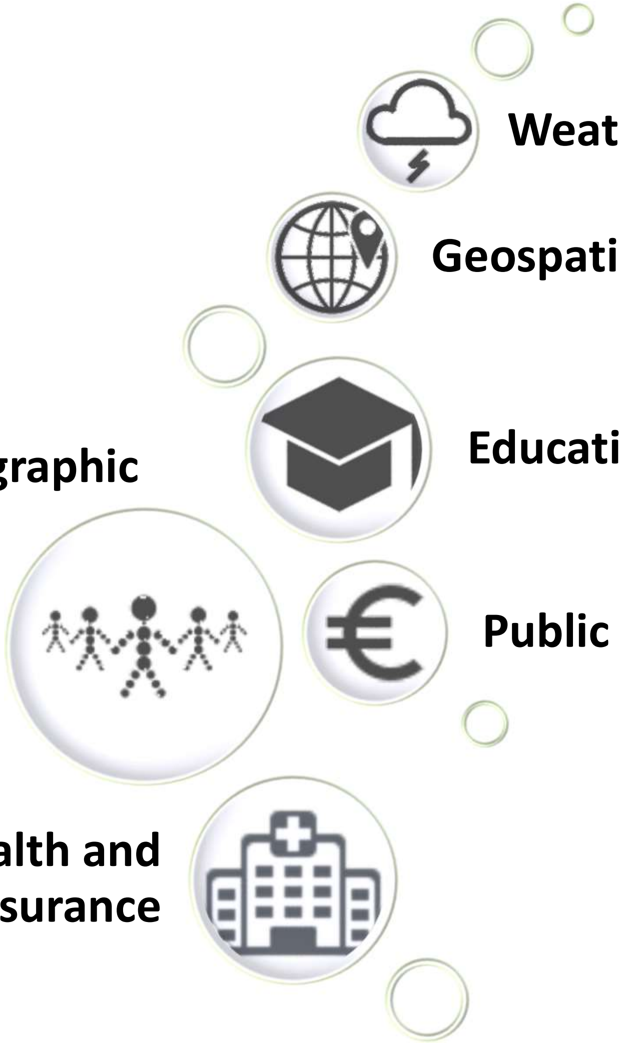
# Preparing and sharing data with the ELRC-SHARE repository
# and what happens next

Maria Giagkou
Institute for Language and Speech Processing / Athena R.C.
ELRC

# The notion of data



Weather / climate

Geospatial

Demographic

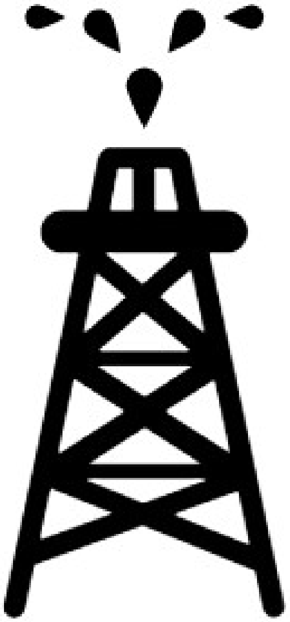
Educational

Public spending

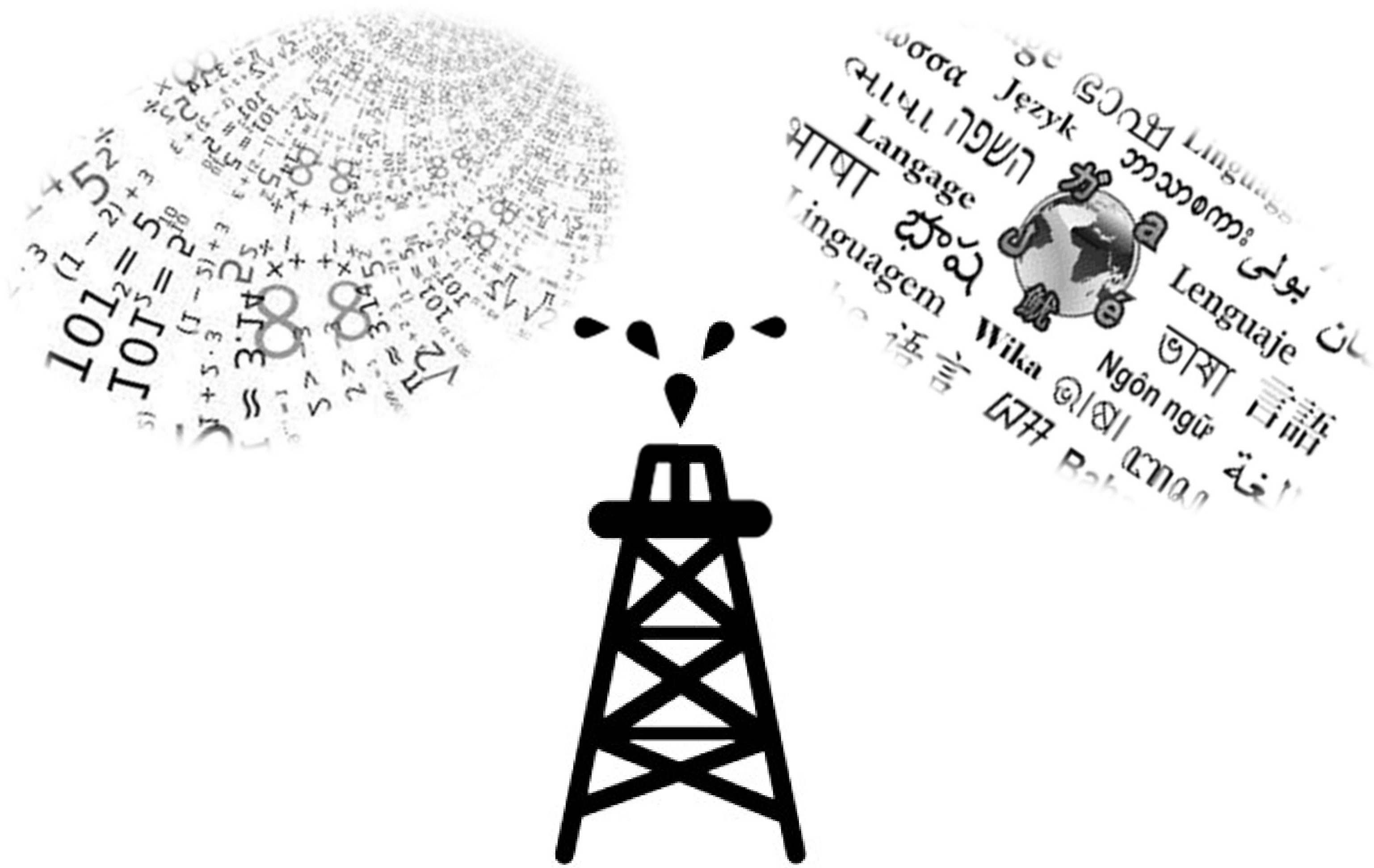
Public health and
insurance

# The notion of data

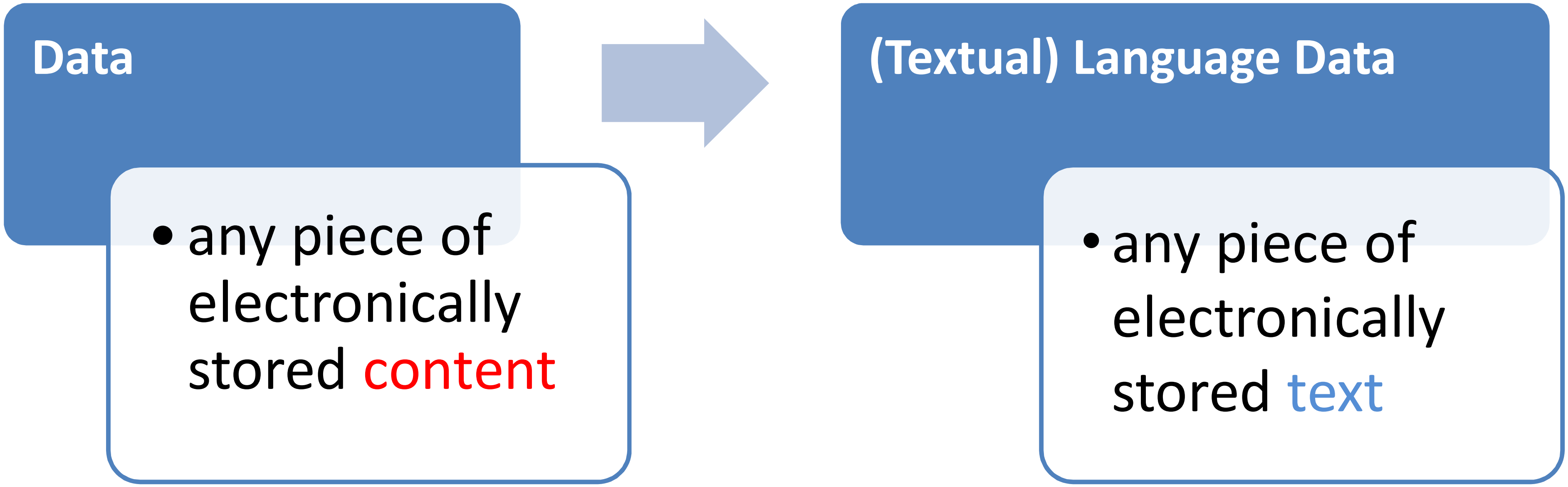
Data: the oil of the 21$^{st}$ century



# The notion of data



# The notion of  language data



**Data**

- any piece of electronically stored content

**(Textual) Language Data**

- any piece of electronically stored text

# Language data for eTranslation

## BG

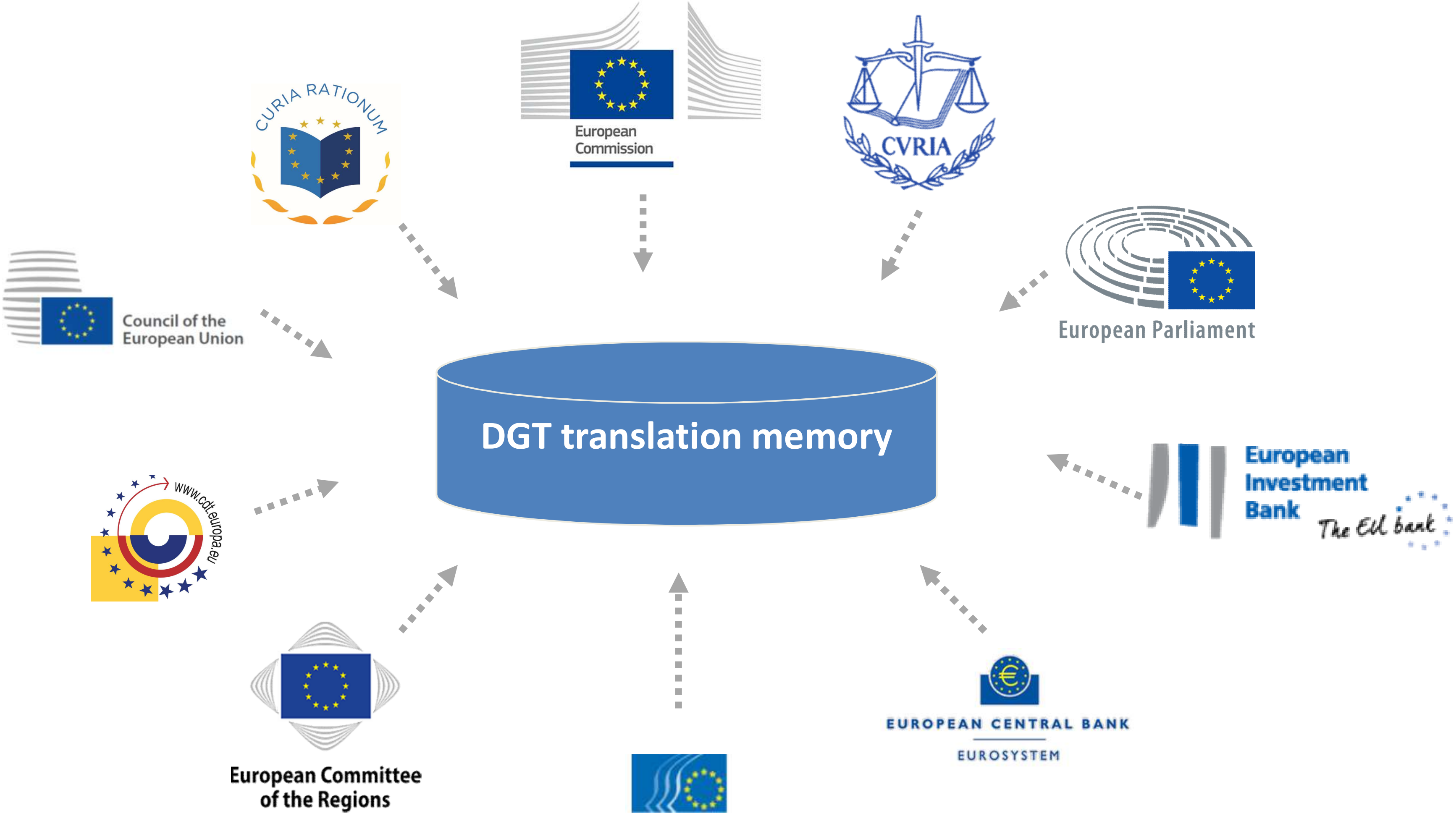
Ако не е осигурена взаимност, може да бъде отказано разрешение за двустранен превоз по редовна линия.

