PANACEA

7FP-ICT-248064

Kick-off Meeting 21/22-01-2010

Barcelona

Núria Bel – Universitat Pompeu Fabra

PANACEA's objective is

to join together a number of interoperable technological tools to build a

factory of Language Resources

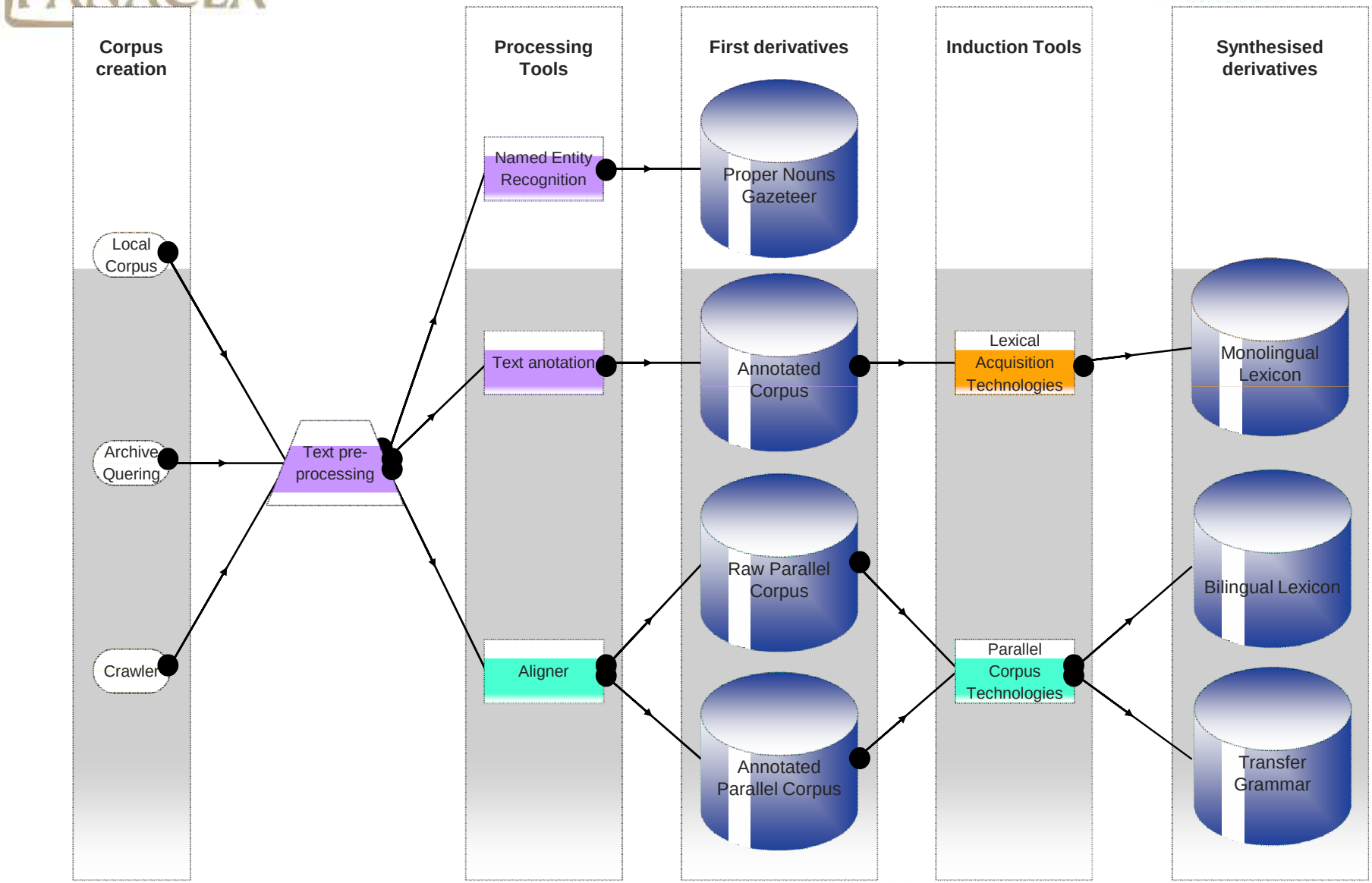
A production line that automates the stages involved in the acquisition, production, updating and maintenance of the LR required by MT and other Language Technologies.

