



Federated Search: Breaking Down the Language Barrier

Abe Lederman, President and CTO
Deep Web Technologies, Inc.
2010 NFAIS Annual Conference
March 1, 2010



About Deep Web Technologies...

- Founded by Abe Lederman, a co-founder of Verity, 2002
- Pioneered federated search technology
- Over \$3M in R&D
- Production applications since 1999
- Based in Santa Fe, New Mexico
- 22 person company with strong executive team





Importance of Multilingual Search

- Increases the value of research output by making it available to a wider audience
- Makes available research from China, Japan, Russia, and other countries prolific in science publication
- Greatly broadens the scope of patent research

Importance of Multilingual Search (cont.)

- Exposes English speakers to diverse perspectives from researchers in foreign countries



English Isn't the Only Language that Matters

Thomson Reuters Research Reveals...

- China's research output far outpacing the rest of the world
- China surpassed Japan, the UK and Germany in 2006 and now stands second only to the USA
- At this pace, China will overtake the USA within the next decade
- Brazil's share of research output is growing rapidly

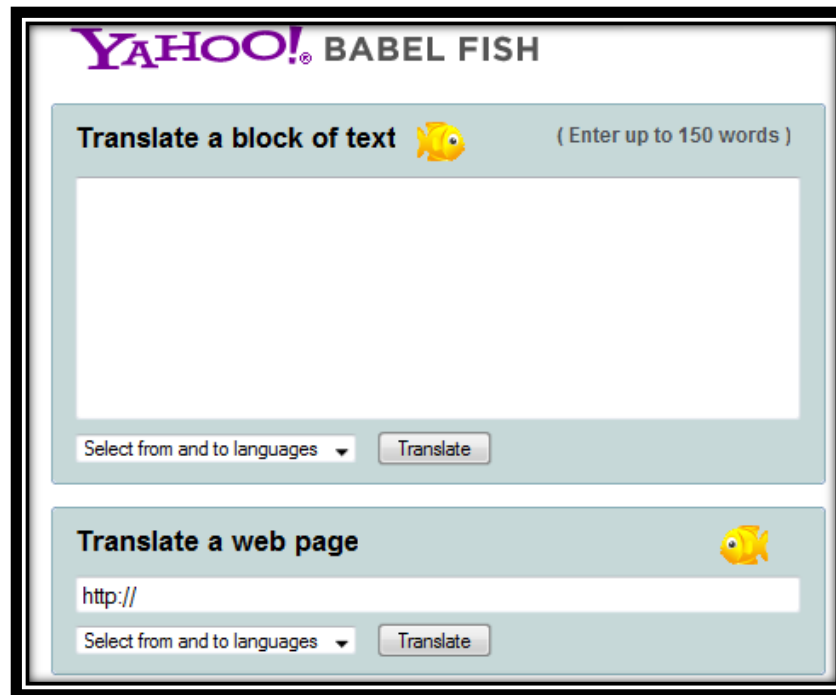
Babel Fish Popularized Machine Translation on the Web

- The first European language translation service for web content
- Launched 12/9/97 by DEC's Alta Vista and SYSTRAN S.A.
- Babel Fish, in "The Hitchhiker's Guide to the Galaxy", is a fish you stick in your ear that allows humans to speak and understand any language
- When released, Babel Fish understood five European languages: French, German, Italian, Portuguese and Spanish



Babel Fish Popularized Machine Translation (cont.)

- SYSTRAN, founded in 1968, leveraged the results of 20 years of military-industrial research



The screenshot shows the Yahoo! Babel Fish web interface. At the top, the Yahoo! logo is in purple, followed by "BABEL FISH" in black. Below this, there are two main sections. The first section is titled "Translate a block of text" and includes a yellow fish icon and the instruction "(Enter up to 150 words)". It features a large white text input area. Below the input area is a dropdown menu labeled "Select from and to languages" and a "Translate" button. The second section is titled "Translate a web page" and also includes a yellow fish icon. It has a text input field with "http://" pre-filled. Below this field is another dropdown menu labeled "Select from and to languages" and a "Translate" button.

Fun Facts About Machine Translation

- In 1954, the Georgetown-IBM experiment, involved fully automatic translation of more than 60 Russian sentences into English and ushered in the era of significant funding for machine translation
- The authors of the Georgetown experiment claimed that within three or five years, machine translation would be a solved problem

Fun Facts About Machine Translation (cont.)