Смесената комисия, създадена в съответствие с член 15, решава какъв вид трябва да има формата на искането за разрешение, реда за съгласуване, както и изискващите се придружаващи документи.

## EN

If there is no reciprocity, an authorization for a bilateral regular service can be refused. The Joint Committee set up under Article 15 hereof decides on the form on the authorization application, the procedure agreement and the supporting documents required.

# Data used by eTranslation



Such data are already available

BUT

they are not enough…

# What data are useful for eTranslation as per type |1

- Any **electronically stored text** in an EU language plus NO and IS
- **Texts and their translations** (i.e. parallel bilingual or multilingual)

## Bulgarian text

1. Компетентните органи на двете Договарящи страни си разменят всяка година договорен определен брой разрешителни.
2. Разрешителните се издават на местните превозвачи от компетентния орган или от организация, определена от споменатия орган.
3. Разрешителните се издават на името на превозвача и не се използват или преотстъпват на други превозвачи.
4. Разрешителните могат да бъдат използвани за едно превозно средство за определения срок.

Разрешителните са валидни до 31 януари на следващата календарна година.

## Translation in English

1. The competent authorities of the two Contracting Parties exchange and agree number of blank permit forms every year.
2. Permits are issued to resident transport operators by the competent authority or by a body designated by the said authority.
3. Permits are personal and are not transferable to third parties.
4. Permits can only be used for one vehicle at a time.

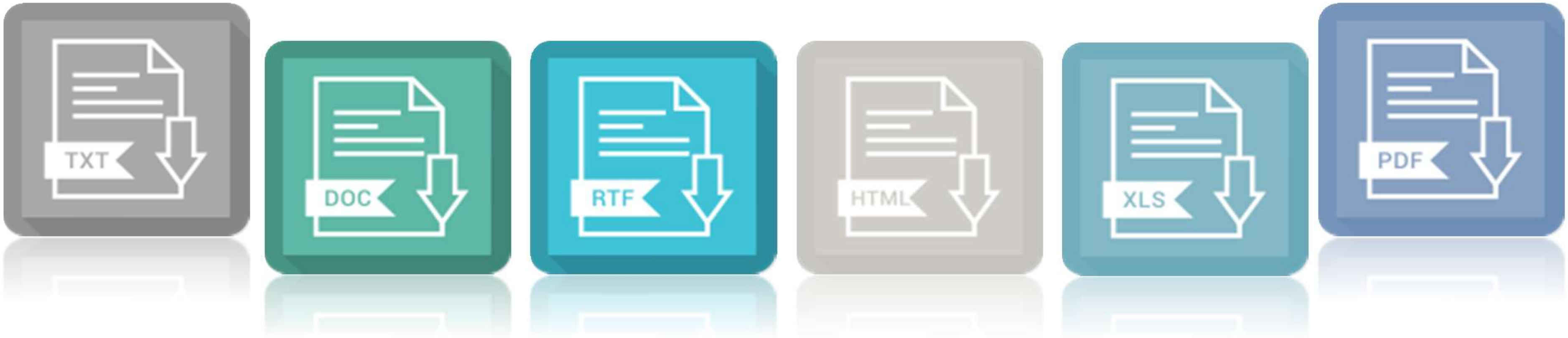
The permits shall be valid until 31 January of the successive calendar year.

# What data are useful for eTranslation as per type |2

- List of terms and their translations, i.e. a **terminology**

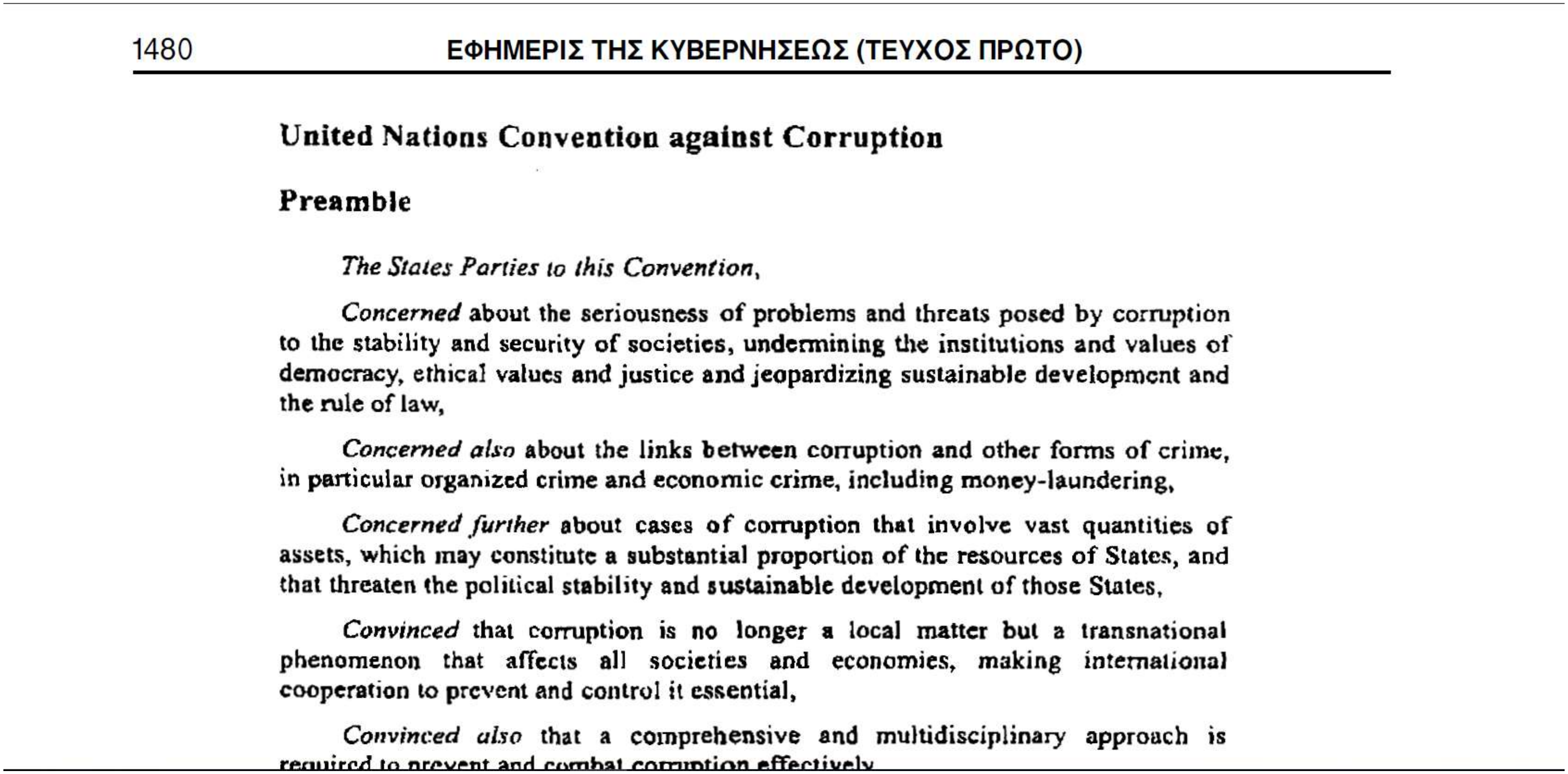
| Bulgarian | English |
|---|---|
| счетоводство | accountancy |
| счетоводен баланс | accounting balance |
| счетоводна конвенция | accounting convention |
| акредитация | accreditation |
| счетоводен принцип на начисляване | accrual accounting concept |
| основа за начисляване | accrual basis |
| начислени разходи | accrued expenses |
| цена за придобиване | acquisition cost |
| искове за отмяна | actions of repeal |
| активен капитал | active capital |
| … | … |

# What data are useful for eTranslation as per format |1



- In principle, any text in machine readable format
- But, some formats are more “MT-ready” than others, i.e. they require less manual or automatic processing
- More processing introduces more errors in the final output, making it less useful for eTranslation

# File formats for parallel texts



1480 ΕΦΗΜΕΡΙΣ ΤΗΣ ΚΥΒΕΡΝΗΣΕΩΣ (ΤΕΥΧΟΣ ΠΡΩΤΟ)

## United Nations Convention against Corruption

### Preamble

*The States Parties to this Convention,*

*Concerned* about the seriousness of problems and threats posed by corruption to the stability and security of societies, undermining the institutions and values of democracy, ethical values and justice and jeopardizing sustainable development and the rule of law,

*Concerned also* about the links between corruption and other forms of crime, in particular organized crime and economic crime, including money-laundering,

*Concerned further* about cases of corruption that involve vast quantities of assets, which may constitute a substantial proportion of the resources of States, and that threaten the political stability and sustainable development of those States,

*Convinced* that corruption is no longer a local matter but a transnational phenomenon that affects all societies and economies, making international cooperation to prevent and control it essential,

*Convinced also* that a comprehensive and multidisciplinary approach is required to prevent and combat corruption effectively

# What data are useful for eTranslation as per format |2

- The following formats are particularly useful (in descending order):
  - For bilingual/multilingual parallel texts
    1. Translation memories (.tmx)
    2. XML translation files (.xliff)
    3. Plain text (.txt, .csv)
    4. Spreadsheets (e.g. xlsx)
  - For terminologies
    1. TermBase eXchange (.tbx)
    2. Plain text (.txt, .csv)
    3. Spreadsheets (e.g. xlsx)
  - For monolingual texts
    1. Plain text (.txt, .csv)

# File formats of parallel texts and their manipulation

# Preparing your data |1

Don’ts

This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English.

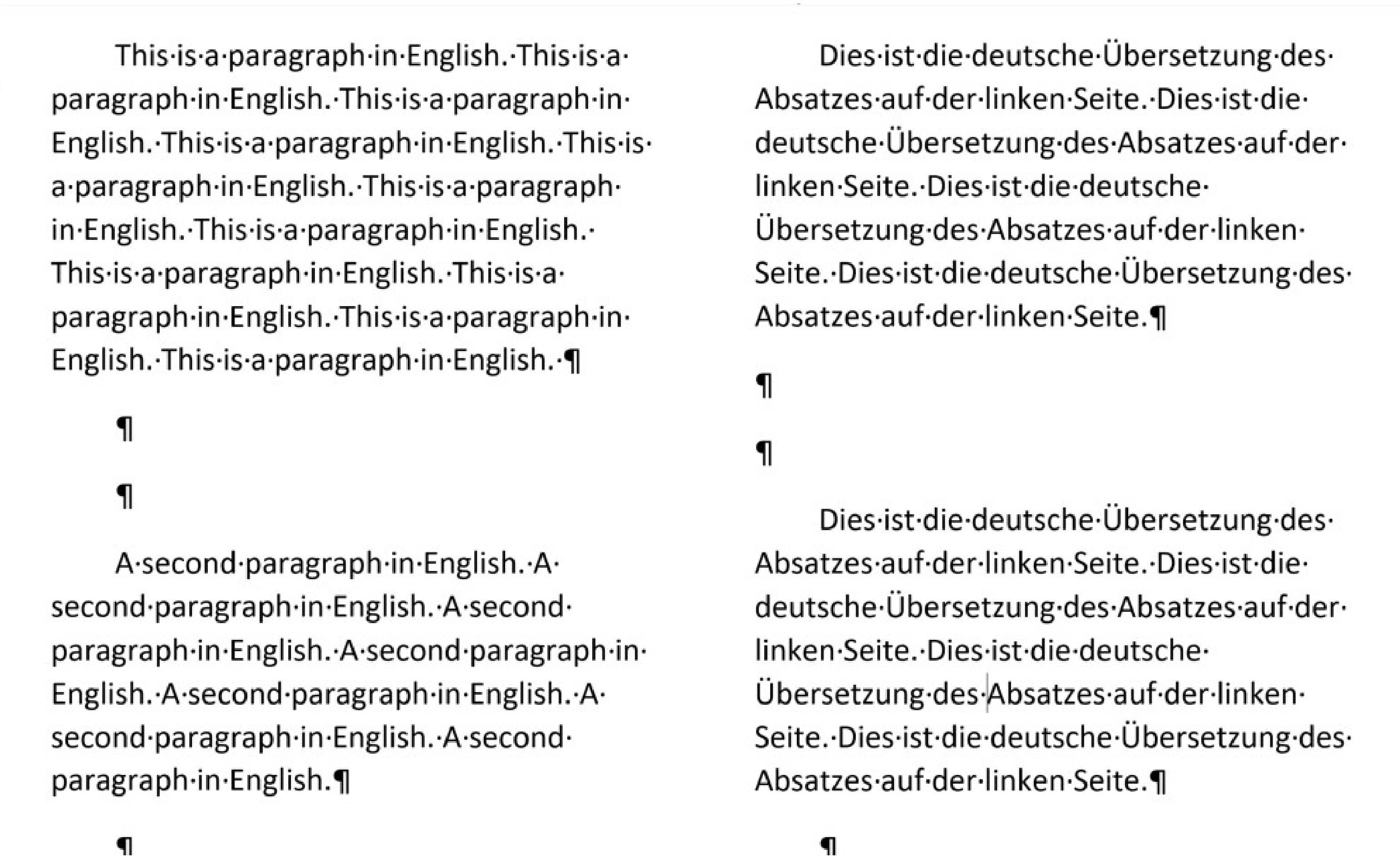
Dies ist die deutsche Übersetzung des vorherigen Absatzes. Dies ist die deutsche Übersetzung des vorherigen Absatzes. Dies ist die deutsche Übersetzung des vorherigen Absatzes. Dies ist die deutsche Übersetzung des vorherigen Absatzes.

