Final project report

Capisco: Semantic Analysis of Documents from the HathiTrust Corpus

Annika Hinze, Craig Taube-Schock, Sally Jo Cunningham, David Bainbridge
University of Waikato, New Zealand

1. Description of the project and its purpose

This project developed a tool to assist scholars in identifying and selecting resources from within the HathiTrust Document Corpus. Current access to this resource is available via text-based search in full-text and metadata. Existing scholarly document search tools use purely lexical analysis, which cannot address the inherent ambiguity of natural language. For example, a simple search on the country name "Niue" will miss references to it by its traditional names ("Nuku-ta-taha") and by its initial English name ("Savage Island"). Our Capisco System analyzes documents by the semantics of their content.

Traditional access to the digitized document collections is available primarily via string-based search in the documents' full-text and metadata. Such a text-based search identifies documents purely according to lexicographical analysis. Most research questions and areas of scholarly interest, however, can rarely be described by simple textual keywords and instead, they encompass larger concepts. Relevant sources remain undetected unless the right keywords are found. Easy identification of appropriate keywords is further hindered when different languages are involved 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.

To facilitate scholarly work on the HathiTrust document set, a clustering of documents by semantic similarity could open up a wealth of further opportunities. We suggest analyzing documents not purely by their text but rather by the semantics of their content. 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. Our project developed the conceptual design and 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.

The project was executed in close collaboration with two humanities scholars from the areas of Māori & Pacific Studies, and Historical Anthropology. The research team like to acknowledge their collaboration with humanities scholars Tom Ryan (Cultural Studies and Historical Anthropology) and Rangi Matamua (Māori & Pacific Development Studies), University of Waikato, New Zealand. The scholars did not only drive this project with research questions based on their scholarly practice, but also provided ongoing input and feedback during the development process.

2. Changes to personnel or project management since initial proposal

No changes were made to the proposed personnel.

3. Summary of project progress and significant accomplishments

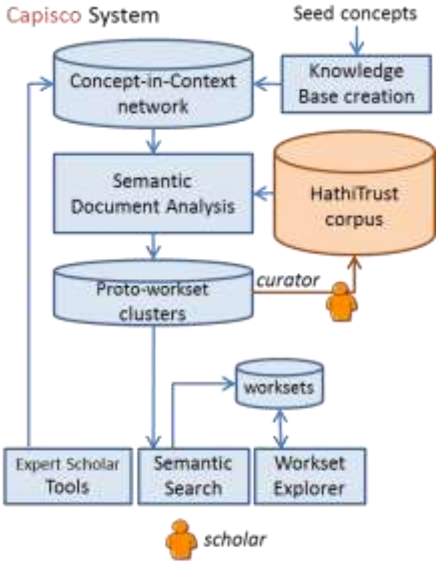
This section describes the project approach, its research contributions and each of the components that were developed.

Capisco Approach

Capisco's semantic analysis is executed in two steps:

Knowledge Base creation: The various meanings of words are encoded in an automatically generated Concept-in-Context (CiC) network seeded by *a priori* analysis of Wikipedia texts and identification of semantics. Our CiC network encodes, for example, that the term "apple" refers to a *fruit* in the context of *nutrition* and to *computers* in the context of *IT*.

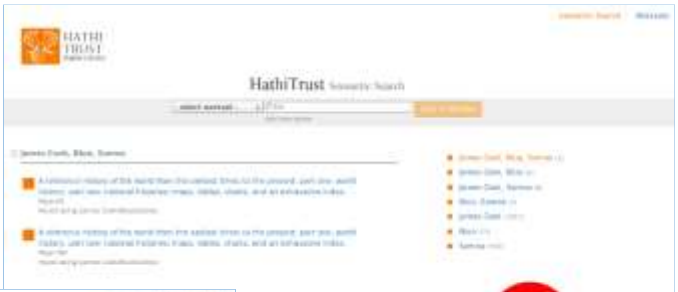
Semantic Document Analysis: We then analyze which concepts and contexts appear in each document in the corpus with the goal of assigning a set of semantic meanings to each document. The documents are then clustered by semantic concepts forming so-called *proto-worksets*.

