# Microsoft® Research FacultySummit 2011

ARE ADDRESS

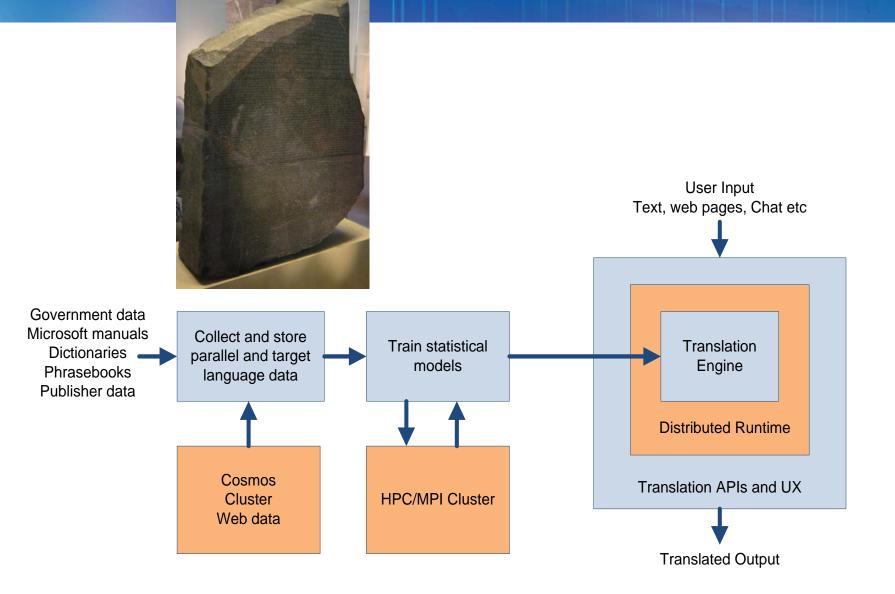
Cartagena, Colombia | May 18-20 | In partnership with COLCIENCIAS Sharing Research Worldwide Advances in Automatic Translation

> Chris Wendt Group Program Manager, Microsoft Research

## Agenda

- Microsoft Translator At A Glance
- Translation Advancing Science
- Quality Considerations
- Collaborative Translations
- Why you should publish in Czech

