Data search and discovery in language data spaces: challenges and solutions

Stelios Piperidis, ILSP/Athena RC
spip@athenarc.gr

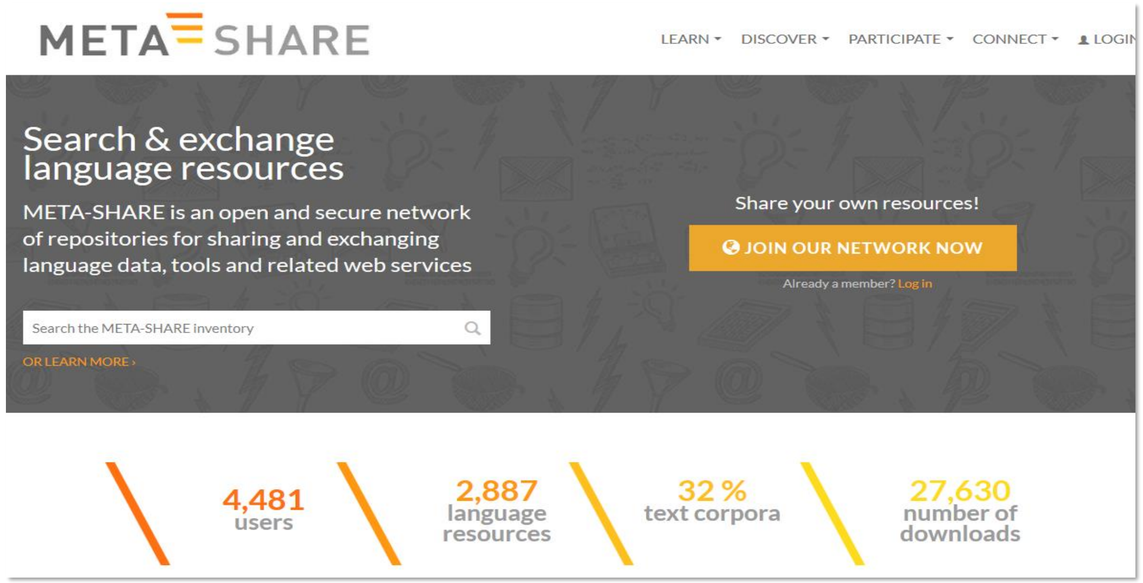
http://www.european-language-grid.eu, http://www.ilsp.gr

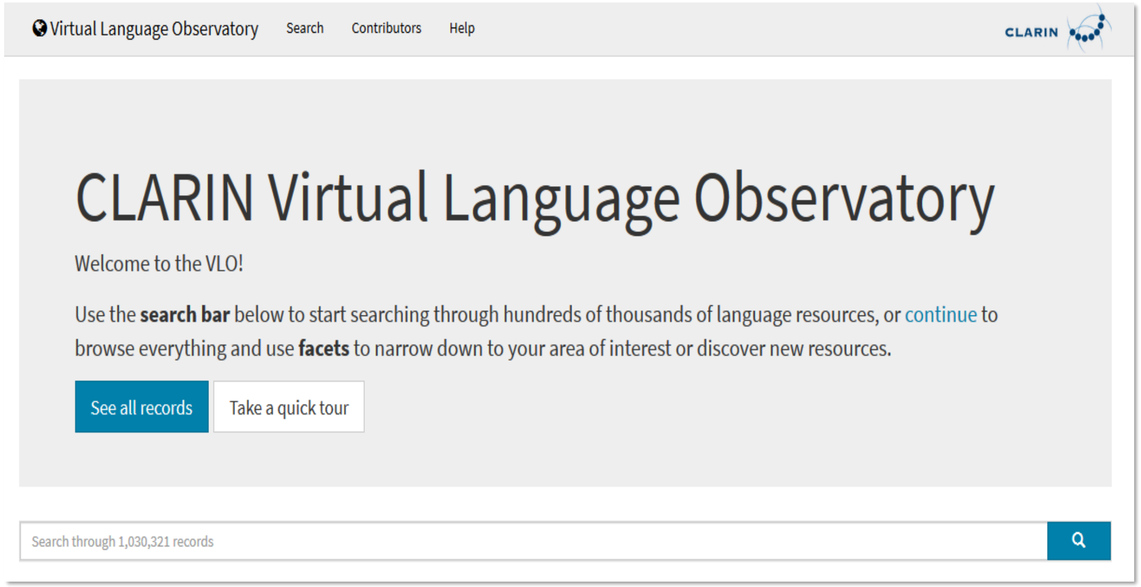
03-06-2022 Data Spaces and Semantic Interoperability

