International Journal of Computer (IJC)

ISSN 2307-4523 (Print & Online)

http://ijcjournal.org/

Metadata Extraction from References of Different Styles

Olugbenga A Madamidola[a]*, Olatunde, Ibikunle[b], Olawale T Adeboje[c], Promise I Ayansola[d]

[a,c,d]Department of Computer Sciences, Dominion University, Ibadan Oyo State Nigeria
[b]Rubber Research Institute of Nigeria, Benin City, Nigeria
[a]Email: a.madamidola@dominionuniversity.edu.ng
[b]Email: olatundeibikunle1@gmail.com
[c]Email: adebojeot@gmail.com
[d]Email: p.ayansola@dominionuniversity.edu.ng

Abstract

Metadata extraction is the process of describing extrinsic and intrinsic qualities of the resource such as document, image, video, including getting data from references. References form an essential part of electronic scholarly publications. A reference is the way of giving acknowledgment to individuals for their creative and intellectual works that one utilized in his or her research work. It can also be used to locate particular sources and combat plagiarism. A reference style dictates the information necessary for a reference and how the information is ordered. Accurate and automatic reference metadata generation provides scalability, interoperability and usability for digital libraries of both public and private institution and their collections. Accurate reference metadata extraction becomes an intriguing task to researchers who want to collect data of scientific publications; therefore, this research work proposes a metadata extraction from references of different styles with the use of regular expression. This work accurately extract metadata such as author, title of article, volume, year of publication and institution from references of different styles limiting it to six referencing style**.**

Keywords: Metadata Extraction; Strings; Research; Reference; Regular Expression.

----

* Corresponding author.

1. Introduction

Metadata is loosely defined as “data about data” most generally, data that describes other data to enhance their usefulness in content explanation. Metadata is descriptive information about an object – not the object itself. Extraction of reference meta-data requires the detection of the segmentation of individual reference strings, and the labeling of single tokens within each string as to which field they belong (e.g., author, title, year, journal), which have constituted an important kind of metadata; valuable for literature search, analysis, and evaluation. These are challenging tasks given the variety of different journal layouts and formatting of reference strings. Though several approaches have been proposed, this work represent pattern matching using regular expression to recognize strings in each of the referencing style which that we shall present. According to different implementation method, the Information extraction technology can be divided into a dictionary-based text information extraction, text information extraction based on the Markov model and hidden Markov models, feature-based rules and semantic-based rules for text information extraction, of which the dictionary-based and Markov model- based information extraction are used mainly in Web field; only the method of feature-based rules and semantic-based rules can be applied not only to Web field but also Word, PDF documents field for text information processing. This work proposes Regular Expression (RE) approach to reference metadata extraction. The propose work presents an all-encompassing system that extracts metadata from references of journals types and identifies the style to which a reference belong. The system streamlined to six (6) different reference styles used in journal citations, and has the capability to represent and match template structures of regular expressions formed for different reference styles from the accepted natural language text after which the set of metadata are extracted from different kinds of reference styles. The six referencing styles implemented include APA Style, IEEE Style, ACM style, MISQ style, JMIS style, and ISR style.

