chains of BT, NT or RT relations. Others do not belong in this thesaurus because they do not make sense within its context. For example, Wikipedia relates the ambiguous term *power* with *sociology*. Agrovoc is concerned with *electrical power* rather than *personal empowerment*, and therefore does not make the same connection. Sense disambiguation is needed to avoid these irrelevant relations. There are many other relations, such as *human* → *ape* and *immune system* → *lymphatic system* that are perfectly valid and relevant, yet do not appear in Agrovoc, even implicitly.

Figure 3a is based on Agrovoc's USE/USE-FOR relations and shows that Wikipedia covers synonymy particularly well: only 5% of relations are absent. Wikipedia's redirect structure is responsible for most of this, covering 75% of Agrovoc's synonymy relations. 20% of related term pairs that Agrovoc deems equivalent are encoded in Wikipedia through other links. Examples indicate that Wikipedia separates such pairs into distinct articles rather than treating them as synonyms, e.g. *aluminum foil* → *shrink film* and *spanish west africa* → *rio de oro*. Agrovoc judges these concepts to be "near enough" in that they do not require separate entries, whereas Wikipedia is more rigorous.

Figure 3b analyzes Agrovoc's hierarchical relations. Wikipedia covers 69% of them, but only 25% appeared in the category structure: the remaining 44% were found in redirects and hyperlinks between articles. The results could be improved by using implicit links. Hierarchical relations are transitive, meaning that *oceania* → *american samoa* is implied by the chain *oceania* → *oceanian countries* → *american samoa*. Coverage doubles when these implicit relations are considered. It is also possible to mine relations found elsewhere, but this would require additional analysis to identify the direction of the relation. For example, a hyperlink between two articles does not say which is broader and which is narrower. This information may be encoded textually (e.g. *South Africa*
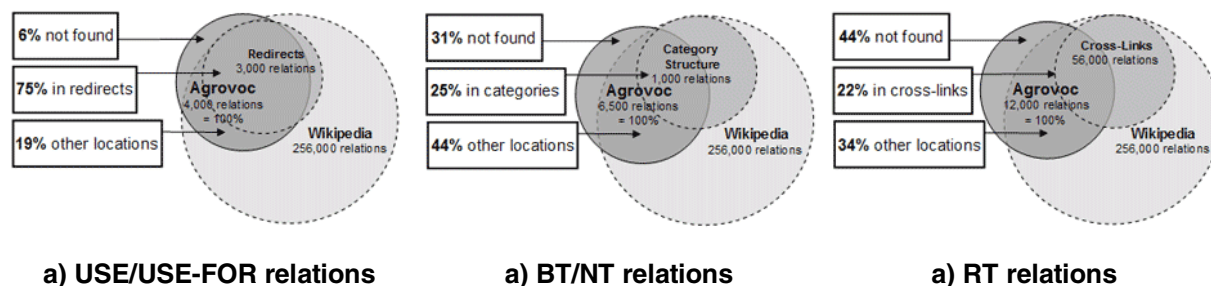


**a) USE/USE-FOR relations**    **a) BT/NT relations**    **a) RT relations**

**Figure 3. Wikipedia's coverage of Agrovoc relations**

is a lexical expansion of *Africa*) or statistically (e.g. *forestry* has many more links than *logging*).

A full 84% of the relations in Wikipedia's category structure are not present as hierarchical relations in Agrovoc. Many are implicitly encoded, while others are irrelevant to Agrovoc's domain. The remainder form a useful increase in connectivity.

Figure 3c depicts associative relations, of which Wikipedia covers 56%. Mutual links between articles were expected to match RT relations closely. However, only 22% were found in this way; the remaining 34% were found within one-way links or the category structure. Also, only 5% of mutual article links correspond to RT relations. Many describe relations that Agrovoc leaves implicit, e.g. all siblings are implicitly RTs. Other mismatches may be caused by inadequate sense disambiguation. As with hierarchical relations, extracting thesaurus-style RTs is a complex procedure that requires sense disambiguation and examination of other link locations in Wikipedia.

## 5. Analysis of corpus coverage

Next we investigate how well Wikipedia provides thesaurus support for a domain-specific document collection—that is, how well it covers the collection's terminology. Statistical comparison with a domain-specific thesaurus produced by human experts specifically for the domain reveals the striking benefits of Wikipedia's immense coverage and contemporary language.

