

PROJECT FINAL REPORT

Grant Agreement number: 248064

Project acronym: PANACEA

Project title: Platform for Automatic, Normalized Annotation and Cost-Effective Acquisition of Language Resources for Human Language Technologies

Funding Scheme: STREP

Period covered: from: 1-1-2010 to: 31-12-2012

Name of the scientific representative of the project's co-ordinator¹, Title and Organisation:

Dr. Núria Bel Rafecas

UNIVERSITAT POMPEU FABRA

Tel: +34 935422307

Fax: +34 935422321

E-mail: nuria.bel@upf.edu

Project website address: www.panacea-lr.eu

¹ Usually the contact person of the coordinator as specified in Art. 8.1. of the Grant Agreement.

Project Final Report

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



This document is licensed under a Creative Commons Attribution 3.0 Spain License. To view a copy of this license, visit <http://creativecommons.org/licenses/by/3.0/es/>.

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

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

Final publishable summary report.

Contents

1.	Executive Summary	3
2.	Project context and objectives	4
3.	Results and Achievements	6
3.1.	The PANACEA platform	6
3.2.	Technologies for Language Resources Production	9
3.3.	PANACEA Catalogue of Resources	12
4.	Dissemination activities, exploitation and potential impact	15
5.1	Dissemination Activities	16
5.2	Exploitation plan	19
6.	PANACEA Impact	20
7.	Contact and further information	22
8.	Use and dissemination of foreground	23
9.	Report on societal implications	35



The PANACEA Team

1. Executive Summary

A strategic challenge for Europe in today's globalised economy is to overcome language barriers through technological means. In particular, Machine Translation (MT) systems are expected to have a significant impact on the management of multilingualism in Europe, making it possible to translate the huge quantity of textual data produced, and thus, covering the needs of hundreds of millions of citizens. PANACEA addressed a critical thread to this vision: the so-called, language-resource bottleneck. Although MT technologies may consist of language independent engines, they highly depend on the availability of language-dependent knowledge for their real-life implementation, i.e., they require Language Resources (LRs). In order to equip MT for every pair of European languages, for every domain, and for every text genre, appropriate LRs covering every language, domain and genre must be produced. Moreover, a Language Resource for a given language can never be considered complete or final. Language change and new knowledge domains emerge at rapid pace. Traditionally, LRs production is done by hand, and its high cost (highly skilled human work and development time) hindered full coverage. A company willing to cover the enlarged Union market needs to produce and maintain 500 bilingual glossaries, for instance.

PANACEA project has focused on the development of a factory of LRs that automates the stages involved in the acquisition, production, updating and maintenance of LRs required by MT systems, and by other based on Language Technologies (LT) applications. This automation is meant to cut down costs significantly, in terms of time and human effort. Such reductions are the only way to guarantee a continuous supply of LRs that MT and other Language Technologies may demand in a multilingual Europe. In order to address this objective, PANACEA has worked in (i) the development of a platform, designed as a dedicated factory for the composition of a number of LRs production lines based on combinations of different web services and (ii) the integration of advanced components for the acquisition and normalization of corpora, monolingual and parallel corpora, their alignment; the derivation of bilingual dictionaries out of aligned corpora; and the production of monolingual rich information lexica using corpus based automatic methods.

The PANACEA factory has been thoroughly evaluated within R&D and industrial settings. The platform and the LRs production lines based on advanced technological components have proved the feasibility of the concept. PANACEA's contribution and potential impact has been demonstrated in an industrial evaluation carried out with the adaptation of Machine Translation products to a specific/specialized domain. In terms of effort, to produce a domain-adapted bilingual glossary of 1000 entries with PANACEA reduces costs from 30 person/hours to 0.5 person/hours. In terms of quality, there were no significant negative effects in the translation quality of the systems using automatically produced resources. A human evaluation showed that PANACEA domain-tuned SMT gained in quality up to a 6% with respect to the not tuned baseline, and that quality was not significantly worse than the achieved by other state-of-the-art systems as Google Translator.

The successful PANACEA results will be sustained by PANACEA partners who intent to exploit them with a business model based on the production of new resources on demand, mainly. Nevertheless, platform exploitation by third parties will also be possible for academic or industrial research purposes with no cost in an attempt to gain visibility and credibility.

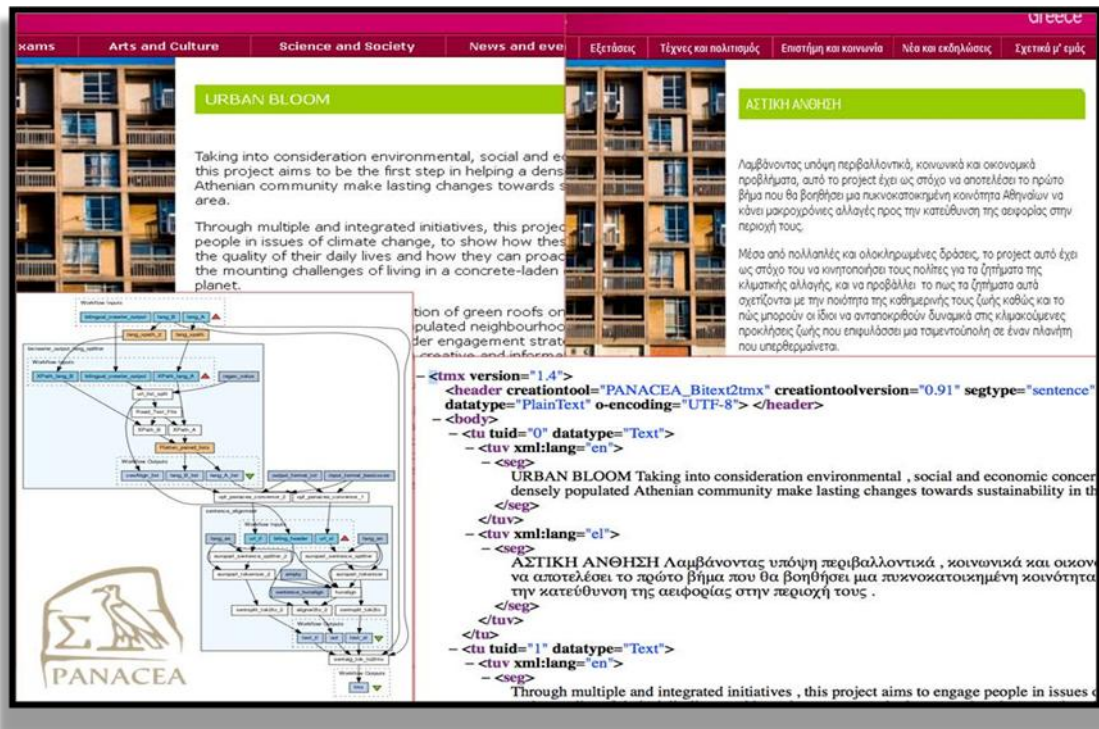
The first objective of the PANACEA project was the use of innovative technologies in the production of LR for the industry of language technologies to break through the problems related to the shortage of language data. Although the industry seems to use some techniques to speed up the creation of the required resources, there is currently no specific framework for this purpose and each industrial must develop their own tools, which are typically more heuristic than scientific. PANACEA has worked on the following different technologies for automating resource production:

- Acquisition and processing technologies: Crawling, monolingual and bilingual, with language identification, cleaning, pre-processing, PoS tagging and parsing.
- Parallel Corpus Technologies: Alignment techniques, extraction of bilingual dictionaries and transfer rules for lexical selection.
- Lexical Acquisition technologies: Verbal Subcategorization frame induction, Selectional Preferences extraction, Multiword extraction, Lexical-Semantic Classification.

For the functional integration of the required components for any possible chain or production line, PANACEA has set up a platform to facilitate interoperability and usability. PANACEA has taken recent developments in the definition and supply of processing components as web services: software systems designed to support interoperable machine-to-machine interaction over a network. The platform is based on an especially dedicated workflow editor and engine for the composition of different possible systems (our production lines) based on available, although remote, distributed web services. PANACEA's second objective was to demonstrate that such a platform, conceived as an interoperability space based on Common Interfaces and standardized input/output object formats (i.e. Travelling Objects), will provide an open, ready to grow and truly interoperable network of services where different components could be easily integrated. This setup is meant to facilitate the exploitation of advanced components by users not particularly interested in installing and maintaining them. Any component could be easily substituted if required for getting better results, tuning to different domains languages, etc.

