Enabling Multilingual Search through Controlled Vocabularies: the AGRIS Approach

Fabrizio Celli[1,*], Johannes Keizer[1]

[1] Food and Agriculture Organization of the United Nations, Rome, Italy
{Fabrizio.celli,Johannes.keizer}@fao.org

Abstract. AGRIS is a bibliographic database of scientific publications in the food and agricultural domain. The AGRIS web portal is highly visited, reaching peaks of 350,000 visits/month from more than 200 countries and territories. Considering the variety of AGRIS users, the possibility to support cross-language information retrieval is crucial to improve the usefulness of the web-site. This paper describes a lightweight approach adopted to enable the afore-mentioned feature in the AGRIS system. The proposed approach relies on the adoption of a controlled vocabulary. Furthermore, we discuss how expanding user queries with synonyms increases the sensitivity of a search engine and how we can use a controlled vocabulary to achieve this result.

Keywords: Cross-language Information Retrieval · Controlled Vocabulary · Query Expansion · Search Engine · Digital Repository · Agriculture

1 Introduction

The debate on the usefulness of controlled vocabularies has been carried out for more than two decades [11], [14], [16], [17], [20], and there are still controversial opinions. On the one hand, there are supporters of the theory of abandoning controlled vocabularies [2], [5], but on the other, there are those who sustain that controlled vocabularies are essentials to ensure the right recall when searching in bibliographic databases [7], [21]. The former base their assertion on the evidence that keyword-based searching has become the preferred method of searching in online information systems [11]. Thus, according to them, a textual search is everything users need; there is no value in using controlled vocabularies, but free keywords are enough to help users in retrieving resources from bibliographic databases. However, several studies emphasize that many resources returned in a keyword-based search would be lost without controlled vocabularies. Gross and Taylor [10] sustain that 35.9% of results would not be found if subject headings were removed from catalog records. In fact, subject fields very often contain terms that are not available in titles and abstracts, since expert cataloguers avoid repetitions [13]. In addition to that, controlled vocabularies can mediate the implementation of advanced features, like semantic search in information retrieval systems, as in the case of the European project INSEARCH [1].

In this paper, we show how the adoption of a controlled vocabulary helps in implementing the multilingual search functionality in the AGRIS information system, in order to retrieve multilingual content whose language may be different from the language of the query. In that way, this functionality refers to cross-language information retrieval. Section 2 introduces AGRIS and AGROVOC multilingual controlled vocabulary. Section 3 presents the problem of enabling multilingual search in AGRIS. We discuss a methodology that relies on AGROVOC to expand user queries in order to retrieve resources in different languages. This methodology can be generalized and applied to other systems that make use of a multilingual controlled vocabulary. In section 4, we analyze how expanding user queries with synonyms may help in improving the recall of a search engine. Section 4 is only analytical, since we have not implemented the proposed solution yet. Finally, in the last section we draw our conclusions.

2 An Overview of AGRIS and AGROVOC

Over the last few years, AGRIS has dramatically changed its shape. AGRIS is the International Information System of Agricultural Science and Technology. It was set up in 1974 as an initiative of around 180 member countries of the Food and Agriculture Organization of the United Nations (FAO). The main objective was to improve access and exchange of information on agricultural research serving the information needs of developed and developing countries on a partnership basis. Now, AGRIS ambition is to be a global hub to agricultural research and technology information.

AGRIS is a collection of more than 8 million multilingual bibliographic references, mainly accessible through the AGRIS website[1]. On the data acquisition side, the AGRIS team collects and publishes data from more than 150 partners all over the world. The data ingestion process includes disambiguation of AGRIS entities, deduplication, and semantic enrichment. Then, data are published as machine-readable RDF triples and become freely downloadable through a SPARQL endpoint or FTP. On the data dissemination side, since 2013 the AGRIS website has been completely revamped as a semantic mash-up that uses formal alignments across many systems to provide a universe of data around each bibliographic record. Users can browse the AGRIS core database, looking for information about a topic in the AGRIS domain. When users select a bibliographic resource, the system shows its associated *mashup page*. A mashup page is a web page that displays an AGRIS resource together with relevant knowledge extracted from external data sources (as the World Bank[2], DBPedia[3], and Nature[4]). The availability of external data sources is not under AGRIS control. Thus, if an external data source is temporary unreachable, it is not displayed in AGRIS mashup pages.

