Improving Access to Large-scale Digital Libraries Through Semantic-enhanced Search and Disambiguation

Annika Hinze
Computer Science
University of Waikato
Hamilton, New Zealand
hinze@waikato.ac.nz

Craig Taube-Schock
Computer Science
University of Waikato
Hamilton, New Zealand
schock@waikato.ac.nz

David Bainbridge
Computer Science
University of Waikato
Hamilton, New Zealand
davidb@waikato.ac.nz

Rangi Matamua
Māori & Pacific Development
University of Waikato
Hamilton, New Zealand
rmatamua@waikato.ac.nz

J. Stephen Downie
Library and Information Science
University of Illinois
jdownie@illinois.edu

ABSTRACT

With 13,000,000 volumes comprising 4.5 billion pages of text, it is currently very difficult for scholars to locate relevant sets of documents that are useful in their research from the HathiTrust Digital Libary (HTDL) using traditional lexically-based retrieval techniques. Existing document search tools and document clustering approaches use purely lexical analysis, which cannot address the inherent ambiguity of natural language. A semantic search approach offers the potential to overcome the shortcoming of lexical search, but—even if an appropriate network of ontologies could be decided upon—it would require a full semantic markup of each document. In this paper, we present a conceptual design and report on the initial implementation of a new framework that affords the benefits of semantic search while minimizing the problems associated with applying existing semantic analysis at scale. Our approach avoids the need for complete semantic document markup using pre-existing ontologies by developing an automatically generated Concept-in-Context (*CiC*) network seeded by *a priori* analysis of Wikipedia texts and identification of semantic metadata. Our Capisco system analyzes documents by the semantics and context of their content. The disambiguation of search queries is done interactively, to fully utilize the domain knowledge of the scholar. Our method achieves a form of semantic-*enhanced* search that simultaneously exploits the proven scale benefits provided by lexical indexing.

1. INTRODUCTION

A few decades ago, scholars of humanities would have had to travel to the Library of Congress and national archives to visually examine many of the documents of interest to their research. Digitized archives such as the one made available in the HathiTrust Digital Library allow scholars to perform these kinds of explorations online, and scholars now regularly analyze large sets of digitized documents. Access to the digitized document collections is available primarily by string-based search, through inverted indexes of both document full-texts and metadata. Within the HathiTrust corpus, the document full-texts have been obtained through optical character recognition (OCR), and the metadata drawn from library catalogues. Such text-based searching identifies documents purely according to lexicographical analysis.

Most research questions and areas of scholarly interest, however, can rarely be described by simple textual keywords; rather they are conceptually-based, such as searching for content on “Māori Astronomy", “Niuean / European encounters" and “mythology in Shakespeare". Taking the first example, a simple text search for the terms “Māori" and “astronomy" obviously cannot identify all documents that may be pertinent to the scholar. Current practice is for the scholar to identify marker terms that might occur in a document set, such as “tohunga ko-ko-rangi" (astronomers), as well as English or Māori names of star constellations. This approach is restrictive and of limited use because large sets of unrelated documents may be included in search results, *i.e.*, where the names of Māori stars are used in a different context with no connection to astronomy. For instance “Matariki" can refer to both the Pleiades star cluster and the Māori New Year. Relevant sources may remain undetected unless the right keyword is found.

Easy identification of appropriate keywords is further hindered when different languages are involved (English and Māori in our example) and when an area contains sources from diverse fields that do not share a common vocabulary. Further problems are introduced through the inherent ambiguity of natural language, *e.g.*, synonyms and homonyms. In all these cases, false negatives (*i.e.*, missed documents) and false positives (*i.e.*, unrelated documents that have to be manually identified and eliminated) have significant adverse effects on the scholar's research.

Focus of this paper. We propose to automatically analyze documents not purely by their text but rather by the semantics of their content and metadata. Our work differs from the concept of semantic search, which we discuss in detail in Section 3. In a first step, the semantic analysis entails text analysis of the whole corpus to generate a specific knowledge structure we have termed a Concept-in-Context (CiC) network. Starting from a network of known concepts, we analyze which of these concepts appear in the document

JCDL'15, June 21–25, 2015, Knoxville, Tennessee, USA.

DOI: http://dx.doi.org/10.1145/2756406.2756920

No	Maori name	English name	Constellation	Cross ref	Prim. source	Type	Quote
003	Kōpū	Venus		Ngā Ihiihi o Kōpū. Matariki.	Andersen, J. (1962). Māori String Figures. Steele Roberts Limites. Wellington. P.5	2	"There are several figures with star names – Kōpū, Nga ihiihi o Kōpū, Matariki. In Polynesia too there are many figures representing stars."
171	Matariki	Pleiades	Taurus	Tohunga Rongoā Toutouwai Tāwera	Tregear, E. (1904). The Māori Race. A. D. Willis Printer and Publisher: Wanganui. P 22	6	"...the wizard doctor could revive a man if there were certain favourable conjunctions at the time; thus if the robin (toutouwai) was to sing for the first time just as the morning star (Tawera) was seen, if also the Pleiades (Matariki) were high in the sky, and the dying man had a shivering fit, then with all these auspicious signs occurring a certain invocation would bring back the departing soul."
274	Puanga	Rigel	Orion	Atutahi, Puanga Maruaonui, Matariki, Whakaahu Rangiāio	Te Tiupiri 1898-1900: Volume 1, No. 43: 4	1	"Waihoki i roto i tenei ra, ko Puanga kua ngaro ki roto o Te Maruaonui, ko tona kai whakahaere ko Matariki; ko Whakaahu kua ngaro ki roto o Te Rangiaio, tona kai whakahaere ko Tautoru; [...]."
Selection of further names : Rēhua (Antares), Takurua (Sirius), Whānui (Vega), Atutahi (Canopus), Maruaonui, Whakaahu (Castor/Pollux), Rangiāio, Mangaroa (Milky Way), Ranginui (Rigel Kentaurus), Tautoru (Orions Belt)							

Figure 1: Word list aiding repeated searches in Māori Astronomy

set, thus assigning meaning to each document. In a second step, the documents can be clustered set into proto work-sets based on the assigned concepts. In the final step, the scholar uses semantic-enhanced search on the data collection using concepts instead of text keywords.

The challenges encountered in this project span technical issues and those of interaction design. The technical challenges lie in the implementation of searching semantically in a text-based environment (*i.e.*, the available information is not encoded using semantic web markup) and in the development of a suitable Context-in-Context network. The challenges in interaction design relate to the development of work-flows and interfaces that are suitable for humanities scholars, who are assumed to be without extensive expertise in semantic querying (technical non-experts). In comparison, traditional interfaces for semantic search and knowledge manipulation are highly complex, and require prior experience and expertise in semantic query languages.

Contributions. This paper provides the new search and interaction concept of *semantic-enhanced search*. We introduce the conceptual design and implementation of our Capisco (Italian for "I understand") system, detail our process of manual disambiguation, and illustrate the effectiveness of our approach by two use cases (a complex one of Māori Astronomy, and a simpler one on that allows a detailed comparison between the different search approaches). We also discuss the insights gained from this project so far, and identify future challenges. Our approach is designed to scale well and has the benefits of semantic capabilities without complex query languages that are unsuitable for technical non-experts.

