The PANACEA platform

Núria Bel (UPF)

COCOSDA/WRITE/FLARENET
May 22, 2010

Consortium

Prof. Núria Bel Universitat Pompeu Fabra - UPF	ES	
Dr. Nicoletta Calzolari Consiglio Nazionale delle Ricerche - Istituto de Linguistica Computazionale – ILC	IT	
Dr. Stelios Piperidis Institute for Language & Speech Processing ILSP	GR	
Dr. Anna Korhonen University of Cambridge – UCAM	UK	
Dr. Gregor Thurmair Linguatec -- LT	DE	
Prof. Andy Way Dublin City University -- DCU	IR	
Dr. Khalid Choukri Evaluations and Language Resources Distribution Agency -- ELDA	FR	

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 to be build as a Web Service-based platform for easy integration of all type of technological components for:

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

Open platform

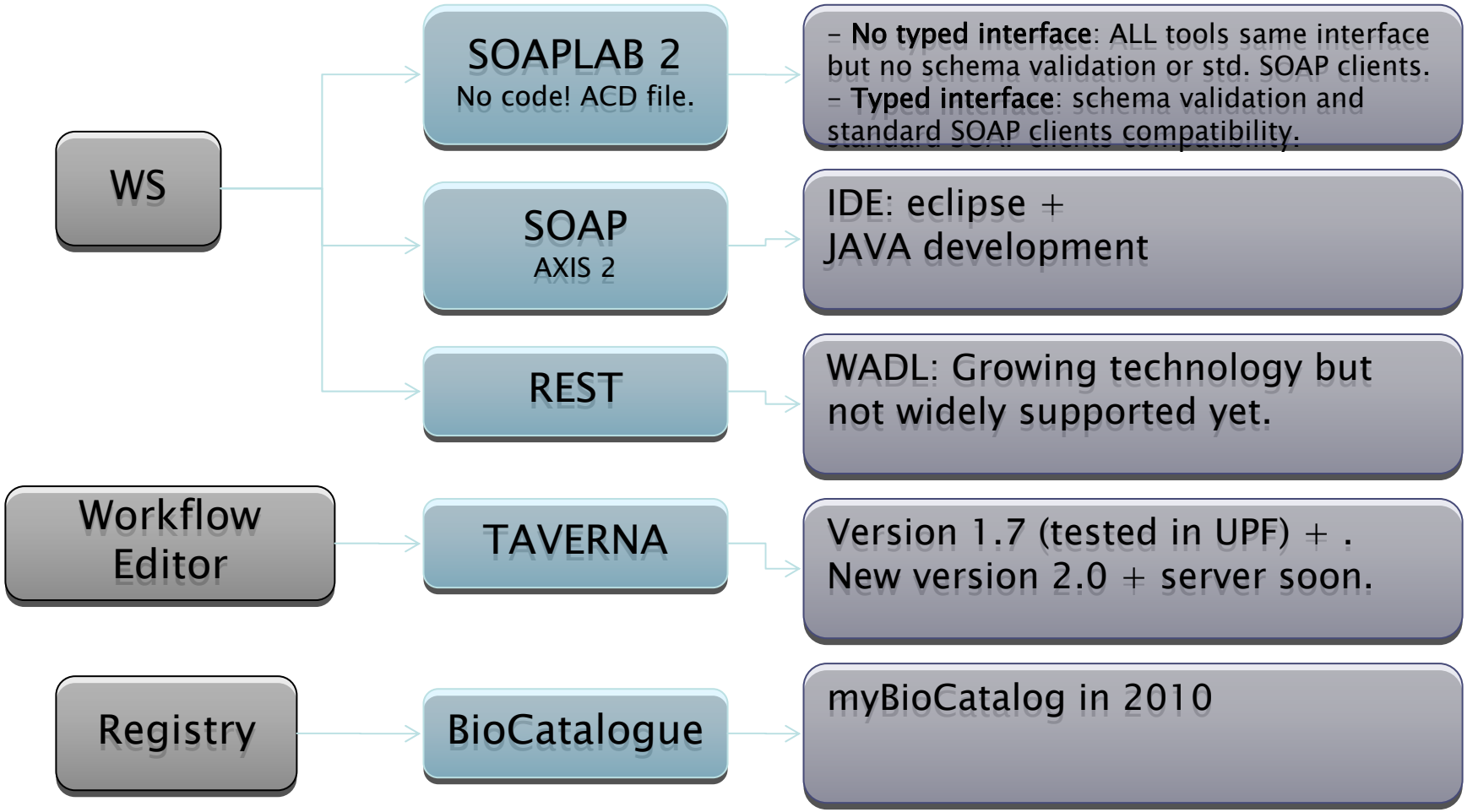
The platform, as a virtual, distributed, production line where different interoperable components can be chained in particular workflows to produce different types of LR's, for different languages.

