TELKOMNIKA, Vol.12, No.1, March 2014, pp. 1~5
ISSN: 1693-6930, accredited A by DIKTI, Decree No: 58/DIKTI/Kep/2013
DOI: 10.12928/TELKOMNIKA.v12i1.1790

Short Communication

The Architecture of Indonesian Publication Index: A Major Indonesian Academic Database

Imam Much Ibnu Subroto*[1], Tole Sutikno[2], Deris Stiawan[3]
[1]Department of Informatics Engineering, Universitas Islam Sultan Agung, Semarang, Indonesia
[2]Department of Electrical Engineering, Universitas Ahmad Dahlan, Yogyakarta, Indonesia
[3]Department of Computer System Engineering, Universitas Sriwijaya, Palembang, Indonesia
Corresponding author, email: imam.mis@gmail.com[1], tole@ee.uad.ac.id[2], deris.stiawan@gmail.com[3]

Abstract

Journal articles are required all researchers as references to improve the quality of research so that the results are better than existing studies. The presence of many journal indexers that collect articles from many publishers and repositories are very helpful to lecturers and researchers to locate articles in their specific areas of interest. The main issues in the indexing include: (i) the selection and collection process articles data, (ii) the management of indexing based on source and relevance of science (iii) the accuracy and speed of the search process, and (iv) the relationship between the articles each other called citation. The desirable by readers from the indexer is to get relevant and good-quality articles easily and accurately. While the interest of the journal managers is to supply their article which reader preferred then hope many cited to their journals. This is where the position of an indexer to bridge publisher and reader. This paper presents the architecture of Indonesian Publications Index (IPI) as the bridge. This architecture is designed with three layers. The layers are the data collection layer, storage layer, and service layer. The functionality of the first layer is IPI collaboration with publishers, the second layer is the index management system, the third layer is a service to the readers. Service layer built on a variety of applications such as web based applications, mobile applications and e-library.

Keywords: *academic database, aggregator, automatic machine, indexer, search engine, IPI*

1. Introduction

From early informal discussions in recent years between the Institute of Advanced Engineering and Science (IAES) Indonesian section members and TELKOMNIKA editors have initiated and been excited to establish a major academic database and search engine for Indonesian academics and researchers which is currently named Indonesian Publication Index (IPI) or formerly known as Portal Garuda Indonesian Publication Index (Portal Garuda IPI). This portal is hosted at: http://portalgaruda.org and is useful in an academic setting for finding and accessing articles in Indonesian academic journals, repositories, archives or other scientific collections [1].

Aggregators, indexers and other decentralized services, such as Web of Science, Scopus, ScienceDirect, Ei Engineering Village, IEEE Xplore, EBSCOHost and ProQuest allow researchers to find relevant articles based on their queries. A large number of collected articles is needed to obtain a high recall result and the method of indexing must be strong to achieve high precision [2, 3, 4]. The problem of collecting data is the growing number of articles and publishers on the internet that are difficult to locate quickly. It is difficult and time wasting to collect data manually one by one. An automated data collection engine is an absolute necessity for the purposes of data collection and data update. This paper presents the development of the architecture of IPI using a semi-automatic machine method which has the capability to collect many appropriate articles from outside original sources.

2. IPI Overview

The IPI is designed for browsing, indexing, abstracting, monitoring and improving the standard of scholarly publications in Indonesia. Currently IPI covers 116,163 articles from 1,864 journals and 190 publishers. It is estimated that there will be over 2,000 Indonesian journals for

Received October 7, 2013; Revised January 26, 2014; Accepted February 11, 2014

inclusion in the IPI database [1]. It is very important that the contents are made visible globally, so that Indonesian academics and researchers can be identified for their expertise and areas of possible collaboration, to stimulate use and provide citations.

The IPI search interface has a number of search options. These are accessed via Tabs across the top of the screen. Using these it is possible to choose between “Simple Search”, "Title Search", "Author Search", "Journal Search" and "Publisher Search" options. All search options lead to a search results form with intuitive refinement options and the ability to link to view, show abstract, download citations/export references with a choice of formats (RIS or Bibtex), save the search and/or results, fulltext, and view from original source. As it is well-known, bibliographical databases lack perfection and standardization. The RIS format is compatible with bibliographic softwares: Endnote, Reference Manager and ProCite. The software tools perform useful information management and bibliometric analysis importing data from them [5].

3. Architecture of IPI

As an academic resource and indexer then the first consideration is the answer to the question of what data are needed and how to collect them. The digital library world has known what is called the open archive initiative (OAI), which is the concept of sharing bibliographic metadata [6]. With this concept it is a source of bibliography that is easy to put together or share with others. This is an opportunity for the indexer journals to use it optimally.

