



Usage and Impact of Controlled Vocabularies in a Subject Repository for Indexing and Retrieval

Timo Borst

German National Library for Economics, Kiel/Hamburg, Germany,
t.borst@zbw.eu

Abstract

Since 2009, the German National Library for Economics (ZBW) supports both indexing and retrieval of Open Access scientific publications like working papers, post-print articles and conference papers by means of a terminology web service. This web service is based on concepts organized as a ‘Standard Thesaurus for Economics’ (STW), which is modelled and regularly published as Linked Open Data. Moreover, it is integrated into the institution’s subject repository for automatically suggesting appropriate key words while indexing and retrieving documents, and for automatically expanding search queries on demand to gain better search results. While this approach looks promising to augment ‘off the shelf’ repository software systems in a lightweight manner with a disciplinary profile, there is still significant uncertainty about the effective usage and impact of controlled terms in the realm of these systems. To cope with this, we analyze the repository’s logfiles to get evidence of search behaviour which is potentially influenced by auto suggestion and expansion of scientific terms derived from a discipline’s literature.

Key Words: information retrieval; controlled vocabulary; authority data

Introduction

For a long time, information retrieval has been a classical topic for information systems not designed for public use, but for special information needs. With the massive digitization of information and the overwhelming development

of the web, publication of libraries' data (metadata, catalogues) has become as common as problematic: library holdings and catalogues in their traditional style are well-structured and pretend to be exhaustive, but they are not designed for being published on the web or for sharing their information (Lagoze, 2010, p. 75). Instead, they provide their own metadata, categories and 'languages' — to their critics' minds hiding rather than exposing information. In particular, with the overwhelming use and acceptance of commercial search engines crawling the web, and more recently with the emergence of social mechanisms for publishing, tagging and sharing documents, the *raison d'être* for special vocabularies pretending to be universal and objective has been put strongly into question. However, we still see some significant potential for developing and using controlled vocabularies in two respects: first, to support retrieval in a certain scientific domain like economics by automatically suggesting search terms which have been used for indexing, including related terms like synonyms and translations, hence bridging the gap between information systems' language and users (Petras, 2006); secondly, to support retrieval *across domains* by automatically suggesting search terms which have an adequate meaning in the complementary domain (Mayr and Petras, 2008).

After a brief outline of the technical infrastructure supporting vocabulary control in repository systems, this paper describes the method and results of our logfile analyses. In conclusion, we suggest adaptations to search interfaces of library applications to support a search and discovery experience which is substantially relying on librarian's work and offering search facilities which go beyond common web search.

Terminology Web Services as Agents for Automatic Search Term Suggestion and Expansion

The general idea is to provide a technical framework for integrating authority data which is both normative and flexible enough to tolerate local idiosyncrasies on a string level — for instance to allow indexing or retrieving of concepts or names in a certain language or a specific notation. According to Semantic Web standards, a concept like '[financial crisis](#)' has a persistent identifier redirecting via HTTP 303 to machine-readable information on translations, broader and narrower terms (Figure 1).

Fig. 1:

uri	surname	forename	variantNames	academicTitle	lifeData	affiliations
http://d-nb.info/gnd/124825109	"Snower"	"Dennis J."	Snower, Dennis James Snower, D.J. Snower, Dennis Snower, D.	Prof. Ph.D.	1950-	http://d-nb.info/gnd/1007681-5



Repräsentation

Similarly, this can be applied to other authority-controlled data, for instance names, as shown in Figure 2.

Fig. 2:

uri	lang_de	lang_en	lang_fr	broaderTerms	narrowerTerms	altLabel	closeMatch
http://zbw.eu/stw/descriptor/19664-4	"Finanzmarkt krise"	"Financial crisis"	"crise financière"	<stw/descriptor/10343-6> <stw/thsys/70187> <stw/thsys/71089>	<stw/descriptor/13688-6> <stw/descriptor/18730-1>	"Financial instability" @en "finanzkris e"@de "Krise der Finanzmär kte"@de	http://dbpedia.org /resource/Financial_crisis



Repräsentation

This information can be constantly delivered by a requested web service and be integrated into the search interface of, e.g., a repository for scientific working papers in two ways: by automatically suggesting single search terms while entering characters, and by suggesting related search terms like narrower terms and translations ('altLabel'). In the first case, the effect will be that search terms are proposed to the user that are used for indexing. In the latter case, the effect will be that the query is expanded by controlled terms, hence normally enlarging the result set.

The technical implementation is based on Semantic Web standards, best practices and components like SKOS, RDFa, SPARQL-Endpoints and triple stores (Neubert, 2009). The use of Semantic Web technology is not at all mandatory, but implies some potential for better dissemination and integration of data into third-party applications and environments. We therefore suggested a 'lightweight' approach for integrating the web services in library applications, and as proof-of-concept did this for our

own repository 'EconStor' (Borst, 2011). Since 2009, the terminology web service is online, automatically suggesting search terms from our STW Thesaurus for Economics which includes about 6,000 subject headings and 18,000 entry terms to support individual key words. Besides freely entered search terms, queries in our repository do not only contain terms from STW, but also about 800 terms from the very common JEL classification for Economics which is integrated into static, preconfigured queries ('browsing').

