Advancing science by overcoming language barriers

Abe Lederman & Darcy Katzman

Scientific discoveries are often restrained by language. The English language may be thought of as the "universal language" in the global scientific community, but only a fraction of scientists actually speak the language, continuing to publish information in their native languages and in non-English databases. Between 2001 and 2011, the number of publications from China in peer-reviewed, non-English journals increased from 3% to 11% of the total world output. In this decade, China will surpass the United States as the most prolific publisher of scientific journal articles. These non-English publications are submerged in what is known as the Deep Web. They aren't likely to be found in places that search engines such as Google or Bing crawl, meaning they will not be easily accessible to English-speaking researchers.

The Deep Web (Figure 1) has garnered a lot of interest lately, partly because the dark web and Deep Web have been referred to interchangeably by many people. In October 2013, the FBI shut down the Silk Road website, a dark web eBay-style marketplace for selling illegal drugs, stolen credit cards and other nefarious items. But the Deep Web is very different from the dark web black market where shifty and illegal activity takes place.

Think of the Deep Web as more academic – used by knowledge workers, librarians and corporate researchers to access the latest scientific and technical reports, gather competitive intelligence or gain insights from the latest government data published. Most of this information is hidden simply because it has not been "surfaced"

Figure 1: Areas of the web.

to the general public through Google or Bing spiders, or is not available globally because of language barriers. If a publication reaches Google Scholar, chances are, it now floats in the broad net of the shallow web, no longer submerged in the Deep Web. A large number of global science publications are located in the Deep Web, only accessible through passwords, subscriptions and only accessible to native language speakers. These publications, hiding in the Deep Web, limit the spread of science and discovery.

Researchers looking for information that is not in their native language(s) find that time works against them. The carrot of scientific collaboration is just before them, yet out of reach

Abe Lederman is the founder and CEO of Deep Web Technologies and has worked in information retrieval for the last 30 years.

Darcy Katzman served as the WorldWideScience.org and Microsoft Translator liaison at Deep Web Technologies from 2007 to 2014.

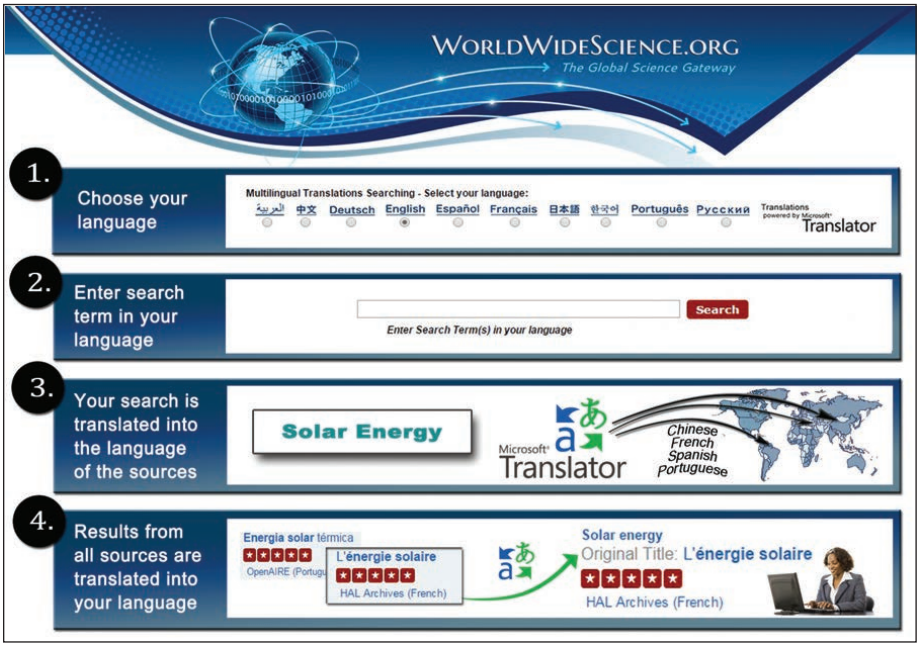

Figure 2: WorldWideScience.org translation flow.

unless they undertake hours and hours of back and forth between a translator and whatever database they hope will have relevant scientific information. In reality, many of these scientists don't even know where to look, nor do they even want to spend the time searching in these repositories. The effort is monumentally time-consuming. In the science and technical industries where timely distribution and access to new knowledge is critical, this process is laborious, painstaking and often unsuccessful, given the sheer effort needed to find information.