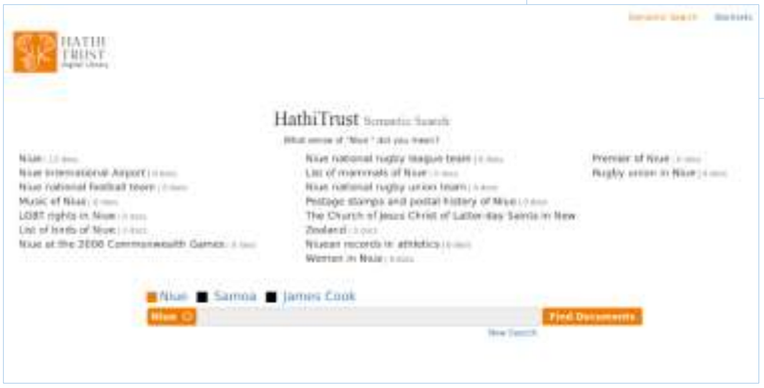


Support for Scholars

Two tools are provided for scholarly search and exploration of worksets.

Semantic Search: This tool allows scholars to search the corpus using concepts instead of keywords. For example, instead of having to search separately for any mentioning of "Niue", "Nuku-ta-taha" or "Savage Island", the scholars search for the concept *Niue*. The results are ordered by semantic clusters.

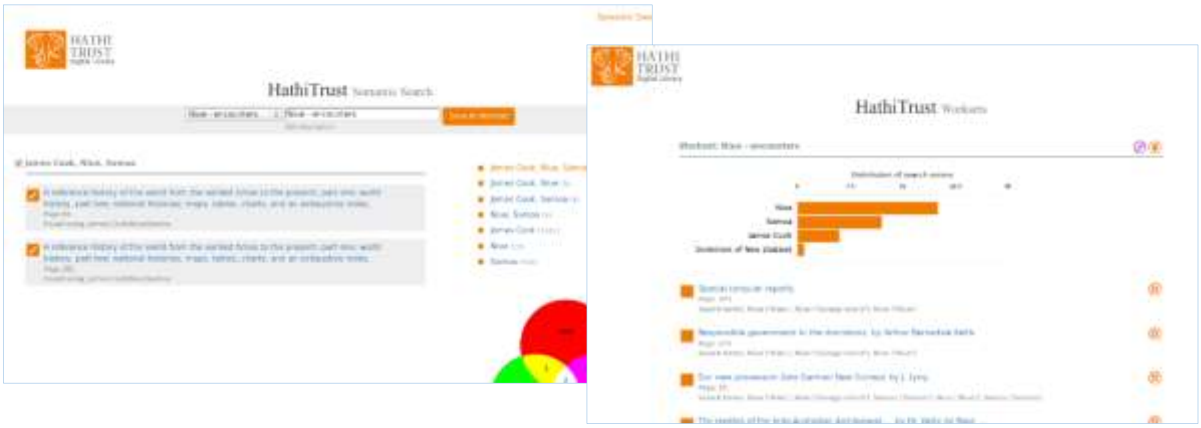




What sense of "Niue" did you mean?

