EUROPEAN
LANGUAGE
GRID

# How to integrate services or data sets into the ELG platform

Dimitris Galanis ("ATHENA" R.C)

23-06-2020 1st Regional ELG Workshop: Switzerland, Austria, Germany – Zürich, Switzerland (virtual workshop)
http://www.european-language-grid.eu

# ELG platform

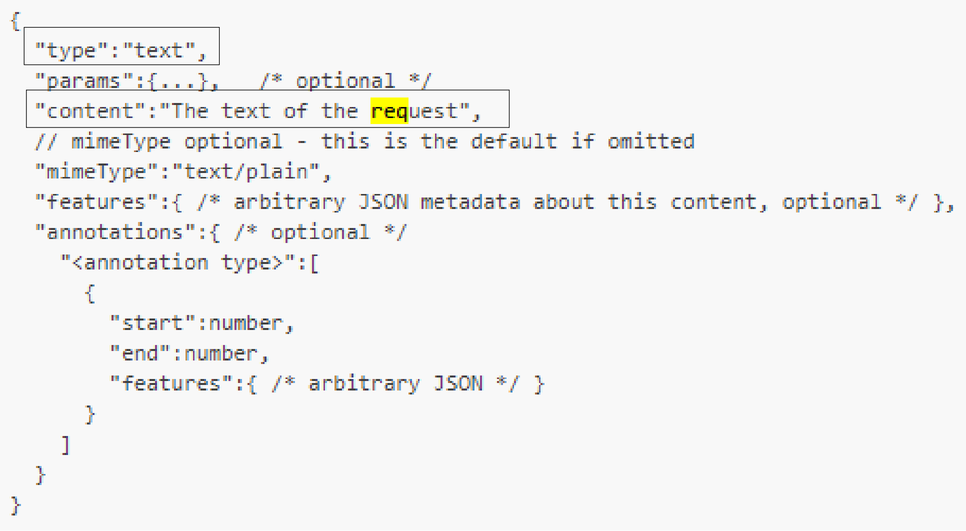
- It is a web-based platform and runs at: https://live.european-language-grid.eu
  - A catalogue for resources: LT services, data sets, lexical/conceptual resources, language descriptions
  - Each resource is documented with the required metadata
  - “Published” resources are visible by anyone; i.e. log-in, is not required
- What is required to be a resource provider in ELG
  - Register to the platform
  - ELG admins should assign you the “provider” role (email to contact@european-language-grid.eu)
  - Upload -> Review process …

# Provide a functional LT service

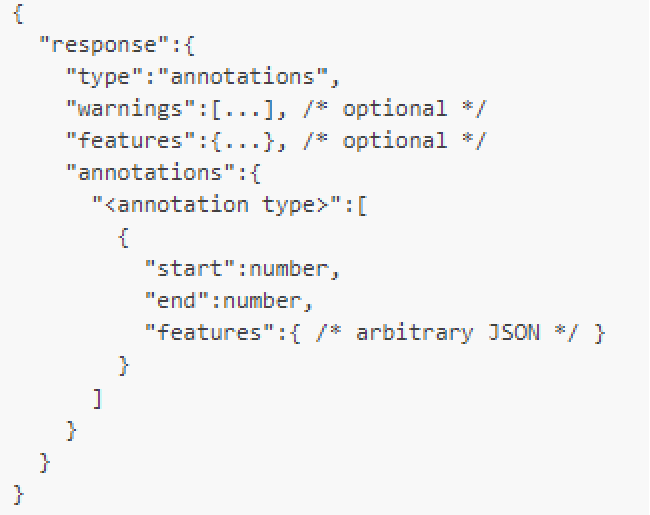
- Currently, ELG supports the integration of tools/services that fall into one of the following broad categories:
  - Information Extraction (IE)
  - Text Classification (TC)
  - Machine Translation (MT)
  - Automatic Speech Recognition (ASR)
  - Text to Speech Generation (TTS)
- Documentation: https://european-language-grid.readthedocs.io/en/release1.0.0/all/RegisterFunc.html

# Provide a functional LT service: Technical Requirements

- **Expose an ELG compatible endpoint:**
  - You MUST create an application that exposes an HTTP endpoint for the LT tool(s).
  - The application MUST consume requests that follow the ELG JSON format, call the underlying LT tool and produce responses again in the ELG JSON format (ELG LT Internal API)
  - E.g. for services that take as input  plain text and return annotations (IE).