Creating the global science gateway

On January 21, 2007, at a ceremony at the British Library, Ray Orbach, then under secretary for science at the US Department of Energy, and Lynne Brindley, then chief executive of the British Library, signed a statement of intent to collaborate in creating a global science gateway that would make science output more easily shareable across the globe.

An initial version of the global science gateway, which became known as WorldWideScience.org, was developed and launched in six months using a federated search product from Deep Web Technologies (DWT) that can quickly and easily search in real-time dispersed science repositories. This initial version of WorldWideScience.org searched 12 databases from ten countries.

Federated search, unlike popular search engines such as Google and Bing, doesn't send spiders out to build an index of information. Instead, when a user clicks the search button, a federated search engine will send that search query out to all of the sources it searches, simultaneously. It's almost as though the user is going to each source of information and searching it directly, rather than searching from a single search box. Once the federated search engine has received all of the results from all of the sources, it blends them together, ranks them, removes any duplicate results and displays them on a single page for the user to see.

On June 12, 2008, at a gathering in Seoul, South Korea, the WorldWideScience Alliance was formed "to formalize their commitment to sustain and build upon the online gateway to the world's science information." The founding document for the WorldWideScience Alliance, signed by 14 representatives of scientific and technical organizations around the world, noted:

"By transitioning WorldWideScience.org from bilateral to multilateral governance, we commit ourselves to a long-term vision for enabling and accelerating scientific discovery through unique and innovative use of federated searching and other technologies. Through our joint efforts, we will sustain and build upon the purpose of providing a single, sophisticated point of access to diverse scientific resources and knowledge from nations and international bodies around the world."

With WorldWideScience.org now firmly established as the go-to gateway for access to a growing collection of the world's most valuable science collections (38 databases from 32 countries) the WorldWideScience Alliance asked DWT to undertake a research effort with the goals of developing a capability for searching non-English science repositories in WorldWideScience.org and making the content in WorldWideScience.org more easily available to non-English speakers in their language.

Choosing a machine translation engine

In this next phase DWT envisioned enhancing WorldWideScience.org with a multilingual search and translation capability for a more comprehensive Deep Web search. Using live translators was simply not an option for the search engine. By its very nature, the search engine required "on-the-fly" translations to work in an on-demand environment, translating a user's query as well as the results brought back. While machine translation is not perfect, it would allow researchers to discover important articles in their field of study. And if an article seemed valuable enough, the researcher could request that the article be translated by a human translator.

Knowing this, DWT investigated a number of machine translation engines, keeping the following requirements in mind:

- *Multiple language support:* Although WorldWideScience.org only needed ten languages for the multilingual launch, the Alliance's plans to grow their membership and their search of diverse science databases required the support of as many language pairs as possible from the translation engine. Only engines supporting a broad number of languages were evaluated.
- *Translation quality:* Acknowledging that the quality of machine translations varies, the Alliance needed to find a machine translator that would perform well across a wide range of languages.

Figure 3: Search results can be easily translated into a variety of languages.

■ *Programmatic interface:* DWT required that the machine translation be accessible programmatically via an application programming interface (API) that would facilitate integrating the machine translation service into its federated search product.

■ *Affordable:* Machine translation software and services can be expensive. The WorldWideScience Alliance needed an affordable solution.

Microsoft Translator, developed and enhanced by Microsoft Research's Natural Language Processing group going back to 1999, fulfilled all of the above requirements. Microsoft Translator uses a statistical translation model to perform translations. This approach is the most common approach today, and requires a large corpus of text and corresponding translated text to build mostly automatically a translator for a new language pair. Microsoft Translator is highly scalable, customizable and supports 50 languages, while continually adding languages to its list. Most importantly, the WorldWideScience Alliance and Deep Web Technologies were able to establish a partnership with the Microsoft Translator group, which supported their development efforts and helped to promote WorldWideScience.org.