- In the 17th century, philosophers Leibniz and Descartes proposed codes to relate words between languages
- The first patents for "translating machines" were applied for in the mid 1930s.
- One patent, issued in 1933, was for a storage device using paper tape to find the equivalent of any word in a foreign language

Approaches to Machine Translation

Rule-based Machine Translation:

- Requires extensive lexicons with morphological, syntactic, and semantic information, and large sets of rules
- Users can improve the out-of-the-box translation quality by adding their terminology into the translation process

Approaches to Machine Translation (cont.)

Statistical Machine Translation

- The most widely studied approach to machine translation
- Utilizes statistical translation models whose parameters stem from the analysis of monolingual and bilingual corpora

Approaches to Machine Translation (cont.)

Statistical Machine Translation (cont.)

- Building statistical translation models is a quick process, but the technology relies heavily on existing multilingual corpora
- A minimum of 2 million words for a specific domain and even more for general language are required

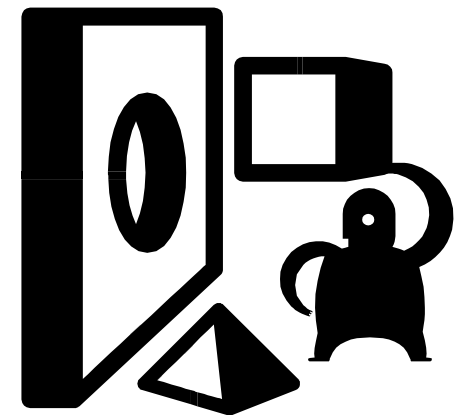
Approaches to Machine Translation (cont.)

Hybrid Machine Translation

- Leverages strengths of rule-based and statistical approaches
- Rules are used to pre-process data in an attempt to better guide the statistical engine
- Rules are also used to post-process the statistical output to perform functions such as normalization

Major Issues with Machine Translation

- Disambiguation - distinguishing between different meanings of a word ("bridging the gap" vs. "dental bridge" vs. "bridge loan" vs. "suspension bridge")
- Harder disambiguation when the text itself is ambiguous



Major Issues with Machine Translation (cont.)

- Idioms - words cannot be translated literally, especially between languages: "hear" vs. "Hear, Hear!"
- Morphology - different word orders
- Words not in the translator's vocabulary
- Translating science has fewer issues

Multilingual Federated Search: State of the Art

- Results merging strategy: Si, Callan, and Others; 2008
- Research into scalable searching of heterogeneous multilingual collections: Powell and Fox; 1998
- Cross-Language Evaluation Forum (CLEF) promotes R&D in multilingual information access