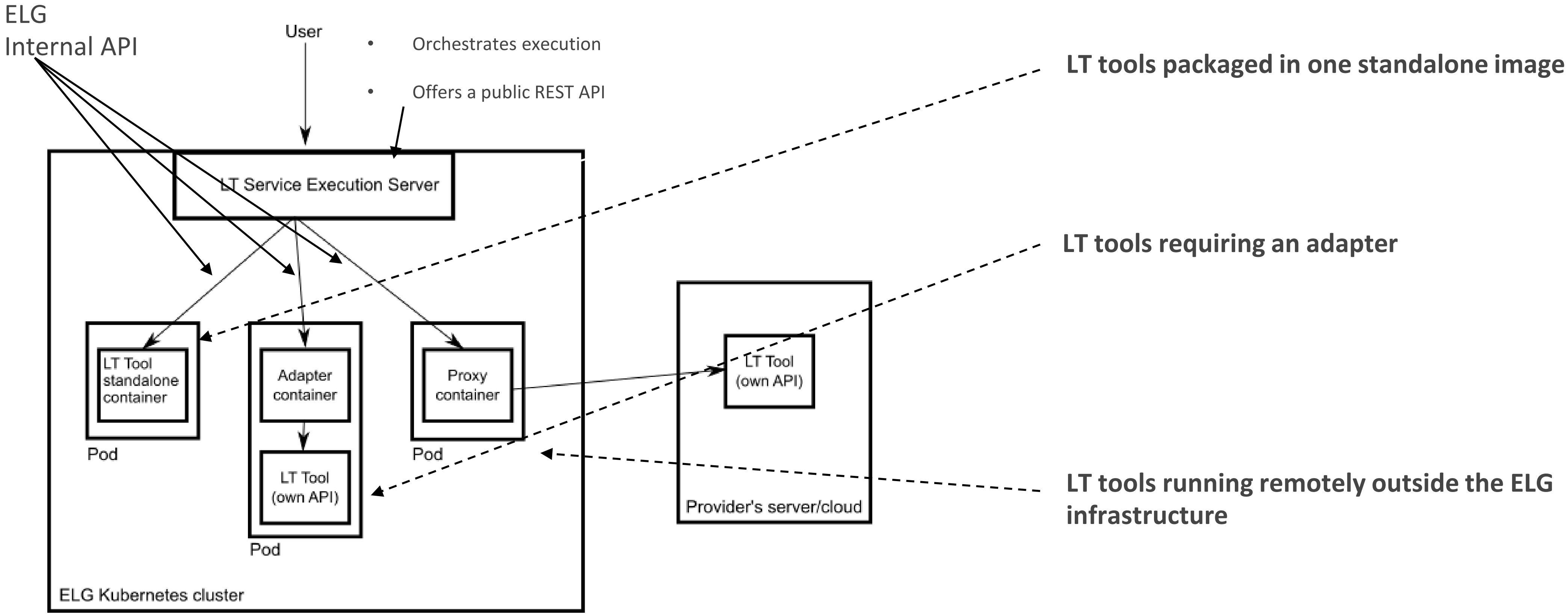
```
{
  "type":"text",
  "params":{...},   /* optional */
  "content":"The text of the request",
  // mimeType optional - this is the default if omitted
  "mimeType":"text/plain",
  "features":{ /* arbitrary JSON metadata about this content, optional */ },
  "annotations":{ /* optional */
    "<annotation type>":[
      {
        "start":number,
        "end":number,
        "features":{ /* arbitrary JSON */ }
      }
    ]
  }
}
```



```
{
  "response":{
    "type":"annotations",
    "warnings":[...], /* optional */
    "features":{...}, /* optional */
    "annotations":{
      "<annotation type>":[
        {
          "start":number,
          "end":number,
          "features":{ /* arbitrary JSON */ }
        }
      ]
    }
  }
}
```

# Provide a functional LT service: Technical Requirements

- **Dockerization**: You MUST Dockerize the application and upload the respective image(s) in a Docker Registry, e.g. GitLab. Three integration options are available.



ELG
Internal API

- Orchestrates execution
- Offers a public REST API

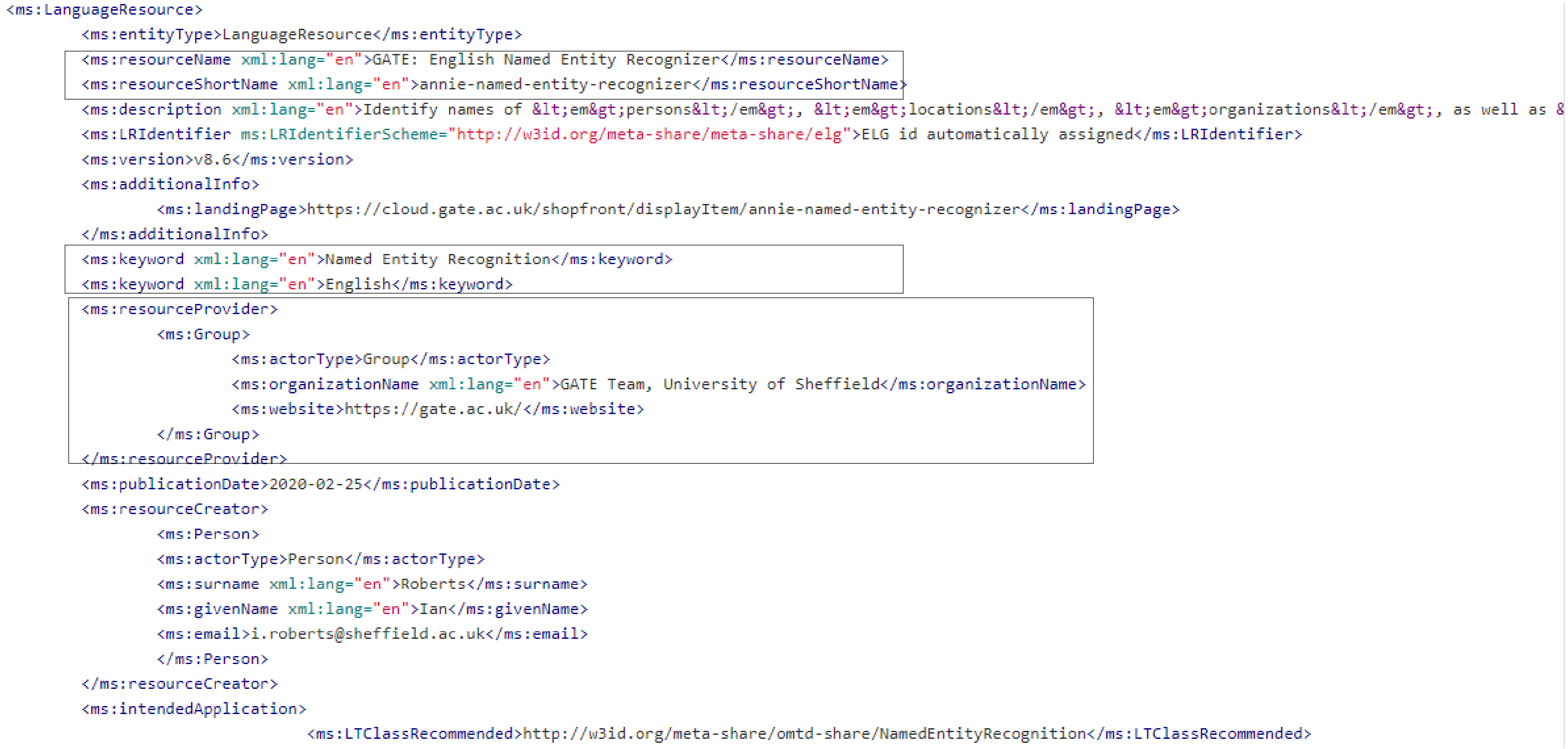
LT tools packaged in one standalone image