1 http://agris.fao.org
2 http://data.worldbank.org/
3 http://wiki.dbpedia.org/
4 http://api.nature.com/

The mediation of AGROVOC[5] makes the generation of mashup pages possible. AGROVOC is a thirty years old multilingual controlled vocabulary containing over 32,000 concepts in 23 languages, and covering all areas of interest of FAO. A community of experts maintains AGROVOC and edits it through VocBench [19], an open source web application for editing SKOS and SKOS-XL thesauri. AGROVOC is aligned with 16 multilingual knowledge organization systems related to agriculture. The AGRIS system relies on those alignments and on the high quality of AGROVOC content to query external web services and interlink AGRIS bibliographic resources to relevant content. In fact, AGRIS records are indexed with AGROVOC descriptors. Sometimes data providers produce records where AGROVOC is already available in their metadata; other times, AGROVOC descriptors are added to AGRIS metadata as a result of the semantic enrichment process. The availability of AGROVOC descriptors in AGRIS metadata represents the backbone for the generation of mashup pages [4].

The AGRIS audience is mainly composed of domain experts, researchers, librarians, information managers, and everyone with an interest in agricultural subjects. According to Google Analytics, every month hundreds of thousands of users from about 200 countries and territories access the system. Considering the high variety of AGRIS users, their needs, and the uniqueness of the AGRIS content, we have the duty to explore new possibilities of usage of AGRIS data. We want to provide AGRIS users with additional features that derive from intrinsic characteristics of AGRIS data. The mediation of AGROVOC controlled vocabulary can be the key of our exploitation of AGRIS data. In another work [4], we have explored the possibility to crawl unstructured web resources, use an automatic indexer to assign AGROVOC descriptors to crawled web resources, and interlink them with AGRIS bibliographic data. In this paper, we show how we can enable multilingual search and other searching features through the usage of AGROVOC controlled vocabulary.

3 Enabling Multilingual Search Using a Controlled Vocabulary

Xian is a Chinese researcher and he wants to retrieve some scientific publications from the AGRIS database. His main research interest is about "rice". Xian performs a keyword-based search using the Chinese keyword ***稻米*** *(which means "rice" in English), but the AGRIS system returns only 14 documents. This result looks strange to Xian, since "rice" is the agricultural commodity with the third-highest worldwide production according to 2013 FAOSTAT data [6]. Thus, Xian is expecting to retrieve a lot of scientific material about this important cereal. He analyzes results and discovers that all of them have Chinese metadata. Xian realizes that he has to query the system in English (and maybe in other languages) to access the international literature. Xian is quite unhappy with AGRIS. He would like to query the system in his native language, which would simplify the choice of further keywords to refine his query, but he would also like to access the international literature.*

[5] http://aims.fao.org/agrovoc

The above paragraph reflects a typical scenario of cross-language information retrieval. In order to understand Xian's problem better, we should provide some background information about the AGRIS default search. When a user queries the system, their query refers to metadata available in the AGRIS database. Thus, if a user searches for 稻米, by default the system returns all bibliographic references containing the word 稻米 in the title, in the abstract, or as a keyword. The problem with this behavior is that the user may be interested in results in all languages or in a subset of them, thus not only in results whose metadata are available in the language of their query. As we show in sections 3.2 and 3.3, a multilingual controlled vocabulary is a valid tool to deal with this scenario. In fact, it can be used to expand user queries by translating keywords in all languages available in the vocabulary. What we want to achieve is to let users searching in their native languages and retrieving scientific publications in all languages.

3.1 Related Work

Several authors have observed that the development of methodologies and tools supporting multilingual information discovery is essential to make non-English content available to end users [8], [12], [15]. Gohrab [8] proposes a framework that performs on-the-fly machine translation of queries and documents. This framework does not rely on controlled vocabularies, but supports automatic translation of queries using external services like "Google Translate[6]" or "Microsoft Translator[7]". It also adopts "OpenMaTrEx[8]" for domain-specific translations. We believe that the usage of a controlled vocabulary for the translation of user queries is important when searching the scientific literature. In fact, it allows searching mediated by concepts, overcoming problems related to synonyms, scientific names, and abbreviations, and increasing the level of precision of the translations. By the way, searching mediated by concepts is still based on words and, in case of polysemy, there is the risk of wrong translations of user queries. Using a domain-specific controlled vocabulary like AGROVOC reduces the impact of polysemy. There is "one sense per discourse" [9]; given a context, there is a high probability that polysemous words are used in a single sense.

