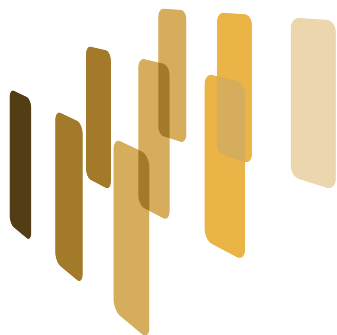


European Language Grid – META-FORUM 2020
Session 4: News from the Language Communities

- Marko Turpeinen (1001 Lakes, Finland)
- Walter Daelemans (University of Antwerp, NCC Belgium)
- Svetla Koeva (Bulgarian Academy of Sciences, NCC Bulgaria)
- Maciej Ogrodniczuk (Polish Academy of Sciences, NCC Poland)
- **Marta Villegas (Barcelona Supercomputing Center, NCC Spain)**
- François Yvon (LIMSI/CNRS, NCC France)



EUROPEAN LANGUAGE GRID

META  NET
META  FORUM 2020

Synergies between the Spanish Plan-TL & ELG

Marta Villegas (Barcelona Supercomputing Center – Centro Nacional de Supercomputación)
marta.villegas@bsc.es

01/02/03-12-2020 META-FORUM 2020 – Piloting the European Language Grid (virtual conference)
<http://www.european-language-grid.eu>

The Plan for the Advancement of Language Technology (Plan-TL)

Objective: to promote the development of NLP and MT for Spanish and co-official languages.

Calendar: the Plan-TL was approved in 2015.

Implementation: Organized into ‘subprojects’ (flagship projects) in strategic domains and collaborations with public administrations, universities and research centers and companies.



Plan-TL / BSC Platform (January 2020)



- **VM** (virtual machines)
 - Openstack (physical structure)
- **Cloud:**
 - Openstack (physical structure)
 - Ranger 2.0 (cloud environment control)
 - Kubernetes (orchestration)
 - Docker (services deployment)
- **HPC:**
 - Nord3 (updated to SLURM & SO), up to 250 nodes (+ 4000 cores)



REQUIREMENTS

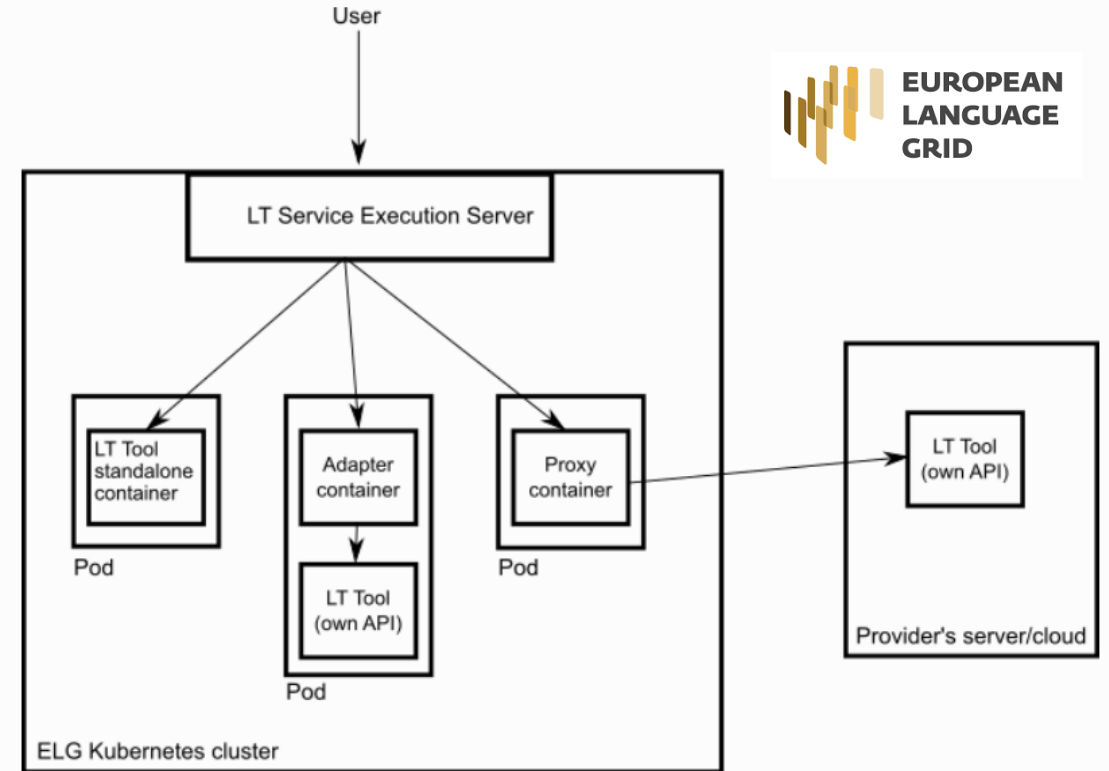
- Multi-user platform
- Able to be used in both cloud and HPC environment.
- Potential to accommodate up to thousands of cores with a good level of scalability.
- High availability of the solution.
- Support for scalable components.
- Support for Docker containers.
- Support for Docker component orchestration through tools such as Kubernetes.
- Monitoring, access control, availability of computing capacity, storage and security of the proposed solution.

