DLSU Research Congress 2022
De La Salle University, Manila, Philippines
July 6 to 8, 2022

A Corpus Management System for CCS Advanced Research Institute for Informatics, Computing and Networking (AdRIC)

Troy Mikael Esguerra[1], Maron Leonard Fabelico[2], Simon Kim[3], Kyle Rafael Lagado[4], Christine Diane Ramos[5]

Information Technology Department, College of Computer Studies, De La Salle University,
2401 Taft Avenue, 1004 Manila, Philippines
troy_mikael_esguerra@dlsu.edu.ph[1], maron_fabelico@dlsu.edu.ph[2], simon_kim@dlsu.edu.ph[3], kyle_lagado@dlsu.edu.ph[4], christine.diane.ramos@dlsu.edu.ph[5]

Abstract: The Advanced Research Institute for Informatics, Computing and Networking is the college unit that pursues the research ideals of the DLSU College of Computer Studies through the identification of its priority research activities in the field of computing and technology. Thus, it is vital for its researchers to have a centralized system when conducting systematic literature review to ease their challenges in gathering, organizing, and synthesizing a knowledge area across multiple research project engagements. With a corpus management system, research gaps are easily identified with the resulting system output of research analytics on a topic searched (including those of which they authored), using information extracts across multiple academic journal databases such as AIS, Scopus, and IEEE. This study concludes with a conceptual framework of a corpus management system that customizes to the needs of the research units, with other notable functions such as research analytics, annotation, and feedbacking.

Key Words: Corpus Management System; Literature Review; Research Analytics

1. INTRODUCTION

A corpus management system is a tool specifically designed to organize and group together components of a specific knowledge area or any linguistic corpora, especially that of a language that has a "complicated agglutinative morphology". It requires semantic analysis and lexical components of a certain query in order to group, extract, and display a set of related and relevant information upon search (Nevzorova, Mukhamedshin, Galieva, & Gataullin,2017). A corpus management system has been applied in several disciplines such as in personal health informatics (Epstein et al., 2015), but also in knowledge management and creation of structured literature review (Coners & Matthies, 2018). A corpus management system thus presents an opportunity for research centers such as the Advanced Research Institute for Informatics, Computing and Networking (AdRIC) to simplify research activities especially in scanning a wide body of knowledge and creating systematic literature reviews. In line with De La Salle University (DLSU)'s mission- vision to be a leading research university, colleges are encouraged to partake in contributing to this mission-vision by participating in research laboratories whose role is to support the multi-disciplinary, collaborative research activities committed to scientific excellence in the areas of information and computing technology. Thus, it is critical to introduce innovations in the field of information technology such as the development of a corpus management system to ease the conduct of research for ADRIC students and faculty in effect, increasing research productivity.

A series of interviews and surveys were conducted among the research center heads to determine their key challenges when doing their systematic literature review. The first issue is the lack of source tracking features in some academic databases. Researchers would like to annotate and bookmark articles to save their search. In the survey, the researchers stated that they want to be able to save the sources they have already gone through if relevant to them. Second, it is difficult to form a synthesis of the references collected due to a large volume of sources in a specific area across different journal/ academic databases, this is most especially difficult for a specific topic belonging to multiple disciplines. This corroborates with the study of Epstein et al. (2020) wherein users find difficulty in making sense of large field of personal informatics which appears in journals of information technology, medicine, nursing, etc. Third, researchers have to utilize multiple sources such as Scopus, ACM Digital Library, and other academic libraries. It would be more ideal if there was just one repository pulling the relevant articles in just one platform. This way, it would be easier to assess which discipline is heavily studied in particular research topic and which needs more contribution on. Gathering these key challenges have prompted for the group to design and develop a corpus management system as an opportunity to meet the desired state of the research centers when writing their systematic literature review. In being more specific, it is the objective of this research:

1. To understand the current process and determine pain points when conducting literature review from data collection to analysis
2. To map key challenges with corresponding key modules and functions to address the pain points enumerated
3. To design a conceptual framework of a corpus management system that is applied in the academic research, and
4. To develop a corpus management system fit to the systematic literature review process of ADRIC researchers.