Kaplan [12] describes a methodology that uses different lexical resources. The proposed query translator module tries to perform the translation using first term networks, then domain specific controlled vocabularies, and finally a general-purpose query translation service. The software component allows querying in English, French, German, or Swedish, and retrieving results in one or more of those languages. Our approach is based on AGROVOC, a multilingual thesaurus covering 23 languages. We are not only interested in translating the source query, but also in extending the query making use of synonyms in the available languages.

---

6 https://cloud.google.com/translate/

7 https://www.microsoft.com/translator/

8 http://www.openmatrex.org/

3.2 The AGRIS Approach to Multilingual Search

AGRIS approach to multilingual search is based on the adoption of AGROVOC as an instrument to translate user search keywords. In this way, we demonstrate that a controlled vocabulary is not only good for document indexing, but it can be applied to other aspects of information retrieval, as enabling multilingual search through automatic query expansion (AQE). AQE has a 50-year history but, as the survey [3] states, only in recent years it has reached a good level of scientific maturity to lose the status of experimental technique.

We have developed a software component for AGRIS that implements the following algorithm. We call this component the *multilingual query expansion module*. This module is responsible for translations of user keywords, but it does not translate titles or phrases. When a user performs keyword searching in the AGRIS database, the system:

- Identifies the query pattern;
- Uses AGROVOC to translate keywords;
- Expand the user query, boosting keywords provided by the user;
- Returns results in all available languages.

The identification of the query pattern is needed to allow the system to expand the query. In fact, users may perform keyword searching or they may perform structured searching. In the second case, the query presents controlled keywords that must not be translated. As an example, if a user wants to search only in the subject field, they can use the query `subject:rice`, where `subject` is the controlled keyword that tells the system in which bibliographic field the user wants to look for the keyword `rice`. In the same example, `rice` is the keyword that the system has to translate. In addition to that, special characters like '*' and '-' have to be discarded, since they are used by the system to build negative and wildcard queries. The special character "+" can be used to define mandatory keywords.

In the current implementation of the algorithm, we have considered a limited set of query patterns. The system expands the source query if:

- The query contains 4 terms or pictograms, without the Boolean operators "AND" and "OR". We have identified this threshold to distinguish keywords searching from title, serials, and author searching. In fact, a study conducted in 2001 [18] affirms that the average length of a search query is 2.4 terms. Furthermore, in March 2016, the average length of a search query in AGRIS was 4.7, but longest queries referred to titles or authors. This parameter affects the retrieval performance, since it can cause very long expanded queries.
- The query has pattern `subject:($keywords)`, `+subject:($keywords)`, `subject:$keywords`, or `+subject:$keywords`[9]. It is the case of searching only in the subject field.

[9] `$keywords` stands for a list of terms or pictograms satisfying constraints expressed in the previous bullet point

The implementation relies on two Apache Solr indexes:

- AGROVOC label index. This index contains all concepts available in the AGROVOC thesaurus. For each concept identified by a URI, the index stores preferred and alternative labels in all languages.
- AGRIS core index, which contains all AGRIS resources. This is the main index used by the AGRIS website to retrieve records after the submission of a user query.

Once the system has identified the query pattern, the *multilingual query expansion module* queries the AGROVOC label index to obtain translations of source keywords. The module matches source keywords against both preferred and alternative labels to identify the AGROVOC concept, but it considers only preferred labels as output of the translation process. In fact, as we show in section 4, alternative labels can mediate query expansion with synonyms. After that, the module expands the source query by building a union of source keywords and their translations. The system boosts source keywords by a factor of 50, since we think that it is important to return to users results of their original query first, and then results of the multilingual query. As an example, if the source query is `+subject:rice`, the system builds the query:

```
+(subject:"rice"^50 OR subject:("चावल" OR "Reis" OR "рис
(зерно)" OR "ငဲ့ါ" OR "벼" OR "Arroz" OR "Riso" OR "Riz" OR
"rizs" OR "稻米" OR "rýže" OR "أر ز" OR "ข้าว" OR "米" OR
"ryža" OR "بـ رنـ ج" OR "Ryż (ziarno)" OR "pirinç"))
```

