Title	Features enhancements and text analytics on institutional repositories
Authors	Jacky Wong Kae Perng
Source	Paper presented at the 4th International Conference of Asian Special Libraries (ICoASL 2015), 22 - 24 April 2015, Seoul, Korea

Features Enhancements and Text Analytics on Institutional Repositories

Jacky Wong Kae Perng
National Institute of Education, Singapore
kaeperng.wong@nie.edu.sg

Abstract

As part of National Institute of Education (NIE) Library's LIBRIS 21 strategic master plans to transform the Library, NIE Library seized the opportunity to engage users and value-add to academics' research ecosystem through the implementation of an Institutional Repository in 2009. With the paradigm shift in research dissemination through institutional repository and the ease of copyright restrictions imposed by publishers, along with the innovative content recruitment efforts by NIE Librarians, the institutional repository become highly popular and significant content growth was observed since the institutional repository was first launched. Within a short span of four years, the NIE institutional repository has gathered more than 15,000 digital objects. Being encouraged by the participation rate, NIE Library seized the opportunity to further enhance users' experience by upgrading NIE Institutional Repository platform (DSpace) to the latest version in September 2014. This paper will share NIE Library's experience in the upgrade, present selected advanced features of the new version and show case how they add value to our users. In addition, with the accessibility of text analytics technologies and the fact that institutional repository have huge amount of valuable textual information, NIE has come up with several text analytics prototypes that could further enhance user experience. This paper will detail the features provided by text analytics and how they are developed using text analytics technologies. Specifically, text analytics applications such as subject terms predictions, trend analysis, articles recommendations and contextual search terms suggestions are explored in this paper.

Keywords: *Articles recommendation, Institutional repository, Subject prediction, Text analytics, Usage analysis*

1. Introduction

As part of National Institute of Education (NIE) Library's LIBRIS 21 strategic master plan to transform the Library, several innovative information services were identified as strategic initiatives to enhance library users experience with the Library (Wan-Yeoh, 2010). One information service that emerged is the implementation of institutional repository.

An institutional repository (IR) is an online archive for collecting, preserving, and disseminating digital copies of the intellectual output of an institution, particularly a research institution such as academic universities (Lynch, 2003). With the maturity of digital repository technologies and the strong demand for an open access publishing platform, NIE Library launched an institutional repository that runs on the open source DSpace platform in 2009.

Riding on the paradigm shift in research dissemination through institutional repository and the ease of copyright restrictions imposed by publishers, along with the innovative content recruitment efforts by NIE Librarians, the institutional repository become highly popular and

significant content growth was observed since the institutional repository was first launched. Within a short span of four years, the NIE institutional repository has gathered more than 15,000 digital objects. Extrapolating from past content growth trend, the NIE Library anticipates a growth rate of 15% per year for the next three years.

Being encouraged by the participation rate, NIE Library seized the opportunity to further enhance users' experience by upgrading NIE Institutional Repository platform (DSpace) to the latest version in September 2014. Moreover, having observed the recent trends in text mining and analytics applications, and knowing that NIE institutional repository contains a valuable source of textual information; several text analytics prototypes were developed and explored. This paper will share NIE Library's experience in enhancing institutional repository, as well as introducing text analytics prototypes for discussion, exploration and future developments.

2. Backgrounds

Founded in 1950 as a Teachers' Training College library, the NIE Library and Information Services Centre is the largest education library in Singapore. It provides resources to support the teaching, learning and research programs of the Institute. It also reaches out to education professionals in the Ministry of Education, schools and other related organizations. The Library serves about 2,500 full-time students and more than 7,000 library members.

NIE Library leverages extensively on technology in providing library services to its patrons. The implementation of DSpace version 1.5 in 2009 as the platform for NIE Library's institutional repository is one case in point. While DSpace version 1.5 served NIE Library's well, there are limitations such as lack of support for mobile devices, rudimentary usage statistics reporting, lack of discoverability features such as facets browsing and inadequate support for Application Programming Interfaces (API) to extend and enhance the core capabilities of that particular version of DSpace.

