

pangea.com.MT

The Statistical MT Customization Company

---

# PangeaMT – putting open standards to work... well

Manuel Herranz – PangeaMT - Pangeanic

[www.pangea.com.mt](http://www.pangea.com.mt)

---

# Unmanageable amounts of data? *The data deluge*

- As of May 2009: 487 Billion gigabytes or  
 $1,000,000,000 * 487,000,000,000 = 4,87 \times 10^{20}$
- Estimates
  - Up 50% a year (Oracle)
  - Doubles every 11 hours (IBM)
- Language translation as a job becoming unmanageable. Increasing demands, increasing volumes, shorter deadlines. Human production is not sufficient.

# Short history

- Pangeanic: LSP. Major clients in Asia, European localization, increasing number of languages and volumes
- Need to produce faster, cheaper, quality
  - Experimenting with some RB systems
  - **TAUS & TDA** founding members (M's of words!)
- Partnering with Valencia's Computer Science Institute (R&D and EU projects: Casacuberta, Och, Vidal, Koehn)

# Short history

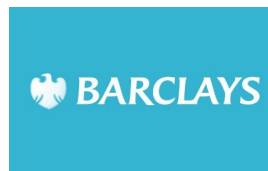
- CHALLENGE: Turn academic development (Moses) into a *commercial application*.
- Limitations: plain text (txt), language model building (first), no reordering, no updating features (always re-start), data availability, Linux-based (server). You need computational linguists (programmers), not translators, to operate it.
- Partnering with Valencia's Computer Science Institute PangeMatic (v1) was developed and then PangeaMT 2009 (web-based)

# Short history

## ■ OBJECTIVES:

1. To provide High Q MT for Post-Editing and save time and cost. No Google-type broad TR but domain-specific, user-centric.
2. To use only community-based **Open standards** → **Oasis / ISO: xliif / tmx, xml**). **NO proprietary formats** (technology independence) so clients are not “locked” in to buying and updating expensive software.
3. To automate as many processes as possible.

# Short history - Implementations



JAVIER GONZÁLEZ  
COO



j.gonzalez@mc-lehm.com  
T +34 91 702 27 97  
García de Paredes, 88  
28010 Madrid | SPAIN



Plus many  
other internal  
engines for ...

- \* Large Japanese Car manufacturing firm
- \* Electronics firms
- \* Technical / Engineering



# How PangeaMT works

Use Open Standards Browser: Mozilla, Safari



We build your engines

About PangeaMT

Partners

Description »

Client Area

Contact

FAQs

Users »

News

Blog

English ▾



## About PangeaMT

PangeaMT builds on SMT technologies that have been tested and applied by [Pangeanic](#) for its clients.

The company is privately owned and offers both consulting and systems integration. We train Statistical Machine Translation engines and customize them to your company's terminology and expressions. We make them fit in your current system so you do not need to purchase additional software.

PangeaMT is a European leader in the implementation of independent, open source statistical



Product Sheet



twitter



Scientists piece together  
EU media structure  
<http://bit.ly/h2am9P>

2 hours ago



Machine Translation -  
How it Works, What Users  
Expect, and What They Get

<http://bit.ly/g4NvAX>

2 days ago



Persian English MT an

# How PangeaMT works



## pangea.com.MT

The Statistical MT Customization Company

Plain Text (without tags)