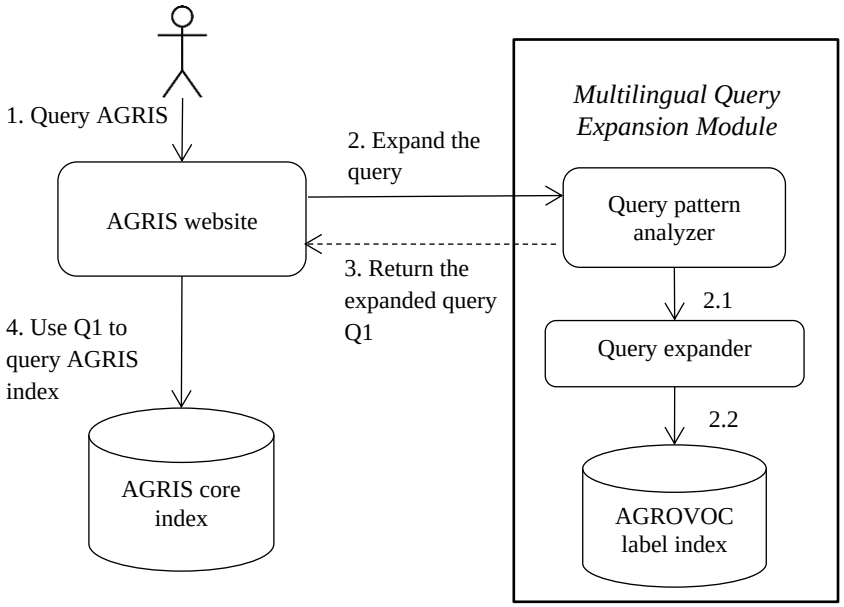

Fig. 1. Workflow of the multilingual query expansion module

As depicted in Figure 1, after query expansion, the system queries the AGRIS core index using the expanded query Q1. The AGRIS website displays results in all languages, boosting results coming from the original query. Overall, the user is not aware that the system has modified their query. In fact, the system never shows the expanded query to the user.

3.3 Analysis of Results

Here follows a sequel to the scenario introduced at the beginning of section 3. *The Chinese researcher Xian has just discovered that AGRIS has implemented the multilingual search functionality. Xian queries the system using the keyword 稻米. The system returns only 14 results, since only 14 AGRIS records contain the keyword 稻米 in title, abstract, or subject field. Xian clicks on the button to enable the multilingual search, and the system returns 166,639 results*[10]*. The new set of results is composed of bibliographic references about the concept "rice", but only 14 of them contains the Chinese word 稻米 actually. Now Xian has a lot of material to analyze and he can apply filters to make his query more specific.*

There is also another interesting scenario to take into account. It concerns the absence of results after searching in a specific language. Let us consider an Indian user who queries the AGRIS system using the Hindi keyword "फसलें" (which means "crops" in English). The system returns zero results. It means that there are no resources in the AGRIS database containing the word "फसलें" in title, abstract, or subject field. Enabling the multilingual search, the user gets access to 474,854 scientific papers in several languages. Unfortunately, the result set does not contain metadata in Hindi, since the AGRIS Indian data provider only produces metadata in English. Be that as it may, our user is now able to query the system in their native language and access scientific literature even when there are no publications in that language.

Query ID	**Source query**	*English translation*[11]	**Number of results**	**Number of results of multilingual search**
Q1	稻米	*rice*	14	166,639
Q2	फसलें	*crops*	0	474,854
Q3	latte	*milk*	8,019	189,475
Q4	Klimaänderung	*climate change*	23	31,028
Q5	"su muhafazası"	*water conservation*	22	15,285
Q6	إذ تظام حراري ل ل تربة	*soil thermal regimes*	21	368
Q7	"forest mensuration"	*forest mensuration*	3,679	3,930

[10] http://agris.fao.org/agris-search/searchIndex.do?enableField=Enable&query=%E7%A8%BB%E7%B1%B3

[11] This column helps in making the source query more comprehensible to the reader

Table 1. Comparison of number of results before and after enabling the multilingual search in the AGRIS website

Table 1 shows the effectiveness of the multilingual search feature, comparing the number of results before and after enabling this functionality in the AGRIS website. Correctness of results depends on the correctness of the AGROVOC thesaurus and AGRIS metadata. A community of domain experts from different countries contributes to the quality and correctness of labels available in AGROVOC. Thus, the multilingual translation based on AGROVOC is highly reliable as far as the agricultural domain concerns. Results of a multilingual search on AGRIS are composed of the union of results of several monolingual searches. The main disadvantage is that there could be too many results; this is why advanced filters are essential to allow AGRIS users to refine their queries reducing the number of results.

