Directorate-General for Translation

Machine Translation
at the European Commission

Translingual Europe
Berlin, 7 June 2010

Spyros Pilos
Head of sector Language applications

EUROPEAN
COMMISSION

Machine Translation at the EC

→ Translation@EC

→ MT@EC

→ What next

Translation@EC

Directorate-General for Translation (DGT)

- Staff:
  1750 linguists and 600 support

- Production (M pages):
  0,9 (1992) 1,2 (2004) 1,8 (2008)

- Cost:
  EC translation: 300 M€
  all translation and interpretation: 2€/y per citizen

BUT

to make europa.eu fully multilingual

- translate almost 6,8 million documents

- 8,500 translators working full-time for one year

The present: ECMT service

- managed by DGT
- rule-based machine translation
- developed between 1975 and 1998
- 28 language pairs available (ten languages)
- since 2006 only maintenance reduced work on dictionary enrichment on a couple of systems
- use (M.pages): 1,5 (2006) more than 2,5 (2009)
- who uses it and what for
    - EU institutions and public bodies for gisting
    - Online services and information systems for raw translation
    - DGT as a CAT tool

The future: MT@EC service

Policy

Commission Communication on "Multilingualism" 2008: *"human and automatic translation is an important part of multilingualism policy"*

Facts

- ECMT is rule-based and costly to develop
- Data-driven systems are cheap and quick to develop….
  if you have the data

Language Technology Watch

- Market and research observation
- Tests of commercial and non commercial tools and MT systems

MT@EC Needs – resources - action

MT@EC strategy

- Adopted in June 2009 by DGT
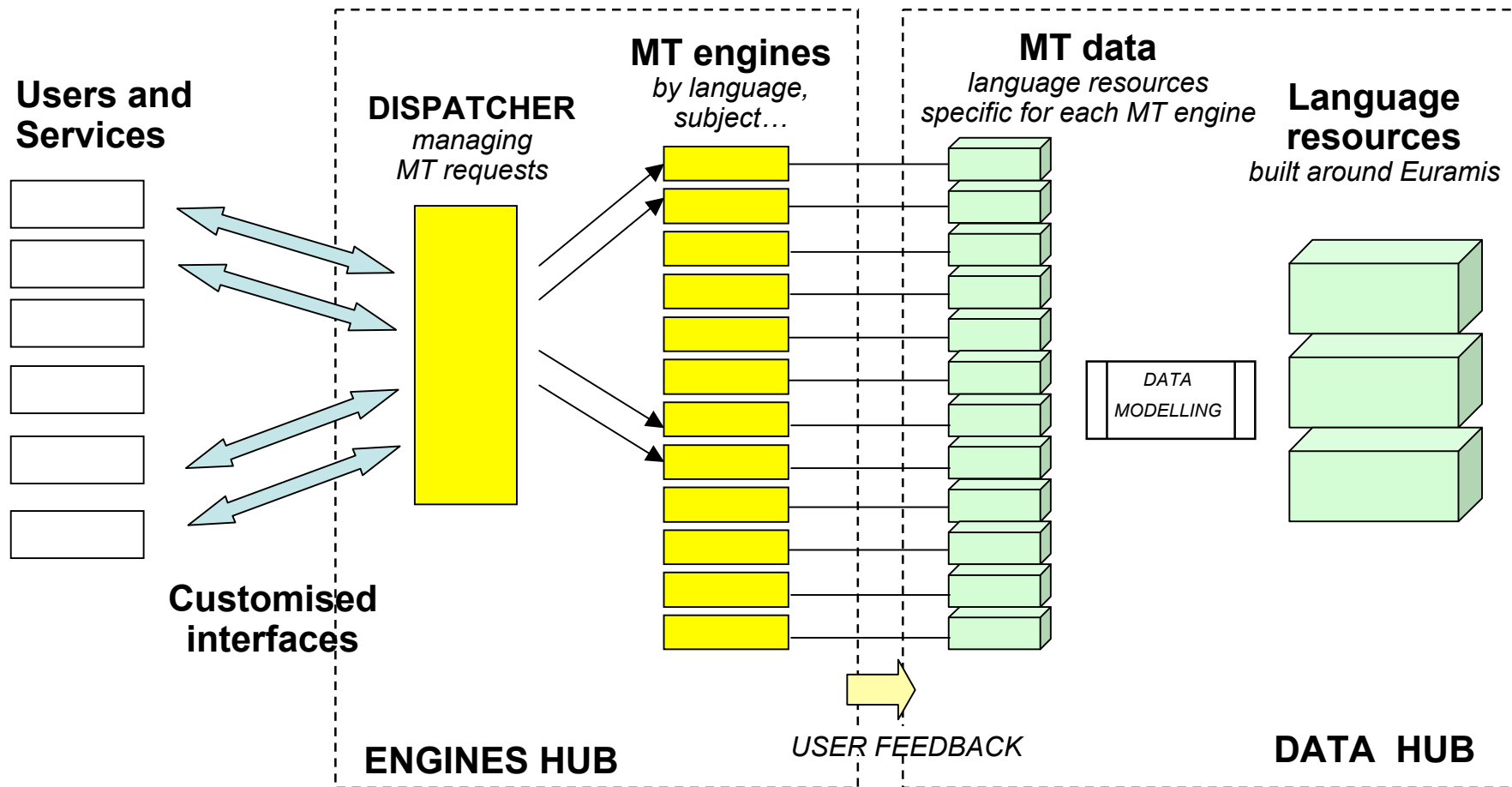
- Task Force created *November 2009*

