

Life science database cross search: A single window system for dispersed biological databases

Jun-ichi Onami¹, Hideki Hatanaka^{1,2}, Shoko Kawamoto^{2,3}, Toshihisa Takagi^{1,4}

¹Japan Science and Technology Agency, National Bioscience Database Center, Japan; ²Research Organization of Information and Systems, Database Center for Life Science, Joint-Support Center for Data Science Research; ³Research Organization of Information and Systems, Department of Informatics, National Institute of Genetics, Japan; ⁴Toyama University of International Studies, Japan

Received December 23, 2019; Revised December 29, 2019; Accepted December 30, 2019; Published December 31, 2019

DOI: 10.6026/97320630015883

Abstract:

A comprehensive search system for the bioscience databases is in progress. We constructed a search service, Life science database cross search system (<https://biosciencedbc.jp/dbsearch/index.php?lang=en>) by integrating numerous biomedical databases using database crawling algorithms. The described system integrates 600 databases containing over 90 million entries indexed for biomedical research and development.

Background:

The cross-search service of bioscience databases is still developing. Because data are considerably dispersed across various organizations and networks, finding required information immediately is difficult. Additionally, conducting comprehensive searches in large bioscience databases using general web search engine such as Google is difficult [1]. Some search-related infrastructure such as BioCaddie [2] and World Wide Science [3], which are dedicated to research, have been constructed. Because of the deep web problem [4], these search results are not comprehensive and efficient. In this project, we collected all the data from selected bioscience web databases that contained entries in the deep web and developed a web search engine that could search the compiled and comprehensive bioscience database.

Methodology:

This web service is constructed in three steps as shown in Figure 1 and the details are described below.

Web data crawling:

URLs of bioscience web database entries were collected from database catalog sites and funding databases. We checked their site policies, terms of uses, and robots.txt and evaluated the pros and

cons of crawling web data by well-known algorithm [5]. Then, biocurators investigated each entry in the database and distinguished the data containing text that was suitable for text search. Additionally, they checked the variation range of URLs. For instance, some URLs comprised sequential numbers in a predetermined number of digits and padding zero or well-known identifiers in the bioscience field such as PDB ID or Uniprot ID. Database crawling scripts were programmed with a compiled URL list. If a database entry contained some useful metadata such as "species name," "gene name," or "date of creating the database entry" the script parsed each metadata for storage. These metadata were classified into bioscience categories, so that the category of each database could be distinguished.

Server construction and application installation:

A web server and a search server were constructed.

[A] Web server:

The input interface containing the search box was written in PHP. If a user inputted some words into the text box, related words were suggested via an internal dictionary. The search query was parsed, converted into JSON format, and properly processed (as described in Section 3) to be inputted to the Elasticsearch application [6].

[B] Search server:

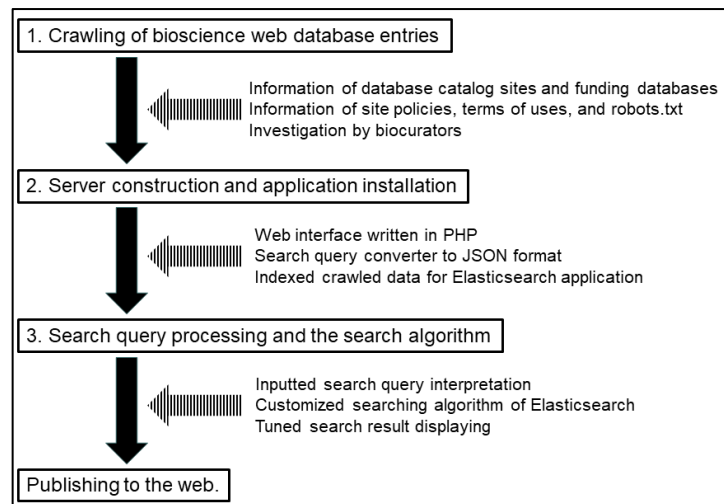


Figure 1: The flowchart of search system construction and the flow of information and technology

The search server parses the outputted JSON data from the Elastic search and displays the search results with a parsed title and snippet fields for visibility. The web interface was published with the page of the target database list and the help page written in English and Japanese. Elasticsearch version 1.7 was installed into the search servers. The crawled data were analyzed by a bigram tokenizer and indexed. The cluster structure that comprised multiple servers accomplished load balancing and data redundancy. The search server receives a search query request from the web server and inputs it into the search process of Elastic search.

