# Identification of User Facility Related Publications

Robert M. Patton, Christopher G. Stahl, Thomas E. Potok, Jack C. Wells
Oak Ridge National Laboratory
PO Box 2008
Oak Ridge, TN 37831
{pattonrm, stahlcg, potokte, wellsjc}@ornl.gov

## ABSTRACT
Scientific user facilities provide physical resources and technical support that enable scientists to conduct experiments or simulations pertinent to their respective research. One metric for evaluating the scientific value or impact of a facility is the number of publications by users as a direct result of using that facility. Unfortunately, for a variety of reasons, capturing accurate values for this metric proves time consuming and error-prone. This work describes a new approach that leverages automated browser technology combined with text analytics to reduce the time and error involved in identifying publications related to user facilities. With this approach, scientific user facilities gain more accurate measures of their impact as well as insight into policy revisions for user access.

## Categories and Subject Descriptors
H.3.7 [**Digital Libraries**]: *User Issues*; H.3.3 [**Information Search and Retrieval**]: *Search Process*;

## General Terms
Algorithms, Design

## Keywords
Scientific user facility, automated browsing, text analytics

## 1. INTRODUCTION
Scientific user facilities such as Spallation Neutron Source (SNS) and European Synchrotron Radiation Facility (ESRF) provide physical resources and technical support that enable scientists to conduct experiments or simulations pertinent to their respective research. Both facility management and sponsors want to know what impact their respective facility has on scientific discovery. Justification for the existence and funding of these facilities drives the need for appropriate performance metrics of the facility. One performance metric is the number of publications that users produce as a direct result of using the facility.

Evaluating this metric faces numerous challenges. Most facilities have a policy that users must self-report their publications that are directly related to the use of the facility. Consequently, this proves problematic in that publications tend to occur after the facility has been used, in some cases, years afterward. As a result, user facilities may not receive a report from the users, or the reporting is not accurately or adequately performed (e.g., submission of the user's entire curriculum vitae rather than just user facility related papers). In addition, many facilities have a policy that an official acknowledgement statement should be included in the publication in order to formally acknowledge the use of the facility as providing a contribution to the research performed. This also proves problematic in that publications tend to have total page limits. As a result, authors may abbreviate the official acknowledgement statement considerably, or simply reference the facility by just its name or acronym. In some cases, an acknowledgement section does not exist, but the facility is mentioned by name somewhere in the content. Furthermore, users of the facility are not necessarily employees of the facility host. This is problematic in that facilities are extremely limited in their ability to enforce their policies regarding facility related papers. In some cases, users may be employees of private companies who do not publish at all. Publications may also be produced by collaborators of the users, but are not users themselves. Finally, most digital libraries do not provide user facility information as part of the meta-data about the publication. Facility information is generally only found in the content of the publication, which is not as easily accessible from digital libraries like the title, author, and abstract information. Clearly, tracing the impact of a facility becomes increasingly difficult.

Currently, one approach to capturing this metric requires a person to perform appropriate keyword searches on various scientific publication websites, download and read the publication as appropriate, and finally evaluate whether the respective publication is, in fact, related to the user facility of interest. This may easily require more than 60 hours of manual effort, even though many if not all of these steps could be automated. Any change to enhance the scope or accuracy of the result will only increase the level of effort required to accomplish this task. Clearly, scientific user facilities need a new approach.

In order to more quickly and accurately measure the number of publications related to user facilities, this work describes an approach that leverages automated browser technology combined with text analytics.

## 2. RELATED WORKS
Other work has been performed in regard to automatically finding publications and understanding scientific impact. In [3], a system called CiteSeer [4] describes an approach to automatically finding scientific publications on the Internet and creating a digital library. While not focused on identifying publications related to user facilities as described here, its efforts are similar in its ability to work with a variety of website formats and collect the data accurately. The work described here focuses on a smaller number of websites, and focuses strictly on the identification of user facility publications, and not publications in general.

In the work of [1], the research impact is evaluated in light of collaborations between scientists. Their work demonstrated that internal collaborations as well as the number of authors resulted in

publications with a higher impact than external collaborations. User facilities could leverage work similar to [1] in order to evaluate the strength of proposals for access to their facility. In contrast, the approach described here seeks to help user facilities that want to examine the role that their respective facility contributes to the publication impact.

# 3. APPROACH

Our initial approach simply automates the current process performed by a person, but allows for future enhancements to support additional metrics or requirements. Figure 1 shows a process control flow of the approach. There are two primary components: automated browsing and text analytics.

With the growth of websites on the Internet, web developers increasingly rely on automated browser technology to enhance testing of their sites. This technology can be used with a variety of programming languages as well as a variety of web browsers, and is used in this process to replicate the work of a person searching for publications on various sites.

