Accelerating Global Science Access:

WorldWideScience.org's Combination of Search, Translations, and Multimedia Technologies

Walter L. Warnick
U.S. Department of Energy
Office of Scientific and Technical Information
Germantown, Maryland, USA
Walter.Warnick@science.doe.gov

Brian A. Hitson
U.S. Department of Energy
Office of Scientific and Technical Information
Oak Ridge, Tennessee, USA
HitsonB@osti.gov

Lorrie Apple Johnson
U.S. Department of Energy
Office of Scientific and Technical Information
Oak Ridge, Tennessee, USA
JohnsonL@osti.gov

Abstract: WorldWideScience.org (WWS.org) is a global science gateway governed by the WorldWide-Science Alliance, with the U.S. Department of Energy Office of Scientific and Technical Information (OSTI) acting as the Operating Agent. WWS.org provides a simultaneous live search of more than 70 scientific and technical databases from government and government-sanctioned organizations representing 71 countries. From its inception in 2007, WWS.org has been at the forefront of transformational technologies, including, since June 2010, the first use of federated searching in conjunction with real-time translations capability for nine of the world's most widely spoken languages. This paper briefly reviews the history of the development of WWS.org and discusses several new features and technologies to be launched in June 2011, including the addition of Arabic to the WWS.org multilingual translations capability, expanded multimedia search functionality, and introduction of innovative mobile access. These ground-breaking technological advances will further increase WWS.org's ability to accelerate scientific discovery throughout the world.

Keywords: WorldWideScience.org, WorldWideScience Alliance, Federated Searching, Multilingual, Machine Translation, Multimedia Scientific Content, Mobile Phone Applications, International Collaboration

1. Introduction

As methods and practices for sharing and communicating scientific information rapidly evolve, WorldWide-Science.org (WWS.org) is, likewise, rapidly adopting new technologies and strategies to accelerate scientific discovery. The history of WorldWideScience.org since its prototype development in 2007 is well documented. While this paper will briefly summarize that history, the primary purpose is to describe the WWS.org technical growth and diversification strategies implemented since 2009, including the specific new features and breakthroughs to be launched in June 2011. These include (a) transition of the beta version of Multilingual WWS.org to the default search for WWS.org; (b) addition of Arabic to WWS.org's multilingual translations capability; (c) integration of multimedia scientific content into WWS.org searches; and (d) development of a mobile web version of WWS.org (and why this is particularly notable). In addition to discussion of these particular technological advances, additional aspects of WWS.org growth and uniqueness will also be explored.

2. Brief History

2.1. The Model – Science.gov

WWS.org was conceived and developed using the model of Science.gov and its underlying federated search tech-

nology. Launched in December 2002, Science.gov provides a single search point to "over 45 databases and 200 million pages of science information" within 14 U.S. federal science agencies [1].

National governments frequently face the challenge of improving the visibility and accessibility of information and other database and website content. In many cases, citizens are indifferent as to which governmental agency conducted research in a particular area; they simply want to access *any* government information on a specific topic. It is quite common that multiple agencies address different aspects of a particular discipline and research topic. Before Science.gov, in the U.S. case, a person would need to be aware of each relevant agency's databases and websites in order to conduct a comprehensive search across the entire government. While having this awareness was possible (albeit unlikely), the follow-on temporal barriers (i.e., the time required to search individual sources sequentially and the time required to sort and order results by relevance) made such searching a practical impossibility for any busy researcher.

Science.gov solved all of these problems by providing users a single point to simultaneously search all of the U.S. government's scientific resources and retrieve relevance-ranked results. Science.gov has been cited by numerous public and private sector information advocates for increasing transparency to government-sponsored scientific information [2].

2.2. WorldWideScience.org – Launch, Governance, and Growth

As a principal figure in the development of Science.gov and subsequently its Operating Agent, Dr. Walter L. Warnick, Director of the U.S. Department of Energy's Office of Scientific and Technical (OSTI), in June 2006, first proposed extending the Science.gov model on an international scale and invited other nations to partner in creating a "Science.world." The rationale behind the concept was fairly obvious and simple: if finding disparate and decentralized scientific databases in a single country is a barrier to information visibility and use, then finding information from geographically-dispersed databases around the world was an even greater barrier to scientific communication. The first "partner" to accept Dr. Warnick's invitation was the British Library, and in January 2007 this partnership was formalized, with other nations invited to join the effort. By June 2007, the "Science.world" concept was realized with the launch of the WorldWideScience.org prototype at the International Council for Scientific and Technical Information (ICSTI) annual conference in Nancy, France. The prototype initially searched 12 national scientific databases in 10 countries.

