

WorldWideScience.org: Bringing Light to Grey

Brian A. Hitson and Lorrie A. Johnson
Office of Scientific and Technical Information
U.S. Department of Energy

Abstract

WorldWideScience.org¹ and its governance structure, the WorldWideScience Alliance², are putting a brighter spotlight on grey literature. Through this new tool, grey literature is getting broader exposure to audiences all over the world. Improved access to and sharing of research information is the key to accelerating progress and breakthroughs in any field, especially science.

WorldWideScience.org has revolutionized access to “deep” web scientific databases. These nationally- and internationally-sponsored databases are comprised of both grey and conventional literature. Consequently, because grey literature is naturally less familiar (and, hence, less accessible) than conventional literature, it receives a disproportionate benefit in terms of usage through its exposure in WorldWideScience.org.

Before expanding on the mechanics and contents of WorldWideScience vis-à-vis grey literature, it is helpful to characterize what is meant by “grey literature.” The term “Grey Literature” can be defined in several ways. Wikipedia³, for example, describes grey literature as “...a body of materials that cannot be found easily through conventional channels such as publishers...” The National Library of Australia⁴ provides a slight variation: “...information that is not searchable or accessible through conventional search engines or subject directories and is not generally produced by commercial publishing organizations.” This description goes further to describe electronic grey literature as constituting the “hidden” or “deep” web. Most laypeople, those outside the professional information community, would think of the color “grey” and may be puzzled as to why a color is used to describe literature. To them, the word “grey” likely brings to mind the Webster⁵ definition, “an achromatic color between the extremes of black and white.”

Traditionally, “white” has been equated with conventional, published literature, but perhaps to better illustrate the point, it could be useful to reverse the “achromatic” color spectrum in this case. The extreme of “black,” for example, could be thought of as traditional black ink printed on paper. It consists of words that are very clear and easily accessible to everyone, and makes up the conventional literature such as journals, books, and published proceedings. “White,” on the other hand, conveys just the meaning of a blank sheet with no words – simply unrecorded ideas, concepts, and thought. So, then, “grey” is between these two extremes. It includes the kinds of literature that information professionals typically associate with “grey,” such as preprints, technical reports, theses and dissertations. More recently, grey literature also includes emerging forms of

information such as numeric data, multimedia, recorded academic lectures, and Web 2.0-generated information.

Looking back at the National Library of Australia's definition for a moment, though, it also implies that grey literature comprises the "hidden" or "deep" web. "Grey" is synonymous with "deep" when it comes to the Internet; grey literature, more than any other type, is a body of information that resides in the "deep web" and is not easily found.

To put this concept in context, there is a distinction between the "surface web" and the "deep web." Generally, major search engines such as Google⁶ and Yahoo!⁷ are searching web pages on the surface web. These are static web pages that are crawled by Google's automatic crawler, where every word on a page is stored in Google's massive index, and the power and sophistication of Google's systems allows it to return millions of hits in milliseconds.

However, the surface web is not where most scientific literature resides. Instead, it resides in databases that typically have their own search interface, and because the contents of those databases do not sit on a static web page, they are not typically indexed by Google. There are ways for databases to expose their contents for Google's crawlers, but by and large, most database owners do not do so. Therefore, this information is firmly planted in the "deep web," only accessible through the database's own search engine. Most experts estimate that the deep web is hundreds of times larger in terms of content than the surface web. Clearly, this situation calls for a solution, which is offered by WorldWideScience.org.

Unfortunately, the perception among a large percentage of internet users is that if it can not be found by one of the big search engines, it must not exist. So, the first challenge of the deep web is a variation on an old cliché, "what you don't know can hurt you, or at least it could help you." For example, if a person with cancer is only searching the surface web to learn about latest clinical trials, she would be missing substantive and possibly helpful information that may reside in key deep web databases. If a scientist wants to explore the latest developments in photovoltaics, he will be missing the most in-depth information if he limits his searches to the surface web. The key challenge here is that most people are unaware of all the rich resources in the deep web.

Making the unrealistic assumption, however, that the world is replete with people who already know about the multitude of deep web databases relevant to their particular field, there is a second key challenge. This challenge is that searching all of these databases individually, one by one, is not physically possible, or at least it will consume precious time needed for actual research and experimentation. Thus, progress will be thwarted.

These challenges can be overcome through the use of federated search technology – essentially becoming a Google or a Yahoo! for the deep web. In a federated search, a single portal is connected to multiple deep web database search engines. A person enters a search query into a single Google-like search box. The query is then sent simultaneously to the many databases that have been previously identified as relevant to

the specialty of the federated search engine. These individual search engines receive the query, perform their own searches, and return results to the federated search engine. The combined results are then ranked using a relevance algorithm (just as Google does) with parameters such as where the query terms appear in the title, how often they appear, and other variables.