# Statistical MT - The Simple View

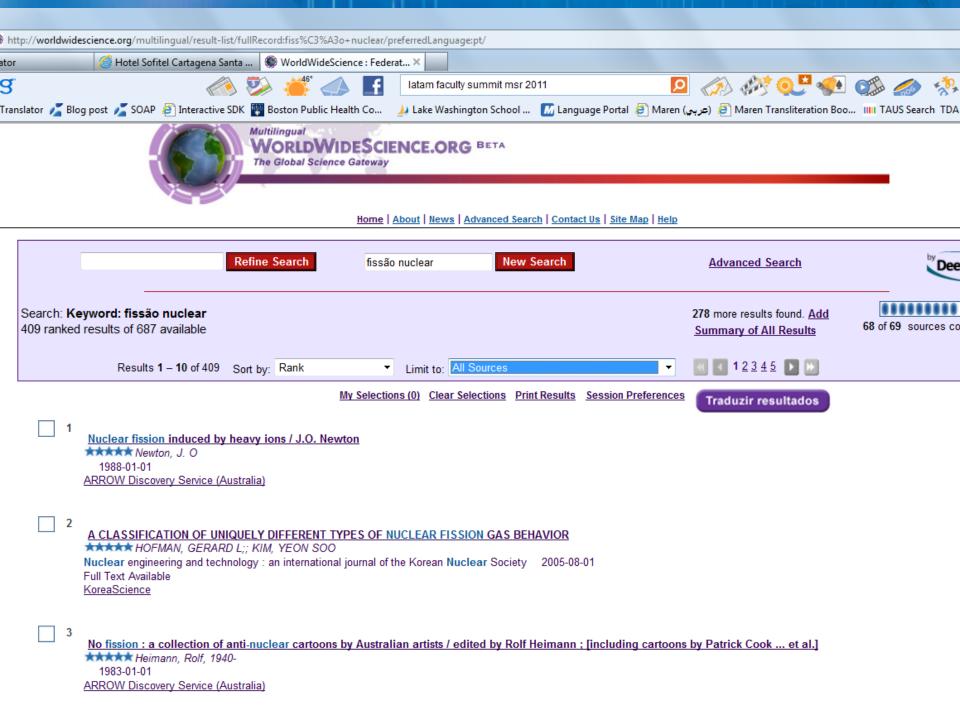


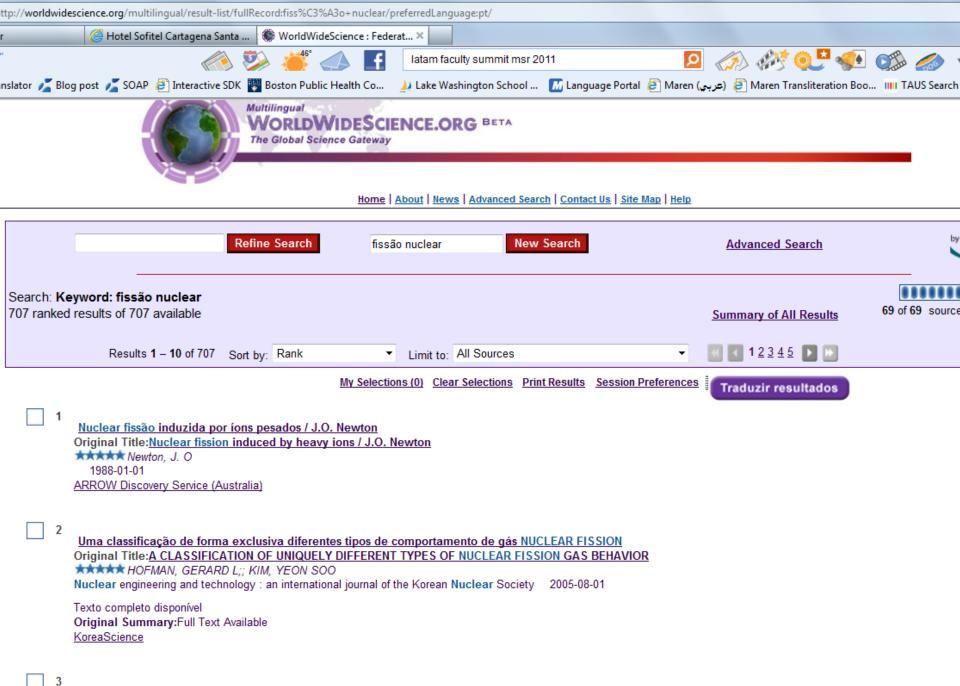


**WorldWideScience.org** is a global science gateway—accelerating scientific discovery and progress through a multilateral partnership to enable federated searching of national and international scientific databases and portals.

Last Modified: 08/19/2010







Nenhum fissão : uma coleção de anti -nuclear desenhos animados por artistas australianas / editado por Rolf Heimann; [incluindo desenhos animados por Patrick Cook... et al.]

20	₩.the hetter	/www.microsofttranslat	or com/D\/ scow?fr	om-on 8ito-ntl	21-2- http://21.0/	2E%2Etroup play	aou 200/25	uerk% 2E101010220	V2Ea field0%26a by	no0%2Dall%26a to	rm/10/2Dmu	
🚰 Bing Tr							-	e.nla.gov.au/work/		peo /os Dan /ozoq-te	1110 /650110	Clear /ozdAND /ozdrissic
-	ing	- Tote	i sonici curugena			f	an an aon	innargo nad, monte	19191		s er	? o.¤ 🐢 🤇
	-	tor 🔏 Blog post 🔏 S	OAP 🗿 Interacti	ve SDK 🕅 Bos	ton Public Hea		ake Washir	aton School 🚺	M Language Portal		1. a	Transliteration Boo
		slator Privacy   Legal			6.55		and trasmi	gton sensor in	congaageronar		P March	Sabiasa
Translator			1-1-V -	XI	X	Micr	rosoft® is no	ot responsible for th	e content below			
Translate U	RL http://tr	ove.nla.gov.au/work/1919	1822?q-fie Engli	sh	*	Portugi	uese	▼ → Vie	ews 🔳 🖼 🗖			
Translating	55%		Mous	e over text to se	e original							
NAT	IONAL	LIBRARY OF	AUSTRALIA	1								
Au	stralia	•								Home Sob	re Site de	notícias Estatísticas
I	<b>UV</b>	e										É grátis e
Todos os	Livros	aith of information Imagens, fotografias,	Revistas, artigo	lornai	s digitalizados	Música, som	Mapas	Diários, cartas,	Sites arquivados	Pessoas	Listas	
10005 05	LIVIUS	objetos	e conjuntos de o	2 XX 1 X 2	173	e vídeo	Mapas	Archives	(1996 – agora)	e organizações	Listas	
Refin	e suas v	versões:	Fissão nu	clear ind	uzida p	or íons p	esado	s / J.O. Ne	wton			Ainda não Taga
<ul> <li>Forma Livro</li> </ul>				Autor	Newton, C	Di						Adicionar uma marc
	istrado (1)			Temas		clear.; Íons pes	ados.					
				Indicador	62	/e.NLA.gov.au/v	work/19191	822				
<ul> <li>Idioma Inglês</li> </ul>				ID do trabalho Edições		do presente tra	abalho é li	stado abaixo				Ainda não estã lista 0
Décado	13					an 1978 series of the fact						
, Desta			Uma versã	o está dis	sponível	a partir de	0					login para criar suas pro
			On-line (0)	leu bibliotec	as (0) Tod	las as bibliote	ecas (1)	Compre uma o	cópia			
			Todos (1) A									Envie para:
												- 99 E F 🕸
			Esta localiz Biblioteca r	Construction of the second		erto ao público]						
			Call Number			1						
									🐵 Link para est	a página.		