To reflect its multinational content, the initial partners in WWS.org established a multilateral governance structure called the WorldWideScience Alliance and elected officers from Africa, Asia, Europe, and North America, further evidence of its geographic diversity and representation. Current Executive Board Members include Richard Boulderstone, British Library (Chair), Pam Bjornson, Canada Institute for Scientific and Technical Information (Deputy Chair), Tae-sul Seo, Korea Institute of Science and Technology Information (Treasurer), Roberta Shaffer, International Council for Scientific and Technical Information (Ex-Officio Member), Walter Warnick, U.S. Department of Energy Office of Scientific and Technical Information, WorldWideScience.org Operating Agent (Ex-Officio Member), and Martie van Deventer, Council for Scientific and Industrial Research of South Africa (At-Large Delegate).

To avoid overlap and duplication of commercially-available scientific information resources, the WorldWideScience Alliance established criteria for the kinds of databases it would consider for WWS.org searches. The over-riding principle was that databases must be produced, sponsored, or endorsed by a national scientific body (for example, Science.gov (U.S.), the Institute of Scientific and Technical Information of China's databases, Japan Science and Technical Agency's J-STAGE and J-EAST databases). Using these primary criteria, WWS.org quickly grew beyond the initial 12 databases via one of two typical pathways: (a) as the WWS.org Operating Agent OSTI actively identified and contacted national databases seeking permission to include them in WWS.org searches or (b) vice versa, and increasingly commonly, national databases contacted WWS.org seeking to be included in WWS.org searches. 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 71 countries and 77 databases by March 2011.

Figure 1 - WorldWideScience.org Home page

3. Measures of Uniqueness and Value

3.1. Uniqueness of Records

A significant quantity of scientific information can be found by commercial and certain publisher-operated search engines. Historically, however, commercial search engines have not been able to perform real-time searching of deep web[1] databases, relying on automated crawls of static web pages to populate their enormous indexes. There are some exceptions to this general model, including, for example, arrangements such as Google Scholar's coverage of scholarly literature and the adoption of sitemap protocols by database owners to expose their content to commercial search engines' crawlers.

A key feature and founding principle of WWS.org is that it searches national scientific databases and nationally-sponsored or –endorsed sources. To avoid overlap and redundancy with commercial and publisher sources, WWS.org does not directly search any commercial scientific database or website. To gauge whether this principle results in unique search results, the WWS.org Operating Agent performed two comparative analyses (September 2009 and February 2011) of search results from Google, Google Scholar, and WWS.org.

In the September 2009 analysis, search results were captured from each of the three search engines, using 33 queries across a wide range of scientific disciplines. A key area of interest was the amount of overlap in the results sets. Overlap was defined as exact title matches of the records. Across the 33 queries and accounting for all results, WWS.org results were uniquely different from Google and Google Scholar 96.5 percent of the time. Assuming that users typically focus on the first 50 results, there was a 9 percent overlap between the first 50 WWS.org hits and the Google and Google Scholar items. In other words, WWS.org results were unique 91 percent of the time within the first 50 hits.

The same analysis was repeated in February 2011. Using the same methodology with the same 33 queries, WWS.org results were 92.7 percent unique. Comparing the first 50 results from each search engine, the overlap had declined to 2.4 percent, or, stated differently, the uniqueness had increased to 97.6 percent.

[1] The Deep Web (also called Deepnet, the invisible Web, DarkNet, Undernet or the hidden Web) refers to World Wide Web content that is not part of the Surface Web, which is indexed by standard search engines. (Wikipedia, 2011)

The primary conclusion from the second analysis was that WWS.org uniqueness remains relatively high and especially high among the first 50 results. In other words, WWS.org is filling a niche by searching national scientific databases.

