HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

The Use of Machine Translation in the Russian Language Information Retrieval: case RUSSIAinfo

"Digital Semantic Content Across Cultures" – symposium. Paris. Louvre. 4-5 May 2006

Tanja Pursiainen, University of Helsinki, Aleksanteri Institute

Content

- New challenges for IR: how to retrieve THE most relevant information, irrespective of its language. Cross-language information retrieval (CLIR)
- Russian language on the Web
- Machine translation in RUSSIAinfo: documents translation into query language, query translation into document language, metadata translation, dictionary coverage
- Next: cross cultural information retrieval (CCIR)

The Google age

- Hard to come up with a search request which will not be TO SOME EXTEND satisfied by Google / a large digital library
- Problem shared by Google and large databases: MOST part of relevant documents remains unretrieved, and the users don't even know about it
- To filter out THE most relevant documents of all relevant:

- Conceptual structures for documents and queries
- Natural Language Processing techniques (NLP)

Top Ten Languages Used in the Web

Internetworldstats, updated March 31, 2006

Language	Internet Users, by Language	% of all Internet Users	World Population 2006 Estimate for Language	Internet Penetration by Language	Internet Growth for Language (2000 - 2005)
English	312,757,646	30.6 %	1,125,664,397	27.8 %	128.0 %
Chinese	132,301,513	13.0 %	1,340,767,863	9.9 %	309.6 %
Japanese	86,300,000	8.5 %	128,389,000	67.2 %	83.3 %
Spanish	80,593,698	7.9 %	429,293,261	18.8 %	229.2 %
German	56,853,104	5.6 %	95,982,043	59.2 %	106.0 %
French	40,974,004	4.0 %	381,193,149	10.7 %	235.9 %
Korean	33,900,000	3.3 %	73,945,860	45.8 %	78.0 %
Portuguese	32,372,000	3.2 %	230,846,275	14.0 %	327.3 %
Italian	28,870,000	2.8 %	59,115,261	48.8 %	118.7 %
Russian	23,700,000	2.3 %	143,682,757	16.5 %	664.5 %

RUSSIAinfo

- Reference database

- Resources: 12% in Finnish, 36% in English; **47% in Russian**, 5% in other languages

- Indexing: English and Finnish

  Finnish query retrieves documents in all languages

  English query retrieves documents in English and Russian

- Russian language proficiency of users:

- None (Russian language documents would be discarded as irrelevant without machine translation)
- Poor (unsure of document relevance)
- Passive skills (Query formulation easier in English)

NLP in RUSSIAinfo

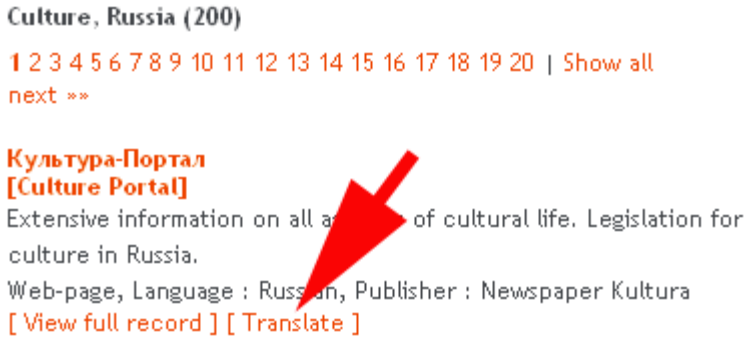
- Ready-made machine translation the only reasonable solution: designing an NLP system can take 100s of man-years

- RUSSIAinfo experience of use:

- Market Analyses: PROMT for best quality of Russian translation
- Hybrid CLIR: Document translation, query translation, dictionaries, metadata translation

Document translation

- Three different ways:
- “Translate” button



- Text translation slot
- Web page translation slot

Query translation

- Translate and search in Google.ru

Quality of result ranking: 93% vs 76% by Google.com

Overlap: 3%

- Translate and search in Yandex.ru

- Translate, search in Yandex.ru and translate results

- Translate and search wherever you want

Dictionary coverage

- Ready-made dictionaries: business dictionary set. ON/OFF option
- Specialized dictionaries for RUSSIAinfo:
  - First: Russian administrative bodies
  - Next: proper names