The remainder of the paper is structured as follows: Section 2 introduces our use case of scholarly search in Māori Astronomy, Section 3 discusses the background of our research, while Section 4 analyses related research. Sections 5 and 6 introduce our semantic-enhanced search and query disambiguation. Section 7 discuses a detailed example and two use cases to highlight the effectiveness of our approach. We conclude with a summary and discussion of future work.

2. USE CASE: Māori ASTRONOMY

Traditionally, Māori held great knowledge of astronomy and their studies of the night sky played an important role in everyday life. Much of this knowledge remains recorded in *Te Reo* (Māori: "the language") and sits within songs, prayers, proverbs and place names. The scholarly field Māori Astronomy is concerned with exploring traditional Māori Astronomy and understanding its practise, application and position within traditional Māori society. It is further concerned with understanding the language of traditional Māori Astronomy, its terms and use. As such, potentially relevant documents are not restricted to the scholarly publications but also other Māori document.

In a current project, scholars are exploring the significance of Māori Astronomy seeking to better understand its importance in traditional Māori society. An additional component to this study is the creation of tools to re-introduce Māori Astronomy into a modern world. One of the problematic issues has been the collection of relevant data. The current process is pain-staking and error-prone. The scholars have to take single Māori words or names related to astronomy, and search through various document collections for each. For example, the term *Matariki* may be first entered into Google search. The scholar then explores each link to judge how related it is to the study. Then the same term is entered into various databases such as the online Māori Newspaper collection *Niupepa*[1] or the Journal of Polynesian Society databases. Again the scholar analyzes each link. Finally, they check the books and papers from the library looking for references to the various star names and terms they have collected. The process is repetitive and tedious. Over the years the scholar has compiled a dataset of several hundred terms referring to star constellations and related concepts, often in both Māori and English language. This expedites the process and serves as a memory aid. The dataset has been extended with references to a primary literature source for each term and a quote linking the term and Māori Astronomy. Figure 1 shows a selection of the terms and their related information. The type information refers to the source of the reference quote (*e.g.*, journal, book, manuscript, online). For an exhaustive search on a new source potentially all terms would need to be checked.

From this simple example it is apparent that this process is very slow and more often than not produces information that is not relevant to their work. As described in an interview by the scholar, the current approach is experienced as being "extremely restrictive".

3. BACKGROUND

Before presenting our semantic-enhanced search, we first need to distinguish between different search approaches, which are often

confused due to unclear nomenclature. We briefly introduce our text corpus of the HathiTrust Digital Library (HTDL). We then discuss existing methods for search enhancement in Section 4.

3.1 Search approaches

Our approach is concerned with the search for (mainly text-based) documents, not knowledge searches that aim to infer answers to a question. Document search approaches can be distinguished by the types of documents and queries (text-based vs SPARQL/RDF), and indexing (see Figure 2).

Text-based (lexicographic) search. Classic text-based search uses keywords (*literals*) as query terms and the target documents are also (predominantly) text-based (*i.e.*, *literals*, see Figure 2, top). In the first phase (indexing), the documents are analyzed for keywords, which are then built into an index. During search, the index of keywords is used to identify matching documents (identified by *docID* in Figure 2).

Semantic search. Traditional semantic search uses a semantic query language (*e.g.*, SPARQL), while the target items are semantic web documents encoded, *e.g.*, using RDF. Each RDF document contains references to its concepts defined in one or more ontologies, and a number of literals. In the first phase (indexing), the documents are analyzed for concepts (as defined in the ontologies), which are then built into an index. Additionally, the ontologies used will be indexed according to the concepts they contain (*e.g.*, via a database). These details are shown in Figure 2, middle. During search, the SPARQL query (also containing concepts and literals) is executed on the ontology index (to identify relevant concepts) and document index (to identify documents referring to these concepts).

Semantic-enhanced search. Semantic-enhanced search as proposed in this paper uses text keywords and the target documents are also (predominantly) text-based (*i.e.*, literals in both queries and documents, see Figure 2, bottom). In this respect, semantic-enhanced search is similar to lexicographic search and can be used in the same settings. Internally an ontology or concept network is used to translate between keywords and concepts (disambiguation). In the first phase (indexing), the documents are analyzed for keywords indicating concepts (using the ontology or concept network), which are then build into an index of concepts. During the search phase, the user's keywords are also first translated into concepts, which are then used to lookup the index for matching documents.

Semantic-enhanced search as proposed here allows technical non-experts to use semantic technology when querying document corpora that do not provide semantic mark-up.

3.2 HathiTrust Digital Library

The HTDL is one such document corpus without semantic markup. It stores over 13,000,000 volumes comprising some 4,500,000,000 pages. Of these volumes, approximately one third are in the public domain [5]. The remainder is under copyright restrictions, which inhibits open access by scholars and researchers. The *non-consumptive research model* developed by the HathiTrust Research Center[1] aims to overcome these limitations by integrating analytic

[1] http://www.hathitrust.org/htrc

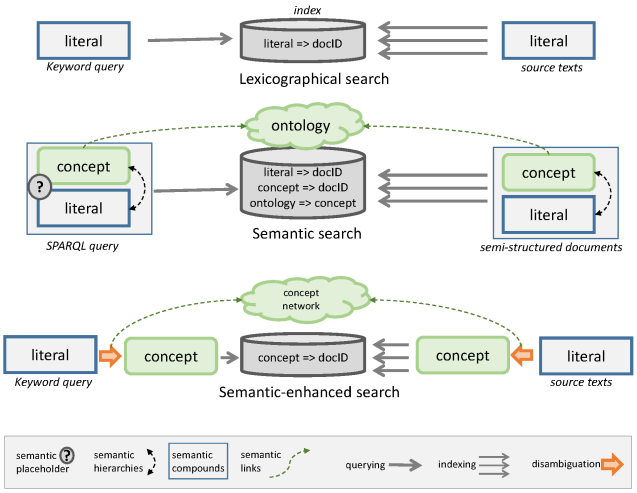

Figure 2: Comparison of search approaches

software into the data collection to allow for analysis of large collections while adhering to the restrictions of the copyright environment. The HTDL has great potential as source and platform for scholarly research. Researchers using digital libraries must be able to find, access and organize collections of materials. However, assisting scholars in building such collection subsets that they can then analyse for their research is challenging due to the scale and diversity of the corpus [10]. The documents in more than 100 languages are represented in the corpus, some of them are in multiple languages. Our research feeds into the project on Workset Creation for Scholarly Analysis, one of those goals is to enrich the available metadata for documents in the HTDL. A full-scale semantic analysis of each document, such as [2], is not feasible for reasons of scale and copyright.

4. TERMINOLOGY AND RELATED WORK

This section introduces the terminology used in this article. It also places our project in the context of related work.

4.1 Lexicographic search enhancement

Semantic-enhanced search requires a translation between text keywords and their respective concepts. A similar need for term clarification is typically encountered in *lexicographic search*, known as the vocabulary problem [12], because users often do not phrase their query in the language and terms of the documents they aim to retrieve [34]. The vocabulary problem is compounded by synonymity (different words, similar meaning) and polysemy (same word, different but related meaning), leading to decreased recall (missed relevant documents) and precision (inclusion of irrelevant documents), respectively [4].

