PANACEA

Núria Bel
Universitat Pompeu Fabra

LREC 2012
EC Projects Session

PANACEA's objective is
to join together a number
of advanced interoperable tools
to build a
factory of Language Resources

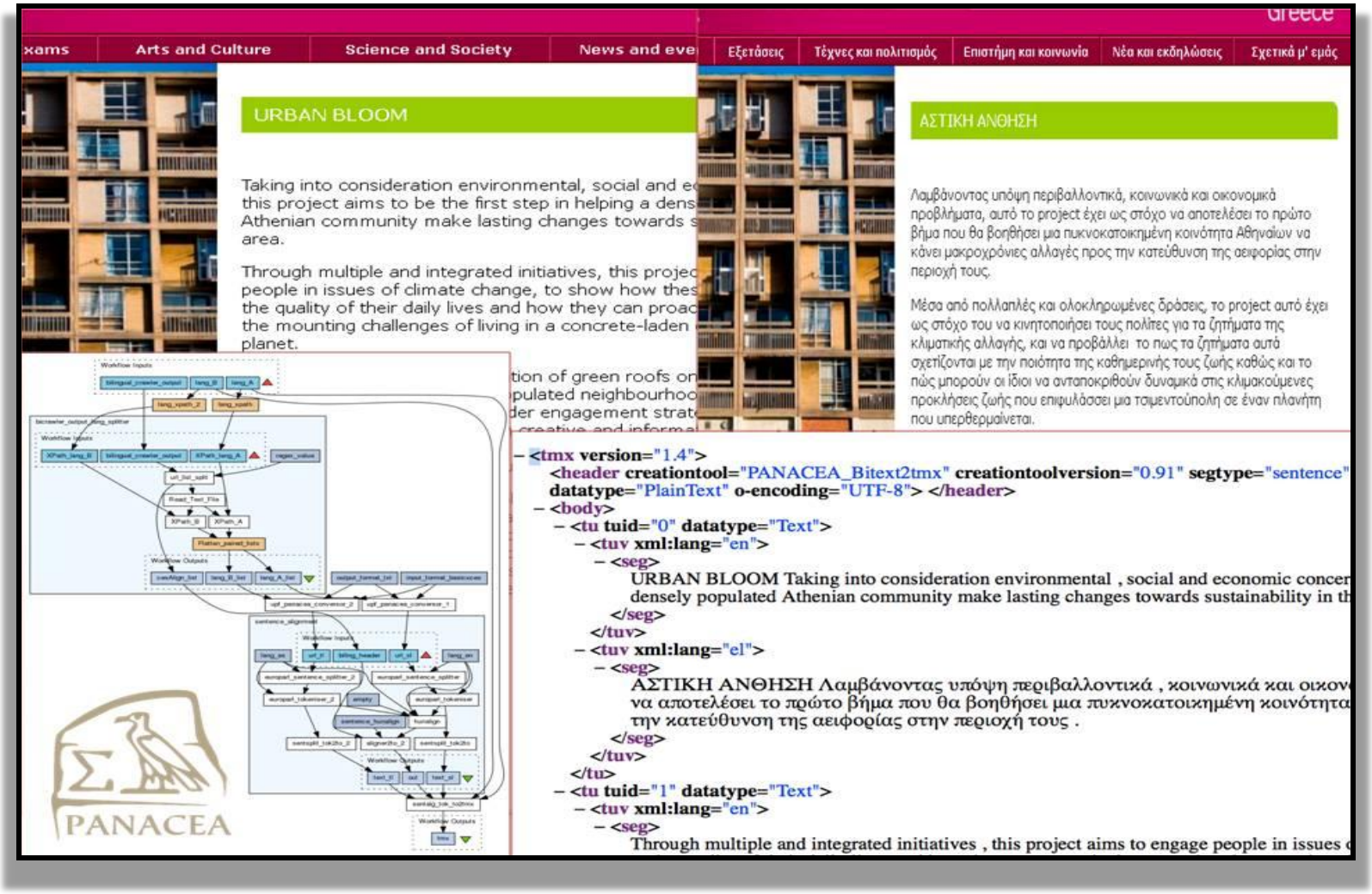
A production line that automates the stages involved in the acquisition, production, updating and maintenance of the LR required by MT and other Language Technologies.

