Indonesian Citation Based Harvester System

Resmana Lim
Electrical Engineering
Petra Christian University
Surabaya, East Java,
Indonesia
resmana@petra.ac.id

Adi Wibowo
Informatics Engineering
Petra Christian University
Surabaya, East Java,
Indonesia
adiw@petra.ac.id

Raymond Sutjiadi
Research Center
Petra Christian University
Surabaya, East Java,
Indonesia
raymondsutjiadi@petra.ac.id

Yustus Eko Oktian
Electrical Engineering
Petra Christian University
Surabaya, East Java,
Indonesia

***Abstract*—This research proposes a harvester system that utilize Indonesian language based parser to capture citations' metadata from papers. Open Harvester System from Public Knowledge Project (PKP-OHS) is used as a base of harvester system. Citation extraction and citation citegraph methods are added to extend the processing capability of PKP – OHS to enable processing citations. Lastly several information output are modified to enable provision of citation information to users.**

Keywords—harvester system, citation extraction, e-journal

I. INTRODUCTION

There is a lack of support in Indonesian scientific repository to provide complete access to Indonesian papers and journals.Usually each repository will maintain each own database, and also its own access methods. There is also still no citation network between papers inside a repository and also between papers among repositories.

This research proposes a harvester system that is part of a larger research project to develop Indonesian Scientific Citation Database (ISCD) System. ISCD system try to provide a database system consists of Indonesian papers and journals from Indonesian researchers, and also to build citation network among papers, and citation analysis that analyse researcher's and journal's impact factor, citation statistics, topic of interest, and other metrics. To achieve this goal the project can not run in solitary but must utilize scientific article databases held by many institutions in Indonesia as content providers. The ISCD will harvest articles' metadata from Google Scholar database and several journal database in Indonesia, parse each article's PDF files, store and then link citation to the original article. The overview of the project is shown at Figure 1.

The harvester part of the project is using Open Harvester System (OHS). OHS is a metadata harvester and indexing system developed by Public Knowledge Project. OHS implements Protocol for Metadata Harvesting from Open Archive Initiative (OAI-PMH) [1]. By using OAI-PMH, OHS collects paper and journal's metadata from journal database that also implement OAI-PMH. Originally OHS doesn't support citation database, so this research try to improve OHS by adding capability to store citations and link them to build citation network.

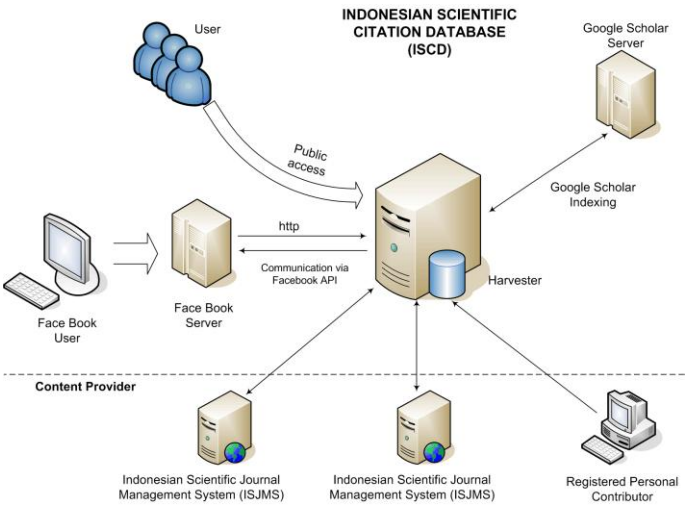

FIGURE I. ISCD PROJECT

II. OPEN ARCHIVE INITIATIVE – PROTOCOL FOR METADATA HARVESTING

OAI-PMH is basically an implementation of REST-based Web services protocols. REST architecture consists of a server and a client. REST client in the OAI-PMH uses GET and POST operations to retrieve metadata collections that are stored by the REST server. Data is sent from the server to the client in the form of XML documents as shown in Figure 2.

OAI-PMH uses verbs to to identify the type of operation requested by the client to the server [2]. Verbs is used to determine the metadata formats that are supported by the repository, to fetch paper metadata from the server, or to know the categories provided by the repository server. Complete verb list is shown in Table 1.