Logfile Analysis as an Approach for Analyzing Search Behaviour

There are several ways to analyze search behaviour in the realm of web applications: logfile analysis is the conventional one, with logfiles generated by web or applications servers and a couple of tools (Hassler, 2010). Other popular approaches are:

- real-time tracking by embedding counter pixels and special tags or JavaScript code into HTML pages, registering each page view (or hits on any other HTML-embedded object) by a remote server. Search queries and keywords are also tracked and counted, but not related to a certain vocabulary;
- questionnaires to learn about users' behaviour, requirements and expectations. This approach reveals more in-depth insights into individual search strategies and preferences, but it is costly and not very much situation-oriented;
- usability studies to investigate users' behaviour in a laboratory on the basis of predefined tasks. This approach is also very costly, but potentially very enlightening with regard to the handling of, e.g., search term suggestion.

Logfiles are automatically generated, for instance, by a web server, and can be persistent and processed at any time by different tools. Filtering of robots, crawlers and other non-human agents is possible, but access through proxy servers or browser caches is not logged. Moreover, in some countries the generation and storage of logfiles is very restricted

because of data privacy rules, for instance by prohibiting IP tracking on an individual level. Nevertheless, we chose logfile analysis, because it is comparatively 'cheap' and on a quantitative level most exhaustive. Our focus is not on individual search patterns or 'user paths', but simply on the role of controlled vocabulary in search queries. For the purpose of our investigations, we took the log entries and processed them through PERL scripts, regular expressions and/or UNIX Shell commands like `grep`, `sed` or `awk`. Moreover, we applied an established linguistic technique by indexing our STW vocabulary via SOLR/Lucene, and by using the stemming from SOLR for mapping uncontrolled terms to STW terms.

Results

What is the current rate of search queries containing controlled vocabulary? Because search terms are suggested, automatically expanded or part of preconfigured queries, one would expect a certain number of occurrences of these terms in queries. And indeed: about one-third from the whole of our repository's queries contain search terms from controlled vocabularies, with a slight majority of STW terms, followed by the more established JEL classification. Term expansion with STW terms is practiced at 7%, with 1% of these queries triggering 'scrolling', meaning that in case of more than 10 results consecutive pages are requested. Scrolling is not performed on any term expansion, but if so, it is done quite exhaustively with an average of 4.1 scrolls and sometimes up to 10 or more scrolls. This may be regarded as an indication that users expect appropriate results from term expansion, although the recall normally increases a lot (Figure 3).

How great is the occurrence of STW terms in non-controlled search queries? This is to get an idea of the potential coverage. Even if a query does not rely on suggested or predefined terms, it may contain terms from a big vocabulary that comprises almost 25,000 terms. The results show that about two-thirds of the free text entries can be mapped to STW terms by means of a SOLR/Lucene index. Eighteen percentage of the entries may not be considered as search terms in a narrower sense, because they originate from other categories: numbers (for instance, ISBN or identifiers from other information systems), names and document titles (Figure 4).

Fig. 3: Controlled terms in search queries.

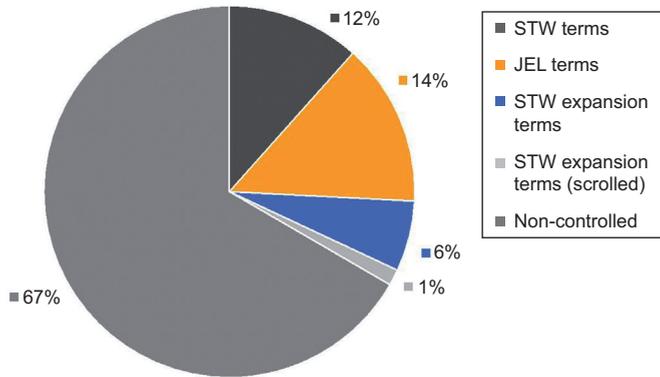
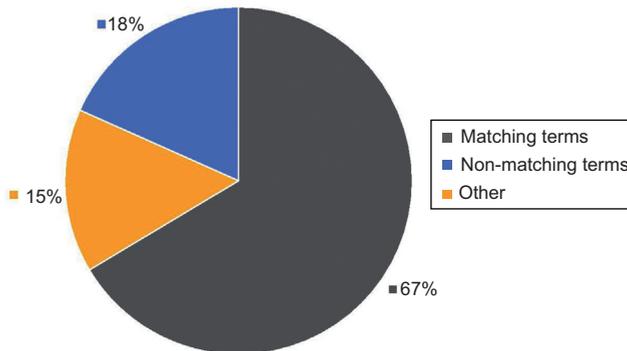


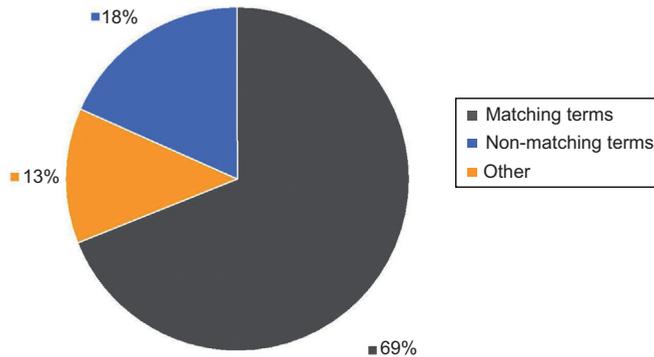
Fig. 4: Occurrence of controlled terms/internal search.