Cost and time reduction by automation
is the only way to ensure
the continuous supply of LR
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 challenges

- Interoperability for the easy integration of **robust**, **scalable** web service-deployed components.
- Handling the impact of dealing with massive data.
- Research for increasing accuracy in automatic acquisition & production methods.
- Gain credibility 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 results.**

First products ..



Also Available …

The Language Resource Factory …

TMX aligned texts

```
– <tmx version="1.4">
    <header creationtool="PANACEA_Bitext2tmx" cr
    datatype="PlainText" o-encoding="UTF-8"> </he
  – <body>
    – <tu tuid="0" datatype="Text">
        xml:lang="en">
        seg>
          URBAN BLOOM Taking into considerati
          densely populated Athenian community n
        /seg>
      /v>
        xml:lang="el">
        seg>
          ΑΣΤΙΚΗ ΑΝΘΗΣΗ Λαμβάνοντας υπ
          να αποτελέσει το πρώτο βήμα που θα
          την κατεύθυνση της αειφορίας στην π
        </seg>
```

Bilingual lexica

hell	Ad	bright	Ad
heller Licht	No	bright light	No
Henne	No	chicken	Ad
Herausforderung des Wettbewerbs	No	challenge of competition	No
Herausforderung Rechnung	No	challenge	No

PANACEA Registry

PANACEA MyExperiment

Intelligent crawling

```
<p  id="p32" topic="sick leave;unfair
  dismissal;employer">… claiming unfair
  dismissal, wrongful dismissal and sex
  discrimination. She claimed
  her sick leave …</p> (LAB_EN)
<p id="p36" type="title" topic
  hielo ártico se derrite más
  predicho</p> (ENV_ES)
<p id="p8" crawlinfo="boilerplate">CDURABLE sur
  twitter</p> (ENV_FR)
```

Rich information lexica

```
- <LexicalEntry id="le_14">
  <feat att="partOfSpeech" val="noun" />
- <Lemma>
  <feat att="writtenForm" val="catástrofe" />
  </Lemma>
- <Sense>
  <feat att="corpusID" val="labour" />
  <feat att="eventive" val="no" />
  </Sense>
  </LexicalEntry>
- <LexicalEntry id="le_15">
    eech" val="noun" />
    rm" val="strike" />
    " val="labour" />
    " val="unknown" />
  </Sense>
```