LT tools requiring an adapter

LT tools running remotely outside the ELG infrastructure

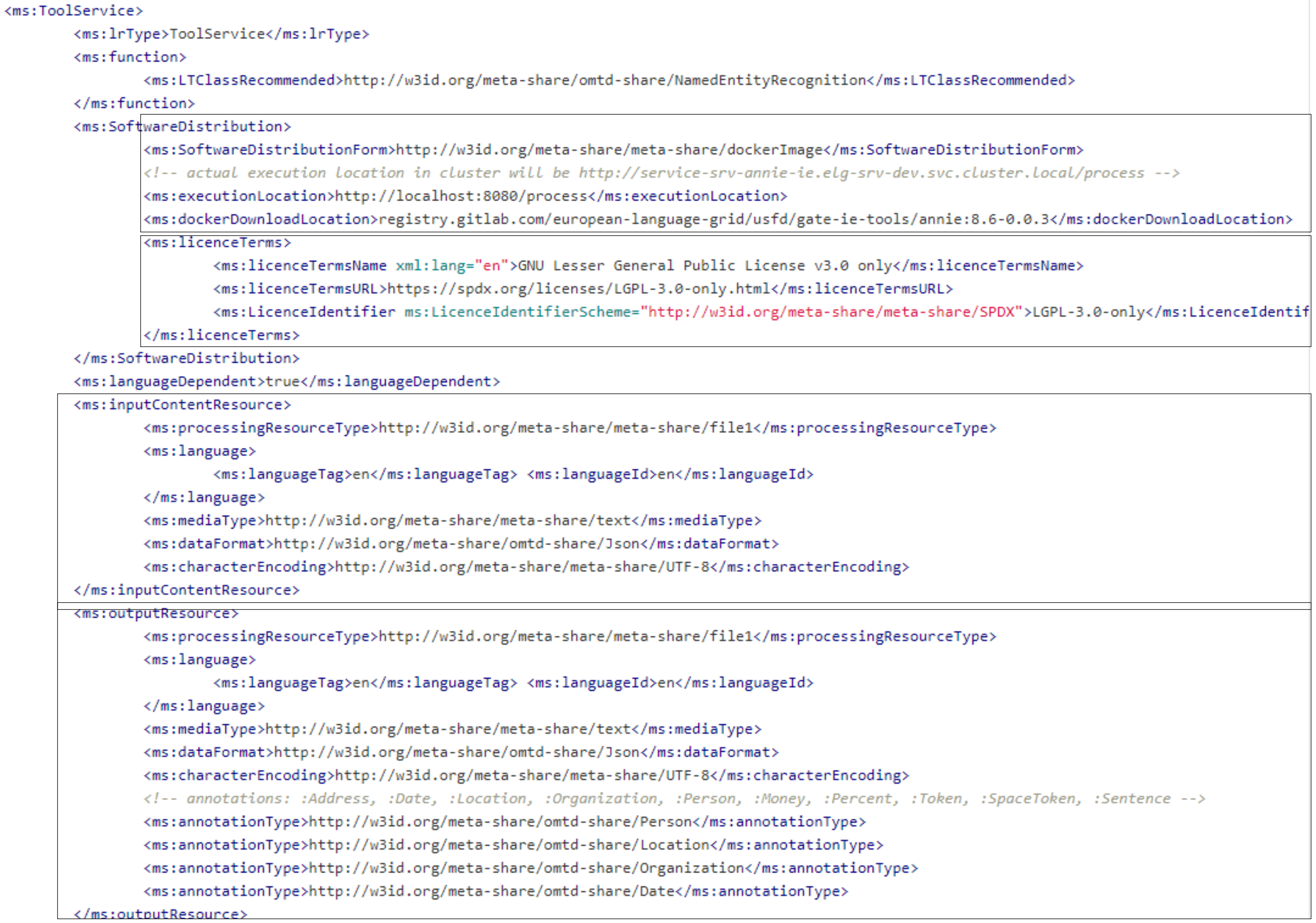
# Describe the LT service according to the ELG metadata schema

- Full metadata record example in XML: https://gitlab.com/european-language-grid/platform/metadatarecords/-/blob/master/R1a/USFD/hosted/annie-named-entity-recognizer.xml