3.2. Usage Growth

A major challenge for any new search product, such as WWS.org, is building brand identity/recognition and user awareness. This is particularly true for a product that essentially has no advertising or marketing expenditures. One method by which WWS.org has addressed this challenge is by leveraging the referral power of major commercial search engines. Since the introduction of search engines' sitemap protocols, this technique has become easier by simply exposing deep web content (e.g., electronic full-text documents) to the sitemaps. The content then gets crawled by the search engines and can subsequently be found in, for example, a Google search, and linked to from the set of search results. This process works quite well for centralized databases, but it is not particularly effective for federated search engines such as WWS.org because WWS.org performs a search of decentralized databases. To benefit from Google's and others' referral power, OSTI developed a program that used results from live searches of WWS.org to create a list of over one million unique scientific query terms. Over five million search results were then searched and categorized by the scientific query terms to create a set of cached, topical search results pages (what OSTI termed "topic pages"). These topic pages were exposed to search engines through a sitemap.

Figure 2 – WWS.org Topic Page

Following the indexing of these topic pages, WWS.org web traffic roughly doubled during the first month after deployment. Traffic has continued to growth at a brisk rate. Current numbers show a ten-fold increase compared to the amount of traffic experienced prior to im-

plementation of the topic pages. Once a user finds a WWS.org-generated topic page, he or she is then on the "premises" of WWS.org and has the option to repeat the search in real-time on WWS.org. Along with growing user awareness and repeat visits, the topic page-generated usage has also contributed to significant direct WWS.org queries. The average number of queries per month has tripled during 2011, as compared to the average number of queries per month during 2010.

4. Expanding Reach and Content Types

4.1 Multilingual Translations

When WWS.org was launched in 2007, it initially 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 non-English-speaking populations and (b) excluding access to the expanding volume of non-English papers in countries such as Brazil, China, France, Germany, Japan, Russia, and many others.

For non-English-speaking populations, this meant that WWS.org had played little role in ameliorating the "digital divide," an issue addressed by developmental initiatives and organizations. Strategies for closing the digital divide have been stated and implemented perhaps most eloquently and effectively by the World Summit on the Information Society (WSIS). The justification for closing the divide was partially captured in this WSIS statement, "The ability for all to access and contribute information, ideas and knowledge is essential in an inclusive Information Society."[3]. Of course, more tangibly, such access provides the intellectual and physical foundations for health systems, scientific facilities, infrastructure, and technology-driven commerce. Considering the ubiquitous nature of globalization (including science), providing access to English-language science to non-English-speaking populations has the potential for accelerating the rate of scientific progress.

Conversely, there is an increasing volume of non-English scientific content, both conventional and non-conventional literature, being produced for national journals, institutional repositories, and regional databases. Addressing the issue of language barriers in science, Meneghini and Packer stated in 2007, "Is there Science beyond English? Initiatives to increase the quality and visibility of non-English publication might help to break down language barriers in scientific communication." [4]. Barany, in 2005, also stated, "As globalization increases, communication between linguistic communities could become a serious stumbling block." [5].

Indeed, of the world's "top 400" institutional repositories, 250, or 63 percent, have some or all non-English content [6]. Japan, France, Germany, Brazil, China, Russia, and other countries all produce vast amounts of science in their respective national languages. Practically all Russian scientific output is in Russian. In the case of China, prior to June 2010, WWS.org searched a small subset of English records from the Institute of Scientific and Technical Information of China (ISTIC). ISTIC holds a much larger collection of Chinese language content. According to Mr. Wu Yishan, Chief Engineer at ISTIC,

> "In China today, we estimate that people who master sufficient English account for at most 5 percent of the urban labor force. On the other hand, more and more people in the world expect to understand China deeply, and the number of Chinese learners outside of China will reach 100 million by 2010, according to an expert estimate. For international community who are eager to keep abreast of the development of China's science and technology, understanding some Chinese is of particular importance. 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."[7].

The continuing growth of such non-English content creates its own 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, but automated/machine translations have historically lacked the precision necessary for scientific translation, and the application has generally been done on a one-to-one basis; that is, translating from a single language to another single language.

