



OpeNER and PANACEA: Web Services for the CLARIN Research Infrastructure

Davide Albanesi
Riccardo Del Gratta
{name}.{surname}@ilc.cnr.it

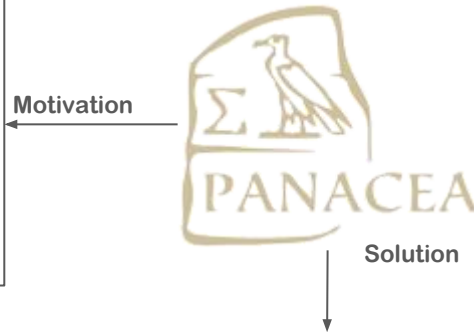
Outline

- Descriptive section
 - PACANEA & OpeNER, An overview of the projects, Panacea & CLARIN + OpeNER & CLARIN, Few Notes on the Projects
- Interlude section
 - On The VLO, Community Involvement
- Technical section
 - Common aspects of the projects:
 - WorkFlow, Interoperability...
 - The projects meet CLARIN
 - Actual develop
 - Future work

PANACEA: an overview of the project (1 / 2)

Project detail Details: [Web](#); Seventh Framework Programme (theme 3) Grant: 248064; Duration: 2010-2012

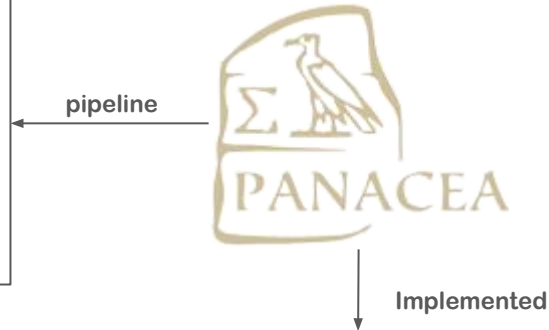
- To overcome language barriers through technological means;
- To address MT systems and their significant impact on the management of multilingualism in Europe w/ related translations of the huge quantity of data produced;
- To cover the needs of hundreds of millions of citizens.



- A factory of different LRs by means of a platform:
 - To automate all the stages involved in the acquisition, production, updating, validation and maintenance of LRs;
 - To foster the interoperability among LTs;
 - To allow less skilled people to use LRTs.

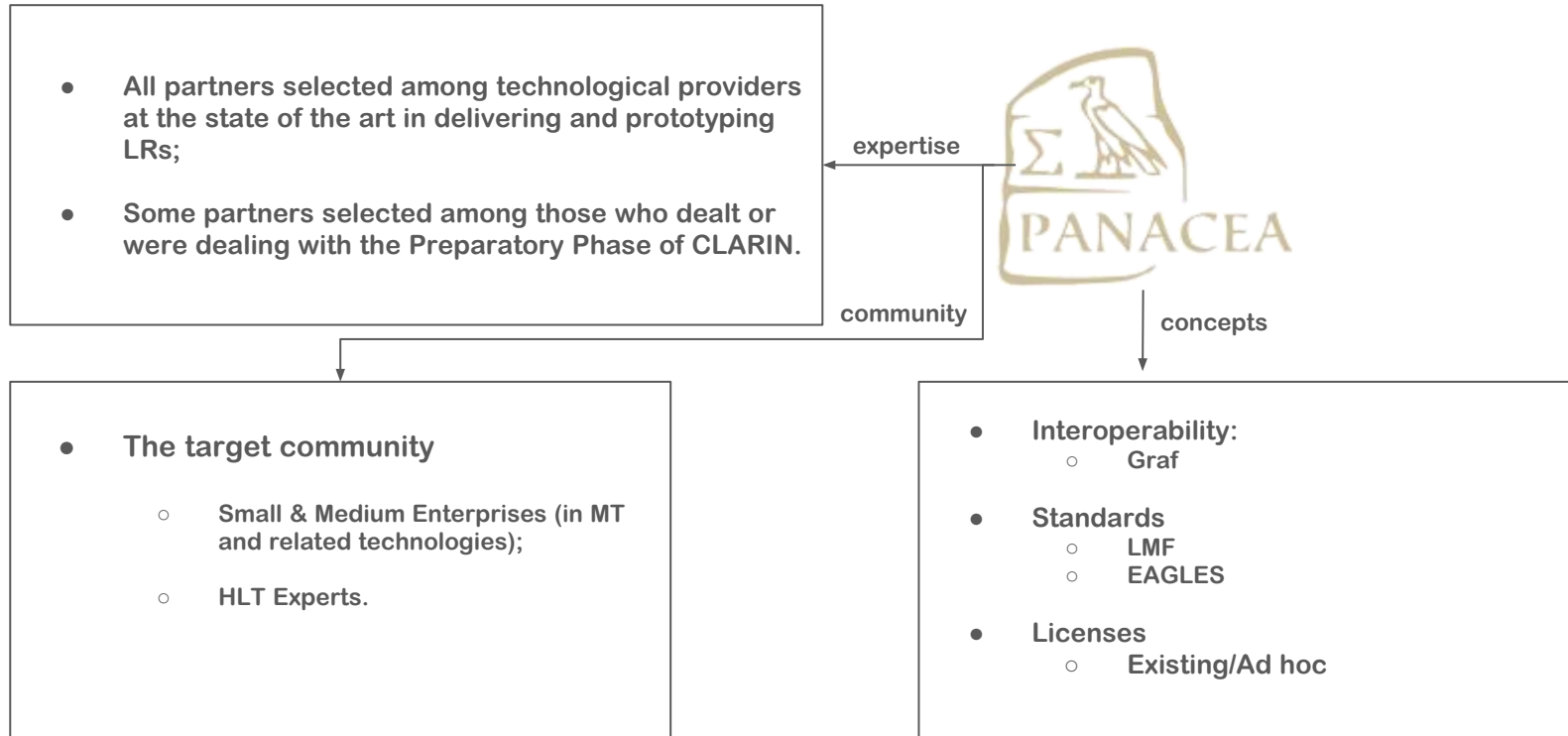
PANACEA: an overview of the project (2 / 2)