Original name	Machine translation	Official translation
Федеральное агентство кадастра объектов недвижимости	**Federal agency of a cadastre of objects of the real estate**	**Federal Agency of Real Estate Cadastre**
Федеральная служба по надзору в сфере природопользования	**Federal service on supervision in sphere of wildlife management**	**Federal Service for Ecology and Natural Resources Supervision**
Федеральное агентство по строительству и жилищно-коммунальному хозяйству	**Federal agency on construction and zhilishchnokommunalnomu facilities**	**Federal Agency for Construction, Housing and Communal Services**

Metadata translation

- The collection of scientific journals, RSL
- Dublin Core metadata, same as in RUSSIAinfo

- Record example:

```
<record id="7">
    <meta name="DC.Title">Web Journal of Formal, Computational &lt; Cognitive
    Linguistics</meta>
   <meta name="DC.Identifier" scheme="URL">http://fccl.ksu.ru/fcclroot.htm</meta>
   <meta name="DC.Alternative">FCCL</meta>
   <meta name="DC.Subject">искусственый интеллект, машинный перевод,
    математическая лингвистика, когнитивная лингвистика, квантитативная лингвистика,
    нейролингвистика, психолингвистика, формальные модели языка</meta>
<meta name="DC.Description.abstract">Журнал ставит своей целью объединить усилия
    лингвистов и специалистов-компьютерщиков в изучении языка с применением
    точных методов.</meta>
   <meta name="DC.Publisher.CorporateName">Казанский университет</meta>
   <meta name="DC.Contributor.CorporateName">Российская ассоциация искусственного
    интеллекта</meta>
   <meta name="DC.Contributor.CorporateName">Казанский унверситет</meta>
   <meta name="DC.Format">text/html</meta>
   <meta name="DC.Language">ru</meta>
 </record>
```

Metadata translation

- Key words:

искусственный интеллект, машинный перевод, математическая лингвистика, когнитивная лингвистика, квантитативная лингвистика, нейролингвистика, психолингвистика, формальные модели языка

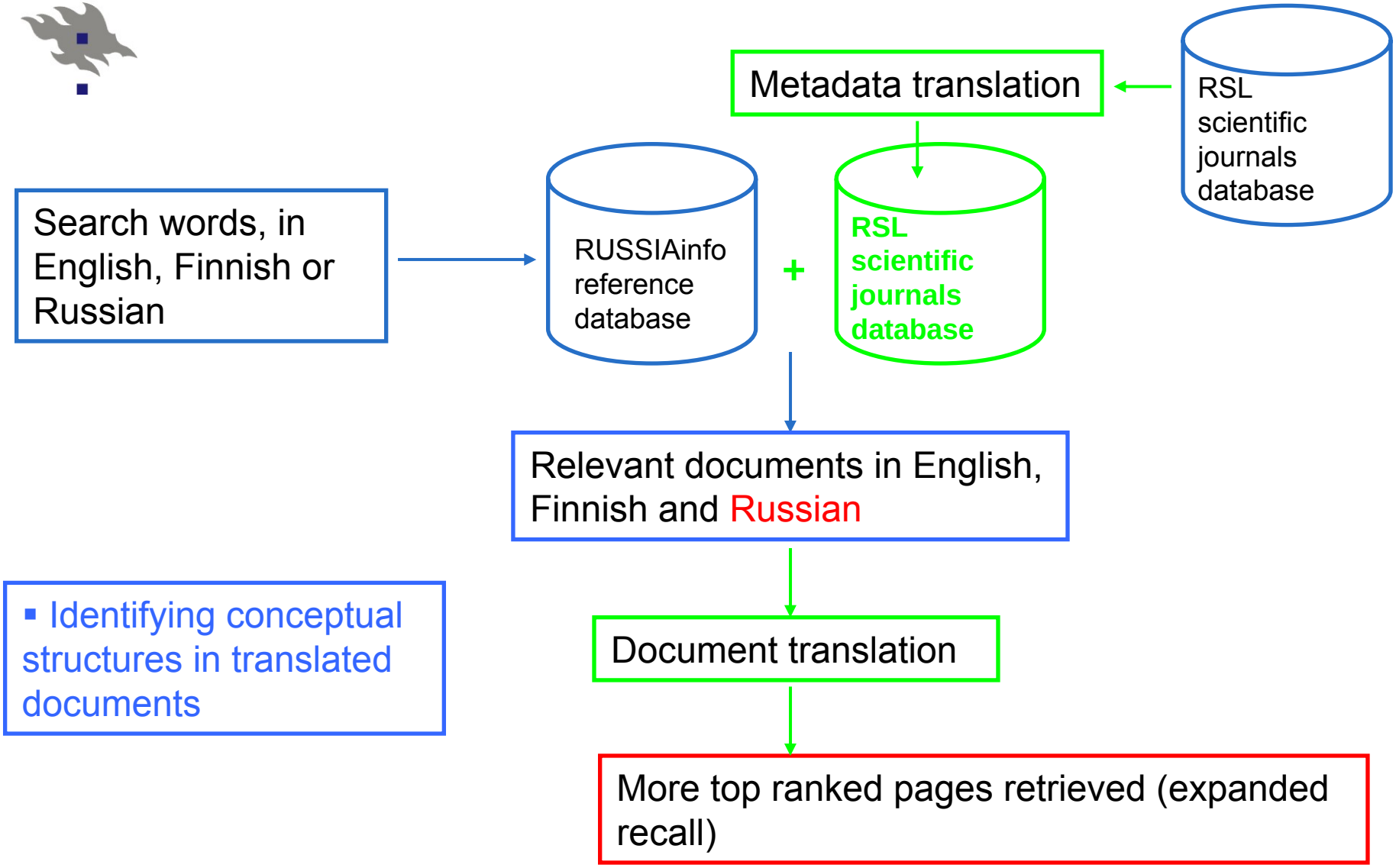
Artificial intelligence, machine translation, mathematical linguistics, kognitivnaja linguistics, kvantitativnaja linguistics, nejrolingvistika, psiholingvistika, formal models of language