Once a publication is found that may be of interest to the user facility, our process then leverages text analytic techniques to search the content of the publication for appropriate acknowledgements. Results of this analysis are then provided as a spreadsheet detailing what was identified for each publication.

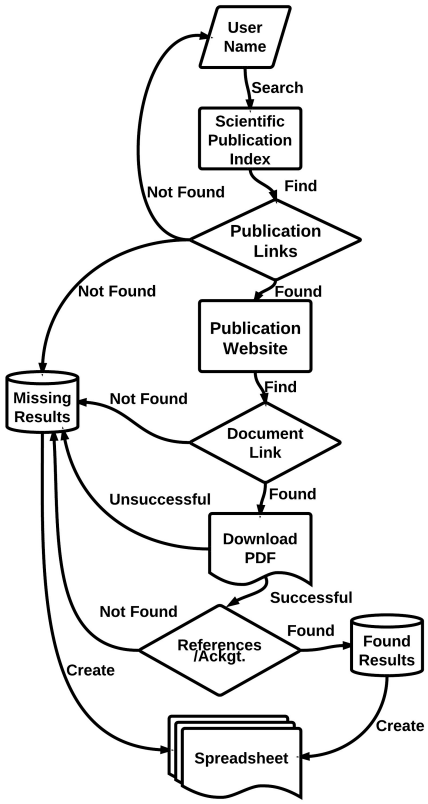


**Figure 1. Automated Process Control Flow**

## 3.1 Automated Browsing

Several automated browsing APIs and tools exist. The primary differences between them include programming language, browser support, and JavaScript support. Depending on the task to be performed, different combinations can also be chosen according to their strengths and weaknesses. The approach described here uses a Perl module called WWW-Mechanize that enables programmatic web browsing and automated interaction with non-JavaScript websites and PhantomJS, a headless Web Kit with JavaScript Application programming interface (API) [7][13]. This combination enables a headless (i.e., window-less) browser to be created providing JavaScript support, Document Object Model (DOM) handling, Cascading Style Sheets (CSS) selector, JavaScript Object Notation (JSON), and a strong API for HyperText Markup Language (HTML) parsing. A headless approach being preferred to full browser emulation because of the speed gains due to the lower amount of data required to be sent across the network (e.g., does not load images on web pages).

Our process begins by using a publication-indexing site. Such sites provide search capability across numerous journals, conferences, and other related websites. These sites generally provide meta-data about the publication such as the title, authors, and abstract, allowing you to search for publications by fields such as user name or affiliate. However, the acknowledgements or actual contents of the publication are not provided. A link to the corresponding journal or conference website is provided by the publication-indexing site where the contents of the publication exists. For the initial search, WWW-Mechanize is used to search the publication-indexing site for names of users associated with the user facility. WWW-Mechanize reads a file containing user names, automatically fills in the correct search information (name, year-range, affiliate, etc.), and finally submits the form and receives the search results. Results are then parsed in order for the link to be found to the actual document. By inspecting the HTML source code of the search results, a parser must be developed to extract the Title, Author Name, and link to the publication. This method has to be tailored for each publication-indexing site due to the custom nature of websites. If the website uses JavaScript, PhantomJS can be used to facilitate the parsing process. Once the publication link is located from the publisher's site, WWW-Mechanize follows the link, searches for a Portable Document File (PDF) file on the publishers website and downloads the corresponding document.

One of the most significant challenges is the use of JavaScript. JavaScript tends to be site specific, resulting in the development of custom code that relies on a specific website and may no longer work if any changes are made to the site. JavaScript also requires different parsing tools to be used than HTML such as PhantomJS or in more advanced cases actual browser emulation. Browser emulation is often used as a result of the complexity introduced by JavaScript such as multiple tabs and/or windows. In a visual environment, this can often be handled in ways that may not be possible in a headless browser (e.g., PhantomJS does not handle JavaScript popup windows).

The inability to find the publications also proves challenging. There are several reasons. Primarily, a license to access the material is unavailable. Secondly, the distinct nature of each journal website plays a role. For most sites, downloading the first PDF file on a page will result in the desired publication although every website may store its publications in a different way. For example, some sites use hybrid PDF+HTML documents that require the +HTML to be trimmed from the link. Multiple PDF files may be stored on a single page and methods must be developed in order to download the desired publication. Multiple links may also have to be followed in order to download the PDF.

A username and password may also be required to access the document, which can be facilitated by WWW-Mechanize but requires custom code to be written.