A second paragraph in English. A second paragraph in English. A second paragraph in English. A second paragraph in English. A second paragraph in English. A second paragraph in English. A second paragraph in English.

Dies ist die deutsche Übersetzung des vorherigen Absatzes. Dies ist die deutsche Übersetzung des vorherigen Absatzes. Dies ist die deutsche Übersetzung des vorherigen Absatzes. Dies ist die deutsche Übersetzung des vorherigen Absatzes.

# Preparing your data |2

**Don’ts**



This·is·a·paragraph·in·English.·This·is·a·paragraph·in·English.·This·is·a·paragraph·in·English.·This·is·a·paragraph·in·English.·This·is·a·paragraph·in·English.·This·is·a·paragraph·in·English.·This·is·a·paragraph·in·English.·This·is·a·paragraph·in·English.·This·is·a·paragraph·in·English.·This·is·a·paragraph·in·English.·¶

¶

¶

A·second·paragraph·in·English.·A·second·paragraph·in·English.·A·second·paragraph·in·English.·A·second·paragraph·in·English.·A·second·paragraph·in·English.·A·second·paragraph·in·English.·A·second·paragraph·in·English.¶

¶

Dies·ist·die·deutsche·Übersetzung·des·Absatzes·auf·der·linken·Seite.·Dies·ist·die·deutsche·Übersetzung·des·Absatzes·auf·der·linken·Seite.·Dies·ist·die·deutsche·Übersetzung·des·Absatzes·auf·der·linken·Seite.·Dies·ist·die·deutsche·Übersetzung·des·Absatzes·auf·der·linken·Seite.¶

¶

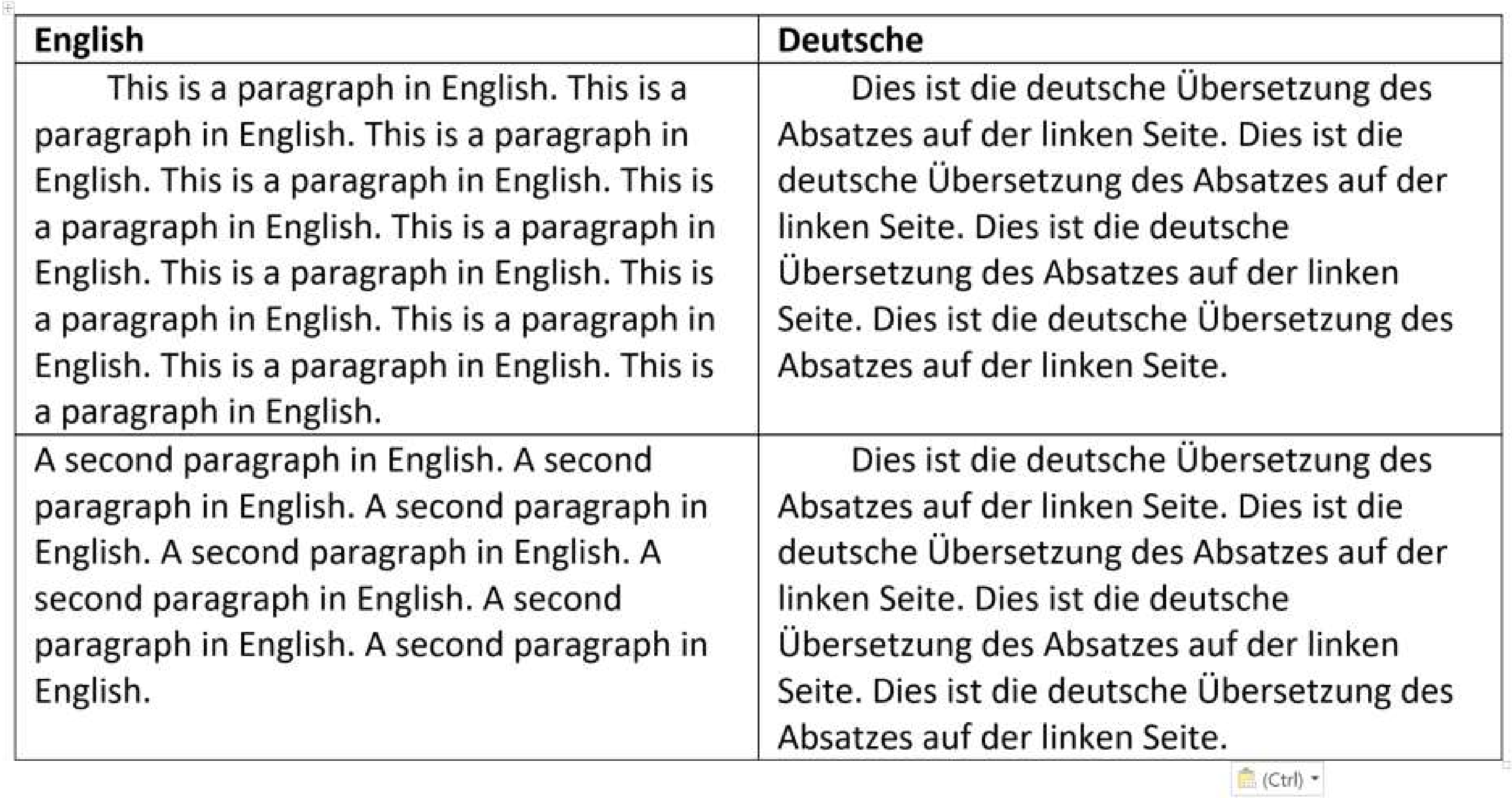
¶

Dies·ist·die·deutsche·Übersetzung·des·Absatzes·auf·der·linken·Seite.·Dies·ist·die·deutsche·Übersetzung·des·Absatzes·auf·der·linken·Seite.·Dies·ist·die·deutsche·Übersetzung·des·Absatzes·auf·der·linken·Seite.·Dies·ist·die·deutsche·Übersetzung·des·Absatzes·auf·der·linken·Seite.¶

¶

# Preparing your data |3

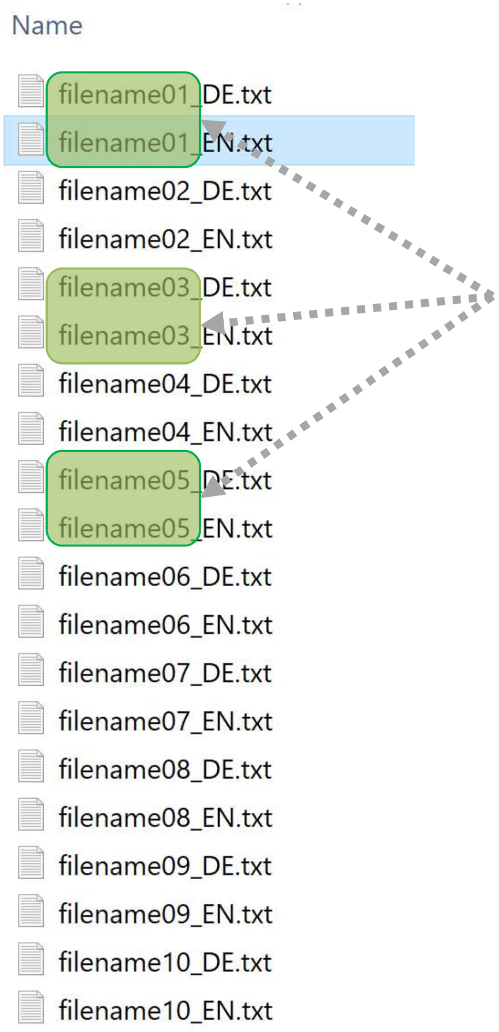
Don’ts



| English | Deutsche |
|---|---|
| This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. | Dies ist die deutsche Übersetzung des Absatzes auf der linken Seite. Dies ist die deutsche Übersetzung des Absatzes auf der linken Seite. Dies ist die deutsche Übersetzung des Absatzes auf der linken Seite. Dies ist die deutsche Übersetzung des Absatzes auf der linken Seite. |
| A second paragraph in English. A second paragraph in English. A second paragraph in English. A second paragraph in English. A second paragraph in English. A second paragraph in English. A second paragraph in English. | Dies ist die deutsche Übersetzung des Absatzes auf der linken Seite. Dies ist die deutsche Übersetzung des Absatzes auf der linken Seite. Dies ist die deutsche Übersetzung des Absatzes auf der linken Seite. Dies ist die deutsche Übersetzung des Absatzes auf der linken Seite. |

# Preparing your data |4

Do’s



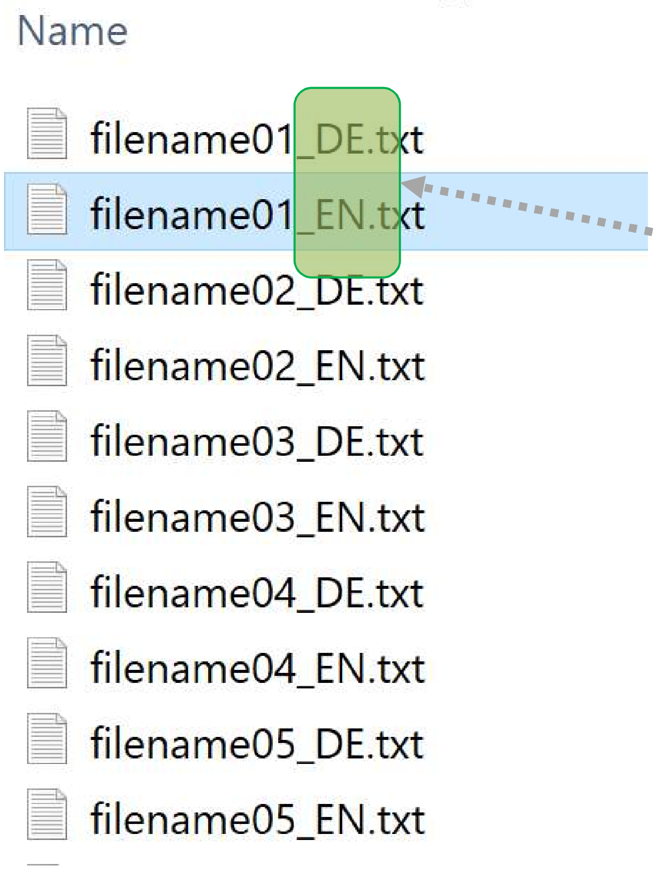
Name

- filename01_DE.txt
- filename01_EN.txt
- filename02_DE.txt
- filename02_EN.txt
- filename03_DE.txt
- filename03_EN.txt
- filename04_DE.txt
- filename04_EN.txt
- filename05_DE.txt
- filename05_EN.txt
- filename06_DE.txt
- filename06_EN.txt
- filename07_DE.txt
- filename07_EN.txt
- filename08_DE.txt
- filename08_EN.txt
- filename09_DE.txt
- filename09_EN.txt
- filename10_DE.txt
- filename10_EN.txt

Use **identical filenames** for each document pair (source – translation)

# Preparing your data |5

Do’s



Name

- filename01_DE.txt
- filename01_EN.txt
- filename02_DE.txt
- filename02_EN.txt
- filename03_DE.txt
- filename03_EN.txt
- filename04_DE.txt
- filename04_EN.txt
- filename05_DE.txt
- filename05_EN.txt

Include **language identifiers** in the filename

# Criteria for grouping your data

- Remember: a dataset is a collection of data **grouped according to certain criteria**
- For the purpose of enhancing and adapting CEF eTranslation, two criteria are critical:
  - **Language(s)**: each collection is defined by the language or language pairs of its data, e.g.
    - *Collection of texts in English – German*
    - *Documents in English – Norwegian - Finnish*
  - **Domain**: each collection ideally belongs to a single domain, e.g.
    - *Collection of texts in English – German in the culture domain*
    - *Social security documents in English – Norwegian - Finnish*

# Preferred domains

- Administrative/regulatory domain and
- Topics relevant to the CEF DSIs

| CEF DSI | Domain |
|---|---|
| Online Dispute Resolution | Consumers' rights, complaints |
| Electronic Exchange of Social Security Information | Social security, insurance |
| eProcurement | Public procurement, contractual agreements |
| European e-Justice Portal | Justice, Law |
| eHealth | Health, Medicine |
| Business Registers Interconnection System | Business, market |
| Safer Internet | |
| Cybersecurity | |
| Public Open Data | |
| Europeana | Culture |