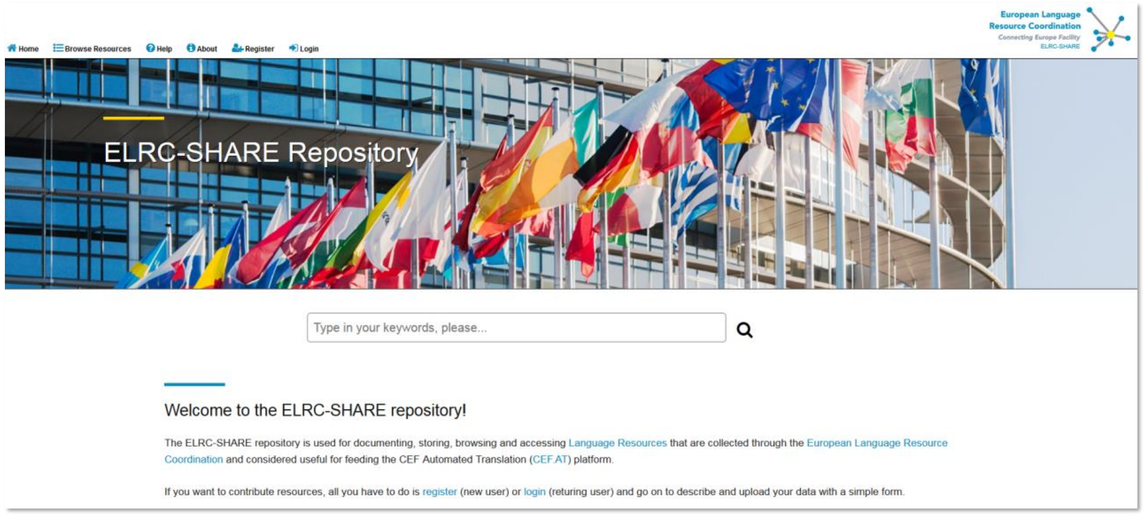
IEA ILSP ATHENA Research & Innovation Information Technologies

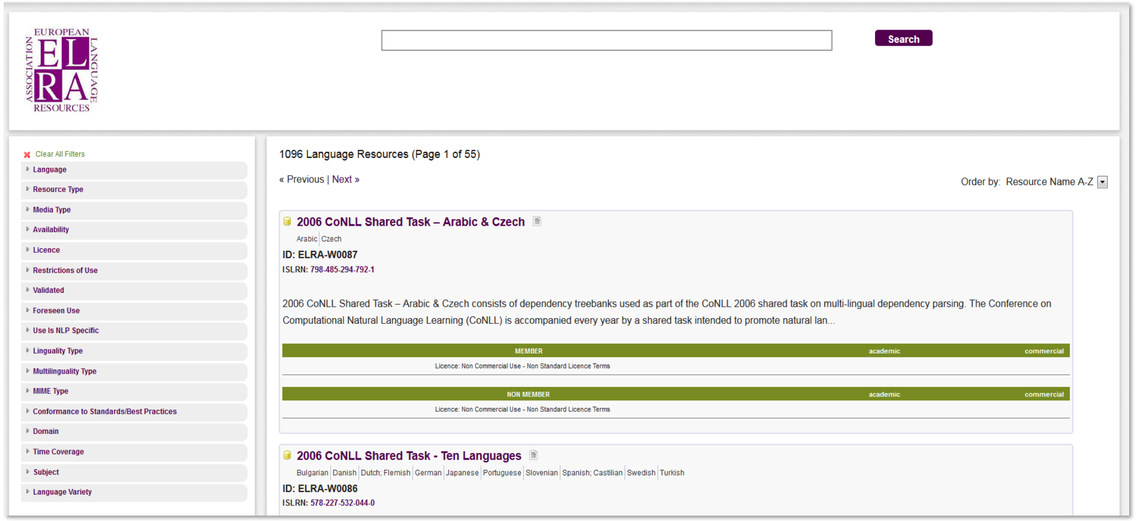
- **Institute for Language and Speech Processing**
  - Part of **Athena R.C.**, the biggest research center dedicated to Computer Science and Information Technology in Greece
  - Single dedicated Language and Speech Institute in Greece
- **Ideas and insights at this meeting from the viewpoint of**
  - Data “producer” – for specific language technology development/applications, e.g. machine translation, aspect-based sentiment analysis
  - LT developer - e.g. text analytics and information extraction, machine translation, speech technologies (recognition/synthesis/analytics), sign technologies
  - Platform and infrastructure design, development and operation

