Index development with the keyword in context method in the online library catalog

M Safii*, Martutik, T Harsiati, T Kurniawan, A Asari, L A Rahmania, A Prasetyawan, and Y G Arraja
Library Science Department, Universitas Negeri Malang

Email: moh.safii@um.ac.id

Abstract. Academic libraries have an essential role in the academic community. As a reference center, libraries provide access to online reference sources or book collections. One of the facilities provided is the library automation which facilitate users to access library resources. With an online catalog that is easy to operate, it will be easier for users to get information. However, libraries generally only provide interface search facilities which make it difficult for users to find information. This study aims to produce an index on the online library catalog using the KWIC (KeyWord in Context) method. The data was collected from the Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH) on the Library server. In the data harvesting process, around 50,000 data were found to be processed. The finding in this study indicates that the library's online catalog metadata index makes it easier for users to retrieve information.

1. Introduction
Libraries as institutions have an essential role in the intellectual empowerment of the community. Aside as a building for storing book collections, libraries also provide librarians who provide services and process library materials to be easily found by visitors. In the past, information needs were fulfilled by simply reading the artifacts in which knowledge was contained. However, with the increasingly modern era, the way to present this information as quickly and efficiently as possible has become a challenge. In the field of library science, there is a term 'library automation' which means the computerization process both on the hardware side and the software side. The influence of the development of information technology on libraries makes it easier for libraries to provide services and users to access information [1].

Libraries in the digital era have transformed into a comprehensive information service supported by technological developments. The application of technology in libraries begins with the transfer of manual processes to an automation process known as library automation. The use of automation applications in Indonesia is dominated by Inlistline and Slims. Both applications have been able to solve routine activity problems in the library (house-keeping library).

When the library user comes to the library and uses the search facility on the computer provided, the user may type in inaccurate keywords that are too narrow or even too broad. If so, the system should offer another way to determine what keywords (variations) should be. In the context of an information retrieval system, users simply type in the desired keywords. However, such a process considered to not provide a two-way system since the information appears because it is triggered by...
keywords [2]. In the concept of the information retrieval process, Information retrieval (IR) is about the collection, representation, storage, organization, accessing, manipulation, and display, of the information items necessary to satisfy a searcher’s information needs [3].

The explanation above expands the library study area to the realm of information science by focusing on the study of information stored in databases. The library’s OPAC (Online Public Access Catalog) database certainly has an extensive record consisting of various metadata according to the 9 ISBD (International Standard Bibliographic Description) fields. Therefore, there are many potentials that can be processed to make it easier for users to find new information that settles in the system. Indexing is one of the tools used to give additional value to the dissemination and retrieval of information [4]. The application of indexing in library automation will undoubtedly provide a new perspective on the concept of library automation. Library automation that previously only led to searching will lead to discovering. The final result of indexing is to fulfill the needs of users in the searching system that does not rely only on keywords but also on the variety of choices provided by the system through a series of discovery and dissemination processes [5].

2. Method
This study produces an index that is structured using the Keyword in Context method. The output of this index is a plugin that will be tested in the library’s online catalog. An index is an additional feature that will make it easier for users to search and make the system able to find information. The research procedure was carried out with the following flow:

2.1 Data Collection
The data used in this research is bibliographic data of Batu City Public Library. Data were collected by harvesting the metadata using the OAI-PMH protocol (Open Archives Initiative Protocol for Metadata Harvesting).

![Figure 1. OAI-PMH protocol scheme.](image)

The result of the harvesting process in Figure 1 uses XML format for data exchange. There are rules in syntax form that must be prepared in advance. Server access for harvesting is located at https://katalogdpkkotabatu.perpusnas.go.id/oai?verb=ListRecords&metadataPrefix=marcxml. The metadata format used in the Batu City Public Library’ online catalog is the standard MARC (Machine-Readable Cataloging) format. The data testing process obtained (before harvesting the metadata) is a tag with fields 001 (Control Number), 005 (Date and Time of Latest Transaction), 008 (Fixed-Length Data Elements), 020 (International Standard Book Number), 035 (System Control Number), 041 (Language Code), 082 (Dewey Decimal Classification), 084, 100, 245, 250, 260, 300, 650 and 856.
Figure 2. The output of the harvesting process is in XML format.

The results of the harvesting process were recorded based on the suitability of the metadata. This study used the title of a book or library collection that was produced through the previous harvesting process. The data was then stored in a MySQL database.

2.2 Index with Keyword In Context

The method of making indexes with keywords in context has been known for a long time by the information sector, especially in libraries and archives. This method is also used in the computer field to speed up access to resources. In order to generate the index, the procedures carried out in this study were:

The bibliographic data obtained at the data collection stage (2.1) was filtered first by removing punctuation and eliminating stop words, or the common words that appear in large numbers and are considered meaningless like the words 'and' and 'or'. From this filtering process, clean words are
obtained. Next, the index was generated using the KWIC method. The clean words then become the keywords which next were rotated with the title to form an index[6].