We compared Wikipedia with Agrovoc on 780 agricultural documents taken from the FAO's document repository. All documents were full text (not abstracts) and had been professionally indexed with at least three Agrovoc terms. From each one we automatically extracted noun phrases using the OpenNLP tool for linguistic analysis. Table 3 shows salient statistics. There are over 700 times more noun phrases than index terms, which is not surprising; index terms represent only the main topics of a document, while the noun phrases it contains cover every concept mentioned in it.

### Table 3. The document corpus

| # of documents | 780 |
| --- | --- |
| Average length in words | 22,000 |
| # of distinct index terms | 1560 |
| # of distinct noun phrases | 1,133,000 |

We learned in Section 4.2 that Wikipedia covers only 50% of the terms in Agrovoc, despite being many times larger. However, many of the remaining Agrovoc terms are rarely used in practice. To assess this, we looked only at the 1560 distinct Agrovoc terms that were actually used by professional indexers to index documents in this corpus—which comprises precisely the kind of material

Agrovoc is intended to be used with. These terms form a small subset of Agrovoc (9.3%), but were manually chosen to be particularly relevant for the corpus.

Encouragingly, Wikipedia's coverage grows from 50% of the full Agrovoc to 72% of the terms actually used. Coverage is still incomplete, however: Wikipedia missed important terms such as *yield forecasting*, *sediment pollution* and *land economics*.
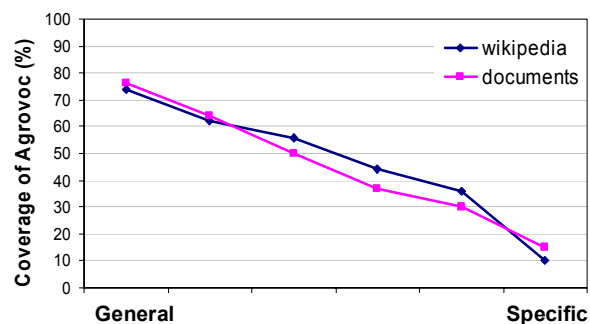


**Figure 4. Specificity of document terminology**

Index terms form a small sample of relevant Agrovoc entries. To gain a more detailed view, we examine the hierarchical distribution of Agrovoc terms that are used anywhere within the document set. As before, we stratify Agrovoc into levels of generality, and intersect these with document noun phrases. Comparing this overlap with that of Wikipedia (Figure 4) reveals a striking trend; both Wikipedia and the document collection cover less of the thesaurus as terms become more specific. Thus the additional detail Agrovoc offers over Wikipedia is predominantly irrelevant for this document set.

Figure 5 shows a three-way comparison between Agrovoc, Wikipedia, and the set of noun phrases extracted from the corpus. Most noun phrases are not found in either source, which probably merely indicates that most noun phrases are not suitable thesaurus terms, syntactically or semantically. The terms found in either structure, however, can be assumed to represent valid concepts mentioned in test documents.
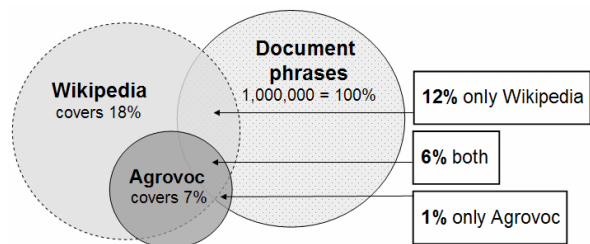


**Figure 5. Wikipedia and Agrovoc coverage of document terminology**

Wikipedia covers approximately three times as many document terms as Agrovoc. Many of these, such as *high school*, *aztec religion*, and *asean free trade area*,

probably lie outside Agrovoc's intended domain. They are, however, distinct concepts that are mentioned in the corpus and should be included in a corpus-specific thesaurus. We conclude that, at least in terms of term coverage, Wikipedia is substantially better suited to describing this document collection than Agrovoc.

## 6. Related work on Wikipedia

Wikipedia has recently been discovered as a vast source of semantic knowledge and a promising tool for natural language processing. NLP systems typically rely on painstakingly created lexical databases like WordNet. Wikipedia articles can easily be accurately matched to entries in these resources, and Wikipedia can be used to extend them [8]. Measures of semantic relatedness computed using Wikipedia are just as accurate as those from WordNet [9]. Both sets of measures performed equally well when applied to the standard linguistic task of co-reference resolution. Like our own research, this suggests that Wikipedia can be considered to be fully-fledged semantic resource in its own right. Bunescu and Pasca [1] apply it to the problem of named entity disambiguation, and obtain promising results.