- Corpus creation
- Text Pre-processing for erasing non-linguistic code, indexing and segmenting the text.
- Text Annotation (part-of-speech tagging, lemmatization NER...)
- “Parallel Corpus Technologies”
- Lexical Acquisition Technologies
- ...



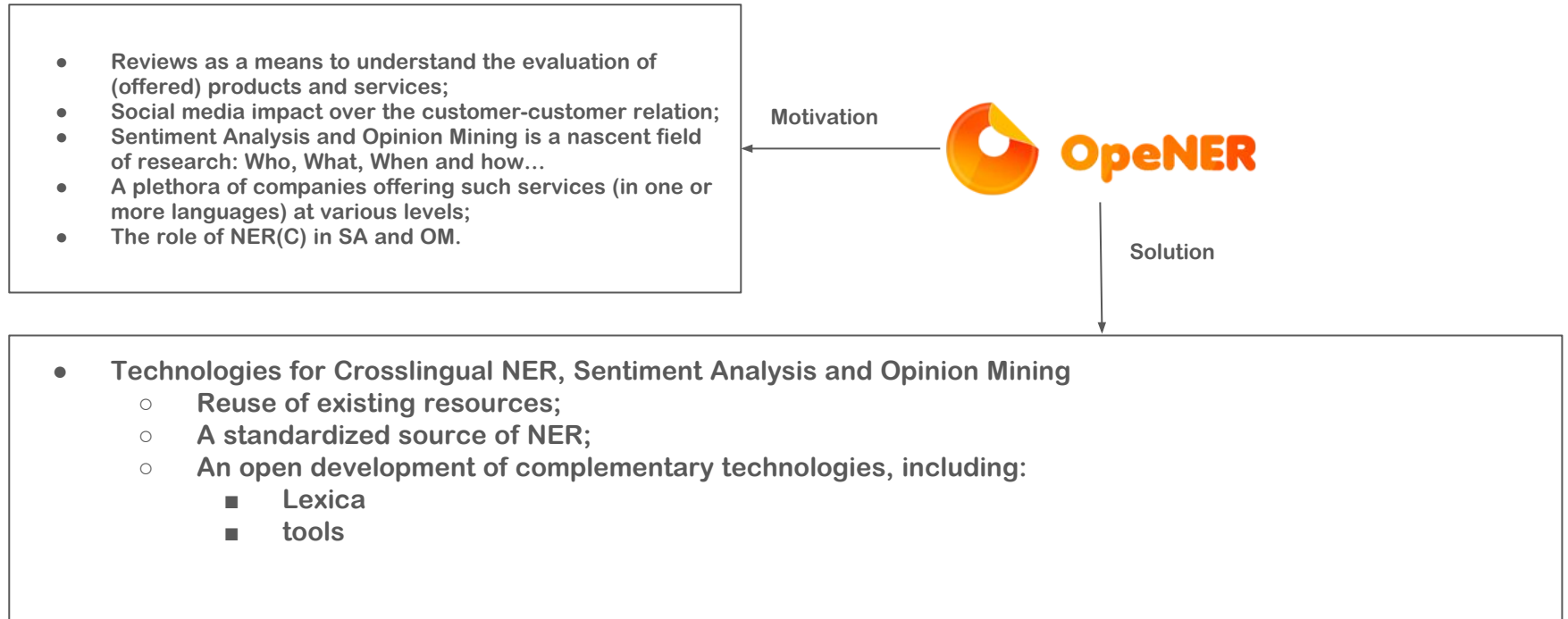
- **As a set of interoperable SOAP Web Services:**
 - based on both available or newly developed tools;
 - To create, transform Language Resources.
- **Through a platform that:**
 - functionally integrates the different components required in the production of LRs;
 - is based on an especially dedicated workflow engine (taverna) for the composition of different web services;