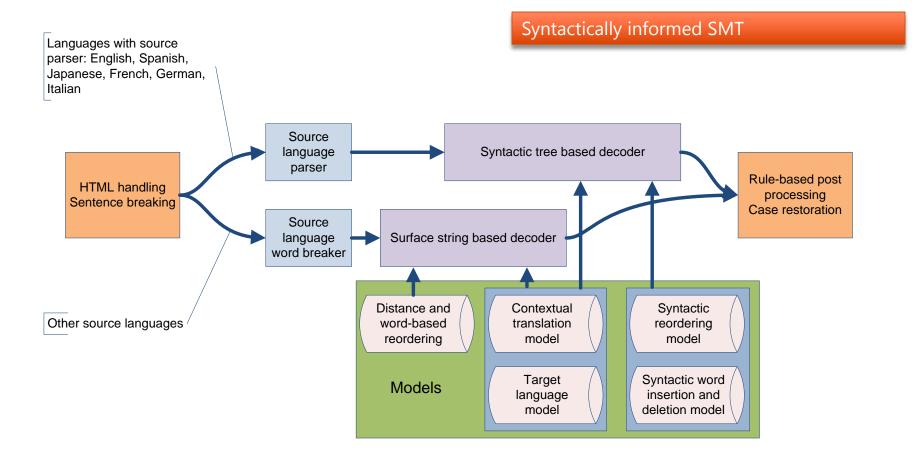


			Refine	Search	fissão	nuclear	New Search	Advar	nced Search	
- 1		eyword: fissão nuclear results of 707 available						<u>Summar</u>	<u>y of All Results</u>	69 of 69
		Results 1 – 10 of 707	Sort by:	Rank	-	Limit to:		<b>-</b>	2345 💽 ≥	
	<ul><li>□ 1</li><li>□ 2</li></ul>	Nuclear fissão induzida po Original Title: <u>Nuclear fissio</u> Newton, J. O 1988-01-01 ARROW Discovery Service (A Uma classificação de form Original Title: <u>A CLASSIFIC</u> HOFMAN, GERAR Nuclear engineering and tech Texto completo disponível Original Summary:Full Text KoreaScience	<u>a exclusi</u> Australia) <u>a exclusi</u> ATION OF 2D L,; KIM, anology : a	sados / J.O. New 1 by heavy ions 1 by heavy ions	/ J.O. Ne pos de co FERENT	ewton omportam TYPES Of	All Sources ARROW Discovery Service (Australia) AsiaJOL Canada Institute for Scientific and Technical II CERN Document Server Electronic Table of Contents (ETOC) (United I German National Library of Science and Tech Indian Academy of Sciences J-STAGE (Japan) (English) J-STORE (Japan) KoreaScience Scientific Electronic Library Online (English) Scientific Electronic Library Online (Portugues Scientific Electronic Library Online (Spanish) State Public Technical Library of Russia (Rus The Russian Union Catalog of Scientific Litera UK PubMed Central (United Kingdom) VINITI Projects Database (Russian)	Kingdom) inology (GetInfo) se) ssian)	resultados	

Nenhum fissão : uma coleção de anti -nuclear desenhos animados por artistas australianas / editado por Rolf Heimann; [incluindo desenhos animados por P Cook... et al.]