Having witnessed healthy content growth and how the institutional repository value-add to Library users, NIE Library made the strategic decision to further enhance users' experience by upgrading the NIE digital repository. In October 2013, a task force was formed to review and explore the opportunity to enhance NIE digital repository. The task force is formed.

The task force reviewed several institutional repositories that had been launched in regional libraries, including institutional repositories from National University of Singapore, Nanyang Technological University, Hong Kong University and Singapore Management University. In addition, the team also reviewed published articles on this topic; *Library Success: A Best Practice Wiki*, which has a listing of libraries with great success in implementing institutional repositories was a particularly useful resource to learn about best practices in other Libraries (Library Success, n.d.).

In early 2014, the latest version of DSpace was identified as the platform that most suit NIE Library's needs and with that, the task force began the arduous task of upgrading NIE digital repository to the latest version. And finally, in September 2014, the enhanced NIE digital repository, running on the DSpace 4.1 platform was launched to the public users.

3. Enhanced institutional repository features

The enhanced institutional repository offers various critical advanced features. One of the most significant enhancements is the ability of the NIE Digital Repository to adopt Responsive Web Design methodology such that the web application will change by adapting to the screen size of devices, regardless of whether it is desktops, tablets, or mobile phone devices.

With the exponential growth in the use of mobile devices, there is increasing demand by users to be able to access information seamlessly at the convenience of their mobile devices. By having this feature enhancement, users can easily access the content seamlessly in NIE Digital

Repository anywhere, anytime, on virtually any devices. Figure 1 provides a snapshot of how the repository interface adapts and changes the layout of various components on desktops, tablets and mobile phone devices.
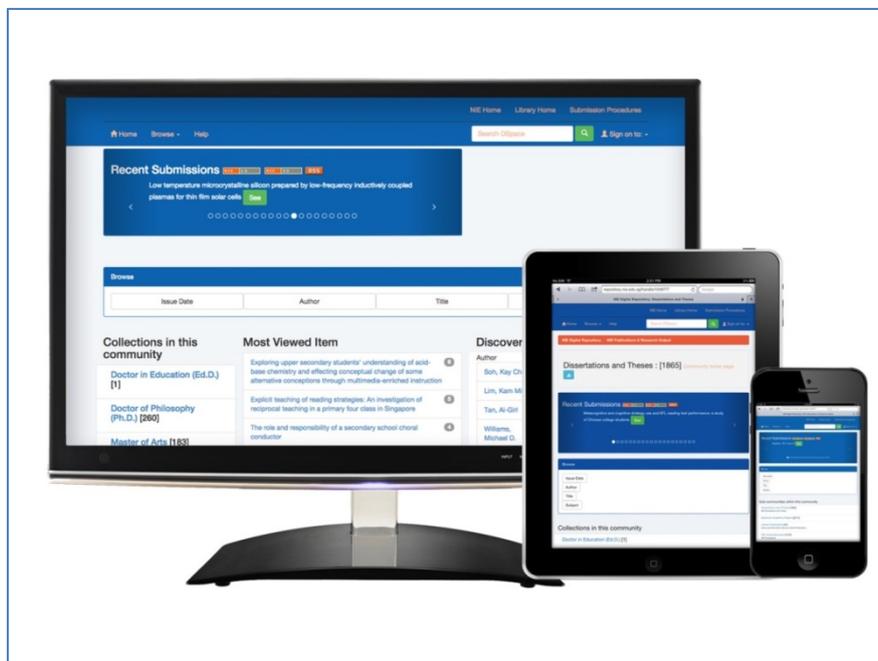