Another challenge is tracking the correct article name with the corresponding PDF file. The reason for this is that the PDF link usually does not download the file with the article title as the file name, and parsing the content to match the article name is not always successful. Developing a method to track the article title during the entire process and handle errors in order to properly match the title with the file solved this. Also this provides the ability to track different reasons a publication may not have been downloaded such as timeout errors, lack of subscription to the publication or the automated browsers inability to find a document on the page. This additional information allows for more complete results to be compiled.

## 3.2 Text Analytics

After retrieving the publication, analysis of the contents begins. A document is first converted to a text format from its current binary state (PDF, Word). Due to the differences between text and binary files, care must be taken to remove all special characters (non UTF-8 (UCS Transformation Format – 8-bit)), trailing line breaks, and excess whitespaces. This is important so that text analytics can be applied on the document properly. As discussed previously, there are several variations of references to user facilities. First, references may be simply by the name of the facility or the equipment of the facility such as a computer name. This is resolved by providing the system with a list of keywords that are relevant to the user facility. These keywords are then searched in the publication. Exact matches are recorded and provided as output to the user. Keyword matches allow the user facility to see if their facilities are being mentioned in articles with or without an official acknowledgment. This will allow the user facility to find publications that used their facility but failed to make a proper acknowledgement. Second, authors may use the official acknowledgement exactly. Like keywords, this is provided as input to the system, and searched in the publication. Exact matches are output to the user.

Finally, the official acknowledgement of the user facility may be paraphrase or reduced in some way by the authors such that it is not identical to the official acknowledgement. To resolve, the system attempts to extract the Acknowledgements section from the publication. This is accomplished by dividing the document into multiple sections based on paragraph divisions. Each paragraph from the publication and the official acknowledgement required by the user facility are converted to a vector space model (VSM) using the term frequency-inverse corpus frequency (TF-ICF) as the term weighting scheme [8][9]. Over the last three decades, numerous term weighting schemes have been proposed and compared [5][6][10][11]. The primary advantage of using TF-ICF is the ability to process documents in $O(N)$ time rather than $O(N^2)$ like many term weighting schemes, while also maintaining a high level of accuracy. A dot product is then calculated between the two vectors to provide a similarity score, which is then output to the user. Paragraphs that have a similarity of 0 are ignored, as they have no similarity to the official acknowledgement. Paragraphs with similarity between 0.15 – 0.40 are likely a condensed version of the official acknowledgment. Higher similarity values are considered strong matches of the official acknowledgment. Similarity values of 1.0 are an exact match.

# 4. RESULTS

The complete automated process was tested on a small-scale in order to show its time saving potential. A short list of 4 user names from a user facility at Oak Ridge National Laboratory was provided to the system. This resulted in 42 links being found, downloading a total of 26 documents (the remaining documents were not available for a variety of reasons). Each document was then analyzed for key terms, official acknowledgments, and finally all of the results were recorded in a spreadsheet for easy readability. This entire process was completed in just slightly more than 6 minutes. A manual attempt by an experienced user on the same dataset required 15 minutes in order to download the documents. Each document was manually checked for key terms and the presence of an official acknowledgment. The results were then recorded in a spreadsheet. This process required approximately 1 hour to complete. Clearly the time saving efforts of automating this process are needed in order to produce results on any large-scale effort in an efficient amount of time. Automating this process also resulted in more accurate results. For the same dataset, the experienced user was only able to collect 22 out of the 26 documents found with automation. The user also erroneously marked one document as having an official acknowledgment when an official acknowledgment was not present. For the other documents the automated process agreed with the results of the manual user.

Table 1 shows a comparison of the timesaving with the proposed approach. Table 2 shows a comparison of the accuracy with the proposed approach. These results are based on 26 publications. Many user facilities may have several hundred publications to identify.

**Table 1. Comparison of time improvement**

| | *Manual* | *Automated* | *% Reduction* |
|---|---|---|---|
| *Collecting* | 15 min. | 6 min. | 60% |
| *Analysis* | 60 min. | 13 sec. | 99% |

**Table 2. Comparison of accuracy & completeness**

| | *Manual* | *Automated* | *Maximum Possible* | *% Increase* |
|---|---|---|---|---|
| *Total Collected* | 22 | 26 | 26 | 15% |
| *Correct Results* | 21 | 26 | 26 | 19% |

# 5. FUTURE WORK

This work provides the first step toward automating and improving the accuracy in evaluating impact metrics for scientific user facilities. A wide range of next steps is now possible.

One such possibility is the use of Apache Solr [2]. Solr is an open source search platform for text documents that enables faceted search. Facets represent meta-data about the document such as authors, title, etc. Facets could be created in regard to the acknowledgement of the user facility. In addition, Solr supports document clustering that enables the ability to find similar documents to one that is already found, but without entering keywords. From a user facility perspective, this enables additional questions to be asked of the data or evaluate performance metrics that may not be otherwise possible.