The problem has been addressed by a number of approaches, including automatic query expansion (AQE), interactive query refinement, relevance feedback, word sense disambiguation and search results clustering, each of which we briefly describe now.

4.1.1 Automatic query expansion

Because large corpora may contain different expressions of the same concept, the vocabulary problem may be addressed by expanding the user query with terms that are related to the keywords provided by the user. Automatic query expansion has long been

used in Information Retrieval to improve search results [4]. The words used for query expansion may be selected using probabilistic methods based on term co-occurrence (*i.e.*, based on their statistic co-occurrence) or based on controlled vocabularies (*i.e.*, thesauri or ontologies). We will refer to these methods as *probabilistic* and *ontological* methods, respectively.

Probabilistic methods can be distinguished into global and local methods depending on the basis on which they calculate the co-occupance of terms (top-n or all documents in the collection). Local methods are executed ad-hoc at query time. It has been shown that a simple approach of using statistical relations has little effect [25]. Vorhees in 1994 used Wordnet[2] as a simplified ontology to expand user queries. In her study, Voorhees manually identified the appropriate concepts—automatic methods still had to be developed. She observed that detailed user queries benefit little from query expansion, but that most users typically enter relatively short queries [34]. She concluded that expanded queries generally are unlikely to outperform well-formulated user-supplied queries. However, it has been observed that search engine users tend to begin by entering short queries, which are gradually modified [28].

4.1.2 Query refinement

The reasons for short search queries might be that users are unfamiliar with the content of a document repository (thus avoiding too specific or narrow queries that might be unsuccessful) [21], or that the user has poorly-defined information needs (redefining their information need as they go along) [3]. This iterative process is called *query refinement* or interactive query expansion [32]. It aims at supporting the user in their decision making process. Efthimiadis found that only about one third of the terms suggested to the user for query expansion was considered useful [7]. In this study, interactive query expansion achieved on average an improvement from originally three highly relevant documents to the inclusion of further nine highly relevant documents. Typically query refinement is done on a syntactic basis.

Query refinement can also be used for semantic queries. Automatic semantic refinement of ontology-based queries aims to incrementally tailor a query to a given ontology and the user's needs [32]. Stojanovic follows an approach related to that of a personal assistant [32, 31]. They analyze the user behaviour during the search process. They refer to the gap between the information need and the current query as *query ambiguity*. Their approach navigates through a neighborhood of similar queries with the aim to decrease the query ambiguity.

4.1.3 Word Sense Disambiguation

None of the approaches described above address the problem faced in our situation. Semantic-enhanced search needs to identify relevant concepts for a given keyword or phrase. Rarely does a keyword belong to exactly one concept alone; the process of identifying the correct concept is called *disambiguation*. The correct concept in a query refers to the intention and query need of the users, while correct concepts for source documents refer to the context and semantics of the document content. Word Sense Disambiguation (WSD) is the task of determining the meaning of a word in a given context [24].

Automatic disambiguation attempts to derive word senses automatically, either as targeted WSD and all-word WSD. Targeted disambiguation aims to identify a selected set of words, whereas all-words disambiguation aims to identify all words in a text. We are interested in targeted disambiguation of keywords. WSD typically uses some kind of knowledge base, which may be ontologies, dictionaries, annotated text corpora, or thesauri. A widely used source is Wordnet [9], which encodes more than 100,000 synonym sets of English words. Kohamban and Lee built a WSD system using a classifier using Wordnet [20]. Most knowledge organization systems like thesauri, classification systems, or DBpedia do not carry information about context and are therefore not suitable for our purpose.

Another approach is using Wikipedia as the originating data collection. WikipediaMiner [23] analyses Wikipedia articles, and based on the results of mining this large text corpus, it exploits word usage statistics to achieve disambiguation. It relies on Wikipedia's prior link probability (likelihood of a link between a word and a concept based on statistical analysis) to determine concepts that match given words. WikipediaMiner uses data mining techniques to disambiguate the semantics of a word. WikipediaMiner therefore does not provide the facility for end-users to influence the disambiguation process, nor does it allow for the manual introduction or modification of concepts.

4.2 Markup for Semantic Search

In order to enable semantic search on existing full-text documents *a-posteriori* semantic mark-up is needed. This is predominantly done with automatic tools, and occasionally manually. Here, semantic search, like lexicographic search, encounters the challenge of disambiguation.

4.2.1 Semantic annotation

Automatic annotation tools, such as OpenCalais,[3] Zemanta,[4] DBpedia Spotlight,[5] and Cohse [35] are services for the semantic web community to increase the volume of interconnected data. Most of these these tools use named entity extraction, also called named entity recognition (NER) and natural language processing. Named entities are " information units such as the name of a person or a location found in a sentence"[13]. Typically the entities are organized in a knowledge base or ontology. The tools use different algorithms and training data, and few comparative evaluations have been conducted to identify the conditions under which each tool is the most appropriate [27, 29, 17]. However, due to the algorithms used, the tools work best on full-text documents but not on a user's search keywords. Furthermore, because the systems use machine learning techniques, the end user cannot explore why these annotations were given or take direct influence.

4.2.2 Semantic keyphrase extraction

Keyphrase extraction aims to identify the most relevant keyphrases of a document, while semantic keyphrase extraction aims to identify the most relevant concepts of a document. Two prominent system for keyphrase extraction have been evaluated and tested in comparison to semantic annotation: KP-miner [8] and Maui [22]. The comparison found that both KP-miner and Maui outperform semantic annotation tools [17]. Maui integrally relies on WikipediaMiner [23] for computing semantic relatedness between phrases and for disambiguation. Our preliminary tests have shown that Capisco outperforms WikipediaMiner's disambiguation.

[2] from `http://wordnet.princeton.edu`

[3] `http://opencalais.com`

[4] `http://www.zemanta.com`

[5] `http://dbpedia.org/spotlight`

4.2.3 *Manual semantic mark-up*

Some argue that automatic annotation is not always of sufficient quality to enable focused search and retrieval: either too many or too few terms are semantically annotated. User-defined semantic enrichment allows for a more targeted approach than automatic semantic enrichment. Manual systems may be distinguished into commenting tools, web-annotation tools, wiki-based systems, content composition systems, digital library tools, and linguistic text analysis. We here discuss as an example system the semantic content authoring tool Loomp, which supports manual markup of text with semantic concepts. A one-click annotator interface is provided for non-expert users to bridge the gap between objective knowledge (as encoded in an RDF data model and ontology) and subjective knowledge of human cognition. The annotator supports this process by presenting labels and contextual information about named entities for identification of semantic identity [15, 16].

Loomp and other manual annotations tools cannot provide semantic markup for large corpora such as the one used in this project. The Loomp authors identified as the most promising annotation approaches those that are semi-automatic, providing automatic suggestions that are manually refined. The resulting documents can be searched using semantic query languages. However, the complexity of such languages still pose a serious impediment for technical non-experts. However, their approach is interesting for the disambiguation of search terms.