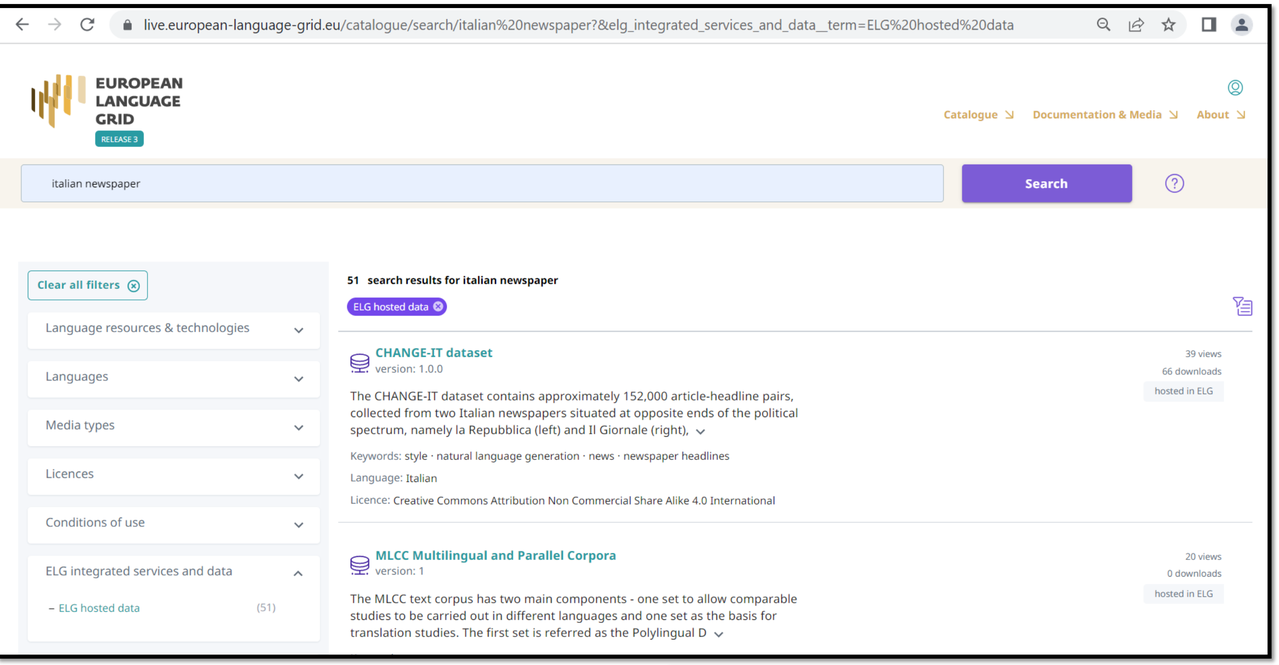
A bit of data spaces/infrastructures/platforms related history