- Abstract:

Журнал ставит своей целью объединить усилия лингвистов и специалистов-компьютерщиков в изучении языка с применением точных методов. Тематика журнала охватывает широкий круг проблем прикладной лингвистики. Статьи публикуются на русском и английском языках.

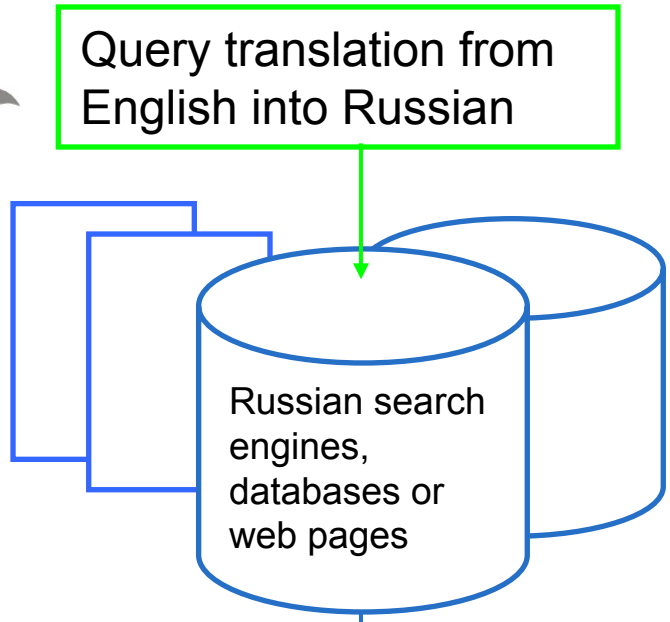
The magazine sets as the purpose to unite effort of linguists and experts-programmers in studying language with application of exact methods. The subjects of magazine covers the broad audience of problems of applied linguistics. Clauses are published in Russian and English languages.