# How to contribute your data to CEF eTranslation
# A step-by-step guide

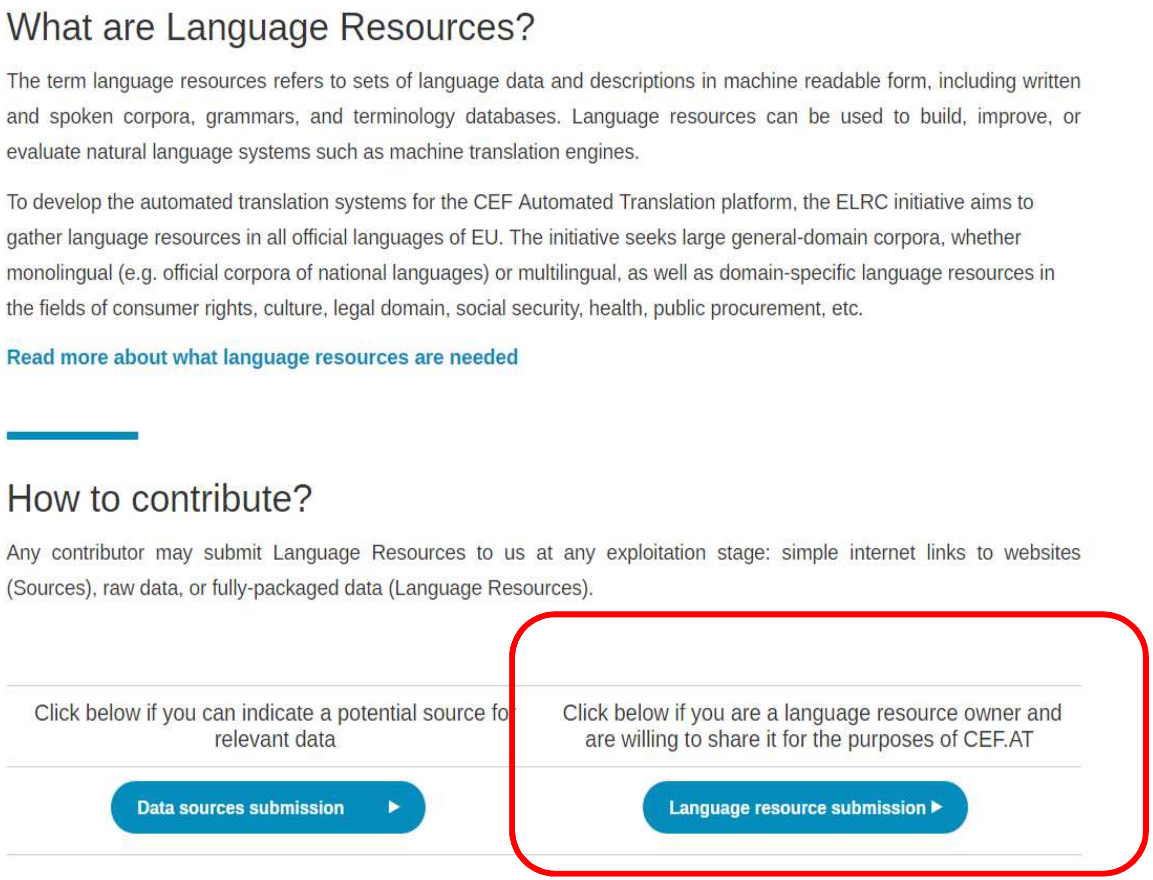
# To ELRC-SHARE

- At the ELRC portal click on the "Language resource submission" button

Or

- Type in the url address:

**elrc-share.eu**



## What are Language Resources?

The term language resources refers to sets of language data and descriptions in machine readable form, including written and spoken corpora, grammars, and terminology databases. Language resources can be used to build, improve, or evaluate natural language systems such as machine translation engines.

To develop the automated translation systems for the CEF Automated Translation platform, the ELRC initiative aims to gather language resources in all official languages of EU. The initiative seeks large general-domain corpora, whether monolingual (e.g. official corpora of national languages) or multilingual, as well as domain-specific language resources in the fields of consumer rights, culture, legal domain, social security, health, public procurement, etc.

Read more about what language resources are needed

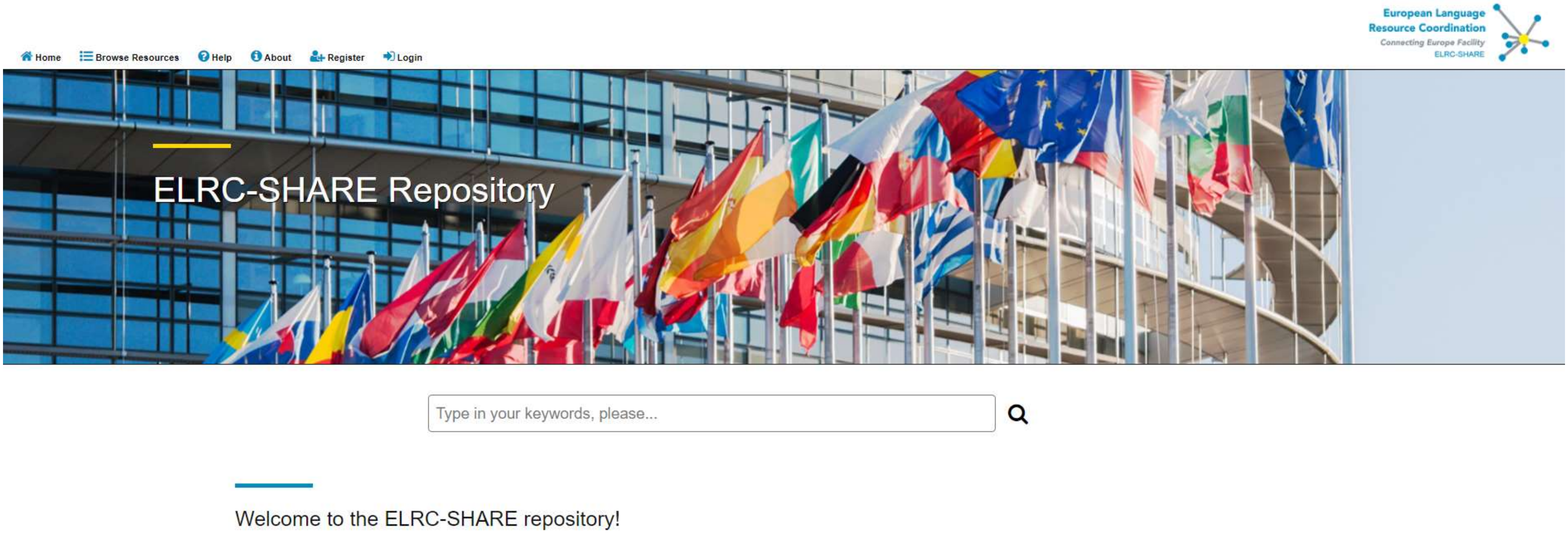
## How to contribute?

Any contributor may submit Language Resources to us at any exploitation stage: simple internet links to websites (Sources), raw data, or fully-packaged data (Language Resources).

Click below if you can indicate a potential source fo relevant data

Click below if you are a language resource owner and are willing to share it for the purposes of CEF.AT

# ELRC-SHARE repository

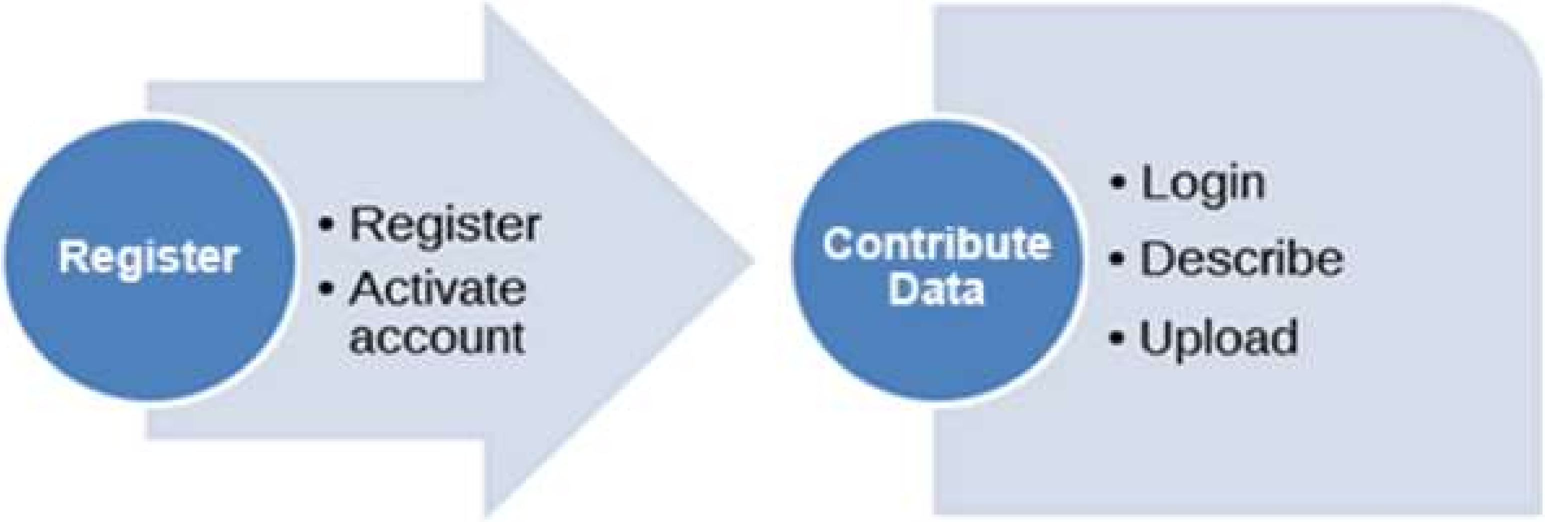


ELRC-SHARE Repository

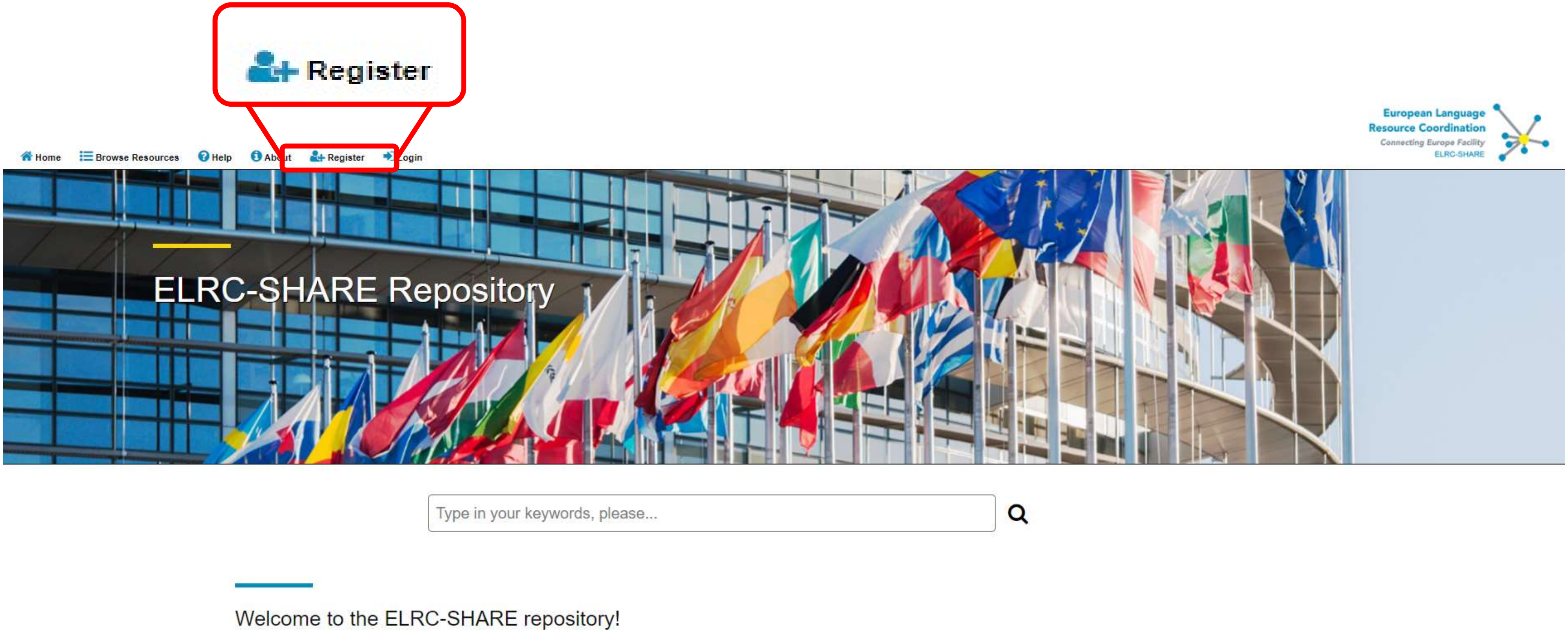
Type in your keywords, please...

Welcome to the ELRC-SHARE repository!