For assessing the achievement of both objectives, PANACEA has dealt with the following concrete aspects:

- To assess the benefits of domain tuning. LR production cost reduction would allow more domain tuning and thus better results for some applications such as Machine Translation.
- To assess for each of the different advanced components the performance when handling different languages.
- To assess that using automatically produced resources has no significant impact in the quality of the application results.
- To assess that costs indeed are reduced with respect to currently used production practices.



PANACEA a production line: a TMX corpora from crawled parallel text? YES!

3. Results and Achievements

In what follows we present the results of PANACEA a three years long Research and Development project. In 3.1, we will first present the resulting platform. In section 3.2 we will review the results and achievements of the technologies that have been integrated in the platform.

3.1. The PANACEA platform

PANACEA's platform is an interoperability space that joins together advanced interoperable tools to build a factory of LRs. These tools are offered as web services that carry out specific tasks. By the selection of the appropriate web services, the user is able to chain different production lines that **automate** the stages involved in the acquisition, production, updating and maintenance of the LRs as required by MT and other Language Technologies.

For complementing the usability of the PANACEA platform, tools for easy searching and finding web services and workflows, accessing to its documentation and web client, have been developed using open available tools: [Biocatalogue](#) and [MyExperiment](#). These applications were selected because they offer:

- Easy to navigate and usable webs
- Search mechanism based on metadata and tags
- Categorization system by *language* and *service*
- Direct link to web clients to test web services
- Monitoring system

The PANACEA Registry finally included 156 services classified into the following 22 categories.

Alignment (12), Chunking/Segmentation (4), Corpus Processing (48), Corpus Workbench (2), Crawling (4), Format Conversion (35), Indexing (1), Language Guessing (1), Lexicon/ Terminology Extraction (23),

Machine Translation (10), Management (1), Morphological Tagging (4), Morphosyntactic Tagging (17), Named Entity Recognition (4), Other (5), Querying (3), Statistics Analysis (7), Stemming/Lemmatization (9), Syntactic Tagging (10), Terminology Management (2), Text Mining (1), Tokenization (12).

These services are built upon existing third party free open tools² and PANACEA developed new tools. Web services, which are offered for unrestricted use, have been documented and annotated with metadata and tags for easy discovery and operation. Most web services also offer a web-based interface for facilitating testing.

As for the workflow editor and engine, PANACEA decided to use TAVERNA 2.4 because it provides:

- Remote web service calls
- Workflow design with lists: to process corpora split in numerous files
- Robust workflow design: retries and polling
- Throughput improvement with parallelization

PANACEA is now a distributed and interoperable platform of web services that can be chained to perform complex operations in the form of workflows.

Interoperability by means of the PANACEA Common Interfaces makes it possible to choose and to select different web services without requiring workflow parameter modifications. This feature is, together with the easy to use workflow editor, an asset for the proliferation and sharing of different processing chains. In addition, interoperability is grounded on the use of existing standards for defining web service input and output formats, i.e. Travelling Objects. PANACEA Travelling Object formats are XCES, GrAF, CoNLL and LMF. PANACEA has implemented converters to reduce the bulk of input/output format problem, a still outstanding interoperability bottleneck. On the one hand, converters were required because of the different formats consumed by third-party tools, as it was not feasible to re-implement the adopted tools. On the other hand, a conversion strategy also fits the general platform strategy of easy integration: by defining PANACEA travelling objects new components could be easily integrated by only writing a N to 1 and 1 to N converters.

Last but not least, PANACEA platform has been tested with respect to scalability issues: it is meant to process large quantities of data. In a specific study (cf. D3.4 “Large data” section) different experiments have shown the capabilities of the platform to process 20,000 files in different workflow configurations. Limitations found are mainly due to external factors: the dependency on the network and possible interruptions, a risk that increases when processing large files that take more time, and the dependency on server dimensions which have a limited amount of RAM, hard-drive and number of processors. Hard-drive restrictions affect the temporal files allocation, for instance. In the actual configuration, a 4CPUs and 8 GB RAM server accepted up to 100 parallel requests, being these of concurrent users or a user executing a workflow based in parallel requests to reduce the execution time. For instance, in a parallelization experiment a particular task could be reduced from 5.4 to 2.2 hours. The study of these restrictions also gave information about feasible future developments that, beyond the use of Cloud resources, could also foster big data processing (cf. WP3 Final Report).

At the end of the project, [PANACEA myExperiment](#) has about [71 workflows](#) combining the different web services for the production of language resources. Workflows are accessible to be shared by users.

Further interesting characteristics of PANACEA web service platform are:

- An automatic temporal files management system

² These include free tools such as Hunalign, FreeLing, MALT parser, DESR parser, MOSES, etc.

- Multiple protocol implementations with SOAP: Axis1 + JAX-WS
- Direct data or URL interface for inputs and outputs
- Taverna, Python and Perl clients
- Polling for long lasting tasks (no timeouts).



Figure 2: myPanaceaExperiment the Social Platform to share NLP workflows

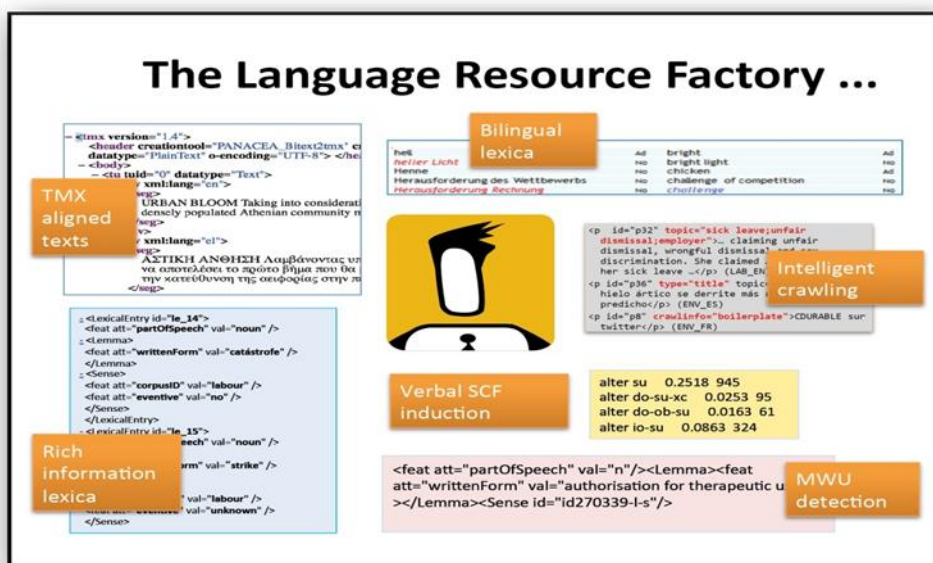
To guide and facilitate user experience, a total of 10 videos are made available at <http://panacea-lr.eu/en/info-for-professionals/tutorials/> offering several introductory scenarios:

- [PANACEA Registry](#)
- [PANACEA myExperiment](#)
- [PANACEA Find and run a workflow](#)
- [PANACEA Building a workflow from scratch](#)
- [Taverna correctly annotating workflows](#)
- [Taverna and lists](#)
- [PANACEA Bilingual Crawler](#)
- [PANACEA Part of Speech Tagging](#)
- [How to register workflows in the PANACEA myExperiment](#)
- [How to register Soaplab web services in the PANACEA Registry](#)

In addition, the following documentation and [tutorials](#) are provided to the user at the web page:

- General PANACEA tutorial: [PANACEA-tutorial_v3.0.pdf](#)
- Specific PANACEA tutorials: [PANACEA-Soaplab-tutorial_v3.0.pdf](#) and [PANACEATaverna-tutorial_v3.0.pdf](#)

- Documentation Index: [PANACEA-Platform documentation_index_v3.0.pdf](#)



PANACEA's Language Resource Factory main results

3.2. Technologies for Language Resources Production

For the production of Language Resources, PANACEA has deployed a set of interoperable Corpus Acquisition, Normalization and NLP web services which are the basis for the production of Language Resources.