CLIR, stage I and II



- Identifying conceptual structures in translated documents

CLIR, stage III



- Concept-based structured queries vs. unstructured queries
- Single words (NL/S) vs. phrase marked (NL/WP) queries

More top ranked pages retrieved + better quality of result ranking

CCIR

- Languages are conceptually incompatible
- from CLIR to CCIR (Cross Cultural Information Retrieval)
  - conceptualizing Universes for different cultures

cognitive science

philosophy

behavioral science

computer science **linguistics** **information science**

cultural studies

sociology

cultural anthropology

history

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

THANK YOU!

www.russiainfo.org
tanja.pursiainen@helsinki.fi

“Digital Semantic Content Across Cultures” –symposium. Paris. Louvre. 4-5 May 2006

University of Helsinki, Aleksanteri Institute

RUSSIAINFO

QUALITY RESOURCES ON RUSSIA SUOMEKSI IN ENGLISH

HOME BROWSE SEARCH TRANSLATOR MY PROFILE
SITEMAP ABOUT US HELP FEEDBACK

Browse by topic

- Public Administration
- Politics
- Economy
- Social Services and Health Care
- Education and Research
- Culture
- Civil Society
- Environment

Browse by region

- Russia
- Central Federal District
- Northwestern Federal District
- Southern Federal District
- Volga Federal District
- Urals Federal District
- Siberian Federal District
- Far Eastern Federal District

- Open map in larger view
- Basics on Russia

Search

- Thesauri
- Transliteration

Subject

All

Geographical area

Russia

Find it

- Advanced search
- Help

Latest updates

Tiedostoja
On site references: 955
Last updated on 2006-04-15.

RUSSIAINFO

QUALITY RESOURCES ON RUSSIA SUOMEKSI IN ENGLISH

HOME BROWSE SEARCH TRANSLATOR ★ MY PROFILE
SITEMAP ABOUT US HELP FEEDBACK ★★★

Browse by topic

- Public Administration
- Politics
- Economy
- Social Services and Health Care
- Education and Research

Culture (200)

- Cultural Policy (3)
- Art (8)
- Languages and Literature (29)
- Cultural Heritage (17)
- Folklore (5)
- Philosophy (6)
- Religion (19)
- History (11)
- Media (7)
- Sports (3)
- International Co-operation (24)

- Civil Society
- Environment

Browse by region

- Russia
- Central Federal District (20)
- Northwestern Federal

Culture, Russia (200)

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 | Show all
next »»

Культура-Портал
[Culture Portal]
Extensive information on all aspects of cultural life. Legislation for culture in Russia.
Web-page, Language : Russian, Publisher : Newspaper Kultura
[View full record] [Translate]

Портал "Культура России"
[Culture of Russia Portal]
Comprehensive information on Russian culture: the history and the present. Photographs of objects of culture and artists, as well as articles about them. Catalogue of cultural organisations, news and links.
Web-page, Language : Russian, Publisher : Culture of Russia Portal
[View full record][English language version]