# How to Contribute Data



# How to Register (1/2)


Register
Home
Browse Resources
Help
About
Register
Login
European Language Resource Coordination
Connecting Europe Facility
ELRC-SHARE

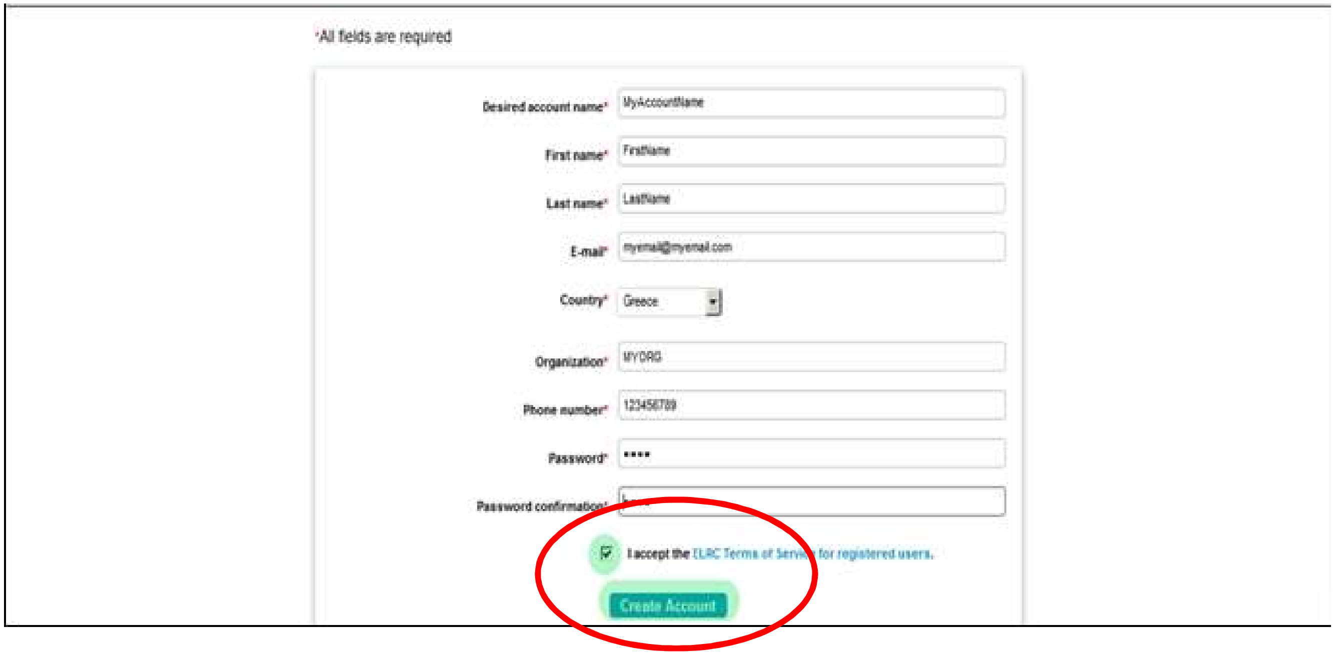
ELRC-SHARE Repository

Type in your keywords, please...

Welcome to the ELRC-SHARE repository!

# How to Register (2/2)

- Fill in the required info
- Read the *Terms of Service* and click *Accept,* if you agree
- Click the *Create Account* button
- Activate your account according to the guidelines emailed to you

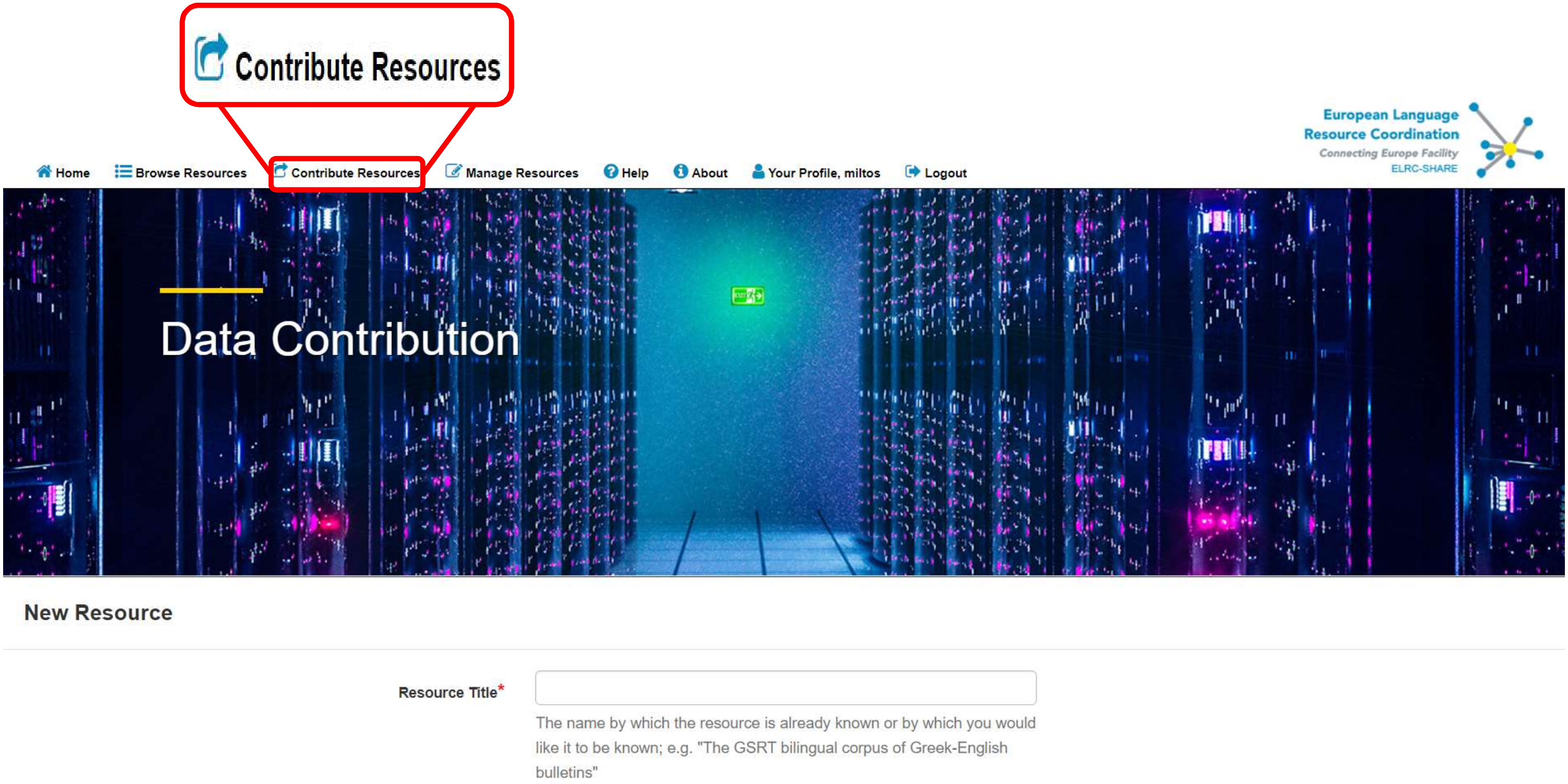


*All fields are required

| | |
|---|---|
| Desired account name* | MyAccountName |
| First name* | FirstName |
| Last name* | LastName |
| E-mail* | myemail@myemail.com |
| Country* | Greece |
| Organization* | MYORG |
| Phone number* | 123456789 |
| Password* | •••• |
| Password confirmation* | •••• |

☑ I accept the ELRC Terms of Service for registered users.

# How to Contribute Data (1/6)



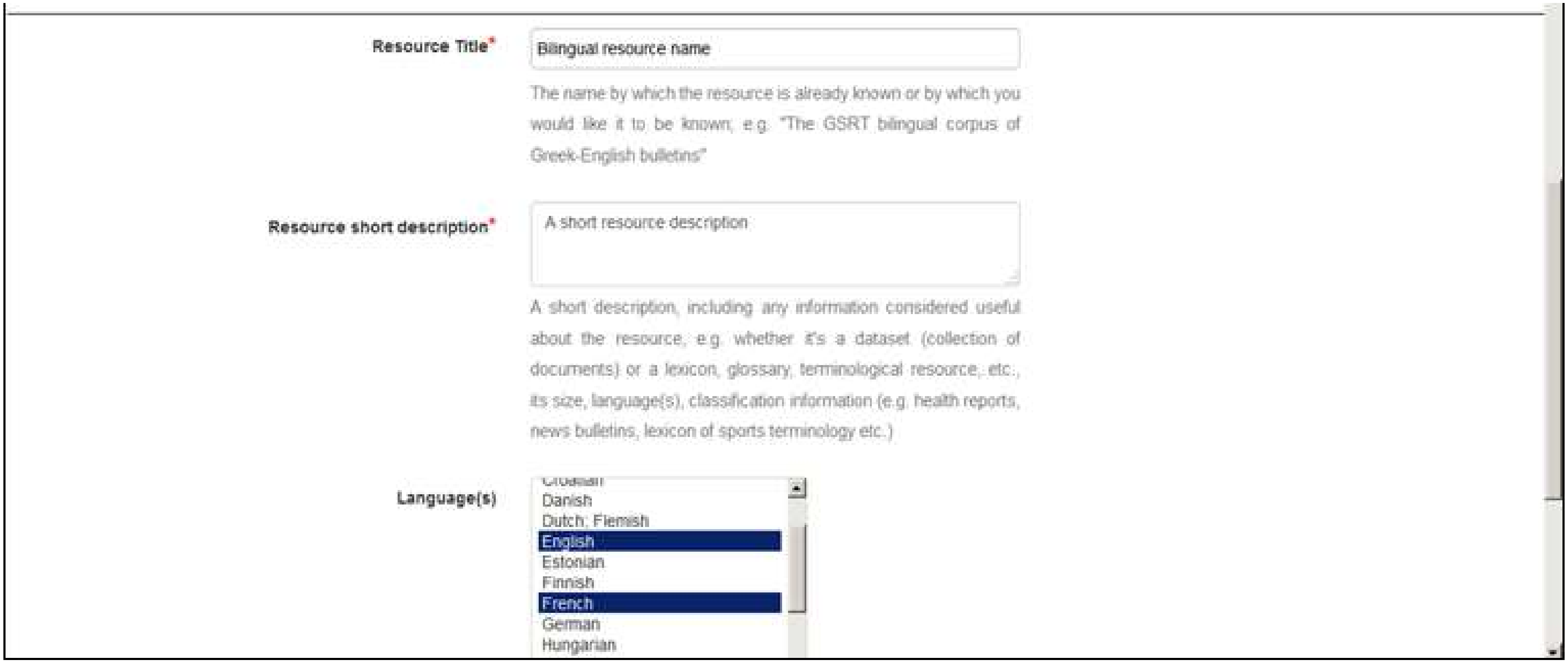
New Resource

Resource Title*

The name by which the resource is already known or by which you would like it to be known; e.g. "The GSRT bilingual corpus of Greek-English bulletins"

# How to Contribute Data (2/6)

- Fill in the details of the dataset



Resource Title*

Bilingual resource name

The name by which the resource is already known or by which you would like it to be known, e.g. "The GSRT bilingual corpus of Greek-English bulletins"

Resource short description*

A short resource description

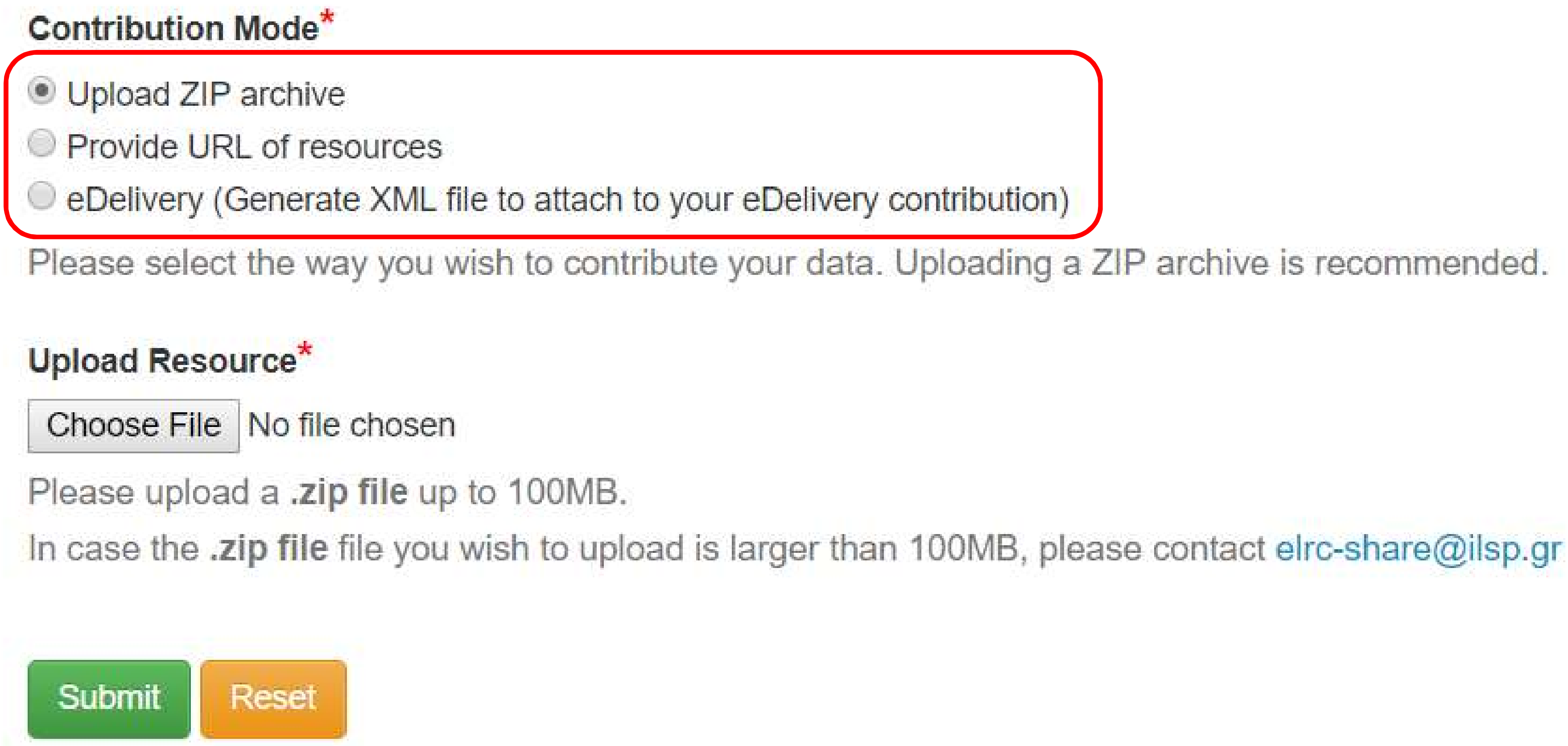
A short description, including any information considered useful about the resource, e.g. whether it's a dataset (collection of documents) or a lexicon, glossary, terminological resource, etc., its size, language(s), classification information (e.g. health reports, news bulletins, lexicon of sports terminology etc.)