3

# Microsoft's Statistical MT Engine



## Haitian Creole

(or How to Build and Ship an MT Engine from Scratch in 4 days, 17 hours, & 30 minutes)

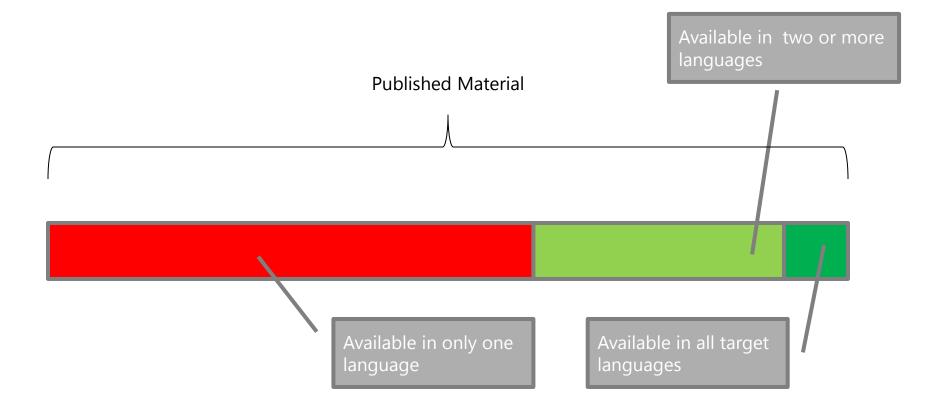
Data -

- Resource Poor Language no corpora, no significant Web presence, not a lot of easily discoverable data
- Finding much of the data manually (lots of volunteer help)
- NLP community started sharing data
  - Carnegie Mellon University (CMU)
  - CrisisCommons
  - 4636 SMS/ushahidi
- Manually translating a small portion
- Targeting content relevant to relief effort
- Giving back to the community through data donations

# Why Microsoft uses Automatic Translation

No. 2014년 1997년 19 1997년 19 1997년 19 1997년 19 1997년 199 1997년 199 1997년 199 1997년 199 1997년 199 1997년 199 1997년 199 1997년 199 1997년 199 1997년 199 1997년 199 1997년 199 1997년 1997년 1997년 1997년 1997년 1997년 1997년 1997년 1997년

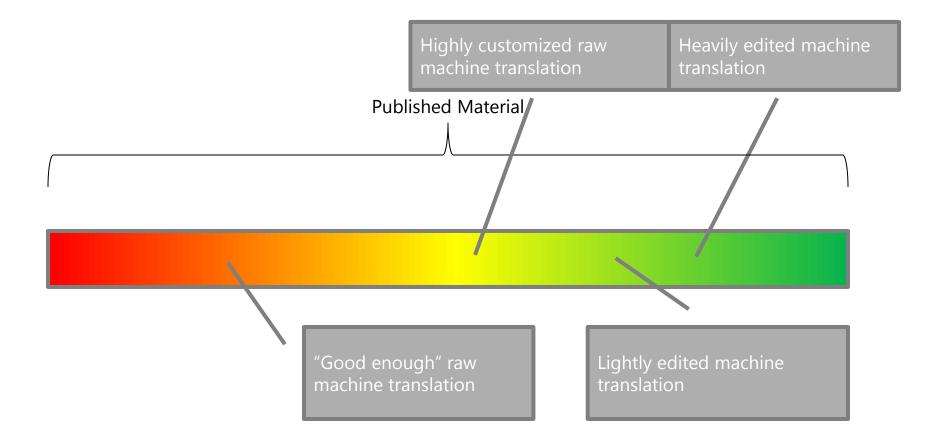
## Situation Today



Fully reviewed high quality human translation

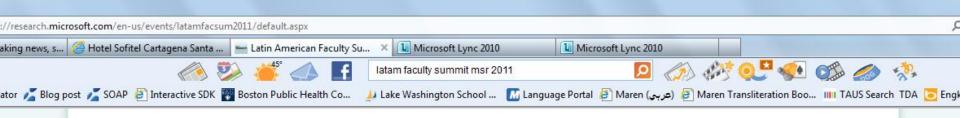
No translation

### Tomorrow



Fully reviewed high quality human translation

No translation



Home Agenda

Logistics DemoFest

### Microsoft Research FacultySummit 2011

#### Cartagena, Colombia | May 18-20 | In partnership with COLCIENCIAS

Microsoft Research will host its seventh annual Latin American Faculty Summit in Cartagena, Colombia, May 18 to 20, 2011. This event is organized in partnership with Colciencias. The theme of the 2011 summit is *Advancing Computing; Advancing Science*.



#### Microsoft Biology Workshop

Join us for the two-day workshop on May 16–17, 2011, in Cartagena, Colombia. Learn more...