A search in a federated search engine is not as fast as Google because live searches of the databases are occurring, but results are generally produced within 30 seconds. Working with other federal science agencies in the United States, the Office of Scientific and Technical Information (OSTI)⁸, first introduced federated searching with Science.gov⁹, which searches practically all federal science databases.

Building on the successful model of Science.gov, OSTI then used this technology to develop other federated search tools for more niche communities.

ScienceAccelerator.gov¹⁰ federates searches of all of OSTI's web systems. The E-Print Network¹¹ specializes in federated searches of e-print databases in the U.S. and several other countries. Science Conference Proceedings¹² federates the search of several professional societies' conference databases. Lastly, the Federal R&D Project Summaries¹³ does the same for databases describing ongoing research projects sponsored by the U.S. government.

Science.gov was a major success as a friendlier way to make government-sponsored science information available to the public, and it won significant praise as a "government-to-citizen" model under the President's e-government agenda. The logical extension of Science.gov as a national federated searching model is that there could be a "Science.world" for a global federated search tool. Nations interested in promoting science globally could allow their individual science databases to be searched by a single portal – something that is not possible with major commercial search engines.

Following the success of Science.gov, Dr. Walter Warnick, OSTI's Director, introduced the concept of a Science.world before the public conference of the International Council for Scientific and Technical Information (ICSTI)¹⁴ in June 2006. Dr. Warnick invited other national libraries to help OSTI implement the concept. The British Library¹⁵, much to its credit and vision, quickly offered a hand of partnership in this effort. In January 2007, the British Library Chief Executive, Dame Lynne Brindley, and the U.S. Under Secretary for Science in the Department of Energy, Dr. Raymond Orbach, signed a statement of intent to partner in the effort, which also invited other nations to join in this partnership.

Between January and June 2007, several other countries participated in offering their databases to demonstrate that federated search could work on an international level. Recognizing that "dot world" was used to simply draw the analogy to Science.gov, a more descriptive and operable web address was needed, and WorldWideScience.org was chosen, with the tag line, "The Global Science Gateway." The first prototype of WorldWideScience.org was demonstrated at the ICSTI public conference in Nancy, France. At that time, twelve databases from ten countries were represented in the searches

of WorldWideScience.org. The successful demonstration of the prototype clearly had the desired effect, as it garnered significant press coverage. In a follow-up ICSTI meeting, it was agreed that ICSTI would play a significant role in helping to form a governance structure for WorldWideScience.org. The formation of the WorldWideScience Alliance was formalized in June 2008 at ICSTI's conference in Seoul. Thirty-eight countries were represented in signing a declaration committing their support to the effort. Completing an international cooperative in a year's time, including terms of reference and governance language, is a reflection of the goodwill and support that this concept received around the world. The Alliance Executive Board is led by Richard Boulderstone of the British Library, who was elected as the Alliance's first Chairperson. A diverse mix of officers from North America, Europe, Asia, and Africa make up the remainder of the Board. The leadership of ICSTI was invaluable in providing a platform to promote this concept to national scientific and technical information officials around the world.

Since the first prototype of twelve databases in ten countries, WorldWideScience.org has now grown to 49 databases in 54 countries (as of December 2008). The scientific content represented in these searches comes from countries accounting for over three-fourths of the world's population. It is estimated, using rough calculations, that these searches cover 375 million pages of science, much of which is obviously grey literature.