Language(s)

Croatian
Danish
Dutch; Flemish
English
Estonian
Finnish
French
German
Hungarian

# How to Contribute Data (3/6)

- Three modes for contributing your data



**Contribution Mode***

● Upload ZIP archive
○ Provide URL of resources
○ eDelivery (Generate XML file to attach to your eDelivery contribution)

Please select the way you wish to contribute your data. Uploading a ZIP archive is recommended.

**Upload Resource***

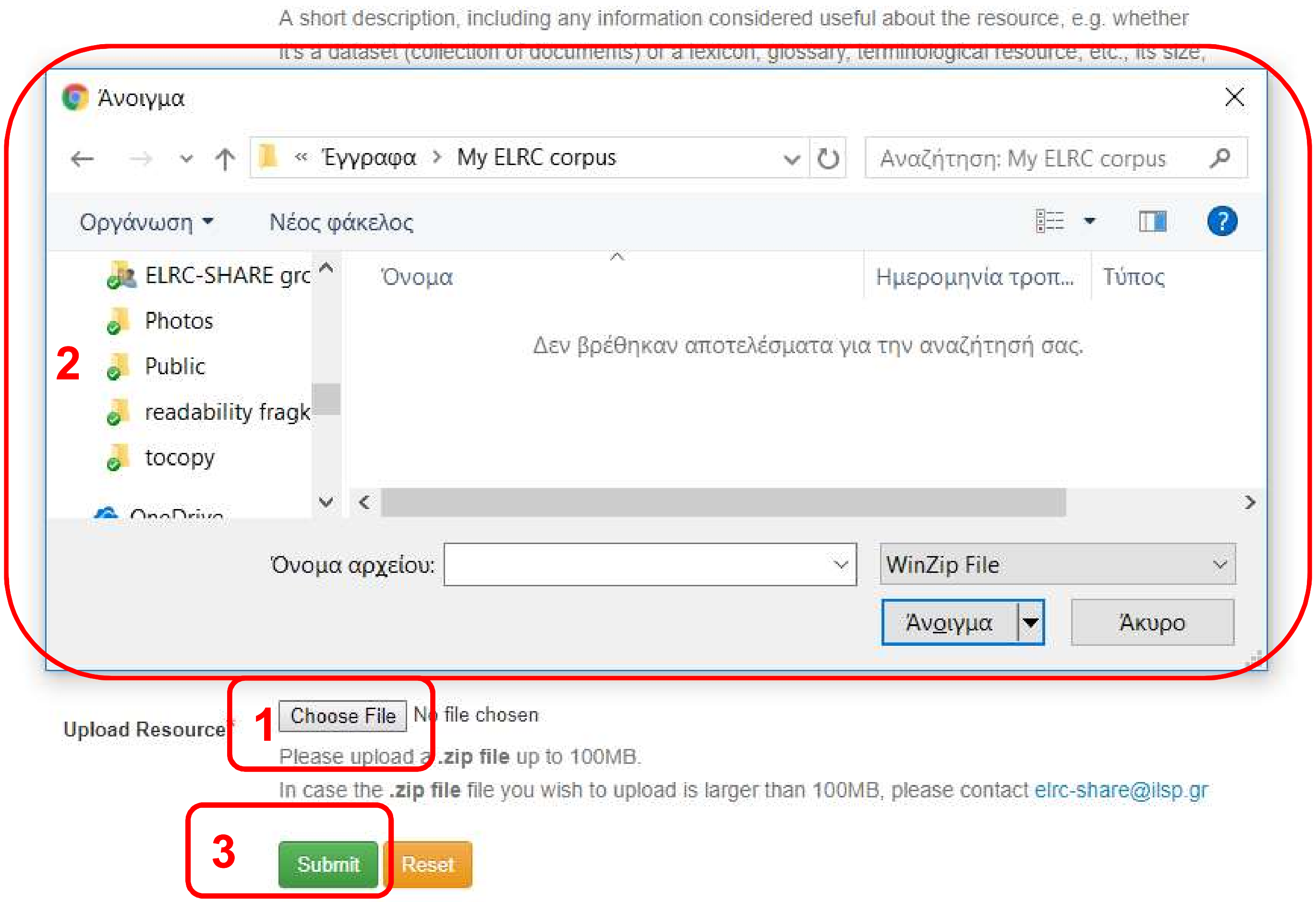
Choose File No file chosen

Please upload a **.zip file** up to 100MB.

In case the **.zip file** file you wish to upload is larger than 100MB, please contact elrc-share@ilsp.gr

# How to Contribute Data (4/6)

1. Click on Choose file
2. Locate your resource in your hard disk
3. Click on Submit



A short description, including any information considered useful about the resource, e.g. whether it's a dataset (collection of documents) or a lexicon, glossary, terminological resource, etc., its size,

Upload Resource* Choose File No file chosen

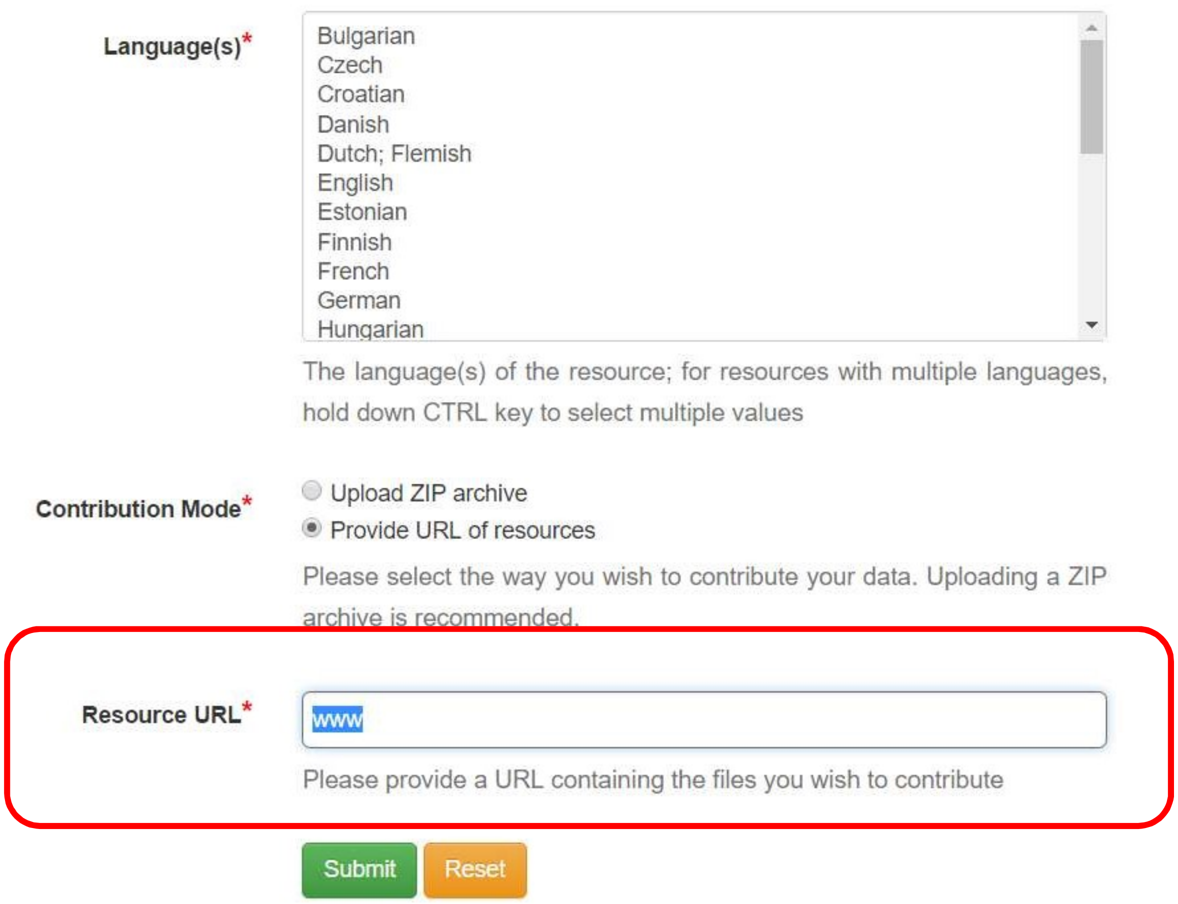
Please upload a .zip file up to 100MB.

In case the .zip file file you wish to upload is larger than 100MB, please contact elrc-share@ilsp.gr

Δεν βρέθηκαν αποτελέσματα για την αναζήτησή σας.

# How to Contribute Data (5/6)

- Alternatively indicate a url (directory listing)



**Language(s)***

Bulgarian
Czech
Croatian
Danish
Dutch; Flemish
English
Estonian
Finnish
French
German
Hungarian

The language(s) of the resource; for resources with multiple languages, hold down CTRL key to select multiple values

**Contribution Mode***

○ Upload ZIP archive
● Provide URL of resources

Please select the way you wish to contribute your data. Uploading a ZIP archive is recommended.