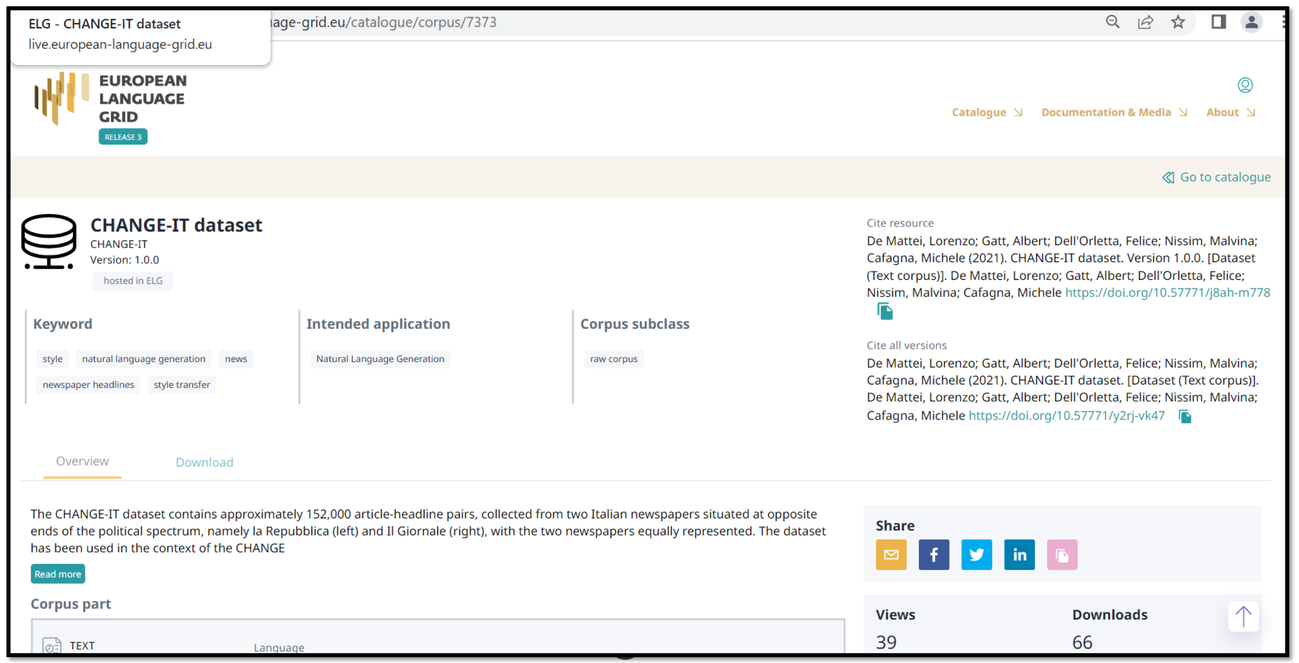


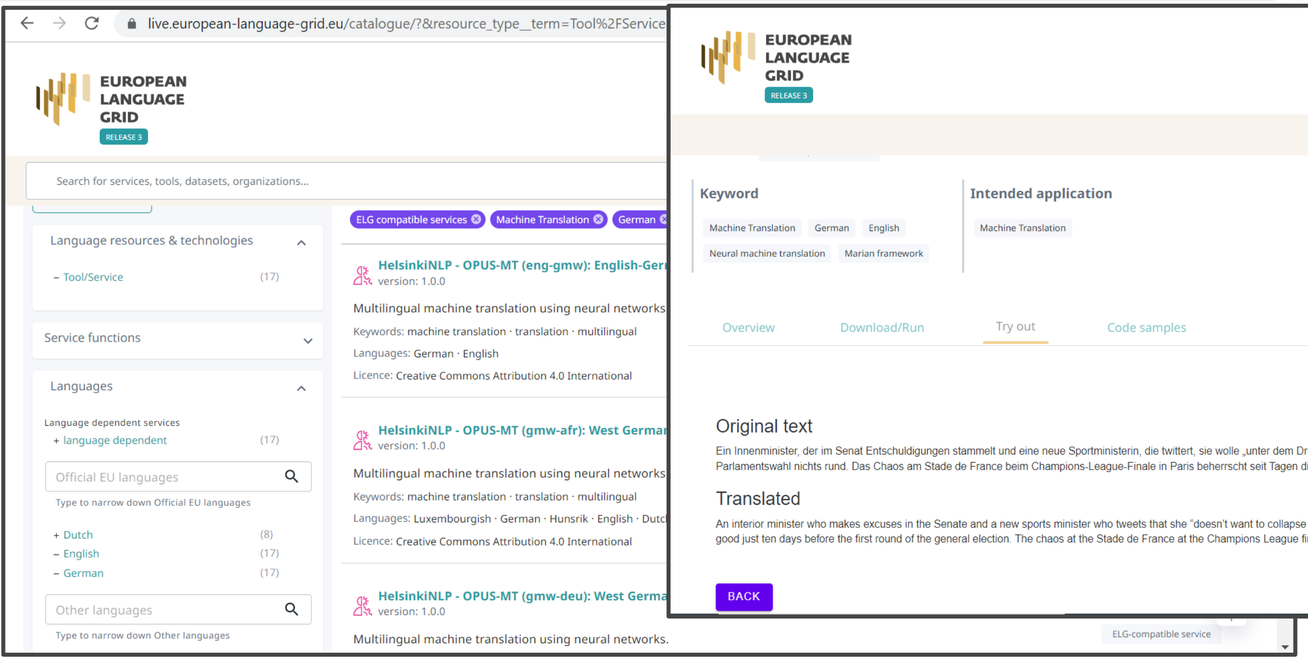


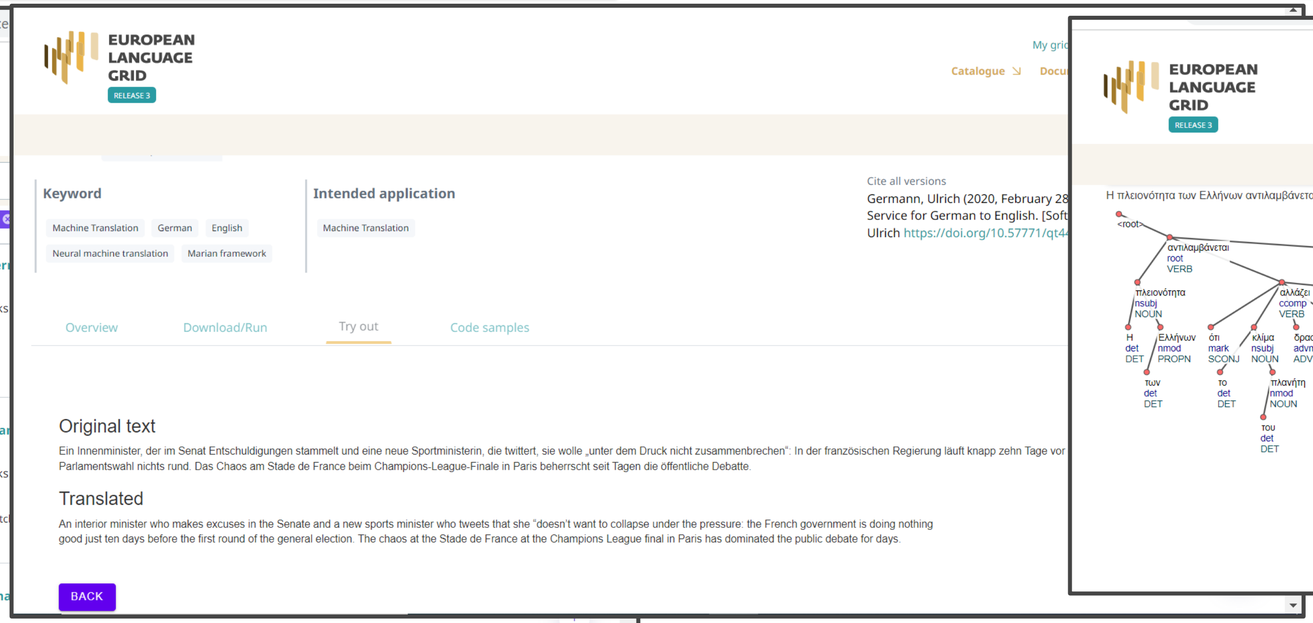


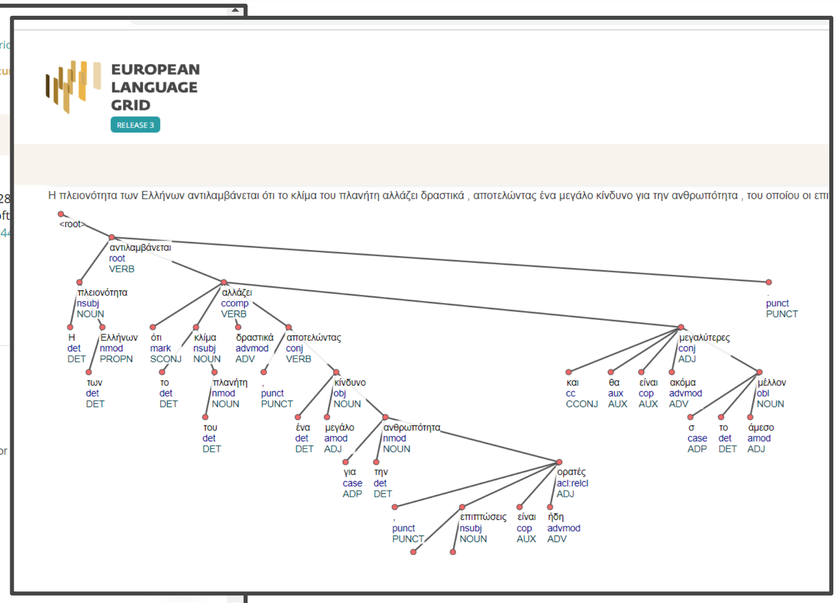



live.european-language-grid.eu/catalogue/search/italian%20newspaper?&elg_integrated_services_and_data__term=ELG%20hosted%20data
EUROPEAN LANGUAGE GRID
RELEASE 3
Catalogue
Documentation & Media
About
italian newspaper
Search
Clear all filters
Language resources & technologies
Languages
Media types
Licences
Conditions of use
ELG integrated services and data
ELG hosted data (51)
51 search results for italian newspaper
ELG hosted data
CHANGE-IT dataset
version: 1.0.0
39 views
66 downloads
hosted in ELG
The CHANGE-IT dataset contains approximately 152,000 article-headline pairs, collected from two Italian newspapers situated at opposite ends of the political spectrum, namely la Repubblica (left) and Il Giornale (right),
Keywords: style · natural language generation · news · newspaper headlines
Language: Italian
Licence: Creative Commons Attribution Non Commercial Share Alike 4.0 International
MLCC Multilingual and Parallel Corpora
version: 1
20 views
0 downloads
hosted in ELG
The MLCC text corpus has two main components - one set to allow comparable studies to be carried out in different languages and one set as the basis for translation studies. The first set is referred as the Polylingual D