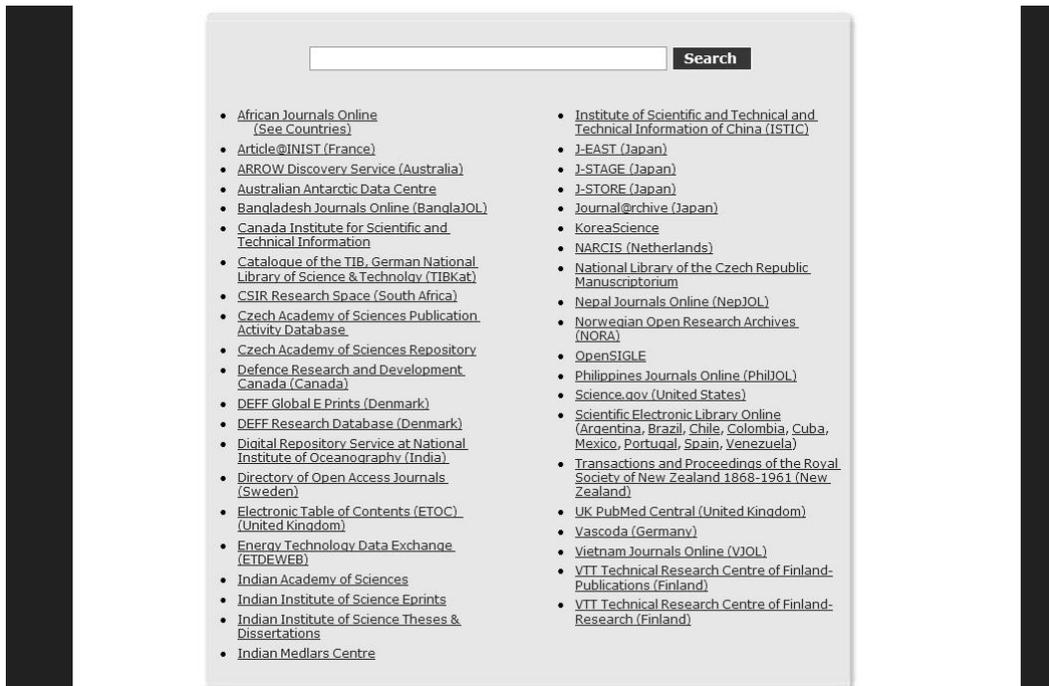
Figure 1 WorldWideScience.org



A map of the world (Figure 1) is used to show which countries have databases represented in WorldWideScience.org. At this stage, these are all databases which have some element of national or international sponsorship rather than commercial databases,

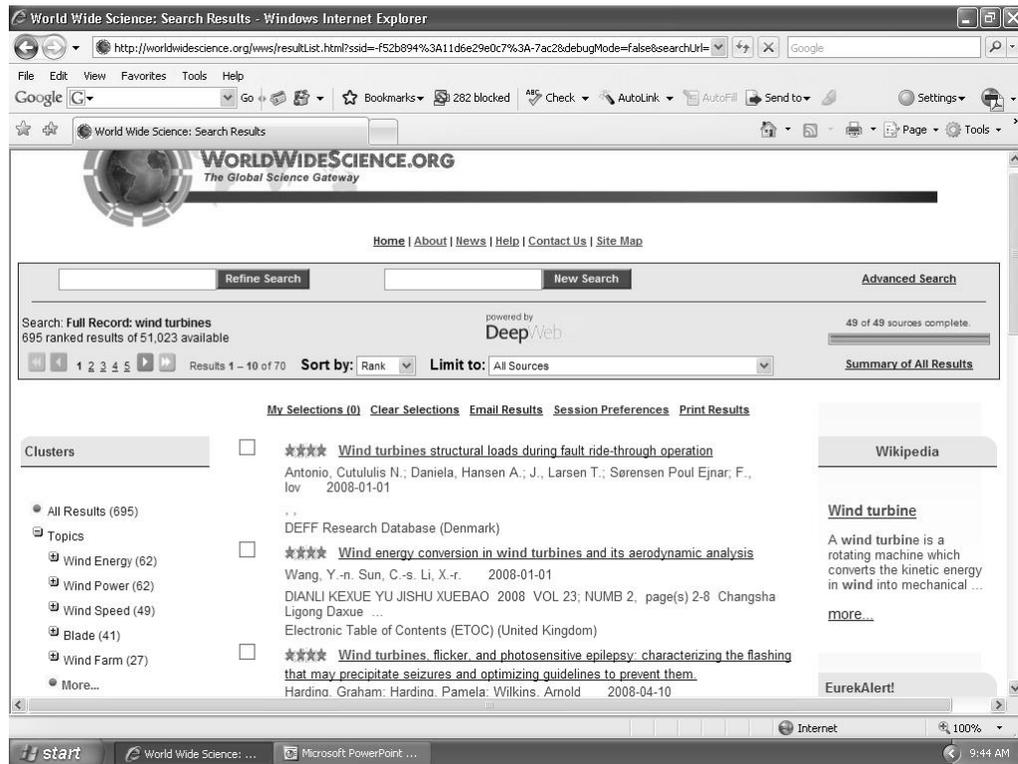
such as publisher databases. As indicated by the map, sources are covered from practically all of North and South America, Australia, a significant portion of Europe, and major segments of Asia and Africa. Some countries have multiple sources. Japan, for example, has four major databases from the Japan Science and Technology Agency¹⁶; India also has four sources. The U.S. source is Science.gov, which is itself a federated search portal of over 30 major databases. In this case, where one federated search engine spawns a search of another federated search engine, it is called nested searching, and it works quite efficiently.