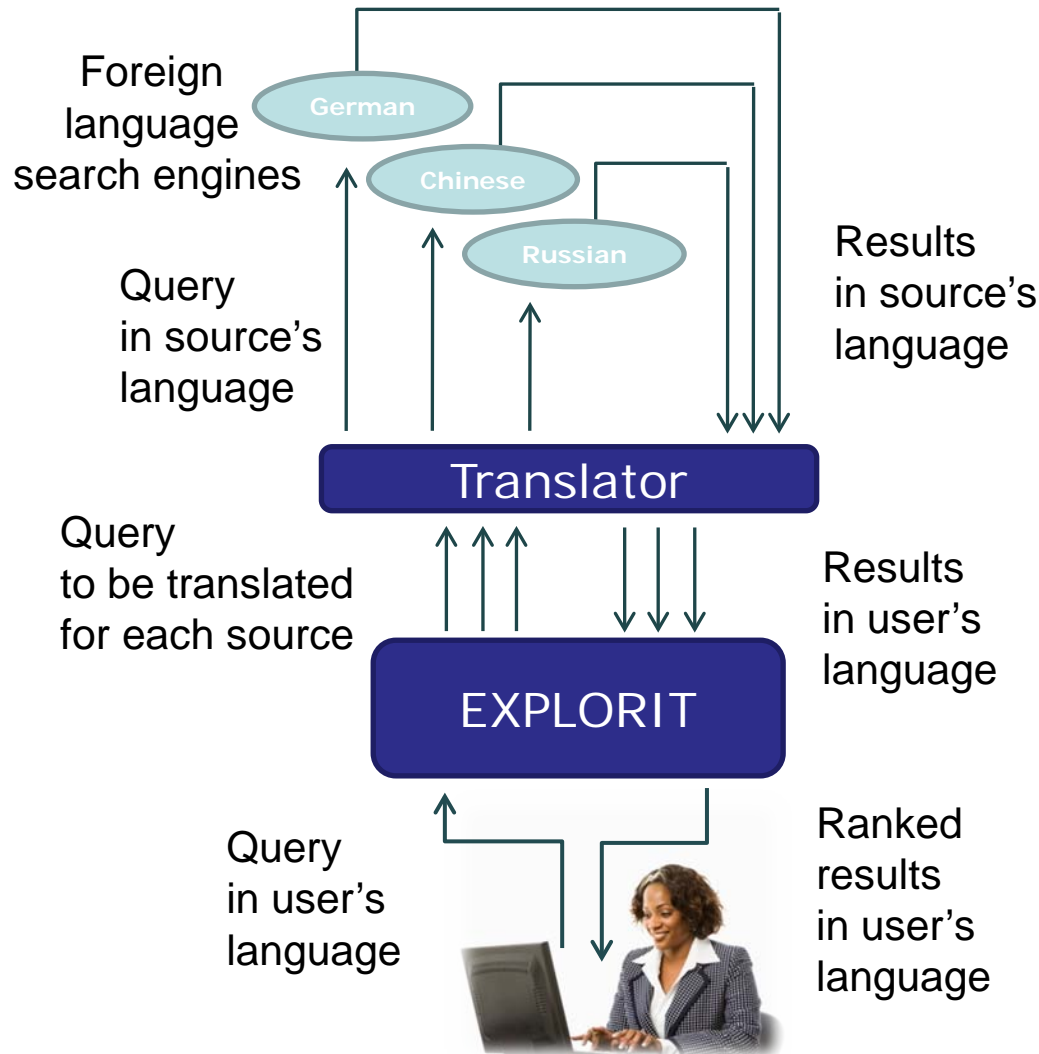
How Multilingual Federated Search Works

1. User enters query in their native language
2. Explorit translator engine translates 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

How Multilingual Federated Search Works (cont.)

5. Explorit translator engine translates the results summaries (title, snippet) into the user's native language
6. Results summaries from different sources are aggregated
7. Results summaries are ranked
8. Results summaries are displayed to the user

How Multilingual Federated Search Works (cont.)



Players in the Machine Translation Space



- One of the oldest machine translation companies, founded in 1968
- Uses hybrid machine translation technology it developed
- Has done extensive work for the US Department of Defense and the European Commission

Players in the Machine Translation Space



- Founded in 2002
- Uses statistical techniques from cryptography
- Applies machine learning algorithms that automatically acquire statistical models from existing parallel collections of human translations

Players in the Machine Translation Space (cont.)

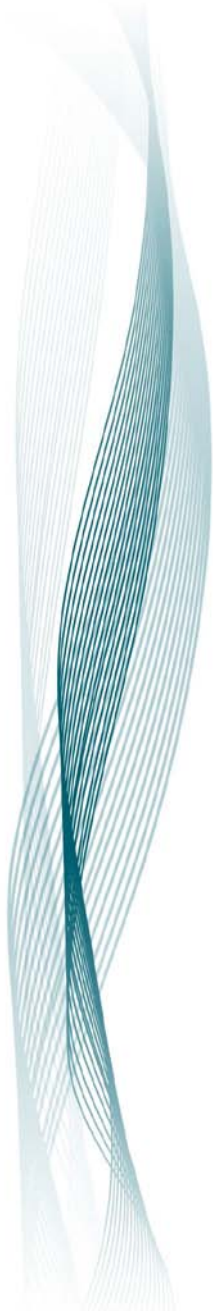
Google translate

- Uses its own translation software, used SYSTRAN until circa 2007
- Based on statistical machine translation
- Google built a 6-language corpus of 20 billion words' worth of human translations from a large set of UN documents, which are normally available in the 6 UN languages

Players in the Machine Translation Space (cont.)