Translating the world's science

Multilingual WorldWide-Science.org Beta officially launched in June 2010 at the International Council for Scientific and Technical Information (ICSTI) annual conference held in Helsinki, Finland. It was the first search engine to search databases in diverse languages and retrieve and translate results. The engine was hailed by Richard Boulderstone from the British Library as "the world's most important scientific resource, where the global science community can share knowledge." WorldWideScience.org Beta searched in nine languages: English, Chinese, French, German, Japanese, Korean, Portuguese, Spanish and Russian, adding Arabic two years later.

Finding information on WorldWideScience.org is simple: a researcher first chooses the language that he or she would like to search in and submits a search query (Figure 2). The query that has been entered in the user's language is then translated into the languages of the sources that will be searched. When results come back, they are translated from the language of the source to the

user's language using Microsoft Translator. The researcher then sees a single page of results (Figure 3) in his or her language, with the most relevant results at the top of the page. The Microsoft Translator service allows users to view on a split-screen an original article and a translated version of the same article side by side.

While multilingual federated search has been well received by WorldWideScience.org users and is in production at a number of other sites, improving federated search is an ongoing process.

In the next couple of years, DWT will add support for additional languages as new members join the WorldWideScience Alliance and will require databases in languages not currently supported by WorldWideScience.org to be searched as well as enable users in these new member countries to search in their languages. We also envision leveraging the Microsoft Translator Hub to create custom translators for languages not available via Microsoft Translator.

Machine translation quality varies among different vendors and may vary depending on the language pair and direction – for example English to Russian vs Russian to English. DWT is working on making its translation modules plug-and-play so that it is easy to use the best available machine translation software based on the languages needed to be translated.

A key capability of our federated search is relevance ranking, determining how closely a search result matches the user's query. Determining relevance is challenging in non-romance languages, particularly East Asian languages such as Chinese, Japanese and Korean. Integrating technologies from companies that are good at parsing and analyzing text in foreign languages will help to improve the quality of the relevance ranking.

Finally, WorldWideScience.org will be enhanced so that it automatically detects the country that the user is coming from, presents an interface in the user's local language and assumes that users are entering their query in their local language. It is also possible to switch the user interface of the federated search application among the languages now supported.

The good news is that the multilingual federated search capability initially developed for WorldWideScience.org is not limited to that site and not even limited to science. Since the initial launch of multilingual WorldWideScence.org in 2010, DWT has developed a number of other multilingual federated search applications.

In 2015, the United Nations Economic Commission in Africa (UNECA) will launch their multilingual federated search portal to enhance cross-pollination of innovative news and ideas for and throughout Africa. The UNECA portal will search in four languages: English, Spanish, Portuguese and French.

Outside of science we see an opportunity to provide global companies with a multilingual competitive intelligence capability so, for example, if a large contractor is bidding on a project in China it will have access to local intelligence in their own language. DWT sees the opportunity to create a website that searches the world's most important news and media sites in their native languages, eliminating some of the sugar coating and bias that occurs when global news is summarized by English-focused news outlets.

In June 2015, WorldWideScience.org will be celebrating its eighth birthday. Over the past eight years, the site has received many accolades. The Science Gateway was promoted to the highest levels of the US State Department as an example of US scientific cooperation with counterparts in China and Russia. WorldWideScience.org is seen as helping to reduce the "digital divide" by providing access to publicly available scientific research output to many in third world countries who do not have access to commercially available scientific literature.

Today, searching 500 million pages of quality science and technology information from around the world, across 100 repositories with 22 of these repositories being non-English, WorldWideScience.org has proven that scientific discoveries can surpass language barriers. **M**

MultiLingual — Your information source for language and business.

Subscribe now and see these benefits:

- Eight issues a year plus an annual editorial index/resource directory
- An online searchable archive of issues beginning January 2006
- Insightful articles on the topics above as well as managing content, standards, language preservation and much more.

Two ways to subscribe:

1. Digital only. Receive an email when each issue is available. Have access to all the digital issues since 2006. Use the indices or search to find topics, companies and people across issues.

2. Print + Digital. Receive a printed copy of the magazine at your address. Keep the archives on your bookshelf plus take advantage of all the digital features outlined in #1.

MultiLingual

Language | Technology | Business

SUBSCRIBE NOW!
www.multilingual.com/subscribe