Verbal SCF induction

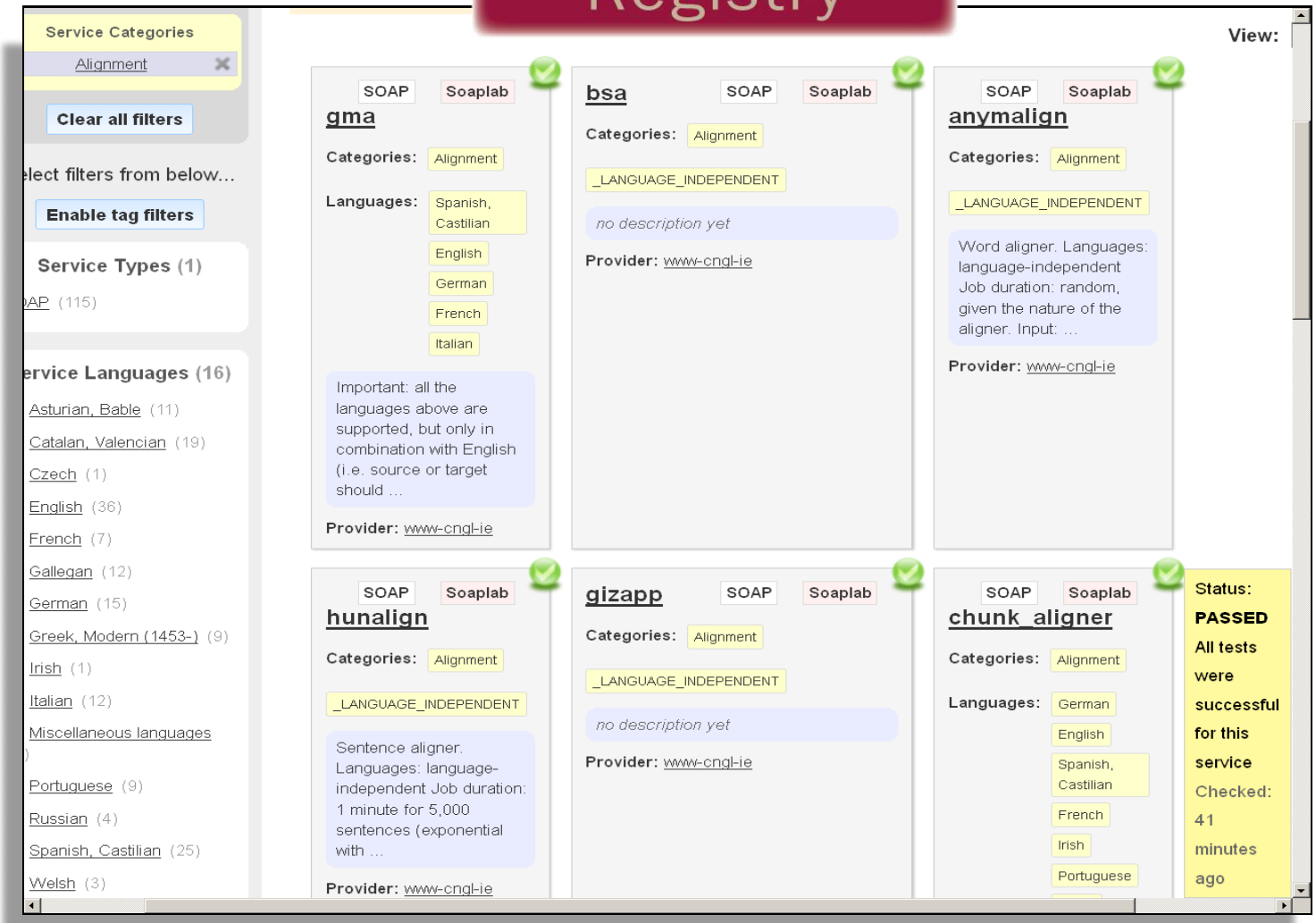
```
alter su     0.2518  945
alter do-su-xc     0.0253  95
alter do-ob-su     0.0163  61
alter io-su     0.0863  324
```