How great is the occurrence of STW terms in Google queries, respectively referrers of document links in our repository? This is a variation of the foregoing investigation. Since the majority of document views and downloads is primarily induced by Google searches, it is obvious to check whether the rate of mapping to STW terms is similar to those of the repository queries. It may come a bit as a surprise that even more terms from Google searches can be mapped to STW terms, where you normally expect more irrelevant and heterogeneous input. But we have to consider that a Google search resulting into a page view of our repository may be preceded by a couple of other Google

searches. Moreover, by indexing both our metadata and documents, the Google search engine may already index most of the terms related to STW (Figure 5).

Fig. 5: Occurrence of controlled terms/Google search.



Conclusions and Suggestions for Improving Search Interfaces of Library Applications

We observed significant use of and potential for controlled vocabularies to support retrieval — if they are big enough and constantly maintained. The mapping from uncontrolled terms still may be improved by linguistic techniques, for instance by lemmatizing in order to embed queries better into document structures and raise the precision of results. In an ideal state, you would expect a total match between uncontrolled search terms and terms from a vocabulary, but in this case the latter would become obsolete. Since the main function of controlled vocabularies is categorizing content by metadata terms which are not a syntactical part of a text, there will always be the requirement of mapping a user's query to these terms and to suggest them. Actually this is already done in our repository by performing term expansion. Nevertheless, a 6% rate of queries with expanded search terms is comparatively low. It may be seen as an indication that search term suggestion is not very much accepted and trusted. On the other hand, once

this option is chosen, the very thorough scrolling of results may be a hint that users rely on the results, hoping to find the relevant documents.

We finally suggest a couple of actions to be taken in order to improve retrieval by means of a controlled vocabulary. These actions mainly affect the user interface, but also imply changes in the modelling and provision of authority data similar to the ones we depicted earlier in the section on terminology web services:

- Search term expansion should be performed in an unobtrusive, but still visible and transparent way. Suggested search terms may be accordingly titled ('Terms which are related to your query and associated with documents') and rendered according to their frequency.
- Too much scrolling of result sets should generally be avoided and tackled by introducing sorting of columns, cascading search and filters.
- Since a significant number of uncontrolled terms belong to other categories like names, numbers and document titles, this should be supported better by the responding information system. For this purpose, we suggest a data infrastructure which is based on authority data from these categories. More concretely, an uncontrolled entry would be simultaneously checked against different sets of authority data, suggesting that the user refine her search in the way of 'Did you mean the person/concept/work/title?' Once she chooses a category, an internal field search would be triggered.

Websites Referred to in the Text

http://en.wikipedia.org/wiki/Financial_crisis

<http://zbw.eu/beta/stw-ws/about>

<http://econstor.eu/>

<http://zbw.eu/stw/versions/latest/about>

http://www.aeaweb.org/jel/jel_class_system.php

References

Borst, T. (2011): 'Improving Library Services with Semantic Web Technology in the Realm of Repository Systems.' In: *Proceedings of the International Conference on Digital Libraries & Knowledge Organization ICDK 2011*. Gurgaon, India.

Hassler, M. (2010): *Web Analytics: Metriken auswerten, Besucherverhalten verstehen, Website optimieren* (Translated Title: Web analytics: evaluating metrics, understanding user behaviour, optimizing webpages). Cambridge, USA.

Lagoze, C. (2010): *Lost Identity: The Assimilation of Digital Libraries into the Web*. Cornell, <http://www.cs.cornell.edu/lagoze/dissertation/dissertation.html>.

Mayr, P. and V. Petras (2008): 'Cross-Concordances: Terminology Mapping and its Effectiveness for Information Retrieval.' In: *74th IFLA World Library and Information Congress*. Québec, Canada, <http://www.ib.hu-berlin.de/~mayr/arbeiten/ifla08-2008-04-15.pdf>.

Neubert, J. (2009): 'Bringing the "Thesaurus for Economics" on to the Web of Linked Data,' In: *Proceedings of the Linked Data on the Web Workshop (LDOW2009)*. Madrid, Spain, http://events.linkedata.org/ldow2009/papers/ldow2009_paper7.pdf.

Petras, V. (2006): *Translating Dialects in Search: Mapping between Specialized Languages of Discourse and Documentary Languages*. Berkeley, <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.134.3444&rep=rep1&type=pdf>.