Genomics & Informatics Vol. 9(1) 37-38, March 2011

ManBIF: a Program for Mining and Managing Biobank Impact Factor Data

Ki Jin Yu[1], Jungmin Nam[2], Yun Her[2], Minseock Chu[2], Hyungseok Seo[2], Junwoo Kim[2], Jaepil Jeon[2], Hyekyung Park[2] and Kiejung Park[1]*

[1]Division of Bio-Medical Informatics, Center for Genome Science, National Institute of Health, Cheongwon-gun 363-951, Korea, and [2]Division of Korea BioBank, Center for Genome Science, National Institute of Health, Cheongwon-gun 363-951, Korea

Abstract

Biobank Impact Factor (BIF), which is a very effective criterion to evaluate the activity of biobanks, can be estimated by the citation information of biobanks from scientific papers. We have developed a program, ManBIF, to investigate the citation information from PDF files in the literature. The program manages a dictionary for expressions to represent biobanks and their resources, mines the citation information by converting PDF files to text files and searching with a dictionary, and produces a statistical report file. It can be used as an important tool by biobanks.

Availability: ManBIF and its manual are available at http://cgs.cdc.go.kr/manbif

Keywords: biobank, Biobank Impact Factor Data

Introduction

Biobanks, as repository organizations, are getting more and more important, not only nationwide but also worldwide. As their roles and functions are being expanded to effectively maintain and distribute biological resources, international cooperation and competition is getting very active.

BIF (Biobank Impact Factor), a comparative index of citation information of biobanks in scientific papers, is used as a major criterion of biobank activity (Zika, 2010).

As a biobank can be represented with several names, including abbreviations and a full name, a dedicated program is required to search for multiple biobanks against multiple literature files.

We have developed ManBIF, a program, to search for Biobank citation information from PDF files in the literature to produce a summary report.

*Corresponding author: E-mail kjpark63@gmail.com
Tel +82-43-719-8850, Fax +82-43-719-8869
Accepted 7 March 2011

Methods

System Structure

ManBIF is composed of three modules for text extraction, indexing, and searching (Fig. 1).

The first module converts PDF files to text files and extracts the full text for searching. The converted files

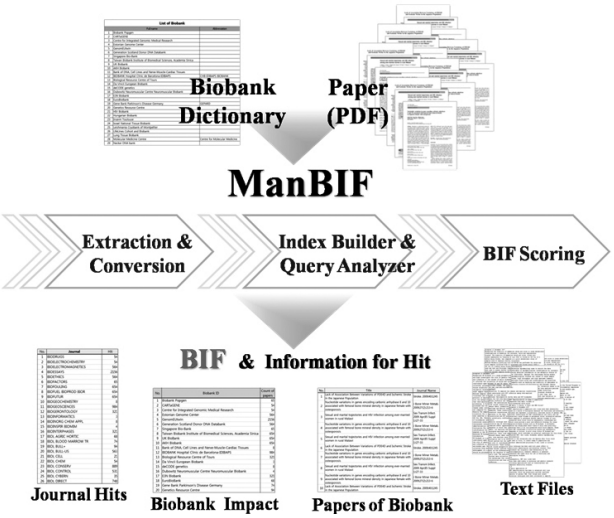

Fig. 1. The system structure of ManBIF.

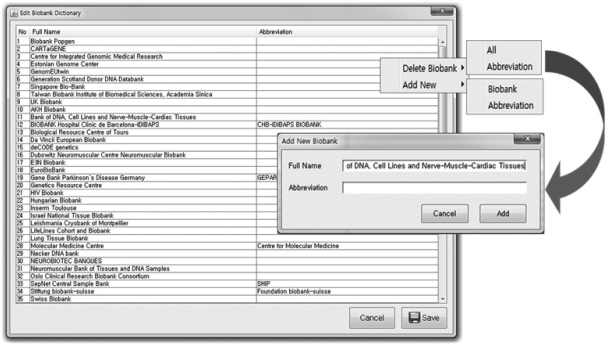

Fig. 2. Editing the ManBIF biobank dictionary.