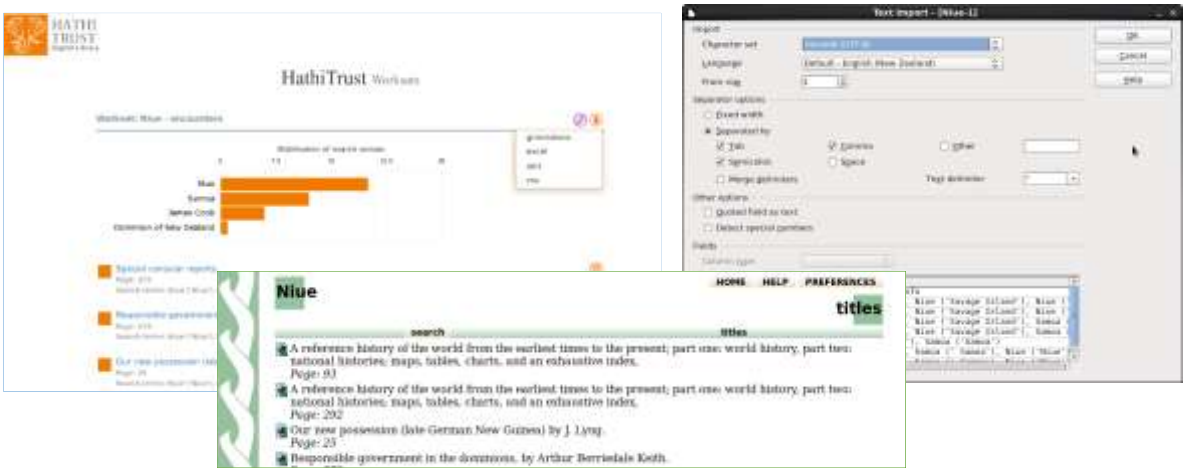
Niue
Niue International Airport
Niue national football team
Music of Niue
LGBT rights in Niue
List of birds of Niue
Niue at the 2006 Commonwealth Games
Niue national rugby league team
List of mammals of Niue
Niue national rugby union team
Postage stamps and postal history of Niue
The Church of Jesus Christ of Latter-day Saints in New Zealand
Niuean records in athletics
Women in Niue
Premier of Niue
Rugby union in Niue

Workset Explorer: The results of the semantic search are presented not only as a list of documents, but also as graphic representation of clusters (proto-worksets). From here scholars can easily explore, create and manipulate worksets, and select them for integration into formal worksets.



Interoperability

Worksets created by a scholar or group of scholars can be exported and joined into existing data sets created from HathiTrust data or other external sources (e.g., Greenstone Digital Library, Excel, XML, CVS). The latter allows the scholar to incorporate worksets into a range of familiar bibliographic tools.



A reference history of the world from the earliest times to the present; part one: world history, part two: national histories; maps, tables, charts, and an exhaustive index.
Page: 93

A reference history of the world from the earliest times to the present; part one: world history, part two: national histories; maps, tables, charts, and an exhaustive index.
Page: 292

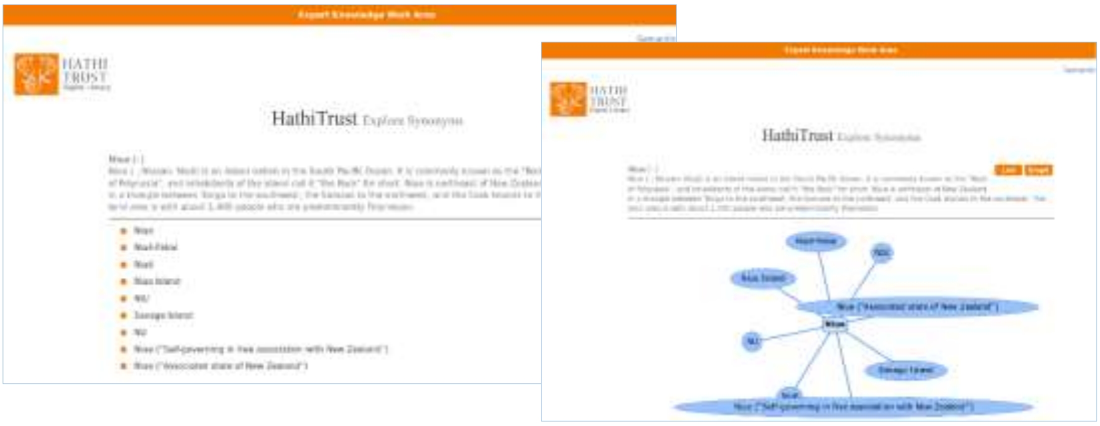
Our new possession (late German New Guinea) by J. Lyng.
Page: 25

Responsible government in the dominions, by Arthur Berriedale Keith.