- **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.**
- **A dedicated Panacea workflow editor for defining different production chains.**

Some challenges, the strategies

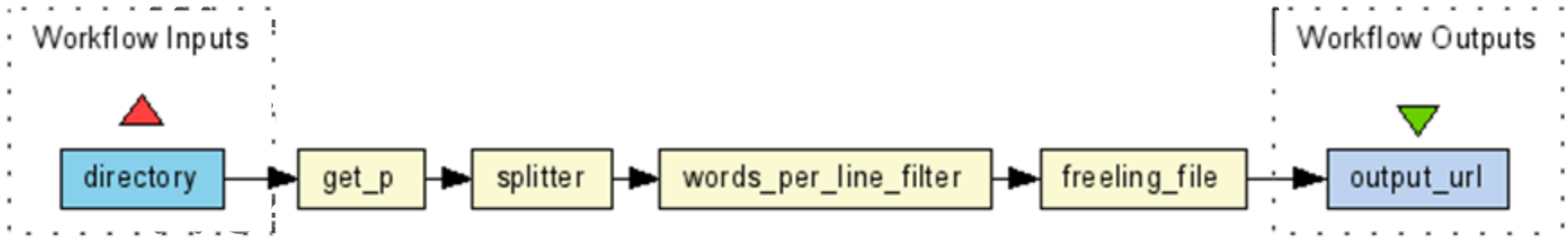
- Handling the impact of dealing with big, massive data (like in other e-sciences).
- Be convincing about the industrial use of available automatic acquisition technologies by introducing ready to use tools (short learning and setting up curves).
- Foster the use of standards motivated by the integration of **robust**, **scalable** web service-deployed components.

PANACEA platform is now based on some of MyGrid tools.