5. SEMANTIC-ENHANCED SEARCH

This section introduces our approach of semantically-enhanced search. One of the central components of Capisco is a Concept-in-Context network, whose principles we explain in Section 5.1. We then introduce the system architecture and data flow in Section 5.2. Our search interfaces including the user-guided disambiguation is introduced in detail in Section 6.

5.1 Concept-in-Context network

Instead of an ontology or knowledge base, we use our Concept-in-Context (CiC) network to capture the semantic concepts and the relationships between them. Typically the development of both knowledge bases or ontologies faces similar challenges to that of the semantic annotation of texts: it is a complex task that often consumes considerable human effort [11]. However, ontology and knowledge base engineering is typically executed by experts in semantic technologies and is not necessarily suitable for end-users. Moreover, often the bottleneck in building lies in the social process rather than in the technology [6].

We follow a different approach here creating a Concept-in-Context network that has been seeded with concepts derived from Wikipedia that was inspired by WikipediaMiner (see top-part of Figure 4). Relationships between symbols and their meaning are extracted from Wikipedia links. Representing the semantics with our CiC approach has several advantages. The representation base has a less-strictly formalized structure compared to an ontology. Instead of classes and relationships defined in RDF/Schema and OWL structures that tightly define the semantics and relationships for each concept, our representation structure encapsulates only two types of relationships: synonyms and contexts. This is expressed through words, concepts and context.

We here follow Sowa's understanding of knowledge representation in which word meaning has two aspects: the intension of a word (or *symbol*) refers to its general principles (*i.e.*, the *concept*) while the extension of a word refers to the existing *expressions* of the concept [30]. For the purpose of writing about symbols and concepts, we introduce the following notation: symbols (words, literals) are marked with "quotes" and concepts (both intensions and extensions) are indicated through [brackets]. For example, the symbol "Matariki" refers to the concept [Matariki star cluster]; the symbol "Pleiades" also refers to [Matariki star cluster]. Internally, the concepts are predominantly identified by concept IDs, but within the paper we will use human readable references ([Matariki star cluster] instead of [id:96471]).

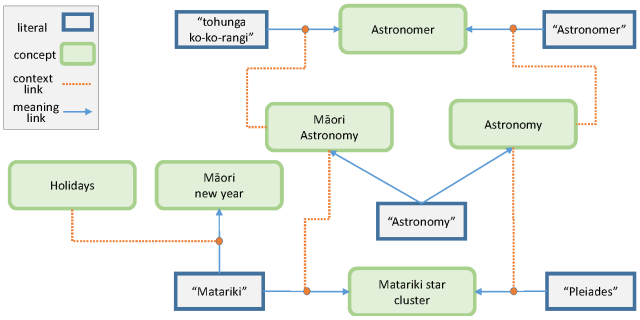

Figure 3: CiC network example

Although Sowa provides a basis for semantic conceptual understanding, he does not cover, or resolve, the problem of ambiguity. Ambiguity is created through the same symbol referring to different concepts. For example, the symbol "Matariki" may also refer to the concept [Māori new year] The way to distinguish between these concepts is by examining their context. Figure 3 shows a small example of the modeling of the relationships between words and concepts and also shows how concepts can serve as context for words.

It can easily be seen that our approach to encoding synonyms as concepts and contexts allows multi-linguality to be treated as a special case of ambiguity because words and phrases of differing languages are synonyms of the underlying concept.

Due to its seeding from Wikipedia, the CiC network may contain noisy data. Because Wikipedia entries are developed through crowd-sourcing, not all entries are of the desired high quality. Furthermore, although the Wikipedia process aims to create a network of information resources, it does not necessarily aim to produce a viable network of semantic concepts.

5.2 Conceptual Architecture

Figure 4 shows the conceptual architecture of the system. Components indicated in dashed lines have been conceptualized but not implemented yet. We now describe each of the components.

CiC Network Seeding. As described above, our knowledge representation is initially seeded through the Capisco system; the relationships between concepts and literals are derived from the links between Wikipedia articles and the anchor terms used in these links (see Figure 5). The location of the link (*i.e.*, information about the originating article) is retained as contextual information. Capisco assigned each concept with an internal concept id (cID). These are stored as triples of the form <*literal,cID,cID*>, which expresses a word or phrase (literal) a meaning in a given context (*i.e.*, <*phrase, context, meaning*>).

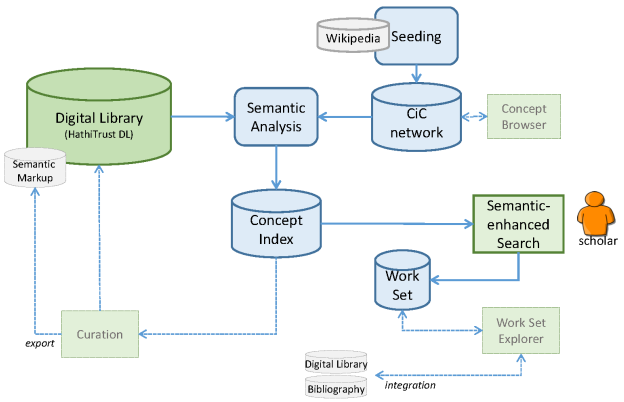

Figure 4: System architecture

Note that both meaning and context are concepts (identified by cID). For example, "Matariki" refers to [Māori New Year] in the context of New Zealand [Holidays]. Additionally we store a short description for each concept. The triples are stored in a MongoDB[6]. Adaptation of the network to reflect current scholarly research and specialized domain knowledge will be done via a concept browser.

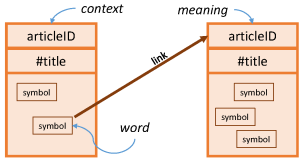

Figure 5: *CiC* triple structure

Semantic analysis and indexing. The documents in the text corpus or digital library are semantically analyzed using the concepts contained in the CiC network. Our disambiguation works on the assumption that it is very likely that terms that appear together in a document will have related meanings. For example, if a sentence mentions "tennis" and the term "set", this most likely refers to the concept [tennis set] not mathematical [set theory]. Thus disambiguation of terms appearing in a document is done through analysing the presence of other terms in the context of the text, which may in turn also be ambiguous. An ambiguous term (*i.e.*, one that refers to a number of concepts in the CC network) may, however, be unambiguous in a given context (that is, it may have only one meaning in a given context). We use those terms that are not ambiguous as anchor points for disambiguation. Conceptually, our disambiguation identifies clusters of all directed graphs between the text of a document and every possible concept in order to identify the most likely semantics. The result of this semantic analysis is a concept index that maintains links between the documents in the corpus and the concepts in CiC network. The semantic information contained in the concept index can be manually refined or exported into semantic markup for the text corpus. The concept index is implemented as Lucene[7] index, for the creation of which auxiliary documents containing the concepts identified for each document in the corpus are used. The document links in the index are redirected from the auxiliary concept documents to the actual text documents.