Figure 2 WorldWideScience.org Databases/Portals



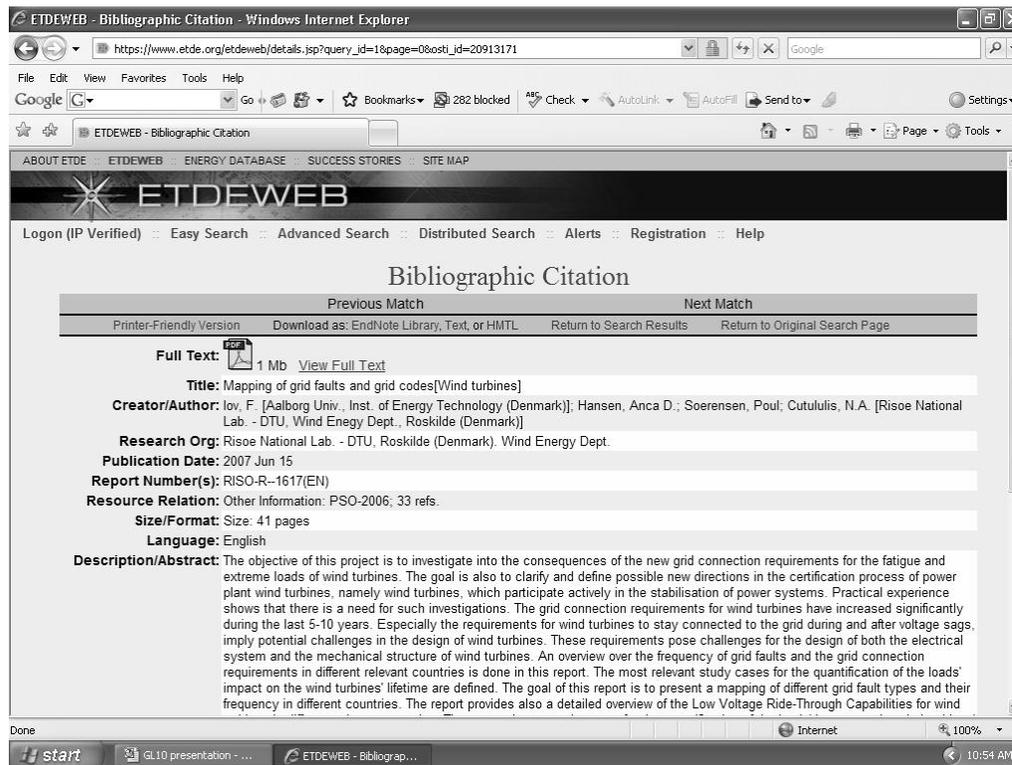
To illustrate with some examples, Figure 3 shows a typical first page of search results from WorldWideScience.org. A search on “wind turbines” has been conducted. All 49 sources were successfully searched, and all together, the sources had over 51,000 records matching this exact phrase. WorldWideScience provides, in this case, the top 695 ranked results. There is a trade-off between showing more results versus the speed of the search; so, typically, the search limits any given source to the top 100 results. A user, if interested in seeing all results, can go to the link “summary of all results” and see which sources have more than 100 results. The user could then go directly to that source for a more in-depth search. Relevance is reflected through the stars (1 through 4) that appear beside each result. Two new enhancements were recently added. On the left side, the user is offered clustering to allow for narrowing results into more refined sets. On the right side, a Wikipedia definition, if one exists, is given for the search term. This is particularly useful for users who simply want to become more familiar with a particular field of science.

Figure 3 WorldWideScience.org Search Results



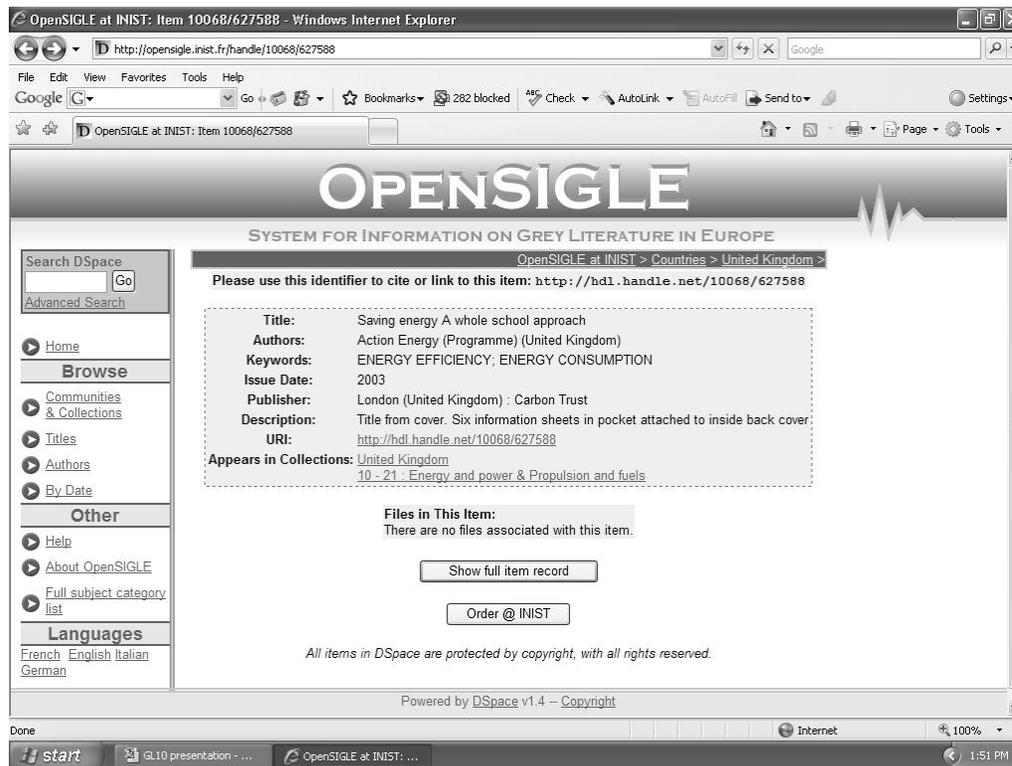
Once the user selects a specific record to view, WorldWideScience.org then takes the user directly to that record within the original database/portal. For example, this record (Figure 4) comes from the Energy Technology Data Exchange (ETDE)¹⁷. ETDEWEB is an international database on energy technology governed by an agreement under the auspices of the International Energy Agency. The agreement is comprised of sixteen member countries, who, along with other partners have built a database of 4 million records. As evident by the record, this is clearly a “grey literature” report emanating from Risoe National Laboratory in Denmark. A link to the full text document in PDF format is provided.