PANACEA: Panacea & CLARIN



OPENER: an overview of the project (1 / 2)

Project detail Details: [Web](#); Seventh Framework Programme. Grant: 296451; Duration: 2012-2014



OPENER: an overview of the project (2 / 2)

- Lexicon creation;
- Basic Text Annotation (part-of-speech tagging, lemmatization NER...);
- Production of advanced annotations (for OM, SA...)
- ...

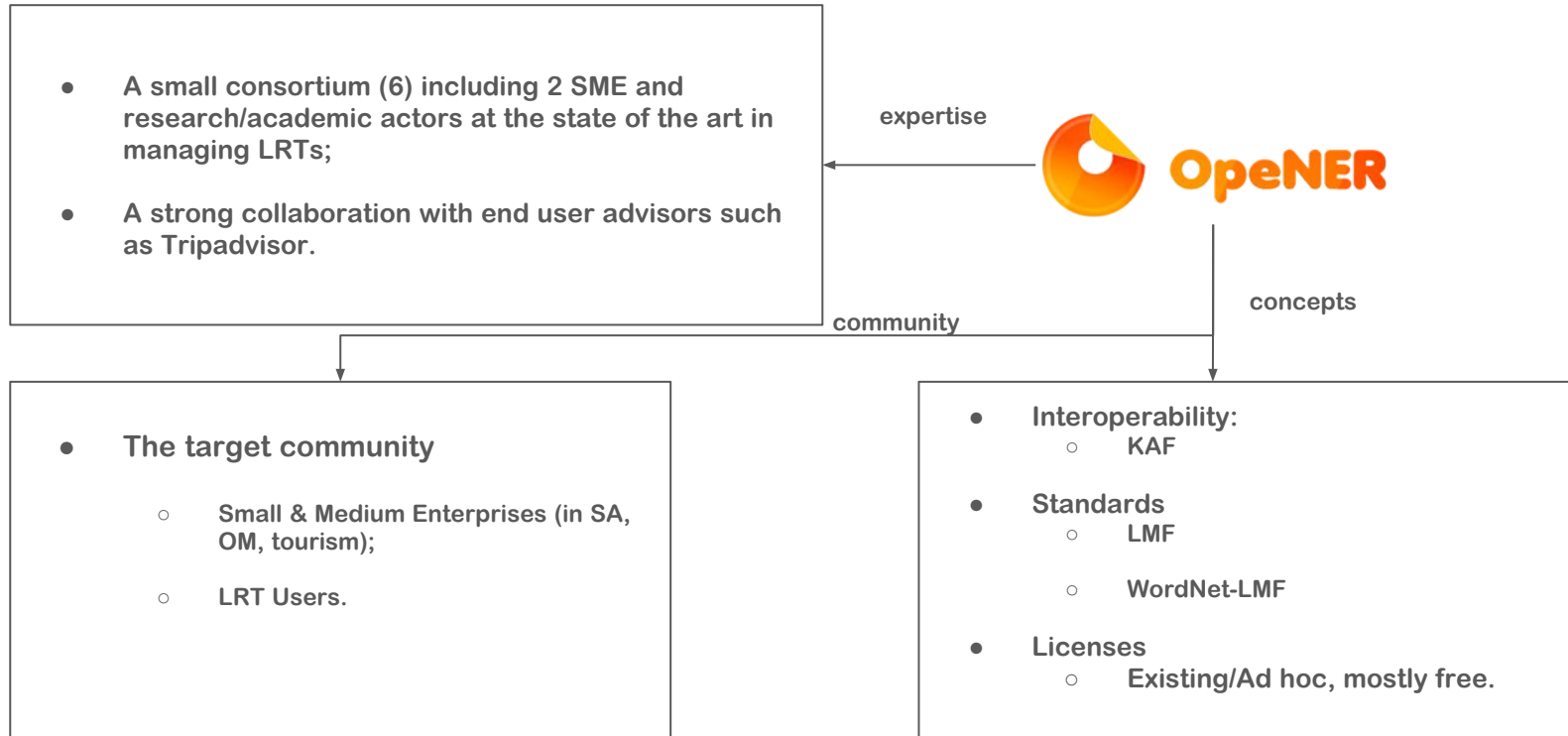
pipeline



Implemented

- **As a set of interoperable REST Web Services:**
 - based on both available or newly developed tools;
 - Adapted to a specific domain.
- **Through a set of APIs that:**
 - functionally integrates the different components required in the production of LR's;
 - is based on an especially dedicated pipeline system which guides the users to compose web services.

OPENER: OpeNer & CLARIN



Few Notes on the Projects

- Both projects implement interoperability and exploit available standards;
- Both projects use a sort of workflow to process linguistic data.

But

- PANACEA involves HLT community more than OpeNer;
 - The consortium of Panacea is closer to “CLARIN stuff” than Opener’s.
-
- So... Opener seems to be more interesting!

On the VLO

Showing 8 results within selection for **sentiment analysis** **software, webservice or lexicalResource** Results per page: 10

Use the categories below to limit the search results to those matching the selected value(s).

Language

Collection

Resource type

lexi

software, webservice **X**
OR lexicalResource **X**

Modality

Format

Czech image captioning, machine translation, and sentiment analysis (Neural Monkey models)

(Part of LINDAT / CLARIN Data & Tools)