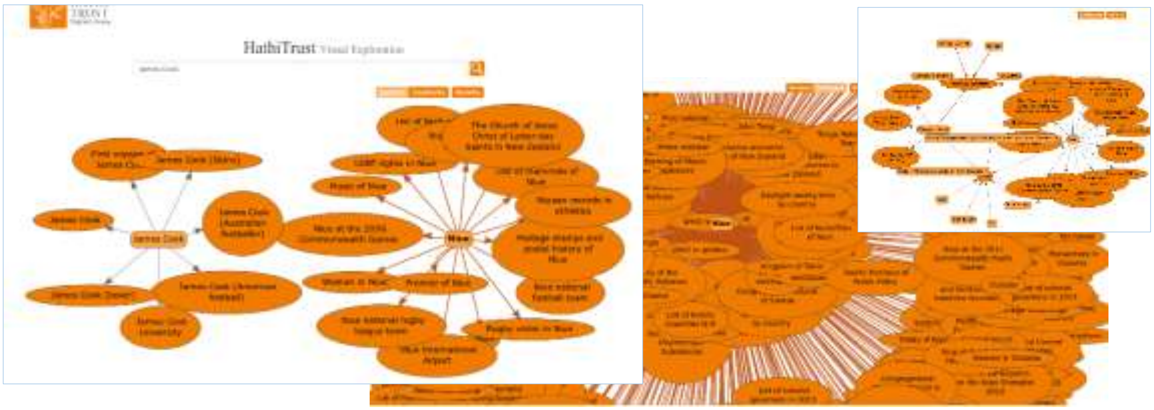
Expert Scholar Tools

High quality semantic analysis requires manual adjustment of concepts such that marker terms and key concepts from a scholar's field are well represented (e.g., adding the name "Nuku-ta-taha" for *concept Niue*). We provide five expert tools to both explore and enrich the knowledge base.

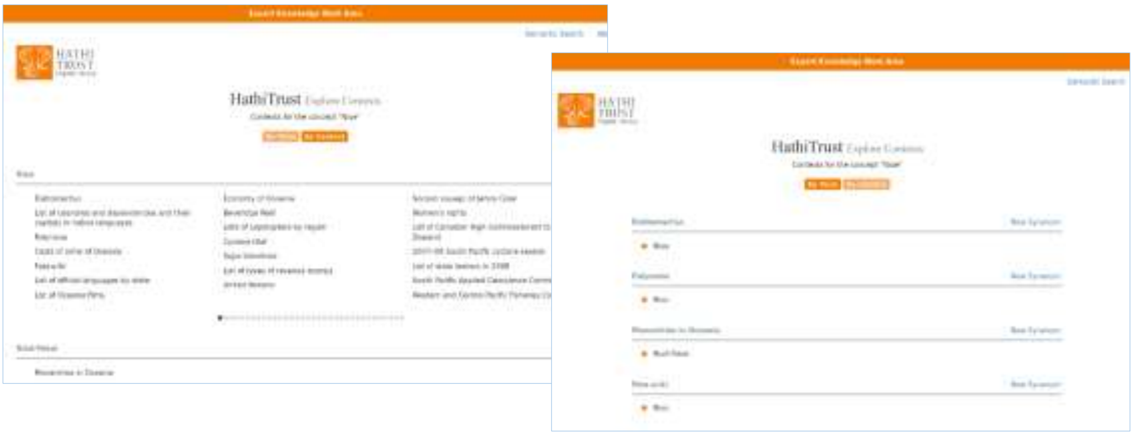
Synonym browser: The Synonym Browser allows scholars to explore synonym words for a given concept (e.g., "Niue" and "Savage Island" for concept *Niue*); it provides a list and a graph view.


Expert Knowledge Work Area
HathiTrust Explore Synonyms
Niue
Savage Island
Niue ("Self-governing in free association with New Zealand")
Niue ("Associated state of New Zealand")

Concept Browser: The Concept Browser allows scholars to explore all links between concepts, contexts and terms. The scholar can interactively walk through the network as an expanding graph.