- Powered by Microsoft Translation
- Based on statistical machine translation
- Once used SYSTRAN, now using system developed by Microsoft Research



WorldWideScience.org is an Excellent Candidate for Multilingual Search

- A global gateway to international science databases and portals
- All content is from national governments or vetted by national governments
- Developed and maintained by the DOE Office of Scientific and Technical Information, OSTI
- One-stop searching
- Will include databases from China, Japan, Korea, Germany, and other non-English countries



WORLDWIDE SCIENCE.ORG

The Global Science Gateway

[Home](#) | [About](#) | [News](#) | [Advanced Search](#) | [Contact Us](#) | [Site Map](#) | [Help](#)

Search

- [African Journals Online \(See Countries\)](#)
- [Article@INIST \(France\)](#)
- [ARROW Discovery Service \(Australia\)](#)
- [Australian Antarctic Data Centre](#)
- [Bangladesh Journals Online \(BanglaJOL\)](#)
- [Canada Institute for Scientific and Technical Information](#)
- [CERN Document Server](#)
- [CSIR Research Space \(South Africa\)](#)
- [Czech Academy of Sciences Publication Activity Database](#)
- [Czech Academy of Sciences](#)
- [Defence Research Canada \(C\)](#)
- [DEFF Global](#)
- [DEFF Research](#)
- [Digital Repository European](#)
- [Digital Repository Institute of](#)
- [Directory of Open Access Journals \(Sweden\)](#)
- [Electronic Table of Contents \(ETOC\) \(United Kingdom\)](#)
- [Energy Technology Data Exchange \(ETDEWEB\)](#)
- [Environment Research Funders' Forum \(ERFF\) Database - United Kingdom](#)
- [German National Library of Science and Technology \(TIBKAT\)](#)
- [Index Scriptorium Estoniae \(Estonia\)](#)
- [Indian Academy of Sciences](#)
- [Indian Institute of Science Eprints](#)
- [Indian Institute of Science Theses & Dissertations](#)
- [Indian Medlars Centre](#)
- [Indonesia Journals Online \(IJO\)](#)
- [Institute of Scientific and Technical Information of China \(ISTIC\)](#)
- [International Development Research Centre \(IDRC\) Digital Library - Canada](#)
- [International Nuclear Information System \(INIS\)](#)
- [International Science & Technology Center \(ISTC\)](#)
- [J-EAST \(Japan\)](#)
- [J-STAGE \(Japan\)](#)
- [J-STORE \(Japan\)](#)
- [Journal@rchive \(Japan\)](#)
- [KoreaMed](#)
- [KoreaScience](#)
- [Philippines Journals Online \(PhilJOL\)](#)
- [Science.gov \(United States\)](#)
- [Scientific Electronic Library Online \(Argentina, Brazil, Chile, Colombia, Cuba, Mexico, Portugal, Spain, Venezuela\)](#)
- [Sri Lanka Journals Online \(SLJOL\)](#)
- [Transactions and Proceedings of the Royal Society of New Zealand 1868-1961 \(New Zealand\)](#)
- [UK PubMed Central \(United Kingdom\)](#)
- [Vascoda \(Germany\)](#)
- [Vietnam Journals Online \(VJOL\)](#)
- [Virtual Library of Lithuania](#)
- [VTT Technical Research Centre of Finland-Publications \(Finland\)](#)
- [VTT Technical Research Centre of Finland-Research \(Finland\)](#)

- [VTT Technical Research Centre of Finland-Publications \(Finland\)](#)
- [VTT Technical Research Centre of Finland-Research \(Finland\)](#)

Milestones in the History of WorldWideScience.org



October 15, 2008
People's Republic of
China joins
WorldWideScience.org
Alliance

June 12, 2008
WorldWideScience.org
Agreement signed in
Korea – formalizes
commitment to
sustain and grow the
service

January 8, 2008
India added to
WorldWideScience.org

June 22, 2007
WorldWideScience.org
Launched

Jan. 21, 2007
Global
Science
Gateway
Agreement
Signed in
London



WorldWideScience.org to Debut Multilingual Searching

- Deep Web Technologies has partnered with OSTI to introduce multilingual searching to WorldWideScience.org
- Free service to be launched in June
- Launch will be at the International Council for Scientific and Technical Information (ICSTI) meeting in Helsinki in June of this year
- ICSTI oversees the WorldWideScience.org Alliance



Refine Search

polishing optical glass pro

New Search

[Advanced Search](#)



Search: **Full Record: polishing optical glass process**

[Create an alert from this search](#)

17 of 17 sources complete

239 ranked results of 548 available

Results 1 – 10 of 239 Sort by:

Rank

Limit to: All Sources

1 2 3 4 5

[My Selections \(0\)](#) [Clear Selections](#) [Alerts](#) [Print Results](#) [Email Results](#) [Bookmark this search](#) [Session Preferences](#)

Clusters

All Results (239)

Topics

Technology (35)

Materials (25)

Study (17)

Процессов (17)

В Конце (15)

More...

