# PANACEA

**Núria Bel**

Universitat Pompeu Fabra

*Language Technology Days*

*Luxembourg, 22/23 March 2010*

**PANACEA's objective** is

to join together a number

of advanced interoperable tools

to build a

**factory of Language Resources**

A production line that **automates** the stages involved in the acquisition, production, updating and maintenance of the LR required by MT and other Language Technologies.

Cost and time reduction by automation
is the only way to ensure
the **continuous supply** of LR’s
that can guarantee a LT industry
covering all languages, all domains,
for current and future needs, and
in the time required by the market.

The factory is build as a Web Service-based platform for easy integration of the **latest** technological components for:

- Monolingual and Parallel Text Acquisition and Pre-processing
- Sentential and sub-sentential alignment
- Bilingual Dictionary and Transfer Grammar production
- Lexical Information Acquisition for rich information dictionary production.

- These resources will feed different applications:
  - Machine Translation (Rule-based and Statistical MT)
  - Multilingual Information Extraction
  - Multilingual Question Answering
  - Event detection and tracking
  - Natural Language Interfaces.

# Project results (1/3)

1. The platform, as a virtual, distributed, production line where different interoperable components can be chained in particular workflows to produce different types of LR's, for different languages.

- The definition of a platform (i.e. an interoperability space built upon the definition of components and objects which are compatible among them)
- A dedicated Panacea Registry, metadata and middleware for the location, searching and documentation of Panacea components.
- Dedicated Panacea workflow editor for defining different production chains.

# Project results (2/3)

2. The automatic acquisition and production components:

- Corpus Acquisition Component
- Corpus clean-up and Normalization Component
- Text Processing Components for sentence splitting, PoS Tagging, lemmatization, chunking and NER
- Sentential and subsentential aligners
- Bilingual dictionary extractor
- Transfer grammar extractor
- Lexical information Induction component
- Lexical classifiers
- Dictionary merger

# Project results (3/3)

3. LR's used as test and proof of the proper functioning of the factory.

- Parallel texts, cleaned and prepared for training-building translational models.
- Large monolingual corpus, PoS tagged and lemmatized for training and modelling language data,
- Monolingual lexica with morpho-syntactic, syntactic and lexical-class semantic information
- Bilingual dictionary and transfer grammar

# Evaluation

PANACEA's contribution & impact
will be demonstrated with a significant
time and cost reduction
in producing LR's.
A real life use case will be used to measure the achievements