ELG - CHANGE-IT dataset
live.european-language-grid.eu
EUROPEAN LANGUAGE GRID
RELEASE 3
Catalogue
Documentation & Media
About
Go to catalogue
CHANGE-IT dataset
CHANGE-IT
Version: 1.0.0
hosted in ELG
Keyword
style
natural language generation
news
newspaper headlines
style transfer
Intended application
Natural Language Generation
Corpus subclass
raw corpus
Cite resource
De Mattei, Lorenzo; Gatt, Albert; Dell'Orletta, Felice; Nissim, Malvina; Cafagna, Michele (2021). CHANGE-IT dataset. Version 1.0.0. [Dataset (Text corpus)]. De Mattei, Lorenzo; Gatt, Albert; Dell'Orletta, Felice; Nissim, Malvina; Cafagna, Michele https://doi.org/10.57771/j8ah-m778
Cite all versions
De Mattei, Lorenzo; Gatt, Albert; Dell'Orletta, Felice; Nissim, Malvina; Cafagna, Michele (2021). CHANGE-IT dataset. [Dataset (Text corpus)]. De Mattei, Lorenzo; Gatt, Albert; Dell'Orletta, Felice; Nissim, Malvina; Cafagna, Michele https://doi.org/10.57771/y2rj-vk47
Overview
Download
The CHANGE-IT dataset contains approximately 152,000 article-headline pairs, collected from two Italian newspapers situated at opposite ends of the political spectrum, namely la Repubblica (left) and Il Giornale (right), with the two newspapers equally represented. The dataset has been used in the context of the CHANGE
Read more
Corpus part
TEXT
Language
Share
Views
39
Downloads
66