Figure 1. Responsive web design of the enhanced institutional repository

In addition, the enhanced repository also provides features that enhance the discoverability and accessibility of digital contents. For instance, users can now easily view recent submissions to collections, subscribe to RSS feeds so as to be alerted when new content is published and perform content filtering by Author, Date and Most Viewed Items facets. Figure 2 provides a snapshot of the enhanced interface and features that allow users to easily access digital contents.
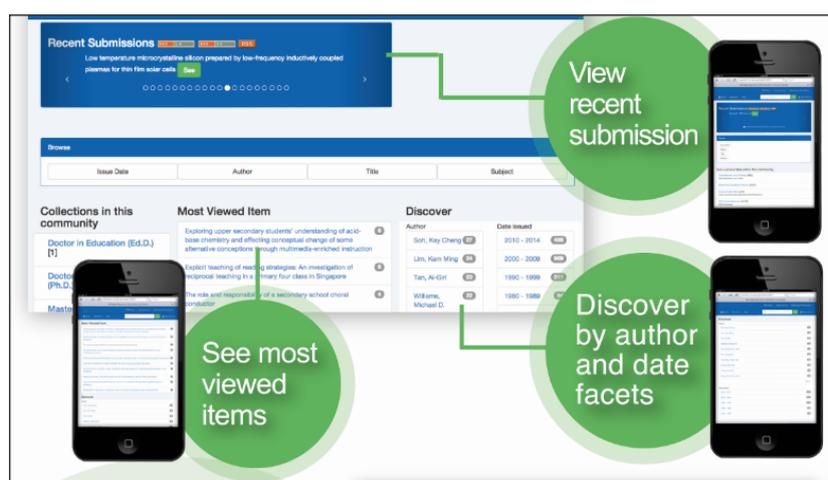

Figure 2. RSS feeds & content filtering by author, date, and most viewed item facets

Moreover, the enhanced Digital Repository also comes with several new features that serve to market and raise awareness on NIE research output. For instance, the digital repository provides the ability for anyone to view the usage statistics of a particular research paper. By having public access to usage statistics, a user will be able to gauge the usage and popularity of the research paper.

In terms of backend workflows, the enhanced digital repository provides the means for metadata librarians to easily perform batch quality checks and bulk editing of metadata. There are also embedded workflows for users to request access to restricted papers and for Librarians to track the approval processes involved in granting users access to such restricted papers. Lastly, the enhanced repository also provides a built in embargo feature that allows Librarian to specify the embargo period, which is the period during which articles published cannot be made publicly available online. The embargo feature works such that access restrictions would be automatically lifted once the embargo period specified is passed.

4. Comparison with other Institutional Repositories

Comparing the enhanced NIE repository with that from other repositories in Singapore, it was observed that the other local institutional repositories in Singapore, such as those from National University of Singapore, Nanyang Technological University and Singapore Management University has long emphasized on usage statistics. For instance, the repository in National University of Singapore repository imported critical citations information from Corpus while Singapore Management University while Singapore Management University repository has begun incorporating Altmetrics in the papers. This is an area where NIE repository can significantly expand on to further value add to researchers.

On the other hand, as NIE repository has upgraded to the latest version, the NIE repository is in a better position to leverage on the enhanced technologies in the latest version on DSpace. For instance, the NIE repository is comparatively superior in rendering the interface on mobile devices, better able to assist users to discover digital objects through facets and search terms suggestions.

5. Exploring Altmetrics (Alternative Metrics)

Evidently, the enhanced repository not only significantly improved user experience through various advanced features but also facilitated productivity improvements by automating several backend workflow processing procedures. However, much more can be achieved by extending the core capabilities of the latest version of DSpace through the flexible and open architecture provided by the Application Programming Interfaces (API) and Web Services.