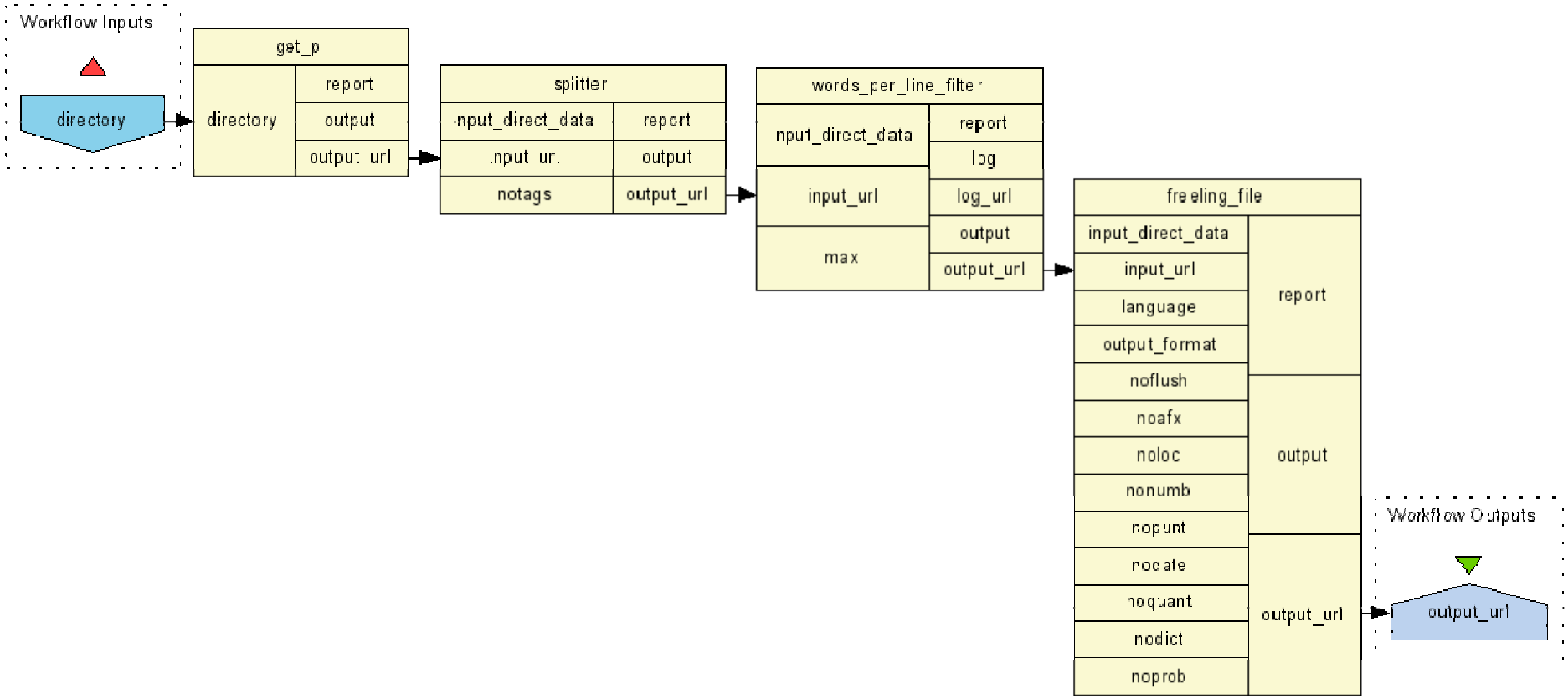


IULA Workflow test

- http://gilmere.upf.edu:9903/soaplab2/
- **Taverna Workflow:**
  - **Get_p: extracts text from JR-Acquis corpus (TEI)**
  - **Splitter: simple sentence splitter**



IULA Workflow test



Analysis

Pros	Cons
- Mature technology (lots of success histories). - Most tools have support. - Tools have nice GUI. - Small learning curve. - Tools are free and open source. - Massive data - Nice semantic model (ontology). - WS oriented.	- No security tools yet.

FYI


myGrid » Home - Mozilla Firefox
Fitxer Edita Visualitza Historial Adreces d'interès Eines Ajuda
http://www.mygrid.org.uk/
mygrid
myGrid » Home
myGrid
omii-uk
www.omii.ac.uk
Home Tools Taverna Projects Research myGrid In Use Outreach About us News
Search
Home
Tools
Taverna
Projects
Research
myGrid In Use
Outreach
About us
News
Presentations
Publications
Home
The myGrid team produce and use a suite of tools designed to "help e-Scientists get on with science and get on with scientists". The tools support the creation of e-laboratories and have been used in domains as diverse as systems biology, social science, music, astronomy, multimedia and chemistry. The tools have been adopted by a large number of projects and institutions.
The team has developed tools and infrastructure to allow:
the design, editing and execution of workflows in Taverna
the sharing of workflows and related data by myExperiment
the cataloguing and annotation of services in BioCatalogue and Feta
the creation of user-friendly rich clients such as UTOPIA
These tools help to form the basis for the team's work on e-Labs.
Taverna Workbench 2.1.2 available for download
The myGrid team have released Taverna Workbench 2.1.2. It supports secure access to data on the web, secure web services and secured private workflows.
Taverna 2.1
DOWNLOAD NOW!

Standards & Travelling Objects

- **Character encoding:**
  - The majority of the tools surveyed (87%) are compatible with the **UTF-8**.
- **Encoding formats:**
  - Around 70% of the tools use some kind of **vertical or inline format**.
  - 13% export their results in both **standoff and inline annotation** files.
  - 28% use encoding format like **XCES** or **UIMA CAS**.
- **Linguistic data:**
  - 3 of 6 use **EAGLES/Parole** compatible tagsets for POS tagging.
  - For EN constituency parsers (and integrating taggers) the **Penn Treebank categories** are clear winners.

Thanks!

