THE WORLDWIDESCIENCE ALLIANCE: AN INTERNATIONAL PARTNERSHIP TO IMPROVE ACCESS TO SCIENTIFIC AND TECHNICAL INFORMATION

Lorrie A. Johnson

U.S. Department of Energy, Office of Scientific and Technical Information

United States

JohnsonL@osti.gov

Abstract

The WorldWideScience Alliance is a strategic partnership, comprised of national and international science agencies, libraries, and information centers, whose goal is to eliminate barriers to finding and sharing scientific and technical information across national boundaries. The Alliance provides the governance structure and sets the direction for WorldWideScience.org (WWS.org), a federated search portal offering users the ability to simultaneously search, in real time, over 100 scientific and technical databases from more than 70 countries. Search results from the various databases, incorporating information in textual, multimedia, and scientific data formats, are then relevance-ranked, and a consolidated results list is presented to the user. Multilingual translations capabilities are available for ten languages, which makes scholarly material more accessible to both developed and developing countries. Through the Alliance partnership, participating members increase access to scientific information by allowing databases in their respective countries to be searched via WWS.org. As the open access movement continues to expand, the Alliance views these resources as a promising area for future growth, as open science and sharing of information are crucial to the advancement of scientific knowledge. WWS.org is an important resource for university libraries because of its unique content and search functionality, not offered by commercial products and search engines.

Keywords: WorldWideScience.org, WorldWideScience Alliance, Federated Searching, Multilingual Translations, Machine Translation, International Collaboration, Multimedia Scientific Content, Scientific Research Data, Public Access

Introduction

As methods and practices for sharing and communicating scientific information have evolved, the rapid development of web-based technologies in the last decade has likewise created a unique opportunity to bring the world's scientists together, by making it easier to find, access, and share research information. WorldWideScience.org (WWS.org) is a global science gateway governed by the WorldWideScience Alliance, with the U.S. Department of Energy's Office of Scientific and Technical Information (OSTI) acting as the Operating Agent. WWS.org provides a simultaneous live search of more than 100 scientific and technical databases from government and government-sanctioned organizations representing over 70 countries. This paper describes the history of WWS.org's development; the formation of the WorldWideScience Alliance; WWS.org's innovative technological features, such as multilingual translations and multimedia search functionality; and projected future growth in the research data and public access arenas.

WorldWideScience.org Launch and the Formation of the WorldWideScience Alliance

WWS.org was conceived and developed using the model of the U.S.-based Science.gov and its underlying federated search technology. Science.gov, launched in December 2002, provides a single search point to over 60 databases and 2,200 selected websites from 15 U.S. federal science agencies. National governments frequently face the challenge of improving access and visibility of government-sponsored research results. Citizens are often unaware of which agency conducted specific research, and in many cases, they simply want authoritative information from any applicable agency. Within the U.S., Science.gov helped solved these problems by providing a single point to simultaneously search all of the U.S. government's scientific resources and retrieve relevance-ranked results. Since 2002, Science.gov has been praised by numerous public and private sector information advocates for its strides in increasing transparency to government-sponsored scientific information.

Considering the early success of Science.gov, it was a logical extension of the concept and technology to move toward international collaboration. At the 2006 annual conference of the International Council for Scientific and Technical Information (ICSTI), the concept for a global expansion of Science.gov was introduced. The British Library agreed to partner with the U.S. Department of Energy in the development of this global science gateway, and on January 21, 2007, a bilateral statement of intent was signed in London between then Chief Executive of the British Library, Dame Lynne Brindley, and then Under Secretary of Energy Dr. Raymond L. Orbach. Other nations were invited to join this partnership.

A prototype of WWS.org was debuted at the June 2007 ICSTI conference in Nancy, France. The inaugural portal performed federated searching of 12 databases and portals across 10 countries. In addition to the United Kingdom and the United States, the WWS.org prototype provided access to scientific information from Australia, Brazil, Canada, Denmark, France, Germany, Japan, and the Netherlands.

Following the prototype's launch, discussions ensued among participating countries, as well as ICSTI, to transition the bilateral U.S.-U.K. governance of WWS.org to a multilateral structure. Terms of Reference for the formation of the WorldWideScience Alliance were ratified during the February 2008 ICSTI meeting. Along with defining the purpose, objectives, terms, conditions, and structure of the Alliance, the Terms also stipulated that OSTI would act as the Operating Agent for WWS.org and Secretariat to the Alliance. The Alliance is also closely affiliated with ICSTI, which serves as a member and primary sponsor. The WorldWideScience Alliance was officially established, and the WWS.org website was launched, on June 12, 2008, at the annual ICSTI conference in Seoul. A signing ceremony marked this milestone, in which organizations representing 38 countries agreed to take part in the governance and funding of WWS.org.