Witnessing this capability, the team further began exploring possibilities of further enhancements through API and Web Services. One key result from the exploration studies is the observation of how the University of Pittsburgh not only incorporates usage statistics but also various elements of Altmetrics (Barnett, Collister, Brittany and Chan, 2014). Altmetrics, or alternative metrics are essentially new measurements for the impact of scholarly content. Altermetrics is "based on how far and wide the article travels through the Social Web (like Facebook & Twitter), social bookmarking (e.g. CiteULike) and collaboration tools (such as Mendeley). What Altmetrics hope to do is provide an alternative measure of impact, distinct from the Journal Impact Factor, which has been categorically misused and is unable to respond to the digital environment that scholarship takes place in today." (Finbar Galligan & Sharon Dyas-Correia, 2013)

By including Altmetrics in institutional repository, users would be able to assess the value of the papers by reviewing the following interactions with the papers:
- Captures – bookmarks of articles via tools such as Delicious, Pinterest and CiteULike.

- Mentions – reviews of articles in blog posts, comments and Wikipedia.
- Social media – discussions of articles in Social Media such as Google Plus, Likes and Shares in FaceBook, Tweets.
- Citations – traditional metrics like citations count on the articles.
- Usage – clicks, downloads, views and holdings information.

Altmetrics therefore gives a more comprehensive and dynamic view on how the paper is being utilized. While this is definitely a good feature to have, the scale of implementing this feature involves sophisticated big data analytics tools such as PlumX from Plum Analytics. Detailed evaluations of the tools have to be carried out before committing investments to acquire the tools.

6. Exploring text analytics

In the continuous efforts to introduce new features to enhance users' experience, one other suggestion was to enhance the taxonomy such that the navigation feature provides the means for users to search and browse the collection by subject area in addition to keyword searching. The benefits of having subjects in the taxonomy are that users can "navigate from need to resource consistently and quickly" and a good taxonomy also "allows an organization to inventory and monitor knowledge based on a structured understanding of user and community needs." (Pack, 2002)

However, a key constraint in implementing such a feature is the resource required by the Librarian to carefully read through the abstract to assign relevant subject headings. The effort to assign subject headings is not merely a one-off exercise, as one need to continuously do so for new papers that got published in the institutional repository. The workload can be quite substantial towards the end of academic semester when huge volume of dissertations is sent to the Library for processing and publication in institutional repository.

To overcome this resource constraint, text analytics can be employed to perform subject terms prediction. Using a software tool called RapidMiner, one can harvest valuable textual information such as Title and Abstract to predict the subject terms to describe the document. Two approaches are explored, automatic subject classification method using supervised machine learning model and unsupervised machine learning model.

6.1 Automatic subject classification (using unsupervised machine learning)

There are two main modes of machine learning algorithms, supervised and unsupervised machine learning. In supervised machine learning, the prediction of new subject terms is based on past assignment of subject terms, which are used to train the machine to understand the underlying pattern. In unsupervised machine learning, the aim is to find underlying patterns in the data using techniques such as cluster analysis (Upton & Cook, 2014).

The use case for using unsupervised machine learning method in institution repository is the case where it is a completely new repository or one in which all the academic papers in the repository do not have any subjects assigned. In this case, to speed up the assignment of subjects, one can use unsupervised machine learning method. Using unsupervised machine learning techniques, we can cluster similar abstracts by computing the similarity scores between abstracts to group similar abstracts together.

To simulate the application, a sample of 120 abstracts from each of the five subjects, namely economics, education, geography, history and mathematics are extracted to run through the unsupervised learning process. Using RapidMiner, the software tool retrieves the 600 abstracts (120 from each of the 5 subjects), runs through some text processing tasks such as transform cases and tokenization, creates word vector using term occurances and makes use of K-Nearest

Neighbor algorithms and Consine Similarity measure to group similar abstracts together into various clusters.

