# M1.1 – Cross-lingual Information Retrieval, Usage Scenarios and PubPsych Structure

Andreas Lüschow[1]

[1]Leibniz Institute for Psychology Information

– v1.0 –
January 2019

**Abstract**

This document describes our motivation and the initital situation at the beginning of the CLUBS project. We explain common issues of multilingual systems and why a multilingual approach for the psychological search engine PubPsych is useful and promising but also which challenges exist in cross-lingual systems. Characteristical user inputs as well as probable usage scenarios are presented. We describe the structure of PubPsych, including currently present problems and necessary adaptations to the system during the project.

# Contents

# 1 Introduction

The aim of the CLUBS project is an empirical evaluation of four different approaches in the field of "Cross-lingual information retrieval" based on the psychological search engine PubPsych.

If written information only exists in a language incomprehensible to the reader, it is often not retrievable or usable for him. This problem poses a particular challenge for technical and scientific publications, with yearly thousands of potentially relevant publications in various languages.

For the field of psychology, the ZPID and various partners have thus developed *PubPsych*[1] in 2013, a vertical search engine for psychology literature, tests, treatment schemes and research data. All information in PubPsych is provided in at least one of the four languages English, Spanish, French and German. However, none of the data sets is available in all of these languages.

With the help of the CLUBS project there will now be an empirical evaluation of four different approaches in order to improve cross-lingual search in bibliographic metadata. This will be done, by way of example, on the basis of the PubPsych platform, to enhance the retrieval with foreign-language results.

This document is organized as follows: First, we describe why a multilingual approach for the psychological search engine PubPsych is useful and promising but also which challenges exists in cross-lingual systems. Characteristical user inputs as well as probable usage scenarios are then presented in Section 3. In Section 4, we describe the structure of PubPsych, including problems faced and necessary adaptations to the system, and a short note concerning interface design concludes the document.

# 2 Cross-lingual Information Retrieval (CLIR)

*PubPsych*, a vertical search engine for literature, tests, treatment programs and research data in the psychological domain, offers multilingual content in four languages (English, Spanish, French, German) and all its metadata is available in at least one of these languages. Nevertheless, the underlying document can still be in a different language (e.g., a Dutch document that has an English abstract). More than 50 languages are available in PubPsych.

Search engine users typically use only their preferred language to look for documents in a database, i.e., they do not translate their query and search again for other documents not found by their previous query. This preferred language does not necessarily have to be the user's first language, though. In fact, many researchers already use English search terms to generate a result list as comprehensive as possible, since international publications usually are written in English.

Finding incomplete search results has a negative effect on research, because the missing of important information can for example lead to duplicate research efforts. Especially in the field of psychology, basing research on partial information bears the risk of drawing conclusions on narrow subpopulations [5] or of needlessly testing humans.

To overcome language barriers and to give users access to these variety of documents, it is necessary to allow users to specify their information needs in their preferred language while still retrieving relevant documents in other languages. This well-known issue still is a complex research topic [2], which is the reason why hardly any digital library or search

[1] https://www.pubpsych.eu

engine offers such a functionality. For instance, the verbal description of key concepts in psychology is extremly language dependent and a translation therefore is difficult [6].

These considerations were affirmed by a survey in 2008, in which psychology researchers considered native language information helpful for access to documents and research findings [8]. In particular, PubPsych users in 2015 stated that the possibility of multilingual search would improve the search engine and their experience with it [9].

To make matters worse, since different databases usually index different information and additionally often use different syntax for querying, users that try to accomplish an exhaustive research need to know how to handle these varying resources. The quite expensive licensing costs add to this problem and as a result, many institutions can not afford granting access to important resources. Thus, users prefer easily available one-stop solutions with well-known and intuitively usable interfaces [4]. Providing such a service is not only crucial for the users' satisfaction but eventually for the success of scientific research.

## 3 Usage Scenarios

In this project, we need to define certain expectations concerning typical PubPsych users and their input. These assumptions are partially based on experience and on an analysis of previous PubPsych log files.