In a federated search environment, such as WWS.org, there was the potential to search databases with content in multiple languages simultaneously, but the WWS.org search engine 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. In

June 2010, a beta version of Multilingual WWS.org was launched at the ICSTI annual conference in Helsinki.

The beta version provided translation capabilities for nine languages (Chinese, English, French, German, Japanese, Korean, Portuguese, Spanish, and Russian). 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 3 – Multilingual WorldWideScience.org[Beta]

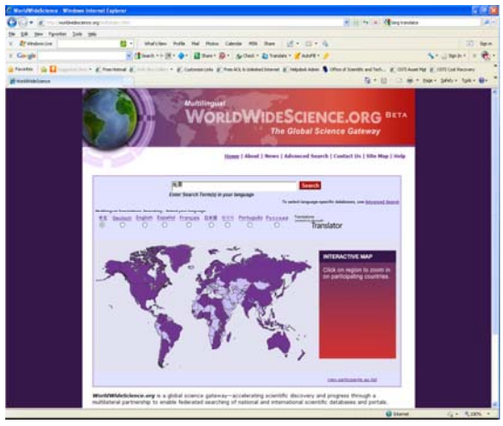


Figure 4 – User's query has been translated and results returned. User clicks on Translate Results button to translate.

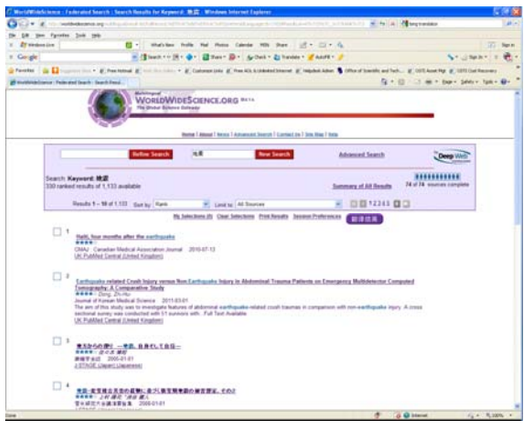


Figure 5 – Results have been translated into user's native language.

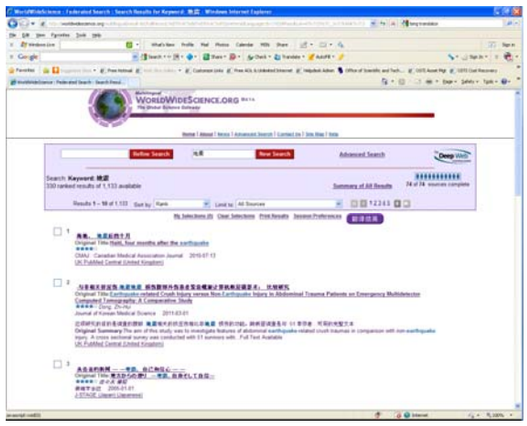
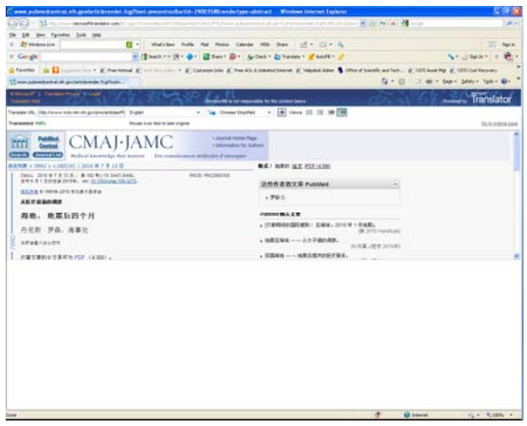


Figure 6 – Users also have the option to translate the record into their preferred language.



Since the beta multilingual version was released in June 2010, the system has proven quite stable, and the WWS Alliance agreed in February 2011 to incorporate multilingual translations into the main WWS.org product by June 2011. Users coming to the main WWS.org search page will be immediately offered the option of searching in one of ten languages. The original nine languages used for the beta version will be retained, and Arabic will be offered. Adding Arabic presented some additional technical challenges, as the browser must accommodate right-to-left text. The WWS Alliance strongly endorsed adding Arabic, the fifth most widely-spoken language in the world, and to provide increased access to scientific and technical resources written only in Arabic.