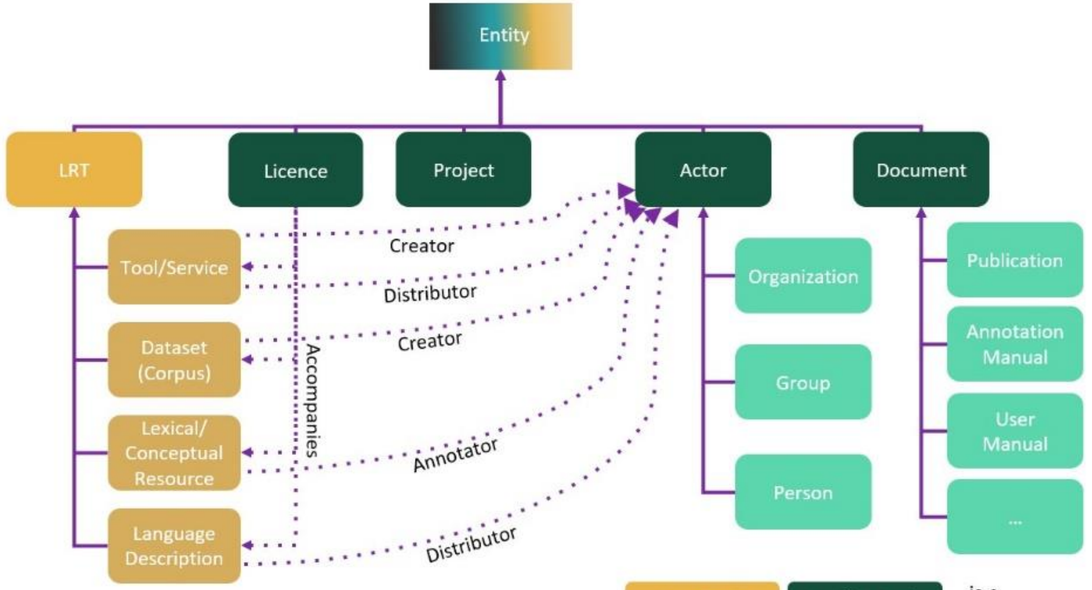

live.european-language-grid.eu/catalogue/?&resource_type__term=Tool%2FService
EUROPEAN LANGUAGE GRID
RELEASE 3
Search for services, tools, datasets, organizations...
ELG compatible services
Machine Translation
German
Language resources & technologies
Tool/Service (17)
Service functions
Languages
Language dependent services
language dependent (17)
Official EU languages
Type to narrow down Official EU languages
Dutch (8)
English (17)
German (17)
Other languages
Type to narrow down Other languages
HelsinkiNLP - OPUS-MT (eng-gmw): English-Ger
version: 1.0.0
Multilingual machine translation using neural networks
Keywords: machine translation · translation · multilingual
Languages: German · English
Licence: Creative Commons Attribution 4.0 International
HelsinkiNLP - OPUS-MT (gmw-afr): West Germa
version: 1.0.0
Multilingual machine translation using neural networks
Keywords: machine translation · translation · multilingual
Languages: Luxembourgish · German · Hunsrik · English · Dutc
Licence: Creative Commons Attribution 4.0 International
HelsinkiNLP - OPUS-MT (gmw-deu): West Germa
version: 1.0.0
Multilingual machine translation using neural networks.
ELG-compatible service

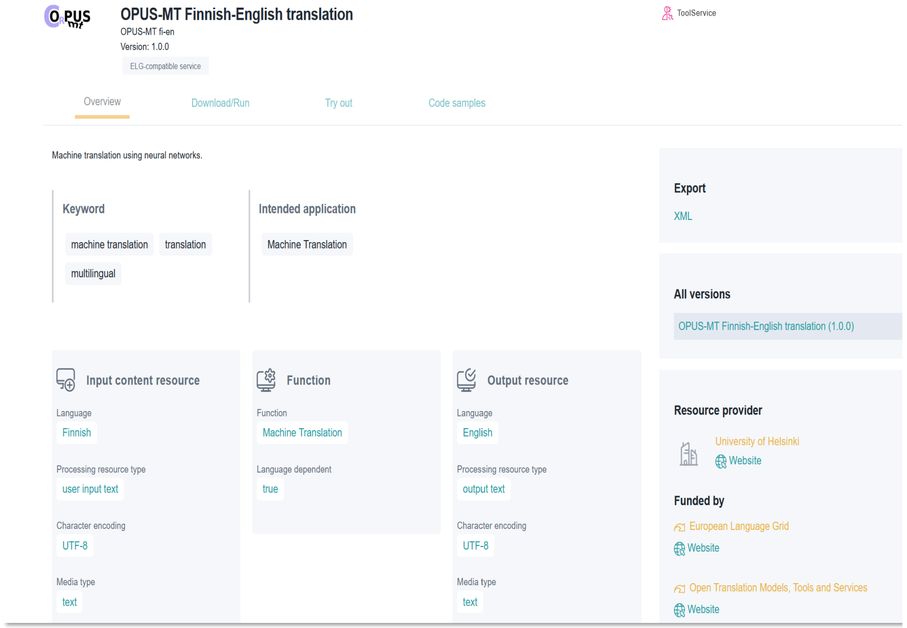

EUROPEAN LANGUAGE GRID
RELEASE 3
Catalogue
Keyword
Machine Translation
German
English
Neural machine translation
Marian framework
Intended application
Machine Translation
Cite all versions
Germann, Ulrich (2020, February 28
Service for German to English. [Soft
Ulrich https://doi.org/10.57771/qt4
Overview
Download/Run
Try out
Code samples
Original text
Ein Innenminister, der im Senat Entschuldigungen stammelt und eine neue Sportministerin, die twittert, sie wolle „unter dem Druck nicht zusammenbrechen". In der französischen Regierung läuft knapp zehn Tage vor
Parlamentswahl nichts rund. Das Chaos am Stade de France beim Champions-League-Finale in Paris beherrscht seit Tagen die öffentliche Debatte.
Translated
An interior minister who makes excuses in the Senate and a new sports minister who tweets that she "doesn't want to collapse under the pressure: the French government is doing nothing good just ten days before the first round of the general election. The chaos at the Stade de France at the Champions League final in Paris has dominated the public debate for days.
BACK