Remarkable achievements in PANACEA are related to the development of new modules:

A **Focused Monolingual Crawler (FMC)** has been implemented that adopts a distributed computing architecture based on Bixo, an open source web mining toolkit that runs on top of Hadoop (a well-known framework for distributed data processing). In addition, Bixo also depends on the Heritrix web crawler and makes use of ideas developed in the Nutch Project. This implementation showed the required scalability of the approach which has been demonstrated by being able to improve 46 times (compared to the data collected with the initial version of FMC developed during the first development cycle) the number of tokens acquired for different languages and topics, ranging from 21 M tokens of Greek Labour regulation topics to 70 M tokens of Italian text for Labour regulation topics. This shows the large differences among languages with respect to contents on the web. The FMC running on a single machine typically processes 40 URLs per minute, with processing time including the assessment of a relevance score for each link included in a visited web page and a boilerplate cleaning step. More information about the performance of FMC is provided in [PANACEA deliverable D7.3](#).

A **Focused Bilingual Crawler** has been implemented too. The Focused Bilingual Crawler (FBC) integrates the Focused Monolingual and a module for detecting pairs of parallel documents from domain-specific collections acquired from the web. The FMC and FBC tools are also available as an open-source Java project named [ILSP Focused Crawler](#) (ilsp-fc).

Focused crawling was used to collect monolingual and bilingual corpora to demonstrate how a number of web services chained in workflows could be applied for the production of aligned corpora and later bilingual dictionaries and transfer selection data. For instance, domain-specific SMT systems with crawled corpora

have been built to show significant improvements in quality with respect to a non-domain tuned baseline. For further information refer to Pecina et al. (2011) [Towards using web-crawled data for domain adaptation in statistical machine translation](#) which introduces the approach in detail and presents first results and Pecina et al. (2012) [Simple and Effective Parameter Tuning for Domain Adaptation of Statistical Machine Translation](#) which explores in depth the use of domain-specific crawled data for tuning SMT systems. An SMT can gain approximately a 50% relative improvement of BLUE if trained with domain tuned crawled data that has been automatically cleaned (boiler plate removal, segmentation and tokenization) and aligned.

Two different modules for bilingual dictionaries extraction out of phrase tables have been developed in PANACEA: LT-BiLex-Extract and DCU-P2G (cf. D5.4 and D5.5). These tools deployed as web services, were used in different workflows (available in PANACEA myExperiment) to extract bilingual glossaries from crawled data MOSES phrase tables with an error³ average of 9.26%. These results were presented and published in Thurmair, and Aleksić, [Creating term and lexicon entries from phrase tables](#). (European Association for Machine Translation: [EAMT 2012](#)).

Tools developed in PANACEA also cover those for building monolingual rich dictionaries out of crawled monolingual corpora. The area of Lexical Information Acquisition is still a research topic and PANACEA encouraging results showed that they can be applied in real situations. Evaluation in WP7 showed that these modules can reduce the amount of human work in more than a 40%. Thus, for Subcategorization Acquisition Components for English, Spanish, and Italian, the evaluation experiments carried on showed that Inductive Subcategorization Acquisition can provide with accurate results if confidence thresholds (MLE, Maximum Likelihood Estimation) are used to separate error. However, as shown in Table 1, recall results are still unsatisfactory and some solutions for future work have been identified (cf. D6.2).

Thresholds	Precision [%]	Recall [%]	F-score [%]
MLE EN - 0.04	67.4	66.4	66.9
MLE ES - 0.1	84.7	40.6	54.9
MLE 0.008 + PVF 2.5% IT	65.3	55.7	60.1

Table 1: Indicative evaluation assessment for the three languages

The Multiword Extraction Component for Italian (cf. D6.2) demonstrated the possibility of extracting collocations and other MWU from Environment crawled corpora with a precision of 81% and a recall of 60%. More details in Quochi et al. (2012) A MWE Acquisition and Lexicon Builder Web Service, published in COLING 2012.

Lexical semantic class information induced from crawled corpus shows also encouraging results when used with confidence thresholds as it allows saving more than 30% of manual work average in the compilation of this information (cf. D6.2).

³ Errors correspond to: (i) translation errors, i.e. the candidates are not translations of each other and (ii) lemmatization and annotation errors (cf. D7.4).

Class	Acc. (%)	Using confidence threshold	
		Acc. (%)	To be revised (%)
HUM	77.29	91.47	68.27
LOC	77.55	89.08	68.73
EVENT	80.90	92.85	66.33
ABSTRACT	73.77	79.90	41.66
PROCESS	78.45	85.42	52.24
ARTIFACT	72.16	80.85	70.63
MATTER	79.33	89.13	41.02
SEMIOTIC	75.09	83.52	67.62

Table 2: Lexical-classifiers results for Spanish, including accuracy and the assessment of entries to be revised.

Other interesting results of the lexical acquisition components can be summarized as follows:

- Automatic lexical analysis to discover OOV items (especially important for German and other languages with frequent compounding) for domain tuning can be done with crawled texts obtaining an annotated morphosyntactic lexicon with an error of a 2% (cf. D4.4).
- Lexical entries coming from different lexica and format, extracted by different components and containing different information, can be automatically merged with a PANACEA developed merging component (cf. D6.5). In an evaluation experiment, 3 different lexica one with morphological, another with verbal SCFs and another with lexical semantic classes information, each entry encoding additional domain information specialization for Labour and Environment was merged. The resulting lexicon, delivered in LMF ISO standards, has a total of 110,316 entries and a potential error estimation of a 3%.

Table 3 sums up the production components developed in PANACEA and their licenses and web service usage conditions (cf. D6.4).

Name	Partner	Program license	Web service Terms & conditions of use
Focused Crawler	ILSP	open-source Java project GPLv3.0	Use for Research purpose
Bilingual dictionary generator P2G	DCU	GPLv3.0	Use for Research purpose
LSF classifiers	UPF	GPL v3.0	No usage restrictions
Inductive subcategorization frame acquisition	UC	GPLv3.0	No usage restrictions

Name	Partner	Program license	Web service Terms & conditions of use
LMF_ML_MERGER	CNR	GPLv3.0	No usage restrictions
lmf_merger	CNR	GPLv3.0	No usage restrictions
Extractor_MW	CNR	GPLv3.0	No usage restrictions
SCF_Extractor_lang_indip	CNR	GPLv3.0	No usage restrictions
SCF_Extractor_IT	CNR	GPLv3.0	No usage restrictions
BiLex-Extract	LT	Apache 2	Use for Research purpose
Decomposer	LT	CC-BY-SA-NC	Use for Research purpose
Tagger	LT	CC-BY-SA-NC	Use for Research purpose

3.3. PANACEA Catalogue of Resources

The different Language Resources produced during the project, the actual test data from acquired corpora to derived lexica have been made public at the [PANACEA](#) web page, ELRA catalogue and at the [META-SHARE](#) platform. The gold standards for intrinsic evaluation of some of the tools have also been made available. Appropriate licences for all of them have been defined. Resources are archived in persistent repositories (UPF e-repository and ELDA's servers) in order to guarantee their persistence. These resources are:

- **Monolingual Corpora raw text** for two domains Environment (ENV) and Labour Legislation (LAB) and for Greek, English, Spanish, French and Italian. The size of the produced Monolingual Corpora ranges from 13K to 28K web pages (26M to 70M tokens) depending on the selected domain (ENV or LAB) and the targeted language (EL, EN, ES, FR, IT). The only exception concerns the Greek data in the LAB domain, where only ~7K web pages were acquired. However, this collection amounts to ~21M tokens, since it consists mainly of large legal documents or lengthy discussions/arguments about Labour Legislation. Detailed information about these corpora can be found in D7.3. The size of the monolingual corpora (excluding tokens from short paragraphs or paragraphs detected as boilerplate) is:

Domain/language	# of documents	# of web sites	# of tokens
ENV_EL	16,073	1,063	27,958,530
ENV_EN	28,071	3,121	50,541,538
ENV_ES	26,009	2,053	46,225,624
ENV_FR	23,514	1,969	47,364,125
ENV_IT	16,159	1,211	40,044,852
LAB_EL	7,124	598	21,077,196

Domain/language	# of documents	# of web sites	# of tokens
LAB_EN	15,197	1,558	46,431,351
LAB_ES	13,188	1,015	53,922,118
LAB_FR	26,675	1,391	56,440,425
LAB_IT	12,706	864	70,563,320

Table 4: Quantitative information for Monolingual Corpus v2

- **Monolingual Corpora n-grams** based on the annotated ENV and LAB monolingual corpora for EL, EN, ES, FR and IT. Word and word/tag/lemma n-grams accompanied by their observed frequency counts have been generated. The n-grams' length ranges from unigrams to five-grams. The size of the monolingual n-grams corpora is as follows:

