Information Services & Use 30 (2010) 125–132
DOI 10.3233/ISU-2010-0617
IOS Press

Breaking down language barriers through multilingual federated search

Abe Lederman [a,*], Walter Warnick [b], Brian Hitson [b] and Lorrie Johnson [b]
[a] *Deep Web Technologies, Santa Fe, NM, USA*
[b] *US Department of Energy, Office of Scientific and Technical Information, Oak Ridge, TN, USA*

Abstract. WorldWideScience.org (WWS) is a global science gateway developed by the US Department of Energy Office of Scientific and Technical Information (OSTI) in partnership with federated search vendor Deep Web Technologies. WWS provides a simultaneous live search of 69 databases from government and government-sanctioned organizations from 66 participating nations. The WWS portal plays a leading role in bringing together the world's scientists to accelerate the discoveries needed to solve the planet's most pressing problems. In this paper we present a brief history of the development of WWS and discuss how a new technology, multilingual federated search, greatly increases WWS' ability to facilitate the advancement of science.

Keywords: Surface web, deep web, federated search, distributed search, multilingual, machine translation, international collaboration

1. Introduction

Discovery drives science. The rapid development of web-based technologies in the last decade has created a unique opportunity to bring together the world's scientists by making it easy for them to share research information. In March 2010, one of the authors of this paper, Abe Lederman, delivered a presentation to the attendees of the 2010 NFAIS Annual Conference explaining the importance of multilingual federated search to advancing science and how the global science gateway, WorldWideScience.org, was enhanced to help realize the vision of bringing together the world's scientists [2]. Much has transpired since the March presentation, most notably the launch of Multilingual WorldWideScience.org$^{\text{BETA}}$ in June 2010. This paper summarizes and updates the important information presented at the conference.

2. Advancing science by accelerating discovery

On February 15, 1676 Isaac Newton, in a letter to fellow British Researcher Robert Hooke, wrote:

> If I have seen further it is only by standing on the shoulders of giants.

While it is common knowledge that science can only be advanced if it is shared, OSTI is on a mission to live the "OSTI corollary":

> Accelerating the sharing of scientific knowledge accelerates the advancement of science.

*Corresponding author: Abe Lederman, Deep Web Technologies, 301 N Guadalupe Street, Suite 201, Santa Fe, NM 87501, USA. Tel.: +1 505 820 0301 ×227; Fax: +1 505 983 7621; E-mail: abe@deepwebtech.com.

The ability for researchers and the public to perform multilingual searches of global science in their native language, irrespective of the language of the source, directly supports the mission of accelerating the advancement of science.

China is now the world's second largest publisher of scientific reports and is on pace to overtake the United States in the next decade. Wu Yishan, Chief Engineer of the Institute of Scientific and Technical Information of China (ISTIC) gives compelling statistics that justify the importance of multilingual search:

> In 2008, while Chinese scholars published 110,000 papers in international journals recorded by Science Citation Index, they also published 470,000 papers in domestic Chinese journals. Without accessing these 470,000 papers, it is impossible to obtain a realistic feeling about the thrust of scientific and technological advancement in China. Therefore, the need for mutual translation between English and Chinese and for cross-language retrieval is increasingly urgent.

Consider the number of speakers of each of the world's major languages listed in Table 1 [7].

Of the nearly four billion speakers of the eleven most spoken languages only 510 million (13%) speak English. Consider also that, according to Cybermetrics Lab, 63% of the world's top 400 institutional repositories have non-English content [9]. These two facts underscore the importance of multilingual search in accelerating the sharing of scientific knowledge as English-only search applications leave very large holes in content coverage. As of September 2010, WWS supports Chinese, English, Spanish, Russian, Portuguese, German, French, Japanese and Korean.

The findings of an OSTI-sponsored study on population modeling of the emergence and development of scientific fields strengthens the case for multilingual search in advancing science [1,11].