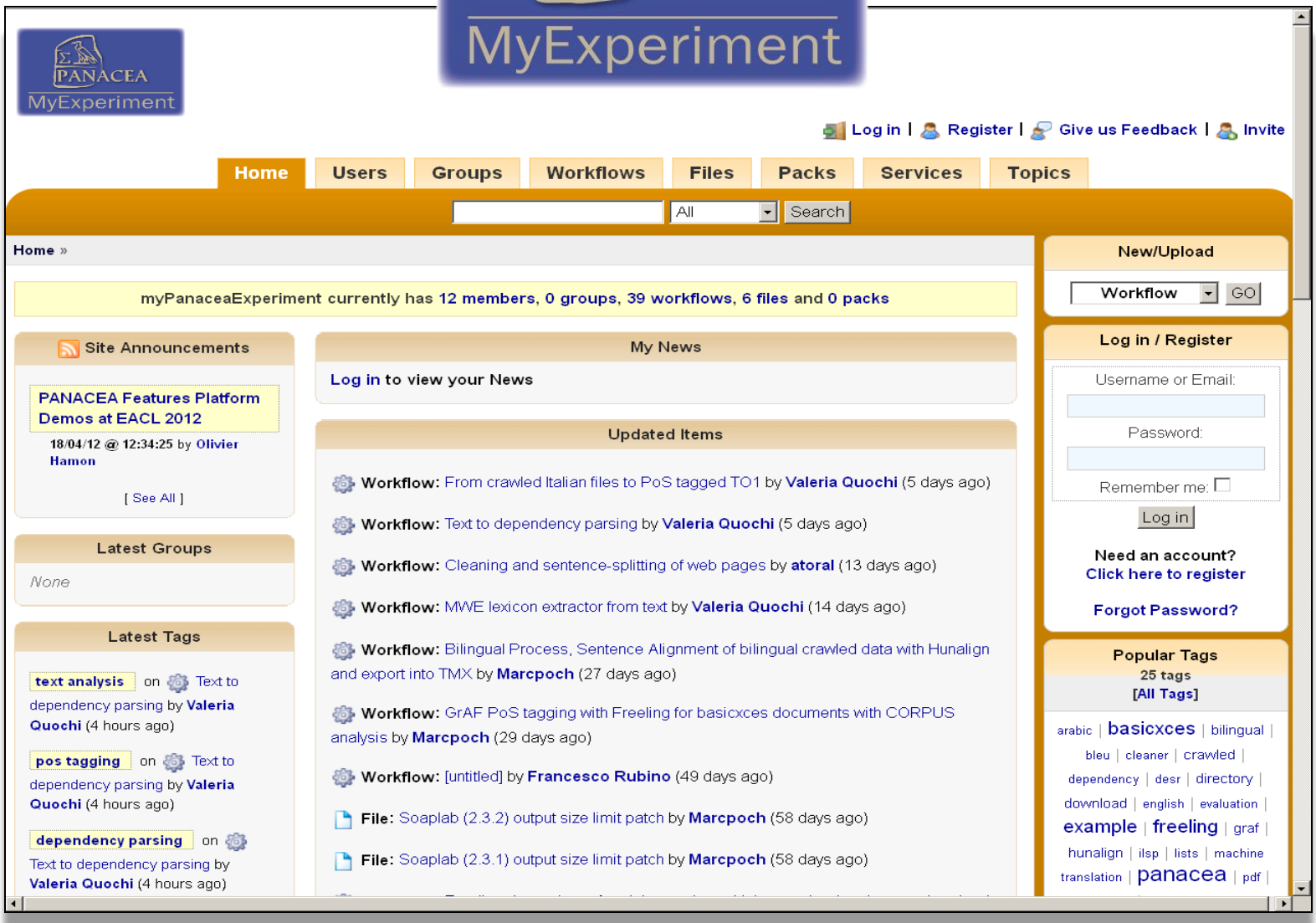
MWU detection

```
<feat att="partOfSpeech" val="n"/><Lemma><feat
att="writtenForm" val="authorisation for therapeutic u
></Lemma><Sense id="id270339-l-s"/>
```

The platform

- More than 110 interoperable web services
- Building workflows with Taverna (Hull et al., 2006)
- The PANACEA Registry is a BioCatalogue instance (Belhajjame et al., 2008)
- The PANACEA sharing space is a myExperiment instance (De Roure et al., 2008)