Once multilingual translations are integrated into the production version of WWS.org, additional topic pages will be generated using terms and phrases from the ten

languages currently available. This technique will continue to increase both accesses and queries, with the additional benefit of expanding potential users' awareness of WWS.org.

4.2 Access to Multimedia-based Science and Technology

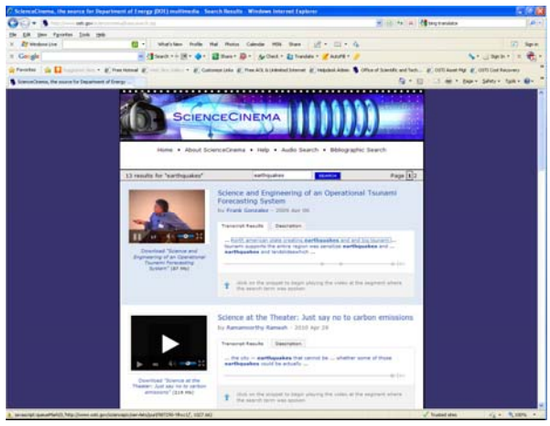
New forms of scientific information, such as numeric data, multimedia, and social media, are emerging rapidly and becoming increasingly prevalent as a primary means of scientific communication. Many scientific and technical conferences and symposia, for instance, are now recorded, and presentations in video format are available for public access. 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 [8].

In February 2011, the WWS.org Operating Agent, OSTI, released a new product called ScienceCinema, which contains approximately 1,000 hours of research-based videos from the U.S. Department of Energy's National Laboratories. In a collaborative partnership with Microsoft Research, this project utilizes Microsoft Research Audio Video Indexing System (MAVIS). MAVIS is a set of software components that use speech recognition technology to enable searching of digitized spoken content [9].

Figure 7 – ScienceCinema Home Page

The end result is that users can search for a precise term within the video and be directed to the exact point in the video where the term was spoken. Once a search is conducted, the results list contains "snippets" and a timeline with links showing the occurrence of the search term. The user simply clicks on the link, and the video will begin to play at the point where the search term was spoken.

Figure 8 – ScienceCinema results list showing snippets where search term was spoken during the videos.



In addition to OSTI's ScienceCinema product, several other WWS.org sources have multimedia content, including, most prominently, CERN and several U.S government agencies. Searches of multimedia content will be integrated into WWS.org in June 2011. Users will continue to have the option of viewing traditional text-based results, along with a separate results list of relevant multimedia items. In the case of ScienceCinema, the actual point in the video where the search terms occur will be identified in the WWS.org results list, and the user will be able to view the video by clicking on the "snippet" links.

Although Microsoft's MAVIS technology has been used in other applications besides ScienceCinema, the integration of ScienceCinema into WWS.org will represent the first use of this audio indexing technology in a federated search environment. It is anticipated that the MAVIS technology will be applied to other WWS.org multimedia sources in the near term. Beyond this, future goals include exploring multilingual translations capabilities for multimedia, in conjunction with the capabilities WWS.org now offers for text-based information.

4.3 Mobile WWS.org – Another First for Federated Search

Growth in mobile phone usage in the last decade has been exponential. A report from the United Nations released in March 2009 indicated that the number of

mobile phone subscriptions throughout the world quadrupled, from 1 billion in 2002 to 4.1 billion by December 2008 [10]. The majority of this growth has emanated from developing countries. In fact, mobile phone usage is growing faster among African countries than anywhere else in the world [11]. Studies indicate that mobile phones allow developing countries to leapfrog old technologies. In the U.S., about 91 percent of the population subscribe to mobile phone services [12].

Beyond the virtual ubiquity of mobile phones, the devices have become increasingly "smart" and capable of rapid web transactions involving significant amounts of data. With the convergence of mobile device availability and capability, the WWS Alliance supported the creation of a mobile version of WWS.org. Although the current production version of WWS.org will work on a mobile phone, the graphics and other content are often difficult to read on smaller devices. A version of WWS.org developed specifically for mobile usage will be released in June 2011.

