Submitted on: 17.07.2015

A Tool for Systematic Visualization of Controlled Descriptors and Their Relation to Others as a Rich Context for a Discovery System

Frank Seeliger
Technical University of Applied Sciences Wildau
Wildau, Germany
fseeliger@th-wildau.de

Abstract:

The discovery service (a search engine and service called WILBERT) used at our library at the Technical University of Applied Sciences Wildau (TUAS Wildau) is comprised of more than 8 million items. If we were to record all licensed publications in this tool to a higher level of articles, including their bibliographic records and full texts, we would have a holding estimated at a hundred million documents. A lot of features, such as ranking, autocompletion, multi-faceted classification, refining opportunities reduce the number of hits. However, it is not enough to give intuitive support for a systematic overview of topics related to documents in the library. John Naisbitt once said: "We are drowning in information, but starving for knowledge." This quote is still very true today.

Two years ago, we started to develop micro thesauri for MINT topics in order to develop an advanced indexing of the library stock. We use iQvoc as a vocabulary management system to create the thesaurus. It provides an easy-to-use browser interface that builds a SKOS thesaurus in the background. The purpose of this is to integrate the thesauri in WILBERT in order to offer a better subject-related search. This approach especially supports first-year students by giving them the possibility to browse through a hierarchical alignment of a subject, for instance, logistics or computer science, and thereby discover how the terms are related. It also supports the students with an insight into established abbreviations and alternative labels. Students at the TUAS Wildau were involved in the developmental process of the software regarding the interface and functionality of iQvoc. The first steps have been taken and involve the inclusion of 3000 terms in our discovery tool WILBERT.

Introduction

We all know that meta search engines like Google are very comfortable and of course, we are familiar with handling it by searching and get an overview of the hit landscape at a glance. However, we know what we are missing there. Our discovery tools promise a multifaceted

navigation ensuing a fuzzy search, auto complete, spell checks and so on. Nonetheless, we might ask if this is enough to easily manage dozens of results after searching for an item? If you look in extensive databases with qualified information of publications such as PubMed you will be inundated with more than three million results, including abstracts, when inserting terms such as cancer! Moreover, after other buzzwords or combinations, for instance, a search for the Alzheimer's disease or gene, it will not be surprising to get more than eleven thousand references. If you use the Scopus database, the number of hits in the same combination are more than twice as much. How can customers, scientists or students keep track of it? Or will our catalogue-centered systems never have the same size or number of items in the database? At our university we expect more than one hundred million items in the near future in our index and backend. Also, we would like to include and integrate all licensed publications, articles and separate chapters of books. What will we do then and how can we manage the diversity and plurality of information at a glance? One way - in and out - we think, could be to introduce a hierarchical, tree-structured system in which terms and buzzwords are connected in relation to others. Our thesaurus is such a system and is integrated in our discovery service named WILBERT.

The Library Search Engine WILBERT

Since March 2013 the library of the TUAS Wildau offers WILBERT parallel to their conventional online catalogue.
The basic search engine technology, which is used in this case, is called ALBERT. ALBERT was developed and hosted by the Head Office of the KOBV (Co-operative Library Network, Berlin-Brandenburg). The library of the TUAS Wildau reserves the data sovereignty and is responsible for providing the data, the data management and the selection of sources to be indexed.
The use of ALBERT technology within the research environment of the library promises, in addition to the general advantages of search engine technology (auto complete search, fast response times, faceting the search result, fuzzy search), the possibility to integrate more information spaces of the library as well.
WILBERT was adapted in an extensive preparatory phase of the profile of the information environment of the library: The index of WILBERT includes not only the metadata from the catalogue, but also Metadata from the Open Access- repository of the UAS Wildau and other relevant data stocks (eg EconStor). Unlike the online catalogue, which allows only a search into approximately 120,000 catalogue records, WILBERT realizes a search in more than eight million items, which are updated daily. These items are partially indexed in full text regardless of their appearance (eg. as a dependent unit works).
However, with our step to integrate the thesaurus into WILBERT, the library team encountered general criticism in dealing with search engines: To introduce semantic structures, which are based on controlled lexica such as a Thesaurus into the environment of a library search engine like WILBERT, has the objective to sharpen the accuracy of the set of hits generated by the machine.

The thesaurus project at the library of the UAS Wildau

The thesaurus project started in the beginning of 2013 and was aimed at creating a bilingual thesaurus in German and English in the fields of logistics, transport, economics and computer science in order to give students an overview about a specific subject area and to offer all users of the library a better orientation within more than 8 million available items. The preferred format for controlled lexical items is the RDF (Resource Description Framework) based formal language SKOS (Simple Knowledge Organization System). This W3C (World Wide Web Consortium)

recommendation allows thesauri to be published on the web and to interconnect with other SKOS vocabulary entries. The following steps have been taken from the thesaurus construction to its integration into the library search engine.