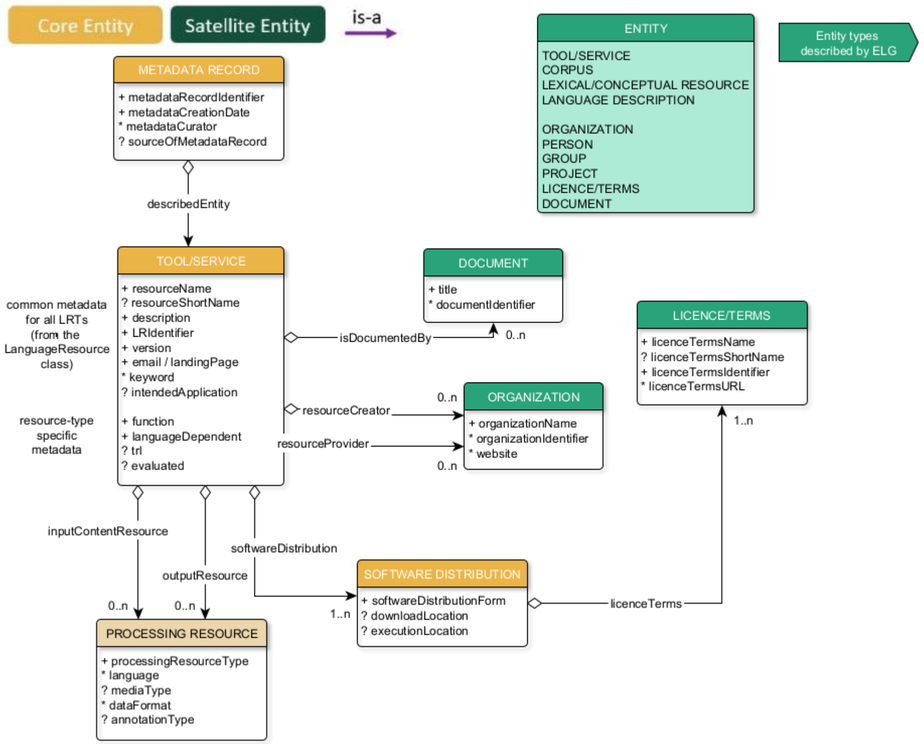

EUROPEAN LANGUAGE GRID
RELEASE 3
Η πλειονότητα των Ελλήνων αντιλαμβάνεται ότι το κλίμα του πλανήτη αλλάζει δραστικά , αποτελώντας ένα μεγάλο κίνδυνο για την ανθρωπότητα , του οποίου οι επι

Asset descriptions

- LT community members
  - can **document** their assets by providing formal descriptions compliant with a dedicated metadata schema
  - describe & provide access (remote or integrated in ELG) **LRTs**
  - describe **organisations** & **projects**
  - **linked with each other**
  - must register and get authenticated as providers