2. Review of Related Works

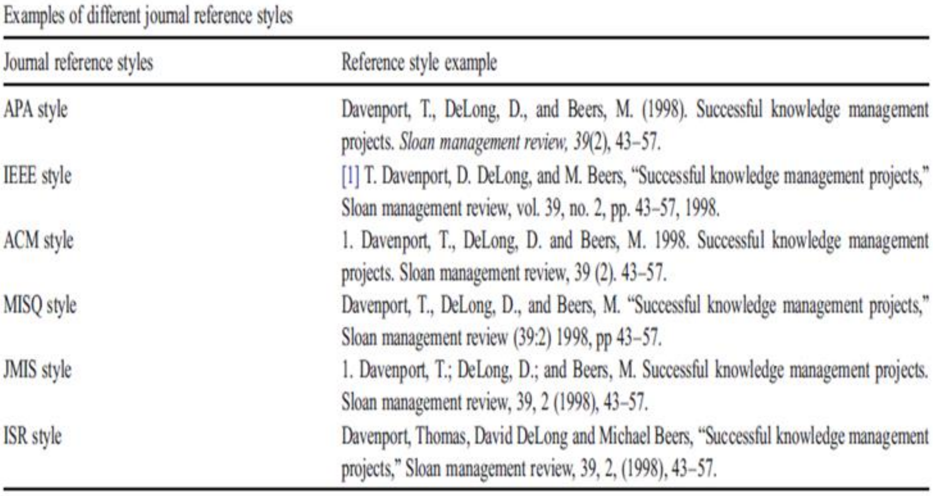
Numerous works on extracting of metadata from reference have been done and implemented. Ojokoh [1] used rule-based approach for the task from a number of reference styles. Gupta and his colleagues [7] used a combination of regular expression-based heuristics and knowledge-based systems to extract metadata from the references of some biological science papers. Ojokoh and his colleagues [8] propose a three-dimensional transition matrix in which the probability of transitioning to a new state depends not only on the current state according to the traditional HMM but also on the previous state. This method improves on those adopted by previous researches, by recommending a new approach for smoothing transition probabilities, a modified shrinkage technique for smoothing emission probabilities and optimization of the emission vocabulary. Papavlasopoulos [9] worked on evaluating the scientific impact of journal. An existing powerful state-of-the-art system for extracting references from a scientific article is **ParsCit**; however, it requires the input document to be converted into plain text, thereby ignoring most of the formatting and layout information. Another is the **metadata extraction tool** built by Sytec Resources for the *National Library of New Zealand Te Puna Mātauranga o Aotearoa (National Library)* to process digital master files and extract metadata about those files. Generally, a reference manager entails larger than metadata extraction. Reference manager supports researchers in performing three basic research steps: searching, storing, and writing (M. H. Fenner 2010). It helps researchers find relevant literature, allows them to store papers and their bibliographic metadata in a personal database for later retrieval, and allows researchers to insert citations and references in a chosen citation style when writing a text. To support those steps, a reference manager should have the following functionalities as identified by Gilmour and Cobus-Kuo (2011):

- Import citations from bibliographic databases and websites
- Gather metadata from PDF files
- Allow organization of citations within the reference manager database
- Allow annotation of citations
- Allow sharing of the reference manager database or portions thereof with colleagues
- Allow data interchange with other reference manager products through standard metadata formats (e.g. RIS, BibTeX)
- Produce formatted citations in a variety of styles
- Work with word processing software to facilitate in-text citation

A reference manager is a software package that allows scientific authors to collect, organize, and use bibliographic references or citations. The terms citation manager or bibliographic management software are used interchangeably.

3. Methodology

Analyzing references of different styles with the view to extract metadata therein requires critical examination of peculiarities of these styles. It is of no doubt that, each possesses it uniqueness, which is used for parsing and matching of contents. The mode to which each of this style is written is shown in the table below and this form the basis of our analysis the implementation.



Examples of different journal reference styles

Journal reference styles	Reference style example
APA style	Davenport, T., DeLong, D., and Beers, M. (1998). Successful knowledge management projects. *Sloan management review, 39*(2), 43–57.
IEEE style	[1] T. Davenport, D. DeLong, and M. Beers, “Successful knowledge management projects,” Sloan management review, vol. 39, no. 2, pp. 43–57, 1998.
ACM style	1. Davenport, T., DeLong, D. and Beers, M. 1998. Successful knowledge management projects. Sloan management review, 39 (2). 43–57.
MISQ style	Davenport, T., DeLong, D., and Beers, M. “Successful knowledge management projects,” Sloan management review (39:2) 1998, pp 43–57.
JMIS style	1. Davenport, T.; DeLong, D.; and Beers, M. Successful knowledge management projects. Sloan management review, 39, 2 (1998), 43–57.
ISR style	Davenport, Thomas, David DeLong and Michael Beers, “Successful knowledge management projects,” Sloan management review, 39, 2, (1998), 43–57.