Figure 4 ETDEWEB Record



WorldWideScience.org also searches OpenSIGLE¹⁸, the system for information on grey literature in Europe. This record (Figure 5) shows how the user could order the full text document from INIST¹⁹, the Alliance member from France.

Figure 5 OpenSIGLE Record



Other examples of records include the sub-element of the Norwegian Open Research Archive²⁰, the Bergen University open research archive (Figure 6). The government of South Africa, through its Council for Scientific and Industrial Research²¹, was one of the earliest supporters of WorldWideScience.org. Its record (Figure 7) also provides the ability to view full text. Working closely with the Alliance member, the International Network for the Availability of Scientific Publications (INASP)²², a number of on-line journal collections from developing countries are available through WorldWideScience.org. These countries include 24 African nations, Bangladesh, Nepal, Philippines, and Viet Nam. A record from Nepal, again providing a link to the full text, is shown in Figure 8. Finally, the last example (Figure 9) shows a record from the Australian open access ARROW²³ system, which covers the repository of over half of all universities in Australia. Again, a thesis is a good example of grey literature, and a link to the full text is provided in this case as well.

Figure 6 NORA Record

Bergen Open Research Archive: Cutpoints for mild, moderate and severe pain in patients with ost - Windows Internet Explorer

https://bora.uib.no/handle/1956/2700

BORAUiB
BERGEN OPEN RESEARCH ARCHIVE

Search BORA

Bergen Open Research Archive > Faculty of Medicine > Department of Public Health and Primary Health Care > Department of Public Health and Primary Health Care >

Please use this identifier to cite or link to this item: <http://hdl.handle.net/1956/2700>

Files in This Item:

File	Description	Size	Format
BIOmed_Cutpoints.pdf		296.46 kB	Adobe PDF View/Open

Title: Cutpoints for mild, moderate and severe pain in patients with osteoarthritis of the hip or knee ready for joint replacement surgery

Authors: Kapstad, Heidi
Hanestad, Berit R.
Langeland, Norvald
Rustøen, Tone
Stavem, Knut

Issue Date: 21-Apr-2008

Publisher: BioMed Central

Figure 7 CSIR Record

CSIR Research Space: Clean coal technology: gasification of South African coals - Windows Internet Explorer

http://researchspace.csir.co.za/dspace/handle/10204/2526

CSIR
our future through science

CSIR Research Space
Open and unrestricted access
to our research outcomes

Search ResearchSpace

CSIR Research Space >
CSIR Conference 2008 >
CSIR Conference 2008 >

A link to the full text document is supplied after the abstract.

Please use this identifier to cite or link to this item: <http://hdl.handle.net/10204/2526>

Title: Clean coal technology: gasification of South African coals

Authors: Engelbrecht, AD
North, BC
Hadley, TD

Keywords: Coal
Gasification
Fluidised bed
Combined cycle
Modelling
Integrated gasification combined cycle
Thermogravimetric analysis

Issue Date: Nov-2008

Citation: Engelbrecht, AD, North, BC and Hadley, TD. 2008. Clean coal technology: gasification of South African coals. Science real and relevant: 2nd CSIR Biennial Conference, CSIR International Convention Centre Pretoria, 17&18 November 2008, pp 12

