IFLA WLIC 2013
SINGAPORE

Submitted on: June 28, 2013

Data Mining for Scholarly Journals: Challenges and Solutions for Libraries

Martha A. Speirs
Azerbaijan Diplomatic Academy
11 Amadbay Agha-Oglu Street
Baku, Azerbaijan AZ 1008
E-mail:mspeirs@ada.edu.az

Abstract:

As our global knowledge environment changes and the information to be found in scholarly journals becomes increasingly available in digital format, it is necessary to employ more and more sophisticated search and retrieval procedures to mine this knowledge. We have large holes in our globally accessible knowledge base as traditional web-crawlers cannot collect and assess all of the serially produced papers, articles and journals that exist. Many search engines only touch the surface and they cannot harvest potentially valuable information in the silos of the "deep web". More comprehensive data mining is therefore essential if we are to effectively tap the knowledge often hidden in scholarly journals and databases. Data-mining models are being developed which aim to search all the global knowledge being produced--an essential goal that will aid in sharing and therefore accelerating global knowledge diffusion. Deep Web Technologies and World Wide Science.org are examples of ongoing efforts to assist in mining the rapidly increasing mass of serially produced scientific information. Knowledge can only be shared, advanced and accelerated if it is accessible and as users expect libraries to be ever more effective in gathering and utilizing knowledge they must serve the global community by offering the best access to and analysis of all information. This paper intends to contribute to a more comprehensive understanding of what information is potentially available and how to access and analyze it using the latest methods of information retrieval.

Keywords: academic journals, information discovery and retrieval, cross-lingual information retrieval, Deep Web, text mining

As one of the key initiatives in accord with IFLA's strategic plan is enabling equitable access to information and since scholarly journals represent a critical body of knowledge compiled within an evolving knowledge economy, then access to this information is essential. Libraries should strive to offer seamless access to valuable information; but many challenges to full global access exist and solutions need to be explored. This paper focuses specifically on information retrieval in a global, multilingual sphere and takes a look at solutions that will enable equitable access for the end-user.

Knowledge may be difficult to access for various reasons; because it has not been published in digital format; because it is not written in a widely known language or script; because it comes from another academic environment, or because it is hidden in the "deep web" which is not accessible by using only simple web searches.

Although many live in a digital world, it is important to be aware of the 34% percent of global scholars who do not have access to digital versions of scholarly output on the Internet. (World Internet Penetration Rates) A digital divide exists between areas of the world where print is still the common format for scholarly journals. The scholars who do not have access to the Internet, either because of lack of sufficient bandwidth or funding for electronic databases cannot join the upsurge now taking place in science and other fields. They are not able to move ahead by researching and by publishing in academic journals, which are now, more often than not, available in digital format in the Western world. Digitized information is invisible to those without Internet access, but scholarship in print format, is often hidden to others because it has not yet been published in digital format and therefore is only available to those who can access the print version, which may be less available on the more developed global scene.

Knowledge contained in scholarly journals is also often hidden because the articles are written in a language or script that is unknown to the researcher. Figure 1 shows the top content languages used in the Internet and even though English is still the most commonly used language for Internet content, the presence of other languages is significant and is growing rapidly. Large well-known databases of scholarly journals are found predominantly in English-speaking areas where the Internet has penetrated most heavily and as most of the journals are written in the most prevalent language, English, these journals do not offer access to scholars who do not understand English but who are actively carrying out research in a different language and on a different level.

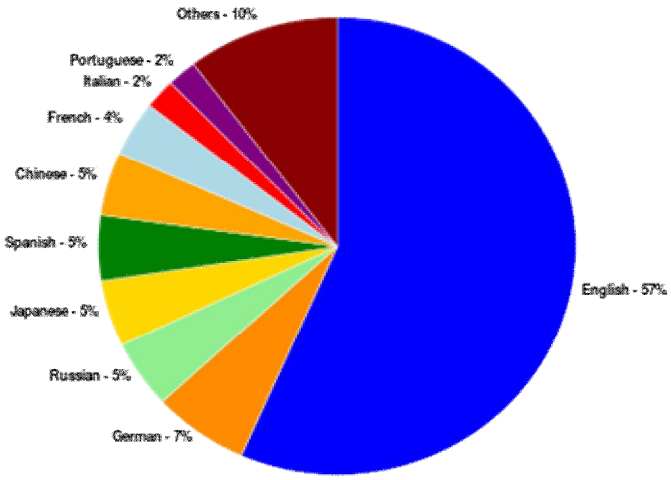

Figure 1
"Usage of content languages for websites", W3Techs.com [retrieved 24 March 2013]

The research and publishing that they are producing may be hidden from many in areas of the globe, which are on the other side of the digital, or language divides. In addition, many academics are carrying out research in their local languages and using print journals or electronic databases that are removed from the mainstream English language favorites. It is necessary to consider the global presence of online scholarly journals and their articles with a multilingual perspective, as scholarship is increasingly being published in many new areas of the world, in languages other than English and in scripts other than Latin.