2. RELATED WORKS

From the related works, corpus management system has been used in cultural studies such as preservation of language(Nevzorova et al., 2017), knowledge management such as the study of (Liu, Yang, & Wang, 2014) and (Coners & Matthies, 2018), and personal informatics such as the work of Li, et. al. wherein corpus management was applied to allow easier retrieval of user records for the purpose of self-reflection and gaining self-knowledge (Li, Dey, & Forlizzi, 2010). Corpus Management Systems have been applied in several domains but rarely in the

learning and education sector. One of the few is the study of the faculty medicine in University of Jaffna where an institutional repository for the faculty of medicine was developed to capture, store, index, and preserve the institution's scholarly output in digital format (Murugathas & Balasooriya, 2015). In the work of Shoeb (2010), an institutional repository (IR) was implemented for the Independent University, Bangladesh (IUB) to allow their members to publish their works and make it available publicly. Among the different repository software options available, and after a comprehensive analysis of case studies, DSpace was the software selected due to its lead in factors like operational analysis, schedule analysis, and economic analysis compared to other options. This system has been adopted in two other universities, namely Imperial College and Georgia Tech. University.

One key difference between a DSpace system and the proposed Corpus Management System is the application of analytical features which can aid researchers in their literature review (See Fig. 1). A DSpace system makes use of simple analytical data which can be viewed by its users. Seen in the table below is the list of analytics that most DSpace systems use today compared to the proposed system.

Features	DSpace	Corpus Management System for CCS AdRIC
Most Viewed	✓	✓
Most Downloaded	✓	
Recent Submissions	✓	✓
Most Searched Terms	✓	✓
Most Bookmarked Publications		✓
Top Country View	✓	
Top City View	✓	
Publications per Author and Area (based on keywords)		✓
Keyword analytics (number of publications w/per keyword)		✓

Fig 1. DSpace and Corpus Management System Features

3. METHODOLOGY

Given the limited studies on corpus management systems, the research was conducted using a qualitative research method for requirements gathering and problem analysis. The group first conducted a series of interview across the different researchers of the ADRIC laboratories to gather pain points in the system. This was then visualized using a the gap analysis as the problem analysis technique. In order to determine and develop the key functions needed for the proposed corpus management system, the researchers needed to map a function obtained from literature review/ related works in addressing the gap raised by the researchers. Once these were determined, the development of the system followed the activities of Agile methodology which is divided into the following phases: the requirements phase, system design phase, implementation phase, integration and testing phase, release phase, and review/iteration phase. Constant changes was also be applied based on the feedback and suggestions of the stakeholders of the system. The solutioning activities was validated against Li et al's study on personal informatics.

3. RESULTS AND DISCUSSION

From the series of interview, survey, solutioning activities and benchmarking done, the researchers have concluded the following main modules needed in the corpus management system: (1) Preparation module, (2) Collection module, (3) Integration module, (4) Reflection module which were aligned in the study of Li et al. (2010) depicting an iterative five-stage process (See Fig. 2). The researchers were then able to design a conceptual framework following the model of Li et al. (2010) which contained the mentioned modules.

