Datenbank-Spektrum manuscript No.
(will be inserted by the editor)

PubPsych: A powerful research tool providing access to a broad supranational body of psychological knowledge

Erich Weichselgartner · Christiane Baier · Roland Ramthun

Author version for publication in Datenbank Spektrum.
The final publication is available at `http://link.springer.com/article/10.1007/s13222-016-0244-3`
DOI: 10.1007/s13222-016-0244-3

Abstract In the scientific domain, millions of research papers in many different languages are published throughout the globe every year. This causes not only information overload, but because of the language barrier, researchers might not understand or even find articles that could be relevant for their work. Therefore, in the field of psychology, the Leibniz Institute for Psychology Information (ZPID) and its partners have created PubPsych (`https://www.pubpsych.eu/`), an open access vertical search engine for psychological literature, tests, treatment programs and research data (metadata, not the content itself). We discuss the motivation for creating the system, design decisions, connected problems and our solutions, especially concerning multilingual information.

Keywords Psychology database · Multilingual access · Domain-specific vertical search engine

1 Introduction

In an ever growing amount of research publications every year throughout the globe, it is difficult for scientists and professionals to keep up with important innovations in their field. Due to the globalization of science and technology, scientific output is not restricted to traditional institutions in the industrialized West any more. In the scientific domain, approximately 2.5 million research papers in many different languages are published throughout the globe every year [11]. As more and more research is published in local languages, researchers might not understand or even find articles that could be relevant for their work. So there is a twofold problem: The information overload (a body of at least 160 million publications in 2015, [10]) and the Tower of Babel language issue. What is needed, is an easily accessible and user-friendly multilingual information system. Unfortunately, finding relevant metadata is only half the job. If the user finds relevant records in a language that he does not speak, he still needs help to comprehend the source document. However, hiring human translators or further progress in machine translation are merely organizational issues. As pointed out earlier, finding relevant information in times of information overflow is the more urgent problem to solve.

In the domain of Psychology, such a system could help avoid duplicate research efforts, provide better counseling, health care, reduce human and animal testing, and so forth. Especially in Psychology, basing research solely on results published in dominant languages such as English bears the risk of drawing conclusions on narrow subpopulations than on broader grounds. For example in social science research, it is criticized that scientific knowledge about human psychology published in the English speaking world's top journals is largely based on findings from a subpopulation of undergraduate students in the industrialized West [6]. This ignores cultural diversity and cross-cultural variability. Draguns [3] encourages monolingual English speaking psychologists to not disregard possible research findings published in other languages. National journals play an important role in providing the essential structure for national research systems

All authors
Leibniz Institute for Psychology Information (ZPID), Trier, Germany
Corresponding author
Erich Weichselgartner, E-mail: weichselgartner@zpid.de

as well as in disseminating research results in those more locally-oriented (sub-)disciplines [2]. The verbal description of key concepts in psychology like emotions is extremely language dependent [8].

Established commercial indexes like Web of Science (Thomson Reuters) or PsycINFO (APA) have a focus on English language material and a U.S. bias [3, 4]; in addition, they are quite expensive and many institutions, even in developed countries, cannot afford to subscribe [9]. In order to carry out effective research, users need to know which indexes contain relevant material for them and how to query them. Thus, there are many barriers to overcome in traditional systems: Language, usage, and cost. Cost for access to commercial bibliographic databases is a growing problem in times of shrinking (university) budgets, not only for developing countries.