The Advantages of WorldWideScience.org's Federated Search Technology

WWS.org employs federated search technology, which allows for the simultaneous search of multiple content repositories from a single query form. Unlike some popular search engines, WWS.org doesn't send spiders out to build an index of information. Rather, when a user initiates a search, the query is sent, simultaneously and in real-time, to all of the sources WWS.org

searches. Once the searches have been completed at all sources, the federated search engine consolidates all of the results and relevance ranks them into a single results list.

Beyond the obvious time-savings for a scientific researcher – he or she can search multiple content sources via a single search – federated search technology also offers the ability to search the "Deep Web." As opposed to web pages that link to one another, the Deep Web consists of documents, typically residing in databases. Federated search applications, such as WWS.org, search this Deep Web, where most scientific content resides. Additionally, federated search technology does not place any requirements or burdens upon the database owners, beyond the ability to handle increased numbers of queries.

The WorldWideScience Alliance established criteria for the databases it searches. The over-riding principle was that the databases must be produced, sponsored, or endorsed by a national scientific body or government organization. As a result, growth in the content and geographic representation in WWS.org has been steady and impressive, going from 10 countries and 12 databases in 2007 to more than 70 countries and over 100 databases today.

Implementation of Multilingual Translations

Originally, WWS.org was limited to search and retrieval of English-language scientific databases. While English is traditionally the *lingua franca* for science, this limitation had the effect of (a) under-serving the non-English-speaking populations, and (b) excluding access to the expanding volume of non-English papers in countries such as China, Germany, Japan, Korea, and many others. Indeed, the list of the world's most widely spoken languages includes Mandarin Chinese, Spanish, English, Hindi, Arabic, Portuguese, Bengali, Russian, Japanese, and Punjabi.

While a large share of scientific literature is published in English, vast quantities of high-quality scientific research are not. An increasing volume of non-English content, both conventional and non-conventional literature, is being produced for national journals, institutional repositories, and regional databases. Japan, France, Germany, Brazil, China, Russia, and other countries all produce scientific literature in their respective national languages. The continuing growth of such non-English content creates a digital divide in that it is not particularly accessible to English-speaking populations and, indeed, any population other than those speaking the language of a particular paper.

These challenges clamored for a multilingual translations solution, and the WorldWideScience Alliance recognized the importance of expanding the network of databases accessible to the worldwide community in an era of increasing globalization of science. To break through these language barriers, the Alliance set out to offer a new, innovative assistive technology.

An automated/machine translations solution for WWS.org needed to be able to translate a user's query into the various languages of constituent databases on the front end, and then translate the search results into the user's language on the back end. Working with the translations team from Microsoft Research, the WWS.org search engine provider, Deep Web Technologies, successfully integrated such translations capability into the front and back end of the user experience.

The initial version of WWS.org with multilingual translations was released in June 2010, at the ICSTI conference in Helsinki. Improvements and enhancements since then have expanded the languages offered to ten: Arabic, Chinese, German, English, Spanish, French, Japanese, Korean, Portuguese, and Russian. Unlike some machine translations that translate from a single language into another single language, the WWS.org application performs a "many-to-many" translation, i.e. a user's search in one language is translated into all nine of the other languages offered.

From the search screen, users simply select their preferred language and enter the search terms, and the software translates the query as appropriate for each database. Users then receive the relevance-ranked results list, with the option to translate the results into their language as well. Upon viewing a specific record, users again have the option to translate the bibliographic record (title, abstract, etc.) into the language of their choice.

Figure 1. User enters search terms and selects language.

Figure 2. Query is translated into languages of all databases, searches are completed, and results are combined into a single, relevance-ranked list.

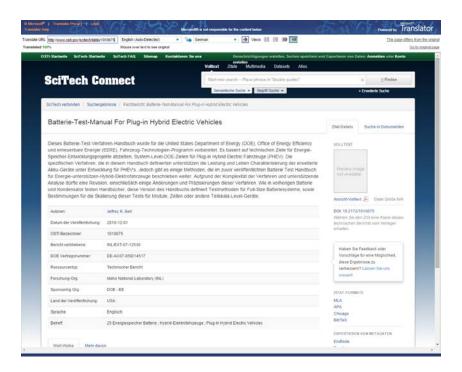Figure 3. User translates results into their preferred language.