Figure 1: Examples of Different Journal reference style

The work defines a Grammar for each of the styles by representing the metadata (tokens) with letters. The metadata to be extracted include Authors, Title of Article, Title of periodical, Volume, Year of publication and

Page.

Grammar Definition for Metadata Extraction

Reference No	N
Author(s)	A
Title of Article	T
Title of Periodical	T1
Volume	V
Year of Publication	Y
Page(s)	P

Metadata extraction Grammar

For APA style, this work present the example reference:

Davenport, T., DeLong, D., and Beers, M. (1998). Successful knowledge management projects. Sloan management review, 39(2), 43-57.

We define the Grammar G → A (Y). T. T1, V, P. | A (Y). T? T1, V, P.

For IEEE Style, we present the example reference:

[1] T. Davenport, D. DeLong, and M. Beers, "Successful knowledge management projects," Sloan management review, vol. 39, no. 2, pp. 43-57, 1998.

We define the Grammar G → [N] A “T,” T1, V, pp. P, Y. | [N] A “T?” T1, V, pp. P, Y.

For ACM Style, we present the example reference:

1. Davenport, T., DeLong, D., and Beers, M. 1998. Successful knowledge management projects. Sloan management review, 39(2). 43-57.

We define the Grammar G → N. A Y. T. T1, V. P. | N. A Y. T? T1, V. P.

For MISQ Style, we present the example reference:

Davenport, T., DeLong, D., and Beers, M. "Successful knowledge management projects," Sloan management review (39:2) 1998, pp 43-57.

We define the Grammar G ⟹ A “T,” T1 (V) Y, pp P. | A “T?” T1 (V) Y, pp P.

For JMIS Style, we present the example reference:

1. Davenport, T.; DeLong, D.; and Beers, M. Successful knowledge management projects. Sloan management review, 39,2 (1998), 43-57.

We define the Grammar G ⟹ N. A T. T1, V (Y), P. | N. A T? T1, V (Y), P.

For ISR Style, we present the example reference:

Davenport Thomas, David DeLong, and Michael Beers, "Successful knowledge management projects," Sloan management review, 39,2, (1998), 43-57.

We define the Grammar G ⟹ A “T,” T1, V, (Y), P. | A “T?” T1, V, (Y), P.

4. Design Architecture

The architecture of the reference metadata extractor is shown in Fig. 1. It consists of three main components

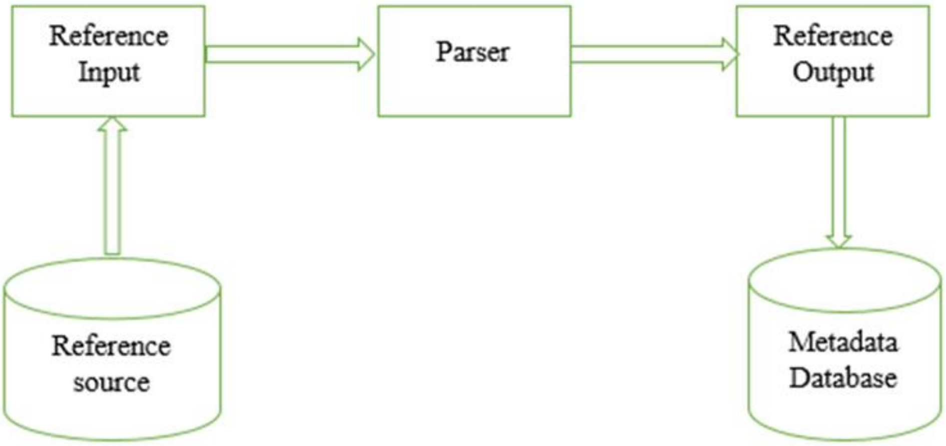


Figure 2: Reference Metadata Extractor Architecture

4.1 Implementation

This research work was implemented exploiting the advantage of Object Oriented Programming (OOP) of C# (C-sharp) programming language on the .net (dot net) framework. The interface and operation are presented below.