To address these issues, the Leibniz Institute for Psychology Information (ZPID) and its partners have created PubPsych, a free-to-use vertical search engine for psychological literature, tests, treatment programs and research data in 2013. Aggregating metadata from various sources in different countries and in different languages, PubPsych is a powerful research tool providing access to a broad body of psychological knowledge containing more than 950,000 references as of September 2016. The extensive data pool is established by filtering and aggregating the information from nine source databases produced by seven institutions. These are the ZPID in Germany (PSYNDEX, PsychData and PsychOpen), the Institut de l'Information Scientifique et Technique in France (PASCAL), the Centro de Ciencias Humanas y Sociales in Spain (ISOC-Psicología), the National Library of Norway (NORART), the U.S. National Library of Medicine (MEDLINE), the Education Resources Information Center in the U.S. (ERIC) and the Data Archiving and Networked Services in the Netherlands (NARCIS). Quality and consistency of the metadata, as well as relevance of the documented publications, are assured by the professional human indexing staff of the institutions. Every partner works independently in his own familiar hardware, software and organizational environment to gather the publication data. Subsequently database dumps are made available on a regular basis to ZPID, where they are collected, processed and finally integrated into PubPsych. Unlike horizontal search solutions, PubPsych contains relevant material solely like monographs, journal articles, dissertations and other quality content. In addition, the sources of the contained information and the update frequency are transparent [16]. It uses a consistent system of terms, based on established terminology. Utilizing the intellectual effort of the indexing staff when creating the database records (e.g. classification codes or population information) precise, directed queries can be carried out.

2 Data processing, refinement and indexing at ZPID

The sources of PubPsych are bibliographic databases. Some of them are multi-disciplinary; since PubPsych is a vertical search engine, their psychology segments have to be extracted. All database producers differ to some degree on how bibliographic records are described. This concerns the order of information (e.g., first name, last name), the labels used for document types or publication types (e.g. article, paper), but also the collation of information into records' fields (e.g. journal title, ISSN, start and end page). For PubPsych, the input needs to be harmonized as much as possible. Harmonization is not only required with respect to formal issues like record structure, but also to semantic issues like controlled vocabulary and classification codes. In library and information science, controlled vocabularies are organized lists of words and phrases, or notation systems, that are used to initially tag content, and then to find it through navigation or search [13]. For example, searching the term "Major Depression" would also retrieve records on Dysphoria or Melancholia. Classification Codes are a system that categorizes content according to its primary subject matter. They can be used to limit a search to a particular subcategory. For example, if you are interested in rehabilitation at the work place, you could use the classification code for Occupational & Vocational Rehabilitation.

In PSYNDEX and PsychData the controlled terms are taken from PSYNDEX Terms [14], the authorized German translation of the Thesaurus of Psychological Index Terms of the American Psychological Association. Terms appear in German and English. However, German thesaurus terms are only found in PSYNDEX and PsychData. ISOC-Psicología uses Spanish terms from their translation of an older version of the APA thesaurus. MEDLINE terms are taken from the U.S. National Library of Medicine's Medical Subject Headings (MeSH) and appear in English, German and French. PASCAL uses trilingual terms (French, English and Spanish) from the TermSciences thesaurus, which has MeSH integrated. ERIC supports English terms from their own thesaurus. Overall each data source supports either indexing with controlled terms and/or subject classifications. Where possible ZPID uses static thesaurus lookups to enrich the mono-lingual terms from the source databases with the multi-language variants of the terms in English, Spanish, French and Ger-

man. In our opinion, the use of a more comprehensive thesaurus or an ontology, which takes psychological, educational and selected medical concepts into consideration would be optimal. The mapping of the already present information in PubPsych and such an information structure is still to be solved. For few subdomains, e.g. age group and origin of population, ZPID uses own mappings to reach a standardization, while maintaining the terms present in the source.