#### Cartagena, Colombia Advancing Computing; Advancing Science

Speakers

Computing stands at a crossroads. The long-discussed movement from singleprocessor to multi-processor computers—whether on a single die or via a cloud—has become a reality that drives the need and opportunity for new paradigms in software engineering and development. The means to command and access computing resources—whether through voice recognition, multi-touch, or gesture—are becoming increasingly technologically and economically feasible. New ways to search, visualize, and organize large data enables new scientific understanding in fields like environmental research, astronomy, and digital humanities. Advances in computing approaches and algorithms are revealing new possibilities in the use of genomics to support innovations in biology, medicine, agriculture, and bio-energy.

Microsoft Research has an ongoing program of collaborations that focus on these research challenges. Attendees of the 2011 Latin American Summit will have the opportunity to hear about the latest research resulting from our collaborations.

#### **Related Links**

- Latin American Faculty Summit 2010
- Microsoft Research Faculty Summits
- Collaborations in Latin America

#### Event Sponsors



Departamento Administrativo de Ciencia, Tecnología e Innovación Colciencias

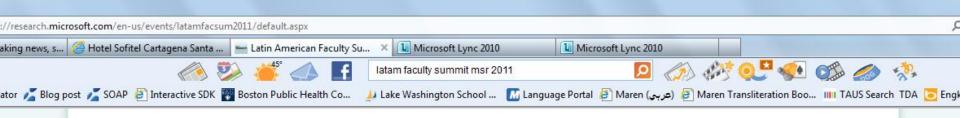
Libertod y Orden República de Colombia

Microsoft Research Connections

# Translator | Widget Beta

The Microsoft Translator web page widget allows you to bring real-time, in-place translation to your web site. Users can see your pages in their own language, without having to go to a separate translation web site, and share your page with friends in multiple languages. You can learn more about how to use the widget, get help or interact with other website owners on the Microsoft Translator forums.

Site info			
* Site address http://www.example.com/ Input the site address where you would be displaying the widget. Learn more here.	* Site language Englis		
Options			
Translation: If you choose to display the widget then any visitor to your site can invoke translation by clicking it	Translation settings		🖱 Auto
(manual). In addition you may choose to have the translation toolbar appear to visitors whose browser language is different than your site language and offer to translate the page (notify) or to do so automatically (auto), in which case displaying the widget is optional.	© Manuar		C Auto
☑ Display widget control			
Color:	Preview:		
	Translate this page		
Widget width: 200	Spanish [		
Collaborative Translations Technology Pievlew			
Enable Community Translations Turning on Community Translations helps you improve the translations provided from your site by editing them selecting from visitor provided edits. Learn more here .		fions 🐱 rtranslation from the options below:	
* Bing Developer AppID: 🕐	4. 📃 💵		Edt
Don't have it? Go to the Bing Developers Portal to get one.			
Generate widget			
* 🔲 I agree to the Microsoft Translator Terms of Use.			



Home Agenda

Logistics DemoFest

### Microsoft Research FacultySummit 2011

#### Cartagena, Colombia | May 18-20 | In partnership with COLCIENCIAS

Microsoft Research will host its seventh annual Latin American Faculty Summit in Cartagena, Colombia, May 18 to 20, 2011. This event is organized in partnership with Colciencias. The theme of the 2011 summit is *Advancing Computing; Advancing Science*.



#### Microsoft Biology Workshop

Join us for the two-day workshop on May 16–17, 2011, in Cartagena, Colombia. Learn more...



#### Cartagena, Colombia Advancing Computing; Advancing Science

Speakers

Computing stands at a crossroads. The long-discussed movement from singleprocessor to multi-processor computers—whether on a single die or via a cloud—has become a reality that drives the need and opportunity for new paradigms in software engineering and development. The means to command and access computing resources—whether through voice recognition, multi-touch, or gesture—are becoming increasingly technologically and economically feasible. New ways to search, visualize, and organize large data enables new scientific understanding in fields like environmental research, astronomy, and digital humanities. Advances in computing approaches and algorithms are revealing new possibilities in the use of genomics to support innovations in biology, medicine, agriculture, and bio-energy.

Microsoft Research has an ongoing program of collaborations that focus on these research challenges. Attendees of the 2011 Latin American Summit will have the opportunity to hear about the latest research resulting from our collaborations.