> We have been working with a group of modelers led by Luis Bettencourt of Los Alamos National Laboratory. They have written an important new paper, currently in press in *Physica A: Statistical Mechanics and Its Applications*, entitled: "The power of a good idea: quantitative modeling of the spread of ideas from epidemiological models". This paper applies a disease model to the spread of Feynman diagrams just after World War II. Feynman diagrams are a central method of analysis in particle physics.
>
> Looking at these models has led us to focus on a parameter called the contact rate. In the disease model, this is the rate at which people come into contact with a person who has the disease . . . . Our

Table 1
Estimated number of speakers of the world's most popular languages as of June, 2010

Rank	Language	Estimated number of speakers
1	Mandarin Chinese	1,051,000,000
2	English	510,000,000
3	Hindi/Urdu	490,000,000
4	Spanish	420,000,000
5	Arabic	280,000,000
6	Russian	255,000,000
7	Portuguese	230,000,000
8	German	229,000,000
9	Bengali	215,000,000
10	French	130,000,000
11	Japanese	127,000,000

focus, therefore, is on increasing the contact rate. To do this we must reduce a huge gap in how the Internet works today [6].

Science advances through increasing the contact rate among global multilingual researchers.

3. A brief history of WorldWideScience.org

The following is a brief summary of the historical events related to the development of WWS and a brief outline of the milestones associated with WWS. Further detail is available in a number of the references listed at the end of this paper.

WWS was conceived in 2005 by one of this paper's co-authors, OSTI Director Dr. Walter Warnick. Dr. Warnick's vision was to utilize the model behind the successful Science.gov federated search portal and to extend it beyond the US federal government databases of Science.gov to encompass science information produced throughout the globe. (See Section 4 for a discussion of federated search and its importance to the development and growth of WWS.) The vision of a global gateway to science was proposed at the annual conference of the International Council for Scientific and Technical Information (ICSTI) in June 2006. The British Library expressed interest in collaborating with the US Department of Energy to develop the application and the two countries formed a partnership in January 2007. Officials from the United States and Great Britain signed a bilateral statement of intent and invited other nations to join what was to become the WorldWideScience Alliance.

OSTI debuted WorldWideScience.org in June 2007 at the ICSTI conference in Nancy, France. By June 2008, when the WorldWideScience Alliance was formed, the number of countries participating in WWS had risen dramatically from ten countries to 44. In October 2008, the People's Republic of China (one of the world's largest producers of scientific research) joined the WorldWideScience Alliance.

WWS has quickly achieved a number of important milestones [4]:

- When first introduced, WorldWideScience.org included only 10 countries and 12 databases and portals, and it represented roughly 12% of the world's population.
- The 38 nations that were represented in the Alliance's founding document, plus 6 others, contributed 32 databases and portals, and represented roughly 53% of the world's population.
- Today, 66 countries contribute 69 databases and portals and represent more than 75% of the world's population, enabling scientists to search these sources with a single query. Results are then collectively ranked in relevance order. Features such as alert services enable scientists to stay abreast of ongoing research in their fields, regardless of international boundaries.
- When WorldWideScience.org launched, it provided searchable access to roughly 200 million pages of science content; today it searches across about 400 million pages of important scientific portals worldwide. That's a lot of science information accessible from one search box – equivalent to a shelf of documents 20 miles long.

4. The role of federated search in surfacing good science

Federated search is the technology that allows for the simultaneous search of multiple content repositories from a single query form; these repositories frequently contain vetted scholarly articles and the output of science research. While, at the surface, federated search applications may appear to operate in

the same way that Google and the other popular "web spiders" operate, the way that federated search applications locate content is quite different.

The Web can roughly be divided into regions, the Surface Web and the Deep Web. The Surface Web comprises perhaps several billion web pages – counting its pages has proven to be notoriously difficult. Many of these Surface Web pages link to one another. Google and other web spiders maintain a large list of web pages and periodically follow the links on these known pages to discover new pages. The spiders index the content of these pages to permit rapid searches of the web pages and documents they locate.

Federated search applications search the Deep Web. The Deep Web consists of documents, rather than web pages that link to one another. Counting the number of documents in the Deep Web is more difficult than sizing the Surface Web, but most estimates place the size of the Deep Web at as much as 500 times the size of the Surface Web. Deep Web documents typically reside in databases; these documents are accessed by humans, via filling out search forms and by federated search applications that simulate the same search form completion process. Federated search applications do not build indexes; they perform live searches of multiple sources and aggregate the results.