A search for articles in English only is not as globally comprehensive, since the Internet has now penetrated many non-English speaking cultures and new databases in many languages increasingly contain new treasures to be mined. Around two-thirds of all scientific publications were written in English in 2005 (Prado 36) but the rate of publications in other languages is steadily increasing. Linguistic diversity of cyberspace is increasingly becoming a priority issue for building inclusive knowledge societies and this diversity will allow for the spread of truly global knowledge once it has been successfully mined. (Pimienta 55) Libraries can be a force for change if they offer access to quality serial publications in a variety of languages and not just those found in the easily available English-language databases.

Researchers need to be aware of how to access databases, which are not in the mainstream, and when they do mine these databases they need to be able to evaluate the quality of the articles. More work and awareness in the field of citation analysis will help in this quest. It is difficult to find statistics with a high level of validity for many extant databases and journals in languages that are less commonly used, but it is

quite clear that those who do not speak or read one of the top ten languages would have difficulty being able to interact with the current pool of global knowledge. One would have to learn one of these more commonly used languages well or be excluded from the information hidden in these databases.

There are other considerations that contribute to knowledge being easily accessible besides the language of the scholarly articles and the existing Internet penetration in a specific area. The academic environment surrounding scholarly journals and the quality of an author's output differs across countries and regions and this can also contribute to a type of digital divide. A closer look at the publishing output and citation analysis of a Russian database as measured against a Western-focused database illuminates this further. Oleinik has carried out a comparative analysis of scholarly articles and citations produced in Russian and Western academic environments and found several findings on both the Russian *eLibrary* and the more Western *Web of Knowledge* databases that reflect the effect that varying environments of science have on the global accessibility of scholarly articles. The Russian impact analysis tool (RINT) measured several factors in the *eLibrary* journal articles and found that Russian scholars tend to publish their work in journals published by their universities, to cite fewer sources, to have fewer and older references in their papers and that they tend to co-author with others in their own institutions. (Olienik 549) These factors exhibit a more closed and less international academic environment where the scholarship is hidden from much of the world, as it is less likely to be contained in the more accessible *Web of Knowledge, Scopus* or in other mainstream Western databases. The Russian *eLibrary* database http://elibrary.ru contains more than 6,700 science journals but only 8% of the *Web of Science* is made up from the 149 Russian journal publications. This raises the question as to whether this knowledge is relatively hidden to the rest of the world because of its lower quality and impact factor or because of the language used. (Wagner 1006)

A growing volume of research publications does not necessarily signal an increase in quality. One key indicator of the value of any research is often seen to be the number of times it is cited by other scientists in their work. Although China has risen in the citation rankings, its performance on this measure lags behind its investment and publication rate. "It will take some time for the absolute output of emerging nations to challenge the rate at which this research is referenced by the international scientific community." (Royal Society 24) If the Royal Society's findings, shown in Figure 2, prove true, then China will soon be producing scientific research papers at a faster rate than the current leader, the United States.

Lederman states that Chinese scholars published 110,000 papers in international journals recorded by *Science Citation Index* but also published 470,000 papers in domestic Chinese journals. (Lederman 126) It is difficult to measure the quality of these Chinese language papers and journals, as language is a problem that calls urgently for the implementation of cross-lingual information retrieval and machine translation advances. China's research output is now far outpacing the rest of the world. In 2006 China's research and development output surpassed that of Japan, the UK and Germany. But non-US innovation is not confined to only Asia and Europe. Brazil's share of research output is growing rapidly also. (Arnold 1) Wagner and