#### **Related Links**

- Latin American Faculty Summit 2010
- Microsoft Research Faculty Summits
- Collaborations in Latin America

#### Event Sponsors



Departamento Administrativo de Ciencia, Tecnología e Innovación Colciencias

Libertod y Orden República de Colombia

Microsoft Research Connections

research.microsoft.com/en-US/events/latamfacsum2011/default.aspx

Agenda

ng news, s... 🖉 Hotel Sofitel Cartagena Santa ... 😑 Latin American Faculty Su... × 💷 Microsoft Lync 2010 🗈 Microsoft Lync 2010 مرابع المعالي الم معالي المعالي المع معالي المعالي الم

Inicio

Conferenciantes Logística

stica DemoFest

### Microsoft Research FacultySummit 2011

#### Cartagena, Colombia | May 18-20 | In partnership with COLCIENCIAS

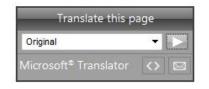
Microsoft Research acogerá su séptima cumbre anual de Facultad de América Latina en en Cartagena, Colombia, 18 a 20 de mayo de 2011. Este evento es organizado en colaboración con Colciencias. El tema de la Cumbre de 2011 es *avance de informática; Avance de la ciencia*.



#### Avanzando la computación; Avanzando la ciencia

Las ciencias de la computación están en una encrucijada. La transición de las computadoras desde un único procesador a multiprocesador, tanto en un único chip o a través de un cloud, se ha convertido en una realidad que impulsa la necesidad y oportunidad de los nuevos paradigmas de desarrollo e ingeniería de software. Los mecanismos de acceso a los recursos informáticos: bien a través del reconocimiento de voz, pantallas táctiles o gestos — se están volviendo cada vez más

tecnológicamente y económicamente viables. Nuevas formas para buscar, visualizar y organizar grandes cantidades de datos permiten avances científicos en campos como la investigación ambiental, la astronomía y las humanidades digitales. Los avances en los enfoques y los algoritmos de la computación están revelando nuevas posibilidades en el uso de la genómica para apoyar las innovaciones en la biología, la medicina, la agricultura y la bioenergía.



#### Taller de biología de Microsoft

Únase a nosotros para el taller de dos días en mayo de 1804, 2011, en Cartagena, Colombia. Más información...

#### Enlaces relacionados

- Cumbre de la Facultad
   Latinoamericana de 2010
- Cumbres de Facultad de investigación de Microsoft
- Colaboraciones en América Latina

#### Patrocinadores del evento



Departamento Administrativo de Ciencia, Tecnología e Innovación Colciencias

Libertud y Orden República de Colombia

 Conexiones de investigación de Microsoft

/research. <b>micr</b>	osoft.com/en-US/events/latam	facsum2011/default.aspx				
king news, s	🧉 Hotel Sofitel Cartagena Sar	nta 🔚 Latin American Faculty	Su 🗙 1 Microsoft Lync 2010	🚺 Microsoft Lync 2010		
		) Ӱ 🍊 🚮 🖪	latam faculty summit msr 2011	🖸 🛷 😡	💆 🐠 🕺	****
tor 🔏 Blog po	ost 🔏 SOAP 🧃 Interactive S	DK 🗱 Boston Public Health Co	🍌 Lake Washington School 🛛 🔝 La	anguage Portal 🗿 Maren (عربی) 🗧 Maren Translit	teration Boo IIIII TAUS Searc	h TDA 🔁

Inicio Agenda Conferenciantes

informática; Avance de la ciencia.

Select a better translation:

2

3

Legal

Avanzando la

computación;

Promoción de

Avanzando la

ciencia

computación; avanzando

informática:

es Logística De

#### DemoFest

### Microsoft Research FacultySummit 2011

Cartagena, Colombia | May 18-20 | In partnership with COLCIENCIAS

colaboración con Colciencias. El tema de la Cumbre de 2011 es avance de

/ Edit

◄ Flag

/ Edit

🝽 Flag

/ Edit

◄ Flag

#### Taller de biología de Microsoft

Translate this page

Únase a nosotros para el taller de dos días en mayo de 1804, 2011, en Cartagena, Colombia. Más información...

### Original Advancing Computing; More Translations \*

Microsoft Research acogerá su séptima cumbre anual de Facultad de América Latina en

en Cartagena, Colombia, 18 a 20 de mayo de 2011. Este evento es organizado en

#### do la ciencia

ncrucijada. La transición de las tiprocesador, tanto en un único chip o realidad que impulsa la necesidad y rollo e ingeniería de software. Los cos: bien a través del reconocimiento volviendo cada vez más

uevas formas para buscar, visualizar y n avances científicos en campos como numanidades digitales. Los avances en n están revelando nuevas posibilidades

en el uso de la genómica para apoyar las innovaciones en la biología, la medicina, la agricultura y la bioenergía.