The RapidMiner software also provides details about the clusters. For instance, by examining the details in each cluster, for example in figure 3, we examine cluster_4.

subject	cluster	text
economics	cluster_4	technical vocational education training tvet understanding nigerian experience descriptive mixed met
economics	cluster_4	essays financial crises foreign direct investment cross national coordination examining ocean value
economics	cluster_4	essays international trade economic development dissertation composed self contained chapters ir
economics	cluster_4	effects globalization panamanian university system global crossroads panama subject effects globa
education	cluster_4	effect globalization multinational corporations primary secondary schools costa rica implications edi
education	cluster_4	policy analysis addressing st century learning project form policy analysis model purpose determini
education	cluster_4	servant leadership predictor student academic growth faculty staff attendance job satisfaction organ
education	cluster_4	misunderstanding miscommunication stakeholders interpret teacher evaluation process louisiana
education	cluster_4	music education reform taiwan beginning music teachers perceptions teacher preparation national
education	cluster_4	organizational leadership teacher work conditions teachers job satisfaction multi country analysis n
education	cluster_4	establishing marketing international branch campus case study savannah college art design hong
education	cluster_4	effects character education incidences physical altercations eighth grade boys dissertation focuses
education	cluster_4	investigating perception efl teachers thailand extent feel prepared teach study examined thai teacher
education	cluster_4	teachers perceptions professional learning community model professional development profession
education	cluster_4	teachers perspectives effective lesson planning comparative analysis study investigated perception:

Figure 3. Results from cluster_4

One can deduce from the abstract text that cluster_4 would have education as the main subject. Although the model clustered several economic abstracts wrongly together with education abstract, the automatic cluster algorithm have for the most parts clustered the abstracts correctly and would serve as a good starting point to automatically assign subjects to abstracts.

One challenge with using the model is with execution time and accuracy of prediction. Using the samples, the execution time took 4 minutes to complete and the accuracy of prediction is only 72%. The execution time would increase exponentially when we add more abstracts and subjects to the cluster. Nonetheless, the automatic classification using unsupervised learning approach would have saved librarians substantial time in assigning subject labels manually.

6.2 Predicting subject terms (using supervised machine learning)

In order to simulate the supervised machine learning approach, dissertations that the librarians had assigned the main subject previously are used as the training data for machine learning. 150 dissertations from each of the eight subjects, namely economics, education, mathematics, geography, history, chemistry, physics and biology are prepared, giving us a total of 1200 dissertations. For each subject, the title and abstract of the dissertations are concatenated to give us a combined text to perform text analytics.

The RapidMiner software harvests the textual information, identifies the most frequent words used in each subject, and use algorithm such as Naive Bayes classifier to predict the subject terms. Using cross validation technique, the RapidMiner software can provide metrics to indicate the accuracy of the subject terms prediction. Using our sample abstracts, the supervised machine learning technique is able to achieve an accuracy of 82.50% as shown in figure 4.

Figure 1. Accuracy metrics of prediction

Evidently, there are instances when the machine learning application predict the subject terms wrongly. This is especially the case when abstracts from the two subjects use very similar terms. One way to improve the accuracy is by detecting abstract that may be on the "borderline" between two similar subjects using weighted scores of the prediction models.

As an example, abstracts from "Geography" subject is compared against abstracts from "History" subject, the RapidMiner is able to compute similarity scores of the abstracts. As shown in figure 5, the normalized score provide important clues on the abstract; scores above 60 have all abstracts with Geography as the subject and scores below 40 have all abstracts with History as the subject. The borderline scores between 40 and 60 then serve as a good starting point for the Librarian to examine the prediction versus the actual subject. By doing so, one can eliminate the need to perform manual quality checks on straight forward cases but instead focus the effort on borderline cases, thus improving the efficiency.

Figure 5. Normalized score for predicting geography vs. history

In addition, RapidMiner software can also compute the weight of words for a particular subject compared to other subject. By aggregating the scores, the RapidMiner software can determine the confidence of predicting particular subject versus other subjects.