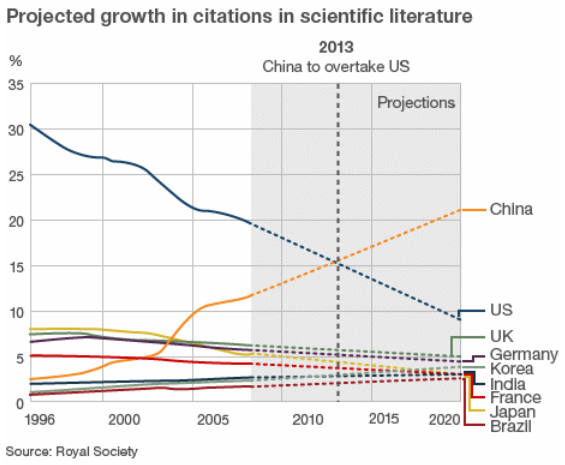

Figure 2

Wong surveyed scientific periodical publications of Brazil, Russia, India and China (BRIC countries) and found that between 2% and 8% were present in Thompson-Reuters *Science Citation Index Expanded* (SCIE), which is more than they expected, and that this bodes well for the future spread of valuable research from these countries. (1007) The translation of the original content into the researcher's literate language still remains as the text these articles are often presented in their original language and script, which keeps them hidden to many who cannot read the presented languages. Automated translation can make research more readily available to researchers who may lack facility in certain languages, but more work remains to make these translation operations as seamless as possible.

Knowledge is also hidden from those who do not have the adequate skills to access the digital information that is often hidden from simple web searches. Academic journals in digital format, which are not readily accessible to all, require knowledge of more sophisticated retrieval methods that can assist librarians, and researchers in finding these valuable texts. Multilingual searching like that which Deep Web Technologies has developed increases the value of research output by making it available to a wider, more global audience. For English speakers, the availability of a multilingual federated search allows diverse perspectives from researchers in foreign countries to be exposed to the world of scholarship. (Arnold 1)

Finding and accessing scholarly journals in other languages is not an easy task as they are not included in many of the aggregators and databases, which are the most popular in the Western world. They are also not well represented in the citation analysis databases such as *Web of Knowledge* and *Scopus*. They may be available and well known in a local area but they remain invisible in an international context. The hidden information is housed in narrow "silos" and these silos are very often language-specific so they are "foreign" to those who do not understand the language.

The silos of information hidden in the deep web also challenge information retrieval. A normal Google search only touches the "surface web" and does not reach the "deep web" where much of the valuable information is stored. Knowledge in the indexed sites that are easily available, which is usually found by using a web crawler or browser like Google, will be harvested from the surface web but this is only a fraction of the total information in existence. In 2001 Bergman's seminal study found that public information on the deep web was 550 times the size of the surface web and 95% of this information was publicly accessible. (Bergman 1)

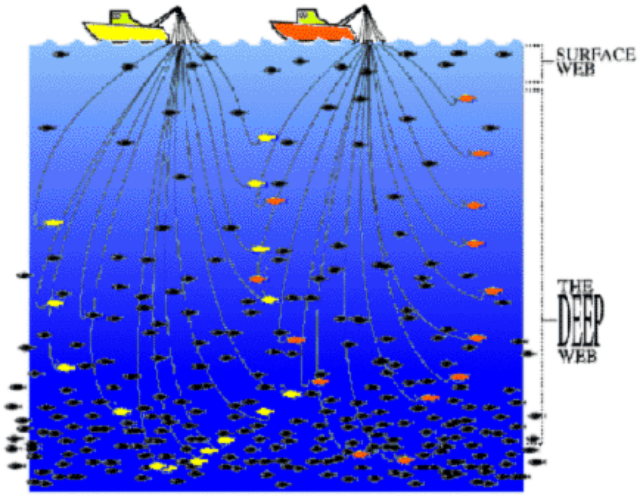

Figure 3
http://beta.brightplanet.com/deepcontent/tutorials/DeepWeb/index.asp

A search engine such as Google or Yahoo can only search the static indexed web pages that make up the smaller surface web, and have been called "a sea of flotsam HTML" by Wright. (4). The method of retrieving this data is not just constrained by language or script but also by the location of valuable information and by the methods used to access this digital information. The deep web is also known as the Academic Invisible Web (AIW) since it contains valuable scholarly information, which is not indexed by search engines such as Google or Yahoo and so is hidden. In order for it to become part of the surface web a Googlebot must be sent out by Google to collect information about documents on the web and to add it to Google's searchable index. If this "spider" does not pick up a resource to index, it remains invisible in any future Google search.

Another option for discoverability is to bring the database to the surface thereby allowing a surface search by a basic search engine to retrieve and index it. The Pacific Rim Library Association (PRL) offers a case study where PRL member digital repositories using the OAI-PMH protocol were made accessible to those

searching the surface web when PRL exposed the metadata to Google for these otherwise hidden repositories. (Palmer 141) This kind of link to the deep web will become more and more common if surface level search engines include OAI-PMH protocol sites in their web crawls.

Methods of retrieving textual data are not just constrained by language or script but also by the location and consequent accessibility of valuable information. Access to online journals generally requires searching with tools that can reach into the Deep Web. Several of these are outlined below.