TABLE I. REQUEST VERB LIST USED IN OAI-PMH

Verb	**Function**
GetRecord	Retrieve one metadata record from the server

Verb	Function
Identify	Getting the OAI-PMH protocol version supported by the server, the email administrator, record removal system, and the level of date detail
ListIdentifiers	Retrieve a list of papers' header.
ListMetadataFormats	Get metadata format supported by server
ListRecords	Retrieve papers' metadata based on date or set criteria
ListSets	Get paper's set (category)

```
<OAI-PMH
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/ http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
<request identifier="oai: 10.1.1.40.5588" metadataPrefix="oai_dc" verb="GetRecord">oai2</request>
<GetRecord>
<record>
<header>
<identifier>10.1.1.40.5588</identifier>
<datestamp>2009-04-11</datestamp>
</header>
<metadata>
<oai_dc:dc
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
<dc:title>A Method for Obtaining Digital Signatures and Public-Key Cryptosystems</dc:title>
<dc:creator>R.L. Rivest</dc:creator>
<dc:creator>A. Shamir</dc:creator>
<dc:subject>the difficulty of factoring the published divisor</dc:subject>
<dc:description>An encryption method is presented ...</dc:description>
<dc:date>2009-04-11</dc:date>
<dc:format>application/postscript</dc:format>
<dc:type>text</dc:type>
<dc:identifier>http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.40.5588</dc:identifier>
<dc:source>http://www.matha.mathematik.uni-dortmund.de/~fv/diplom_i/ars78.ps</dc:source>
<dc:language>en</dc:language>
<dc:relation>10.1.1.116.2833</dc:relation>
<dc:relation>10.1.1.115.3569</dc:relation>
<dc:rights>Metadata may be used without restrictions as long as the oai identifier remains attached to it.</dc:rights>
</oai_dc:dc>
</metadata>
</record>
</GetRecord>
</OAI-PMH>
```

FIGURE II. XML DOCUMENT AS SERVER'S RESPOND FOR CLIENT'S GETRECORD REQUEST

The harvester system in this research works as a client to other journal database system to retrieve papers' metadata from Indonesian journal database systems. OAI PMH support Dublin Core metadata. Dublin Core metadata element consists of paper's contributor, coverage, creator, date, description, format, identifier, language, publisher, relation, rights, source, subject, title, and type. PKP-OHS that is used as a basis for this harvester system is also support those 15 metadata elements.

Although PKP-OHS support all OAI-PMH metadata elements, the harvester system will only use title, creator, and journal name including journal number or edition metadata elements from harvested XML. Additional information about paper (bibliographic and citations metadata) are provided by the parser used by the harvester.

III. HARVESTER SYSTEM

A. *Harvesting Process*

Harvester system works by extending PKP-OHS to support citation processing and storing. Harvester system has three steps in collecting, parsing and storing papers' metadata.

1. Metadata Extraction

   Harvester first has to have a list of content providers which are Indonesian online journal databases that support OAI-PMH. By using PKP-OHS, harvester retrieves bibliographic metadata of papers from content providers. As mentioned in chapter 2 harvester will only store title, creator, and journal name from harvested metadata into database.

   This step also try to grab PDF files location for each paper from html files acquired from content providers' web server. Harvester uses regular expression to locate PDF files' URL address.

   Metadata and files location are then stored in two tables, i.e. puslit_papers as shown in Table 7 and puslit_files tables in Table 5.

2. Citation Extraction

   By using PDF files location from step 1, this step download the files and parse them to extract bibliographic and citation metadata. The harvester system utilizes other part of research project which is a parser that extract citations from paper's PDF files. The parser is an enhanced ParsCit system [3] that able to identify indonesian language based bibliographic metadata and citations from papers. For bibliographic metadata, the parser will provide author's name, affiliation, email, paper's title, abstract, and keywords. For citation the paper can provide paper's title from citations, and also the name of all authors, journal name and edition. How the parser works is already explained in other project research publication so it will not be explained in this paper.

   Bibliographic and citation metadata are stored in two tables i.e. puslit_papers as shown in Table 7 and puslit_citations in Table 3.