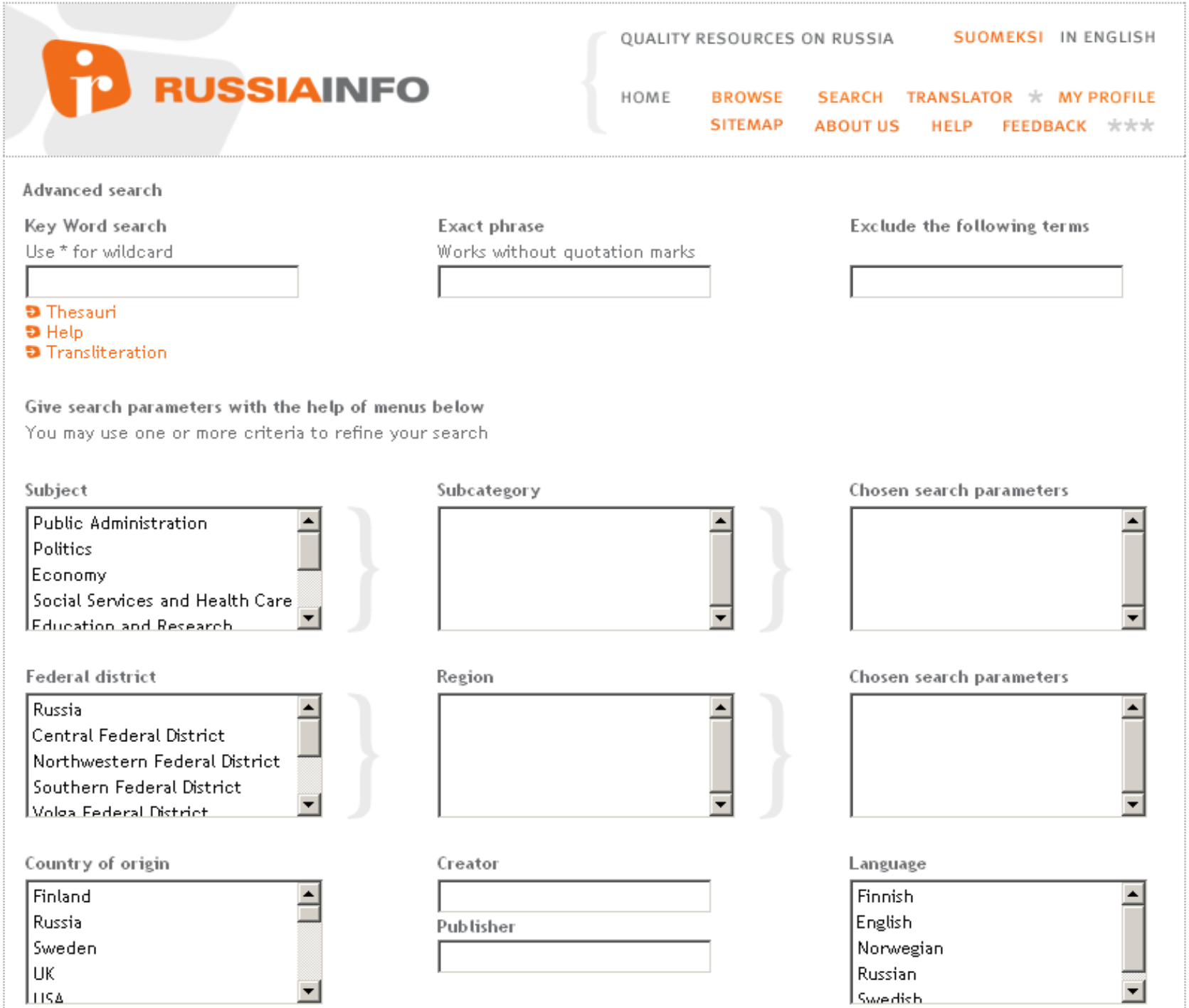
Kulttuuriuutiset. Poimintoja Venäjän lehdistöstä
Verkkolehden tuorein numero. Kaikkien artikkeleiden arkisto.
Web-page, Language : Finnish
[View full record]

Humbul Humanities Hub. Slavonic and East European Studies

Search

Find it

Advanced search


RUSSIAINFO
QUALITY RESOURCES ON RUSSIA
SUOMEKSI
IN ENGLISH
HOME
BROWSE
SEARCH
TRANSLATOR
MY PROFILE
SITEMAP
ABOUT US
HELP
FEEDBACK
Advanced search
Key Word search
Use * for wildcard
Exact phrase
Works without quotation marks
Exclude the following terms
Thesauri
Help
Transliteration
Give search parameters with the help of menus below
You may use one or more criteria to refine your search
Subject
Public Administration
Politics
Economy
Social Services and Health Care
Education and Research
Subcategory
Chosen search parameters
Federal district
Russia
Central Federal District
Northwestern Federal District
Southern Federal District
Volga Federal District
Region
Chosen search parameters
Country of origin
Finland
Russia
Sweden
UK
USA
Creator
Publisher
Language
Finnish
English
Norwegian
Russian
Swedish


RUSSIAINFO
QUALITY RESOURCES ON RUSSIA
SUOMEKSI
IN ENGLISH
HOME
BROWSE
SEARCH
TRANSLATOR
MY PROFILE
SITEMAP
ABOUT US
HELP
FEEDBACK
Enter the Russian language text here (max. 1000 characters) and press `Translate`
TRANSLATION POWERED BY @PROMT
Help
Translate
Direction
Russian to english
Enter the URL of the Russian language web page here and press `Translate`
Web address
Translate
Enter search words in English, press `Search`, and browse Russian Internet.
If you want your search results in English, press `Search and translate`.
Google
Search
Yandex
Search
Search and translate

LENTA·RU
издание Rambler Media Group

Tuesday, 02.05.2006, 14:03:36
The Main thing

 Нефтяники Боливии ждут убытков

 Японцы остались без Coca-Cola

 Создан внеш

The Latest news

02.05 12:10

The Ministry of Foreign Affairs of Iran has declared support from Russia and China

The Iranian Minister for Foreign Affairs Manusher Mottaki has declared, that both Russia, and China have officially informed Teheran that these countries will not support introduction of economic sanctions or carrying out of military operation against Iran in connection with continuation of its nuclear program.