The OAI working principle is to use a URL query to retrieve metadata through OAI that can be shared by the system. From the results of the query, the server will provide a response in the form of metadata in XML format. Currently there are two widely used standard protocols, OAI-PMH version 2.0 and MARC 21. For managers of journals and journal indexers OAI-PMH 2.0 models are preferred because they are easier to understand [7]. IPI also utilizes this protocol as the main engine for collecting data on articles. However, in reality there are many managers who do not provide OAI journals for its management system so that the OAI is not the only engine in the architecture of IPI data collection. Overall IPI is not simply a matter of collecting data, but also data management and presentation of information to the end user. Everything needs to be managed systematically in order to satisfy the publisher and the end user.

Figure 1 shows the general architecture of the IPI divided into three layers. The first layer is a device/machine to collect data from various sources that the article refers to as "Collecting Data". The second layer is called the Storage to manage metadata in a structured article. Included in this layer is a citation counting machine. The third layer is the "Services" that is the bridge to serve applications to the end user. For more details we will explain one by one below.

3.1. Data Collecting

IPI data source is from the journal publishers, especially the managers and also the organizers of the proceedings. In today's world almost all published journals are available on websites, but there is open access and there is limited access. Most open access journals have provided OAI (Open Archive Initiative) services where the aggregator may retrieve data citations in standard XML-based format. Most sources are not open access journals, so need different methods of data processing. Based on these two sources of data, the technique of data collection is achieved in three ways:

a. Harvesting

This method is used to retrieve data from journal citations that support the OAI repository. This kind of data is already in a form that is structured so that it takes a harvester machine to retrieve new data or update data that never existed. Due to the dynamic nature of the data source that is created automatically the harvester machine is periodically update.

b. Metadata Importing

For journals or proceedings which do not have an online version then the manager shall deliver the metadata of all articles that have been published. This kind of data is already in a structured form so that only the necessary adaptations to the format of the data in storage are required.

c. Crawling

This method is adopted when all the information about the journal is on the website only, without the support of OAI. This can be done if all the source PDFs are placed on a page and can be retrieved at any time by anyone. This means that all articles are open access and can be taken by the IPI without having to ask permission from the owners but the records do not remove the copyright origin. The workings of the crawler are to analyse all the links on the website and then open any page on the link that still exists in the same domain name. Each page will also be analysed for each link. This process will stop when a file is found for the article in PDF format or go back to the pages that have been visited. Furthermore, the PDF file will be copied to the server IPI.

The next process is to extract citation metadata of PDF files that conform to the structure of the data standardized by IPI. The trick is to use a parse method on the first two pages of a PDF as we see in Figure 2. To extract title, authors, journal name, publisher, issued year, volume and issue numbers they can practically be taken from the pages in each paper 1 and 2.