Federated search plays a key role in science discovery because the landscape of federated search (the Deep Web) consists of many more high quality scientific documents than does the Surface Web. In other words, Google and the other crawlers find a much smaller proportion of the documents that scientists care about than do federated search applications. The Deep Web is a much better landscape for science researchers because publishers of scientific papers are much more likely to place their documents in databases than in linked Surface Web pages.

Federated search is also key to science discovery because it bypasses the high noise to signal ratio inherent in Surface Web searching. Google searches a large variety of content, much of which is not of interest to serious researchers. Even with Google's relevance ranking mechanism the researcher is left to separate the wheat from the chaff. Federated search applications usually search content repositories that contain only vetted content, greatly reducing the number of non-relevant records to sift through.

OSTI recognizes the importance of federated search in facilitating the rapid growth of WorldWideScience.org which directly leads to the acceleration of discovery.

> Federated search technology is of particular strategic value to OSTI in that it does not place any requirements or burdens on owners of databases. This means that when an agreement is made with a scientific organization to make its content searchable by one or more OSTI applications, setting up access to the organization's content is a rapid and straightforward process. If the organization's content is already searchable, via the web or some other mechanism, then the organization has no responsibility other than to keep its database accessible, a responsibility it already has [4].

OSTI also notes that "in comparison of search results from identical queries on WWS, Google and Google Scholar, only 3.5% overlap (i.e., WorldWideScience is 96.5% unique)" [5]. Thus, searching only Google or Google Scholar, will miss almost all of the content that WWS will locate.

5. Implementation of multilingual federated search

Multilingual federated search (MLFS), developed in partnership with Deep Web Technologies (DWT), extends DWT's Explorit federated search solution. MLFS breaks down language barriers of traditional federated search by enabling search and results retrieval of foreign-language sources in the researcher's native language.

The DOE Small Business Innovation Research (SBIR) program provided the funding necessary to perform the R&D work. DWT partnered with Microsoft Research to provide the translation capabilities of multilingual federated search. Microsoft Research's Natural Language Processing group has focused on developing and improving machine translation technology since 1999 [10].

Microsoft's translation approach is to perform statistical machine translation [8]. In this model, which is the most prevalent model implemented today, the translation software learns to translate text automatically, without rules, by analyzing large corpora of text in a number of pairs of languages. The Microsoft translation service operates on the combination of text and a pair of languages (source plus target). A MLFS search including sources in multiple languages will generate several translation requests to the Microsoft translation service.

While, conceptually, MLFS appears simple – translate search terms and search results between the user's native language and the source's language – the mechanics are complex. The MLFS search and translation process includes eight steps [3]:

1. User enters query in their native language.
2. Explorit uses translator service to translate the query into the right language for each source.
3. Explorit submits query to each source.
4. Each source returns results in the source's native language.
5. Result metadata from different sources is aggregated.
6. Result metadata is ranked in the source's language.
7. Result metadata is displayed to the user in potentially multiple languages.
8. Results page is translated to user's language when the user presses the "translate" button.

Figure 1 illustrates the MLFS process [3]. Figure 2 shows some of the contents of a translated search results page.

Translation occurs at three points in the MLFS process. First, the user's query is translated into the languages of the sources. Second, the search result titles and abstracts (or summaries) that are in a language foreign to the researcher are translated to the researcher's native language on request. Third, full text documents, when available from the content source, are translated when the user clicks a button to view a document's full text. All three types of translation requests are sent to the Microsoft translation service, which displays the document in its original and translated form side by side in two panes of a browser window.

Relevance ranking and clustering (faceting) of results works best when many search results are translated to a single language to be compared and organized. A single user query can easily provide the multilingual federated search application with thousands of results to translate. Given that a user may only browse one or two dozen results and given the resources required to perform large numbers of translations it was decided that, for the first release of multilingual WWS, translations would only be performed upon user request (i.e., by the user clicking a "translate" button.) Future work will seek to determine how many results can realistically be translated to provide ranking and clustering features.