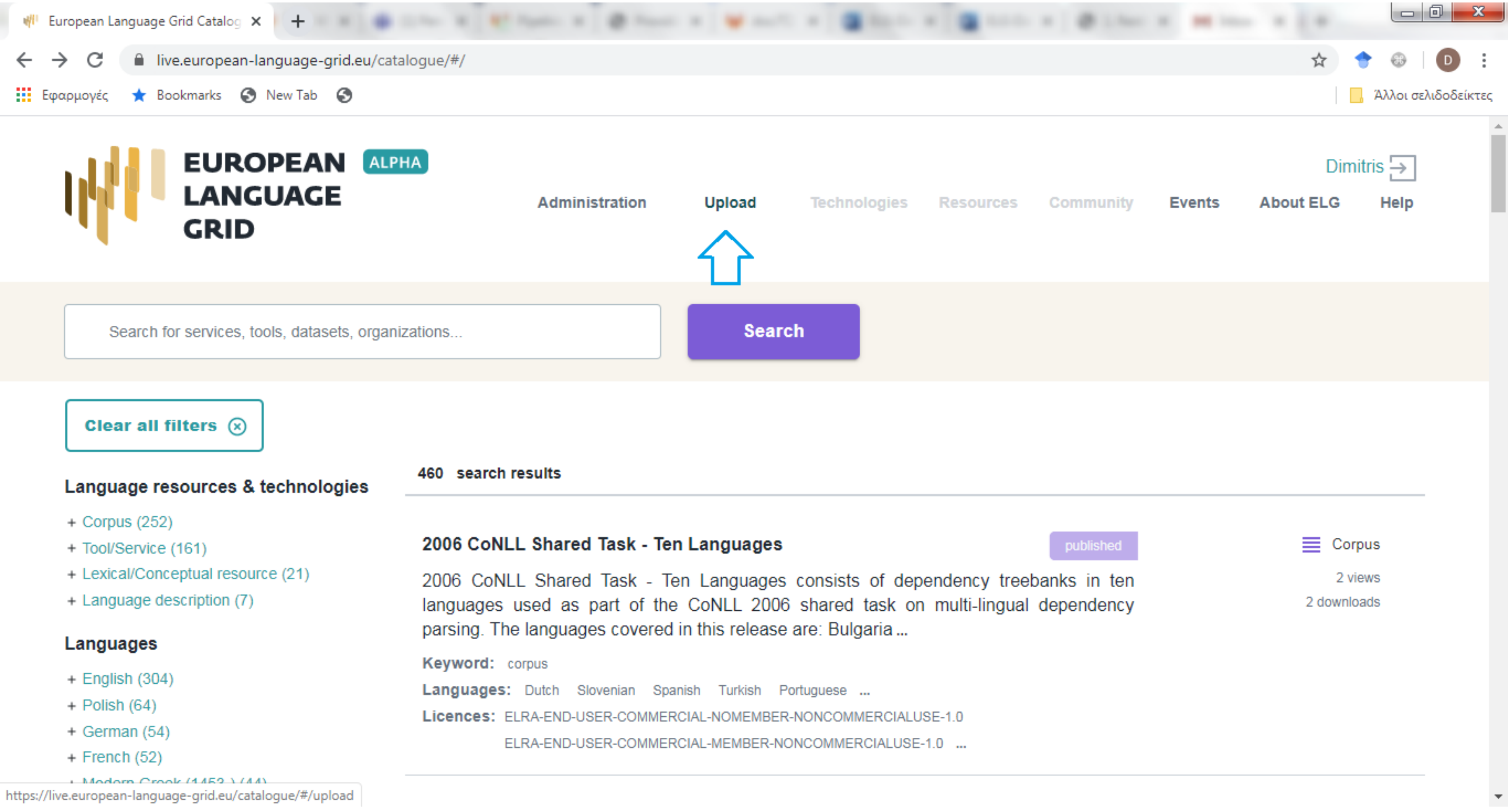
```
<ms:LanguageResource>
        <ms:entityType>LanguageResource</ms:entityType>
        <ms:resourceName xml:lang="en">GATE: English Named Entity Recognizer</ms:resourceName>
        <ms:resourceShortName xml:lang="en">annie-named-entity-recognizer</ms:resourceShortName>
        <ms:description xml:lang="en">Identify names of &lt;em&gt;persons&lt;/em&gt;, &lt;em&gt;locations&lt;/em&gt;, &lt;em&gt;organizations&lt;/em&gt;, as well as &
        <ms:LRIdentifier ms:LRIdentifierScheme="http://w3id.org/meta-share/meta-share/elg">ELG id automatically assigned</ms:LRIdentifier>
        <ms:version>v8.6</ms:version>
        <ms:additionalInfo>
                <ms:landingPage>https://cloud.gate.ac.uk/shopfront/displayItem/annie-named-entity-recognizer</ms:landingPage>
        </ms:additionalInfo>
        <ms:keyword xml:lang="en">Named Entity Recognition</ms:keyword>
        <ms:keyword xml:lang="en">English</ms:keyword>
        <ms:resourceProvider>
                <ms:Group>
                        <ms:actorType>Group</ms:actorType>
                        <ms:organizationName xml:lang="en">GATE Team, University of Sheffield</ms:organizationName>
                        <ms:website>https://gate.ac.uk/</ms:website>
                </ms:Group>
        </ms:resourceProvider>
        <ms:publicationDate>2020-02-25</ms:publicationDate>
        <ms:resourceCreator>
                <ms:Person>
                <ms:actorType>Person</ms:actorType>
                <ms:surname xml:lang="en">Roberts</ms:surname>
                <ms:givenName xml:lang="en">Ian</ms:givenName>
                <ms:email>i.roberts@sheffield.ac.uk</ms:email>
                </ms:Person>
        </ms:resourceCreator>
        <ms:intendedApplication>
                                <ms:LTClassRecommended>http://w3id.org/meta-share/omtd-share/NamedEntityRecognition</ms:LTClassRecommended>
```