In addition, the ability to automatically and accurately collect publications that can be associated to a user facility creates the ability to automatically identify indirect impacts through citation analysis. For example, a publication is produced that directly acknowledges the user facility. Over time, the citations of this paper can be monitored to observe its impact on other publications not related to the user facility. While much work has been performed in citation analysis, the ability to automatically monitor hundreds or thousands of user facility related publications over many years would provide a new capability as well as new metrics that many user facilities currently do not have. A capability such as this would then enable the next one: enhancing policy.

Currently, scientific user facilities have limited ability to enforce policies regarding acknowledgements of the facility in publications. However, user facility managers and sponsors armed with the knowledge provided by previous user facility acknowledgements and citation analysis would be able to add or modify policies regarding access to the facility. For example, a scientist who has previously published and used the official acknowledgement may receive additional or extended access to the facility in the future in comparison to a scientist who did not formally acknowledge the facility in their publications. Furthermore, a scientist who formally acknowledged the facility and whose paper was highly cited would receive even more access to the facility. It is expected that such policy implementations would encourage scientists to formally acknowledge the user facility in their publications, thus improving the original problem.

Other future work may include leveraging the work described in [12] to identify emerging research at a user facility. In [12], the authors use citation and text analysis to identify emerging research in regenerative medicine explicitly for R&D managers and policymakers. From a user facility perspective, this would provide insight into the needs of the respective community that it serves.

## 6. SUMMARY

In this work, an approach was developed and tested to solve the challenge of identifying publications related to scientific user facilities. Managers and sponsors of user facilities need to understand and measure the scientific impact that their respective facility provides. One such metric for that impact is the number of publications produced by users of the facility as a result of having used it. Unfortunately, the current approach to evaluating this metric is manual, and in some cases, error-prone. To address this challenge, an automated approach was developed using automated browser technology and text analytics that replicates the results of the manual process. An automated approach not only saves time and money, but also enables additional analysis and metrics to be evaluated as well as influence policy changes as needed.

## 7. ACKNOWLEDGMENTS

This manuscript has been authored by Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, Tennessee 37831-6285; managed by UT-Battelle, LLC, and used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory under contract DE-AC05-00OR22725 for the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains non-exclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

## 8. REFERENCES

[1] Abbasi, A., Hossain, L., and Owen, C., "Exploring the Relationship between Research Impact and Collaborations for Information Science," System Science (HICSS), 2012 45th Hawaii International Conference on , vol., no., pp.774-780, 4-7 Jan. 2012

[2] Apache Solr, http://lucene.apache.org/solr/, Current April 2012.

[3] Bollacker, K.D., Lawrence, S., and Giles, C.L., "Discovering relevant scientific literature on the Web," Intelligent Systems and their Applications, IEEE , vol.15, no.2, pp.42-47, Mar/Apr 2000

[4] CiteSeerX, http://citeseerx.ist.psu.edu, Current April 2012.

[5] Jones, K.S. and Willett, P.: Readings in Information Retrieval, Chap. 3. Morgan Kaufmann Publishers, San Francisco, CA, pp. 305-312 (1997)

[6] Lan, M., Sung, S-Y., Low, H-B, and Tan, C-L.: "A comparative study on term weighting schemes for text categorization," In Proc. of the 2005 IEEE International Joint Conference on Neural Networks, vol.1, no., pp. 546- 551, (2005)

[7] PhantomJS, http://phantomjs.org/, Current April 2012.

[8] Reed, J.W., Jiao, Y., Potok, T.E., Klump, B.A., Elmore, M.T., and Hurson, A.R.: "TF-ICF: A new term weighting scheme for clustering dynamic data streams," In Proc. of the 5th International Conference on Machine Learning and Applications, pp. 258-263 (2006)

[9] Salton, G., Wong, A., and Yang, C.S.: "A Vector Space Model for Automatic Indexing," Communications of the ACM, 18(11), pp. 613620, (1975)

[10] Salton, G., and McGill, M.J.,: Introduction to Modern Information Retrieval, Mc-Graw Hill Book Co., New York, (1983)

[11] Salton, G. and Buckley, C. Term-weighting approaches in automatic text retrieval. Journal of Information Processing and management, 24(5), pp. 513-523, (1988)

[12] Shibata, N., Kajikawa, Y., Takeda, Y., Sakata, I., and Matsushima, K., "Detecting emerging research fronts in regenerative medicine by citation network analysis of scientific publications," Management of Engineering & Technology, 2009. PICMET 2009. Portland International Conference on, vol., no., pp.2964-2976, 2-6 Aug. 2009

[13] WWW-Mechanize, http://search.cpan.org/~jesse/WWW-Mechanize-1.72, Current April 2012.