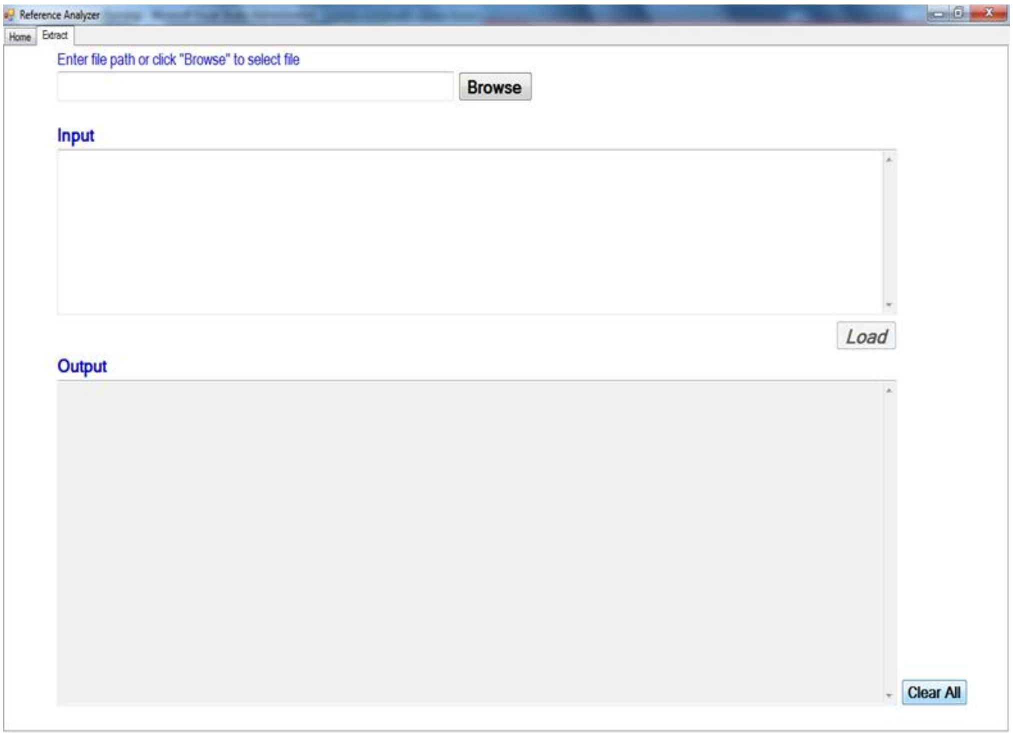