# Describe the LT service according to the ELG metadata schema



```
<ms:ToolService>
        <ms:lrType>ToolService</ms:lrType>
        <ms:function>
                <ms:LTClassRecommended>http://w3id.org/meta-share/omtd-share/NamedEntityRecognition</ms:LTClassRecommended>
        </ms:function>
        <ms:SoftwareDistribution>
                <ms:SoftwareDistributionForm>http://w3id.org/meta-share/meta-share/dockerImage</ms:SoftwareDistributionForm>
                <!-- actual execution location in cluster will be http://service-srv-annie-ie.elg-srv-dev.svc.cluster.local/process -->
                <ms:executionLocation>http://localhost:8080/process</ms:executionLocation>
                <ms:dockerDownloadLocation>registry.gitlab.com/european-language-grid/usfd/gate-ie-tools/annie:8.6-0.0.3</ms:dockerDownloadLocation>
                <ms:licenceTerms>
                        <ms:licenceTermsName xml:lang="en">GNU Lesser General Public License v3.0 only</ms:licenceTermsName>
                        <ms:licenceTermsURL>https://spdx.org/licenses/LGPL-3.0-only.html</ms:licenceTermsURL>
                        <ms:LicenceIdentifier ms:LicenceIdentifierScheme="http://w3id.org/meta-share/meta-share/SPDX">LGPL-3.0-only</ms:LicenceIdentif
                </ms:licenceTerms>
        </ms:SoftwareDistribution>
        <ms:languageDependent>true</ms:languageDependent>
        <ms:inputContentResource>
                <ms:processingResourceType>http://w3id.org/meta-share/meta-share/file1</ms:processingResourceType>
                <ms:language>
                        <ms:languageTag>en</ms:languageTag> <ms:languageId>en</ms:languageId>
                </ms:language>
                <ms:mediaType>http://w3id.org/meta-share/meta-share/text</ms:mediaType>
                <ms:dataFormat>http://w3id.org/meta-share/omtd-share/Json</ms:dataFormat>
                <ms:characterEncoding>http://w3id.org/meta-share/meta-share/UTF-8</ms:characterEncoding>
        </ms:inputContentResource>
        <ms:outputResource>
                <ms:processingResourceType>http://w3id.org/meta-share/meta-share/file1</ms:processingResourceType>
                <ms:language>
                        <ms:languageTag>en</ms:languageTag> <ms:languageId>en</ms:languageId>
                </ms:language>
                <ms:mediaType>http://w3id.org/meta-share/meta-share/text</ms:mediaType>
                <ms:dataFormat>http://w3id.org/meta-share/omtd-share/Json</ms:dataFormat>
                <ms:characterEncoding>http://w3id.org/meta-share/meta-share/UTF-8</ms:characterEncoding>
                <!-- annotations: :Address, :Date, :Location, :Organization, :Person, :Money, :Percent, :Token, :SpaceToken, :Sentence -->
                <ms:annotationType>http://w3id.org/meta-share/omtd-share/Person</ms:annotationType>
                <ms:annotationType>http://w3id.org/meta-share/omtd-share/Location</ms:annotationType>
                <ms:annotationType>http://w3id.org/meta-share/omtd-share/Organization</ms:annotationType>
                <ms:annotationType>http://w3id.org/meta-share/omtd-share/Date</ms:annotationType>
        </ms:outputResource>
```

# Describe the LT service according to the ELG metadata schema



European Language Grid Catalog
live.european-language-grid.eu/catalogue/#/
Εφαρμογές
Bookmarks
New Tab
Άλλοι σελιδοδείκτες
EUROPEAN LANGUAGE GRID
ALPHA
Dimitris
Administration
Upload
Technologies
Resources
Community
Events
About ELG
Help
Search for services, tools, datasets, organizations...
Search
Clear all filters
https://live.european-language-grid.eu/catalogue/#/upload

**Language resources & technologies**

- + Corpus (252)
- + Tool/Service (161)
- + Lexical/Conceptual resource (21)
- + Language description (7)

**Languages**

- + English (304)
- + Polish (64)
- + German (54)
- + French (52)
- + Modern Greek (1453-) (44)

460 search results

**2006 CoNLL Shared Task - Ten Languages** published

Corpus

2 views

2 downloads

2006 CoNLL Shared Task - Ten Languages consists of dependency treebanks in ten languages used as part of the CoNLL 2006 shared task on multi-lingual dependency parsing. The languages covered in this release are: Bulgaria ...

**Keyword:** corpus

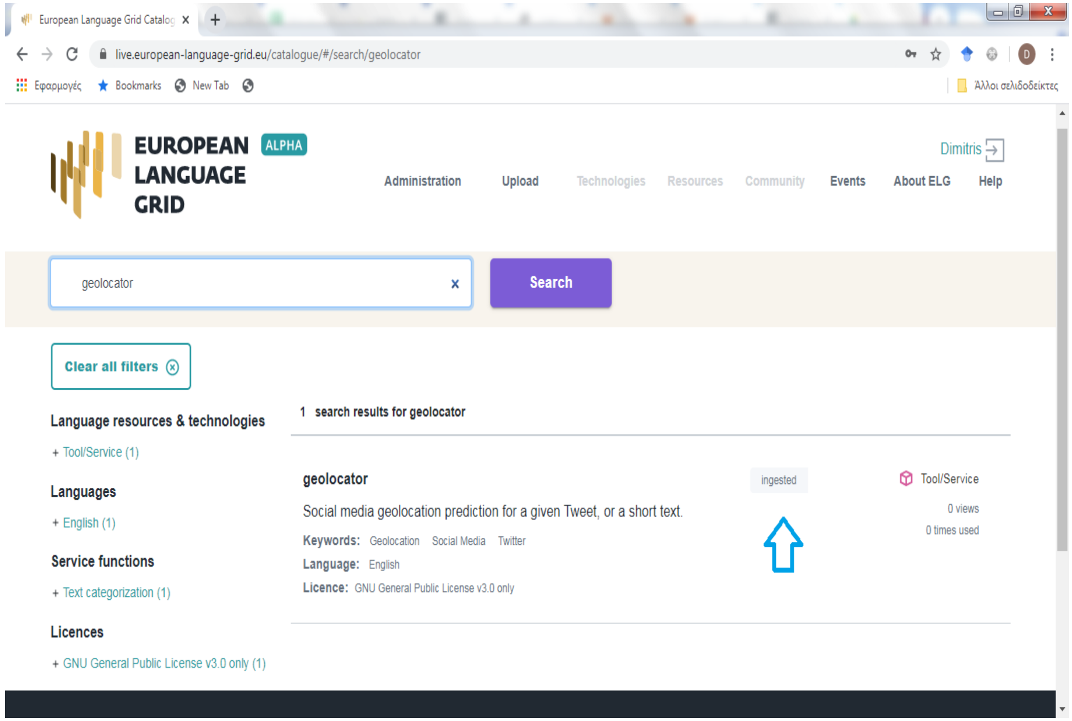
**Languages:** Dutch Slovenian Spanish Turkish Portuguese ...