RESOURCE (CC BY SA Licence)	EL		EN		ES		IT		FR	
	ENV	LAB	ENV	LAB	ENV	LAB	ENV	LAB	ENV	LAB
Tokens	31,713	24,069	46,553	45,134	49,860	58,067	35,998	70,438	42,780	46,992
Sentences	1,185	948	1,700	1,407	1,882	1,969	525	1,175	1,235	1,232
1-grams	435	363	910	606	652	582	459	547	824	664
2-grams	3,860	3,104	17,548	12,561	4,829	4,856	4,109	5,247	13,738	11,776
3-grams	9,767	7,725	75,302	57,544	13,777	13,993	11,773	16,698	60,196	54,793
4-grams	13,683	10,650	148,085	120,877	21,883	22,815	17,243	26,014	127,183	121,160
5-grams	14,954	11,513	207,570	177,588	25,791	27,262	19,268	29,778	187,590	185,045

Table 5: Data sizes (in thousand units) for the n-grams generated from PANACEA monolingual data

- **Monolingual Dependency Parsed Corpora** for two domains ENV and LAB and three languages EL, ES and IT. Text has been automatically pre-processed, and annotated with PoS and lemma and dependency relations. Further details can be found in D4.4 report and in the associated readme files.

RESOURCE (CC BY NC SA)	TOKENS	SENTENCES
Annotated Corpus ENV EL	34,668,532	1,554,313
Annotated Corpus LAB EL	26,046,343	1,255,073
Annotated Corpus ENV ES	30,972,685	1,367,518
Annotated Corpus LAB ES	60,956,181	2,388,141
Annotated Corpus ENV IT	36,000,000	1,431,914
Annotated Corpus LAB IT	70,000,000	2,975,818

Table 6: Data sizes for Dependency Parsed Corpora

- **Monolingual Lexica** for two domains ENV and LAB. Automatic acquired information is Verb Subcategorization frames for EN, ES and IT; Lexical semantic classes for ES and EN nouns, and a merged Lexicon with Subcategorization and Lexical semantic classes for EN and ES. Further details about these lexica can be found in D6.2 and D6.3 reports and in the associated readme files.

RESOURCE (CC BY Licence)	ENTRIES
Monolingual Lexicon V-SUBCAT ENV ES	1,543
Monolingual Lexicon V-SUBCAT LAB ES	1,015

RESOURCE (CC BY Licence)	ENTRIES
Monolingual Lexicon V-SUBCAT ENV IT	26
Monolingual Lexicon V-SUBCAT LAB IT	27
Monolingual Lexicon V-SUBCAT ENV EN	895
Monolingual Lexicon V-SUBCAT LAB EN	1,063
Monolingual Lexicon lexical sem classes ENV ES	4,199
Monolingual Lexicon lexical sem classes LAB ES	5,037
Monolingual Lexicon lexical sem classes ENV EN	3,641
Monolingual Lexicon lexical sem classes LAB EN	3,762
Monolingual Subcat & Lex Sem classes ENV ES	5,742
Monolingual Subcat & Lex Sem classes LAB ES	6,052
Monolingual Subcat & Lex Sem classes ENV EN	4,704
Monolingual Subcat & Lex Sem classes LAB EN	4,657
Multilevel, multi-domain lexica ES	110,316

Table 7: Data sizes for Monolingual Lexica

- **Monolingual Lexica Gold Standard** for three domains: General, Environment and Labour with Verb Subcategorization and Lexical semantic classes for Spanish; Environment and Labour with Verb Subcategorization for Italian and Lexical semantic classes for English. Further details about these lexica can be found in D7.4 report and in the associated readme files.

RESOURCE (CC-BY-SA Licence)	ENTRIES	SENTENCES
Monolingual Gold Standard: V-SUBCAT ENV EN	28	200
Monolingual Gold Standard: V-SUBCAT LAB EN	29	200
Monolingual Gold Standard: V-SUBCAT ENV ES	30	200
Monolingual Gold Standard: V-SUBCAT LAB ES	30	200
Monolingual Gold Standard: V-SUBCAT OPEN IT	30	200
Monolingual Gold Standard: V-SUBCAT LAB IT	30	200
Monolingual Gold Standard: Lex Sem Class GRAL ES	5,068	Na.

Table 8: Data sizes for Monolingual Lexica Gold Standards

- **Monolingual Multiword Units** for Italian. The ENV MW is a lexicon of nominal multiword expressions automatically extracted from a 37Mio words newspaper corpus. The LAB MW is a lexicon of nominal multiword expressions automatically extracted from a 66Mio words newspaper corpus. The former contains 14,109 and the latter 15,332 Lexical Entries, of which 10,000 in each have entryType = Multiword. Further details about these lexica can be found in D6.2 and D6.3 reports and in the associated readme files.

RESOURCE (CC-BY-NC Licence)	ENTRIES	MWU
Monolingual MultiWord Units ENV IT	15,491	10,000
Monolingual MultiWord Units LAB IT	15,279	10,000

Table 9: Data sizes for IT Multiword Units

- **Bilingual Aligned Parallel Corpora.** The initial version of the Focused Bilingual Crawler (FBC) was used to construct domain-specific parallel corpora (aligned on sentence level) in the EN-FR and EN-EL language combinations for Environment and Labour Legislation domains. Table 11 provides the size in aligned sentences, after filtering. Details about these collections are reported in PANACEA deliverable [D5.3](#) and in the associated readme files.

RESOURCE (CC-BY Licence)	EN-FR	EN-EL
Bilingual Aligned Parallel Corpus ENV	13,840	13,253
Bilingual Aligned Parallel Corpus LAB	23,861	9,764

Table 10: Data sizes (sentences) for Bilingual Aligned Parallel Corpora

- **Bilingual Glossaries.** From the already mentioned parallel corpora, bilingual glossaries were automatically produced. Details about these collections are reported in PANACEA deliverable D5.4 and D5.7 and in the associated readme files.

RESOURCE (CC-BY Licence)	ENTRIES
Bilingual Glossary ENV EL-EN	18,584
Bilingual Glossary LAB EL-EN	13,220
Bilingual Glossary ENV FR-EN	17,690
Bilingual Glossary LAB FR-EN	27,798

Table 11: Data sizes for Bilingual Glossaries

4. Dissemination activities, exploitation and potential impact

Dissemination activities together with the study of the legal issues that could affect the exploitation of PANACEA results were the object of WP2. A first Dissemination plan (D2.2) was put in motion already at the very beginning of the project. The main goals were:

- (i) sharing the experiences that the project would bring about,
- (ii) promoting the need for research and development in the field of Automatic LR acquisition and
- (iii) demonstrating the existence and benefits of using the new developed technologies in the framework of PANACEA.

An addendum to the Dissemination Plan was produced and delivered 13 of December 2011. It mainly addressed an update of the dissemination activities according to the event agenda in 2012, as well as the plans for disseminating already available results. Following reviewers recommendations, more effort and resources were spent in 2012 for dissemination activities carrying out specific actions addressed to professional profiles

of possible users: PANACEA participated in professional exhibitions and an introductory video and different tutorial materials were produced with professional assistance in order to better communicate with them.

During the project, PANACEA has produced different dissemination materials. A professionally designed web site was made available at the very beginning of the project offering different profiled sections in order to assist in finding relevant information to the different aimed at audiences. The web site was the main communication channel: news have been periodically issued reporting on achievements, dissemination activities, publications, etc. In average, more than one new per month has been issued. Deliverables and incoming results (web services available, and workflows and LR available) have been advertised. Analytics of web site helped to assess the impact of PANACEA dissemination activities, for instance:

1. PANACEA was presented at the 15th Annual Conference of the European Association for Machine Translation, which took place on 30-31 May, 2011 in Leuven, Belgium. The presentation, made by DCU, addressed specifically MT developers and presented PANACEA developments. [143 visits]
2. PANACEA was an exhibitor of the META-FORUM 2011 held in Budapest, 27 and 28 of June. The PANACEA users workshop was held immediately after the 29 of June as a satellite event. [117 visits]
3. The PANACEA platform has been presented at the Workshop on Language Resources, Technology and Services in the Sharing Paradigm, 12 November 2011, at IJCNLP 2011 (Chiang Mai, Thailand). [111 visits]
4. “EC Village”, “EU Projects Track” and different presentations held on May 2012 at the 8th International Conference on Language Resources and Evaluation, LREC 2012. [156 visits]
5. PANACEA was exhibitor at the 2012 [Localization World](#), (Paris 4-6 June) a conference and networking organization dedicated to the language and *localization* industries. [100 visits].