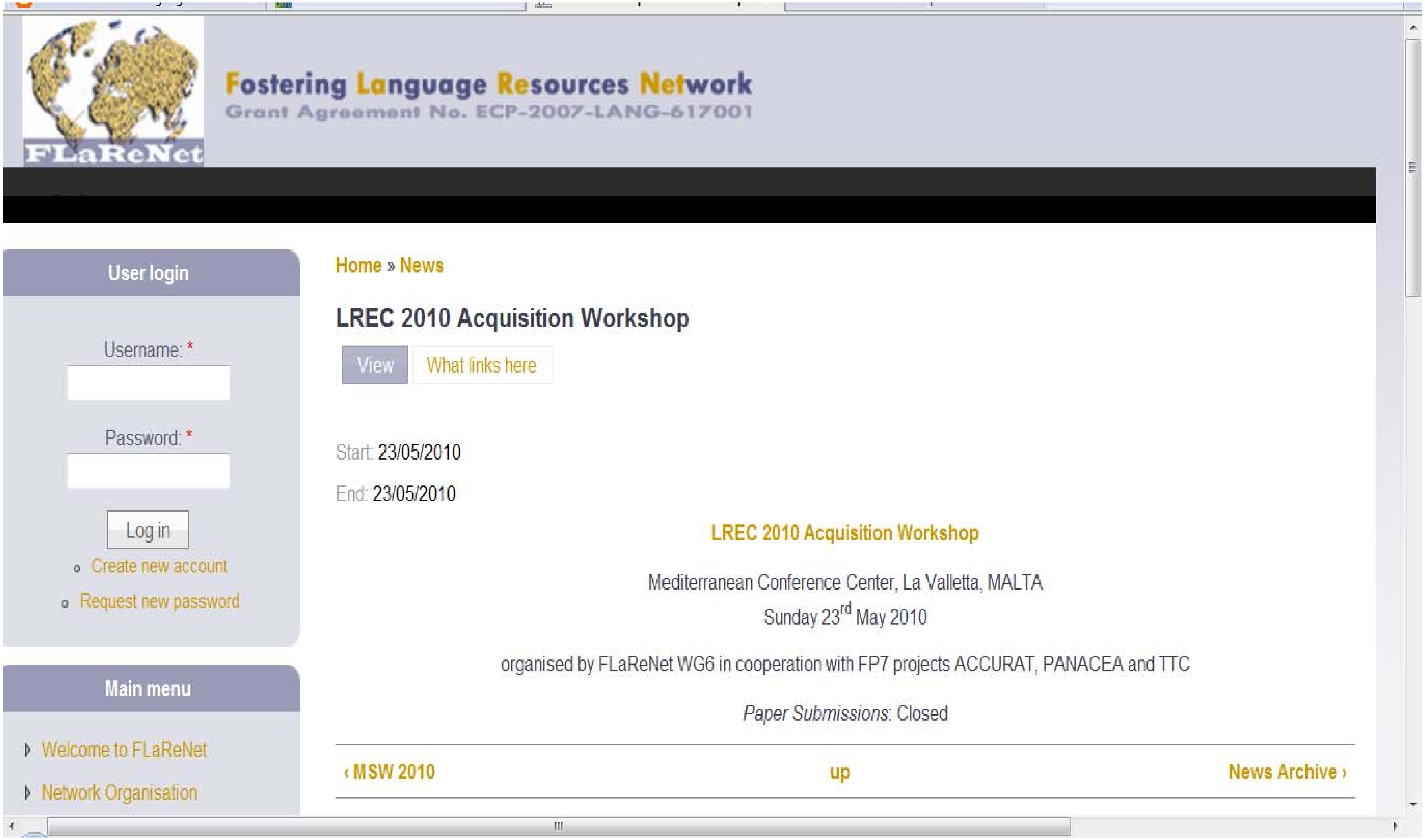
# The challenges

- Standards for the integration of robust, scalable web service-deployed components.
- Handling the impact of dealing with massive data
- Be convincing about the industrial use of available automatic acquisition technologies by introducing ready to use tools, with **confidence** indicators and which give priority to high **precision**
- Research for increasing accuracy in automatic acquisition & production

# Activity & Results will be disseminated

- As scientific papers submitted to conferences and journals
- In workshops addressed to specific profiles: researchers, professionals and industry.
- In the web page www.panacea-lr.eu
- Harvesteable metadata and active subscription to catalogues and harversters

# Our first WS…



Fostering Language Resources Network

Grant Agreement No. ECP-2007-LANG-617001

## LREC 2010 Acquisition Workshop

Start: 23/05/2010

End: 23/05/2010

LREC 2010 Acquisition Workshop

Mediterranean Conference Center, La Valletta, MALTA
Sunday 23rd May 2010

organised by FLaReNet WG6 in cooperation with FP7 projects ACCURAT, PANACEA and TTC