This submission contains trained end-to-end models for the Neural Monkey toolkit for Czech and English, solving three NLP tasks: machine translation, image captioning, and **sentiment analysis**. The models are trained on standard datasets and achieve state-of-the-art or near state-of-the-art performance in the tasks. Th...

See this record and its resources at the record's landing page

Slovene sentiment lexicon JOB 1.0

(Part of CLARIN.SI data & tools)

The JOB lexicon for **sentiment analysis** of Slovenian texts contains a list of 25,524 headwords from the List of Slovenian headwords 1.1 (<http://hdl.handle.net/11356/1038>) extended with **sentiment** ratings based on the AFINN model with an integer between -5 (very negative) and +5 (very positive). The ratings are derived fr...

See this record and its resources at the record's landing page

Emoji Sentiment Ranking 1.0

A simple VLO-search performed by non-experts shows:

The two communities are differently represented.

Showing 1 to 10 of 45 results within selection for **machine translation** **software, webservice or lexicalResource** Results per page: 10

Use the categories below to limit the search results to those matching the selected value(s).

Language

Collection

Resource type

Type to filter or search for more

software, webservice **X**
OR lexicalResource **X**

text (127)
corpus (53)
session (27)
audio (19)
Data (2)
software (2)
Article (1)
Tools (1)
annotation (1)
dataset (1)
more...

<< < 1 2 3 4 5 > >>

English-Lithuanian Machine Translation Service

(Part of LRT + Open Submissions Data & Tools)

On-line freely accessible **machine translation** tool for translating English webpages or texts into Lithuanian.