Figure 4. Translated bibliographic record (English to German).

Expanding Access to Multimedia Content and Scientific Research Data

New forms of scientific information, such as numeric data, multimedia, and social media, are emerging rapidly and becoming increasingly prevalent as a primary means of scientific communication. Many scientific conferences and symposia, for instance, are now recorded, and presentations in video format are available to the public. Multimedia information introduces some special challenges, such as the lack of written transcripts, minimal metadata (no abstracts or keywords), and complex scientific/technical/medical terminology. Additionally, many of these videos are long, up to an hour or more in length. For a scientist interested in only one particular part of a video or experiment, locating it could represent a substantial time burden.

In addition to the U.S. Department of Energy's (DOE) ScienceCinema product, which contains over 3,500 videos from DOE national laboratories and from CERN, several other WWS.org sources have multimedia content, including sources from other U.S. government agencies and from the German National Library of Science and Technology. Searches of multimedia content were integrated into WWS.org in 2011, and users have the option of viewing a separate results list of relevant multimedia items. In the case of ScienceCinema, the actual point in the video where the search terms occur is identified in the WWS.org results list, and the user is able to view the video by clicking on the "snippet" links.
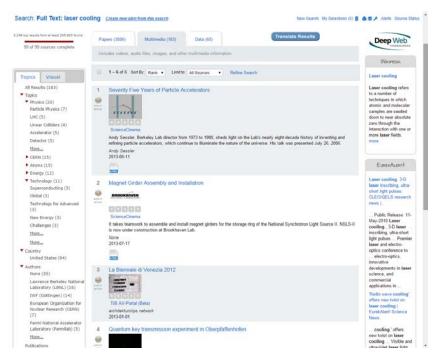

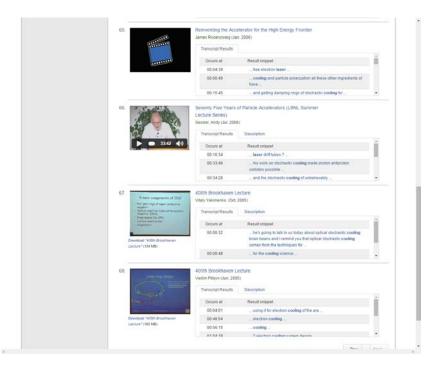Figure 5. Multimedia search results.

Figure 6. Snippets identify search terms where they occur in the video. Users can play the exact segments of the videos where the terms were spoken.

The inclusion of the audio indexing technology in ScienceCinema, also made available via a collaborative partnership with Microsoft Research, marks the first usage of such speech recognition technology in a federated search environment. As multimedia content continues to grow, WWS.org plans to expand its offerings by adding new sources.

Recognizing the emergence and importance of data accessibility, WWS.org began including scientific data collections in 2012. Users seeking scientific datasets can conduct a real-time, one-stop search and immediately gain access not only to the metadata, but to the actual scientific data itself – representing a significant milestone in improving access to scientific data from the world.

WWS.org's unique federating searching capability meets many of the challenges users face in the discovery of scientific and numeric data. For instance, unless users are very familiar with a particular data center, or know that specific datasets exist, it is very difficult to identify and locate scientific data. WWS.org enables users to simultaneously search across many databases, and receive consolidated, ranked results. In most cases, links direct the user to the actual underlying datasets.

Figure 7. The "Data" tab shows results for databases containing research and numeric datasets.
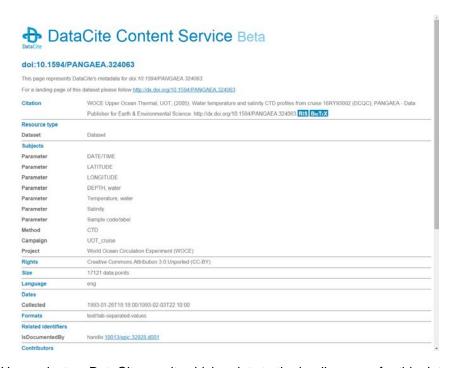

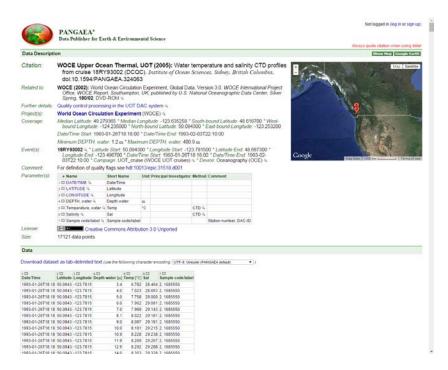Figure 8. User selects a DataCite result, which points to the landing page for this dataset.