Most visitors that use our service come from German-speaking countries, followed by France, Spain, and English-speaking countries. Therefore, many queries will be in German, albeit a significant amount of users already use English queries to find more documents.

We can distinguish three types of queries that are used to find information in a search engine (following [1] and refined by [7]):

- *Informational* queries: looking up topics, specific authors, places, etc.
- *Navigational* queries: finding publications based on identifiers (e.g., ISSN or DOI).
- *Transactional* queries: interacting with the website, e.g., downloading PDF files or buying a product in a web shop.

A clear majority of queries will be informational since we offer no transactional services for the documents. Of course, people can export and download result sets, but this is not something that they are searching for in our database.

Moreover, many queries will use technical terminology because many users are used to psychology terminology and querying databases. Examples are searches for a specific treatment program or a certain author.

Informational queries comprise semantically meaningful terms that normally can be translated. In contrast, navigational queries do not need to be translated since they often are made up of non-linguistic content, e.g. numeric identifiers.

However, there will remain many documents that can not be accessed by using English keywords (estimated 20 %) and thus will hardly be found. Our approach in translating queries and/or database entries should therefore allow for a significant improval of the retrieving process.

# 4 PubPsych Structure

## 4.1 Basic Structure

*PubPsych* comprised more than 950,000 references at the beginning of the project in 2016. These metadata are aggregated by using various sources from data providers in different countries and thus different languages: PSYNDEX, PsychData, PsychOpen (Germany), PASCAL (France), ISOC-Psicología (Spain), NORART (Norway), MEDLINE, ERIC (USA), and NARCIS (The Netherlands). Metadata from these sources are processed, refined and converted before ingestion into the PubPsych index. For example, MEDLINE documents are already enriched with German translations of English keywords to facilitate the retrieval of multilingual results. Hence, some fields of the search engine's index are already able to hold data in different languages and to support multilingualism.

PubPsych's backend is based on the public Apache Solr and Lucene projects [3] and the frontend, written in the Java programming language, therefore allows for different search tasks. Common tasks like search by keyword or using an advanced search with multiple, specialized fields are available. Results lists can be narrowed with a variety of filters and field searches and they can be refined by facets. Many terms (such as author names or keywords) in the bibliographic document descriptions can directly be used to start another search with these terms. Additionally, results can be exported or saved in a personal list and a RSS feed is provided. For each entry there is a link to check the availability of the document and if applicable, links to full texts or further information are displayed.

The interface can be used in all four languages; other than that, no additional features for multilingualism (e.g., showing abstracts according to the user's interface language) are implemented.

## 4.2 Problems and Improvements

The PubPsych system at the start of the project was not up to date (technically and conceptually) and needed major adaptations to be useful in a multilingual context.

**(Meta)Data.** A full list of all currently available metadata fields from then index needs to be prepared so that it is clearly documented what data are provided by which source database and how this information is processed, manipulated, converted, and finally stored in the Lucene index. Some errors that are not present in the source data but in PubPsych's index need to be removed and the affected data needs to be corrected. This concerns e.g. encoding errors like double encoded HTML entities.

Additionally, all datasets need to be addressed in a unique way, i.e., by a unique record identifier, that also facilitates the detection of duplicates in data from different sources. Other inconsistencies, e.g. concerning the heterogeneous use of subtitle fields, also need to be looked at.

**Backend.** After splitting the PubPsych source code into two main parts – frontend and backend – each part can be reworked individually. This allows to apply the MVC (model-view-controller) model of software development to the code base and simplifies the separation of associated code parts.

For instance, some query processing is currently made in the frontend and user input is directly passed to the Lucene index. These parts have to be shifted to the backend logic of the application. Deprecated code fragments need to be updated, replaced, or removed. Index fields not used in the software anymore can be removed; other fields that are necessary or at least helpful for PubPsych need to be added. If these fields should be searchable, this has to be implemented as well; e.g., it is currently not possible to search for specific documents using an ISO language code.