The full text

02.05 12:42

Foreigners are afraid to deal with Bolivia after nationalization of oil

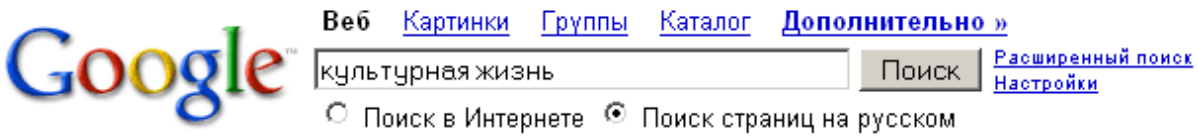
02.05 12:18 Embassy of France continues to altercate with the Russian tour agencies

02.05 13:43 " Red October " has lighted up because of casual handling of fire

02.05 13:07 Moscow militia prepares to zachistkam ethnic groupings

02.05 13:52 " the Nightmare before Christmas " will get the third measurement

02.05 13:52 Municipal Departments of Internal Affairs of Moscow suggests murderers of militiamen to surrender voluntary



Веб Результаты: **1** - **10** из приблизительно **8 010 000** с

Культурная жизнь г.Старая Русса
Ансамбль скрипачей и детская школа искусств, художники, фотографии города.
russa-art.narod.ru/ - 7k - Сохранено в кэше - Похожие страницы

SibMama Форумы о детях и семье :: Просмотр форума - **Культурная ...**
Нет новых сообщений, В чем заключается Ваша **культурная жизнь**? [На страницу На страницу: 1, 2, 3, 4 #], 50, Wild, 1607, Вс Мар 26, 2006 7:47 pm **...**
forum.sibmama.ru/viewforum.php?f=45 - 115k - Сохранено в кэше - Похожие страницы

Культурная жизнь Нижнекамска с Альфией Камиловой | **Культурная** Столица
Культурная жизнь Нижнекамска с Альфией Камиловой; III Нижнекамский фестиваль телерадиокомпаний "Культура в эфире" · Город Нижнекамск · Искусство в городской **...**
www.culturecapital.ru/capitals/nizhnekamsk - 19k - 30 апр 2006 -
Сохранено в кэше - Похожие страницы

Культурная жизнь Чебоксар с Ритой Кирилловой | **Культурная** Столица
Культурная жизнь Чебоксар с Ритой Кирилловой; «День **культурной** столицы Поволжья» 2 марта 2004 года в Чебоксарах · МАРАФОН **КУЛЬТУРНЫХ** СОБЫТИЙ Программы **...**
www.culturecapital.ru/capitals/cheboksary/46 - 19k - Сохранено в кэше - Похожие страницы
[Дополнительные результаты с www.culturecapital.ru]

Культурная жизнь
Культурная жизнь. Достопримечательности; **Культурная жизнь**. Драматические театры · Музыкальные театры · Концертные залы и площадки · Кинотеатры **...**
www.st-petersburg.ru/entertainment/culture/ - 13k - Сохранено в кэше - Похожие страницы

Болгария.**Культурная** жизнгь.
КУЛЬТУРНАЯ ЖИЗНЬ. Болгария славится своими многочисленными деятелями культуры и артистами, известными не только в стране, но и за ее пределами. **...**
www.ctel.msk.ru/btlnf/turism/land/bulgar/bulg_cult.htm - 8k -
Сохранено в кэше - Похожие страницы

Почта

Яndex
Найдётся всё

культурная жизнь

в найденном в регионе: Финляндия

Везде | Новости | Маркет | Адреса | Словари | Блоги | Картинки | Все службы...

Результат поиска: страниц — **1 069 369**, сайтов — не менее **1 891**, в каталоге — **165**
Запросов за месяц: культурная — 22 504, жизнь — 208 164. Купить эти слова.

Хотите найти рядом? Поищите «культурная жизнь» на сайтах Финляндии.