Authors

Горелик В.С.,
Рахматуллаев И.А.
(2)

Бондарь, Е.А. (2)

Шадрина, Л.П. (2)

Соловьев, Виктор
Сергеевич (2)

A., Tervonen (2)

Publications

Journal Of The
National Science
Foundation Of Sri
Lanka (13)

Applied Optics (5)

Sri Lanka Journal Of
Bio-Medical
Informatics (3)

Ceylon Medical



1 [Physico-chemical processes of polishing optical glass](#)

Original Title: Физико-химические процессы полирования оптического стекла

★★★★★ Ходаков Г.С., Кудряцева Н.П.

1985-01-01

Moscow: Mashinostroenie, 1985 Kolich.harakteristiki: 220 pp. Price: np GRNTI.; 61

Original Snippet: Москва: Машиностроение, 1985 Колич.характеристики :220 с. Цена : Б.ц. ГРНТИ : ; 61

[The Russian Union Catalog of Scientific Literature](#)



2 [Evaluation of mechanical properties in polyurethane polisher is used as abrasives in the polishing process of optical glass](#)

Original Title: Evaluation des caracteristiques mecaniques du polissoir en polyurethane utilise comme porte abrasifs durant le processus du polissage du verre

optique

★★★★★

Journal de physique. IV 2005 , Vol : 124 , p. : 123 - 128 2005-01-01

[Article@INIST \(France\)](#)



3 [Eighth China International Optoelectronic Fair \(CIOE\) - Antwerp Antwerp Electronic Ansett long Chengdu Electronic Science and Technology](#)

Original Title: 第八届中国国际光电博览会(CIOE)-安特电子安捷隆科技成都安特电子

★★★★★

Colored / colorless optical glass, special optical glass, ... optical processing testing equipment: polishing machine, lens molding equipment, electric cast-type ...

[Google China](#)



4 [Automated optical inspection system - Optical Design - Lens Design - Laser Design - Film Design - lighting design ...](#)

Original Title: 自动光学检测系统-光学设计-透镜设计-激光设计-薄膜设计-照明设计 ...

★★★★★

2007 Nian 3 Ri Yue 4 ... In many cases, process engineers ... optical glass lens molding technology, polishing Common Defects Causes and methods to overcome ...

[Google China](#)



5 [Why do some glass polishing powder polishing does not shine? - Has been answered - questions and answers horizon](#)

Original Title: 为什么有些玻璃抛光粉会抛不亮? - 已回答-天涯问答

★★★★★

Rare earth polishing powder because of its unique chemical-mechanical action principle and the ... in the process of asymmetric war ... activator LaOBr: Br (blue), to enhance



Refine Search

agua caliente

New Search

[Advanced Search](#)

Search: **Full Record: agua caliente**

284 ranked results of 13,509 available

[Create an alert from this search](#)

[Summary of All Results](#)

11 of 11 sources complete

Results 1 – 10 of 284 Sort by: Rank

Limit to: All Sources

1 2 3 4 5

[My Selections \(0\)](#) [Clear Selections](#) [Alerts](#) [Search Builder](#) [Print Results](#) [Email Results](#) [Summary of All Results](#) [Bookmark](#) [Session Preferences](#)

Clusters

All Results (284)

Topics

- Efectos (42)
- Grados (34)
- Actividad (31)
- Vapor Y Agua Caliente (29)
- Temperatura (27)

[More...](#)

Authors

- 2008 (4)
- Российская Федерация.Гос.ком.по Надзору За Безопасным Ведением Работ В Промышленности И Горном... (4)
- Hwang, Seokhwan (2)
- Chen, Jiann-Chu (2)
- Oswald, Dirk (2)
- [More...](#)

Publications

- Bioresource Technology (13)
- Journal Of Hazardous Materials (7)

