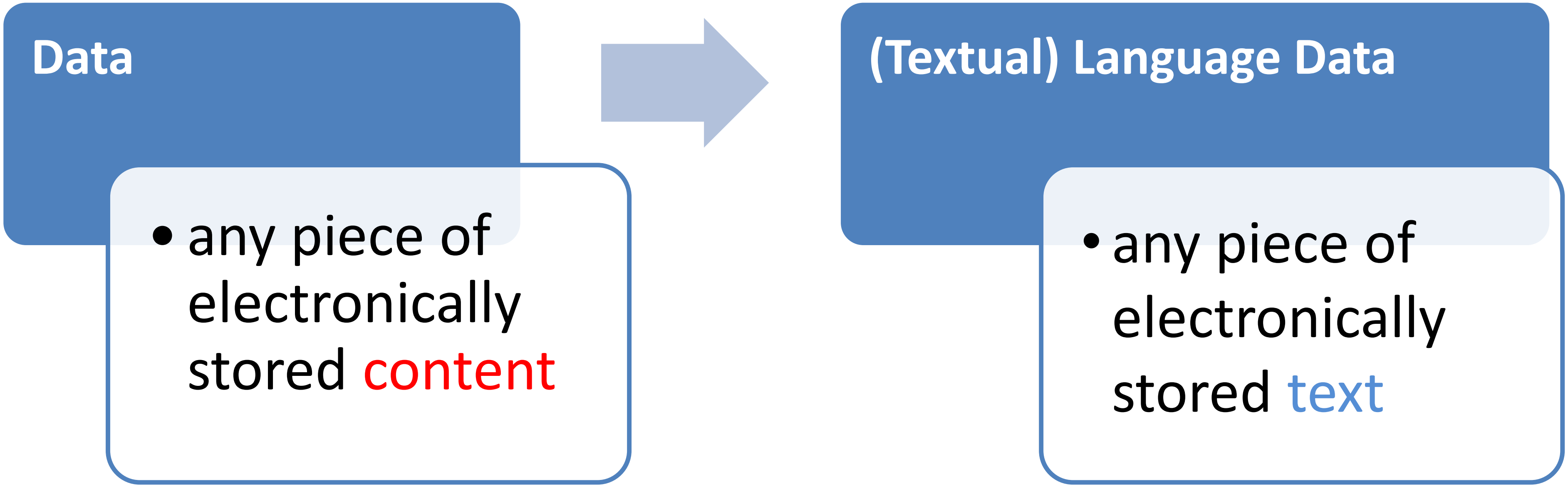
Preparing and sharing data via the ELRC-SHARE repository
ELRC on site Services and what happens next