PANACEA
Registry
Service Categories
Alignment
Clear all filters
lect filters from below...
Enable tag filters
Service Types (1)
AP (115)
ervice Languages (16)
Asturian, Bable (11)
Catalan, Valencian (19)
Czech (1)
English (36)
French (7)
Gallegan (12)
German (15)
Greek, Modern (1453-) (9)
Irish (1)
Italian (12)
Miscellaneous languages
Portuguese (9)
Russian (4)
Spanish, Castilian (25)
Welsh (3)
View:
SOAP Soaplab
gma
Categories: Alignment
Languages: Spanish, Castilian English German French Italian
Important: all the languages above are supported, but only in combination with English (i.e. source or target should ...
Provider: www-cngl-ie
SOAP Soaplab
bsa
Categories: Alignment
_LANGUAGE_INDEPENDENT
no description yet
Provider: www-cngl-ie
SOAP Soaplab
anymalign
Categories: Alignment
_LANGUAGE_INDEPENDENT
Word aligner. Languages: language-independent Job duration: random, given the nature of the aligner. Input: ...
Provider: www-cngl-ie
SOAP Soaplab
hunalign
Categories: Alignment
_LANGUAGE_INDEPENDENT
Sentence aligner. Languages: language-independent Job duration: 1 minute for 5,000 sentences (exponential with ...
Provider: www-cngl-ie
SOAP Soaplab
gizapp
Categories: Alignment
_LANGUAGE_INDEPENDENT
no description yet
Provider: www-cngl-ie
SOAP Soaplab
chunk_aligner
Categories: Alignment
Languages: German English Spanish, Castilian French Irish Portuguese
Status: PASSED
All tests were successful for this service
Checked: 41 minutes ago


PANACEA
MyExperiment
Log in | Register | Give us Feedback | Invite
Home
Users
Groups
Workflows
Files
Packs
Services
Topics
All
Search
Home »
myPanaceaExperiment currently has 12 members, 0 groups, 39 workflows, 6 files and 0 packs
Site Announcements
PANACEA Features Platform Demos at EACL 2012
18/04/12 @ 12:34:25 by Olivier Hamon
[See All]
Latest Groups
None
Latest Tags
text analysis on Text to dependency parsing by Valeria Quochi (4 hours ago)
pos tagging on Text to dependency parsing by Valeria Quochi (4 hours ago)
dependency parsing on Text to dependency parsing by Valeria Quochi (4 hours ago)
My News
Log in to view your News
Updated Items
Workflow: From crawled Italian files to PoS tagged TO1 by Valeria Quochi (5 days ago)
Workflow: Text to dependency parsing by Valeria Quochi (5 days ago)
Workflow: Cleaning and sentence-splitting of web pages by atoral (13 days ago)
Workflow: MWE lexicon extractor from text by Valeria Quochi (14 days ago)
Workflow: Bilingual Process, Sentence Alignment of bilingual crawled data with Hunalign and export into TMX by Marcpoch (27 days ago)
Workflow: GrAF PoS tagging with Freeling for basicxces documents with CORPUS analysis by Marcpoch (29 days ago)
Workflow: [untitled] by Francesco Rubino (49 days ago)
File: Soaplab (2.3.2) output size limit patch by Marcpoch (58 days ago)
File: Soaplab (2.3.1) output size limit patch by Marcpoch (58 days ago)
New/Upload
Workflow
GO
Log in / Register
Username or Email:
Password:
Remember me:
Log in
Need an account?
Click here to register
Forgot Password?
Popular Tags
25 tags
[All Tags]
arabic | basicxces | bilingual | bleu | cleaner | crawled | dependency | desr | directory | download | english | evaluation | example | freeling | graf | hunalign | ilsp | lists | machine translation | panacea | pdf |

DEMO's 2012

- ~~EACL 2012 (Avignon, 23-27 April)~~
- LREC 2012 (Istambul, 23-25 May)
- Localization World 2012 (Paris, 4-6 June)
- LT-Innovate (Brussels, 19 June)
- META-FORUM (Brussels, 20-21 June)

Ongoing …

Before the end of the project in December 2012, PANACEA wants to address more industrial needs and is seeking collaboration

with LT industries.

Let us devise a solution tailored to your needs!

info@panacea-lr.eu

PANACEA

SEVENTH FRAMEWORK PROGRAMME

Keep visiting www.panacea-lr.eu

to access PANACEA's results and available web services and workflows

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).

This documented 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:
iulatrl@upf.edu

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

Barcelona, 2012