One can infer whether an abstract should have multiple subjects from the confidence score. For instance, if an abstract scores highly in both geography and history but not the other subjects, one can infer that the abstracts have both geography and history as the subjects. Another alternative is to set a threshold confidence level; for instance if the confidence of predicting a subject exceed a certain threshold value, the subject will be assigned to the abstract. Doing so allow one to automatically select all subjects, including multiple subjects that exceed the confidence threshold for a particular abstract.

7. Other enhancements

7.1 Contextual search terms suggestions and trend analysis

Information from the search terms and predicted subject terms can be further fed into a controlled vocabulary dictionary that the institutional repository used to provide contextual search terms suggestions. Figure 6: provides a screenshot of the contextual search terms suggestions provided by the system as user types "e" in the text box.
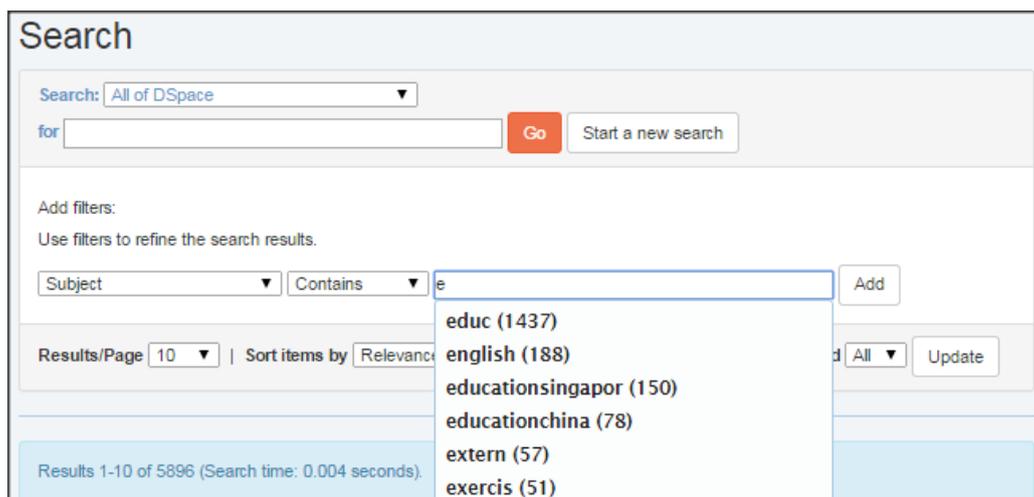
Figure 6. Contextual search team suggestions

Similarly, the subject terms and the search and browse patterns of the subject terms can be further analysed using standard analytics software such as Google analytics to observe the trending subject areas that the users are interested in over a period of time.

7.2 Check duplicates and detect plagiarism

In addition to using text mining to predict subject terms, librarians can also use text mining applications to check for duplicated records in the institution repository so as to ensure that the same papers don't get published again. The same check can also be used to clean up existing duplicate records in the institution repository. This can be achieved by using the RapidMiner software which can compute similarity scores between the papers. Duplicated papers would have a similarity score of 100%. Using similar process also enables Librarians to detect signs of plagiarism. Librarians can set a cutoff, say a similarity score of 50% and above to flag out abstracts which are so similar that they demand further investigation to check if there is any plagiarism involved.

7.3 Article recommendation

While the above processes would serve well to improve the productivity of Librarians to process and publish the dissertations in institution repository, the key point of having the institution repository is to have users access them. To further enhance access, the system can make use of text mining techniques to provide an article recommendation feature. For instance, when a user is reading the abstract of an article in the institution repository, the system can compute the top 10 articles related to the abstract and show the user the links to these articles. This can be easily achieved by using RapidMiner software, which can compute the similarity scores of the abstracts. The similarity scores can then be sorted such that the top 10 articles related to a particular abstract are flagged out as recommended articles for the users.

8. Conclusion

With the introduction of the enhanced features in NIE institutional repository, the Library is optimistic that the value-add to users and academics' research ecosystem would make a significant impact. The Library is currently gathering and monitoring the usage statistics of the enhanced institutional repository to gain insights on users' usage patterns.