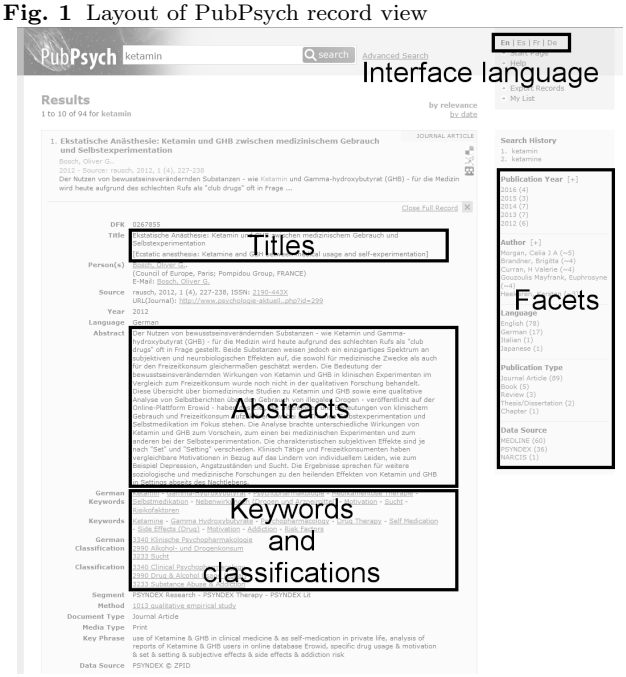
The backend of PubPsych is based on the public Apache Solr and Lucene projects. Solr is an open-source fulltext indexing and searching platform from the Apache Lucene project and very popular due to its ease of use [12]. Lucene is a Java based fulltext indexer with powerful features like ranked searching, phrase and wildcard queries, fielded searching, flexible faceting, highlighting and result grouping. Many libraries have created so called next-generation library catalogs with Solr to aid enhanced discovery (for example VuFind from Villanova University [7]). To enable use of the multilingual data contained in PubPsych a complex field structure is utilized, which can hold data in all four supported languages for textual fields like title, abstract, keyword or classification. Each index field that supports multilingualism consists of language-specific sub-fields with specific tokenizers and stemmers to increase access despite different morphological realizations in query and record.

3 Interface

Retrieval is designed to be intuitive and along the lines of common search engines. You can search by keyword, narrow the results with a variety of filters and look at related items. Users can switch between simple and advanced search. The simple search matches the query in all available search fields [16]. The advanced search allows control of different queries for different fields. A Boolean combination of queries is possible in both searches. Range operators, wildcards, phrase searching and grouping of search terms enable the user to build complex queries, which enlarge or reduce the desired result set. To enhance usability and refine searches many terms in the result output are linked to predefined queries. With the same objective facets are calculated to further refine search results. As internal statistics show, faceting is the fourth most used search type function of PubPsych, behind searching, looking at entry details and browsing result pages.

The interface is a custom development by ZPID and iSearch IT Solutions GmbH (Hannover) in the Java programming language. It is completely available in the four languages English, French, German and Spanish of

Fig. 1 Layout of PubPsych record view



which at least one is spoken by more than 70% of the European population. The interface translations have been supplied by the project partners. For the same four languages multilingual display of the available information for each record on titles, abstracts, classifications and keywords is implemented, if this information is available from the source database or from ZPID enrichment (see fig. 1).

Several further service functions are present in the interface. RSS feeds of queries are available and allow easy monitoring of new information for a specific query. Relevant records can be saved in a watch list and exported in plain text, RIS format or sent by email. Records are annotated with hidden bibliographic metadata according to OpenURL COinS, so they can be stored in popular reference management applications like Zotero and Citavi. Users want full text (content) and not only metadata. Whenever possible, full text retrieval is provided by either direct links, DOI resolution or OpenURL resolution.

4 Experiences and outlook

Users very much appreciate one-stop retrieval solutions where they don't need to learn a new query syntax for each of the search engines or accustom themselves with a new interface design. As opposed to the non-standard custom interfaces of the large bibliographic database vendors use for untrained users is much easier [5].

In 2016 most visitors came from German-speaking countries, followed by France, Spain and English-speaking countries. The user activity is steadily increasing year by year since 2013. In a survey among 351 early adopters from 25 countries, 75% considered PubPsych a valuable complement to existing research tools in psychology [1]. In April 2015 a German translation of the MeSH thesaurus has been integrated into the MEDLINE data of PubPsych, which then enabled retrieval of English records using German keywords. This instantly resulted in an increasing number of requests for full entry display of English MEDLINE records, compared relatively to all other eight data sources. The relative number of MEDLINE full entry display requests, meaning these records were looked at in detail, increased by up to 9.7% when comparing each month with data from the previous year [15]. We consider this to be a first hint that retrieval of records in languages different from the query language is actually useful and interesting for PubPsych users.