*Paper Submissions*: Closed

# The project

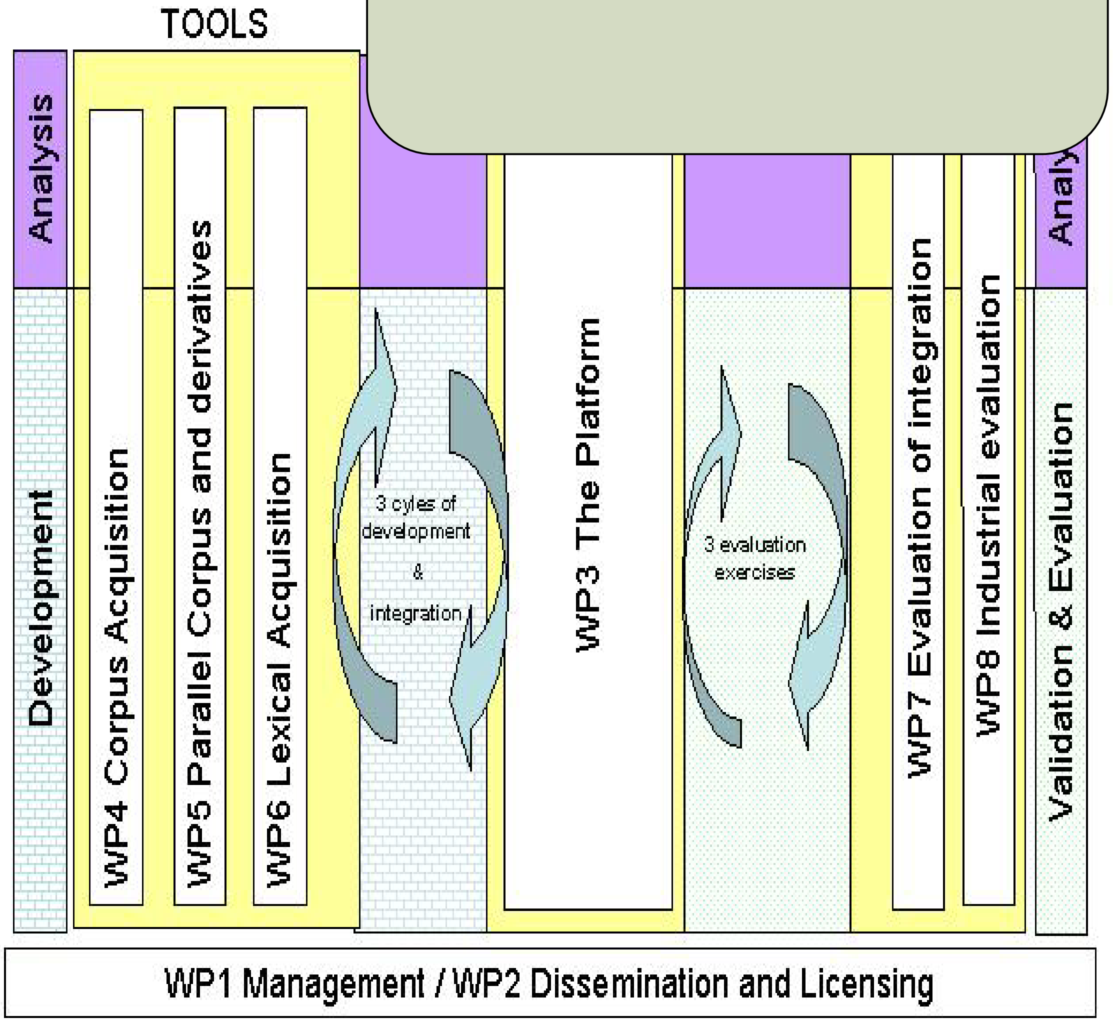
# Consortium

| | | |
|---|---|---|
| **Prof. Núria Bel**<br>**Universitat Pompeu Fabra - UPF** | ES | UNIVERSITAT UPF POMPEU FABRA |
| **Dr. Nicoletta Calzolari**<br>**Consiglio Nazionale delle Ricerche - Istituto de Linguistica Computazionale – ILC** | IT | ILC |
| **Dr. Stelios Piperidis**<br>**Institute for Language & Speech Processing ILSP** | GR | IEΛ ILSP |
| **Dr. Anna Korhonen**<br>**University of Cambridge – UCAM** | UK | UNIVERSITY OF CAMBRIDGE |
| **Dr. Gregor Thurmair**<br>**Linguatec -- LT** | DE | linguatec language technologies |
| **Prof. Andy Way**<br>**Dublin City University -- DCU** | IR | DCU |
| **Dr. Khalid Choukri**<br>**Evaluations and Language Resources Distribution Agency -- ELDA** | FR | eLDa |

# PANACEA WP's

- WP1 – Coordination (UPF)
- WP2 – Dissemination and Exploitation (ELDA)
- WP3 – The Platform (UPF)
- WP4 – Corpus Acquisition & Annotation (ILSP)
- WP5 – Parallel corpus & derivatives (DCU)
- WP6 – Lexical Acquisition (UCAM)
- WP7 – Integration & resource evaluation (ILC)
- WP8 – Evaluation in industrial environment (LT)

# First results in t14


TOOLS
Analysis
Development
WP4 Corpus Acquisition
WP5 Parallel Corpus and derivatives
WP6 Lexical Acquisition
3 cyles of development & integration
WP3 The Platform
3 evaluation exercises
WP7 Evaluation of integration
WP8 Industrial evaluation
Analy
Validation & Evaluation
WP1 Management / WP2 Dissemination and Licensing

2 Big Phases:
Analysis & Development

3 Cycles of development, integration and evaluation

1 Final Industrial Evaluation

# Summary

PANACEA is to build

a Language Resource factory

that will ensure the supply that Language Technology industry needs to break through problems such as Machine Translation systems covering all languages, all domains, for current and future needs, and in the time required by the market.

PANACEA will open new challenges:

- Automation of the production of resources for dialogue, interaction commands, and new demands.
- Deployment of broker web services dedicated to convert formats, add specialized information, and many others …

Keep informed at

www.panacea-lr.eu

Thanks!

This document is part of dissemination material generated in the PANACEA Project, Platform for Automatic, Normalized Annotation and Cost-Effective Acquisition (Grant Agreement no. 248064).

Please send feedback and questions on this document to:
iulatrl@upf.edu

TRL Group (Tecnologies dels Recursos Lingüístics), Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra (IULA-UPF)

Barcelona, 2010