Search query processing and the search algorithm:

The search query that is inputted to the search box is subjected to Boolean interpretation, word distinction, and stop word removal in the web server. In this process, if the search query contains words that are often used in the field of bioinformatics (e.g., gene and database), the ranking score of the related database category is boosted. The search results are displayed from Elasticsearch in the order of a decreasing score.

Availability:

Users can retrieve comprehensive information of a matched text by inputting a keyword query into the search box from Life science

database cross search system [7]. Detailed information is available via the link of the search result snippet. The index is automatically updated using the batch script when the updates of the original site have been detected or the running annual updates have been recognized on the basis of RSS or update history. This extensive service would be helpful as an academic search infrastructure for researchers who need to access comprehensive entries in the bioscience database and for users from the intellectual property department. This type of domain-specific search infrastructure has been expensive to construct, and the legal decision of crawling each database has been difficult to make. However, search infrastructure has recently become ubiquitous for global commercial search engines and has become prevalent in ordinary society. Search engine optimization technologies such as “robots.txt” have been well understood. The importance of crawling is well known because data-driven analytic studies are becoming common. Japanese copyright law has been revised and the legitimacy of data crawling has been clarified. In this situation, Life science database cross search is the only service that provides high quality search capabilities to access large quantities of bioscience data. The expectation that the bioscience database must continuously distill to this infrastructure is reasonable.

Conclusion:

We describe the development and use of a comprehensive search system integrating 600 databases containing over 90 million entries indexed for biomedical research and development using database crawling algorithms.

Acknowledgment:

The dictionary data in the Life science database cross search system [7] is supported by the Life Science Dictionary Project [8] at Kyoto University. The literature data of Protein, Nucleic Acid, and Enzyme are provided by Kyoritsu Shuppan Co., Ltd. [9]. The Life science database cross search system [7] was originally developed at the Database Center for Life Science (DBCLS) and was then transferred to the National Bioscience Database Center (NBDC). In this project, we collaborated with the National Institutes of Biomedical Innovation, Health, and Nutrition (NIBIOHN) to develop biomedical searching optimization. Search indexes have been shared with NIBIOHN’s web-based search engine, Sagace [10]. The authors would like to thank Mr. Nobutaka Mitsuhashi and Mitsui Knowledge Industry Co., Ltd. for technical assistance.

References

- [1] Henzinger M *Science* 2007 **317**:468 [PMID: 17656714]
- [2] <https://biocaddie.org/>
- [3] <https://worldwidescience.org/>

- [4] Bergman M *JEP*. 2001 7:1
- [5] Calishain T & Hemenway K *Spidering Hacks* 2009 4:164
- [6] <https://www.elastic.co/jp/products/elasticsearch>
- [7] <https://biosciencedbc.jp/dbsearch/index.php?lang=en>
- [8] <https://lsd-project.jp/cgi-bin/lsdproj/ejlookup04.pl>
- [9] <https://www.kyoritsu-pub.co.jp/contents/index?page=about>
- [10] Morita M *et al. BMC Res Notes*. 2012 5:604 [PMID: 23110816]

Edited by P Kanguane

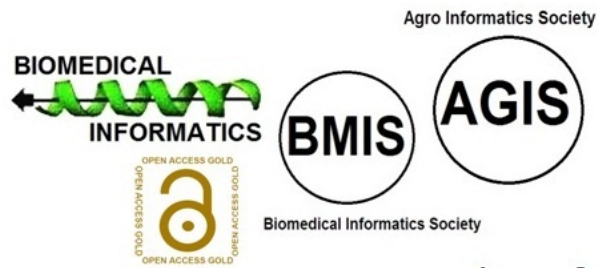
Citation: Onami *et al.* *Bioinformation* 15(12): 883-886 (2019)

License statement: This is an Open Access article which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited. This is distributed under the terms of the Creative Commons Attribution License

Articles published in BIOINFORMATION are open for relevant post publication comments and criticisms, which will be published immediately linking to the original article for FREE of cost without open access charges. Comments should be concise, coherent and critical in less than 1000 words.

BIOINFORMATION

Discovery at the interface of physical and biological sciences



since 2005

BIOINFORMATION

Discovery at the interface of physical and biological sciences

indexed in