Query ID	**Execution time**	*Time for query expansion*	*Execution time of the expanded query*	**Total multilingual search execution time**
Q1	350ms	*25ms*	*390ms*	415ms
Q2	340ms	*30ms*	*370ms*	400ms
Q3	360ms	*20ms*	*400ms*	420ms
Q4	400ms	*30ms*	*430ms*	460ms
Q5	390ms	*25ms*	*440ms*	465ms
Q6	370ms	*30ms*	*430ms*	460ms
Q7	360ms	*30ms*	*390ms*	420ms

Table 2. Comparison of execution time (in milliseconds) before and after enabling the multilingual search in the AGRIS website

Table 2 compares the execution time of the default search with the execution time of the multilingual search. The execution time of the multilingual search is composed of two parts:

- The time to expand the user query with translations. This is the time that the multilingual query expansion module needs to translate the source query.
- The time to execute the expanded query. This is longer than the time to execute the source query, since the expanded query contains more terms.

On average, the execution of multilingual search requires 68.75 milliseconds more than the default search in our implementation. This delay is highly acceptable.

An analysis of usage of this functionality show that 2% of AGRIS active users enable the multilingual search. Let us focus on the expression “active users” to define it better. According to Google Analytics, in March 2016 AGRIS received around 513,000 unique page views. 80% of them come from Google.com and Google Scholar, while 20% of them represent activity of users in the AGRIS website. The latter percentage is about users who rely on AGRIS to actively search for scientific literature, i.e. the “active users”; on the other hand, users coming from Google access a bibliographic record directly, without using the AGRIS website search feature. On

average, among active users, 2% of them enable multilingual search. This number is quite satisfying, since multilingual search is an advanced functionality and we expect a small percentage of usage. In addition to that, the multilingual search is a new AGRIS functionality and it needs time to reach the public. It is highly likely that the percentage will increase over the time and after we will promote the multilingual search in webinars and events.

In order to improve the multilingual search usefulness, we have to explore possible extensions. First, we should allow users to select a subset of languages when enabling the functionality. In fact, it may well be that a user wants to retrieve results only in a couple of languages and not in all possible languages. Second, we need to solve the problem of singular/plurals, abbreviations, and misspellings. Currently the translation of the user query relies on exact match of strings; if AGROVOC contains a keyword in the exact way the user has written it, the system can translate the keyword in all available languages. At present, the system can manage singular/plural variations only for English terms. Finally, we have to explore the possibility to expand user queries to synonyms. In this way, the system can increase the recall, including all resources about the same topic in a specific language. The combination of this functionality with the multilingual search option is a valuable tool for end users, as we argue in the next section.

4 The Synonyms Problem with Recall

Methodologies for multilingual information retrieval through controlled vocabularies can be applied to different scenarios. In the previous section, we have described the implementation of the translation of user keywords through AGROVOC. In this way, users can access the AGRIS scientific literature in any languages, querying the system in any languages too. There are other situations where we can adopt the same methodology. For instance, it can be used to implement the expansion of user queries with synonyms in a given language, or even in all languages.

We can consider the following example. Peanut is a crop that mainly grows in the tropics and subtropics. People with a general background usually know this crop by the name of *Peanut*, but people in the field know it as *Groundnut* (this is the technical name, while the scientific name is *Arachis hypogaea*). If an AGRIS user queries the system with the keyword *Peanut*, the system returns only results containing such keyword in their metadata, but not results containing the keyword *Groundnut*. A search mediated by concepts returns all results related to the crop Peanut, containing both *Peanut* and *Groundnut* in the metadata.

We can apply the methodology described in section 3.2 to implement this behavior. The *Synonyms Query Expansion Module* works exactly as the Multilingual Query Expansion Module described in Figure 1. The new module identifies the query pattern and then uses the AGROVOC index to retrieve all synonyms of a keyword in a given language. The difference is that now the expansion module includes both preferred and alternative labels in a specific language as output of the translation process, while the Multilingual Query Expansion Module considers only preferred labels in all lan-

guages. In fact, the union of preferred and alternative labels in a language compose the set of available synonyms for that language.