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

Main features of PANACEA

The factory is build upon a web service-based platform for easy integration of the latest technological components for:

- Monolingual and Parallel Text Acquisition and Preprocessing
- Sentential and subsentential alignment
- Bilingual Dictionaries and Transfer Grammars production
- Lexical Acquisition for rich information lexica production.


Corpus creation
Processing Tools
First derivatives
Induction Tools
Synthesised derivatives
Local Corpus
Archive Quering
Crawler
Text pre-processing
Named Entity Recognition
Text anotation
Aligner
Proper Nouns Gazeteer
Annotated Corpus
Raw Parallel Corpus
Annotated Parallel Corpus
Lexical Acquisition Technologies
Parallel Corpus Technologies
Monolingual Lexicon
Bilingual Lexicon
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 validated platform (i.e. an interoperability environment built upon the definition of components which are compatible among them)
   – Dedicated Panacea Registry, metadata and middleware for the location, searching and information 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
- 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 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 morphosyntactic, syntactic and lexica-class semantic information
   - Bilingual dictionary and transfer grammar

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

PANACEA's challenges

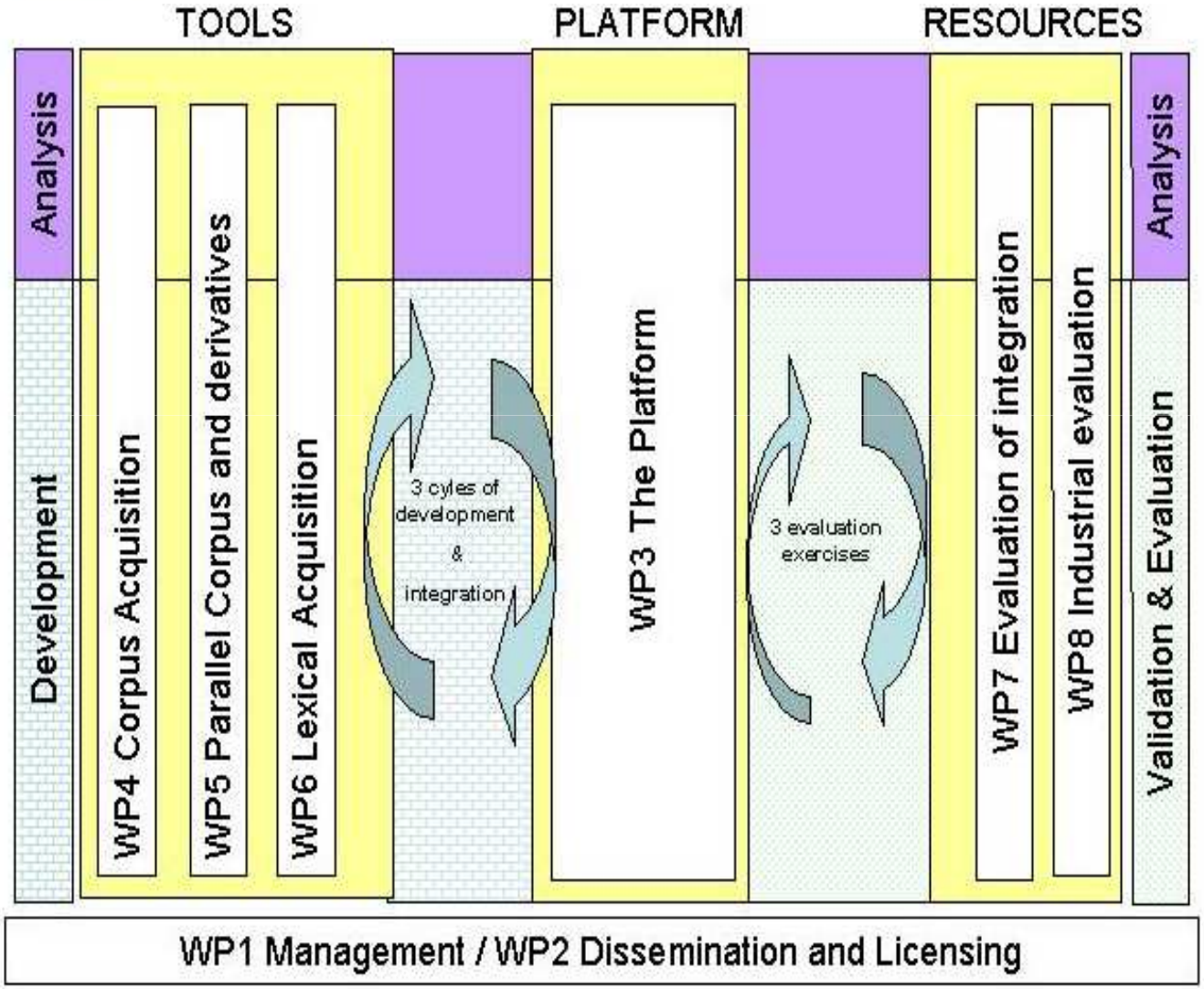
- Deployment of robust, scalable web service-deployed components assembled in a LR factory. Unknown impact of massive data handling.
- Be convincing about the industrial use of available acquisition technologies by introducing ready to use tools and workflows, with confidence indicators and which give priority to high precision results
- Research for improving accuracy in acquisition technologies