Figure 9. User can then link to the actual data, via the landing page.

As access to scientific data becomes increasingly important, WWS.org offers the ability to easily identify, search, and access this information, contributing to the spread of scientific knowledge and advancements worldwide.

WorldWideScience.org's Future Role in Public Access to Research Data and Scholarly Publications

As the public access movement continues to expand within the United States, United Kingdom, European Union, Canada, and many other countries, the WorldWideScience Alliance envisions an important role for WWS.org. Namely, it will be possible to offer federated searching of these "public access" resources and portals, allowing users to perform one-stop searching of publicly funded research output from around the world.

Within the U.S., in February 2013, the White House Office of Science and Technology Policy (OSTP) issued a requirement to Federal agencies with over $100 million in annual conduct of research and development expenditures to develop plans to support increased public access to the results of research funded by the Federal government. Public access plans were required to address both scientific publications and digital scientific data. Up to this point, only the National Institutes of Health had a public access mandate, which it fulfills through PubMed Central. In response to the OSTP requirement, the Department of Energy released its public access plan on July 24, 2014, and in August, its corresponding public access portal, the DOE Public Access Gateway for Energy & Science (DOE PAGES[Beta]) was launched. Although it is still in a Beta version, DOE PAGES[Beta] has been added to WWS.org. While only a small subset of DOE-funded journal articles and accepted manuscripts are currently available in this initial prototype, when full implementation is underway and the product reaches maturity, DOE PAGES[Beta] is expected to grow by 20,000-30,000 articles and accepted manuscripts per year.
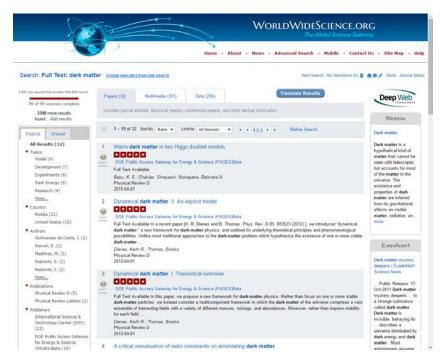
Figure 10. Public access results from the U.S. Department of Energy's PAGES[Beta]

Figure 11. A bibliographic record from PAGES[Beta]. The Accepted Manuscript will be available to public 12 months after publication.

Along with the Department of Energy, several other U.S. federal agencies have recently released public access plans, including the U.S. Department of Agriculture, Department of Defense, National Aeronautics and Space Administration, National Institute of Standards and Technology, Centers for Disease Control, Food and Drug Administration, Agency for Healthcare Research and Quality, National Institutes of Health, National Oceanic and Atmospheric Administration, National Science Foundation, and the Office of the Assistant Secretary for Preparedness and Response. Canada announced its "Tri-Agency Open Access Policy on Publications" in February 2015, and the Research Councils United Kingdom have had an Open Access policy in place for several years. Other countries are developing similar plans and policies for open and public access.

As new public and open access portals and databases become available, the WorldWideScience Alliance views these resources as a promising area for future growth, and indeed, plans are already underway to incorporate U.S. agencies' sources via Science.gov. Alliance members are highly encouraged to stay abreast of public/open access movements within their respective countries, and databases will be added to WWS.org as they are developed.

Summary – A Unique Combination of Technologies

WWS.org is an important resource for libraries around the world because of its global content and search functionality, not offered by commercial products and search engines. With the extension of federated searching on an international scale, combined with its multilingual translations capabilities, WWS.org enables access to a diverse array of scientific and technical content, in both traditional and non-traditional formats. WWS.org's pattern of continuous growth in content and technological capabilities has resulted in significantly increased usage, and as public and open access content grows, the WorldWideScience Alliance looks forward to incorporating new resources into WWS.org. By filling a unique niche in the scientific information landscape, WWS.org plays a leading role in accelerating scientific progress.