We can further extend this process by combining the Synonyms Query Expansion Module and the Multilingual Query Expansion Module. If the user looking for *Peanut* enables the synonyms retrieval, they obtain results including also the keyword *Groundnut*. If the user also enables the multilingual retrieval, they obtain results in all languages, including synonyms for each different language. This is a very important step, since it is not a mere translation of strings. Different languages may have different synonyms, which are not the direct translation of one another. Relying on a controlled vocabulary like AGROVOC solves this issue. In fact, we do not translate the main keyword and all its synonyms in other languages, but we search for the AGROVOC concept, and we extract all preferred and alternative labels of the concept for all the available languages.

Even if we have not implemented the Synonyms Query Expansion Module yet, we can provide some numbers to demonstrate its power. We can use the AGRIS website to simulate the synonyms expansion manually. We start with three fulltext queries, using the keywords *Peanuts*, *Groundnuts*, and their combination. This is the number of results:

1. `Groundnuts`: 2,824 results
2. `Peanuts`: 6,750 results
3. `Peanuts OR Groundnuts`: 9,222 results

The third query is exactly the query that the *Synonyms Query Expansion Module* would generate. As we can see, enabling the synonyms retrieval improves the recall. Another observation is that the sum of results of the first two queries is 9,574 while the synonyms expansion returns 352 results less; this means that results sets 1 and 2 have a small overlapping, because 352 AGRIS records contain both *Peanut* and *Groundnut* in their metadata.

Now, we simulate the combination of the synonyms expansion with the multilingual search. We enable the multilingual search for the query `Groundnuts`, and then for the synonyms expanded query `Peanuts OR Groundnuts`:

4. `Groundnuts` (multilingual query): 4,713 results
5. `Peanut OR Groundnut`[12] (multilingual query): 10,842 results

The fourth query shows that the multilingual retrieval allows obtaining 1,889 records more than the default query with the keyword *Groundnuts* (query number 1). On the other hand, the combination of synonyms expansion and multilingual query (query number 5) returns 6,129 results more than the multilingual expansion of the keyword *Groundnuts* (query number 4) and only 1,620 results more than the synonyms expansion performed by the third query.

[12] We have manually built the multilingual expansion for this query. In fact, the current system does not recognize a query pattern including the OR operator, as we have explained in section 3.2. We have executed a union of the two expanded queries generated for the keywords *Peanuts* and *Groundnuts*.

The major impact of the synonyms expansion with respect to the multilingual one is due to the fact the AGRIS has an high coverage of English metadata, thus synonyms has more impact than translations when the source keyword is in English. AGRIS resources cover 64 languages, and there is a coverage of at least 10,000 resources for 28 languages. However, many data providers translate metadata also into English. For example, Chinese resources have titles, abstracts, and keywords both in English and in Chinese.

5 Conclusions

Multilingual search and synonyms expansion have a profound impact on searching in online repositories. While multilingual search allows users to search in their native language and to retrieve documents in several languages, expanding user queries with synonyms allows retrieving more resources about the given topic. In this paper, we have proposed a methodology that relies on a controlled vocabulary to implement the aforementioned features. We have implemented the methodology in the AGRIS website to enable the multilingual search; our implementation has required AGROVOC controlled vocabulary and a software component that detects query patterns and translates a query through AGROVOC. We have also discussed how the same methodology can be adopted to expand user queries with synonyms. Experimental results demonstrate significant improvements of recall in both cases. The high amount of retrieved resources can be reduced by an effective advanced search that helps users in refining and filtering out results.

The current implementation in the AGRIS system can be improved. First, we have to provide the possibility to select a subset of languages when enabling the multilingual search. This feature would allow users to retrieve results only in their favorite languages, reducing the number of undesired results. Second, we have to implement the Synonyms Query Expansion Module. Finally, we have to work on homonyms and variations of keywords, like abbreviations and misspellings, especially for non-Latin characters.

There are also further scenarios to explore. As future work, it would be useful to study which additional expansions of queries can be useful to users. For instance, a controlled vocabulary like AGROVOC allows generalizing or restricting the topic of a query by navigating the hierarchy of concepts. Even more useful would be a system that automatically performs different query expansions and combinations of them, presenting to end users alternative subsets of results. In this case, users can select the desired result set by considering the number of results and the specific mechanism under the retrieval of different result sets.