**Licences:** ELRA-END-USER-COMMERCIAL-NOMEMBER-NONCOMMERCIALUSE-1.0
ELRA-END-USER-COMMERCIAL-MEMBER-NONCOMMERCIALUSE-1.0 ...

- Logged-in
- Assigned the “provider”

# Test LT service and publish (review process)

- The LT Service is assigned to a “reviewer”
- The LT tool is deployed to the ELG platform
  - The respective YAML files for kubernetes have to be created
- The LT Service is tested by the reviewer and the LT provider …
  - The service is “ingested” not “public”
  - Test via UI
  - Check container logs …
  - Troubleshoot …
- Finally it is published to the catalogue



geolocator

1 search results for geolocator

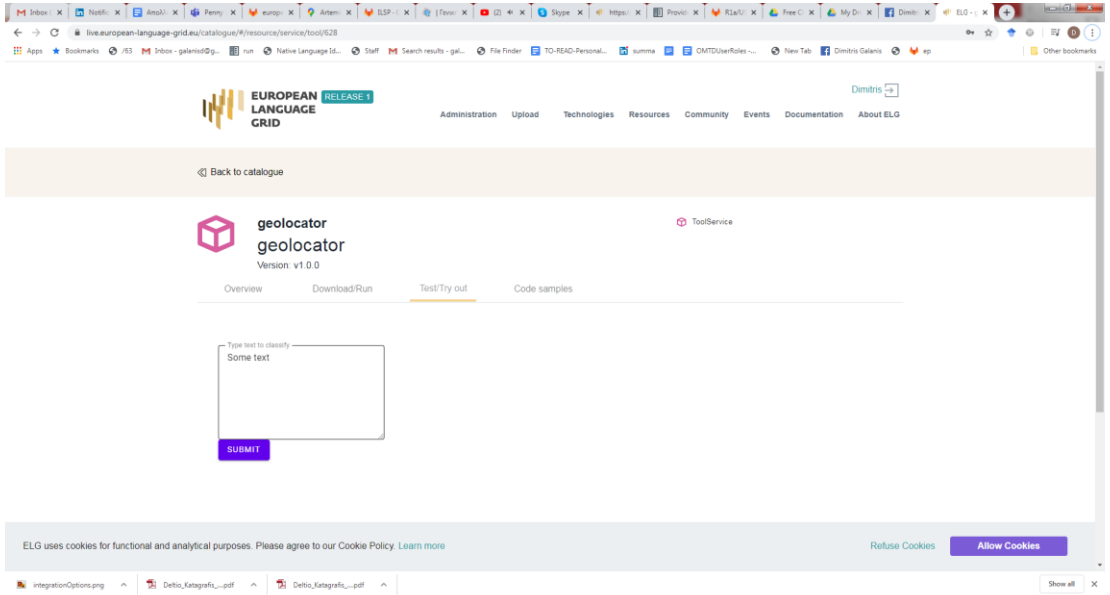
**geolocator**

Social media geolocation prediction for a given Tweet, or a short text.

**Keywords:** Geolocation Social Media Twitter

**Language:** English

**Licence:** GNU General Public License v3.0 only



**geolocator**

geolocator

Version: v1.0.0

Type text to classify: Some text

ELG uses cookies for functional and analytical purposes. Please agree to our Cookie Policy. Learn more

# Provide a functional LT service: FAQs

- The task of packaging, deploying and registering an LT service to ELG is not “trivial”
- Many questions and several difficulties arise
- FAQs: https://european-language-grid.readthedocs.io/en/release1.0.0/all/RegisterFunc.html#frequently-asked-questions
- E.g.
  - Q: How many resources will be allocated for my LT container in the k8s cluster?
  - A: By default, 512MB of RAM and half a CPU core….
  - Q: What is a k8s namespace and when should an LT Provider ask for one?
  - A: A k8s namespace is a virtual sub-cluster, which can be used to restrict access to the respective containers that run within it. You should ask for a dedicated namespace when you need to ensure isolation and security