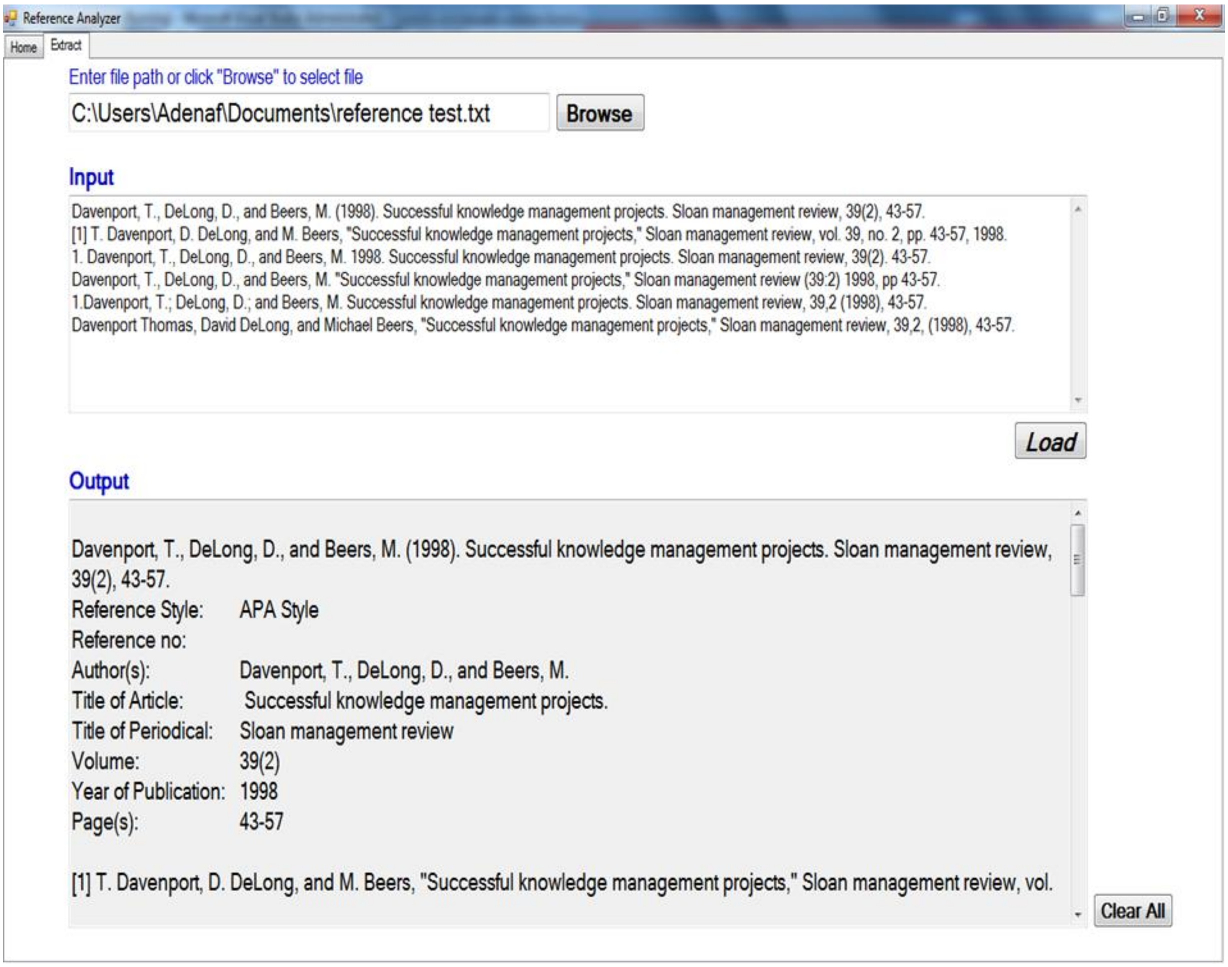