Current techniques for extracting and using semantic knowledge from Wikipedia tend to consider the category structure as the only source of relations. We have found many useful relations elsewhere. The redirect structure seems to describe synonymy particularly well, and links between articles encode important semantic information. To our knowledge, the quality and utility of these relationships has not been investigated elsewhere.

## 7. Discussion

We have evaluated Wikipedia's quality as a semantic resource by examining the extent to which it replicates the high-quality domain-specific thesaurus Agrovoc, and comparing the extent to which both cover the vocabulary of a relevant document set. Comparisons of both terminology and relations yielded promising results.

While Wikipedia covers only 50% of Agrovoc's terminology, it tends to cover terms that are more likely to be used. Wikipedia covered the vocabulary of the specialized document corpus even better than Agrovoc, which was specifically designed to support it. Given the sheer breadth and size of Wikipedia (and its rate of expansion), it seems likely that similar coverage will be obtained for all but the most technical document sets.

Wikipedia covers most Agrovoc relations, and is a good source of semantic relations between terms. Its redirect structure represents a complete and accurate mapping of Agrovoc's synonyms. Hierarchical and associative relations are covered to a lesser extent and in a less organized fashion; the two types are intermingled with the category structure and hyperlinks between

articles. More work is required to separate these.

## 7.1 Applications

As a verified source of topics and semantic relations, Wikipedia has three main areas of application: improving access to documents, extending existing thesauri, and producing new thesauri.
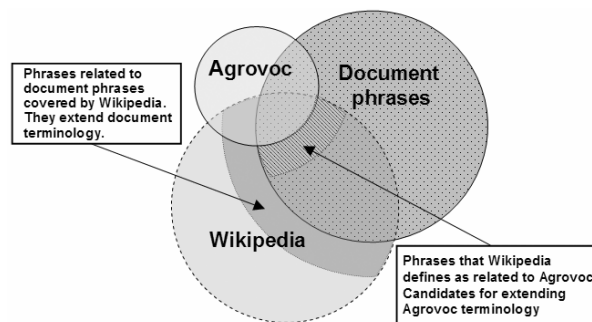


**Figure 6. Two applications of Wikipedia's topics and relations**

**Improving access to documents.** Users often require a bridge between their own vocabulary and that of the documents they seek. Wikipedia, which is produced by both experts and novices, can provide this. Figure 6 illustrates how the terminology of a particular corpus could be extended by including terms related to phrases in its documents. In our corpus users could access material on *salvelinus fontinalis* and *african trypanosomiasis* through Wikipedia terms such as *brook trout* and *sleeping sickness*, which do not appear in the documents verbatim.

**Extending existing thesauri.** Thesaurus maintainers could benefit from Wikipedia's broad and contemporary coverage. They could systematically extend the vocabulary by examining extra-thesaurus terms that relate to domain terms, and phrases from relevant documents, as Figure 6 shows. They could augment non-descriptors by mining Wikipedia's redirects. For example, *backbone* could be added to Agrovoc as a redirect for *spine*, *mainstream media* for *mass media*, and *M'sia* for *Malaysia*. Using cross-links and the category structure suggest new concepts such as *biochemicals*, *subsistence economy*, *natural abundance* and *money* for Agrovoc maintainers to consider. Furthermore, terms for which Wikipedia has corresponding articles in other languages could be used to enhance Agrovoc's multi-lingual features.

**Mining corpus-specific thesauri.** Wikipedia is a valuable thesaurus in its own right and not merely a means of improving existing ones. For our test collection it surpassed Agrovoc, a traditional thesaurus. If this holds for other collections and domains, one must question the need for domain specific thesauri at all: they merely approximate the topics that corpora are expected to discuss.

More exact matches can be obtained by intersecting document terminology with Wikipedia to produce truly corpus-specific thesauri—Wikisauri, if you will.

## 7.2 Concerns

The controversial nature of Wikipedia [4] raises definite concerns about using it as a thesaurus substitute. Although in principle its open editing policy renders it vulnerable to inaccuracy, we believe that in practice this will have little effect on extracted thesauri. They are unlikely to suffer from vandalism, self promotion, or large scale misinterpretation, because obvious errors are quickly detected and corrected within Wikipedia [5]. More subtle errors such as poorly worded statements and factual inaccuracies are restricted to the articles' prose, which does not affect derived thesauri.