Semantic-enhanced search. To simplify access to the digital library, Capisco provides a text-based search interface that scholars would find familiar. Search terms entered, however, are processed quite differently to lexicographic search. First, terms are automatically mapped to a list of concepts by Capisco, that are returned to the scholar for inspection. This is the point where manual disambiguation can occur (described in more detail in Section 6). By selectively clicking on the terms returned, a traditional lexicographic search is constructed and – once the scholar is satisfied with their selection – executed against our Lucene index. Returned by this step is a traditional result set with links to matching documents. The resulting set of documents can be saved as a workset to be explored and annotated. Worksets can also integrate documents and references from other sources (digital libraries, corpora or reference systems), or be exported in full or as metadata into other formats suitable for integration into the workflow of the scholar (*e.g.*, endnote or zotero).

Test corpus. The HathiTrust test corpus we are using for our current tests contains over 2 million OCR pages (2,348,172) that are each indexed separately. The pages refer to 8,489 volumes. The corpus' size is 14GB. Metadata about the pages is held in pair tree format,; metdadata about the volumes is available from the HathiTrust. Capisco's CiC network as seeded from Wikipedia contains 4,562,497 concepts that are involved in 101,115,481 triples. The index between HathiTrust test corpus and the CiC network is 256MB (2,942,698 terms link to 2,331,194 documents — missing documents due to unindexable OCR results). We observe that when ranked by frequency in the index, only the top 1115 concepts appear more than 10,000 times. For example, rank 1 is the concept [Shilling] with 268,914 document links, and rank 10 is the concept [water], which appears in 144,916 links. Most terms, however, appear only a few times, such as [St. Stephen's day] with 21 times (rank 100,000).

6. INTERACTIVE DISAMBIGUATION

Manual semantic disambiguation is known to be a challenging task for users that are not familiar with the intricacies of semantic web or ontology concepts [15, 16]. The vocabulary needs to be kept simple and as close as possible to terms with which users are familiar [18, 19]. Our semantic-enhanced search has two main interfaces for scholars without technical knowledge in semantic web technology: one for inserting the query keywords and disambiguation, and one for display of the result list.

Query interface and disambiguation. Figure 6 shows the initial query interface (all screen-shots have been cropped, but retain high resolution for detailed digital reading). The scholar is prompted to insert their search keywords (literals), divided by "|". Note that for simplicity in the Capisco interface we refer to our semantic-enhanced search as 'semantic search'. Figure 7 shows the inserted query containing the two literals "Fox River" and "Public Lands".

Figure 8 shows the list of candidate concepts returned in response to the query entered in Figure 7. The black dot next to the literals indicated that they have not yet been disambiguated. The user is prompted to select a concept for the first literal "Fox River" (top part of screen). Again, for simplicity the term 'concept' is not used but instead the user is prompted for the "sense of 'Fox Rover'". The list of concepts shows all those that are connected with the literal "Fox River" in the CiC network.

Figure 9 shows the screen after the scholar has disambiguated the concept [Fox River (Illinois River tributary)] as the relevant semantics of the literal "Fox River". The concept has been inserted into

[6] http://www.mongodb.org
[7] http://lucene.apache.org

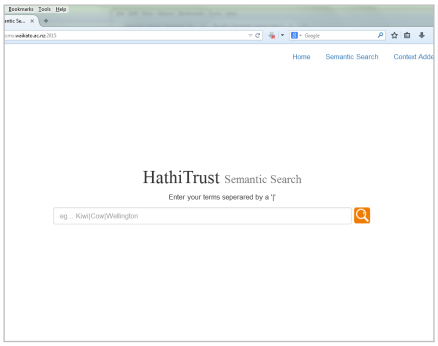

Figure 6: Query interface (start)