- ☐ 1 [AGUA CALIENTE Y CALEFACCIÓN SOLAR](#)
Original Title: [L'ÉNERGIE SOLAIRE POUR LE CHAUFFAGE ET L'EAU CHAUDE](#)
★★★★★
Revue technique du bâtiment et des constructions industrielles 2009 , Num : 250 , p. : 28 - 30 2009-01-01
[Article@INIST \(France\)](#)

- ☐ 2 [Agua](#)
Origin
★★★★★
CFP.
[Article](#)

- ☐ 3 [Agua](#)
Origin
[\(dagg\)](#)
★★★★★
The journal of physical chemistry: A : 2010-02-10
[PubMed](#)

- ☐ 4 [Agua caliente de baterías de tanques metal-manual de A/c.1/52/7: m³ de utv.Roskommunènergo-WA - communes.HOZ - WA RSFSR 11.11.86.](#)
Original Title: [Типовая инструкция по эксплуатации металлических баков-аккумуляторов горячей воды : Утв. Роскоммунэнерго М-ва жил.-комму. хоз-ва РСФСР 11.11.86](#)
★★★★★
1986-01-01
B.m., sijercic de 1986.características: 66, [1] con: IL.Price: b.c.Gmti.; 67.53
[The Russian Union Catalog of Scientific Literature](#)

- ☐ 5 [Cálculo de calentadores de agua caliente](#)
Original Title: [Berechnung von Warmwasser-Heizungen](#)
★★★★★ Böhme, Hubert
1982-01-01
[Catalogue of the TIB. German National Library of Science & Technology \(TIBKat\)](#)

Wikipedia

[Calentamiento de agua](#)

agua es
amigo
energía
arriba

Full Record: agua caliente

Title:

Author:

Match: All Field(s)

Date Range: Pick Year to Pick Year

Translate from/to: Spanish

Search

Clear All

Help

[cuesta arriba, aunque prácticamente vuela en superficie nueva](#)

Los investigadores de ingeniería han diseñado una superficie plana que se niega a mojarse. Skitter de gotitas de agua a través de ella como rodamientos de bolas arrojó sobre hielo.

[Dispositivo de hábito-learning bajará facturas de energía bajo el nuevo esquema de reembolso de energía limpia](#)

Unidades de control inteligente que aprenden



Web ИРБИС

Библиотека Estatal de científico y técnico de pública de Rusia

Base de datos

- Catálogo electrónico gontb
- Nuevas adquisiciones en el procesamiento (libro)
- Tesis doctorales
- Catálogo consolidado de Rosisky en literatura
- Ediciones raras
- Ecología: Recursos de Internet
- Ecología: Ciencia y tecnología
- Directorio de recursos electrónicos

Catálogo de Rosisky en los resultados de la búsqueda de la literatura

Palabras clave

Buscar

Presentación de los documentos encontrados:

completa [información](#) [Resumen](#)

El número total de documentos encontrados: 1

1. **Inventario PSC N-97 13763**
Agua caliente de baterías de tanques metal-manual de A/c.1/52/7: m³ de utv.Roskommunènergo-WA - communes.HOZ-WA 11.11.86 -b.m., -1986.-66, [1] con: IL-Price: b.c.Gmtti: 67.53
Edición 800 copias.
Gmtti 67.53
Menudo 697.432.8.004.54 (083.133)
Sin perjuicio de las rubricas: backy-baterias agua caliente-operar-tèrmicos eléctricos es aplican-en-... explotación

Signas:

- 19011032 (Academia de Ciencias de la biblioteca de Rusia (prohibición))
- 10017011 (Biblioteca de Estado ruso)
- 19017073 (Biblioteca Nacional Rusia)

Búsqueda de vista

- Descripción de la base de datos
- Estándar
- Advanced
- Professional
- Distribuido
- Diccionario
- Conti-Navigator
- El menueno-Navigator
- BBC-Navigator

Web ИРБИС

Государственная публичная научно-техническая библиотека России

Российский электронный каталог по научно-технической литературе - результаты поиска

Вид поиска

Ключевые слова

Искать

Формат представления найденных документов

полный [информационный](#) [документ](#)

Общее количество найденных документов: 1

1. **Инвентарный N РКП 87-13763**
Типовая инструкция по эксплуатации металлических баков-аккумуляторов горячей воды. Утв. Роскоммунэнерго М-ва жил.-комму. хоз-ва РСФСР 11.11.86. -b.m., -1986.-66, [1] с: ил. - На рус. яз. - ГРНТИ 67.53
УДК 697.432.8.004.54(083.133)
Предметные рубрики: Баки-аккумуляторы горячих вод; Эксплуатация; Тепловые электрические станции; Оборудование; Эксплуатация

См. также:

- 19011032 (Российская академия наук. Библиотека РАН (БАН))
- 10017011 (Российская государственная библиотека)
- 19017073 (Российская Национальная библиотека)

© Векторная Ассоциация пользователей и разработчиков электронных библиотек и новых информационных технологий (Москва, 2007)

Caliente (29)

► Temperatura (27)

► More...