Evaluation Visibility

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

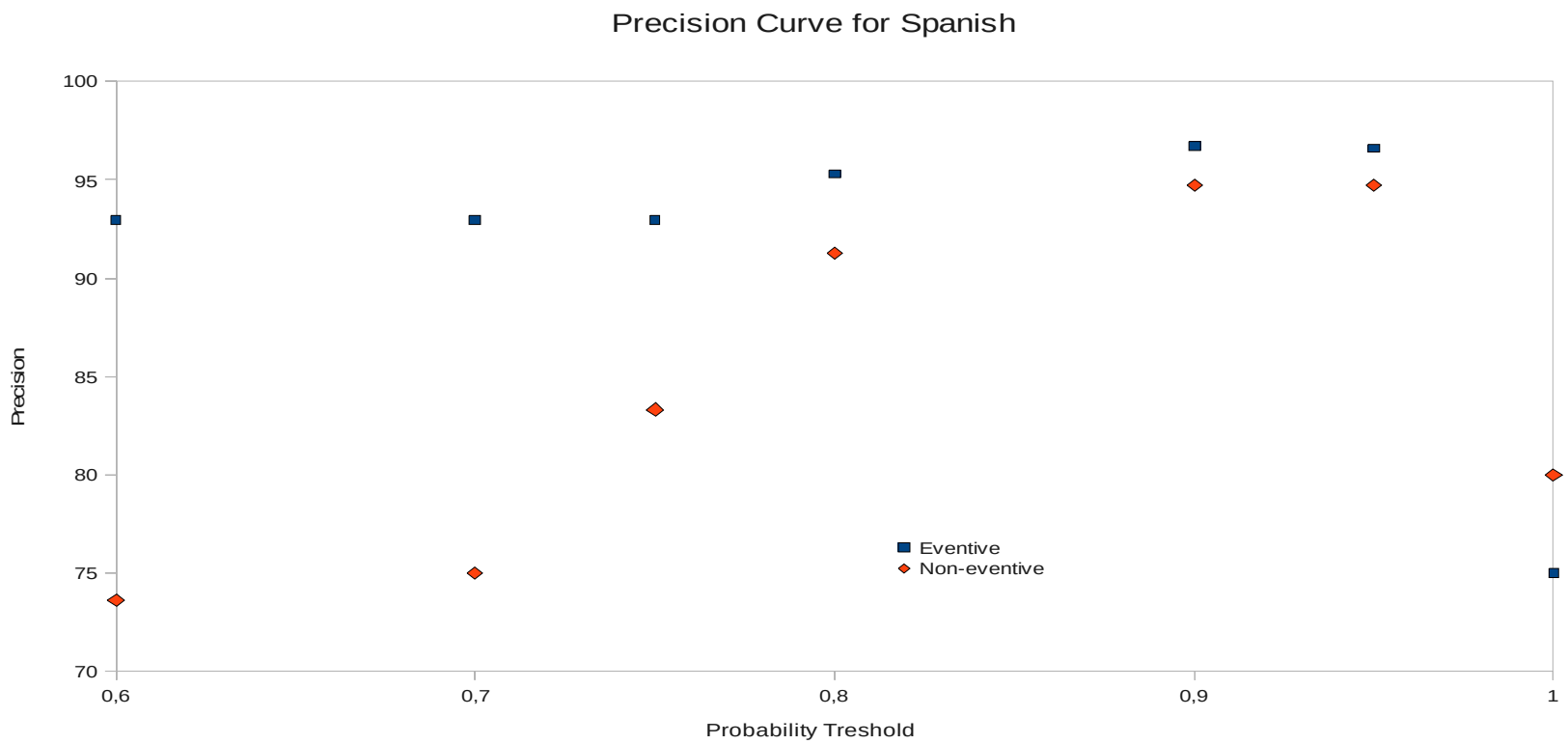
Evaluation Methods

- PANACEA's aim is to reduce the costs that imply human intervention (time/labour costs) then we have to measure our results accordingly.
- Current metrics based on accuracy, precision and recall are difficult to be interpreted as something related to cost reduction
- Besides, manual validation of all the results is mandatory with accuracy results below 90%, again

PANACEA's proposal

- Is it possible to separate the results of automatically produced resources that need human revision of those that do not?
- Can we separate the ones that should not be revised (acceptable error rate) as a measure of human intervention reduction, hence cost reduction?

An example: eventive nouns detection



Can we do this for all our resources?