Other dissemination materials include flyers for being distributed in different events, posters, presentations and an introductory video. Most of these materials were produced with professional support. All these materials can be found in the PANACEA web site.

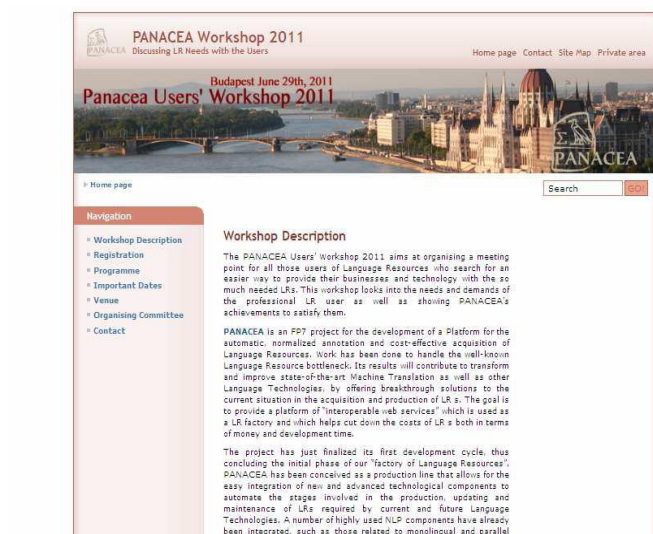
5.1 Dissemination Activities

Dissemination activities were planned and executed taking into account the different target audiences: Researchers, Professionals and General Audience, although we mainly concentrated in the two first groups.

For PANACEA, academic communication has been very important because the technologies addressed are mainstream research areas, and also because the need to become convincing about the feasibility and reliability of the LRs production technologies in development. In order to find interested researches, two scientific workshops were organized collocated with two different editions of the International Conference on Language Resources and Evaluation (LREC 2010 and LREC 2012). The topics were “**Methods for the automatic acquisition of Language Resources and their evaluation methods**” and “Merging of Language Resources”. In both cases, high participation and interest was achieved. Workshop Proceedings were published in both cases.

The PANACEA internal policy of promoting scientific articles as the documentation of the project developments has been very positive resulting in more than 50 specialized publications (mainly in International Conference proceedings including, ACL, EACL, EAMT, COLING, LREC). This scientific communication effort gave already fruits with more than 100 citations (according Google Scholar, in

September 2012) of PANACEA related published papers. This impact is indeed a validation of the soundness of PANACEA results by the scientific community and contributes to the goal of being convincing about PANACEA approach feasibility and appropriateness. We expect more citations in the coming years. A complete list of PANACEA related publications is available at the web site⁴.



User's workshop web page and opening session

Another important PANACEA dissemination objective was to attract the professional audience. For that purpose, a specific user's workshop for presenting first available results was planned and took place in Budapest on June 29, 2011, as a satellite event of the META-FORUM 2011. The choice for this collocated event was indeed strategic. META-FORUM is an event organized by another FP7 project (T4ME/META-NET) that shares a strong interest for the industrial world with PANACEA. This seemed very interesting in order to attract potential users' interest and encourage them to join us for the workshop. Targeting two events was certainly a plus for some of the attendees. Besides the 2 invited speakers and 11 project members (speakers included) 21 other people registered for the workshop. It should be said that both the fact of being announced within the META-Forum program and holding a demo during the Forum helped us get in touch with some further people, such as members of newly funded projects. The participants' background can be divided as follows: Industry, 10, out of which 1 was a journalist and Academia, 11.

Also to attract Language Technology Industries' attention and because of the special interest shown by a number of professionals, PANACEA made available a platform beta version together with beta-testing tutorial material and call for the participation in a beta-testing exercise during 2012. Worth mentioning here is Qualia, a business intelligence studio based in Athens, who reported to be interested in commercially applying the software after the end of the project. Qualia tracks and measures the millions of on-line conversations on web sites, blogs, microblogs, social networks, social news and forums and also provides near real-time search on television and radio content. The company's advanced technologies analyze and extract knowledge and actionable insights from the wealth of data, so that clients can detect threats and opportunities, develop more creative strategies, understand what is being said and decide what they should do next. Qualia used the Focused Monolingual Crawler in order to discover web sources on male skincare products and they concluded that FMC is very useful software that can be valuable for Qualia's intelligence and analytics services. Already available LR's also attracted the attention of the industry and the demand for licensing data sets used in the project already started in 2011.

⁴ <http://www.panacea-lr.eu/en/project/publications/>

PANACEA has participated in the following events in the period 2010-2012

Events	Presentations	Demonstrators/ Exhibitors
EC Related/Promoted activities (LT-days 2010; META-FORUM 2010, 2011, 2012; LREC EC-Village 2010, 2012; EAMT 2010)	3	5
International Research & Professional Conferences (LREC 2010, 2012; EAMT 2011, 2012; IJCNLP 2011; COLING 2010, 2012; BUCC 2012)	4	2
Professional Conferences (Localization World 2012)		1
National Research & Professional Events (SEPLN 2010; Innovation Dublin Festival 2010; UPF Industry Day 2011; CNGL Localisation innovation showcase 2011; International Conference on Greek Linguistics 2011)	3	3
TOTAL	10	11

Table 12: Summary of Dissemination activities

Also, as envisaged in the Dissemination Plan, PANACEA has established a liaison with other EU Projects with the objective of taking maximum advantage of collaboration between projects.

PANACEA has kept promoting the collaboration with other projects with related objectives, both scientifically and strategically: ACCURAT, LET's MT, TTC, ACCURAT, KYOTO, MONET, MEDAR and BOLOGNA. In particular, the following activities have taken place:

- META-SHARE. PANACEA has made its resources available at the [META-SHARE](#) platform. All data sets used during the project, and eventually the results of the factory, will be accessible from this distribution platform.
- UCOMPARE, the platform developed by the Centre of the National Center for Text Mining and made available by the [METANET4U project](#), is able to run PANACEA web services from their platform.
- Special contacts with the EUA-NSF project "The Language Application Grid" (Nancy Ide and James Pustejovsky) are maintained in order to transfer knowledge acquired and lessons learnt to this project which is meant to build a web service based infrastructure for Language Applications in the USA.
- After using [MyGrid Project](#) tools, close relations have been established with developers in both sites. PANACEA is advertised as one [TAVERNA](#) user at MyGrid web pages.

- Collaboration with the CNGL - Centre for Next Generation Localisation (Ireland). The existing CMS-L10n architecture has been used as an interface between content creators and content translators, post-editors and MT deployments. Panacea has been integrated into the architecture allowing post-edits, performed on previously MT'd content, to be re-incorporated into the MT engine, re-training the MT engines, based on the post-edits made on its previous translations. Panacea provided a platform for the management/automatic re-training of domain-bespoke MT engines using post-edits.
- PANACEA started contacts with the LT-INNOVATION team in order to use this ambitious initiative as a dissemination channel.

During the last year, PANACEA has specially monitored the different ongoing initiatives working with Web Services and Workflows. In particular, PANACEA was invited to join the initiative of creating an ISO international working group on Web Service Exchange Protocols. The main concern of this group is to go a step further and to define in addition to syntactic interoperability protocols, semantic interoperability. PANACEA has attended two meetings of this group and presented its results as input for the future group task definition. UPF has shown its interest in participating in this working group. Other participants are: Language Grid (University of Kyoto); LinguaGrid (CELI), EXCITEMENT EU Project (FBK); Australian National Corpus Project; NLP Interchange Format and INTEROP project (Brandeis University).



PANACEA has been presented at different international conferences

5.2 Exploitation plan

PANACEA D2.4 has analysed the exploitation of the PANACEA project assets. These assets have been clustered into a few items (a) the PANACEA Factory/Platform, (b) the web services integrated within the platform, (c) the associated workflows to manage the sequencing of web services (d) the tools developed during the project and last but not least (e) the data sets, i.e. Language Resources (LRs) produced mainly for Machine translation/localization, but not only, within the project, exploiting the platform, web services, and the workflows.