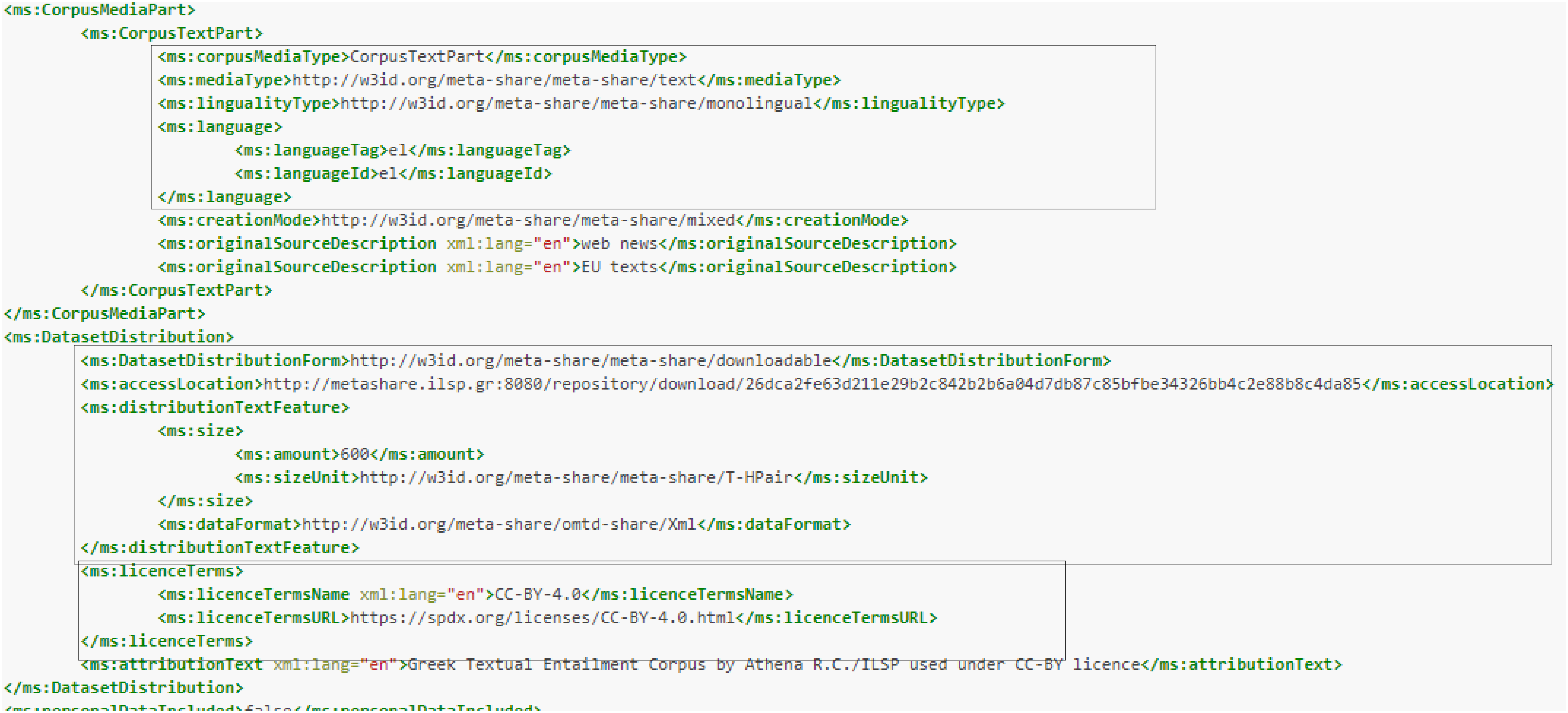
# Provide a data set

- Documentation:
  - https://european-language-grid.readthedocs.io/en/release1.0.0/all/RegisterNonFunc.html
  - https://european-language-grid.readthedocs.io/en/release1.0.0/all/RegisterCorpus.html
- Steps:
  - Register to the platform and ask (by email to contact@european-language-grid.eu) to be granted “provider” permissions;
  - Describe your dataset according to the ELG metadata schema -> upload
  - Provide access to it (?)

# Provide a data set (corpus)

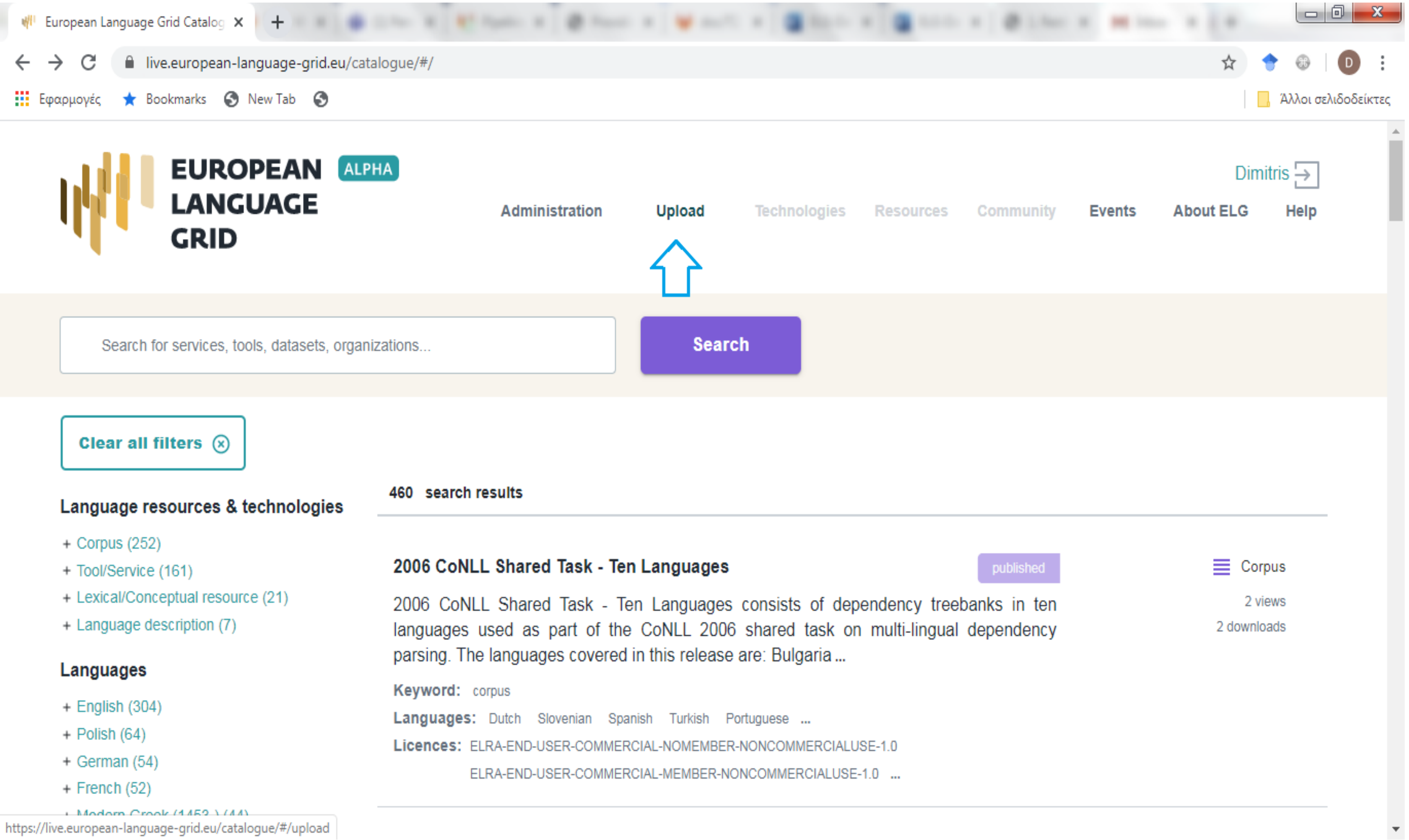
- Corpora are collections of data selected according to specific criteria:
- text corpora: monolingual, bilingual or multilingual collections of texts in a specific domain, e.g. news articles,
- corpora of audio recordings, e.g., lists of sentences recorded by individuals from a specific region with a dialect accent, etc.
- collections of videos, such as interviews with politicians
- corpora combining all of the above