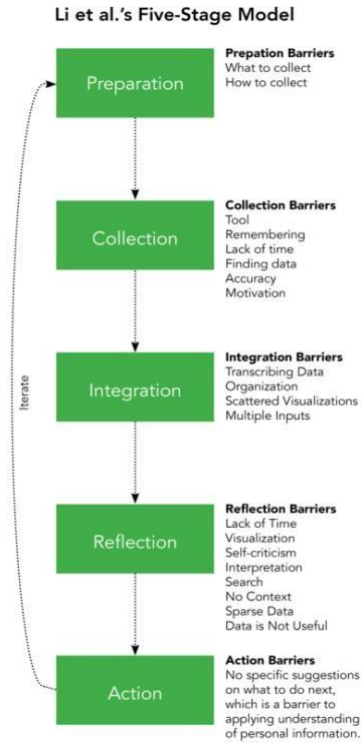

Fig. 2 Li et al's Five-stage Model

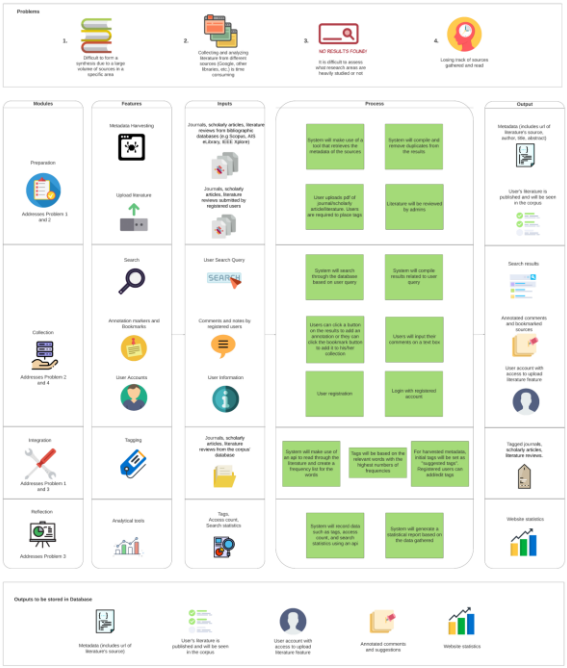

Fig. 3 Conceptual Framework

The Preparation module (See Fig. 4) contains features of the harvested metadata and the uploading of literature. Metadata of journals, scholarly articles, and literature reviews from Scopus, AIS eLibrary, and IEEE Xplore were harvested through Publish or Perish and csv exports from the corresponding databases (See Fig. 3). The output metadata was stored into the MySQL database, specifically in the publications table where each will have a unique primary key.

Fig. 4 Conceptual Framework Preparation Module

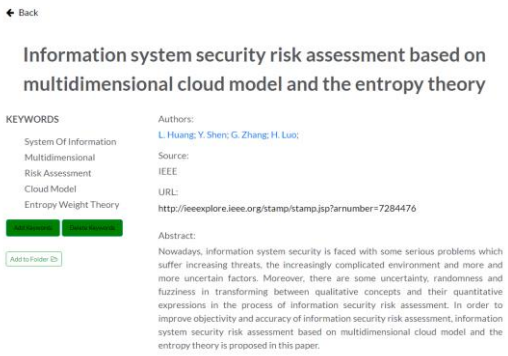

Fig. 5 Publication Page with Metadata

After getting the .csv extract from the academic databases, the researchers will select which among which of the articles are relevant and to be uploaded in the system. The criteria as to which articles are to be uploaded is based on the expert judgement of the researcher. For instance, in the game development research center, they have set research tracks such as "team development" or "3D modeling". This becomes their basis of filtering which articles to upload in the system. Some may filter also based on discipline (only including works found in Computer Science or Technology). This workaround addresses the system issue on loading time based on the high volume return of articles extracted from all three academic database which may reach up to more than 50,000+ articles. To

be able to access and upload the literature extract, users must register and login. Registering a user account requires information such as username, first name, last name, email, and password. When logging in, users need to input their email and password for verification.

The researchers can also opt to upload their own publications or articles by adding necessary information such as title, author, year, keywords, and the file (See Fig. 6). The uploaded literature will then have to be approved an administrative user to validate credibility of the article before it can be accessible through the system. This will be later on useful if the researcher would like to see research statistics (to be discussed further in the Reflection module) based on data on his or her own publications.

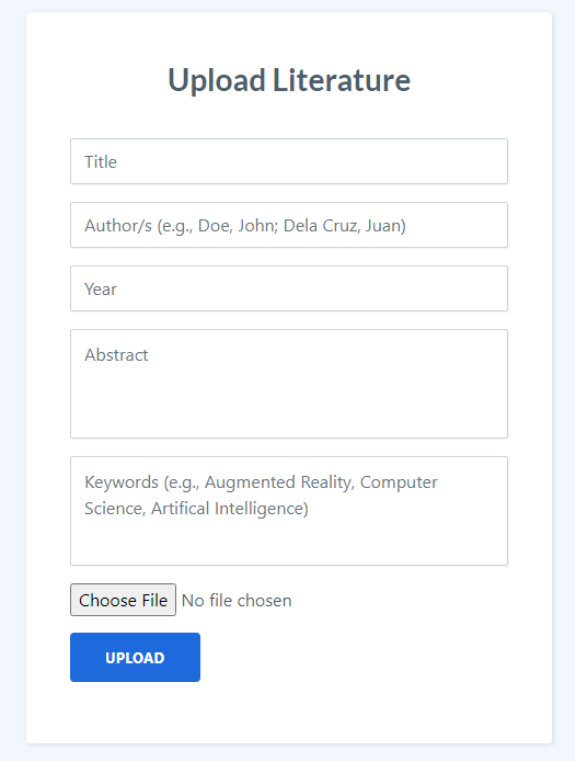

Fig. 6 Upload Literature Page

Another key module is the Collection module (See Fig. 7) which contains features such as search, annotations and bookmarks, and user accounts. Once confirmed, users will be able to search for a specific topic, author, name, or any key word in search bar and this would return articles pulled from the uploaded articles. (See Fig. 5). From the uploaded articles, the researchers can do annotations and bookmarks in their created folders (See Fig. 9). Collaborators can also be added into the created folders which will allow the collaborator to view the publications inside and add further annotations to the publications (See Fig. 10)