1. Культура-Портал - актуальная информация о значительных событиях в **культурной** ...
   Культура-Портал: электронная версия газеты Культура, архив, форум, культура российских регионов, законодательство в сфере культуры, рынок антиквариата, актуальная информация о музыке, кино, книгоиздании
   www.kultura-portal.ru (24 КБ)
   Найденные слова · Еще с сайта (759) · Рубрика: Культура

2. Мир досуга (подробная информация о **культурной жизни** Москвы - театральные ...
   www.mirdosuga.ru (1 КБ) · 19.04.2006
   Найденные слова · Еще с сайта (486)

3. WeekEnd.Ru
   текст ссылок: Weekend ежедневная информация о **культурной жизни** Москвы... **Культурная жизнь** Москвы...
   www.weekend.ru (48 КБ) · 27.04.2006 — найден по ссылке
   Еще с сайта (4) · Рубрика: Культура

4. Болгария.
   КУЛЬТУРНАЯ ЖИЗНЬ.
   www.telegraf.ru/btInf/turism/land/bulgar/ (6 КБ) · 22.01.2001
   Найденные слова · Еще с сайта (4)

5. Чешская культура - новости **культурной жизни** Праги и Чехии | Пражские музеи ...
   " Серия "**Культурная жизнь** Праги".
   в **культурной жизни** Праги (апрель 2002)
   kultura.prag.ru (28 КБ) · 01.10.2002
   Найденные слова · Еще с сайта (33) · Рубрика: Культура

6. НИС-РЕВЮ ::
   текст ссылок: Нис ревю о **культурной жизни** Нижнего Новгорода... Новости **культурной жизни** Нижнего Новгорода Афиша...
   www.nisrevue.ru (21 КБ) — найден по ссылке
   Еще с сайта (100) · Рубрика: Культура

Mail

Яndex
There will be all

культурная жизнь

In found In region: Finland

Everywhere | News | The Market | Addresses | Dictionaries | Blogi ✔ | Pictures | All services ..

Result of search: pages - **1 069 369**, sites - not less **than 1 891**, in the catalogue - **165**
Inquiries for a month: cultural - 22 504, a life - 208 164. To buy these words.

Wish to find beside? Look " a cultural life " on sites of Finland.

1. Culture-portal - the actual information on significant events in **cultural**...
   Culture-portal: the electronic version of the newspaper Culture, archive, a forum, culture of the Russian regions, the legislation in of culture, the market of antiques, the actual information on music, cinema, book publishing
   www.kultura-portal.ru (24 KB)
   The Found words | From a site (759) | The Heading: Culture

2. The World of leisure (the detailed information on **a cultural life** of Moscow - theatrical...
   www.mirdosuga.ru (1 KB) 19.04.2006
   The Found words | From a site (486)

3. WeekEnd. Ru
   The text of references: Weekend the daily information on **a cultural life** of Moscow... **The Cultural life** of Moscow...
   www.weekend.ru (48 KB) 27.04.2006 - it is found under the reference
   From a site (4) | The Heading: Culture

4. Bulgaria.
   THE CULTURAL LIFE.
   www.telegraf.ru/btInf/turism/land/bulgar/(6 KB) 22.01.2001
   The Found words | From a site (4)

5. The Czech culture - news **of a cultural life** of Prague and Czechia | the Prague museums...
   "Series" **the Cultural life** of Prague ".
   In **a cultural life** of Prague (April 2002)
   kultura.prag.ru (28 KB) 01.10.2002
   The Found words | From a site (33) | The Heading: Culture

Enter the Russian language text here (max. 1000 characters) and press `Translate`

cultural heritage

Translate Direction Russian to english

культурное наследие

Результаты поиска для: 'культурное наследие'

Совпадения: Все слова | Формат вывода результатов: Подробный

Поиск: культурное наследие | Искать

- (•) по всему серверу
- ( ) исключить газету "Карелия" из поиска
- ( ) в газете "Карелия"
- ( ) в архиве фотографий
- ( ) в новостях все месяцы | все года

* - поиск по месяцам доступен с января 2004 года.

Расширенный поиск

Документы 1 - 10 из 87. У лучших совпадений больше ☆.