See this record and its resources at the record's landing page

Spanish to English Machine translation module

(Part of PORTULAN CLARIN)

Technical Description: http://qt leap.eu/wp-content/uploads/2015/05/Pilot1_technical_description.pdf http://qt leap.eu/wp-content/uploads/2015/05/TechnicalDescriptionPilot2_D2.7.pdf http://qt leap.eu/wp-content/uploads/2016/11/TechnicalDescriptionPilot3_D2.10.pdf

See this record and its resources at the record's landing page

English to Basque Machine translation module

(Part of PORTULAN CLARIN)

Technical Description: http://qt leap.eu/wp-content/uploads/2015/05/Pilot1_technical_description.pdf http://qt leap.eu/wp-content/uploads/2015/05/TechnicalDescriptionPilot2_D2.7.pdf http://qt leap.eu/wp-content/uploads/2016/11/TechnicalDescriptionPilot3_D2.10.pdf

See this record and its resources at the record's landing page

Community Involvement

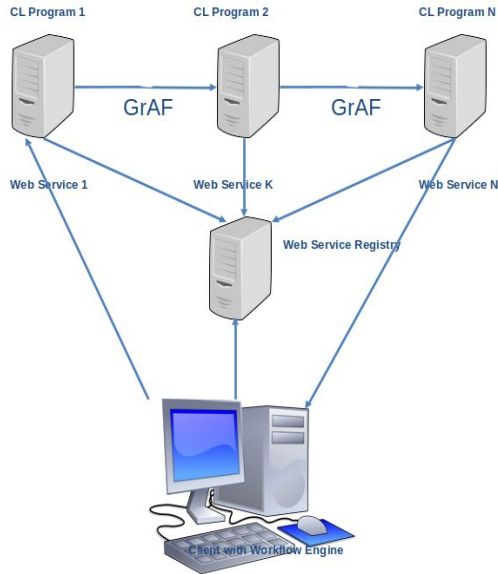
- From a double point of view:
 - Political/organizational:
 - User Involvement;
 - Workshop;
 - Dedicate events;
 -
 - Technical
 - Include the web services into CLARIN:
 - REST APIs;
 - Ready to be inserted into Workflow engines;
 - Documented into repositories.

The Projects: Common Aspects

PANACEA and OpeNer services have some common aspects:

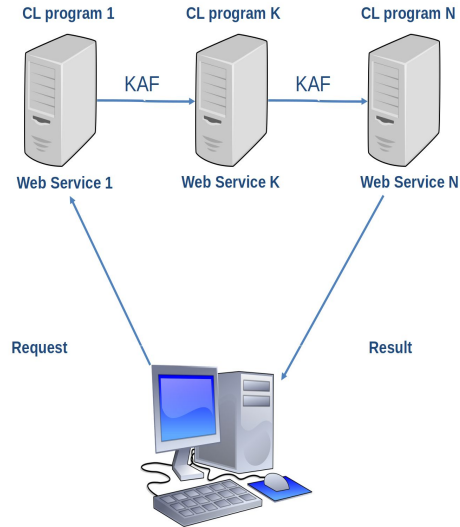
- I. Many tools in OpeNer and PANACEA are command line tool. OpeNer uses Ruby to wrap command line tools and to build REST Web Services, while PANACEA uses Soaplab and offers SOAP Web Services;
- II. OpeNer offers both POST and GET API;
- III. PANACEA offers SOAP Web Services through a Web Interface;
- IV. Simple pipelines are available in OpeNer, while a workflow engine is used in PANACEA;
- V. Kyoto Annotation Format (KAF), Lexical Markup Framework (LMF) and Graph Annotation Format (GrAF) guarantee the interoperability among data and services at different levels.

PANACEA: WorkFlow sketched



- Each Web Service is created from a CL program using SOAPLab;
- Each SOAP WS is embedded into a GUI (for human-machine interaction);
- Each WS is registered into a registry;
- GRAF is used as *lingua franca*;
- Native CL Program output as output format.

OpeNer: WorkFlow sketched



- Each Web Service is created from a CL (o existing tools) program using Ruby and offers REST APIs;
- KAF is used as *lingua franca*;
- KAF(JSON) as output format(s).

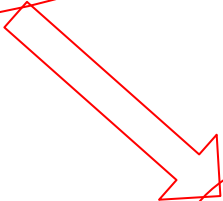
OpenNer & PANACEA Meet CLARIN

- **Motivations**
 - **Attract more users;**
 - **Ingest new services into Weblicht and Language Resource Switchboard, which (should)**
 - **Attract new users.**

OpeNer & PANACEA Meet CLARIN

- Motivations

- Involve more users;
- Ingest new services into Weblicht and Language Resource Switchboard, which (should)
 - Attract new users.