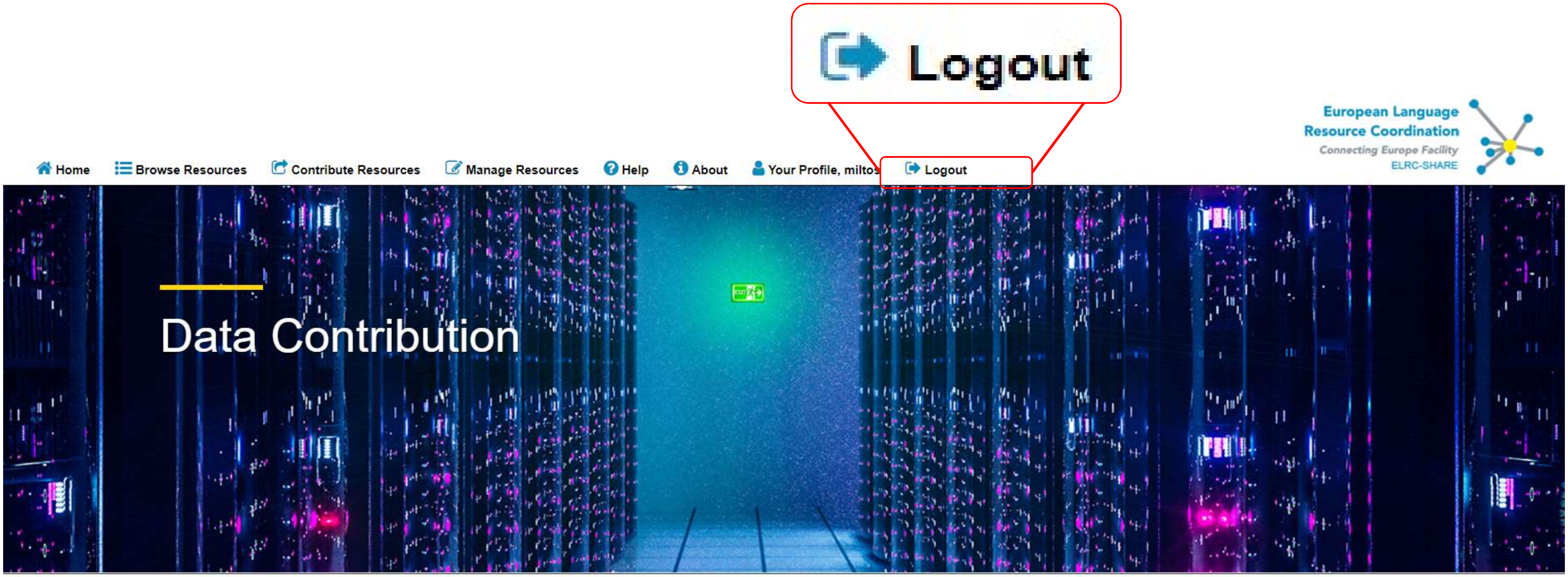
**Resource URL***

www

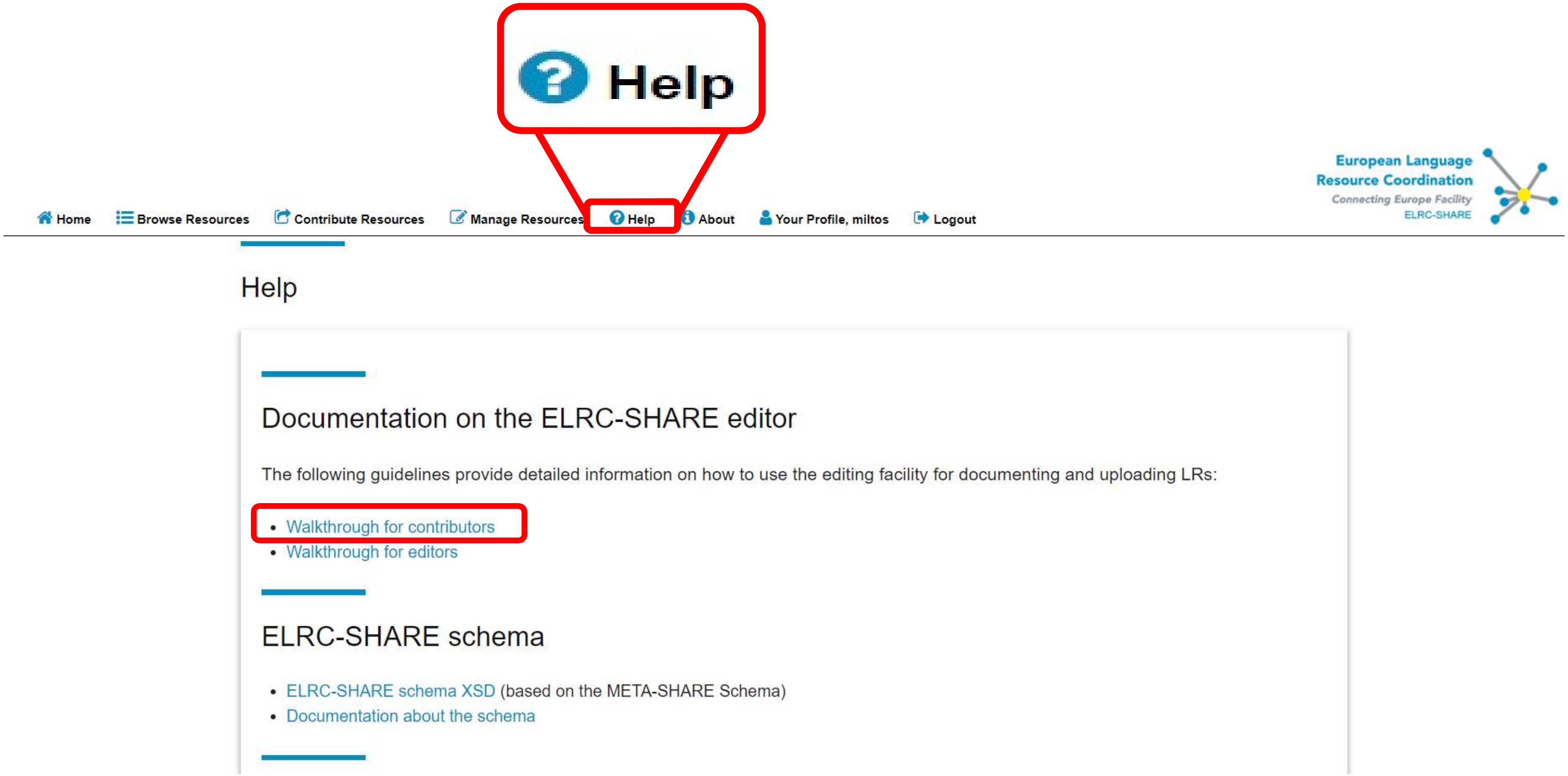
Please provide a URL containing the files you wish to contribute

# How to Contribute Data (6/6)

- Repeat the process if you want to contribute another resource, or log out



# Guidelines for contributors



Help

## Documentation on the ELRC-SHARE editor

The following guidelines provide detailed information on how to use the editing facility for documenting and uploading LRs:

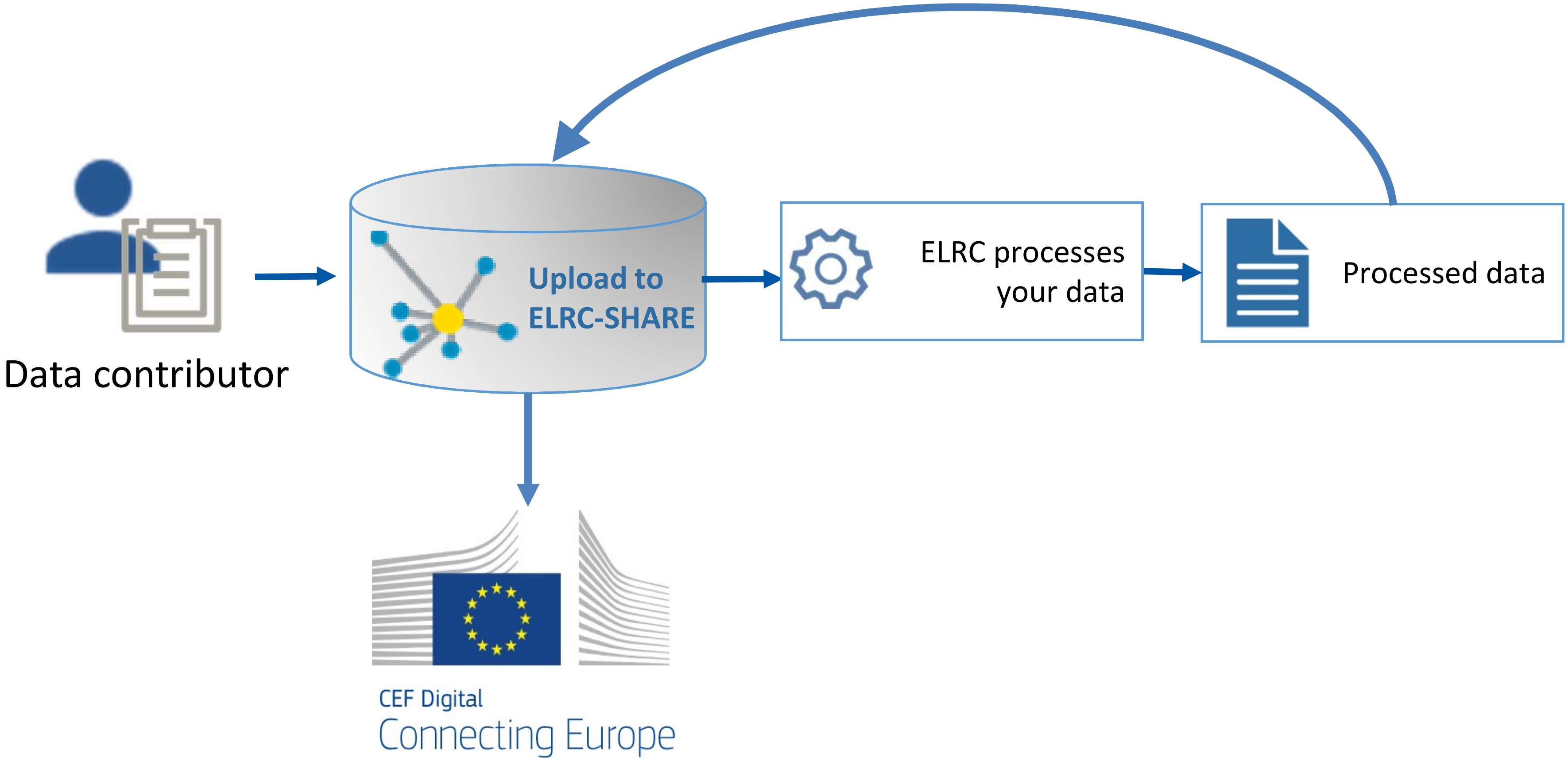
- Walkthrough for contributors
- Walkthrough for editors

## ELRC-SHARE schema

- ELRC-SHARE schema XSD (based on the META-SHARE Schema)
- Documentation about the schema

# What happens next?

# What happens to your data?



# Language processing services

## Data extraction

If your data is trapped in archives and databases, we can help extract it

## Anonymisation

Does your data contain private info? We can help to anonymise

## Cleaning

If your data is messy (i.e., lots of noise), we will clean it up

## Re-formatting

Need to re-format DOCX to XML, or PDF to WORD? Let us do it for you!

# Language processing services

## Data conversion

If your data isn't converted to the proper formats, we can help convert it

## Tag removal

Does your data contain unneeded tags? We can assist in removing them!

## Alignment

Translations aren't aligned? We'll do it for you with our tools!

## Metadata

Metadata are crucial! We can organise and validate metadata for your team

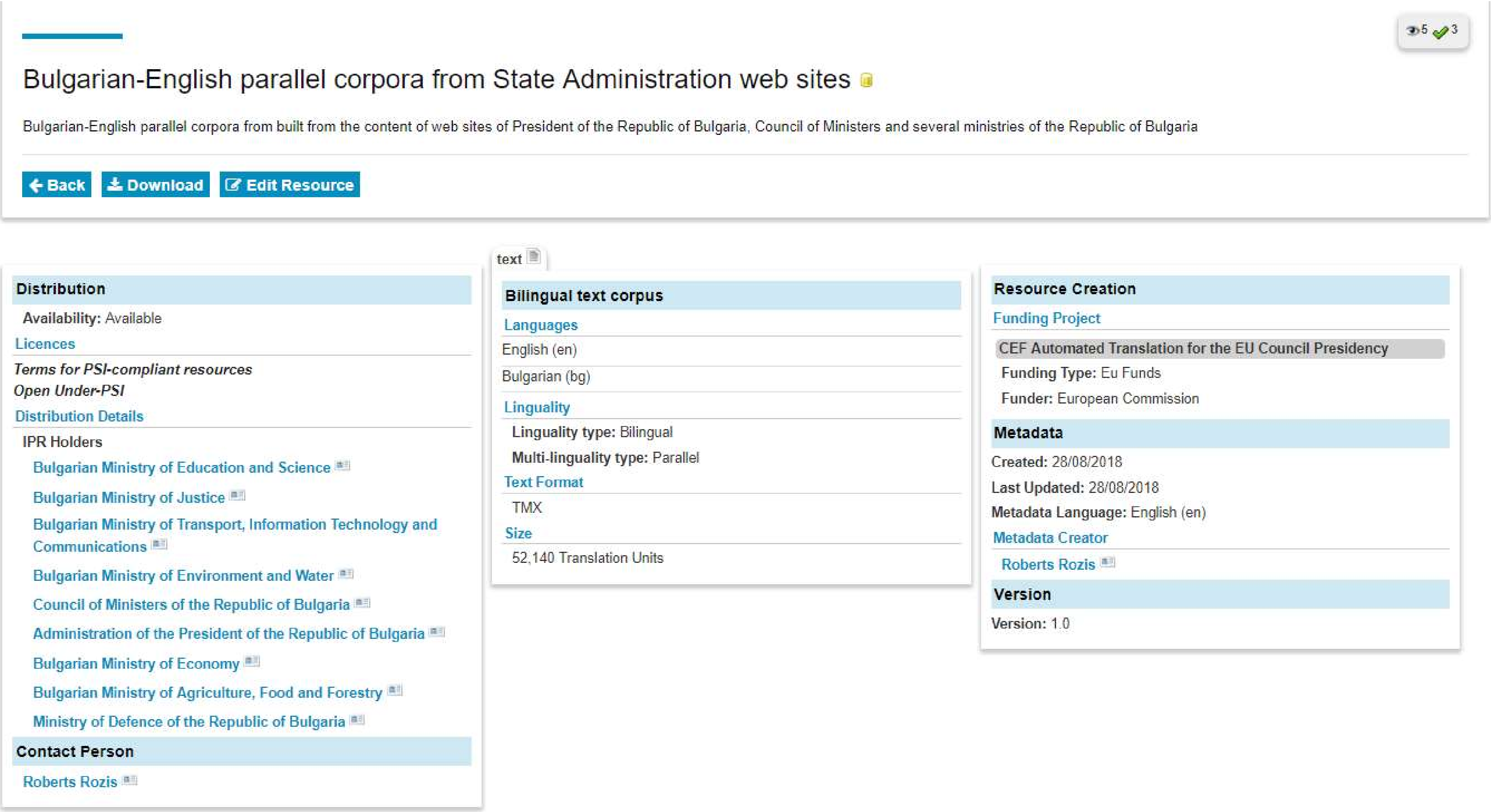
# What has happened to your data?

File01_bg.txt
File01_en.doc
File02_bg.pdf
File02_en.txt
File03_bg.doc
File03_en.doc
…

After processing

```
<tu tuid="20">
    <tuv xml:lang="bg">
       <seg> ''Страна на регистрация'' означава територията на договарящата страна, в която са регистрирани превозвачът и превозното средство.</seg>
      </tuv>
     <tuv xml:lang="en">
       <seg> “Country of establishment” means the territory of a Contracting Party within which the transport operator is established and the vehicle registered.</seg>
      </tuv>
</tu>
<tu tuid="21">
      <tuv xml:lang="bg">
       <seg> ''Пътнически превоз по редовна линия'' означава превозна услуга, при която се превозват пътници по определени маршрути, разписание и установени цени.</seg>
      </tuv>
      <tuv xml:lang="en">
       <seg> “Regular passenger service” means a service which carries passengers over a specified route, according to a timetable and for which set fares are charged.</seg>
      </tuv>
</tu>
```

# The notion of data in the context of eTranslation



## Bulgarian-English parallel corpora from State Administration web sites

Bulgarian-English parallel corpora from built from the content of web sites of President of the Republic of Bulgaria, Council of Ministers and several ministries of the Republic of Bulgaria

**Distribution**

**Availability:** Available

Licences

*Terms for PSI-compliant resources*
*Open Under-PSI*

Distribution Details

IPR Holders

- Bulgarian Ministry of Education and Science
- Bulgarian Ministry of Justice
- Bulgarian Ministry of Transport, Information Technology and Communications
- Bulgarian Ministry of Environment and Water
- Council of Ministers of the Republic of Bulgaria
- Administration of the President of the Republic of Bulgaria
- Bulgarian Ministry of Economy
- Bulgarian Ministry of Agriculture, Food and Forestry
- Ministry of Defence of the Republic of Bulgaria

**Contact Person**

Roberts Rozis

**Bilingual text corpus**

Languages

English (en)

Bulgarian (bg)