6. Launch of multilingual WorldWideScience.org in Helsinki

Multilingual WorldWideScience.org^{BETA} was officially launched on June 11, 2010 in Helsinki, Finland at the International Council for Scientific and Technical Information's (ICSTI) Annual Conference entitled *From Information to Innovation*. Dr. Warnick presented the keynote session "Multilingual

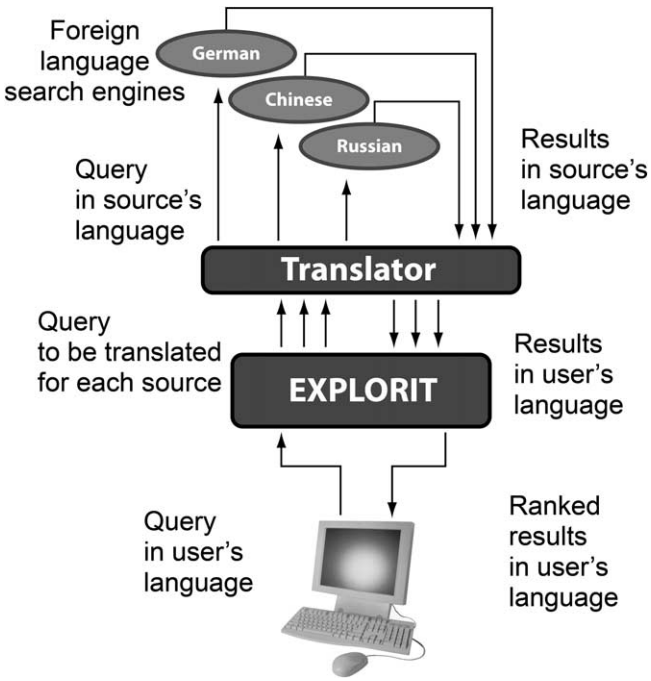

Fig. 1. Multilingual federated search steps.

WorldWideScience.org: accelerating discovery through multilingual translations" [5], during which he emphasized the significant role that multilingual search can play in providing both non-English speakers translated access to research in languages other than their own and English speakers with access to the ever-increasing body of non-English scientific content. A brief demonstration of Multilingual WorldWideScience.org^{BETA} followed. Several other speakers also took part in this special session including Mr. Richard Boulderstone, British Library and Chair of the WorldWideScience Alliance; Dr. Tony Hey, Vice-President of External Research for Microsoft; Dr. Wu Yishan, Chief Engineer of the Institute of Scientific and Technical Information of China (ISTIC); and Dr. Yuri Arskiy, Director of the All-Russian Institute of Scientific and Technical Information. Each of these gentlemen remarked upon the importance of multilingual translation capabilities within the global science community.

7. What's next?

Multilingual federated search is in its infancy. While WWS has been well received, there is still much to do. In this section we consider next steps.

The two highest priorities are to increase the number of languages and content sources supported. WorldWideScience.org Alliance members have indicated that Arabic is an important language to add. The needs of the Alliance members and the languages of new databases will dictate the priorities for further language support. We are also developing multilingual Topic Pages (pages of pre-computed search results) in a number of foreign languages as a next step towards promoting and increasing use among non-English speakers.

We addressed the limitations of translation capacity (bandwidth) in Section 5. Looking ahead we will experiment with the middle ground between translating all search results (most desirable) and translating

Public library in century digital Pulman project information: recommendations of the European Commission
Original Title: Публичные библиотеки в век цифровой информации : Рекомендации проекта Pulman Европейской Комиссии
★★★★☆
2004-01-01
M.: CTI Grand: Fair-press, 2004 Grand features: 410 pp. Series: Special publishing project library ISBN, price 5-8183-0645-3: B.c. SSTI:; 13.20.31
Original Summary: М.: ИТК Гранд: Фаир-Пресс, 2004 Колич.характеристики :410 с. Серия: Специальный издательский проект для библиотек ISBN, Цена 5-8183-0645-3: Б.ц. ГРНТИ : ; 13.20.31
The Russian Union Catalog of Scientific Literature (Russian)

Electronic libraries, digital and virtual: three entities defined
Original Title: Bibliotecas electrónicas, digitales y virtuales: tres entidades por definir
★★★★★
2002-12-01
Technological development has brought a revolution in the work of libraries, develop the electronic libraries, digital and virtual. Currently there are dissimilar considerations in this regard. The work shows some of these in the study of works of different authors
Original Summary:El desarrollo tecnológico ha traído consigo una revolución en el trabajo de las bibliotecas, desarrollándose las bibliotecas electrónicas, digitales y virtuales. En la actualidad hay disímiles consideraciones al respecto. En el trabajo se muestran algunas de estas por medio del estudio de trabajos de diferentes autores
Scientific Electronic Library Online (Spanish)