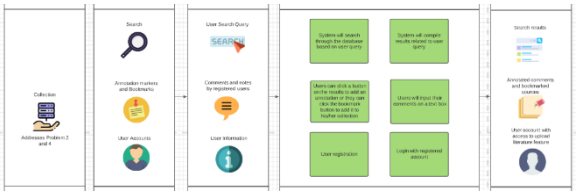

Fig. 7 Conceptual Framework Collection Module

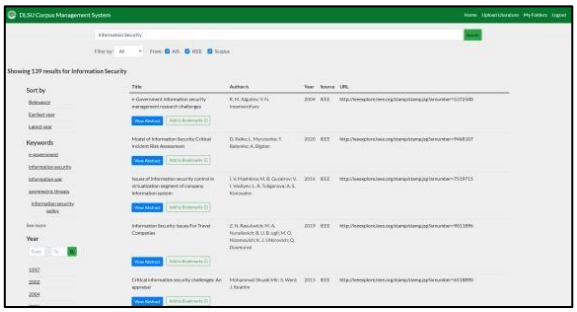

Fig. 8 Search Page with Results

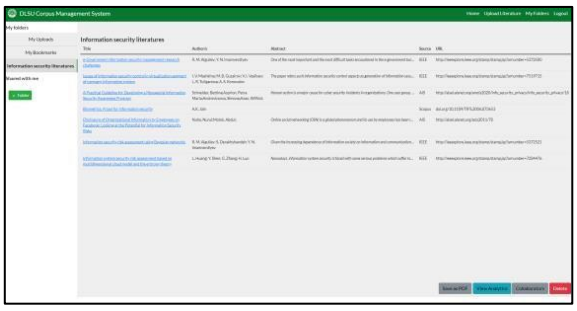

Fig. 9 My Folders Page

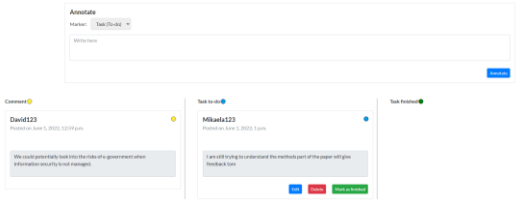

Fig. 10 Annotations Made by User and Collaborator

The third module is the integration module (See Fig. 11) which consists of the tagging feature allowing the auto generation of keywords for each publication through web scraping. The article's keywords will be used as the main source but if it does not exist, then the most frequent words from the article will be returned. The Natural Language Toolkit (NTLK) was used to remove stop words to ensure irrelevant frequently occurring keywords such as articles and pronouns will be omitted. For their own articles uploaded which are usually not found in the academic databases in scope, researchers are also able to log in their own keywords with validation from the administrator.

Fig. 11 Conceptual Framework Integration Module

Finally, the reflection module consists of the research analytics of the selected articles uploaded in the system from both the academic databases and researcher upload (See Fig. 12). At a glance, information such as the number of publications published per year, the number of publications present from each academic database, and the count for each keyword found in the folder are displayed (Fig. 13). The system also tracks user queries to display trending titles or keywords which can be found in the home page of the system. The most viewed keywords can also generate a page which will contain publications that share the same keyword.