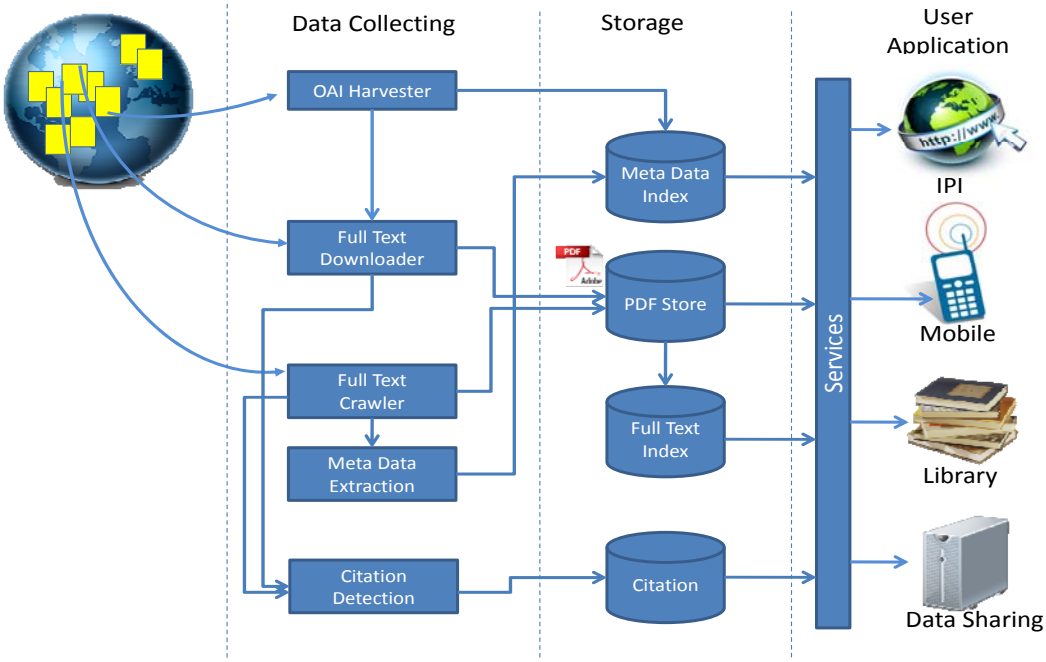

Figure 1. IPI Architecture

3.2. Storage

The database is the most important component of an indexer. Data that is accurate and up to date represents the epitome of confidence from the customer database. IPI is broadly divided into four groups: Metadata Index, PDF Store, Full Text Index, and Citation.

Metadata Index

In a journal indexer, the metadata of each article could be considered the most crucial. This database contains the essential fields in an article that are often called a bibliography or citation. This section includes the most sought after people. Search in general involves a profile article, author, journal name, publisher, abstract and field area.

PDF Storage

After conducting a search process with tools that have been provided, it is usual to download the paper as a whole. IPI in 2014 shows that the number of downloads is higher than the number of page views as can be seen from Figure 2. The PDF data can be said to be very large and consume a lot of bandwidth so that the reliable data management and storage infrastructure are important.

PDF documents are also needed for the calculation of citations from each paper to another paper. With an accurate parsing method, each list of references in PDF metadata can be extracted into a citation that will be compared with other woods citation metadata (see Figure 3). It is necessary to calculate the impact factor of a journal.

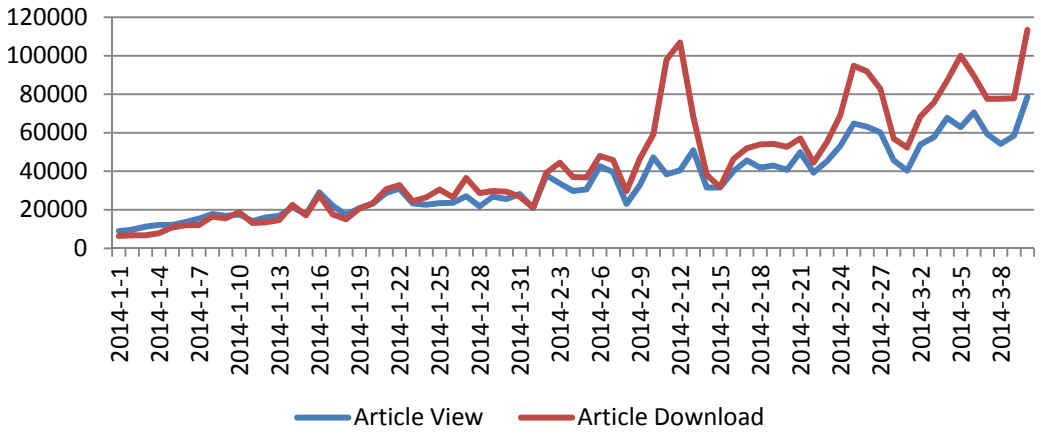

Figure 2. Number of page views and downloads article 2014

Fulltext Index

PDF fulltext extracted from the storage system is very useful for more advanced search engines. With a good indexing system this database is very effective as a plagiarism detection engine. Many methods have been developed in this kind of plagiarism detection using similarity techniques and machine learning techniques. One of promises techniques is using hybrid machine learning of SVM [8].

Citation

Citation is a correlation database comparing one paper with other papers. If a paper is cited by other papers then it means that there is a direct relationship between these papers or also between the authors. This map describes the relationship between the researchers throughout the world who are still closely related to the same field of research.