For the future ZPID wishes to further increase the multilingualism of the available bibliographic metadata. Beside the already present multilingual data in titles and abstracts as delivered by the data source providers and further enhancements by ZPID with term and classification mappings, efforts will be taken to evaluate contributions of machine translation to metadata enhancement.

In a joint project between ZPID, Humboldt University of Berlin and Saarland University, four approaches to integrating automatic machine translation will be investigated that require different workloads either in the frontend or in the backend of PubPsych. These four approaches will be implemented and tested for quality of translation and evaluated for quality of information retrieval. It is hoped that a system can be built and sustainably maintained that allows effective cross-lingual search and thus, provides access to relevant local resources otherwise not found.

References

1. Bittermann I, Clerc A, Naescher S, Schui G, Waeldin S, Weichselgartner E (2015) Ergebnisse der Befragung zum Produktstart von PubPsych: Bewertungen des Suchportals PubPsych und weiterer Produkte des ZPID. ZPID Science Information Online 15(3):1–10, URL `https://www.zpid.de/pub/research/2015_Ergebnisbericht_PubPsych_Befragung.pdf`, [Online; accessed 14.12.2016]
2. Bordons M, Gómez I (2004) Towards a single language in science? A Spanish view. Serials: The Journal for the Serials Community 17(2):189–195, DOI 10.1629/17189
3. Draguns JG (2001) Toward a truly international psychology. Am Psychol 56(11):1019–1030, DOI 10.1037/0003-066X.56.11.1019
4. Fingerman S (2006) Web of Science and Scopus: Current Features and Capabilities. Issues in Science and Technology Librarianship (48), DOI 10.5062/F4G44N7B
5. Hearst MA (2009) Search user interfaces. Cambridge University Press, Cambridge [u.a.]
6. Henrich J, Heine SJ, Norenzayan A (2010) Most people are not WEIRD. Nature 466(7302):29–29, DOI 10.1038/466029a
7. Houser J (2009) The VuFind implementation at Villanova University. Libr Hi Tech 27(1):93–105, DOI 10.1108/07378830910942955
8. Kornadt HJ, Trommsdorff G, Kobayashi RB (1994) "Mein Hund hat mich bestorben" : sprachlicher Ausdruck von Gefühlen im deutsch-japanischen Vergleich. In: Kornadt HJ, etal (eds) Sprache und Kognition: Perspektiven moderner Sprachpsychologie, Spektrum Akad. Verl., Heidelberg [u.a.], pp 233–250
9. Larivière V, Haustein S, Mongeon P (2015) The oligopoly of academic publishers in the digital era. PLOS ONE 10(2), DOI 10.1371/journal.pone.0127502
10. Orduna-Malea E, Ayllón JM, Martín-Martín A, Delgado López-Cózar E (2015) Methods for estimating the size of Google Scholar. Scientometrics 104(3):931–949, DOI 10.1007/s11192-015-1614-6
11. Plume A, van Weijen D (2014) Publish or perish? The rise of the fractional author.... Trends Journal of Sciences Research 38:16–18
12. The Apache Software Foundation (2016) Solr. URL `https://lucene.apache.org/solr/`, [Online; accessed 14.12.2016]
13. Warner AJ (2002) A Taxonomy Primer. URL `https://www.ischool.utexas.edu/~i385e/readings/Warner-aTaxonomyPrimer.html`, [Online; accessed 14.12.2016]
14. ZPID (ed) (2011) PSYNDEX Terms: Deskriptoren/Subject Terms zur Datenbank PSYNDEX (Lit & AV, Tests), 9th edn. Trier, URL `https://www.zpid.de/pub/info/PSYNDEXterms2011.pdf`, [Online; accessed 14.12.2016]
15. ZPID (2015) Tätigkeitsbericht, Trier, URL `https://www.zpid.de/pub/profil/report2015.pdf`, [Online; accessed 14.12.2016]
16. ZPID (2016) PubPsych Help Guide. URL `https://www.pubpsych.eu/Guide_PubPsych.pdf`, [Online; accessed 14.12.2016]