The seventh International Conference on Asian digital libraries held in Shanghai
Original Title: 第七届亚洲数字图书馆国际会议在沪召开
★★☆☆☆ 庄琦
河北科技图苑
From Shanghai Jiaotong University and Shanghai Library jointly organized the 7th International Conference on Asian digital libraries (The 7th International Conference of Asian Digital Libraries, hereinafter referred to as ICADL2004), on 15 December 2004 at the Grand opening of Shanghai Everbright Convention and Exhibition Center. From 25 countries and regions of more than 350 delegates ask for a period of Exchange and discussion. The Conference theme was "digital: international cooperation and mutual development." Thesis: (1) technologies and standards; (2) services and management; (3) cooperation and localization.
Original Summary:由 上海交通大学和上海图书馆联合举办的第七届亚洲数字图书馆国际会议(The 7th International Conference of Asian Digital Libraries, 简称ICADL2004), 于2004年12月15日在上海光大会展中心隆重开幕。来自25个国家和地区的350多位与会代表在会议 期间进行了交流和探讨。本次会议的主题是“数字图书馆：国际合作与相互发展”。论文主题内容为：(1)技术与标准；(2)服务和管理；(3)合作和本地化。
Institute of Scientific and Technical Information of China (Chinese)

Fig. 2. Screenshot of translated MLFS search results page.

the fewest required to present a single results page. In order to provide better relevance ranking and a clustering capability we will need to translate a larger subset of search results than we currently do. We may, for example, choose to translate all results whose rank exceeds a particular threshold.

To improve relevance ranking we will implement ranking and stemming of foreign-language terms. Research and development is required to make foreign-language ranking as effective as English-

language ranking. Currently, only English-language terms are stemmed. This leads to less effective search of foreign-language terms. Development of multi-lingual stemming algorithms will improve the precision of mixed-language and foreign-language searches.

To improve usability we will be incorporating a foreign-language spell checking capability.

To increase the sharing of ideas we will add an alert capability, where a user can save terms for one or more searches; the system will perform those searches on the user's behalf on a regular basis and email new summaries of results translated into the user's native language.

Multilingual WorldWideScience.org^{BETA} is live, but there is still much to do. Work is ongoing to smooth out the rough edges and to ensure scalability of WWS for researchers for years to come.

References

[1] L. Bettencourt et al., Report for the office of scientific and technical information: population modeling of the emergence and development of scientific fields, available at: https://www.osti.gov/innovation/research/diffusion/epicasediscussion_lb2.pdf.
[2] A. Lederman, Federated search: breaking down the language barrier, in: *2010 NFAIS Annual Conference*, Philadelphia, PA, March 2010, available at: http://deepwebtech.com/talks/NFAIS.ppt.
[3] A. Lederman, Federated search: breaking down the language barrier, in: *NFAIS Workshop*, Philadelphia, PA, May 2010, available at: http://deepwebtech.com/talks/NFAIS Workshop.ppt.
[4] W. Warnick, Federated search as a transformational technology enabling knowledge discovery: the role of WorldWideScience.org, http://www.osti.gov/ILDS_38_2Warnick2010.pdf.
[5] W. Warnick, Multilingual WorldWideScience: accelerating discovery through multilingual translations, in: *International Council for Scientific and Technical Information (ICSTI) Annual Conference*, June 2010, Helsinki, Finland, available at: http://worldwidescience.org/speeches/June2010/warnick_multi.html.
[6] W. Warnick, Science depends on the diffusion of knowledge, available at: http://www.osti.gov/ostiblog/home/entry/science_depends_on_the_diffusion.
[7] http://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers.
[8] http://en.wikipedia.org/wiki/Statistical_machine_translation.
[9] http://repositories.webometrics.info/top800_rep_inst.asp.
[10] http://research.microsoft.com/en-us/projects/mt.
[11] http://www.sciencedirect.com/science/article/B6TVG-4HC15T8-1/2/9017ca224bb58c4dc129c2b92647c826.