Mobile WWS.org features a streamlined search screen and results list. Users may view records directly at the original source database by clicking on the links or have the option to email the results list for later viewing on a PC or MAC. Mobile WWS.org works with major popular devices and operating systems, such as the iPhone, Android, and Blackberry.

Figure 9 – Mobile WWS.org on an iPhone

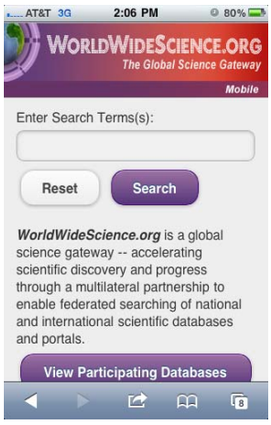
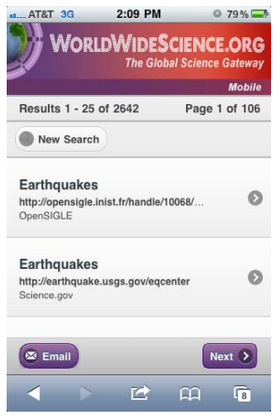


Figure 10 – Mobile WWS.org results on an iPhone



While a mobile version of a web product is not new, Mobile WWS.org offers unique capabilities, as it enables federated searching of nearly 80 databases, many of which are not individually optimized for mobile web searching. Access to scientific and technical information provided by national and international entities, such as those represented in WWS.org, is important, particularly within developing countries. Because mobile phone usage within these areas is expanding rapidly, mobile applications may be the most successful means of reaching these users. After release, usage of Mobile WWS.org will be monitored and additional features will be added according to user feedback and demand.

5. Summary – A Unique Combination of Technologies

At its inception in 2007, WorldWideScience.org represented the unique extension of federated searching technology on an international scale. Its growth to searching nearly 80 databases across all inhabited continents represents a novel scaling of this technology. In 2010, federated search was combined with multilingual translations capability, another first in the blending of technologies to accelerate access to scientific information. And, in 2011, WWS.org builds on this innovation with the addition of Arabic to multilingual translations; the integration of speech-indexed multimedia search and retrieval; and a mobilized version of international federated search. This pattern of continuous growth in content and technological capability has resulted in significantly increased usage and unique search results. As a result, WWS.org is filling a niche in the scientific information landscape and playing a leading role in accelerating scientific progress.

References

[1] Science.gov website, "About Science.gov," http://www.science.gov/about.html, March 2011.

[2] Science.gov website, "Communications Archive Page," http://www.science.gov/communications/archive.html#photo, March 2011.

[3] Report of the Geneva Phase of the_World Summit on the Information Society," GENEVA-PALEXPO, 10-12 December 2003.

[4] R. Meneghini and A. L. Packer, "Is there science beyond English? Initiatives to increase the quality and visibility of non-English publications might help to break down language barriers in scientific communication," in EMBO reports (2007), vol. 8, pp. 112-116.

[5] M. J. Barany, "Science's Language Problem," in Bloomberg Business Week, 16 March 2005.

[6] "Ranking of World Repositories," Cybermetrics Lab http://repositories.webometrics.info/toprep.asp.

[7] Wu Yishan, Remarks at "From Information to Innovation," International Council for Scientific and Technical Information (ICSTI) Annual Conference, Helsinki, Finland, June 2010.

[8] B. Chitsaz and L. A. Johnson, "ScienceCinema: Multimedia search and retrieval in the sciences," Multimedia and Visualisation Innovations for Science, ICSTI Winter Workshop, Redmond, WA, USA, February 2011.

[9] Microsoft Research Audio Video Indexing System (MAVIS) website, http://research.microsoft.com/en-us/projects/mavis/.

[10] J. Fermoso, "U.N. world report picks up massive growth in mobile phone ownership," Wired.com, March 2009.

[11] "Mobile phone growth fastest in Africa," BBC News website, March 2005 http://news.bbc.co.uk/2/hi/business/4331863.stm.

[12] C. Foresman, "Wireless survey: 91% of Americans use cell phones," Ars Technica website, March 2010, http://arstechnica.com/telecom/news/2010/03/wireless-survey-91-of-americans-have-cell-phones.ars.