Guest | Sign in

#### Enlaces relacionados

Original

- Cumbre de la Facultad
   Latinoamericana de 2010
- Cumbres de Facultad de investigación de Microsoft
- Colaboraciones en América Latina

#### Patrocinadores del evento



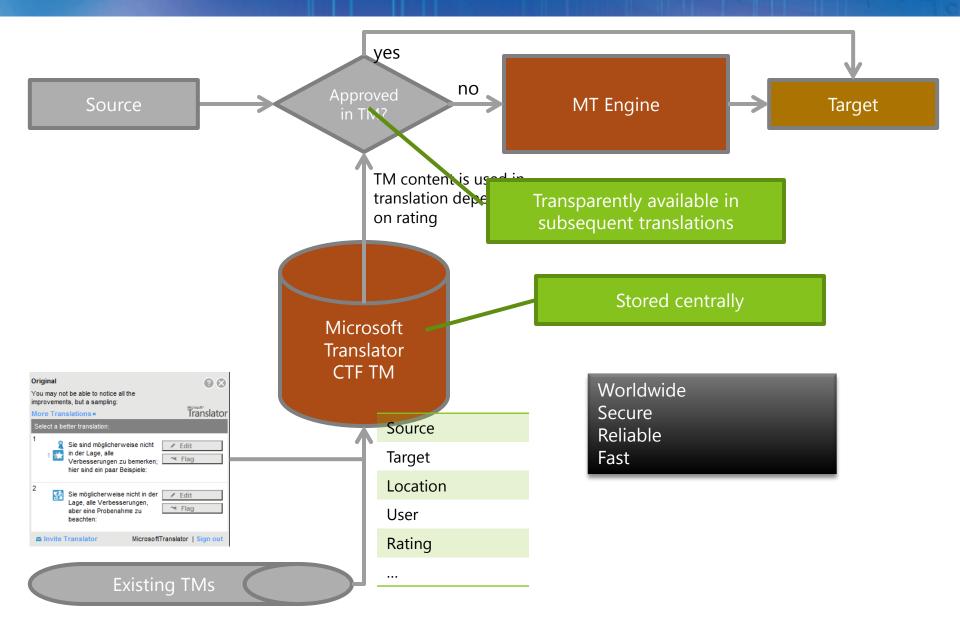
Departamento Administrativo de Ciencia, Tecnología e Innovación Colciencias

Libertud y Orden República de Colombia

 Conexiones de investigación de Microsoft

Microsoft Research tiene un programa continuo de colaboraciones que se centran en

### **Collaborative Translations Framework (CTF)**



# CTF in the API

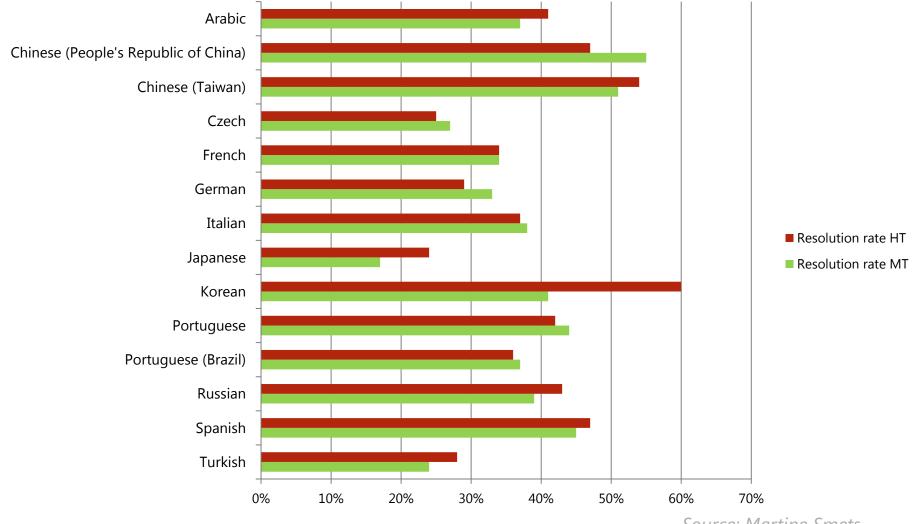
- AddTranslation() adds a translation pair (source, target) with metadata
- AddTranslation() adds a rating, which reflects the authority of the submitter
- All votes and approvals are private to this appID
- Any edit can show up somewhere else as a suggestion in GetTranslations(), but won't have approved status
- Every action is a new record
- GetTranslations() provides a list of alternative translations, at least one.