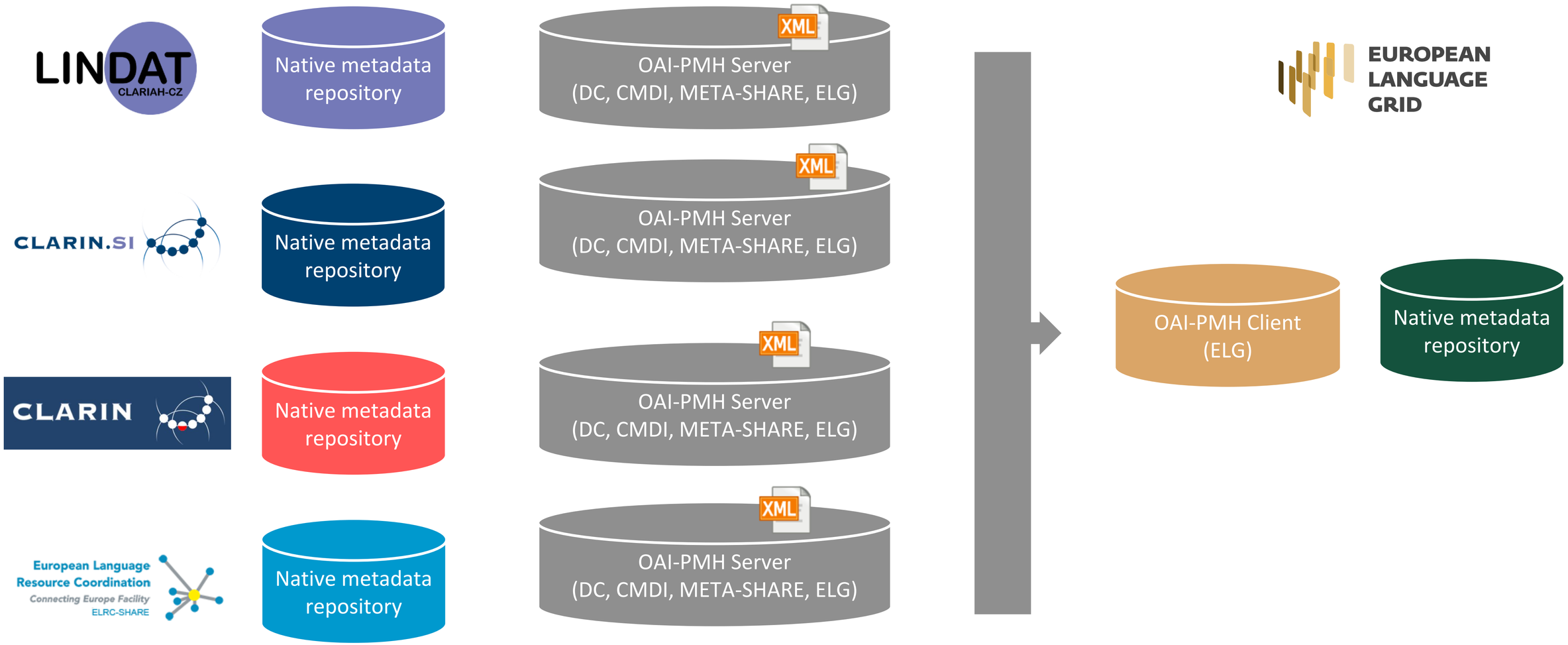




ELG in the wider LT and AI ecosystem

- ELG is **building bridges** to existing platforms/infrastructures
  - Mainly in terms of metadata-based descriptions
  - Based on **open protocols** (OAI-PMH), or **APIs** offered by the platform or infrastructure providers
  - Respecting their own policies
- ELG also as infrastructural arm of ELE
  - using a mixture of automatic and collaborative population of the ELG catalogue

Community repositories



ML/NLP repositories

EUROPEAN LANGUAGE GRID

Native metadata repository

Metadata API (Hugging Face)

Native metadata repository

Collection

Download from API

Harmonisation

Treatment of mandatory elements and controlled vocabularies

Ingestion

Loading of resulting metadata records to ELG

01

02

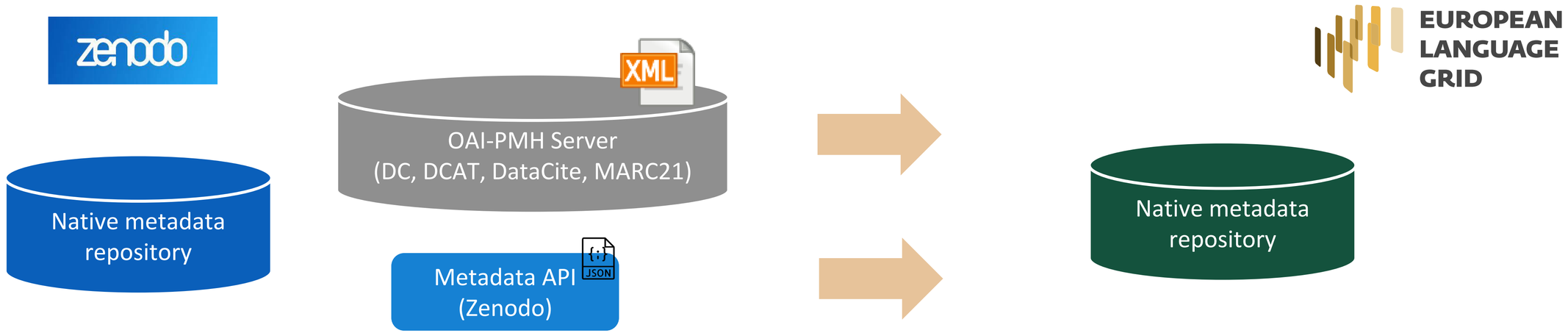
Mapping

Mapping of elements to ELG

Conversion

Conversion to ELG metadata

General repositories



Collection

Download from OAI-PMH server

Selection

Filtering of entries

Conversion

Conversion from DCAT to ELG relaxed

Validation

Targeted inspection and metadata enrichment

Ingestion

Loading of resulting metadata records to ELG

Metadata aggregation - interoperability challenges

- Different **types** (goals/roles) of repositories: institutional / disciplinary / general repositories
- Different metadata **schemas** due to different purposes (e.g., preservation, sharing, processing, …), different resource types, different user needs & practices
- Different platform/infrastructure **architectures** – from fully federated to fully centralised
- Different **policies**, access rights, user access scenarios
- Different **data types** & sets of values for similar elements (e.g. "language" value as free text vs different **controlled vocabularies**)
- Different **granularity level** of metadata schema and unit of data (e.g. language data for anthropological studies vs language tech development)
- Different community/application **needs** for specific elements (e.g. "language" and "resource type")
- ...

Elements of solutions?

- "Open shared semantic space" for vocabularies
  - A common place for vocabularies and their concepts (cf. Fair semantics recommendations)
  - Linking vocabulary concepts (properties, classes, instances) with semantic relations clearly documented
    - Starting from elements of standard vocabularies (DCAT, schema.org, DC, etc.) linked between them
  - Elements/Values from community-specific vocabularies
    - maintained and curated by communities
    - links when possible with other elements (in other domain specific or general vocabularies)
- Technically, cross-platform sharing of metadata records
  - cross-platform search instead of a central single endpoint –pros & cons
  - platforms/repos expose their full metadata and each search API selects only those of interest
  - metadata converted (search API – exposing catalogue) through the open shared vocabularies

European Language Grid

Thank you!

Stelios Piperidis, ILSP/ Athena RC

spip@athenarc.gr

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

03-06-2022, Data Spaces and Semantic Interoperability

http://www.european-language-grid.eu