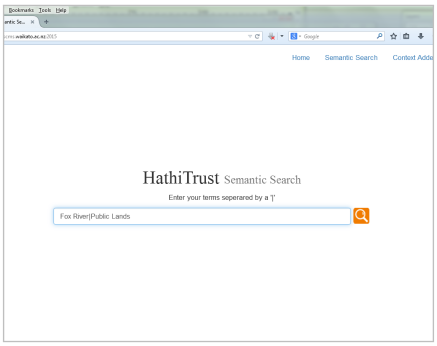

Figure 7: Inserted query of literals

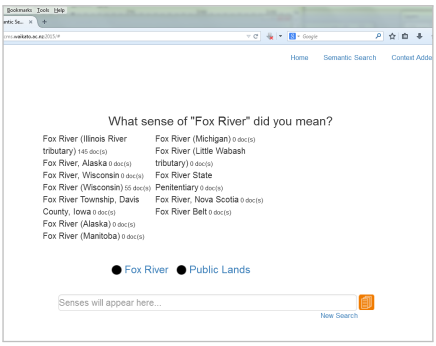

Figure 8: Disambiguation of first term

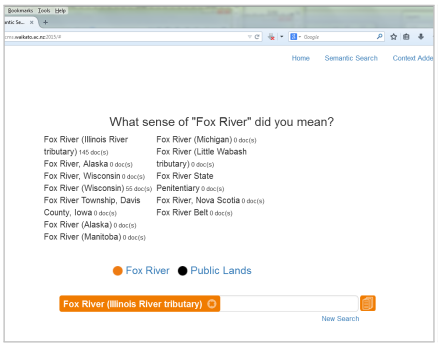

Figure 9: Concept has been identified

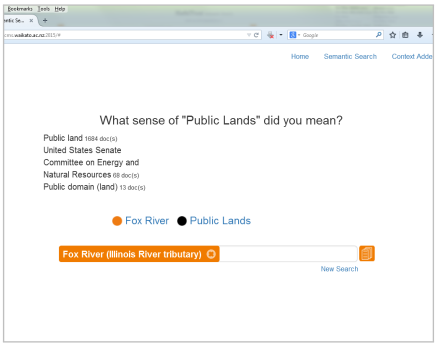

Figure 10: Disambiguation of second term

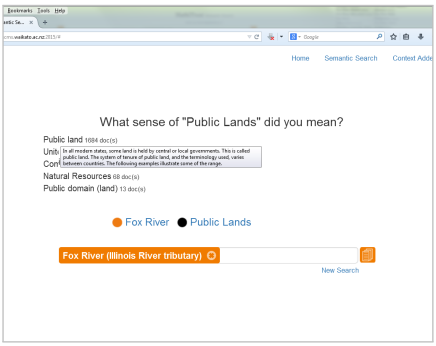

Figure 11: Pop-up for information

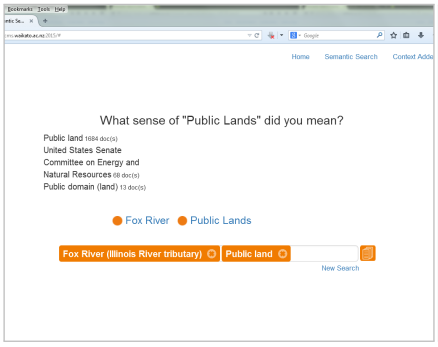

Figure 12: Final query

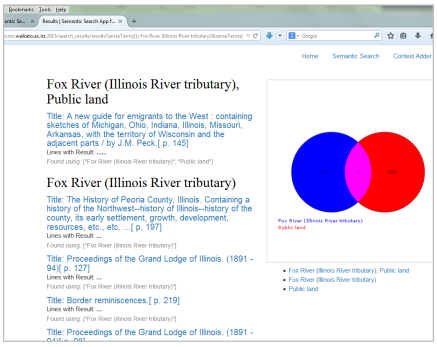

Figure 13: Result listing

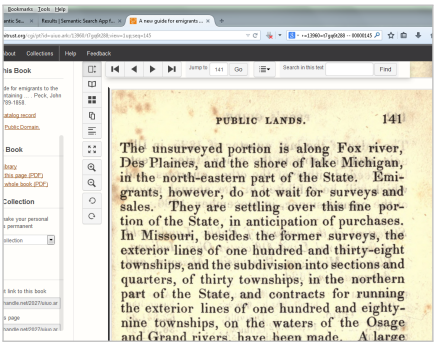

Figure 14: Document view in HTDL

the list of query concepts (in orange), and the orange dot at the literal indicates that this term has been disambiguated. The second literal that needs disambiguating is “Public Lands" (see Figure 10). The scholar switched to this term by selecting the literal next to the black dot. Again, an array of concepts are offered for disambiguation. As the scholar hovers the mouse over each concept, a pop-up window provides the context to the concept through scope notes (see Figure 11). The scholar selected the concept [Public Land], which is inserted into the concept query as shown in Figure 12. The query is now ready to be submitted.

Result list and Document view. The result list shows the documents found and also a Venn Diagram of how many documents were found for each of the concepts, and their overlap. Figure 13 shows the results for our query using the two concepts [Fox River (Illinois River tributary] and [Public Land]. The list of resulting documents is ordered by the number of query concepts that are found in the document. We are currently exploring a secondary ranking by strength of connection between document and concept. The document titles are pulled from the HTDL metadata (via document ID) and are not held in Capisco’s index.

By selecting a link in the result list shown in Figure 13 takes the scholar directly to the HTRC resource as held at the HTDL. Figure 14 shows an example page from the returned result in the HTDL.

7. USE CASES AND DISCUSSION

We show two use cases, one running on the original Wikipedia-seeded CiC network and HathiTrust corpus subset, and a second one running off a subject-specific CiC network and test set, seeded from scholar submitted information.

7.1 “Puck" on the original CiC network

We executed a search for Puck, for which our system offered 14 concepts for disambiguation, among them the concepts [Puck (Mythol-

ogy)] and [Hockey Puck]. Because we were interested in the mythological figure Puck, we selected this concept. The current subset of HathiTrust documents contained in our corpus returned a list of 116 documents. One of the documents returned is “Sylvia", by Adeline Adams (1859-1948) notwithstanding that the term “Puck” appears nowhere in its 120 pages. However, as shown in Figure 15, it contains two references to “Robin Goodfellow". Analysing our Wikipedia-seeded CiC network (see Figure 16), we find that “Robin Goodfellow" is one of the labels for the concept [Puck (Mythology)]. It is, in fact, one of Puck’s euphemistically disguised names.

We also entered the same query “Puck" into the online HTDL interface that uses lexicographic search. 281,407 results are returned. Within the first 100 documents (4 Spanish, 1 Swedish, 1 Polish, 1 Latvian, 1 Italian, 1 German), 44 referred to the concept of Ice Hockey. Explicitly excluding the term “Hockey" (i.e, by inserting -Hockey) results in an appropriately reduced result set. However, documents that refer only to Robin Goodfellow (such as the above cited “Sylvia") were not included.

Entering the same query into Google yields also a mixture of results. Prominently displayed is “Puck, also known as Robin Goodfellow, is a character in William Shakespeare’s play". However of the first 100 hits provided results referring to a mix of the mythological Puck, Hockey, a game, a comic, a pub, and a programmer. All 5 images shown were of Hockey pucks and the News section was about the Yahoo Sports’ NHL blog “Puck Daddy" and “Puck Headlines". Excluding the term “Hockey" removed the results referring to Ice Hockey but not the images of Hockey pucks. Most importantly, Adeline Adams’ “Sylvia" was not included and could only be found by explicitly asking for “Robin Goodfellow".

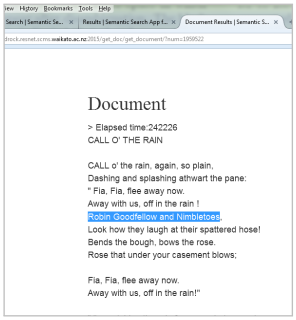

Figure 15: Document view of “Sylvia" (cropped)

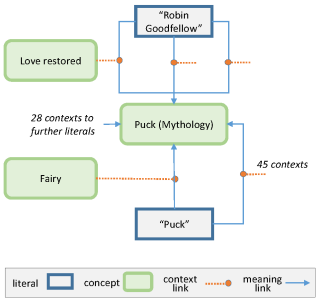

Figure 16: *CiC* sub-network

7.2 “Māori Astronomy" on subject-specific CiC network

There are certain subjects where Wikipedia does not contain sufficient content to generate a useful CiC network that adequately represents those subjects. One such example is Maori Astronomy. For our second use case, we therefore took the tabulated content on Astronomy (see Figure 1) compiled by an internationally recognized scholar on Maori Astronomy—and co-researcher—and developed Content-in-Context triples that were ingested into Capisco to provide the necessary coverage.

Similarly we created a test corpus composed of documents deliberately chosen to confound the meanings of terms that could be associated with Maori Astronomy. We selected documents from three different groups: (Set I) Māori documents that mention the phrase “Māori astronomy, (Set II) documents about the Matariki star cluster without the phrase “Māori Astronomy", and (Set III) documents about Matariki, the Māori New Year.

We use some simple queries here to illustrate the effect of the system. When one searches for:

1) “Māori Astronomy" and disambiguates to [Māori Astronomy] the result set contains the documents from both Set I and Set II.

2) “Matariki" and disambiguates to [Matariki star cluster] the result set contains the documents from Set II and Set I.

3) “Matariki" and disambiguates to [Māori New Year] the result set contains the documents from Set III.

These examples illustrate that on this small test set and the subject-specific CiC network, the system addresses our use case as introduced in Section 2. Documents in Set II are about Māori Astronomy but never mention the term explicitly. However, this meaning is inferred from the words and phrases for which Māori Astronomy defines the context. 5 We illustrate this for a document from Set II: the document contains Chapter 17 (“Heavenly Bodies") from Edward Tregear’s 1904 book titled “The Māori Race"[33]. This is one of the historic books referenced in the word list mentioned in Section 2 and in Figure 14. The document contains the word “Astronomical" once, which may disambiguate to both [Astronomy] and [Māori Astronomy]. It further contains terms such as “Matariki", “Tautoru", “te Kakau", “Pou-ta-te-rangi", “Makahea", “stars", all of which use the context of [Māori Astronomy] for disambiguation. Finally, the text also refers to [Māori people], which provides the contextual support for the link between “Astronomical" and [Māori Astronomy]. We could say that all these star names form the *context support set* of [Māori Astronomy], *i.e.*, the presence of these terms indicates that the concept [Maori Astronomy] acts as a context for this document.