▼ Authors

- 2008 (4)
- Российская Федерация. Гос. ком. по Науч.
- Бел.
- Ведением Факультета Промышленности 14
- Го.
- Hwang, Seokhwan (2)
- Chen, Jiann-Chu (2)
- Oswald, Dirk (2)
- More...

▼ Publications

- Bioresource Technology (13)
- Journal of Hazardous Materials (7)

3 **Agua caliente de fría.El disociativas recomb**
Original Title: **Hot Water from Cold. The Dis**
(dagger).
★★★★★ Thomas, R. D.; Zhaunerchyk, V.; Hel
The journal of physical chemistry. A 2010-02-
[PubMed](#)

4 **Agua caliente de baterías de tanques metal-manual de A/c.1/52/7: m³ de utv.Roskommunènergo-WA - communes.HOZ - WA RSFSR 11.11.86.**
Original Title: **Типовая инструкция по эксплуатации металлических баков-аккумуляторов горячей воды : Утв. Роскоммунэнерго М-ва жил.-комму. хоз-ва РСФСР 11.11.86**
★★★★★
1986-01-01
B.m., sijercic de 1986.características: 66, [1] con: IL.Price: b.c.Gmtti: 67.53
[The Russian Union Catalog of Scientific Literature](#)

5 **Calculo de calentadores de agua caliente**
Original Title: **Berechnung von Warmwasser-Heizungen**
★★★★★ Böhme, Hubert
1982-01-01
[Catalogue of the TIB. German National Library of Science & Technology \(TIBKat\)](#)

practicamente vola en superficie nueva

Los investigadores de ingeniería han diseñado una superficie plana que se niega a mojarse. Skitter de gotitas de agua a través de ella como rodamientos de bolas arrojó sobre hielo.

[Dispositivo de hábito-learning bajará facturas de energía bajo el nuevo esquema de reembolso de energía limpia](#)

Unidades de control inteligente que aprenden

References

An effective and efficient results merging strategy for multilingual information retrieval in federated search environments

<http://portal.acm.org/citation.cfm?id=1331574>

Babel Fish

<http://www.infotektur.com/demos/babelfish/en.html>

China's Research Output More than Doubled Since 2004, Thomson Reuters Study Reveals

http://science.thomsonreuters.com/press/2009/China_Research_Output/

Comparison of machine translation applications

http://en.wikipedia.org/wiki/Comparison_of_Machine_translation_applications

Cross Language Evaluation Forum

<http://www.clef-campaign.org/>

Deep Web Technologies Developing Multilingual Translator for Federated Search

<http://www.ereleases.com/pr/deep-web-technologies-developing-multilingual-translator-federated-search-25166>

Deep Web Technologies to unveil multilingual federated search in June

<http://federatedsearchblog.com/2009/12/23/deep-web-technologies-to-unveil-multilingual-federated-search-in-june/>

References (cont.)

Deep Web Implements the Multilingual Search that Google Imagines

<http://www.globalwatchtower.com/2009/12/17/multilingual-search-deepweb-google/>

History of machine translation

http://en.wikipedia.org/wiki/History_of_machine_translation

Machine translation

http://en.wikipedia.org/wiki/Machine_translation

Multilingual Federated Searching Across Heterogeneous Collections

<http://www.dlib.org/dlib/september98/powell/09powell.html>

SYSTRAN: What is Machine Translation?

<http://www.systransoft.com/systran/corporate-profile/translation-technology/what-is-machine-translation>

Thomson Reuter Global Research Report Series

<http://researchanalytics.thomsonreuters.com/grr/>

WorldWideScience.org News/Press Releases

<http://worldwidescience.org/news.html>

Thank you!



Abe Lederman

abe@deepwebtech.com

Online Presentation:

<http://deepwebtech.com/talks/NFAIS.pdf>