Tools for searching the Deep Web

DeepDyve
http://www.deepdyve.com
DeepDyve is one of the newer search engines, which is being used by scholars to retrieve information from 7,000 leading scholarly journals using Natural Language Processing (NLP) in the academic sphere. DeepDyve is a research engine for the Deep Web or invisible web that is that part of the Web that is not index-able by search engines like Google. DeepDyve goes beyond keyword indexing used by most current search engines. As the site explains, "DeepDyve's KeyPhrase technology extracts substantially more information from documents than typical keywords. It indexes every word, as well as every phrase in each document, and weighs their informational impact using advanced statistical computation. Steve Wozniak has joined DeepDyve's Advisory Board and says that DeepDyve's KeyPhrase technology has the potential to transform Deep Web searching. (Priemesberger 1)

DeepWebTechnologies (DWT)
http://www.deepwebtech.com/
Deep Web Technologies' multilingual search capability translates a user's search query into the native languages of the collections being searched, aggregates and ranks these results according to relevance, and translates result titles and snippets back to the user's original language. Researchers in the English-speaking world have mostly been restricted to searching only English language sources since the tools for simultaneously searching foreign language sources and for performing the translations haven't existed until recently. Thus, opportunities to search scholarly journals in Chinese, Japanese, Portuguese and other languages associated with countries producing a great volume of science output are being missed. DWT has developed a multilingual search version of its *Explorit* federated search application that integrates the search and translation technologies making for a seamless and productive research environment for scientists, engineers, and researchers in business, science, and technology. (Arnold 1)

WorldWideScience.org
http://worldwidescience.org
A publicly searchable deployment of Deep Web Technologies' multilingual federated search is WorldWideScience.org, which is a global science gateway, comprised of national and international scientific databases and portals. WorldWideScience.org accelerates scientific discovery and progress by providing one-stop searching of databases from around the world. On behalf of the WorldWideScience Alliance, WorldWideScience.org was developed and is maintained by the Office of Scientific and Technical Information (OSTI), within the U.S. Department of Energy. WorldWideScience.org allows you to find science from participating nations of every inhabited continent. Currently, approximately 95 databases and portals from over 70 countries are searchable through WorldWideScience.org. http://worldwidescience.org

DeepWeb Harvester from BrightPlanet
http://www.brightplanet.com
Our Deep Web Harvester™ is the most comprehensive tool available for Deep Web data acquisition. It lets you find Big Data that standard searches simply cannot access. By giving you the power to tap directly into content sources throughout the Deep Web, the Harvester returns relevant Big Data regardless of language, location, or source – and makes it all readily available for analysis. The

Harvester boasts virtually limitless advanced document filtering capabilities, customizable Boolean query development, and the ability to customize queries that are simultaneously submitted to literally thousands of Deep Web sources. This version can be tightly integrated into a custom or enterprise solution through our OpenPlanet® Enterprise Platform, which gives you much deeper analytical capabilities. http://www.brightplanet.com/solutions/deep-web-harvester/
BrightPlanet also powers a search engine directory CompletePlanet that can mine over 70,000 databases in the Deep Web. http://aip.completeplanet.com

Text mining

This paper focuses on information retrieval in a global, multilingual sphere and takes a look at several information retrieval solutions for the end-user. Text mining is often the next step for researchers once they have gathered the specific texts they need to serve as data. "Text mining is a new interdisciplinary field, which combines data mining, information extraction, information retrieval, text categorization, machine learning and computational linguistics to discover structure, patterns and knowledge in large textual corpora."(Lee 1) Once corpora of texts are gathered through information retrieval then text mining of the retrieved texts can take place. The content information that is contained in many academic databases can be further structured into contextual metadata, which allows for exact matches to keyword strings and therefore presents more of a direct search and retrieval process. This sort of mining and clustering of semantic natural language terms can lend itself to a new way to group or cluster scholarly themes within journals. This work is still evolving, but is something that will provide more accurate textual searching for specific and similar subject areas. When using information retrieval (IR) it is assumed that the query language and the retrieved document will be in the same language, but in cross-lingual information retrieval (CLIR) they are generally different and text mining becomes even more complex when it is dealing with multilingual texts. The first step is to find commonalities in texts and therefore enable the discovery of important hidden information. Work is being done to develop the Semantic Web, which integrates structured data from many sources and aids in resource interoperability, which is necessary for cross-lingual information retrieval. (Witt 11) Work is also ongoing by the Global WordNet Grid initiative where cross-language information retrieval will be simplified because of the multilingual lexicons linking various different wordnets in an effort to create a shared multilingual knowledge base. As various languages become more prevalent on the Internet, multilingualism will become a major challenge in web-based knowledge management and the work on the WordNet models should prove to be a valuable tool in this future challenge. (Soria et al. 94,95).