A further example illustrates the result of the disambiguation of this example document corpus. The document is one from Set I, an article titled “A review of Māori Astronomy in Aotearoa-New Zealand" [14]. Figure 17 illustrates the disambiguation decisions taken, and also possible future improvements. We here show the concepts identified in the disambiguation step (and encoded in the index) as a concept map for the document. The number next to each literal indicates how often this phrase appeared in the document. The literal “Māori Astronomy" is unambiguous and can serve as an anchor point for the disambiguation. For example, the term “Atutahi" appears four times in the document (see Figure 17, bottom right). It might refer to the [Atutahi Island], but since there is no context of [Cook Island] in the document to support this disambiguation, this link cannot be established. Rather the context [Māori Astronomy] supports a disambiguation of “Atutahi" to the star [Canopus]. A number of other star names appear, for most of which [Māori Astronomy] provides context. The document also contains “Matariki" 17 times, and additionally the phrase “New Year" seven times. Both readings of [Matariki star cluster] and [Māori New Year] are included. The granularity of our current contextual analysis can be set to whole documents or single pages. The case of “Matariki" is a good motivation for a more fine-grained analysis, which is planned for the future. In this case it would allow us to identify that “Matariki" should only once be disambiguated to [Māori New Year] when the terms appear within the same paragraph. This information can then be used for result ranking.

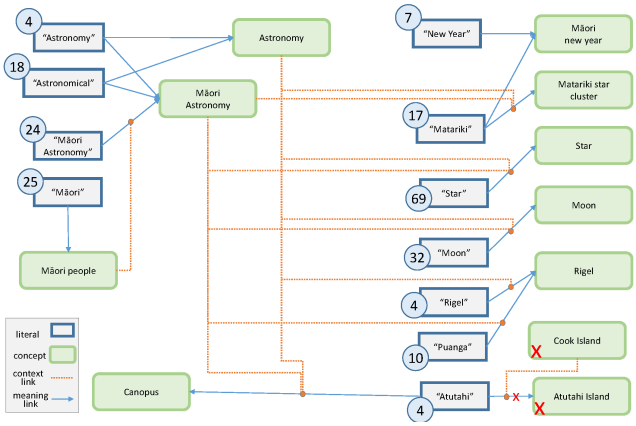

Figure 17: Concept map for document after disambiguation

As a last example, we show how our method can also be used to analyze parts of documents (or to cluster semantically similar documents). Due to human error we used the complete proceedings of the 2008 conference on Traditional Knowledge [26]. We had intended to index one of the articles, which contains a discussion of the phrase "tohunga tātai arorangi" for [Māori Astronomer] and the context of [Māori Astronomy]. The resulting network of concepts for the book shows only a small cluster of terms connected with [Māori Astronomy] (each appearing up to four times), and an overwhelming number of occurrences of Māori (1077), "New Zealand" (334) and "world" (171). Also many of the occurring concepts to not link together well, and the document seems to fall into a number of clusters. It is clear from analysing the concept map for the document that only a fraction of the full document refers to [Māori Astronomy], and that the planned extension to more fine-grained context capturing is appropriate.

7.3 Quality of disambiguation

For in-depth manual analysis, we also built a small corpus of 23 documents. The documents were short articles or introductions into topics (*e.g.*, honey, twitter, family, microbrews), collected from online archives of newspaper articles and the Encyclopedia Britannica; each article was between 1 and 2 pages long. We processed this test corpus in three ways: (1) using WikipediaMiner for identification of concepts, (2) semantic annotation using Capisco, and (3) manual semantic mark-up of concepts. For the manual mark-up each article was independently read by two researchers, using the same concept set seeded by Wikipedia, and the resulting mark-up was checked and by a third researcher and aligned into a single set of concepts. As both WikipediaMiner ad Capisco are seeded by Wikipedia, the set of concepts that can potentially be identified is the same in all three cases.

The first observation is that Capisco did not miss any concepts that the human annotators wished to include. WikipediaMiner had few miss-identified concepts, for example the Africa article contained the phrase "world's land mass", which was wrongly disambiguated by WikipediaMiner to [Mass] and correctly identified by Capisco as [Land Mass].

Overall, the agreement between human annotators and WikipediaMiner per article ranged between 71% and 92% for the 23 articles (avg. of 82%). Out of overall 1531 identified concepts, 1215 were confirmed by the human annotators (78.9%). Out of 1232 concepts identified by Capisco, 98.62% were confirmed by human annotators. The agreement between human annotators and Capisco was 100% for nine of the 23 documents. For 14 documents, Capisco included between one and three additional concepts that the human annotators did not include (agreement on terms 88.8% and above). For example, Capisco identified the concept [Canning] for the article about beer brewing and canned beer. The human annotators excluded this term as the article did not focus on the process of canning beer. None of the additionally identified concepts were wrong, but the human annotators felt that they did not describe well the main focus of the article.

However, we noticed that concepts for which no Wikipedia page had yet been created, could not be included in any of our three methods. This particularly applied to concepts describing recent events ("Ebola virus epidemic") and names of people or organisations. For example, Wikipedia Miner wrongly disambiguated the aid group SIM, which is not contained in Wikipedia, to the concept [Subscriber Identity module]; Capisco omitted this concept. The concept sets in both WikipediaMiner and Capisco could be updated once Wikipedia has been updated. In Capisco, these cases would additionally be able to be addressed through manual extension of the CiC network.

When comparing the number of concepts the three methods produced, we found that the human annotators included the fewest and WikipediaMiner the most concepts. The concepts that WikipediaMiner suggested, which were not included by Capisco, were almost only those concepts that the human annotators did not wish to include. The reason is that often WikipediaMiner's analysis is too literal. For example, "General Information About Africa" was disambiguated by WikipediaMiner into [General officer]. It was disambiguated by Capisco into [General Knowledge]. Similarly WikipediaMiner identified the concept [Run (baseball)] from "Nile River, which runs through ..."; Capisco found no context for baseball in the text and discarded this concept. The same article about Africa also contained information about the hight of Mt. Kilimanjaro (19,340 feet). Both WikipediaMiner and Capisco correctly identified the concept [Foot (unit)], but Capisco (correctly) did not include the concept because no further context supported the notion that the article could be *about* this concept. Capisco correctly identified the Seinfeld episode [Comedians in Cars Getting Coffee], while WikipediaMiner here identified the concept [Coffee]. Overall, Capisco excluded 123 concepts that Wikipedia had identified (*e.g.*, as in the [Coffee] example). A further 186 concepts were excluded because no supporting context was found. Capisco's strength is thus in the indexing of concepts in context, which led to the correct exclusion of 12.15% of WikipediaMiner's concepts.

8. CONCLUSION

Summary. This paper introduced our method of *semantic-enhanced search*, which provides a bridge between lexicographic and semantic search. Our Capisco system implements this multi-lingual approach, which was also used for initial evaluations. We used a Concepts-in-Context network of over 4 million concepts sourced from Wikipedia, combined with the largest scholarly DL to enable scholarly access to OCR'ed documents.

Different to other systems providing disambiguation, Capisco stores context information to ensure quality disambiguation. Our approach thus extends and improves the effective, low-cost measure of semantic relatedness from Wikipedia. On the Digital Library side, this is a simpler and faster approach than building dedicated on-

tologies. On the scholars' side, it avoids complex semantic queries while addressing scholars' needs.