First Step: Infrastructure

The beginning of the project in 2013 started with the establishment of the organizational and technical infrastructure for project management, project documentation and vocabulary management. As the usage of Wiki software is well known at the library of the TUAS Wildau, MediaWiki was selected for managing and documenting the project.

The search for a vocabulary management system started from scratch with the evaluation of available software. The selection criteria for vocabulary management systems in a linked data environment are described in Morshed and Dutta's (2012) article *Machine Learning based Vocabulary Management Tool Assessment for the Linked Open Data*. For providing an orientation about the functional range of vocabulary management tools, the authors examined criteria like functionality, complexity, maintainability, learnability, availability, flexibility, configurability, multilinguality, authentication and others. In addition to these particular criteria, other project specific demands such as the open source availability with an active user community were also considered.

After different tools had been tested, the vocabulary management system iQvoc was selected. Once installed, it is an easy-to-use, web-based tool that creates SKOS vocabulary entries and considers the basic linked data principles.

Second Step: Vocabulary work

After the installation of the required software we started the construction of the thesaurus for different subject areas such as logistics and transport, taking international and German standards and specifications, such as ISO 25964-1 and DIN 1463-1, into consideration.

As the thesaurus has predominantly been created for and oriented towards student's needs, the fundamentals with regard to content of the thesaurus are the module descriptions for the degree courses. They contain the most important topics of a subject with degree program goals and references. By extracting these topics from the descriptions and by using the Table of Contents (TOC) from the references, a basic overview for a course of study was created and put into hierarchical order. For completing the modeling of a specific subject area and for adding alternative labels and relations, other available sources like glossaries, terminologies, dictionaries or other thesauri were also included.

Another additional value is the use of definitions from subject specific reference work for providing a closer explanation of the concepts. For many sources a licence agreement with the copyright holder for reusing their content is required.

Third Step: Feedback

Receiving feedback on the thesaurus is important for quality assurance issues. The constructed thesaurus is reviewed by faculty and staff of the UAS Wildau in different ways. One important aspect is to provide an easy-to-use and quick feedback platform in order to get support from colleagues.

Available online tools such as card sorting experiments provide a simple approach for getting feedback about the given structures. Originally, these tools were used in user experience design. Users categorize predetermined items for checking the plausibility of website structures or menus.

For the thesaurus context, a list of concepts from the computer science domain had to be assigned to predefined groups (applied computing, computer engineering, programming and software, theoretical computer science) or freely selectable groups by faculty and staff of the UAS. The outcome reflects the staff's view of the categorization of the most important concepts for a specific field.

An alternative way of approaching the problem is to get a qualified overview of an unfamiliar subject. This can be obtained in an old-fashioned way: each concept that you would like to bring into hierarchical order is written on a piece of paper and experts are asked to arrange them for you. By listening to their thoughts while arranging them, one can get a good impression of how concepts belong together. The results of the arrangement are available straightaway and can be integrated into the thesaurus.

Fourth Step: Integration

Making the thesaurus available in the university's search environment through integration into the local library search engine is the main objective of the project. After the creation and review processes, the thesaurus is exported from iQvoc as RDF/XML document for further processing. Three thesaurus applications have been implemented so far:

1. The thesaurus hierarchies can be browsed via an intuitive tree structure with many branches in order to get a quick overview of a specific subject domain.
2. When looking for specific concepts, a search with autocomplete functionality for thesaurus concepts and their alternative labels can be used. After choosing a specific concept, a more detailed view with broader, narrower and related terms as well as definitions and examples is displayed. With the selected concept you can initiate a search in the library search engine Wilbert.
3. Furthermore, the thesaurus is applied within the library search engine's hit list. When searching via the simple search or advanced search, the generated hit list makes thesaurus concepts visible for the user. Available thesaurus concepts and alternative labels are highlighted in a different colour. If you click on one of them, a pop-up window opens, containing detailed information about the concept. This functionality enables the user to expand or limit the search with broader or narrower terms or to search for related terms instead.

Conclusion

The project has been a success for the university library. Our analytic tool Piwik shows that the thesaurus tool is used and accepted, which also means that we have to teach more about this foreign-word-centered tool in our information literacy courses. Surprisingly, we discovered other thesaurus applications like a topic-structured overview for freshmen to get a first orientation by their choice of degree course. One of our next goals is to integrate the thesaurus in organizational structures of the university and into the descriptions of degree programmes.

References

Borchert, F., Keidel, P. (2014): Und sie bewegt sich doch! Der Einsatz eines Thesaurus zur Unterstützung der Sacherschließung in einem Discoverysystem. b.i.t. online 17(5):456-463.

Morshed, A., Dutta, R. (2012). Machine Learning based Vocabulary Management Tool Assessment for the Linked Open Data. *International Journal of Computer Applications* 60(9):51-58, December 2012. Available from http://ijcaonline.org/archives/volume60/number9/9724-4197