Linguality

**Linguality type:** Bilingual

**Multi-linguality type:** Parallel

Text Format

TMX

Size

52,140 Translation Units

**Resource Creation**

Funding Project

CEF Automated Translation for the EU Council Presidency

**Funding Type:** Eu Funds

**Funder:** European Commission

**Metadata**

**Created:** 28/08/2018

**Last Updated:** 28/08/2018

**Metadata Language:** English (en)

Metadata Creator

Roberts Rozis

**Version**

**Version:** 1.0

# On-site assistance

**All these services can also be offered on-site to all data contributors free of charge**

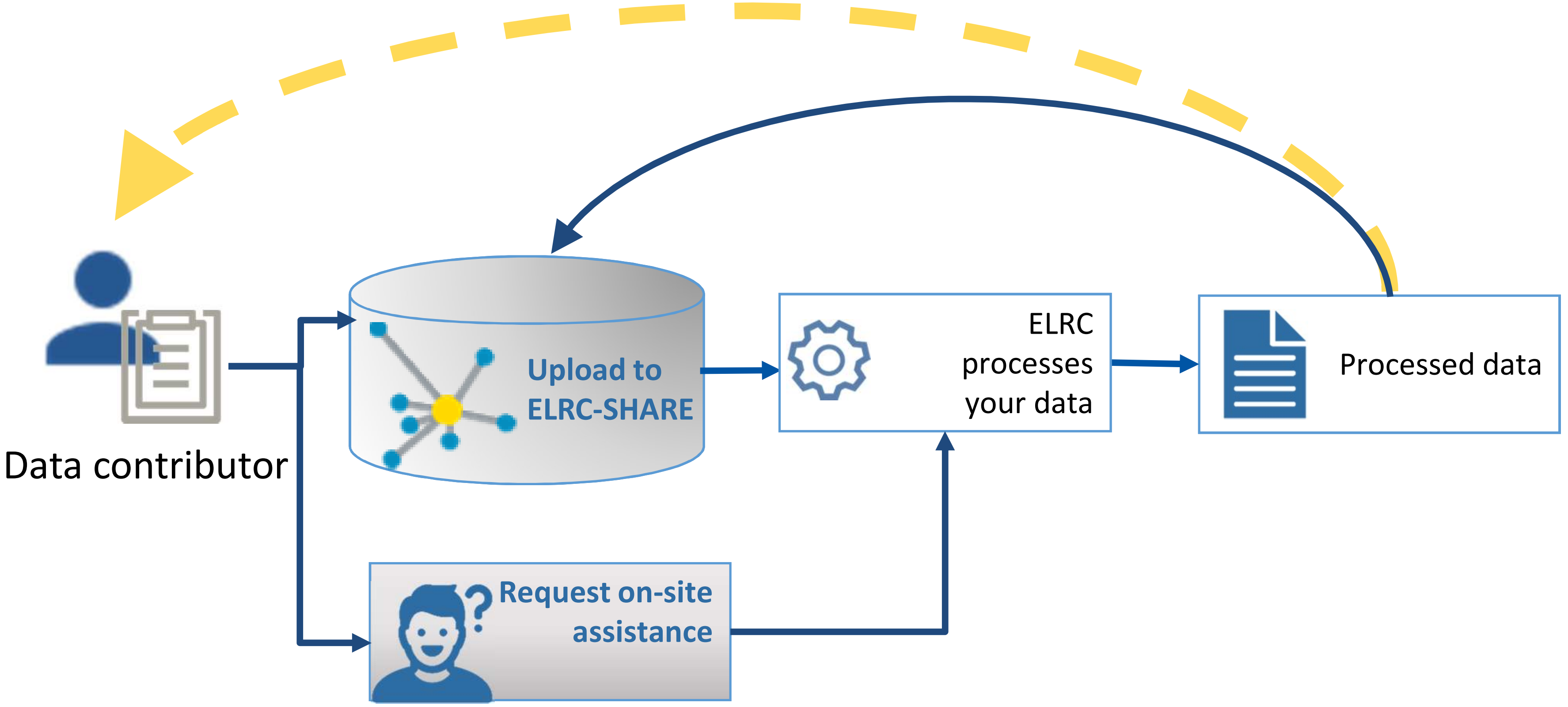
# On-site assistance

**Our team of experts will travel directly to assist you at your own offices**

# On-site assistance

**We will fix your data issues and return the processed data directly to you. We can also help to improve your data management processes. Just ask!**

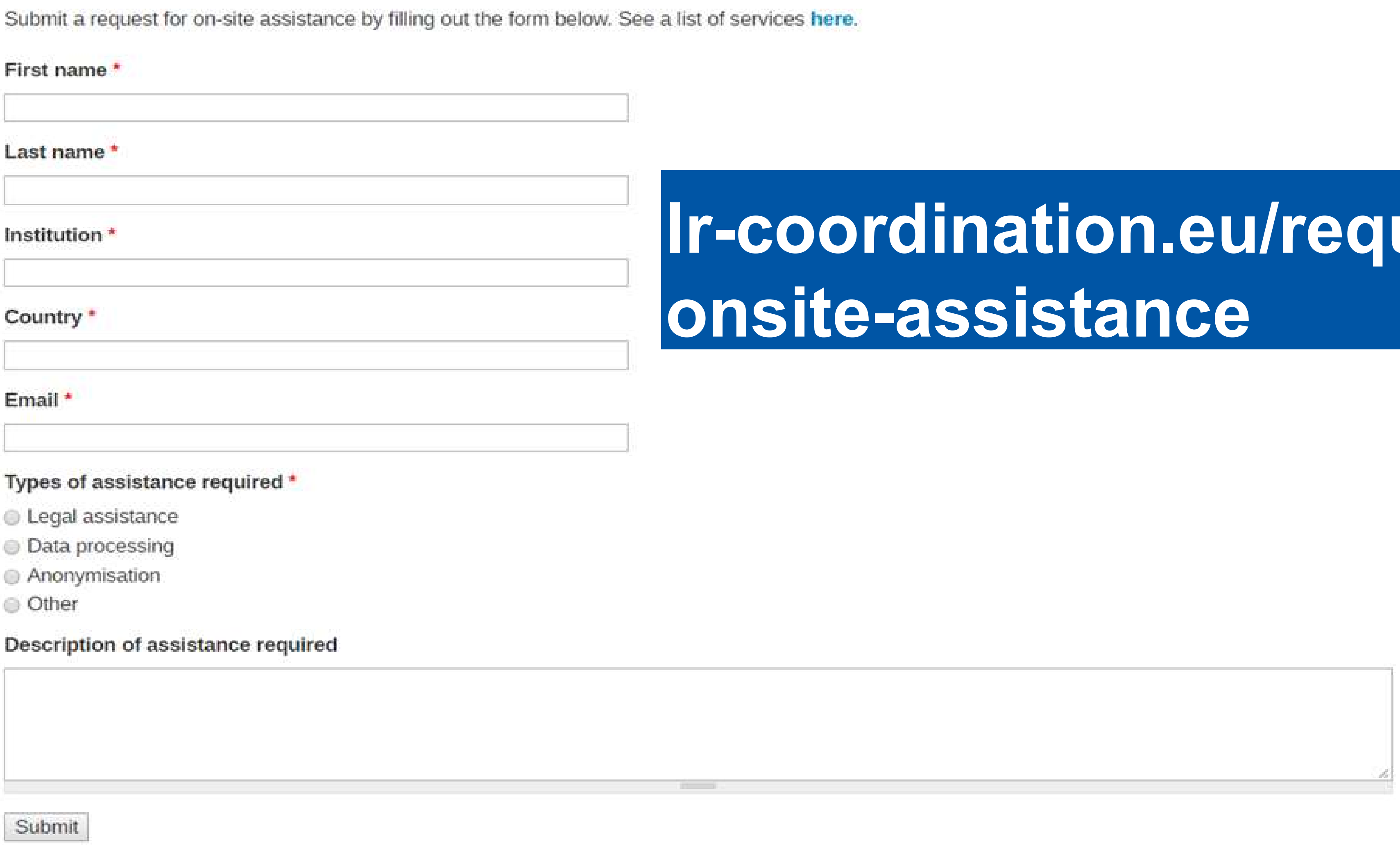
# What happens to your data?



# How to request services and help

European Language
Resource Coordination
*Connecting Europe Facility*

# ELRC onsite assistance



Submit a request for on-site assistance by filling out the form below. See a list of services here.

First name *

Last name *

Institution *

Country *

Email *

Types of assistance required *

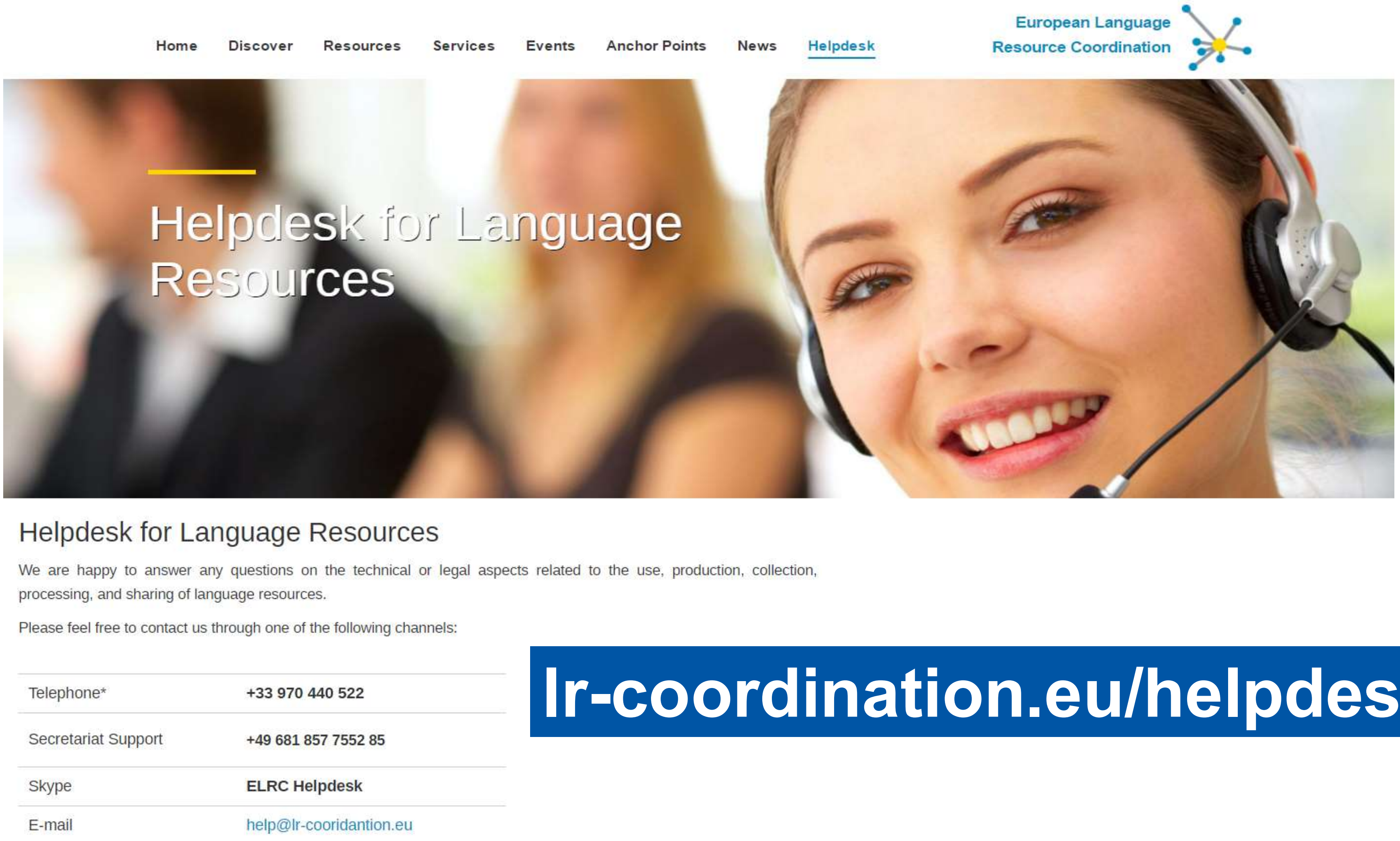
○ Legal assistance
○ Data processing
○ Anonymisation
○ Other

Description of assistance required

Submit

**lr-coordination.eu/request-onsite-assistance**

# ELRC Helpdesk



Helpdesk for Language Resources

## Helpdesk for Language Resources

We are happy to answer any questions on the technical or legal aspects related to the use, production, collection, processing, and sharing of language resources.

Please feel free to contact us through one of the following channels:

| | |
|---|---|
| Telephone* | **+33 970 440 522** |
| Secretariat Support | **+49 681 857 7552 85** |
| Skype | **ELRC Helpdesk** |
| E-mail | help@lr-cooridantion.eu |

**lr-coordination.eu/helpdesk**

# ELRC consortium – come talk to us!

# Благодаря ви!

# Icons used in this presentation

- By Michael Mellon, GB, , CC-BY 3.0 US
- By Joana Pereira, BR, CC-BY 3.0 US
- By Becca O'Shea, NZ, CC-BY 3.0 US
- By Creative Stall, Basic licence www.iconfinder.com
- By Creative Stall, PK, CC-BY 3.0 US
- By Arthur Shlain, IL, CC-BY 3.0 US
- By Shmidt Sergey, US, CC-BY 3.0 US
- By Gregor Cresnar, CC-BY 3.0 US
- By anbileru adaleru, CC-BY 3.0 US
- By Vectors Market, CC-BY 3.0 US