Fig. 12 Conceptual Framework Reflection Module

Fig. 13 Word Cloud on Top Keywords

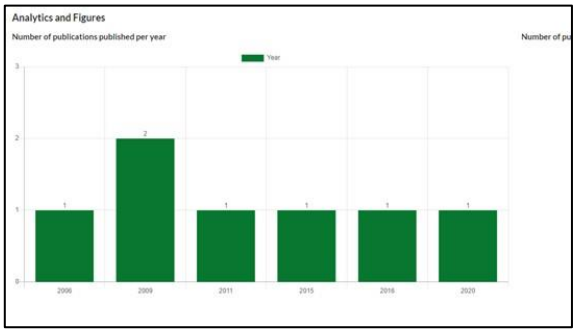

Fig. 14 Publication Trend on Keyword Search within the past 5 years

The analytics will also help give the researchers an understanding of what the body of knowledge and know the trends within the topic searched. In the above example, the researcher wanted to know about the literature on the topic "security. Based on the returned results on selected uploaded articles, we can conclude that security research is prominent on risk assessments and information security. With this information, researchers can explore writing on limited security topics such as that of hardware protection or security violation threats. Furthermore, the same information can be concluded with author information by creating a folder specifically analyzing all articles written by the researcher. Other reports for this module include knowing which author is most prominent in the field, and which discipline the topic is mostly written about.

To run the system the following hardware and software are needed: 8GB of ram, Quad-core processor, Windows Operating System Version 8.0 or higher, and an up-to-date internet browser (Google Chrome, Safari, Mozilla Firefox, or Microsoft Edge). The system is designed to be accessed through web browsers for desktops and not mobile. The design follows a responsive approach for desktop web browsers to cater to different resolutions. The system

was also programmed using Django, a Python-based web framework as the developers are more familiar with the Python language.

5. CONCLUSIONS & RECOMMENDATIONS

With the use of the proposed system, researchers in the academic community of the AdRIC and the College of Computer Studies will be able to conduct their literature reviews more systematically and efficiently. The repository of the system will be able to help researchers gather publications from three different academic databases at once. The system will also be able to help researchers organize their literature by allowing them to group them through folders that they can create. Keywords for each publication will also be available for easier navigation. Lastly, the analytics of the system will be able to help generate ideas for possible research areas that could be explored.

The group would like to recommend future researchers to address system limitations such as direct collection from the databases and to consider further technical solutioning to reduce processing time. Moreover, the Action module can also be extended given that in this study, the researchers sought the action module not needed as it entailed integration to multiple other business processes that interfaces with other external stakeholders.

6. ACKNOWLEDGMENTS

The authors would like to express its gratitude to Mr. Oliver Malabanan for his valuable input in forming the conceptualization of a corpus management system applied in an academic setting. Moreover, we would also like to thank our technical panels' Ms. Marivic Tangkeko and Dr. Michelle Renee Ching for their expertise in evaluating and providing feedback in terms of the system's functionality, relevance and usability.

7. REFERENCES

Coners, A., & Matthies, B. (2018). Perspectives on reusing codified project knowledge: A structured literature review. International Journal of Information Systems and Project Management, 6(2), 25–43. https://doi.org/10.12821/ijispm060202

Epstein, D. A., Ping, A., Fogarty, J., & Munson, S. A. (2015). A lived informatics model of personal informatics. UbiComp 2015 - Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, (July), 731–742.

Epstein, D. A., Caldeira, C., Figueiredo, M. C., Lu, X., Silva, L. M., Williams, L., Lee, J. H., Li, Q., Ahuja, S., Chen, Q., Dowlatyari, P., Hilby, C., Sultana, S., Eikey, E. V., & Chen, Y. (2020). Mapping and taking stock of the Personal Informatics Literature. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 4(4), 1–38. https://doi.org/10.1145/3432231

Li, I., Dey, A., & Forlizzi, J. (2010). A stage-based model of personal informatics systems. Conference on Human Factors in Computing Systems - Proceedings, 1(January), 557–566. https://doi.org/10.1145/1753326.1753409

Liu, Y., Yang, D., & Wang, Y. (2014). A semanticbased knowledge management platform. Proceedings - Pacific Asia Conference on Information Systems, PACIS 2014

Murugathas, K., & Balasooriya, H. (2015). Developing an institutional repository: Experiences at the library, Faculty of Medicine, University of Jaffna. Journal of theUniversity Librarians Association of Sri Lanka, 18(1), 39. doi:10.4038/jula.v18i1.7860

Nevzorova, O., Mukhamedshin, D., Galieva, A., & Gataullin, R. (2017). Corpus management system: Semantic aspects of representation and processing of search queries. 2016 7th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications, SETIT 2016, 285–290. https://doi.org/10.1109/SETIT.2016.7939881

Shoeb, Z. (2010). Developing an institutional repository at a private university in Bangladesh. OCLC Systems & Services: International digital library perspectives, 26(3), 198-213. doi:10.1108/10650751011073634