The following is one example with a book entitled "Library and Information Research Methods". By eliminating the stop words, the result is "Library Information Research Methods". The word 'and' is removed because it is considered as a stop word. There is no punctuation elimination process as punctuation is not found in the title. Then, from the title "Library Information Research Methods", the keywords are determined. The representative keywords taken from the title are methods, research, libraries, and information. The index arrangement obtained by the KWIC method are:

a. Library and Information Research Methods
b. Information Methods Research Library and
c. Research Library and Information Methods
d. Library and Information Methods Research
e. Information Methods Research Library

From the example above, the number of indexes using the KWIC method will be in accordance with the number of original words. If the sentence consists of n words, then the index result will be n.

2.3 Plugins in the library's online catalog
The next stage was testing the index by making it a plugin to the library's online catalog. The programming language used was PHP using the MySQL database.

3. Results
Based on the research results, the data collection process obtained 50,000 metadata through OAI PMH harvesting metadata on the Batu City Public Library server. Data recapitulation is shown in table 1 below

| Year | The Number of Data |
|------|-------------------|
| 2002 | 192               |
| 2003 | 1.868             |
| 2004 | 2.654             |
| 2005 | 4.864             |
| 2006 | 2.588             |
| 2007 | 2.656             |
| 2008 | 2.390             |
| 2009 | 4.268             |
| 2010 | 2.608             |
| 2011 | 873               |
| 2012 | 1.285             |
| 2013 | 3.061             |
| 2014 | 8.427             |
| 2015 | 2.378             |
| 2016 | 2.695             |
| 2017 | 2.820             |
| 2018 | 1.988             |
| 2019 | 2.042             |
| 2020 | 343               |

The next step is to eliminate the stop words in Indonesian and stop words in English. The data obtained through the metadata harvesting process turns out to have stop words and punctuation on each metadata. One example is a book entitled "The origin and purpose of Human (Theory of Evolution) / Franz Dahler, Julius Chandra". In the rules of library collection management, the sign / is commonly used to separate the title and the main author. Moreover, the names of the main authors are reversed. From Table 1 above, after going through the indexing process with KWIC, the index obtained are:
Table 2. The Total of KWIC Index.

| Year | The Number of Data | Year | The Number of Data |
|------|--------------------|------|--------------------|
| 2002 | 8.724              | 2011 | 15.909             |
| 2003 | 21.035             | 2012 | 42.461             |
| 2004 | 30.111             | 2013 | 108.485            |
| 2005 | 56.414             | 2014 | 27.475             |
| 2006 | 30.950             | 2015 | 31.394             |
| 2007 | 31.029             | 2016 | 32.096             |
| 2008 | 27.611             | 2017 | 26.221             |
| 2009 | 51.951             | 2018 | 24.021             |
| 2010 | 30.099             | 2019 | 4.498              |

In Table 2 above, the total index produced is 610,861. It means that the index creation results in a lot of new data. The comparison between the amount of data and the index results is shown in the following graph:

![Figure 4. The comparison between the amount of metadata and the index.](image)

The next stage was creating a plugin with the PHP programming language to be applied to the library's online catalog. The results that have been applied are:
Figure 5. The Front display of the library's online catalog.

An index menu has been added to the online catalog, which will lead to an index in the form of keywords that have been arranged alphabetically.

Figure 6. KWIC index and index results.

4. **Discussion**

This research has resulted in a study of indexes using the Keyword in Context method. Since ancient times, the use of indexes in documentation both in digital form has eased the users in the searching process. Even in the database itself, if a field is set as an index, then the search process will be faster [7]. However, behind that convenience, there are data that need to be generated, and this requires a very long process, especially when the index was first created. The use of indexes can also be developed for keyword suggestion on search engines, spell checkers on documents, and other text mining which does provide easy and fast access. [8]. There is indeed a high price to pay because the indexes created are very large in number [9].
5. Conclusion

This study contributes to the library's online catalog in order to provide an index for its users. If there is no index that can help users, the information search process is limited to true and false in the database. If the index is available, then the discovery process will very likely be carried out for more advanced stages.

Acknowledgement

This research is funded by the State University of Malang, and we would like to thank everyone who is involved in LP2M and the lecturers of the Library Science study program.

References

[1] Barbuti N 2020 Thinking digital libraries for preservation as digital cultural heritage: by R to R4 facet of FAIR principles Int J Digit Libr
[2] Black S 2020 The implications of digital collection takedown requests on archival appraisal Arch Sci 20 91–101
[3] Chen L, Shang S, Yang C et al 2020 Spatial keyword search: a survey Geoinformatica 24 85–106
[4] Fu Z, Liu Y, Sun X, et al 2020 Confusing-Key word Based Secure Search over Encrypted Cloud Data Mobile Netw Appl 25 125–132
[5] Habib M N, Jamal W, Khalil U et al 2020 Transforming universities in interactive digital platform: case of city university of science and information technology Educ Inf Technol
[6] Cho J 2020 Intellectual structure evolution of open access research observed through correlation index of keyword centrality Scientometrics 125 2617–2635
[7] Hauswedell T, Nyhan, J, Beals M H et al 2020 of global reach yet of situated contexts: an examination of the implicit and explicit selection criteria that shape digital archives of historical newspapers Arch Sci 20 139–165
[8] Lekshmi K K, Prem M V 2020 Multi-keyword score threshold and B+ tree indexing based top-K query retrieval in cloud Peer-to-Peer Netw. Appl. 13 1990–2000
[9] Manasrah A M, Abu Nasir M, and Salem M 2020 A privacy-preserving multi-keyword search approach in cloud computing Soft Comput 24 5609–5631