II. Анализ состояния приграничного сотрудничества (сильные, слабые стороны, возможности и угрозы)☆☆ ☆☆

. . . по их подготовке, научных центров. * Многолетние тесные связи между городами и районами Республики Карелия и городами и коммунами Финляндии. * Общее **культурное наследие**. Расположение на границе Российской Федерации и Европейского Союза Географические и геополитические факторы имеют важнейшее влияние . . .

http://www.gov.karelia.ru:8083/gov/Power/Ministry/Relations/Boundary/02a1.html 30.04.2006, 24000 bytes

Международный семинар "Культурный туризм и устойчивое развитие"☆

. . . в начале, и в середине, и уже в конце XX века уникальные по красоте и исторической значимости памятники архитектуры, природные памятники, выдающееся **культурное наследие**, превосходные туристические маршруты, места лечения и отдыха в Карелии - все это было практически недоступно для организации полноценного . . .

http://www.gov.karelia.ru:8083/gov/Power/Ministry/Culture/Tourism/seminar_tur3.html 30.04.2006, 30421 bytes

Организации и учреждения, получившие в 2003 году целевое финансирование или добровольные безвозмездные пожертвования от Совета Министров Северных Стран☆

. . . климата" с 29 мая по 4 июня 2003 г. 22 145 Петрозаводский государственный университет, г.Петрозаводск, республика Карелия 5. Проект "**Культурное** и природное **наследие** Карельского Беломорья (перспективы развития туризма и образования)", 104266 Карельский научный центр РАН. Республика Карелия . . .

http://www.gov.karelia.ru:8083/gov/Different/Region/Document/sovet_ministr03.html 30.04.2006, 17880 bytes

Results of search for: ' a cultural heritage '

Concurrences: All words | The Format of a conclusion of results: Detailed

Search: Cultural heritage | To search

- (•) On all server
- ( ) To exclude the newspaper "Kareliya" from search
- ( ) In the newspaper "Kareliya"
- ( ) In archive of photos
- ( ) In news | All the months long | All of year

* - search on months is accessible since January, 2004.

The Expanded search

Documents 1 - 10 from 87. The best concurrences have more ☆.

II. The analysis of a condition of frontier cooperation (strong, weaknesses, opportunities and threats) ☆☆☆☆

. . . On their preparation, centres of science. * long-term close communications between cities and areas of Republic Kareliya and cities and communes of Finland. * the general **cultural heritage**. An arrangement on border of the Russian Federation and the European Union Geographical and geopolitic factors have the major influence . . .
http: // www.gov.karelia.ru:8083/gov/Power/Ministry/Relations/Boundary/02a1.html 30.04.2006, 24000 bytes

The International seminar " Cultural tourism and steady development " ☆

. . . In the beginning, and in the middle, and already in the end of XX century unique on beauty and monuments of architecture, the natural monuments, an outstanding **cultural heritage**, excellent tourist routes, places of treatment and rest in Kareliya - all this was practically inaccessible to the historical importance to the organization high-grade . . .
http: // www.gov.karelia.ru:8083/gov/Power/Ministry/Culture/Tourism/seminar_tur3.html 30.04.2006, 30421 bytes

The Organizations and the establishments which have received in 2003 target financing or voluntary gratuitous donations from Ministerial council of Northern Countries ☆

. . . A climate " from May, 29th till June, 4th, 2003 22 145 Petrozavodsk state university, Petrozavodsk, republic Kareliya 5. The project " **the Cultural** and natural **heritage** Karelian Belomorja (prospects of development of tourism and formation) ", 104266 Karelian centre of science of the Russian Academy of Science. Republic Kareliya . . .
http: // www.gov.karelia.ru:8083/gov/Different/Region/Document/sovet_ministr03.html 30.04.2006, 17880 bytes

The Epos "Kalevala" ☆

. . . Languages, and among still more many such which in a science are called unwritten or mladopismennymi as people and languages. For such people their oral **cultural heritage** represents those greater and rather actual value as the developed literary-book traditions on the native language at them still are not present . . .
http: // www.gov.karelia.ru:8083/gov/Different/Kalevala/karhy1.shtml 30.04.2006, 28881 bytes

Kareliya N 64 (on June, 27th 2002): the SCIENCE: Belomorskaja Kareliya eyes of the scientific two countries ☆