Two major aspects have been addressed: (i) the technical (and hence market) value and (ii) the implied legal issues, both to ensure that PANACEA results exploitation is done in a cleared legal framework as well as to ensure that such assets can be licensed to third parties under a clean, easy to understand and implement,

licensing schema. In order to do so, we also describe the potential users and their needs, leading to the design of our exploitation plan.

PANACEA uses tools implemented by third parties, who are licensed under various particular licenses: mostly GPL v3 or BSD. PANACEA has confirmed that none of them had restrictions for its use in the PANACEA platform. All PANACEA partner's tools are offered by their owners as public services, accessible to all too and only few of them include restrictions for uses with "non-commercial purposes". Usage conditions are mentioned at the PANACEA registry for every web service.

The consortium agreed to ensure that the assets of the PANACEA project are maintained for at least 24 months after the end of the project. This includes a commitment for keeping the whole platform in place and running (the platform including Registry, myExperiment currently located at ELDA, the workflow engine and the existing workflows and related web services, currently located in UPF, CNR-ILC, ILSP, UC, DCU, LT and ELDA). Such agreement will help better assess the future. The Language Resources developed within the project will be exploited as part of the ELRA catalogue but also using the META-SHARE infrastructure (in addition to the UPF e-repository).

The business model envisaged in the exploitation plan is based mostly on exploiting the platform to produce new resources on demand, design workflows on demand, connect to third parties new web services, and also to make available the platform for internal use by commercial organizations. Resources produced by the Project will be made freely available for research purposes. Although, this will not generate revenues, it is expected that such use will make PANACEA (and the factory that produced them) more visible and may be trigger more commercial projects for similar productions.

6. PANACEA Impact

PANACEA proposal mentioned a number of potential impacts. We now review them as to assess the impact that the project has indeed brought already about, as well as the potential for future impact.

PANACEA has contributed to demonstrate that the "LR bottleneck" problem can be effectively addressed by automation. Automatic production lowers the cost and time required for producing basic LR for languages which are currently not well covered. For instance, the META-NET White Paper Series reports "Europe's Languages in the Digital Age" identified in its section "Key Results and Cross-Language Comparison" that LRs for all languages addressed by PANACEA, but for English, have a moderate support which PANACEA approach might have help to normalize, in particular for Greek that suffers from a "fragmentary" support.

Automatic production of BLARK resources can be addressed now with the PANACEA factory producing all the non-multimodal resources identified as critical for LT applications, that is: unannotated corpora, annotated corpora, aligned parallel corpora, monolingual lexica and bilingual lexica. Annotated corpora produced in PANACEA include PoS tagging and dependency parsing. Monolingual lexica include rich information such as verbal subcategorization frame information, multiword identification and lexical semantic information for nouns. Automatic production of resources has been demonstrated to achieve a quality level that, for resources such as aligned corpora and bilingual dictionaries for Machine Translation, do not affect significantly the quality of the results (cf. D8.3). The PANACEA factory approach will have a very important impact for the

proliferation of resources also domain tuned, achieving thus a better results of applications such as MT, but not only.

As for quantities, PANACEA has shown the scalability of the technical solutions proposed and has made evident where the limitations are (mainly related to network reliability and hardware configurations, cf. WP3 Final Report). Possible solutions to be studied in the near future have also been pointed to. The problem analysis and the drafted solutions will also impact the current situation by providing sound guidelines that are expected to speed up future developments.

Some of them have been already acknowledged by the META-NET (the European Network of Excellence that supports META the Multilingual Europe Technology Alliance) Strategic Research Agenda (SRA) where more investment in the automatic production of resources technologies is considered a priority. This SRA acknowledgment can be considered already an impact of PANACEA successful results. Hopefully, this agenda will lead the community towards broader societal impacts including the solution to the identified language barriers which are hindering the free flow of information, goods, knowledge, thought and innovation.

LT cannot be made independent of the available LRs for each language and thus are crucial for making Europeans able to communicate with one another, with their governments and with web services in their native mother tongue.

On the more technical part, the PANACEA successful approach based on web services that easy the operations related to the production of resources has contributed to the raising of interest in the community. For instance, ISO TC37 is now addressing the standardization of the protocols for language technology related web services in a new working group. This is indeed a contribution to the expected impact of PANACEA that aimed at contributing to creating an interoperability space that fosters the use and re-use of different tools and resources maximizing thus their revenue and saving the community from duplicated work.



7. Contact and further information






For more information, please visit PANACEA web site at www.panacea-lr.eu.



The screenshot shows the PANACEA website interface. At the top, there is a navigation bar with the PANACEA logo, a search bar, and links for 'Contact Us' and 'Members Login'. Below the navigation bar, there is a main content area with several sections:

- Welcome to PANACEA:** A paragraph describing the project as a STREP Project under EU-FP7, developed a factory of Language Resources (LRs) in the form of a production line that automates all steps involved in the acquisition, production, maintenance and updating of the LRs required by Machine Translation and other Language Technologies. It also mentions that the factory is a Web Service-based platform that integrates advanced technological components for:
 - Monolingual and Parallel Text Acquisition and Pre-Processing
 - Parallel corpora Alignment
 - Bilingual Dictionary Production
 - Monolingual Rich Information Lexica Production
- News:** A section with two news items:
 - 8 PANACEA Papers accepted at COLING 2012:** A list of accepted papers at the forthcoming 24th International Conference on Computational Linguistics COLING 2012, December Mumbai, India. (23 Nov 2012)
 - Progress regarding PANACEA articles' citations:** A link to refer to PANACEA's Publications. (10 Sep 2012)
 - Proceedings of PANACEA's LR Merging Workshop now available:** Proceedings of PANACEA's Successful Workshop held in May, 2012 in Istanbul, Turkey. (05 Sep 2012)
- PANACEA Web Portals:** A section with four icons: Registry, MyExperiment, Tutorials, and Documentation.
- PANACEA ? Watch the video...:** A section with a video player.
- Latest Services Available:** A list of services with their timestamps:
 - delic4mt 2013-01-29, 06:49 am
 - LMF_ML_MERGER 2012-12-11, 05:46 am
 - post_filter_reorder 2012-11-28, 10:44 am
 - filter_averagefreq_on_mwe 2012-11-28, 10:44 am
 - extractor_mw 2012-11-28, 10:44 am
 - converter_mwe_tsv_2_lmf 2012-11-28, 10:44 am
 - post_filter_remove_substring 2012-11-28, 10:44 am
 - post_filter_remove_stopwords 2012-11-28, 10:44 am
 - post_filter_merge_overlapping_strings 2012-11-28, 10:44 am
 - treeagger2to 2013-01-29, 06:49 am

<p>Núria Bel & Marc Poch</p> <p>Universitat Pompeu Fabra – UPF</p> <p>info@panacea-lr.eu</p>	<p>ES</p>	
<p>Nicoletta Calzolari & Valeria Quochi</p> <p>Consiglio Nazionale delle Ricerche - Istituto de Linguistica Computazionale – CNR-ILC</p>	<p>IT</p>	

Stelios Piperidis & Prokopis Prokopidis Institute for Language & Speech Processing - ILSP	GR	
Anna Korhonen & Thierry Poibeau University of Cambridge – UCAM	UK	
Gregor Thurmair , Reinhard Busch Linguattec -- LT	DE	
Antonio Toral Dublin City University -- DCU	IE	
Victoria Arranz & Olivier Hamon Evaluations and Language Resources Distribution Agency – ELDA	FR	

8. Use and dissemination of foreground

▪ Section A

This section should describe the dissemination measures, including any scientific publications relating to foreground. **Its content will be made available in the public domain** thus demonstrating the added-value and positive impact of the project on the European Union.

[LIST of PUBLICATIONS and DISSEMINATION ACTIONS]

▪ Section B

Section A (public)

This section includes two templates

- Template A1: List of all scientific (peer reviewed) publications relating to the foreground of the project.
- Template A2: List of all dissemination activities (publications, conferences, workshops, web sites/applications, press releases, flyers, articles published in the popular press, videos, media briefings, presentations, exhibitions, thesis, interviews, films, TV clips, posters).

These tables are cumulative, which means that they should always show all publications and activities from the beginning until after the end of the project. Updates are possible at any time.