In addition, as illustrated in this paper, through the combination of API and Web Services, one can extend the core capabilities of the DSpace platform. For instance, one can embed Altmetrics into the articles and employ text mining techniques to improve the work processes of Librarian in managing the institutional repository. As described, text mining techniques can be employed to predict key subject terms, enhance contextual search terms suggestions and trends analysis, detect duplicates and plagiarism, and finally, improve users' experience by enabling the article recommendation feature.

While the machine learning techniques investigated did not achieve 100% accuracy, the prediction models we explored do achieve an accuracy of more than 70%. The prediction models can therefore reduce some of the manual work such as subjects tagging and quality check work that is to be done by the Librarians. Moreover, by examining the weighted scores of the terms generated by the prediction models, Librarians would have some clues on where to look for possible prediction errors, again cutting down the time to perform manual quality checks substantially. All of these enhancements would greatly help to reduce the bottlenecks to publish dissertations and theses in institution repository significantly, giving users timely access to valuable research. In order to improve the accuracy further, perhaps the institutional repository can provide a feature for users to provide manual feedback on the predicted subject, doing so allow one to tap on the wider pool of resource to further improve the system. With these strategies in place, Libraries would truly be able to do more with less when it comes to managing the institution repository.

The NIE Library is committed towards meeting the needs and demands of NIE's stakeholders as well as their changing priorities and lifestyles. As opportunities for further developments in institutional repositories are plentiful, we will continuously strive to increase accessibility, discoverability and usability of information by leveraging on the latest technologies. This is in line with the LIBRIS[21] Initiative, our five-year library strategic plan to transform the NIE Library into a teacher education library relevant to the 21st century.

References

Barnett, John H., Collister, Lauren B., and Chan, Jennifer. (2014). Something Widget This Way Comes: Using Altmetrics to Illustrate the Impact of Open Access on Graduate Student Research. In Western Pennsylvania-West Virginia Chapter of the Association of College and Research Libraries (WPWVC/ACRL) Spring Conference 2014, 6 June 2014, Bethany, West Virginia.

Dill, E. & Palmer, K. L. (2005). What's the Big IDeA? Considerations for Implementing an Institutional Repository. *Library Hi Tech News,* 22(6).

Finbar Galligan & Sharon Dyas-Correia. (2013). Altmetrics: Rethinking the Way We Measure, *Serials Review*, 39(1), 56-61.

Hofmann, M. & Klinkenberg, R. (2014). RapidMiner: Data Mining Use Cases and Business Analytics Applications. CRC Press.

Library Success (n.d.). Institutional Repositories. In Library Success: A Best Practices Wiki. Retrieved July 9, 2014. from http://www.libsuccess.org/Institutional_Repositories

Lynch, Clifford. (2003). Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age. Association of Research Libraries Report. Retrieved November 11, 2014. from http://www.arl.org/storage/documents/publications/arl-br-226.pdf

North, M. (2012). Data Mining for the Masses. A Global Text Project Book.

Pack, T. (2002). Taxonomy's Role in Content Management. *EContent,* 25(3), 26-31.

RapidMiner Documentation. (n.d.). Retrieved November 11, 2014. from http://rapidminer.com/documentation/

Upton, G. & Cook, I. (2014). Machine Learning. In A Dictionary of Statistics. Oxford University Press. Retrieved November 17, 2014. from http://www.oxfordreference.com.ezlibproxy1.ntu.edu.sg/view/10.1093/acref/9780199679188. 001.0001/acref-9780199679188-e-2380.

Wan-Yeoh, S. K. (2010). LIBRIS[21]—Transforming LIBRIS for the 21st Century. Library and Information Services Centre, National Institute of Education, Singapore.

Weiss, S. M., Indurkhya, N., & Zhang, T. (2010). Fundamentals of Predictive Text Mining. London: Springer-Verlag.