One unavoidable drawback is that derived thesauri would be only available for areas that interest contributors. This is mitigated by Wikipedia's tendency to describe domains that traditional thesauri are hard pressed to cover, and by Wikipedia's continued exponential growth [10]. Of more concern is the bias toward more general topics. Most contributors are enthusiasts rather than professional experts, and thus produce broad but shallow coverage. Derived thesauri may therefore be of limited use for highly technical document collections.

A fundamental concern is that Wikisauri are based on a structure that was never intended to be used in this way. There could be profound differences between the way that articles are organized and the way that semantic terms are related. However, our work indicates that this is not the case; the theoretical similarities described in Section 3.1 and the quantitative ones uncovered by comparing with Agrovoc indicate that the two goals are compatible.

## 7.3 Advantages

Using Wikipedia as a platform for constructing thesauri has substantial advantages over traditional domain-specific thesaurus construction. The most obvious is cost. Another is currency: Wikisauri will evolve at a rapid pace. They excel in swiftly changing domains that capture the interest of contributors: current affairs, entertainment, and new technologies. The panels of professional indexers that construct traditional thesauri find it impossible to keep abreast of turbulent subject matter.

Another advantage is multilingualism. Wikipedia exists in 125 different languages. Although different versions are only lightly tethered to each other, in future they will be systematically mirrored across different languages. Versions for popular languages overlap significantly, and thus could produce multilingual thesauri.

Wikipedia is a source of useful statistics about terms and relations. Term occurrence and co-occurrence frequencies can be extracted from Wikipedia articles just as they can from conventional corpora. However, Wikipedia also reflects the relevance and popularity of concepts based on frequency of visits, number of article edits, and contributions to the discussion forums that accompany each article. Such statistics are attractive for the many information retrieval and natural language processing tasks to which Wikisauri could be applied.

## 8. Conclusions and Future Work

We have shown how to construct domain- and corpus-specific thesauri from Wikipedia. Comparing terms and semantic relations to those in a manually created thesaurus demonstrates excellent coverage of domain terminology, and of synonymy relations between terms. Wikipedia is a good source of hierarchical and associative relations, with scope for improvement in coverage and accuracy. Surprisingly, we have found that Wikipedia outperforms a professional thesaurus in supporting a domain-specific document collection.

Wikipedia, with its interwoven tapestry of articles in many languages, is a huge mine of information about words and concepts. Its exploitation is just beginning. Still unexplored are applications such as support for document retrieval, maintenance of existing thesauri and derived thesauri that match corpora for practically any domain. While there are serious concerns surrounding Wikipedia, these are for most part irrelevant for our purposes and are far outweighed by many advantages that traditional resources cannot possibly offer.

## 9. References

[1] Bunescu, R. and Paşca, M. "Using Encyclopedic Knowledge for Named Entity Disambiguation," *Proc. EACL,* 2006.

[2] Clark, P., Thompson, J., Holmbeck, H. and Duncan, L. "Exploiting a Thesaurus-Based Semantic Net for Knowledge-Based Search," *Proc. Innovative Applications of AI*, 2000.

[3] Curran, J. and Moens, M. "Improvements in automatic thesaurus extraction," *Proc. ACL Workshop on Unsupervised Lexical Acquisition*, 2002.

[4] Denning, P., Horning, J., Parnas, D., and Weinstein, L. "Wikipedia Risks," *Communications of the ACM* 48(12), 2005.

[5] Giles, J. "Internet encyclopedias go head to head," *Nature 138(15),* 2005.

[6] Leuf, B. and W. Cunningham, *The Wiki Way*, Addison Wesley Longman. 2001.

[7] Porter, M. "An algorithm for suffix stripping," *Program* 14(3), 1980.

[8] Ruiz-Casado, M., Alfonseca, E., Castells, P. "Automatic Assignment of Wikipedia Encyclopedic Entries to WordNet Synsets," *Proc. AWIC 2005*.

[9] Strube, M. and Ponzetto S.P. "WikiRelate! Computing Semantic Relatedness Using Wikipedia," *Proc. AAAI 2006*.

[10] Voss, J. "Measuring Wikipedia," *Proc. ISSI 2005*.