3. Citation Citegraph

   This step try to match citation metadata records with original paper records already stored in database. If there is a match the link is stored in table puslit_citegraph_citations. This step also try to find if the

```
if (trim($row_citation['authors'])!='' and trim($row_author['name']) != '')
{
	$author_citation = explode(" ", remove_mark(strtoupper(trim(str_replace("\n", " ", $row_citation['authors'])))));
	$author_paper = explode(" ", remove_mark(strtoupper($row_author['name'])));

	$count=count($author_citation);

	for($counter=0;$counter<$count;$counter++){

		if (strlen($author_citation[$counter])<=1)
		{
			unset($author_citation[$counter]);
		}
	}
	$author_citation=array_values($author_citation);

	$count=count($author_paper);
	for($counter=0;$counter<$count;$counter++){
		if (strlen($author_paper[$counter])<=1)
		{
			unset($author_paper[$counter]);
		}
	}
	$author_paper=array_values($author_paper);


	$match_author = array_intersect ($author_citation, $author_paper);
	echo "Author Data from Citation: ";
	print_r ($author_citation);
	echo "<br />";
	echo "Author Data from Paper: ";
	print_r ($author_paper);
	echo "<br />";
	$num = count($match_author);
	if (count($match_author)>0)
	{
		$found++;
		$common_words = implode("~", $match_author);
		echo "Matching author = " . $common_words . " (" . $num . " words) <font color='blue'>VALID</font><br />";
	}
}
```

FIGURE III. MATCHING PROCESS BETWEEN CITATION RECORDS AND PAPER RECORDS

paper is self cited, that means that the authors cite their own paper. If self cited then field "self" will be given a value of 1.

The code to match citation metadata records and original paper records can be seen in Figure 3.

B. *Harvester Database*

To enhance the capability of PKP-OHS to store bibliographic and citations metadata several additional tables need to be created to store citation parsing results. The tables are:

a. puslit_authors as shown at Table 2 contains authors' name parsed from paper's pdf files.

b. puslit_citations as shown at Table 3 contains parsed and raw citations from pdf files.

c. puslit_citegraph_citations as shown at Table 4. If original paper metadata is already stored in puslit_papers then the pointer from citation to original paper stored in this table.

d. puslit_files as shown at Table 5.

e. puslit_keywords as shown at Table 6.

f. puslit_papers as shown at Table 7 contains paper indexed by harvester.

TABLE II. PUSLIT_AUTHORS TABLE

Field	Type	Description
id	bigint(20)	Primary key of table (auto increment)
name	varchar(100)	Author's name
affil	varchar(255)	Affiliation
address	varchar(255)	Address
email	varchar(100)	Email
ord	int(11)	Order of authors
paperid	varchar(100)	Id of paper writen by this author (reference to primary key of puslit_papers)

TABLE III. PUSLIT_CITATIONS TABLE

Field	Type	Description
id	bigint(20)	Primary key of table (auto increment)
authors	text	Author's name
title	varchar(255)	Citation's paper title
venue	varchar(255)	Publication venue (name of journal / conference, etc.)
venueType	varchar(20)	Type of venue (Journal Article, Proceeding, dll)
year	int(11)	Publication year
pages	varchar(20)	Page number in journal
editors	text	Editor name
publisher	varchar(100)	Publisher name
pubAddress	varchar(100)	Publisher address

volume	int(11)	Volume of journal
number	int(11)	Number of journal
tech	varchar(100)	Method/type of writing (research, thesis, dll)
institution	varchar(255)	Institution name
note	varchar(255)	Additional note
raw	text	Raw format of citation as a result of citation parsing
paperid	varchar(100)	Id of original paper of this citation (reference to primary key of puslit_papers)
self	tinyint(4)	1 if self-cited, 0 if not.
ord	int(11)	Order of citation

TABLE IV. PUSLIT_CITEGRAPH_CITATIONS TABLE

Field	Type	Description
id	bigint(20)	Citation ID (reference to primary key of puslit_citations)
paperid	varchar(100)	Paper ID (reference to primary key of puslit_papers)

TABLE V. PUSLIT_FILES TABLE

Field	Type	Description
id	bigint(20)	Primary key dari tabel (auto increment)
paperid	varchar(100)	Paper ID (reference to primary key of puslit_papers)
file	varchar(200)	File Name

TABLE VI. PUSLIT_KEYWORDS TABLE

Field	Type	Description
id	bigint(20)	Primary key of table (auto increment)
keyword	varchar(100)	Keywords
paperid	varchar(100)	Paper ID (reference to primary key of puslit_papers)