# Provide a data set (example)



```
<ms:CorpusMediaPart>
        <ms:CorpusTextPart>
                <ms:corpusMediaType>CorpusTextPart</ms:corpusMediaType>
                <ms:mediaType>http://w3id.org/meta-share/meta-share/text</ms:mediaType>
                <ms:lingualityType>http://w3id.org/meta-share/meta-share/monolingual</ms:lingualityType>
                <ms:language>
                        <ms:languageTag>el</ms:languageTag>
                        <ms:languageId>el</ms:languageId>
                </ms:language>
                <ms:creationMode>http://w3id.org/meta-share/meta-share/mixed</ms:creationMode>
                <ms:originalSourceDescription xml:lang="en">web news</ms:originalSourceDescription>
                <ms:originalSourceDescription xml:lang="en">EU texts</ms:originalSourceDescription>
        </ms:CorpusTextPart>
</ms:CorpusMediaPart>
<ms:DatasetDistribution>
        <ms:DatasetDistributionForm>http://w3id.org/meta-share/meta-share/downloadable</ms:DatasetDistributionForm>
        <ms:accessLocation>http://metashare.ilsp.gr:8080/repository/download/26dca2fe63d211e29b2c842b2b6a04d7db87c85bfbe34326bb4c2e88b8c4da85</ms:accessLocation>
        <ms:distributionTextFeature>
                <ms:size>
                        <ms:amount>600</ms:amount>
                        <ms:sizeUnit>http://w3id.org/meta-share/meta-share/T-HPair</ms:sizeUnit>
                </ms:size>
                <ms:dataFormat>http://w3id.org/meta-share/omtd-share/Xml</ms:dataFormat>
        </ms:distributionTextFeature>
        <ms:licenceTerms>
                <ms:licenceTermsName xml:lang="en">CC-BY-4.0</ms:licenceTermsName>
                <ms:licenceTermsURL>https://spdx.org/licenses/CC-BY-4.0.html</ms:licenceTermsURL>
        </ms:licenceTerms>
        <ms:attributionText xml:lang="en">Greek Textual Entailment Corpus by Athena R.C./ILSP used under CC-BY licence</ms:attributionText>
</ms:DatasetDistribution>
<ms:personalDataIncluded>false</ms:personalDataIncluded>
```

# Provide a dataset
# Upload metadata


European Language Grid Catalog
live.european-language-grid.eu/catalogue/#/
Εφαρμογές
Bookmarks
New Tab
Άλλοι σελιδοδείκτες
EUROPEAN LANGUAGE GRID
ALPHA
Dimitris
Administration
Upload
Technologies
Resources
Community
Events
About ELG
Help
Search for services, tools, datasets, organizations...
Search
Clear all filters
Language resources & technologies
+ Corpus (252)
+ Tool/Service (161)
+ Lexical/Conceptual resource (21)
+ Language description (7)
Languages
+ English (304)
+ Polish (64)
+ German (54)
+ French (52)
+ Modern Greek (1453-) (44)
https://live.european-language-grid.eu/catalogue/#/upload

460 search results

**2006 CoNLL Shared Task - Ten Languages** published

Corpus
2 views
2 downloads

2006 CoNLL Shared Task - Ten Languages consists of dependency treebanks in ten languages used as part of the CoNLL 2006 shared task on multi-lingual dependency parsing. The languages covered in this release are: Bulgaria ...

**Keyword:** corpus

**Languages:** Dutch Slovenian Spanish Turkish Portuguese ...

**Licences:** ELRA-END-USER-COMMERCIAL-NOMEMBER-NONCOMMERCIALUSE-1.0
ELRA-END-USER-COMMERCIAL-MEMBER-NONCOMMERCIALUSE-1.0 ...

The dataset will be checked and published to the catalogue by a reviewer.

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).