Microsoft Sup	port	Search Microsoft Support	<mark>9</mark> bing
Support Home Solution C	enters Advanced Search Sl	hop	
		🖂 🚔 🌌 📑 📴 ≫ [+]	Další zdroje
ID článku: 294893 - Poslední akt	ualizace: 3. prosince 2007 - Revize: 5.5		Další stránky podpory
		dálostí služby Directory a události	Komunita
na řadiče domény s	mimo doména systému V	Windows XP	Získat pomoc teď
<ul> <li>UPOZORNĚNÍ: TENTO ČLÁNE</li> <li>Produkty, které se vztahují k tom</li> <li>Systémový tip</li> <li>Tato verze článku se vztahu,</li> </ul>	uto článku. je k jiné verzi operačního systému Windo í být relevantní. Navštivte Centrum řešen	× wws, než právě používáte. Obsah v tomto tohoto ú pro Windows 7	Free antivirus software for your small business Messett Security Essentials Download Now 🕑
🖃 Souhrn			Angličtina
	Popis události ID ( <i>number</i> ) ve zdroji (r	y událostí, může se zobrazit následující zpráva: name) nelze nalézt. Místní počítač možná nemá, pory knihovny DLL zpráv pro zobrazení zpráv ze	Související střediska odborné pomoci • Windows XP • Windows Server 2003 • Windows Server

# Knowledge base feedback

Provide feedback on this information	
Did this information solve your problem?	
O Yes	
O No	
🔘 I don't know	
Was this information relevant?	
O Yes	
○ No	
What can we do to improve this information?	
	~
	~
To protect your privacy, do not include contact information in your feedback.	
Submit	

### Knowledge Base Resolve Rate



Source: Martine Smets, Microsoft Customer Support

# Why you should publish in Czech

- Example: CSS Knowledge Base Czech
- Data from Kai Gehrlach, Martine Smets, and Chris Moore

### Effect of publishing translated content

### Machine Translating the Czech Knowledge Base

### October 2009

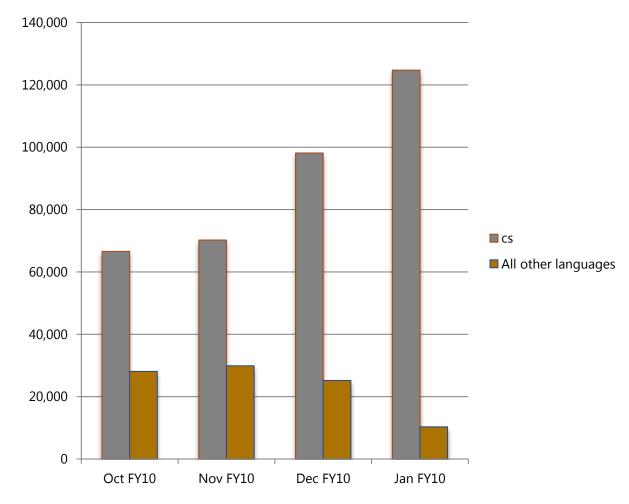
- 2.5% of content of the English KB is human translated to Czech, ranked by page view.
- The top 2.5% cover an estimated 50% of the page views.
- The remaining content is **untranslated**.

### January 2010

- 2.5% of content of the English KB is human translated to Czech, ranked by page view.
- The top 2.5% cover an estimated 50% of the page views.
- The remaining content is **machine translated**, starting December 5 and completed over the next 10 days.

## Referrals from the Czech Republic

- Referrals to the CSS KB site from the top 2 search engines in the Czech Republic (google.cz and seznam.cz
- to the Czech KB (blue)
- to the KB in other languages (green)



# **Microsoft**<sup>®</sup>

© 2010 Microsoft Corporation. All rights reserved. Microsoft, Windows, Windows Vista and other product names are or may be registered trademarks and/or trademarks in the U.S. and/or other countries. The information herein is for informational purposes only and represents the current view of Microsoft Corporation as of the date of this presentation. Because Microsoft must respond to changing market conditions, it should not be interpreted to be a commitment on the part of Microsoft, and Microsoft cannot guarantee the accuracy of any information provided after the date of this presentation.

MICROSOFT MAKES NO WARRANTIES, EXPRESS, IMPLIED OR STATUTORY, AS TO THE INFORMATION IN THIS PRESENTATION.

### References

- API documentation: http://api.microsofttranslator.com
- Forums (Linked from http://api.microsofttranslator.com)
- Widget Adoption Portal: http://www.microsofttranslator.com/widget
- mtcont@microsoft.com
- Chris Quirk, Arul Menezes, and Colin Cherry, Dependency Treelet Translation: Syntactically Informed Phrasal SMT, in Proceedings of ACL, Association for Computational Linguistics, June 2005
- TAUS Data Association: www.tausdata.org