**Make CLARIN growing
and increase its
importance.**

CLARIN @ ILC

- ILC4CLARIN
 - The first and leading CLARIN B-centre of CLARIN-IT
 - Already offers services as web interfaces:
 - <https://ilc4clarin.ilc.cnr.it/en/services/> (mostly Panacea's)
 - <http://opener.ilc4clarin.ilc.cnr.it/tokenizer>
 - <http://opener.ilc4clarin.ilc.cnr.it/pos-tagger>
 - <http://opener.ilc4clarin.ilc.cnr.it/kaf2json>
 - No offered service is WebLicht AND Language Resource Switchboard

Strategy

- **Goal:**
 - **Create REST Web Services (and APIs).**
- **Considerations:**
 - **OpeNer offers REST Web Services (through APIs)**
 - **Easily wrapped into a REST context**
 - **PANACEA offers SOAP Web Services ((through GUI)**
 - **We need to play w/ some SOAP APIs before inserting in a REST context.**

Strategy

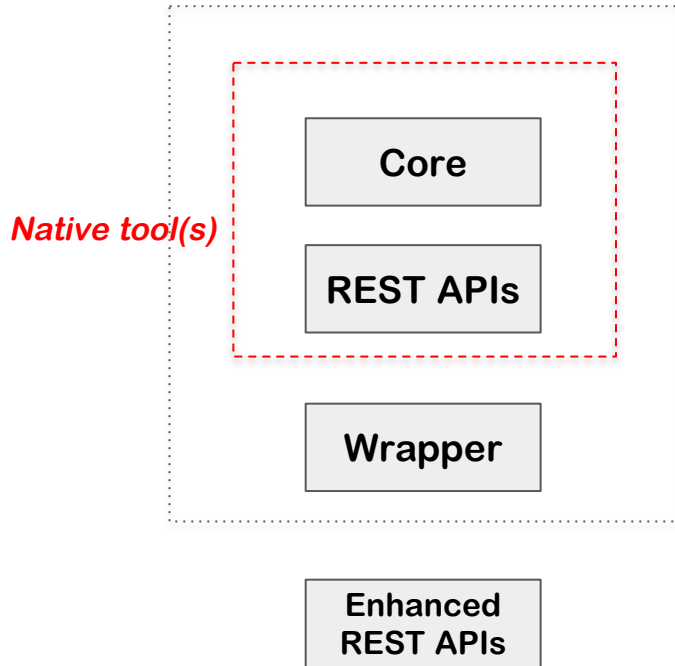
- **Two alternatives for PANACEA:**
 - (i) Use SOAP APIs (thanks to SoapLab);
 - (ii) Start from the original CL programs and use a different framework to transform these scripts into REST APIs.
- **Pros & Cons (i).**
 - Develop *a shell* around SOAP Web Services s.t. they can be “executed” by a software program and not just from a web interface
 - Fortunately the SoapLab APIs help a lot! (It’s a pro)
 - But:
 - Additional pieces of software to manage (It’s a con)

Strategy

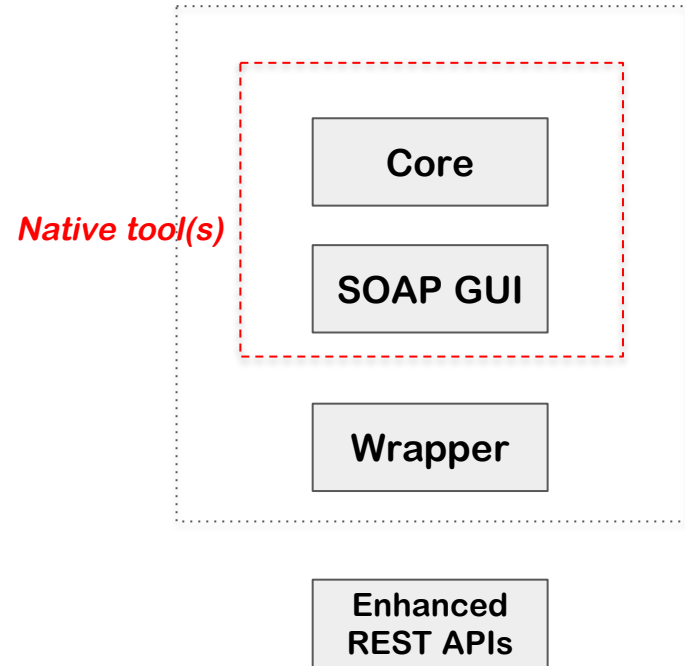
- **Pros & Cons (ii).**
 - **Replicate (operatively) what was already obtained through SoapLab**
 - **Using a state-of-the-art tool and technologies implies to be part of a wider community and to be aware of new possible solutions (not strategies) (It's a pro)**
 - **But:**
 - **Replication of endpoints (same service with more than one endpoint)**
 - **Duplication of server (file systems) (cons)**
- **We consider (ii) not economical (from our perspective)**