Figure 3: Input and Output Interface

Input: This application is flexible to accommodate three different input approaches. Although the input is a text file (.txt), its entry can be done by entering the path of a text file directly to the first textbox on the interface i.e. typing a link to where it can be found on the system. This automatically checks within the system if the file exists and loads the content to the second textbox, which display the content of the file as input. It can also be done by clicking on the "BROWSE" button to open a dialog box where our input file is located and uploaded. This also loads the content of our text file into the second textbox labeled "Input" on the interface. Lastly, users can directly type references into the textbox labeled "Input" or paste a copied reference.

Parser: This is where the system compares the input reference strings with the predefined grammars to identify which reference type is a particular reference string before extracting the metadata. This is done by a simple click on button "LOAD".

Output: Here, the result obtained from the parser is displayed via the textbox labeled "Output". It displays the complete reference string, reference style to which it belongs, reference no (if required), Author(s), Title of Article, Title of periodical, Volume, Year of publication and referenced page each on a line in order. The "Output" textbox is enabled with scrollbars should the results to be displayed grow beyond the size of the box.



C:\Users\Adenaf\Documents\reference test.txt

Davenport, T., DeLong, D., and Beers, M. (1998). Successful knowledge management projects. Sloan management review, 39(2), 43-57.
[1] T. Davenport, D. DeLong, and M. Beers, "Successful knowledge management projects," Sloan management review, vol. 39, no. 2, pp. 43-57, 1998.
1. Davenport, T., DeLong, D., and Beers, M. 1998. Successful knowledge management projects. Sloan management review, 39(2). 43-57.
Davenport, T., DeLong, D., and Beers, M. "Successful knowledge management projects," Sloan management review (39:2) 1998, pp 43-57.
1.Davenport, T.; DeLong, D.; and Beers, M. Successful knowledge management projects. Sloan management review, 39,2 (1998), 43-57.
Davenport Thomas, David DeLong, and Michael Beers, "Successful knowledge management projects," Sloan management review, 39,2, (1998), 43-57.

```
Davenport, T., DeLong, D., and Beers, M. (1998). Successful knowledge management projects. Sloan management review,
39(2), 43-57.
Reference Style:     APA Style
Reference no:
Author(s):           Davenport, T., DeLong, D., and Beers, M.
Title of Article:     Successful knowledge management projects.
Title of Periodical: Sloan management review
Volume:              39(2)
Year of Publication: 1998
Page(s):             43-57

[1] T. Davenport, D. DeLong, and M. Beers, "Successful knowledge management projects," Sloan management review, vol.
```

Figure 4: Result Interface

5. Conclusion

Extraction of metadata references from references of all styles is hard, but not impossible. This work has been able to accurately extract metadata from six different reference styles of journal citations. This work was tailored on journals for it role and importance in the research field. Although, reference metadata extraction problems have been solved successfully over the years with various method and model, the need for more improvements keeps arising to achieve better accuracy.

5.1 Limitations of the Study

The limitation in this work is the use of only Six (6) referencing style without adding more styles like American Meteorological Society styles, Chicago styles, Harvard referencing, OSCOLA referencing, Oxford referencing to mention few has there are still many more. The work might not be scalable enough to be utilized for generic solution

6. Recommendation

The Metadata extraction of different referencing styles can be implemented in research institutes, private and public Libraries and also in higher institution of learning. Also, additional referencing styles can be incorporated to have a system which can be scalable for generic applications

References

[1]. B.A. Ojokoh, “Rule-based metadata extraction for heterogeneous references”, Oriental Journal of Computer Science and Technology 2 (2009).

[2]. Houssam Nassif, Ryan Woods, “Information Extraction for Clinical Data Mining: A Mammography Case Study”, *in 2009 IEEE International Conference on Data Mining Workshops (ICDMW). FL, USA*, pp.37-42, December 2009.

[3]. Bin Zhou, Yan Jia, “A Distributed Text Mining System for Online Web Textual Data Analysis”, *in Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC). Huangshan, China*, pp.1-4, October 2010.

[4]. Sushain Pandit, *Ontology-guided extraction of structured information from unstructured text: Identifying and capturing complex relationships*, Ames, Iowa: Iowa State University, 2010.

[5]. D. Carrell, D. Miglioretti, “Coding free text radiology reports using the cancer text information extraction system (caTIES)”, *In American Medical Informatics Association Annual Symposium Proceedings (AMIA). Rochester, USA,* pp.889-893, September 2007.

[6]. L. Rokach, O. Maimon, “Information retrieval system for medical narrative reports”, *In Proc. of the 6th International Conference on Flexible Query Answering Systems (FQAS). Lyon, France,* pp.217–228, June 2004.

[7]. D. Gupta, B. Morris, T. Catapano, G. Sautter, A new approach towards bibliographic reference identification, parsing and inline citation matching, in: Proceedings of the International Conference of Contemporary Computing, India, 2009, pp. 93–102.

[8]. Bolanle Ojokoh, Ming Zhang, Jian Tang, A trigram hidden Markov model for metadata extraction from heterogeneous references, Information Sciences 181 (2011) 1538–1551.

[9]. S.H. Papavlasopoulos, M.S. Poulos, N.T. Korfiatis, G.D. Bokos, A non linear index to evaluate a journal’s scientific impact, Information Sciences 180 (2010) 2156–2175