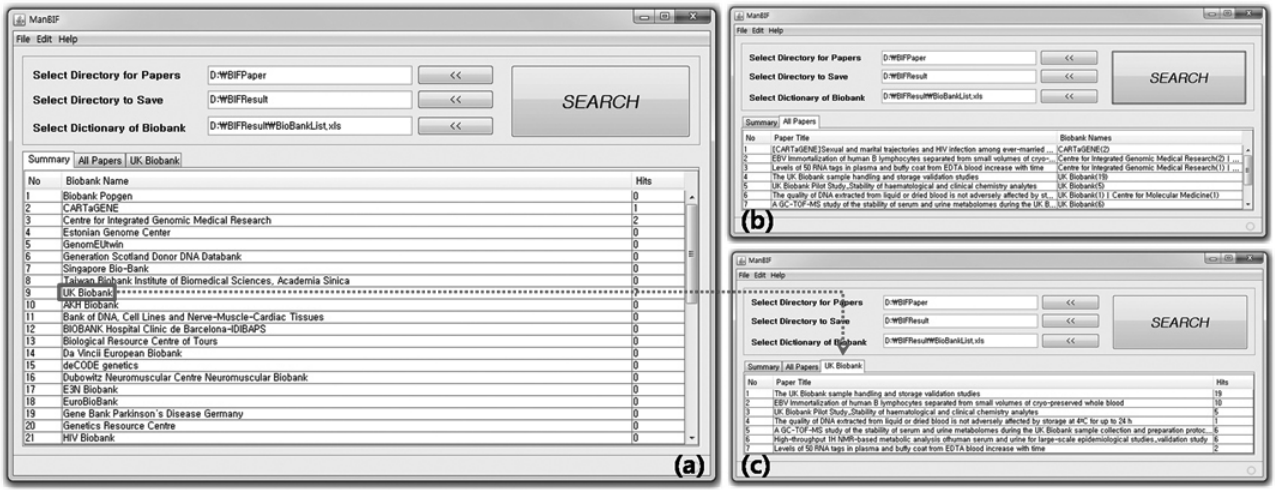

Fig. 3. Screenshots of ManBIF search results that show a list of biobanks and their paper hits (a), a list of all hit papers stored in the database (b), and a list of hit papers for a biobank (c).

are stored in the ManBIF database. The second module analyzes each full text into tokens after removing unnecessary words, and the tokens are indexed and stored in the database. The last module searches the indexed tokens for biobank expressions in the ManBIF biobank dictionary (Fig. 2), which can be edited for managing biobank name databases. The search results of all hits against all biobanks are summarized into a report.

Implementation

ManBIF was implemented with Java, and many libraries and tools were used. iText (Lowagie and Soares, 2010) was used to treat and convert PDF files. The PDF files are converted with library functions. Lucene (Gospodnetic, 2010) was used as a text search engine to support high-performance indexing and searching. jxl (GPL, 2009) was used for generating Excel files. ManBIF can read Excel files and produce search result files in Excel format.

The search results are summarized as a 'biobanks vs paper hits' table, and more detailed results are also shown, including the detailed hit information for each biobank (Fig. 3).

Result and Discussion

ManBIF analyzes PDF files to search for diverse expressions of all biobanks with high-performance indexing and a user-friendly interface. The program can be used practically for evaluating the activity of biobanks in a given interval.

The program can be improved for more precise and categorical searches. Searching by sections of papers, that is, 'Abstract,' 'Introduction,' 'Methods,' and 'Results,' can be more helpful to analyze precisely. Showing marked hit regions directly on PDF files can be more powerful in confirming and analyzing hit information. The search feature for resource names or IDs related to biobanks could also be implemented for improvement. Such an evolution could make the program an important tool for biobanks.

Acknowledgments

This work was supported by the Korea Biobank Project from the Korea Centers for Disease Control and Prevention.

References

GNU General Public License (GPL) (2009). Java Excel API - A Java API to read, write and modify Excel spreadsheets. http://www.jexcelapi.org.

Gospodnetic, O. and Hatcher, E. (2010). Lucene in Action, Second Edition (Manning Publications).

Lowagie, B. and Soares, P. (2010). iText in Action, Second Edition (Manning Publications).

Zika, E., Paci, Daniele., Schulte in den Bäumen, T., Braun, A., RijKers-Defrasne, S., Deschénes, M., Fortier, I., Laage-Hellman, J., Scerri, C. A. and Ibarreta, D. (2010) Biobanks in Europe: Prospects for Harmonisation and Networking, JRC scientific and Technical Reports.