Overview

OpeNer



PANACEA

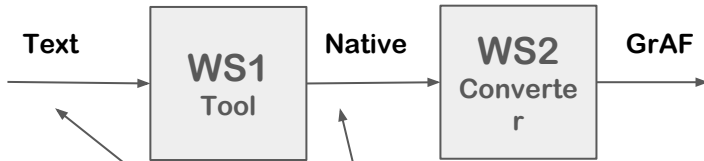


Recipe

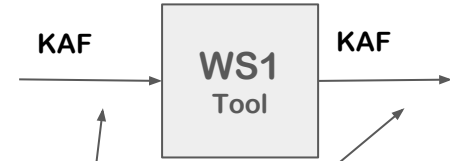
- Use [Dropwizard](#)
 - To separate web-ish part from core-ish part
 - Servlet Engine (Jetty)
 - JAX-RS (Jersey)
 - Lingua franca (JSON)
- Create multiple endpoints for
 - Language Resource Switchboard
 - WebLicht
- (A couple of) Dockers

Details of the wrapper

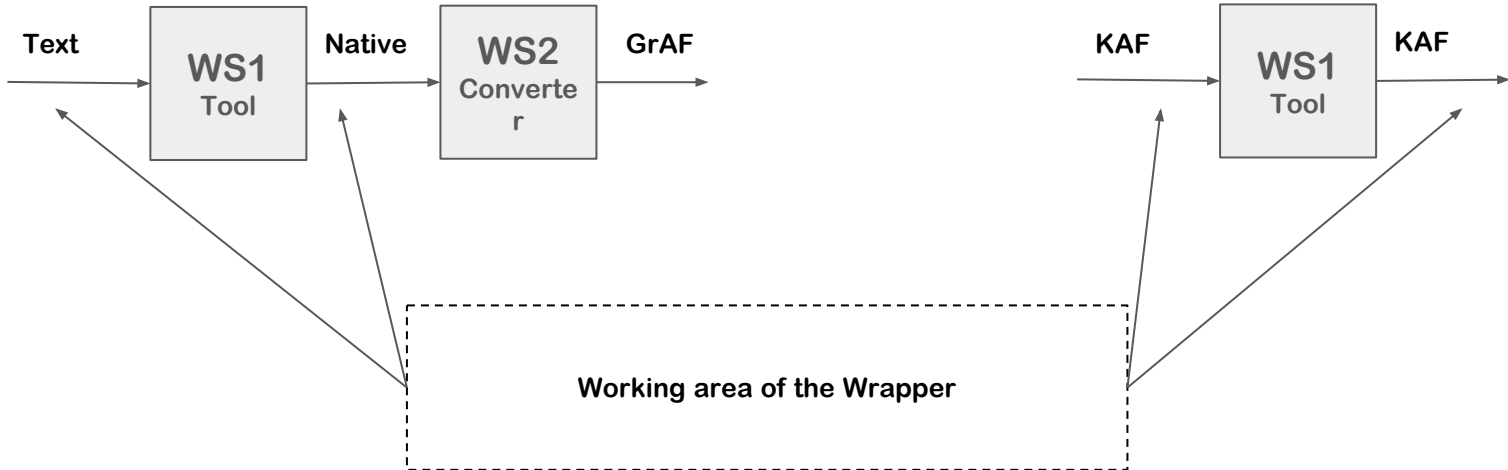
PANACEA



OpeNer



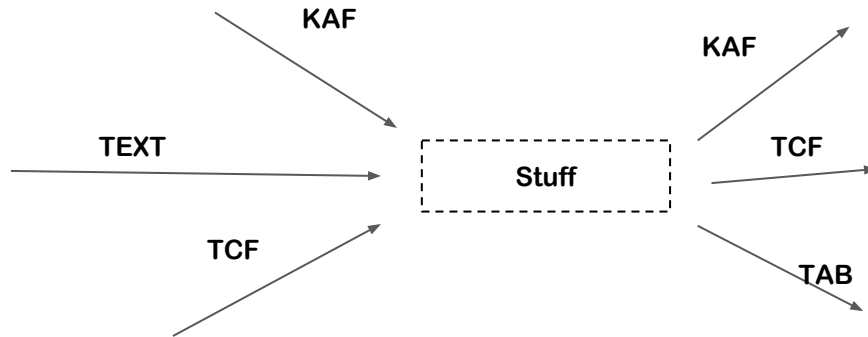
Working area of the Wrapper



Details of the wrapper

The wrapper consumes different formats in input and produces different formats in output.

- **Input**
 - Text & TCT for integration in WebLicht
 - KAF for back compatibility with the OpeNer chain
 - Text for integration with Language Resource SwitchBoard
- **Output**
 - Valid KAF, TCF and tabbed formats



From the OpeNer service readme

From the readme documentation of OpeNer <http://ilc4clarin.ilc.cnr.it/services/opener/readme>

We get some examples from

<http://ilc4clarin.ilc.cnr.it/services/opener/tokenizerhr>

The service

- manages plain texts to produce a valid TAB, TCF or KAF document;
- manages TCF document to produce a valid TAB, TCF or KAF document;
- manages KAF document to produce a valid TAB, TCF or KAF document.

Example

```
curl -H 'Content-Type: multipart/form-data' -F 'file=@myfile' -F 'form={"language":"it","iformat":"raw","oformat":"tcf"}' -X POST  
http://ilc4clarin.ilc.cnr.it/services/opener/tokenizer/runservice
```

From the PANACEA service readme

From the readme documentation of PANACEA <http://ilc4clarin.ilc.cnr.it/services/panacea/readme>

The service

- manages plain texts to produce a valid TAB, TCF or KAF document;
- manages TCF document to produce a valid TAB, TCF or KAF document;