Figure 8 Nepal Journals Online Record

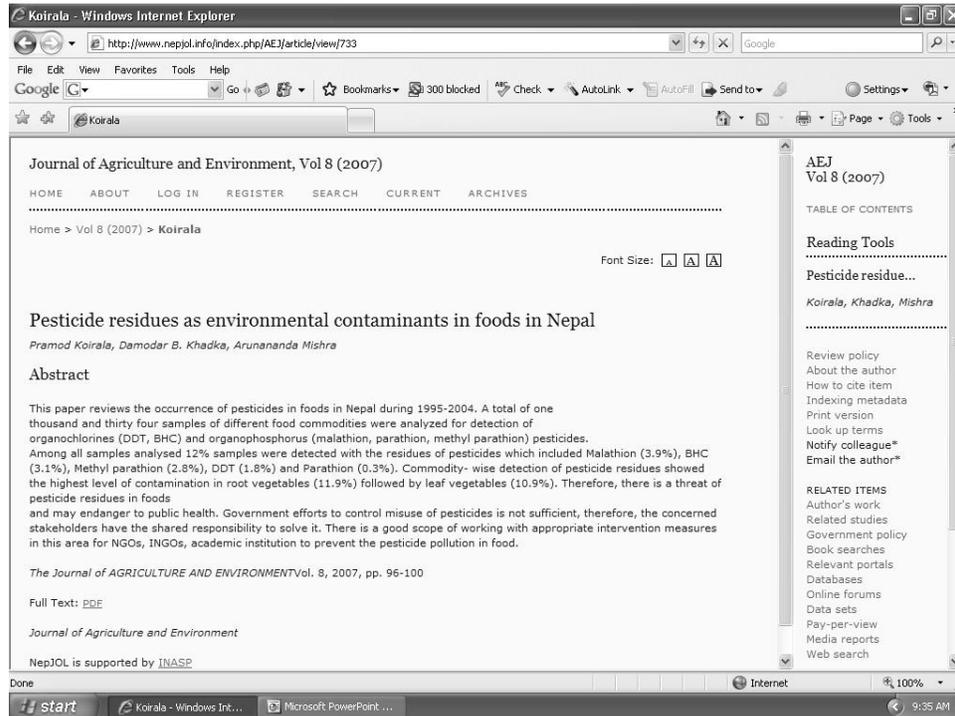
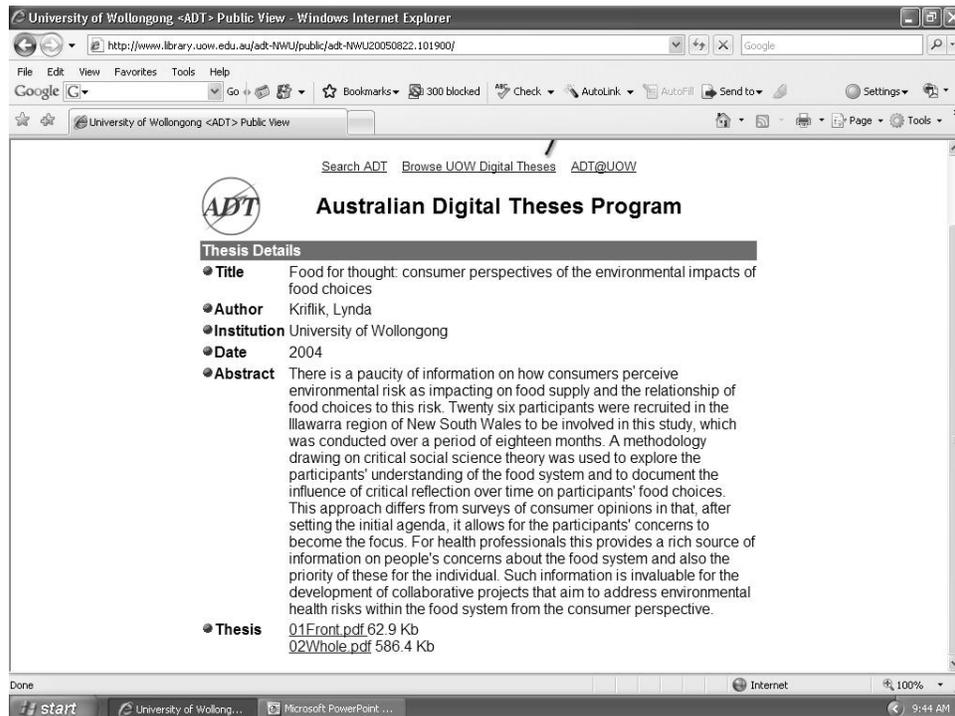


Figure 9 ARROW Record



WorldWideScience.org is continuing to grow consistently both in terms of content and usage. On the content front, the most notable recent addition is an English-language journal source from China. The symbolic significance of opening this access to Chinese science cannot be overstated, and the cooperation of the staff at the Institute of Scientific and Technical Information of China²⁴ was much appreciated.

No one really knows ultimately how many sources exist that would make WorldWideScience.org the most comprehensive gateway to nationally- and internationally-sponsored science research, but at around 100 sources, the speed and efficiency of the search engine may start to degrade. A vast amount of computing is involved in processing so many results from so many sources. One of the challenges WorldWideScience.org faces in the future will be overcoming this scalability issue. Strategies have been defined for overcoming this challenge. At a simpler level, one planned enhancement is to offer an alerts service. A user will be able to create a profile and be alerted when any of the WorldWideScience.org sources has added new materials matching that profile.