Conclusion

With advances in information retrieval, semantic web development and natural language and cross-lingual processing, the full content of valuable textual resources, many of which are hidden in the deep web, will be more accessible to the world. New methods for increasing access to these publications may come via classification and organization of articles using a specific taxonomy, interlinking common themes using a semantic web, or crowdsourcing in order to develop more annotations or metadata for articles. The goal is to find solutions for researchers and the librarians who work with them, to locate the most appropriate research resources, especially the often hidden nuggets in the global mass of digital information. Libraries need to be aware of what advances are being made to bring all global knowledge to the user, whether it is information that is being freed from the deep web or that which is made available via translation or inclusion via multi-lingual searching solutions. WorldWideScience's federated search is a good model and a step towards the realization of true global knowledge mining for information from all over the globe. Eventually, we can hope that there will be many more tools like this so that we are able to seamlessly mine critical information hidden in the journals in the deep web. Further areas to research will focus on the text mining of these corpora, which have been found to be valuable additions to our global knowledge base.

More specific descriptions of various search engines and search strategies for text mining will be included in the presentation.

References

Arnold, S. (January 30, 2012) Deep Web Technologies: Cracking Multilingual Search. *Beyond Search.* Retrieved from http://arnoldit.com/wordpress/2012/01/30/deep-web-technologies-cracking-multilingual-search/

Bergman, M.K. (Sept. 21, 2001) The Deep Web: Surfacing Hidden Value. *Deep Content.* Retrieved from http://brightplanet.com/wp-content/uploads/2012/03/12550176481-deepwebwhitepaper1.pdf

About DeepDyve (2013) http://www.deepdyve.com

Lederman, A. (2010) Breaking Down Language Barriers through Multilingual Federated Search. *Information Services & Use.* 30 125-132.

Lee, C-H, & Yang, H-C (2000) Towards Multilingual Information Discovery through a SOM Based Text Mining Approach. *PRICAI 2000 Workshop on Text and Web Mining*. p.81.

Oleinik, A. (2012, March 9). Publication Patterns in Russia and the West Compared. *Scientometrics*. pp. 533-551.

Palmer, D. (2009) The Pacific Rim Library: A Surprising Pearl. *Serials Review.* Vol. 35, No. 3. pp.138-141.

Pimienta, D. (2009) Twelve Years of Measuring Linguistic Diversity on the Internet: Balance and Perspective. *UNESCO. Global Symposium on Promoting the Multilingual Internet.* Retrieved from http://unesdoc.unesco.org/images/0018/001870/187016e.pdf

Prado, D. Political and Legal Context. (2005) *Measuring linguistic diversity on the Internet*. *UNESCO Institute for Statistics* Montreal, Canada – UNESCO.

Preimesberger, C. (2009) Wozniak Joins Another Company, This Time Search Engine DeepDyve. *E-week.* Retrieved from http://www.eweek.com/c/a/Search-Engines/Wozniak-Joins-Another-Company-This-Time-a-Search-Engine-849454/

Royal Society. (2011) *Knowledge, Networks and Nations, Global Scientific Cooperation in the 21st Century*. The Royal Society. Retrieved from: http://royalsociety.org/uploadedFiles/Royal_Society_Content/policy/publications/2011/4294976134.pdf

Soria, C., Monachini, M., Bertegna, F., Calzolari, N., Huang, C-R., Hsieh, S-K., …Tesconi, M., (February 11 2009) Exploring Interoperability of Language Resources: the Case of Cross-lingual Semi-automatic Enrichment of Wordnets. *Language Resources & Evaluation*. 43:87-96

Wagner, C. & Wong, S. (2012) Unseen Science? Representation of BRICs in Global Science. *Scientometrics*. 90:1001-1013.

Witt, A. et al. (2009, March) Multilingual Language Resources and Interoperability. *Language Resources and Evaluation*. Vol. 43, No. 1 1-14.

Worldwide Science Alliance. *WorldWideScience.org. About page.* (2011) Retrieved from http://worldwidescience.org/about.html

World Internet Penetration Rates-By Geographic Regions Q2 2012. Retrieved from http://www.internetworldstats.com/stats.htm

Wright, A. (March 9, 2004) In Search of the Deep Web. *Salon.com.* Retrieved from http://www.salon.com/2004/03/09/deep_web.