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

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.**
   - **A 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 modeling language data,
- Monolingual lexica with morpho-syntactic, syntactic and lexical-class semantic information,
- Bilingual dictionary and transfer grammar

Summary

PANACEA is to build

a Language Resource factory

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

The case of lexica of event nouns

- We based our proposal in the use of the confidence measure
- We assessed the level of confidence that ensured maximum precision in results although decreasing coverage
- We identified a number of entries that should not be reviewed

Activity & Results will be disseminated

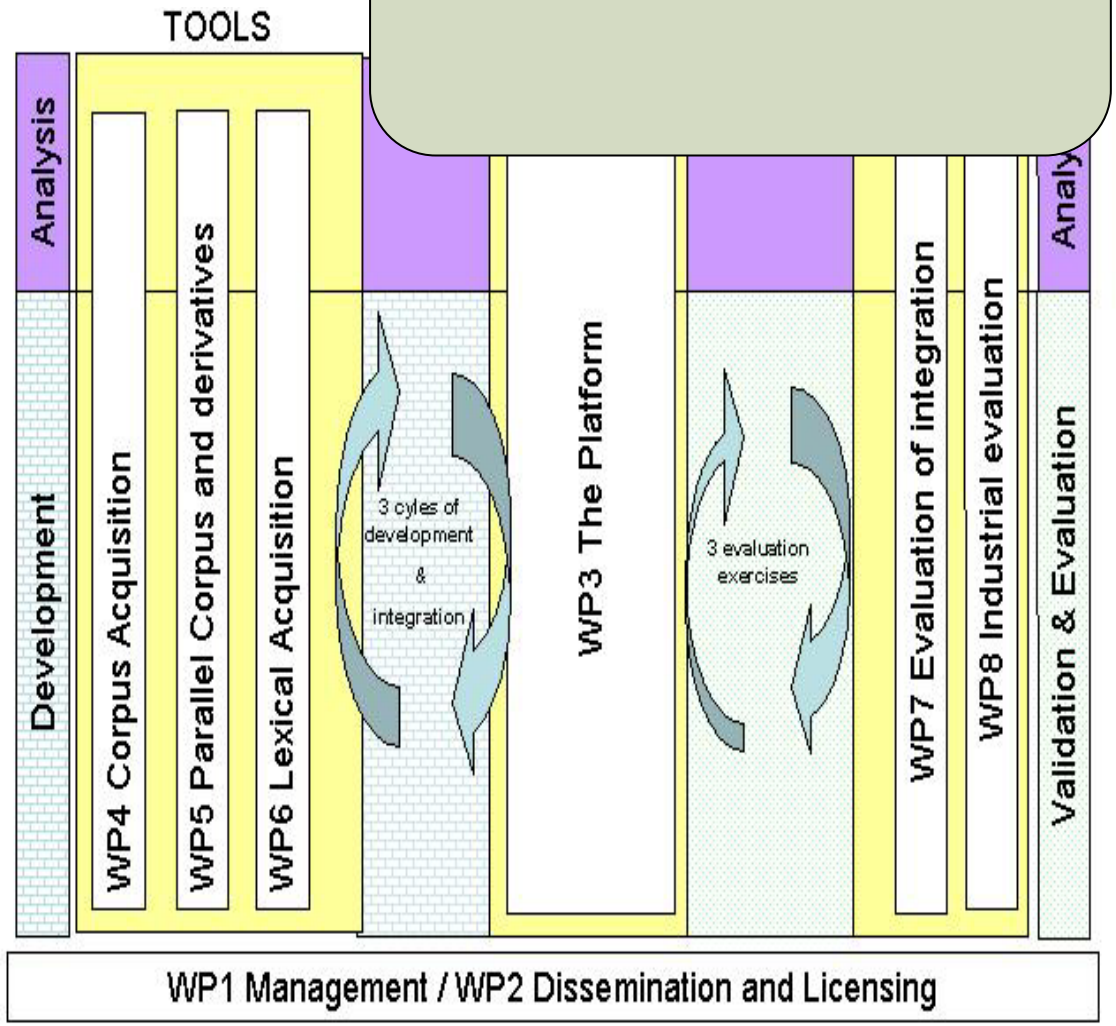
- As scientific papers submitted to conferences and journals
- In workshops addressed to specific profiles: researchers, professionals and industry.
- In the web page www.panacea-lr.eu
- Harvesteable metadata and active subscription to catalogues and repositories.
- Active collaboration with ongoing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

The project

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



2 Big Phases:
Analysis & Development

3 Cycles of development, integration and evaluation

1 Final Industrial Evaluation

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 …

PANACEA

SEVENTH FRAMEWORK PROGRAMME

Keep informed and
get in touch with us at

www.panacea-lr.eu