Another challenge WorldWideScience.org hopes to address in the future is providing access to non-English sources. A few of the Alliance members have experience in this area, particularly INIST in France. The WorldWideScience.org team will begin exploring translation modules that will open access to sources that only exist in a native language, such as the Chinese record in Figure 10.

Figure 10 Chinese Language Record

中国科学技术信息研究所
国家工程技术数字图书馆

注册 | 登录 | 忘记密码 | 帮助

服务中心 | 资源导航 | 馆藏检索 | 科技信息 | 研究报告 | 科学评价 | 刊物出版 | 学术研究 | 院士著作馆 | 统一检索

您现在的位置是:

摘要信息

Characterization of a Maize Retrotransposon Introgressed into The Wheat DH Plant Genome through Wheat*Maize Cross and Its Transmission in the Progenies; Identification and Chromosomal Location of a New Tandemly Repeated in Maize; A Comparative Mapping of Two Wheat Markers Linking to Pm20 Gene on Rice Chromosomes

陈纯霞
中国科学院遗传学研究所 博士论文 1997
指导老师: 朱立煌

原文下载

摘要: 该论文的研究结果包括三个部分,第一部分是博士论文已获得的研究结果基础上的继续,主要通过运移杂种导入小麦基因组中的玉米特异的重复序列的功能特性及其在DH3后代的遗传传递进行了分析,为了便于读者了解这些新结果,特在该部分简单叙述了以前取得的有关结果(见2.1.4.1,小字体。)第二部分则是从玉米的随机基因组文库中鉴定了一个新的串联重复序列,第三部分是与该实验室的八六三课题"黑麦中的抗小麦白粉病基因Pm20的分离"有关的小麦与水稻基因组比较作图研究的初步结果。

There are also challenges, not just for WorldWideScience.org, but for all in the grey literature community, with emerging formats such as YouTube videos, podcasts, and other audio and visual sources. There has been a proliferation recently of sites offering access to these types of files. For example, there is a small database of video files of academic lectures from the Fermi National Laboratory²⁵ in the United States (Figure 11), but files such as this are truly in the deep web and are not accessible beyond this interface.

Figure 11 Fermi National Laboratory Video Archive

The screenshot shows a web browser window titled "Video - Search - Streaming Video - Windows Internet Explorer". The address bar contains a URL from the Fermi National Laboratory VMS site. The page content includes a navigation menu on the left, a "Video Search Results" section, and a table of search results.

Video Search Results

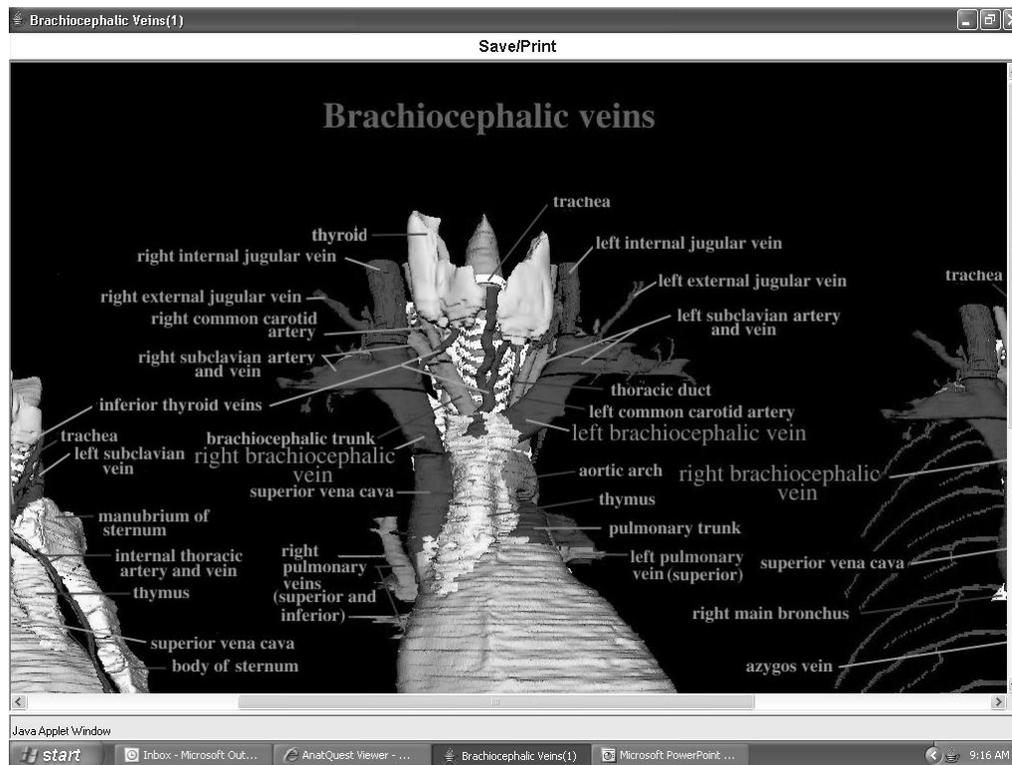
Displaying records 1 through 25 of 85 records found. (25 records displayed).
[Next page of records.](#)