Publication types are always in English, thus they can not be found using other languages. Some index fields should have a higher influence on the relevance of a document than others, which at the moment is not the case. Ideally, this weighting takes multilingual aspects into account, e.g., if a document has an original title and an automatically translated title, the latter should have less influence on the ranking since it may be not as accurate as the original one.

**Frontend.** Changes in the backend that affect the presentation of results or search options need to be reflected in the frontend as well. New searchable index fields and filtering options have to be added to the user interface. Translated content needs to be presented in a reasonable way. This affects for example translated titles and abstracts. Currently, all abstracts are shown in English, even if a manually translated abstract in the user's interface language exists.

Finally, by implementing web tracking code, it should be possible to document and analyze the usage of PubPsych and its functions in the future. By this we can for example document how users access PubPsych, where users come from, which functionalities are used frequently and how users interact with the interface of the search engine.

## 5 Interface Design

In addition to conceptual, logical and structural changes in the PubPsych code base that allow for integration of machine translated content into the search index, the presentation of multilingual information is also a main factor for the success and usefulness of such a system. Even if perfectly translated content is available, it is of no use to researchers or students if they can not search for this data or understand its presentation. Therefore, the exploration of different presentation formats and the evaluation of diverse usage scenarios is a crucial aspect.

However, this project focuses on creating, implementing and evaluating different machine translation approaches for search engine contents. We can not thoroughly address research questions concerning the interface design since this would require a closer look into this research area which can not be done within the scope of the CLUBS project. Nevertheless, we are aware that all effort in producing high quality translated data is of hardly any use if it is not well integrated in the user interface and the application logic.

## 6 Conclusions

The search engine PubPsych, as it currently stands, needs to undergo some major improvements to be able to ingest and handle multilingual data. This includes upgrading backend functionalities (search index, query processing, etc.) as well as allowing the user to take advantage of these functionalities in the frontend (searching, filtering, etc.).

After finishing these adaptations, we will be able to implement the machine translation systems developed during the course of the CLUBS project and to finally evaluate the impact of the enhanced functionalities on retrieval results.

Determining the best approach in translating queries and/or database content will allow for PubPsych users not only to enhance their retrieval experience but also to improve the generated search result lists by incoporating documents in languages other than the query language.

We expect that this approach is not just limited to the field of psychology information systems because we do not establish domain-specific assumptions. We rather use techniques for the translation of queries and index terms and their integration into the software that can be adapted in different domains and aid infrastructure institutions in improving their retrieval services.

## References

[1] Andrei Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.

[2] Anne Diekema. Multilinguality in the digital library: A review. *The Electronic Library*, 30(2):165–181, 2012.

[3] The Apache Software Foundation. Solr. 2017.

[4] Marti Hearst. *Search user interfaces.* Cambridge University Press, 2009.

[5] Joseph Henrich, Steven J. Heine, and Ara Norenzayan. Most people are not WEIRD. *Nature*, 466(29), 2010.

[6] Hans-Joachim Kornadt, Gisela Trommsdorff, and Ryozo B. Kobayashi. "Mein Hund hat mich bestorben" : sprachlicher Ausdruck von Gefühlen im deutsch-japanischen Vergleich. In Hans-Joachim Kornadt, editor, *Sprache und Kognition : Perspektiven moderner Sprachpsychologie*, pages 233–250. Spektrum Akad. Verl., Heidelberg, 1994.

[7] Xinyi Li, Bob J.A. Schijvenaars, and Maarten de Rijke. Investigating queries and search failures in academic search. *Information Processing & Management*, 53(3):666 – 683, 2017.

[8] Martin Uhl. Survey on european psychology publication issues. *Psychology Science Quarterly*, 51(1):19–26, 2009.

[9] Sandra Waeldin. Results from the PubPsych launch survey: Short report. *ZPID Science Information Online*, 15(2):3, 2015.