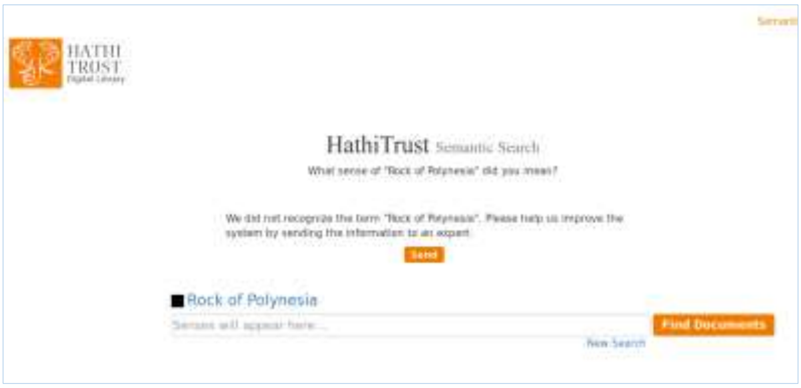

HathiTrust Visual Exploration
James Cook
Niue
Niue at the 2006 Commonwealth Games

Context Browser: The Context browser shows synonyms and their context, for a selected concept (e.g., the concept *Niue*). They can be viewed through two different facets: by terms and by context.



Synonym Adder: For a given concept and a given context, the scholar can add new synonyms to an existing term. This provides the scholar with a first step into easy tailoring of the knowledge base. The Synonym Adder is integrated with the Context Browser (see above).

Basket of Knowledge: When a scholar uses terms or phrases that are currently not recognized through the semantic analysis, these can be submitted as a request into a *Basket of Knowledge*. This request is then submitted to expert scholars (knowledge workers), who then can include the terms and concepts into the Knowledge Base (the Concepts-in-Context network). Every scholar is assumed to be an expert in their specialization—submitting a request may therefore be effectively a `note to self’.

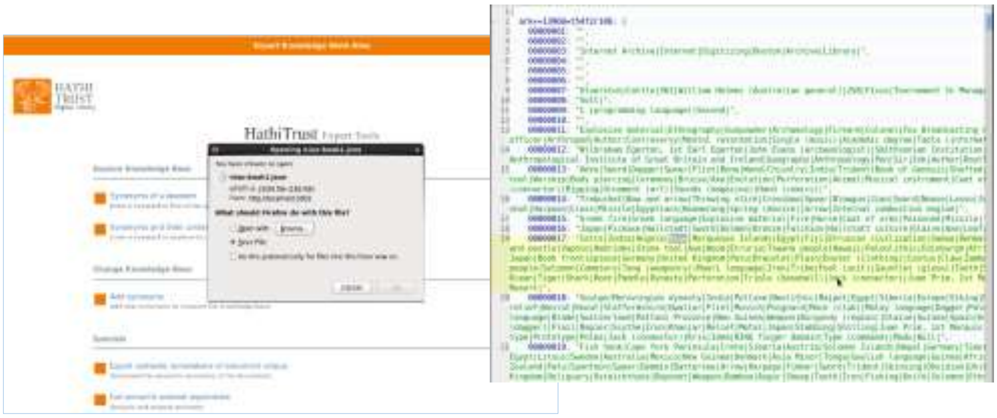

HathiTrust Semantic Search

What sense of "Rock of Polynesia" did you mean?

We did not recognize the term "Rock of Polynesia". Please help us improve the system by sending the information to an expert.

Curation: Corpus Enrichment

Integration of the semantic links between concepts and documents as annotations or mark-up, e.g., into HathiTrust metadata, is possible via the curator export function of semantic information.



Contributions of the project to Humanities Scholars

The Capisco System provides a number of benefits for scholarly search and workset creation:

- More documents are found through semantic search (fewer missed matches)
- Complex searches are made easy through avoiding repeated lexical searches
- Exclusion of results that match at word level but do not match desired semantics
- Workset creation and exploration by semantics (discover more about selected documents)
- Incorporation of scholarly knowledge through expert tools
- Semantics export into HathiTrust metadata (semantic enrichment of metadata)
- Interoperability through workset export
- Use cases of worksets: Small Nations (Niue) and Maori Astronomy
- Parallelizable software for non-consumptive document analysis (respects copyright)

4. Explanation of any setback or challenges

The current Capisco version is a proof-of-concept implementation, which fulfills the research aims for this project.