Plan-TL / BSC Platform



- **VM** (virtual machines)
 - Openstack (physical structure)
- **Cloud:**
 - Openstack (physical structure)
 - Ranger 2.0 (cloud environment control)
 - Kubernetes (orchestration)
 - Docker (services deployment)
- **HPC:**
 - Nord3 (updated to SLURM & SO), up to 250 nodes (+ 4000 cores)

ELG contribution

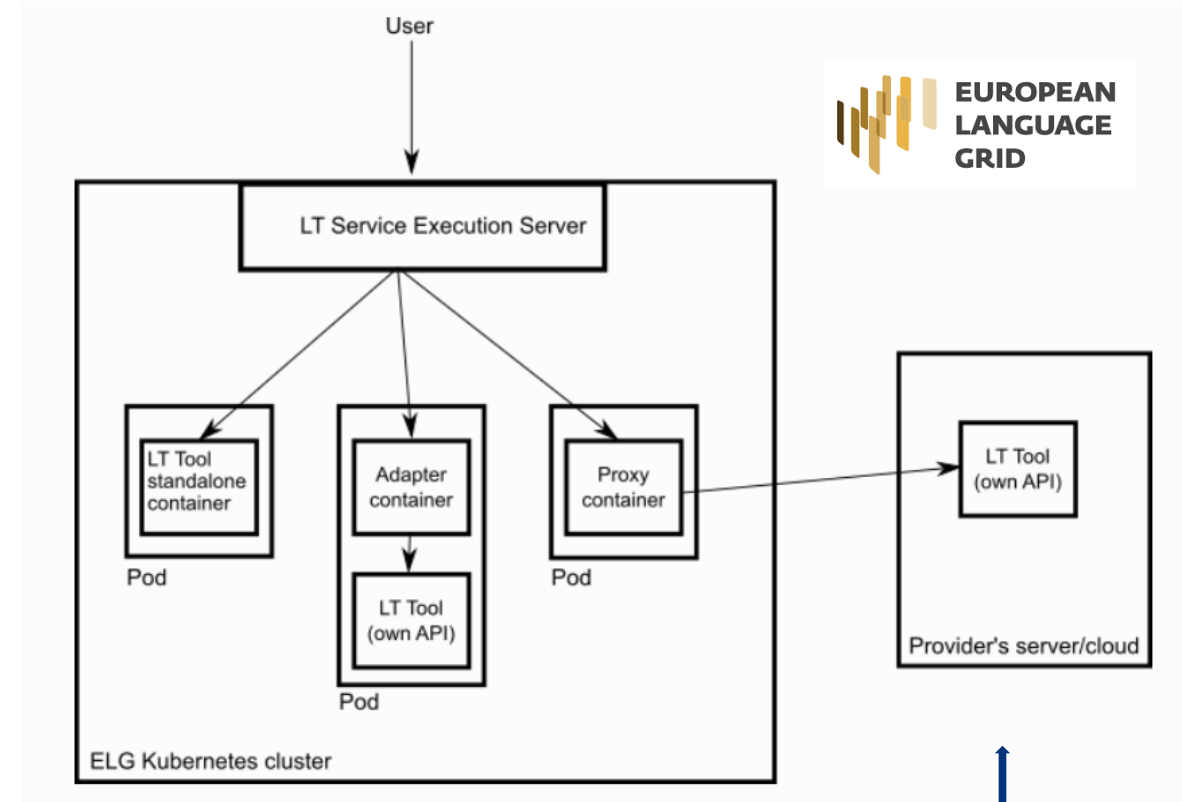


Plan-TL / BSC Platform



- **VM** (virtual machines)
 - Openstack (physical structure)
- **Cloud:**
 - Openstack (physical structure)
 - Ranger 2.0 (cloud environment control)
 - Kubernetes (orchestration)
 - Docker (services deployment)
- **HPC:**
 - Nord3 (updated to SLURM & SO), up to 250 nodes (+ 4000 cores)