The project
Work Plan

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)

Methodology



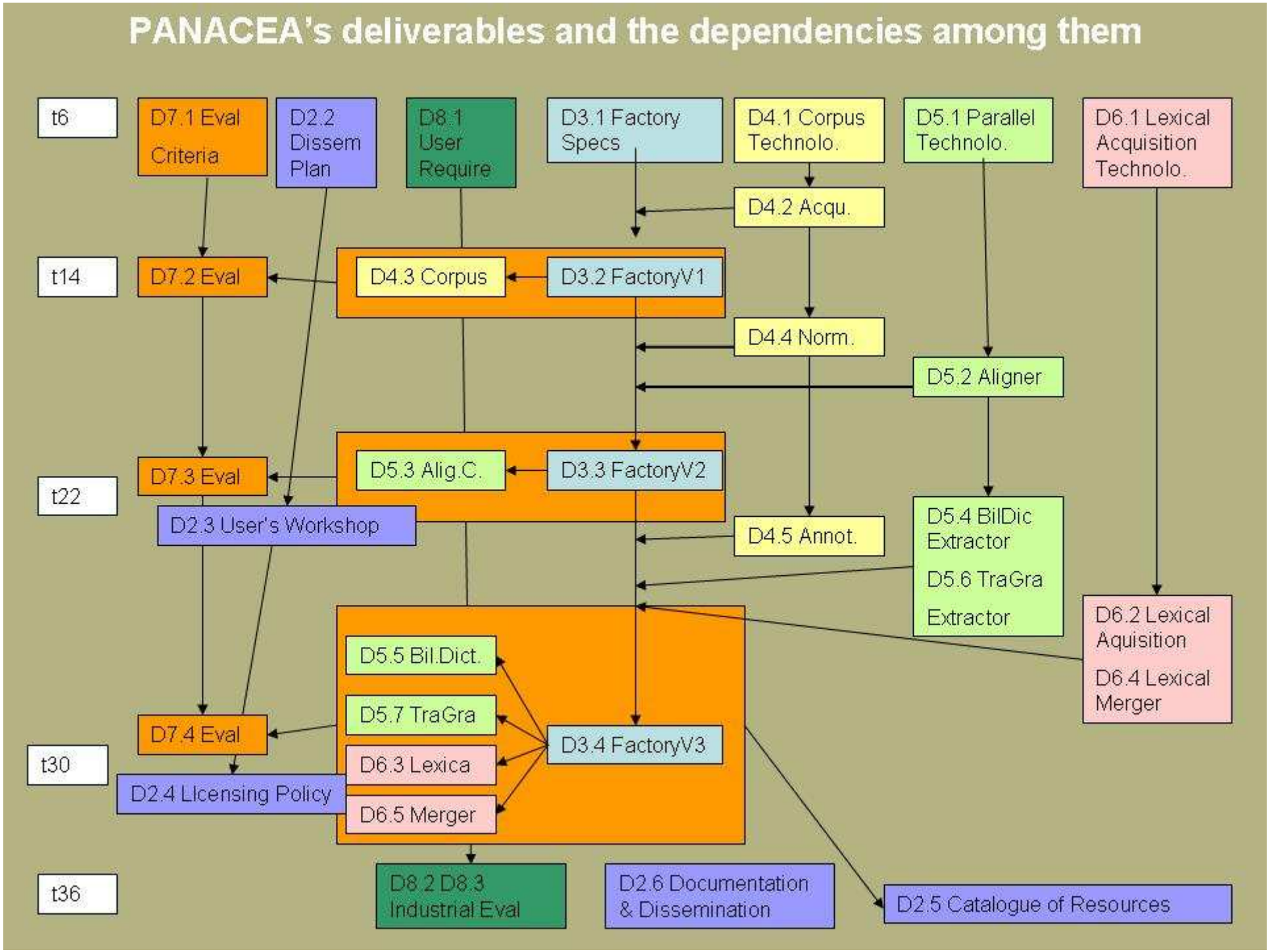
2 Big Phases: Analysis & Development

3 Cycles of development, integration and evaluation

1 Final industrial evaluation

Calendar

	Workpackage Description		1st year												2nd year												3rd year												Start	End
			M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20	M21	M22	M23	M24	M25	M26	M27	M28	M29	M30	M31	M32	M33	M34	M35	M36		
3	**WP1**	**COORDINATION**																																					M1	M36
4	T1	Project Coordination																																					M1	M36
5	**WP2**	**Dissemination, exploitation**																																					M1	M36
6	T2.1	Exploitation																														D4							M1	M30
7	T2.2	Legal Framework																														D4							M18	M30
8	T2.3	Dissemination			D1,2										D6.1									D3								D6.2						D5	M1	M36
9	**WP3**	**The platform**																																					M1	M30
10	T3.1	Architecture and design of the platform						D1																															M1	M6
11	T3.2	Work Flow editor and engine																					id							id	D4								M15	M30
12	T3.3	Common interfaces, middleware, etc.													id	D2							id	D3						id	D4								M7	M30
13	T3.4	The Registry																					id							id	D4								M15	M30
14	T3.5	Deployment of web services													id	D2							id	D3						id	D4								M7	M30
15	**WP4**	**Corpus Acquisition and annotation**																																					M1	M30
16	T4.1	Corpus Acquisiton component						D1						id	D2,3																								M1	M14
17	T4.2	Clean-up and normalization						D1							id	D3							id	D4															M1	M30
18	T4.3	Text processing components						D1															id								D5								M1	M30
19	**WP5**	**Parallel corpus derivatives**																																					M1	M30
20	T 5.1	Aligners						D1						id	D2								id	D3															M1	M22
21	T 5.2	Bilingual Dictionaries Induction Tech's						D1															id							id	D4,5								M1	M30
