The construction of a bilingual Hebrew-English thesaurus for archives in Israel based on information and knowledge presentation technology / Assaf Tractinsky

Abstract

During the last three decades the importance of accessing Archival databases according to the Subject has been increasingly recognized. This is especially true since the tools that libraries usually provide lack the necessary depth and are found inadequate for archival search. Although the subject approach for archival search was recognized, the issue of information retrieval was hardly discussed in the archival research field until the mid1990s. The first to study this subject in the archival world was Richard Lytle (l980), followed by Fernanda Riberio (I996).

In his research Lytle defined two methods for searching archives. The first method is based on the principal of its archival origin, meaning a search according to its archival arrangement and description. The second method is based on the subject found in the archival finding aides, meaning a search using controlled vocabularies. He concluded that by themselves each of the methods was not sufficient and recommended using both methods in collaboration. Ribeiro compared two methods for information retrieval, free retrieval and controlled vocabularies. Her recommendation was also to combine both search methods in Archival information retrieval systems.

In Israel, only two thesauri were developed for archives that were aimed at serving special areas of the information system: The first for the social security system and the other for the Dr. Zerah Warhaftig Institute for the Research on Religious Zionism. Other archives use thesauri that were developed for libraries and adjusted to their own use. Many archives in Israel hold material from several common areas such as, Judaism, Zionism and Holocaust, which may facilitate the creation of key terms for common retrieval. Therefore, considering information retrieval for Israeli archives, it is only natural to construct a bilingual thesaurus. The Bilingual thesaurus will enable users who do not command one of languages to use the archive with their own language. In addition, a bilingual thesaurus will improve the information retrieval within the archives and in the archival information networks.

Research goals and their application

There are many definitions for "thesaurus". In the current research the term refers to a collection of controlled and methodological terms of the indexing language that include hierarchical and associative connections between them (Aitchison, Gilchrist, and Bawden, 2001, p. 1). One of the most prominent examples for a multi lingual thesaurus it the UNESCO thesaurus (2002), originally written in English and later translated into several other languages. The efficiency of the search using the bilingual thesaurus constructed for the use of Israeli archives was tested on an information retrieval system, and its performance was compared to a free text search. For reaching its goals the current study examined the following questions.

1. Can a bilingual thesaurus that fits all the Israeli archives be constructed?
2. What would be the parameters of such a thesaurus, as for its subject and structure?
3. Would the bilingual thesaurus improve the information retrieval in an archival system?

For studying these questions the research was divided into four parts:

1. The first part included a theoretic-research literature review for studying the methodologies and practices involved in the wide variety of. Subjects that are related to thesaurus creation and their application to informational systems and archival systems. The review focused on information systems and classification methods, information retrieval, indexing and abstracting, arranging the description and archival information retrieval while examining the theoretical elements that can contribute knowledge to the goals of our research, the construction of a thesaurus and preparing the experiment.
2. The second part included a survey of Israeli archives. The survey was performed in thirteen archives. in nine archives both the institutes and their internet sites were examined. In the remaining four only their internet sites were examined for the purpose of backing up the results and adding more terms when necessary_ The content of the archives and the description system they use were surveyed and also their compatibility with the ICA-ISAD(G) standard. The survey also examined the existence of controlled vocabularies and their degree of use, The Purpose of the survey was to determine whether the contents, the description systems and the controlled vocabularies of the archives allow the creation of a common thesaurus for Israeli archives.
3. The third part included the construction of a bilingual thesaurus model for Israeli archives. The purpose was to determine theoretical and practical principals for the Creation of a bilingual thesaurus by using international standards. A full thesaurus skeleton was constructed, and two specific micro thesauri for Eretz Israel (Palestine) and Zionism and the State of Israel and Zionism were fully developed These two subjects are not developed in the UNESCO thesaurus, but constitute the main content of the Israeli archives. The remainder of the subjects in the UNESCO thesaurus fit the Israeli archives. It is important to note that the thesaurus was constructed manually and not by machines that are unable to reach the required quality level.
4. The fourth part included an experiment in information retrieval that included an archival database and a thesaurus system using the two micro thesauri developed in the framework of this research. The database included only textual records, most in Hebrew and some were bilingual, English and Hebrew. The two micro thesauri also served as a basis for indexing the database's records. No difference was found in the precision of the search results when comparing the free text search and the thesaurus search methods. All other measures indicated that the two methods complement each other. The results also indicated that if the material is described according to international standards, it is difficult to expect a reasonable simultaneous retrieval of all the levels. The research results suggested that the difference between free text and thesaurus searches is small both in precision and relative retrieval.

Several measures were examined for comparing the search methods. The precision measure showed little difference between the free (62.67%) and the thesaurus (60.75%) search. In the relative retrieval measure the thesaurus search scored 22.39% and the free search 17.14%. Additional measures showed that the thesaurus search retrieved more records while using less searches than the free search. This result suggests that the thesaurus search was more efficient than the free search. The number of searches that retrieved no results was smaller when using a thesaurus search when compared to the free search. This measure too points to the advantage of "Sing a

thesaurus search method. In addition, while examining the search strategies it was found that users of both methods preferred Boolean search operators.
The search in the various archival levels indicates that a search in a certain level does not necessarily fit other levels. This can be the result of the different description systems used for the different archival levels, meaning that the indexing and abstracting of a certain level does not fit other levels. The conclusion is that the vocabulary has to be adjusted to the specific level. On the fonds level terms broader meaning have to be used while on the file level a more specific vocabulary. These level specific vocabularies will of course affect the indexing. Consequently, an important conclusion that stems from the research is that it may be necessary to develop a flexible thesaurus that allows the descriptors' hierarchy levels to be changed according to the archival database. This way, it will be possible to index the hierarchy levels in different databases without the need to build a new thesaurus. This issue requires further research.

Conclusions

The current research indicates that it is advisable to include a thesaurus in the process of describing and retrieving in an archival information system, in combination with free text search. However, clear rules have to be determined for archival indexing and abstracting for all these components to allow efficient information retrieval. In addition, it was found difficult to retrieve all the archival levels by using a system organized by the International standards. However, in order to resolve this issue additional research will have to be conducted.

System No.
002367871