Task Force results *April 2010*

- MT@EC is necessary for the Commission

  (trust, confidentiality, continuity)

- Data-driven systems: a major technological breakthrough

- User requirements have been collected

- An outline of an "architecture" has been elaborated

  (flexible, sustainable, ensuring technological independence)

- Recommendations on organisational and financial arrangements

Machine Translation Service
Outline of the proposed MT@EC architecture

Users and Services

DISPATCHER
managing MT requests

MT engines
by language, subject…

MT data
language resources specific for each MT engine

Language resources
built around Euramis

DATA MODELLING

Customised interfaces

ENGINES HUB

USER FEEDBACK

DATA HUB

Machine Translation Service
A number of projects within a "MT@EC programme"

"MT Engines - baseline" project **(EC)**

   IT infrastructure for the core of the "MT Engines Hub"

"MT data management hub" projects **(DGT)**

   Language resources (LR) underlying the MT system

"Customised MT solutions" projects **(clients)**

   "Client" requesting development of (examples) :

   – a domain specific MT engine

   – a specific interface to external services

Exodus

- Internal DGT experimentation with Moses toolkit

- Using Euramis (internal) TM data

- With temporary redeployment of existing IT and human resources by the DGT IT unit

- With the active contribution of :

  – the Portuguese language department of the EC

  – the EuromatrixPlus

  – the European Parliament

Exodus

What was done

- Development of the EN->PT engine
- Corpus preparation and cleaning
- Human evaluation by the PT LD (more than 30 translators involved)

What has not been done

(due to time and resource limitations)

- No iterative process for improving corpus quality.
- No incremental updates of translation and language models
- No engineering interventions

Exodus

First conclusions

- Quality evaluation of MT output for EN->PT results very encouraging

- Dedicated analysis on IT engineering work required for production ready system for all EU languages

- Quality of data cleaning and preparation: the main "comparative" advantage of DGT

Note:

More Exodus pairs are currently being evaluated
by the European Parliament, who also submitted an Exodus pair
(EN-to-FR) to the WMT 2010 competition

Translingual Europe - MT@EC

Next: putting pieces together

- Action plan

- Set up governance for MT@EC and MT@DGT

- Work on data in preparation of the "MT data hub"

- **A key challenge:** compare alternative systems (both commercial and non-commercial) in terms of:
  - quality of output
  - price (total cost of ownership)
  - feasibility
  - language coverage

- In parallel EC is preparing to continuously update the DGT Multilingual Translation Memory of the Acquis Communautaire (DGT-TM) - autumn 2010

Thank you