ELG contribution



ELG API



Plan-TL / BSC Data

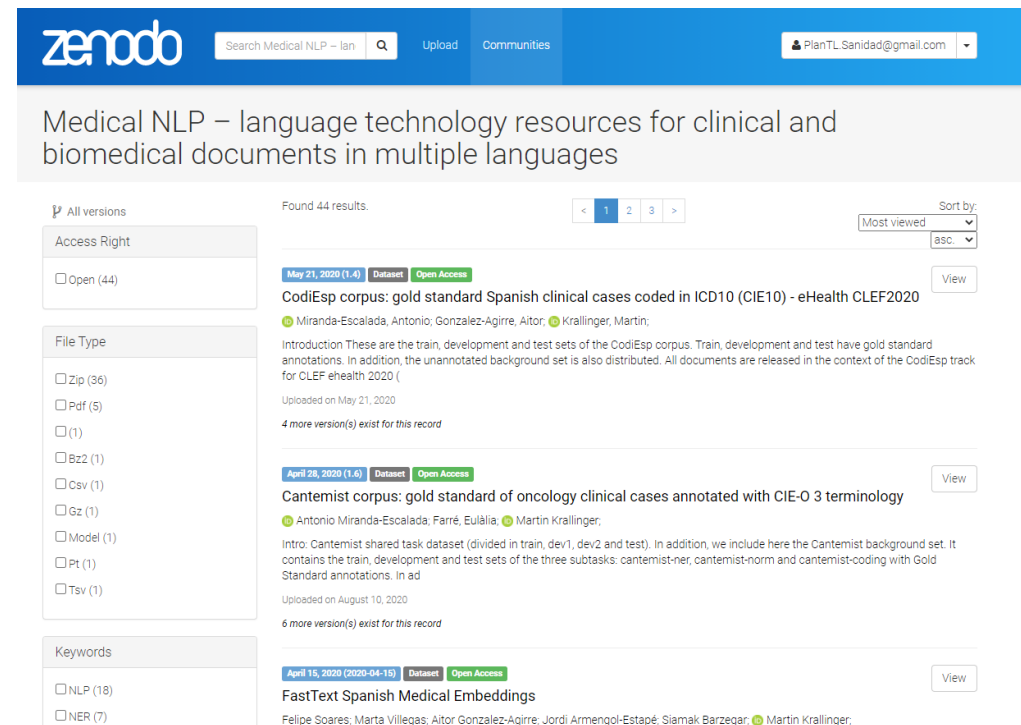
Up to now, all our data is in Zenodo (44 resources)

Zenodo works well for us as,

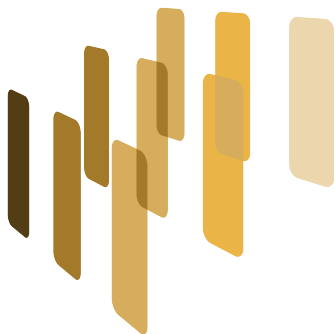
- we get a DOI
- it allows large datasets (we are about to publish huge corpora and models)
- it has version control system
- includes Github, OpenAire linking
- ...

But we can also expose and share the resources in **ELG**
(maybe via OAI-PMH)

https://zenodo.org/oai2d?verb=ListRecords&set=user-medicalNlp&metadataPrefix=oai_dc



The screenshot shows the Zenodo website interface. At the top, there's a blue header with the Zenodo logo, a search bar containing 'Search Medical NLP - lan', and buttons for 'Upload' and 'Communities'. A user profile 'PlanTL.Sanidad@gmail.com' is visible in the top right. Below the header, a grey banner reads 'Medical NLP – language technology resources for clinical and biomedical documents in multiple languages'. The main content area shows search results for 'Medical NLP'. On the left, there are filters for 'All versions', 'Access Right' (with 'Open (44)' selected), 'File Type' (listing various formats like Zip, Pdf, Bz2, etc.), and 'Keywords' (listing 'NLP (18)' and 'NER (7)'). The search results list 44 items. The first result is 'CodiEsp corpus: gold standard Spanish clinical cases coded in ICD10 (CIE10) - eHealth CLEF2020' by Miranda-Escalada, Antonio; Gonzalez-Agirre, Aitor; and Krallinger, Martin. It includes an introduction, upload date (May 21, 2020), and a link to view. The second result is 'Cantemist corpus: gold standard of oncology clinical cases annotated with CIE-O 3 terminology' by Antonio Miranda-Escalada; Farré, Eulàlia; and Martin Krallinger. It includes an introduction, upload date (August 10, 2020), and a link to view. The third result is 'FastText Spanish Medical Embeddings' by Felipe Soares, Marta Villegas, Aitor Gonzalez-Agirre, Jordi Armengol-Estapé, Siamak Barzegar, and Martin Krallinger. It includes a link to view.



European Language Grid

Thank you!



The European Language Grid has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement № 825627 (ELG).

Marta Villegas (Barcelona Supercomputing Center – Centro Nacional de Supercomputación)
marta.villegas@bsc.es

01/02/03-12-2020 META-FORUM 2020 – Piloting the European Language Grid (virtual conference)
<http://www.european-language-grid.eu>