22	T 5.3	Transfer Grammar Induction Tech's						D1															id							id	D6,7								M1	M30
23	**WP6**	**Lexical Acquisition**																																					M1	M30
24	T 6.1	Subcat, SP and MW Acquisition						D1							id								id							id	D2,3							M1	M30
25	T 6.2	Lexical-Semantic Classes Classification						D1							id								id							id	D2,3							M1	M30
26	T 6.3	Merging of dictionaries						D1							id								id							id	D4,5							M1	M30
27	**WP7**	**Evaluation**																																					M1	M30
28	T7.1	Analysis and definition						D1																															M1	M6
29	T7.2	Evaluation of integration of components														D2								D3								D4							M13	M30
30	T7.3	Evaluation of resources produced														D2								D3								D4							M13	M30
31	T7.4	Evaluation of resources for MT														D2								D3								D4							M13	M30
32	**WP8**	**Evaluation in industrial environment**																																					M1	M36
33	T8.1	Description of industrial set-up						D1																															M1	M6
34	T8.2	Tool oriented evaluation																																				D2	M31	M36
35	T8.3	Task-based evaluation																																				D3	M31	M36


PANACEA's deliverables and the dependencies among them
t6
t14
t22
t30
t36
D7.1 Eval Criteria
D2.2 Dissem Plan
D8.1 User Require
D3.1 Factory Specs
D4.1 Corpus Technolo.
D5.1 Parallel Technolo.
D6.1 Lexical Acquisition Technolo.
D4.2 Acqu.
D7.2 Eval
D4.3 Corpus
D3.2 FactoryV1
D4.4 Norm.
D5.2 Aligner
D7.3 Eval
D5.3 Alig.C.
D3.3 FactoryV2
D2.3 User's Workshop
D4.5 Annot.
D5.4 BilDic Extractor
D5.6 TraGra Extractor
D6.2 Lexical Aquisition
D6.4 Lexical Merger
D5.5 Bil.Dict.
D5.7 TraGra
D7.4 Eval
D6.3 Lexica
D3.4 FactoryV3
D2.4 Licensing Policy
D6.5 Merger
D8.2 D8.3 Industrial Eval
D2.6 Documentation & Dissemination
D2.5 Catalogue of Resources

PANACEA is to build

a Language Resource factory that will ensure the LR continuous supply that Language Technology industry needs to break through the problem for LT's covering all languages, all domains, for current and future needs, and in the time required by the market.