TEMPLATE A1: LIST OF SCIENTIFIC (PEER REVIEWED) PUBLICATIONS, STARTING WITH THE MOST IMPORTANT ONES										
NO.	Title	Main author	Title of the periodical or the series	Number, date or frequency	Publisher	Place of publication	Year of publication	Relevant pages	Permanent identifiers ⁵ (if available)	Is/Will open access ⁶ provided to this publication?
1	<i>Economic transformation in Hungary and Poland'</i>		<i>European Economy</i>	<i>No 43, March 1990</i>	<i>Office for Official Publications of the European Communities</i>	<i>Luxembourg</i>	<i>1990</i>	<i>pp. 151 - 167</i>		yes/no
2										
3										

⁵ A permanent identifier should be a persistent link to the published version full text if open access or abstract if article is pay per view) or to the final manuscript accepted for publication (link to article in repository).

⁶ Open Access is defined as free of charge access for anyone via Internet. Please answer "yes" if the open access to the publication is already established and also if the embargo period for open access is not yet over but you intend to establish open access afterwards.

TEMPLATE A2: LIST OF DISSEMINATION ACTIVITIES

NO.	Type of activities ⁷	Main leader	Title	Date/Period	Place	Type of audience ⁸	Size of audience	Countries addressed
1	Conference		European Conference on Nanotechnologies	26 February 2010				
2								
3								

⁷ A drop down list allows choosing the dissemination activity: publications, conferences, workshops, web, press releases, flyers, articles published in the popular press, videos, media briefings, presentations, exhibitions, thesis, interviews, films, TV clips, posters, Other.

⁸ A drop down list allows choosing the type of public: Scientific Community (higher education, Research), Industry, Civil Society, Policy makers, Medias, Other ('multiple choices' is possible).

Section B (PUBLIC)

Part B1

No applications for patents, trademarks, registered designs, etc. have result of PANACEA.

Part B2

Type of Exploitable Foreground ⁹	Description of exploitable foreground		Foreseen embargo date dd/mm/yyyy	Exploitable product(s) or measure(s)	Sector(s) of application ¹⁰	Timetable, commercial or any other use	Patents or other IPR exploitation (licences)	Owner & Other Beneficiary(s) involved
	<i>Ex: New superconductive Nb-Ti alloy</i>			<i>MRI equipment</i>	<i>1. Medical 2. Industrial inspection</i>	<i>2008 2010</i>	<i>A materials patent is planned for 2006</i>	<i>Beneficiary X (owner) Beneficiary Y, Beneficiary Z, Poss. licensing to equipment manuf. ABC</i>
exploitation of results through (social) innovation	Research prototype for acquiring domain-specific monolingual and bilingual corpora			ILSP Focused Crawler (ILSP-FC)	J63 - Information service activities	2013	open-source Java project. GPLv3.0	ILSP
exploitation of results through (social) innovation	Bilingual dictionary creation from factored phrase tables that include part of speech tagged text for EL-EN and FR-EN language pairs			DCU Bilingual dictionary generator P2G	J63 - Information service activities	2013	GPLv3.0	DCU

¹⁹ A drop down list allows choosing the type of foreground: General advancement of knowledge, Commercial exploitation of R&D results, Exploitation of R&D results via standards, exploitation of results through EU policies, exploitation of results through (social) innovation.

¹⁰ A drop down list allows choosing the type sector (NACE nomenclature) : http://ec.europa.eu/competition/mergers/cases/index/nace_all.html

Type of Exploitable Foreground ⁹	Description of exploitable foreground		Foreseen embargo date dd/mm/yyyy	Exploitable product(s) or measure(s)	Sector(s) of application ¹⁰	Timetable, commercial or any other use	Patents or other IPR exploitation (licences)	Owner & Other Beneficiary(s) involved
exploitation of results through (social) innovation	Bilingual dictionary creation from factored phrase tables that include part of speech tagged text for EL-EN and FR-EN language pairs			DCU Bilingual dictionary generator P2G	J63 - Information service activities	2013	GPLv3.0	DCU
exploitation of results through (social) innovation	They perform the classification of a list of nouns as belonging or not belonging to the given class, for instance EVENTIVE nouns.			UPF LSF classifiers	J63 - Information service activities	2013	GPLv3.0	UPF
exploitation of results through (social) innovation	It takes parsed text as input and produces a subcategorization frame lexicon			UC Inductive subcategorization frame acquisition	J63 - Information service activities	2013	GPLv3.0	UC

Type of Exploitable Foreground ⁹	Description of exploitable foreground		Foreseen embargo date dd/mm/yyyy	Exploitable product(s) or measure(s)	Sector(s) of application ¹⁰	Timetable, commercial or any other use	Patents or other IPR exploitation (licences)	Owner & Other Beneficiary(s) involved
exploitation of results through (social) innovation	Multi-level merger for LMF lexica			CNR LMF_ML_MERGE R	J63 - Information service activities	2013	GPLv3.0	CNR
exploitation of results through (social) innovation	Merges to LMF lexicons for (syntactic) subcategorisation information,			CNR lmf_merger	J63 - Information service activities	2013	GPLv3.0	CNR
exploitation of results through (social) innovation	A language independent tool that implements statistics-based methods for the acquisition of multi-word expressions			CNR Extractor_MW	J63 - Information service activities	2013	GPLv3.0	CNR
exploitation of results through (social) innovation	A language independent tool that implements and inductive method for the acquisition of verb subcategorisation frames			CNR SCF_Extractor_language_indip	J63 - Information service activities	2013	GPLv3.0	CNR

Type of Exploitable Foreground ⁹	Description of exploitable foreground		Foreseen embargo date dd/mm/yyyy	Exploitable product(s) or measure(s)	Sector(s) of application ¹⁰	Timetable, commercial or any other use	Patents or other IPR exploitation (licences)	Owner & Other Beneficiary(s) involved
exploitation of results through (social) innovation	Implements an inductive method for the acquisition of verb subcategorisation frames for Italian			CNR SCF_Extractor_IT	J63 - Information service activities	2013	GPLv3.0	CNR
exploitation of results through (social) innovation	Prototype of a bilingual lexicon extractor from phrase tables			LT-BiLex-Extract	J63 - Information service activities	2013	Apache 2 ¹¹	LT
exploitation of results through (social) innovation	Prototype for decomposing German compounds into their parts			LT-BiLex-Extract	J63 - Information service activities	2013	CC-BY-SA-NC	LT
exploitation of results through (social) innovation	Prototype of a rule-based POS tagger for German			LT-BiLex-Extract	J63 - Information service activities	2013	CC-BY-SA-NC	LT
exploitation of results through (social) innovation	Language Resources: lexica, corpora, etc. (described			Language Resources	J63 - Information service activities	2013	CC-BY-SA-NV	

¹¹ The LT-BiLex-Extract component is already released in the ACCURAT toolbox framework (under the name LT-P2G).

Type of Exploitable Foreground ⁹	Description of exploitable foreground		Foreseen embargo date dd/mm/yyyy	Exploitable product(s) or measure(s)	Sector(s) of application ¹⁰	Timetable, commercial or any other use	Patents or other IPR exploitation (licences)	Owner & Other Beneficiary(s) involved
	below)							

Language resources description

- **Monolingual Corpora raw text** for two domains Environment (ENV) and Labour Legislation (LAB) and for Greek, English, Spanish, French and Italian. The size of the produced Monolingual Corpora ranges from 13K to 28K web pages (26M to 70M tokens) depending on the selected domain (ENV or LAB) and the targeted language (EL, EN, ES, FR, IT). The only exception concerns the Greek data in the LAB domain, where only ~7K web pages were acquired. However, this collection amounts to ~21M tokens, since it consists mainly of large legal documents or lengthy discussions/arguments about Labour Legislation. Detailed information about these corpora can be found in D7.3. The size of the monolingual corpora (excluding tokens from short paragraphs or paragraphs detected as boilerplate) is:

Domain/language	# of documents	# of web sites	# of tokens
ENV_EL	16,073	1,063	27,958,530
ENV_EN	28,071	3,121	50,541,538
ENV_ES	26,009	2,053	46,225,624
ENV_FR	23,514	1,969	47,364,125
ENV_IT	16,159	1,211	40,044,852
LAB_EL	7,124	598	21,077,196
LAB_EN	15,197	1,558	46,431,351
LAB_ES	13,188	1,015	53,922,118
LAB_FR	26,675	1,391	56,440,425
LAB_IT	12,706	864	70,563,320

Table 4: Quantitative information for Monolingual Corpus v2

- **Monolingual Corpora n-grams** based on the annotated ENV and LAB monolingual corpora for EL, EN, ES, FR and IT. Word and word/tag/lemma n-grams accompanied by their observed frequency counts have been generated. The n-grams' length ranges from unigrams to five-grams. The size of the monolingual n-grams corpora is as follows:

RESOURCE (CC BY SA Licence)	EL		EN		ES		IT		FR	
	ENV	LAB	ENV	LAB	ENV	LAB	ENV	LAB	ENV	LAB
Tokens	31,713	24,069	46,553	45,134	49,860	58,067	35,998	70,438	42,780	46,992
Sentences	1,185	948	1,700	1,407	1,882	1,969	525	1,175	1,235	1,232
1-grams	435	363	910	606	652	582	459	547	824	664
2-grams	3,860	3,104	17,548	12,561	4,829	4,856	4,109	5,247	13,738	11,776
3-grams	9,767	7,725	75,302	57,544	13,777	13,993	11,773	16,698	60,196	54,793
4-grams	13,683	10,650	148,085	120,877	21,883	22,815	17,243	26,014	127,183	121,160
5-grams	14,954	11,513	207,570	177,588	25,791	27,262	19,268	29,778	187,590	185,045

Table 5: Data sizes (in thousand units) for the n-grams generated from PANACEA monolingual data

- **Monolingual Dependency Parsed Corpora** for two domains ENV and LAB and three languages EL, ES and IT. Text has been automatically pre-processed, and annotated with PoS and lemma

and dependency relations. Further details can be found in D4.4 report and in the associated readme files.

RESOURCE (CC_BY_NC_SA)	TOKENS	SENTENCES
Annotated Corpus ENV EL	34,668,532	1,554,313
Annotated Corpus LAB EL	26,046,343	1,255,073
Annotated Corpus ENV ES	30,972,685	1,367,518
Annotated Corpus LAB ES	60,956,181	2,388,141
Annotated Corpus ENV IT	36,000,000	1,431,914
Annotated Corpus LAB IT	70,000,000	2,975,818

Table 6: Data sizes for Dependency Parsed Corpora

- **Monolingual Lexica** for two domains ENV and LAB. Automatic acquired information is Verb Subcategorization frames for EN, ES and IT; Lexical semantic classes for ES and EN nouns, and a merged Lexicon with Subcategorization and Lexical semantic classes for EN and ES. Further details about these lexica can be found in D6.2 and D6.3 reports and in the associated readme files.

RESOURCE (CC BY Licence)	ENTRIES
Monolingual Lexicon V-SUBCAT ENV ES	1,543
Monolingual Lexicon V-SUBCAT LAB ES	1,015
Monolingual Lexicon V-SUBCAT ENV IT	26
Monolingual Lexicon V-SUBCAT LAB IT	27
Monolingual Lexicon V-SUBCAT ENV EN	895
Monolingual Lexicon V-SUBCAT LAB EN	1,063
Monolingual Lexicon lexical sem classes ENV ES	4,199
Monolingual Lexicon lexical sem classes LAB ES	5,037
Monolingual Lexicon lexical sem classes ENV EN	3,641
Monolingual Lexicon lexical sem classes LAB EN	3,762
Monolingual Subcat & Lex Sem classes ENV ES	5,742
Monolingual Subcat & Lex Sem classes LAB ES	6,052
Monolingual Subcat & Lex Sem classes ENV EN	4,704
Monolingual Subcat & Lex Sem classes LAB EN	4,657
Multilevel, multi-domain lexica ES	110,316

Table 7: Data sizes for Monolingual Lexica

- **Monolingual Lexica Gold Standard** for three domains: General, Environment and Labour with Verb Subcategorization and Lexical semantic classes for Spanish; Environment and Labour with Verb Subcategorization for Italian and Lexical semantic classes for English. Further details about these lexica can be found in D7.4 report and in the associated readme files.

RESOURCE (CC-BY-SA Licence)	ENTRIES	SENTENCES
-----------------------------	---------	-----------

Monolingual Gold Standard: V-SUBCAT ENV EN	28	200
Monolingual Gold Standard: V-SUBCAT LAB EN	29	200
Monolingual Gold Standard: V-SUBCAT ENV ES	30	200
Monolingual Gold Standard: V-SUBCAT LAB ES	30	200
Monolingual Gold Standard: V-SUBCAT OPEN IT	30	200
Monolingual Gold Standard: V-SUBCAT LAB IT	30	200
Monolingual Gold Standard: Lex Sem Class GRAL ES	5,068	Na.

Table 8: Data sizes for Monolingual Lexica Gold Standards

- **Monolingual Multiword Units** for Italian. The ENV MW is a lexicon of nominal multiword expressions automatically extracted from a 37Mio words newspaper corpus. The LAB MW is a lexicon of nominal multiword expressions automatically extracted from a 66Mio words newspaper corpus. The former contains 14,109 and the latter 15,332 Lexical Entries, of which 10,000 in each have entryType = Multiword. Further details about these lexica can be found in D6.2 and D6.3 reports and in the associated readme files.

RESOURCE (CC-BY-NC Licence)	ENTRIES	MWU
Monolingual MultiWord Units ENV IT	15,491	10,000
Monolingual MultiWord Units LAB IT	15,279	10,000

Table 9: Data sizes for IT Multiword Units

- **Bilingual Aligned Parallel Corpora.** The initial version of the Focused Bilingual Crawler (FBC) was used to construct domain-specific parallel corpora (aligned on sentence level) in the EN-FR and EN-EL language combinations for Environment and Labour Legislation domains. Table 11 provides the size in aligned sentences, after filtering. Details about these collections are reported in PANACEA deliverable [D5.3](#) and in the associated readme files.

RESOURCE (CC-BY Licence)	EN-FR	EN-EL
Bilingual Aligned Parallel Corpus ENV	13,840	13,253
Bilingual Aligned Parallel Corpus LAB	23,861	9,764

Table 10: Data sizes (sentences) for Bilingual Aligned Parallel Corpora

- **Bilingual Glossaries.** From the already mentioned parallel corpora, bilingual glossaries were automatically produced. Details about these collections are reported in PANACEA deliverable D5.4 and D5.7 and in the associated readme files.

RESOURCE (CC-BY Licence)	ENTRIES
Bilingual Glossary ENV EL-EN	18,584
Bilingual Glossary LAB EL-EN	13,220
Bilingual Glossary ENV FR-EN	17,690
Bilingual Glossary LAB FR-EN	27,798

Table 11: Data sizes for Bilingual Glossaries

In the tables above, the programs developed by the project as well as the resources produced have been described and listed. In addition some other assets are considered for exploitation (cf. D2.4. Exploitation Plan):

The Platform: the factory for the production of language resources will be exploited as such to produce R&D resources on demand. It could be also exploited through creative workflows and with raw data, obtained with the adequate rights, to produce resources for industry, in particular Machine Translation and Localization players, but not only. Such exploitation service will be rendered by the Consortium members under market conditions.

The Language Resources (see the exhaustive list on <http://www.PANACEA-lr.eu/en/info-for-researchers/data-sets/>): the test sets used during the project, that is the language resources will be made freely available for research purposes. Although, this will not generate revenues, it is expected that such use will make PANACEA (and the services) more visible and may be trigger more commercial projects for similar productions.

Web services: integrated into the final version of the platform will be exploited by the consortium (as individual partners), for instance to generate new resources on demand.

Workflows: The workflows are intrinsically associated with the Platform and hence will be part of the platform exploitation. They may be exploited as standalone pieces if the users have adopted similar environment (Taverna, etc.).

The tools developed during the life of the project (therefore foreground material as described in the Consortium Agreement) will be exploited by their owners, for instance to generate new resources on demand.

The business model envisaged in the PANACEA exploitation plan is based mostly on exploiting the Platform to produce new resources on demand, design workflows on demand, connect to third parties new web services, and also to make available the platform for internal use by commercial organizations.

The consortium agreed to ensure that the assets of the PANACEA project, described in this report, are maintained for at least 24 months after the end of the project. This includes a commitment for keeping the whole platform in place and running (the platform including Registry, myExperiment currently located at ELDA, the workflow engine and the existing workflows and related web services, currently located in UPF, CNR-ILC, ILSP, UC, DCU, LT and ELDA). Such agreement will help better assess the future exploitation after the implementation of what is described in this report.

The Language Resources developed within the project will be exploited as part of the ELRA catalogue but also using the META-SHARE infrastructure (in addition to the UPF e-repository). Such agreement will help better tune the exploitation plan as described in this report through this "sustainability" period.

9. Report on societal implications

[Adding tables]