We explored two use cases and explored an extended example. The example showed how our semantic-enhanced search uses keyword queries and text documents but provides the quality of semantic search. The first use case showed the improvements of searches for scholars in the area of Māori Astronomy. We discussed in detail how our search worked for three different document sets. The second study compared the disambiguation of human annotators, WikipediaMiner and Capisco. We found that Capisco further improved the already excellent results of WikipediaMiner through the use of concepts.

Future work. Although Capisco supports multiple languages, the current CiC network uses English anchor terms for concepts (based on English Wikipedia). We plan for the next version of Capisco to become fully language-transparent. The component interfaces for CiC network manipulation, workset exploration and curation are currently under development. Once these are completed, we plan to perform large-scale tests for usability and integration into scholarly workflows, as well as detailed scalability and performance tests.

9. REFERENCES

[1] M. Apperley, S. J. Cunningham, T. T. Keegan, and I. H. Witten. Niupepa: a historical newspaper collection. *Communications of the ACM*, 44(5):86–87, 2001.

[2] V. Basile, J. Bos, K. Evang, and N. Venhuizen. Developing a large semantically annotated corpus. In *LREC*, volume 12, pages 3196–3200, 2012.

[3] I. Campbell. *The Ostensive Model of Developing Information-Needs*. PhD thesis, University of Glasgow, 2000.

[4] C. Carpineto and G. Romano. A survey of automatic query expansion in information retrieval. *ACM Comput. Surv.*, 44(1):1:1–1:50, 2012.

[5] J. S. Downie, T. Cole, B. Plale, K. Fenlon, K. Wickett, and M. Senseney. The workset creation for scholarly analysis (wcsa) prototyping project: Background and goals. In *Proceedings of the Chicago Colloquium on Digital Humanities and Computer Science*, Chicago, IL, December 5–7 2013.

[6] A. Duineveld, R. Stoter, M. Weiden, B. Kenepa, and V. Benjamins. Wondertools? a comparative study of ontological engineering tools. *International Journal of Human-Computer Studies*, 52(6):1111–1133, 2000.

[7] E. N. Efthimiadis. Interactive query expansion: A user-based evaluation in a relevance feedback environment. *J. Am. Soc. Inf. Sci.*, 51(11):989–1003, Sept. 2000.

[8] S. R. El-Beltagy and A. Rafea. Kp-miner: A keyphrase extraction system for english and arabic documents. *Information Systems*, 34(1):132–144, 2009.

[9] C. Fellbaum. *WordNet*. Wiley Online Library, 1998.

[10] K. Fenlon, M. Senseney, H. Green, S. Bhattacharyya, C. Willis, and J. Downie. Scholar-built collections: A study of user requirements for research in large-scale digital libraries. In *Proc. of the Association for Information Science and Technology*, 2014.

[11] G. Flouris, D. Manakanatas, H. Kondylakis, D. Plexousakis, and G. Antoniou. Ontology change: Classification and survey. *The Knowledge Engineering Review*, 23(02):117–152, 2008.

[12] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais. The vocabulary problem in human-system communication. *Commun. ACM*, 30(11):964–971, Nov. 1987.

[13] R. Grishman and B. Sundheim. Message understanding conference-6: A brief history. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 1*, COLING '96, pages 466–471, Stroudsburg, PA, USA, 1996. Association for Computational Linguistics.

[14] P. Harris, R. Matamua, T. Smith, H. Kerr, and T. Waaka. A review of Māori Astronomy in Aotaora-New Zealand. *Journal of Astronomical History and Heritage*, 16(3):325–336, 2013.

[15] A. Hinze, R. Heese, M. Luczak-Rösch, and A. Paschke. Semantic enrichment by non-experts: usability of manual annotation tools. In *The Semantic Web–ISWC 2012*, pages 165–181. Springer, 2012.

[16] A. Hinze, R. Heese, A. Schlegel, and M. Luczak-Rösch. User-defined semantic enrichment of full-text documents: Experiences and lessons learned. In *Theory and Practice of Digital Libraries*, pages 209–214. Springer, 2012.

[17] L. Jean-Louis, A. Zouaq, M. Gagnon, and F. Ensan. An assessment of online semantic annotators for the keyword extraction task. In *PRICAI 2014: Trends in Artificial Intelligence*, pages 548–560. Springer, 2014.

[18] D. Karger. Unference: Ui (not ai) as key to the semantic web. Panel on Interaction Design Grand Challenges and the Semantic Web, at the 3rd International Semantic Web User Interaction Workshop, 2006.

[19] D. Karger and mc Schraefel. The pathetic fallacy of rdf. In *The 3rd International Semantic Web User Interaction*, September 2006.

[20] U. S. Kohomban and W. S. Lee. Learning semantic classes for word sense disambiguation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 34–41. Association for Computational Linguistics, 2005.

[21] M. Lytras, M. Sicilia, J. Davies, V. Kashyap, and N. Stojanovic. On the conceptualisation of the query refinement task. *Library Management*, 26(4/5):281–294, 2005.

[22] O. Medelyan, E. Frank, and I. H. Witten. Human-competitive tagging using automatic keyphrase extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, pages 1318–1327. Association for Computational Linguistics, 2009.

[23] D. Milne and I. H. Witten. An open-source toolkit for mining wikipedia. *Artificial Intelligence*, 194:222–239, 2013.

[24] R. Navigli. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41(2):10, 2009.

[25] H. J. Peat and P. Willett. The limitations of term co-occurrence data for query expansion in document retrieval systems. *Journal of the American Society for Information Science*, 42:378–383, 1991.

[26] J. S. T. Rito and S. M. Healy, editors. *Proceedings of the Traditional Knowledge Conference 2008: Traditional Knowledge and Gateways to Balanced Relationships*. New ZealandŠs Mäori Centre of Research Excellence, 2008.

[27] G. Rizzo and R. Troncy. Nerd: evaluating named entity recognition tools in the web of data. In *ISWC'11, Workshop on Web Scale Knowledge Extraction (WEKEX'11)*, 2011.

[28] C. Silverstein, M. Henzinger, H. Marais, and M. Moricz. Analysis of a very large altavista query log. *ACM SIGIR Forum*, 33, 1998.

[29] R. Sinkkilä, O. Suominen, and E. Hyvönen. Automatic semantic subject indexing of web documents in highly inflected languages. In *The Semantic Web: Research and Applications*, pages 215–229. Springer, 2011.

[30] J. F. Sowa. *Conceptual structures: information processing in mind and machine*. Addison-Wesley Longman Publishing Co., Inc., 1984.

[31] N. Stojanovic. Information-need driven query refinement. *Web Intelli. and Agent Sys.*, 3(3):155–169, July 2005.

[32] N. Stojanovic, R. Studer, and L. Stojanovic. An approach for step-by-step query refinement in the ontology-based information retrieval. In *Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence*, WI '04, pages 36–43, Washington, DC, USA, 2004. IEEE Computer Society.

[33] E. Tregear. *The Maori Race*. AD Willis, 1904.

[34] E. M. Voorhees. Query expansion using lexical-semantic relations. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '94, pages 61–69, New York, NY, USA, 1994. Springer-Verlag New York, Inc.

[35] Y. Yesilada, S. Bechhofer, and B. Horan. Cohse: dynamic linking of web resources. Technical report, Sun Microsystems, Inc., 2007.