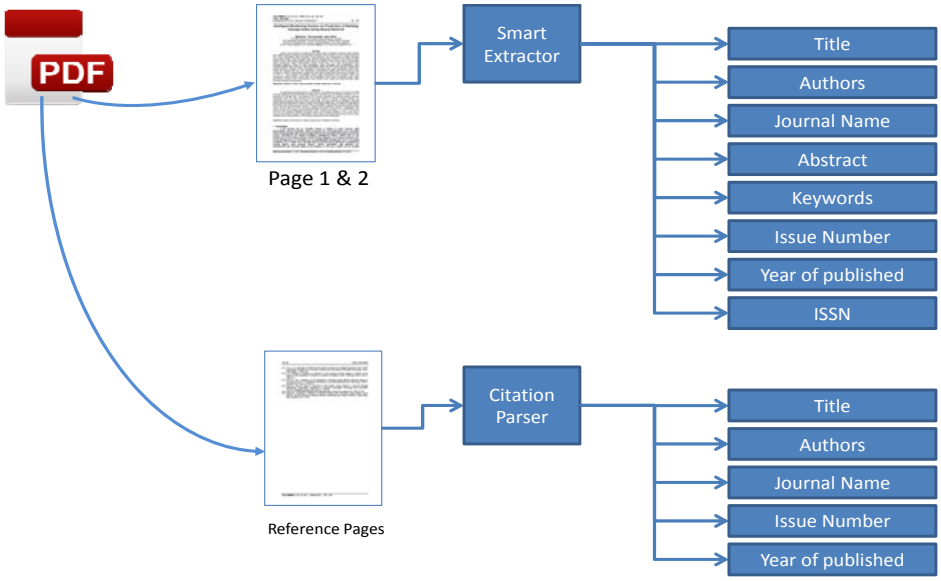

Figure 3. Metadata Extraction from PDF Paper

3.3. Services

As part of the service to end users, a wide range of applications is needed. The primary users of IPI academic resources are faculties, students, researchers, librarians and fellow providers for the purposes of sharing academic resources. With these considerations, the main services are web-based IPI article search (portalgaruda.org), mobile application for smart phones and tablets, integration with the campus library system, and also a bridge for sharing the source with another servers outside the IPI. The latter is a service to open the doors of cooperation with other parties to extend and enlarge the scope of the data.

The current WWW protocols are the most widely used in the world today. So IAES built the first service which is a web-based portal called Garuda Indonesian Publication Index (IPI). This portal is placed on the domain http://portalgaruda.org. The main service provided is search and search articles, authors, journals and publishers. Other additional services include journal profile, visitor statistics and downloader journals, articles based on the statistical spread of the city and country, top journals, a top publisher, accessed top journal, download citations to various formats (RIS, BibTex), download the PDF articles, and so forth. Some examples of service are as follows.

3.3.1. Simple Search

“Simple Search” is designed for those who want to carry out a basic search, with new searchers in mind. There is a single dialogue box into which search terms are entered. Controlled terms, free text terms, author names, title source names and publisher names can all be searched. All fields are searched in all the databases in this option with free access.

3.3.2. Quick Searches

“Quick Search” allows greater flexibility in searching than “Simple Search”. It is possible to specify the fields to be searched via "Title Search", "Author Search", "Journal Search" and "Publisher Search" options and to restrict the answer set using the criteria.

4. Conclusion

In this paper, we have presented an overview of the architecture of IPI. This portal is designed for browsing, indexing, abstracting, monitoring and improving the standard of scholarly publications in Indonesia. The portal has the capability to obtain appropriate articles via a number of search options with high speed and accuracy due to being developed using an automatic machine method.

References

[1] IPI: Indonesian Publication Index. Institute of Advanced Engineering and Science (IAES) Indonesia Section. http://portalgaruda.org
[2] Wu J. Towards a Decentralized Search Architecture for the Web and P2P Systems. http://wwwis.win.tue.nl/ah2003/proceedings/ht-6/towardsah.html
[3] Gomez-Jauregui V, Gomez-Jauregui C, Manchado C, Otero C. Information management and improvement of citation indices. *International Journal of Information Management*. 2014; 34(2): 257-271.
[4] Collins SL. In the aggregate: A quantitative study of the impact of interdisciplinary full text databases on historical research. *Journal of the Association for History and Computing.* 2003; 6(1): 22.
[5] Fernández-Sáez AM, Bocco MG, Romero FP. *SLR-Tool a tool for performing systematic literature reviews*. Proceedings of the 5th International Conference on Software and Data Technologies. 2010; 2: 157-166.
[6] Jackson AS, Han MJ, Groetsch K, Mustafoff M, Cole TW. Dublin core metadata harvested through OAI-PMH. *Journal of Library Metadata*. 2008; 8(1): 5-21.
[7] Seára EFR, Sunye MS, Bona LCE, Vignatti T, Vignatti AL, Doucet A. Extending OAI-PMH over structured P2P networks for digital preservation. *International Journal on Digital Libraries*. 2012; 12(1): 13-26.
[8] Subroto IMI, Selamat A. Plagiarism Detection through Internet using Hybrid Artificial Neural Network and Support Vectors Machine. *TELKOMNIKA Telecommunication, Computing, Electronics and Control.* 2014; *12(1): 209-218*.