Date	Title	Presenter	Series	Length	Tech. Level	Podcast ? MP3 ? CD ?
11/08/2007	Experimental Signatures for Extra Dimensions in Space - Part 2	Greg Landsberg	Academic Lectures	01:30:00	Physicist	NONE NONE CD
11/06/2007	Physics in Extra Dimensions - Part 4	Bogdan Dobrescu	Academic Lectures	01:30:00	Student	NONE NONE CD
11/01/2007	Experimental Signatures for Extra Dimensions in Space - Part 1	Greg Landsberg	Academic Lectures	01:30:00	General Public	NONE NONE CD
10/30/2007	Physics in Extra Dimensions - Part 3	Bogdan Dobrescu	Academic Lectures	01:30:00	Student	NONE NONE CD
10/18/2007	Physics in Extra Dimensions - Part 2	Bogdan Dobrescu	Academic Lectures	01:30:00	Physicist	NONE NONE CD
10/16/2007	Physics in Extra Dimensions - Part 1	Bogdan Dobrescu	Academic Lectures	01:30:00	Physicist	NONE NONE CD
03/06/2007	QCD effects in B decays - Lecture 3	Thomas Becher	Academic Lectures	01:00:00	Physicist	NONE NONE DOWNLOAD CD
03/01/2007	Lattice QCD with applications	Andreas	Academic	01:30:00	Student	NONE NONE

The browser's taskbar at the bottom shows the Start button, several open applications (Inbox - Microsoft Outlook, Video - Search - Streaming Video, Microsoft PowerPoint), and the system tray with the Internet icon, 100% zoom, and the time 9:08 AM.

Beyond sound and video files, there are some fascinating image databases (photographs, drawings, illustrations) that need to be more accessible. This highly-detailed medical illustration (Figure 12) resides in a National Library of Medicine images database²⁶, but the terms on the drawing would not be indexed by a major search engine, leaving this significant resource potentially under-utilized by the public and medical communities.

Figure 12 National Library of Medicine Image



With the prominence of computational sciences, simulation, and the use of measurements in so many fields, numeric data sets are also critical to advancing science. Yet they are hardly integrated at all into traditional textual search engines, let alone in any meaningful federated way across data sets. This, too, is a rich opportunity for expanding and improving access to valuable information.

ICSTI, who provided invaluable leadership for WorldWideScience.org, is sponsoring a number of technical projects that address some of these challenges. In the area of numeric data, TIB-Hannover in Germany is leading a multinational project to demonstrate the integration of access to numeric data sets from within grey literature textual reports.

On the multimedia front, ICSTI is leading a project to demonstrate how indexing of spoken words in audio and video files can result in profoundly improved search precision. Another ICSTI project is exploring how Web 2.0 technology can be used to improve scientific communication.

Conclusion

Through the demonstration of WorldWideScience.org, it is clear that *grey is global*, and it has benefited from a global solution. Second, *grey is growing*, both in traditional formats and media but also in emerging forms, which need to be offered the same level and ease of access as textual literature. Finally, *grey is good*, and must be treated as an essential commodity for progress in all fields, especially science, medicine, and technology.

References

1. <http://worldwidescience.org/>
2. <http://worldwidescience.org/alliance.html>
3. http://en.wikipedia.org/wiki/Gray_literature
4. <http://www.nla.gov.au/padi/topics/372.html>
5. <http://www.websters-online-dictionary.org/definition/grey>
6. <http://www.google.com/>
7. <http://www.yahoo.com/>
8. <http://www.osti.gov/>
9. <http://science.gov/>
10. <http://www.scienceaccelerator.gov/>
11. <http://www.osti.gov/eprints/>
12. <http://www.osti.gov/scienceconferences/>
13. <http://www.osti.gov/fedrnd/>
14. <http://www.icsti.org/>
15. <http://www.bl.uk/>
16. <http://www.jst.go.jp/EN/>
17. <http://etde.org/>
18. <http://opensigle.inist.fr/>
19. <http://international.inist.fr/>
20. <http://www.ub.uio.no/nora/noaister/search.html?siteLanguage=eng>
21. <http://www.csir.co.za/>
22. <http://www.inasp.info/>
23. <http://search.arrow.edu.au/>
24. <http://www.istic.ac.cn/>
25. http://vms-db-srv.fnal.gov/fmi/xsl/VMS_Site_2/000Search/video/f_streaming.xsl
26. <http://anatquest.nlm.nih.gov/AnatQuest/ViewerApplet/aqrendered.html>