6 References

1. Basili, R., Stellato, A., Daniele, P., Salvatore, P., & Wurzer, J. (2012, September). Innovation-related enterprise semantic search: the INSEARCH experience. In Semantic Computing (ICSC), 2012 IEEE Sixth International Conference on (pp. 194-201). IEEE.

2. Bibliographic Services Task Force of the University of California Libraries (2005). Rethinking How We Provide Bibliographic Services for the University of California: Final Report.
3. Carpineto, C., & Romano, G. (2012). A survey of automatic query expansion in information retrieval. ACM Computing Surveys (CSUR), 44(1), 1.
4. Celli, F., Keizer, J., Jaques, Y., Konstantopoulos, S., & Vudragović, D. (2015). Discovering, Indexing and Interlinking Information Resources. F1000Research.
5. Deanna B. Marcum (2005). The Future of Cataloging: Address to the Ebsco Leadership Seminar. Boston, Massachusetts.
6. FAOSTAT Food and Agriculture commodities production, http://faostat3.fao.org/browse/rankings/commodities_by_regions/E
7. Gardner, S. A. (2008). The changing landscape of contemporary cataloging. Cataloging & Classification Quarterly, 45(4), 81-99.
8. Ghorab, M. R., Leveling, J., Lawless, S., O'Connor, A., Zhou, D., Jones, G. J., & Wade, V. (2011). Multilingual adaptive search for digital libraries. In Research and Advanced Technology for Digital Libraries (pp. 244-251). Springer Berlin Heidelberg.
9. Gale, W. A., Church, K. W., & Yarowsky, D. (1992, February). One sense per discourse. In Proceedings of the workshop on Speech and Natural Language (pp. 233-237). Association for Computational Linguistics.
10. Gross, T., & Taylor, A. G. (2005). What have we got to lose? The effect of controlled vocabulary on keyword searching results. College & Research Libraries, 66(3), 212-230.
11. Gross, T., Taylor, A. G., & Joudrey, D. N. (2015). Still a lot to lose: the role of controlled vocabulary in keyword searching. Cataloging & Classification Quarterly, 53(1), 1-39.
12. Kaplan, A., Sándor, Á., Severiens, T., & Vorndran, A. (2014). Finding Quality: A Multilingual Search Engine for Educational Research. In Assessing Quality in European Educational Research (pp. 22-30). Springer Fachmedien Wiesbaden.
13. Lu, C., Park, J. R., & Hu, X. (2010). User tags versus expert-assigned subject terms: A comparison of LibraryThing tags and Library of Congress Subject Headings. Journal of information science, 36(6), 763-779.
14. McCutcheon, S. (2009). Keyword vs controlled vocabulary searching: the one with the most tools wins. The Indexer, 27(2), 62-65.
15. Peters, C., Braschler, M., & Clough, P. (2012). Cross-Language Information Retrieval. In Multilingual Information Retrieval (pp. 57-84). Springer Berlin Heidelberg.
16. Rowley, J. (1994). The controlled versus natural indexing languages debate revisited: a perspective on information retrieval practice and research. Journal of information science, 20(2), 108-118.
17. Scicluna, R. (2015). Should libraries discontinue using and maintaining controlled subject vocabularies?.
18. Spink, A., Wolfram, D., Jansen, M. B., & Saracevic, T. (2001). Searching the web: The public and their queries. Journal of the American society for information science and technology, 52(3), 226-234.
19. Stellato, A., Rajbhandari, S., Turbati, A., Fiorelli, M., Caracciolo, C., Lorenzetti, T., ... & Pazienza, M. T. (2015). VocBench: a web application for collaborative development of multilingual thesauri. In The Semantic Web. Latest Advances and New Domains (pp. 38-53). Springer International Publishing.
20. Voorbij, H. J. (1998). Title keywords and subject descriptors: A comparison of subject search entries of books in the humanities and social sciences. Journal of documentation, 54(4), 466-476.
21. Zavalina, O.L. (2010). Collection-Level Subject Access in Aggregations of Digital Collections: Metadata Application and Use. PhD dissertation, University of Illinois at Urbana-Champaign.