TABLE VII. PUSLIT_PAPERS TABLE

Field	Type	Description
id	varchar(100)	Primary key of table (Paper ID from harvester)
auth_code	varchar(10)	OAI metadata format used
title	varchar(255)	Paper title
abstract	text	Paper abstract
month	varchar(15)	Publication month
year	int(11)	Publication year
venue	varchar(100)	Publication venue (name of journal / conference, etc.)
venueType	varchar(20)	Type of venue (Journal Article, Proceeding, dll)
pages	varchar(20)	Page number in journal
volume	int(11)	Volume of journal
number	int(11)	Number of journal
publisher	varchar(100)	Publisher name
pubAddress	varchar(100)	Publisher address
pub_date	varchar(50)	Publication date
file_type	varchar(50)	Format of file
language	varchar(15)	Bahasa yang digunakan
note	varchar(150)	Additional note
public	tinyint(4)	1 if public access, 0 private access.
crawlDate	timestamp	Processing / crawling date
versionTime	timestamp	Last version update

C. *Citation Information Output*

To display citation information gathered by harvesting process PKP-OHS must be modified. There are several modifications to PKP-OHS:

1. Main page is able to display 10 most cited paper, and also link to see paper in more detail as can be seen at Figure 4.

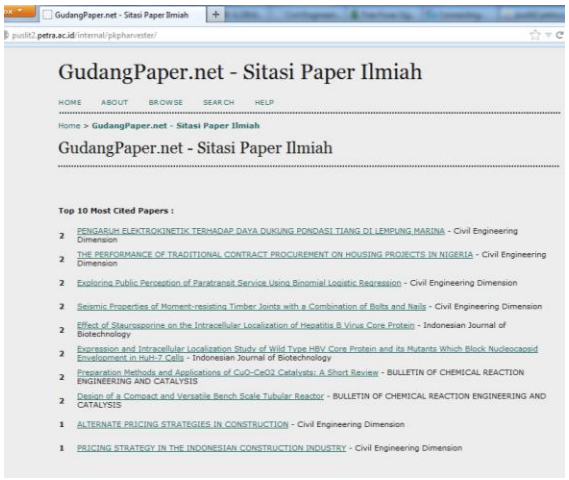

FIGURE IV. TEN MOST CITED PAPERS

2. Search result page can show number of citations to the paper and also name and link of other papers that cite the paper as can be seen at Figure 5.

FIGURE V. LIST OF PAPERS THAT CITES DISPLAYED PAPER

3. Record Details page can show number of paper that cites, number of self-citation, and a link to download the paper as can be seen at Figure 6.

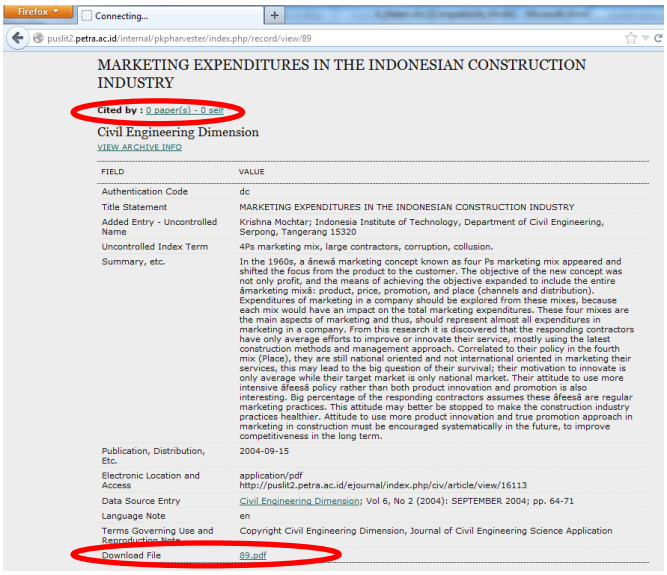

FIGURE V. RECORD DETAILS PAGE

IV. CONCLUSION

This research is able to extend the capability of PKP-OHS to process, store, and provide information about paper citations. Extended PKP-OHS needs parser that able to identify Indonesian language based citation and bibliographic metadata from papers. This research also suggests several tables to supplement PKP-OHS database structure to provide citation network between papers.

REFERENCES

[1] Public Knowledge Project, “Open Harvester System”. Retrieved July 4, 2012 from http://pkp.sfu.ca/?q=harvester

[2] Open Archive Initiative, “The Open Archives Initiative Protocol for Metadata Harvesting”. Retrieved July 4, 2012 from http://www.openarchives.org/OAI/openarchivesprotocol.html

[3] Councill, Isaac G., Giles, C. Lee., Kan, Min-Yen., “ParsCit: An open-source CRF reference string parsing package”, The Pennsylvania State University, National University of Singapore. Retrieved July 4, 2012, from http://aye.comp.nus.edu.sg/parsCit/lrec08/lrec08.pdf