1. Select File to  
translate



TMX / XLIFF / TXT File

Seleccionar archivo

Articul\_dtxlf

2. Select language pair



Language Pair

Spanish -> English

Engine

Legal 2010

Electronics 2010

Automotive 2009

Legal 2010

LSM 2010

MEB 2009

McLEHM Specific

McLEHM Extended

3. Select type of  
engine



Translation

PangeaMT – putting open standards to work... well



# How PangeaMT works

Users get an email with the translation minutes later

<input type="checkbox"/>	★ Elena, Manuel (3)	»	Recibidos	datos de rodrigo - hola deb he hablado con e	2 dic
<input type="checkbox"/>	★ gestion tecnologica	»	m.herranz@pangeanic.es	Proyectos I+D - Convocatoria	2 dic
<input type="checkbox"/>	★ pangeamt (2)	»	Recibidos	mherranz@pangea.com.mt [Pangea.com.MT] T	2 dic
<input type="checkbox"/>	★ pangeamt	»	Recibidos	mherranz@pangea.com.mt [Pangea.com.MT] T	2 dic
<input type="checkbox"/>	★ Translation Automation G.	»	TAUS_news	From Bob Donaldson and other Translation	1 dic
<input type="checkbox"/>	★ Pascal CANTAT via Linked.	»		RE: Manuel Herranz wants to connect on LinkedIn - Pas	1 dic
<input type="checkbox"/>	★ Antonio, Manuel, Elia (4)	»	Recibidos	mherranz@pangea.com.mt Paquete Cygwin para	1 dic
<input type="checkbox"/>	★ Elia Yuste_PangeaMT	»	Recibidos	mherranz@pangea.com.mt URG - nueva cuenta c	1 dic
<input type="checkbox"/>	★ Nataly Kelly (Common Sen.	»	Recibidos	What good deeds has your company done in	1 dic
<input type="checkbox"/>	★ Automated Language Trans.	»	TAUS_news	From Claudio Porcellana and other Automa	1 dic
<input type="checkbox"/>	★ Simon Stepney	»	Recibidos	GILTs_Xarxa_ ACT_Circular_22d_Recordat	1 dic

# Post-editing

En cualquier caso, hay una gran divergencia en la situación patrimonial y en los indicadores de rentabilidad y morosidad de estas entidades.

Así, mientras algunas siguen destacando por su sana situación, otras atraviesan serias dificultades y dos de ellas han requerido la intervención de la autoridad supervisora.

Así, en el primer semestre del año 2010 se han producido intensos procesos de concentración que, en algunos casos, se han formalizado mediante la constitución de sociedades participadas por las cajas de pensiones de los SIPs, lo que abre la puerta a futuras ampliaciones de capital o "privatizaciones" que permiten la entrada de capital de inversores institucionales.

En este sentido, la estructura jurídica puede reducirse a la gestión de la actividad que caracteriza a estas entidades, por lo que se plantea una estructura dual que contempla la actividad financiera enmarcada en una compañía por acciones y, paralelamente, pervive la esencia de la caja de pensiones como forma jurídica fundacional para la

In any case, there is a wide variation in the balance sheet, profitability indicators and bad debts of these organisations.

TM Databases

* Database	%
<input type="checkbox"/> ECH_Carlos_autogen	0%
<input type="checkbox"/> ECH_Carlos_web	0%
<input type="checkbox"/> ecofin	0%
<input checked="" type="checkbox"/> emma	0%
<input type="checkbox"/> ENESECHTECH2	0%
<input type="checkbox"/> internal5	0%
<input type="checkbox"/> internal6	0%

Write-enabled Database: ecofin

Set Match Penalty

Accept Cancel

Penalty: 20

Accept Cancel

algunos casos, se han formalizado mediante la

Thus, in the first half of 2010 there have been intensive process of concentration which, in some cases, have been formalized through the

GTranslate

Auto-Translation

SET MT MATCH  
PENALTY



Terminology

1	al año	annually
2	Nº procesos	No. processes
3	CAJAS DE AHORRO	SAVINGS BANKS

# Future Work

- “on the fly” MT training (minutes, not manually) – April 2011 !!
- pick and match sets of data: “extreme customization” – April 2011 !!
- objective stats for post-editors (calculate effort)
- confidence scores for users (→ translators or readers) with CAT integration (web-based / desktop)
- Web samples

---

# Thank you !

## QUESTIONS ?

[mherranz@pangea.com.mt](mailto:mherranz@pangea.com.mt)