- manages KAF document to produce a valid TAB, TCF or KAF document.

Example

```
curl -H 'content-type: text/xml' --data-binary @kaf-file.xml -X POST http://ilc4clarin.ilc.cnr.it/services/panacea/freeling_it/kaf/runservice?format=tcf
```

Github & Docker(s)

- Code available at
 - <https://github.com/cnr-ilc/linguistic-tools-for-weblicht/tree/master/OpeNerServices>
 - <https://github.com/cnr-ilc/linguistic-tools-for-weblicht/tree/master/PanaceaServices>
- Docker Images
 - <https://cloud.docker.com/u/cnrilc/repository/docker/cnrilc/ltfw>

Few Comments on Github & Docker(s)

- The github code contains more than the services described. It is also a skeleton to wrap other services. What is needed to do is to intercept the output of a service and convert it into an internal format which will be further managed.
- At [ILC4CLARIN](#) we use [rancher](#) to manage docker images;
- The official organization *cnrilc* at [dockerhub](#) is constantly updated.

Finally, to do list

- Update the list of services offered by ILC4CLARIN so that they can be used and distributed;
- Contact Language Resource Switchboard to test our services;
- Build-up a CMDI file for WebLicht.



Conclusion



We told a story.

A story of two Projects and what they created for researchers in HLT, OM, SA; and to what extent such outcomes can be useful for CLARIN too.

We digressed a bit on some technical aspects and made a lot of promises for the times to come. Promises that must be fulfilled.

“And now it ends”. Absolutely not. “No, now it begins”

Credits

Monica Monachini, National Coordinator

Alessandro Enea, Head of ICT @ ILC & Centre Technical Manager

Paola Baroni & Valeria Quochi, User Involvement & Web Content & Communications @ ILC4CLARIN

Federico Boschetti, sort of... Linguistic Counselor