However, in order to transfer the software into a production system, a number of challenges need to be addressed to transfer the insights of this project into production software.

These were discussed during the project presentation (1 April 2015) and have been begun to be addressed after the project finished. They are listed as current limitations below.

Current limitations

The semantic-enhanced search in Capisco provides user interfaces geared towards Humanities Scholars. The underlying software is divided in three main functions, some elements of which are proof of concepts.

1. CiC knowledge network creation: The knowledge network creation is a lengthy process but only needs to be executed once.

2. Disambiguation & indexing: Disambiguation and indexing was explored using two approaches: one building on a pre-existing software (faster but lower quality) and another one with improved semantic analysis quality. For this second software package, we explored semantic quality but not yet automation and scalability in a data-heavy environment.
3. Semantic-enhanced search of documents: The search function relies on the higher-quality indexing process in order to ensure best results.

In order to explain the details of current ongoing investigations into scalability and performance, we list the software elements of the Capisco backend. This part of the software (1.+2. above) is structured into three main pieces.

- Sentence parser: This component attempts to break blocks of text into what it can best identify as sentences.
- Context identification: This component attempts to identify the context (or context) of the body of text. This is the most computationally expensive component because it attempts to reconcile between context and concepts from a very large number of ambiguous terms. Performance can be dramatically improved if unambiguous context is provided as a parameter along with the body of text being analyzed.
- Disambiguation: This component disambiguates terms based on context/concept relationships. These relationships can be strong (if they are bidirectional) or weak if they are unidirectional.

Ongoing research is dealing with the computational complexity of the context identification. Among other issues, this complexity is increased by the noise in Wikipedia (reducing the performance of high-quality automatic identification of context). The various pieces outlined above are currently independent processes that are combined sequentially. This is currently done manually due to the investigative nature of the current research as this allows flexible testing of combinations of processes as part of both analysis and synthesis.

Future research directions

A project like this always opens possibilities and avenues of future research. We outline here some of these that are particularly relevant to digital humanities.

Performance:

- Incremental indexing after knowledge base extension
- Improving query performance for very large corpora

Functional:

- Integration of semantic search into scholarly workflow,
- In-depth semantic workset analysis for scholars
- Scholarly expert tools for knowledge base manipulation

5. List of project disseminations and final deliverables

Current project disseminations

- Annika Hinze, Craig Taube-Schock, David Bainbridge, Rangi Matamua, J Stephen Downie: "Improving access to large-scale Digital libraries through Semantic-enhanced Search and Disambiguation", *Proceedings of the International Joint Conference of Digital Libraries*, June 2015
- Sally Jo Cunningham, Annika Hinze, David Bainbridge, Craig Taube Schock, Thomas Ryan: "Building heritage document collections for Pacific Island nations using semantic-enriched search", *Proceedings of the Samoa III Conference*, March 2015

Further publications are in preparation.

Final project deliverables

The following list gives an overview of the proposed project deliverables as per contract and the references to the respective deliverables.

1. Final report: this document
2. Well-documented source code for software for semantic clustering, concept browser, semantic search interface, work set explorer, integration tool, and curator tool:

   The software has been uploaded onto github (*github.com/HTrustProject/SemanFinal2015*)

   - Capisco backend: knowledge network (CiC network) creation
   - Capisco indexer: disambiguation, document indexer
   - Capisco frontend: semantic search interface, workset explorer, synonym browser, concept browser, context browser, synonym adder, basket of knowledge)
3. Sample output demonstrating enhancements for workset creation:
   Sample outputs have been documented
   - in this report
   - in the two peer-reviewed publications, and
   - in the demo video (available at *youtu.be/2LiW_4X_6iU*)
4. New worksets:
   - We created smaller worksets for Niue Cultural Heritage and Samoan Cultural Heritage.
   - The Niue Cultural Heritage workset features in the demo video.
   - Larger worksets are in preparation (available after scalability has been addressed).
5. Copies of or links to any scientific publications:
   - The copies are attached and further publications will be made available as they are published

6. Projected vs actual expenses

There are no differences between projected and actual expenses for the project.

15 June 2015 A. (Annika Hinze)