Dutch Workshop

Khalid Choukri, Hélène Mazo
ELRA/ELDA

The notion of language data



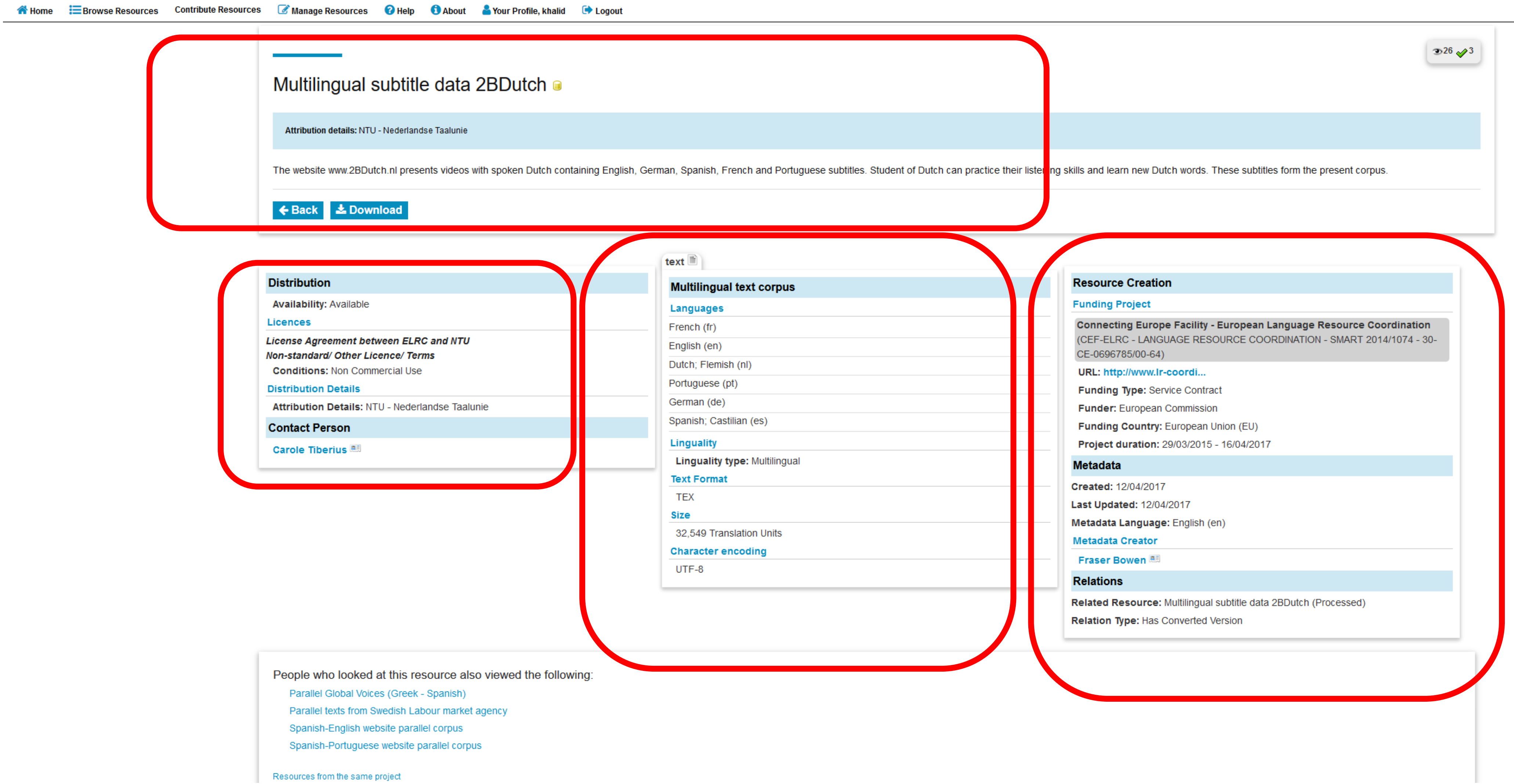
Data

- any piece of electronically stored content

(Textual) Language Data

- any piece of electronically stored text

The notion of data in the context of eTranslation



Multilingual subtitle data 2BDutch

Attribution details: NTU - Nederlandse Taalunie

The website www.2BDutch.nl presents videos with spoken Dutch containing English, German, Spanish, French and Portuguese subtitles. Student of Dutch can practice their listening skills and learn new Dutch words. These subtitles form the present corpus.

Distribution

Availability: Available

Licences

License Agreement between ELRC and NTU
Non-standard/ Other Licence/ Terms

Conditions: Non Commercial Use

Distribution Details

Attribution Details: NTU - Nederlandse Taalunie

Contact Person

Carole Tiberius

text

Multilingual text corpus

Languages

French (fr)
English (en)
Dutch; Flemish (nl)
Portuguese (pt)
German (de)
Spanish; Castilian (es)

Linguality

Linguality type: Multilingual

Text Format

TEX

Size

32,549 Translation Units

Character encoding

UTF-8

Resource Creation

Funding Project

Connecting Europe Facility - European Language Resource Coordination (CEF-ELRC - LANGUAGE RESOURCE COORDINATION - SMART 2014/1074 - 30-CE-0696785/00-64)

URL: http://www.lr-coordi...
Funding Type: Service Contract
Funder: European Commission
Funding Country: European Union (EU)
Project duration: 29/03/2015 - 16/04/2017

Metadata

Created: 12/04/2017
Last Updated: 12/04/2017
Metadata Language: English (en)

Metadata Creator

Fraser Bowen

Relations

Related Resource: Multilingual subtitle data 2BDutch (Processed)
Relation Type: Has Converted Version

People who looked at this resource also viewed the following:

- Parallel Global Voices (Greek - Spanish)
- Parallel texts from Swedish Labour market agency
- Spanish-English website parallel corpus
- Spanish-Portuguese website parallel corpus

Resources from the same project

The notion of data
in the context of eTranslation

```
File01_it.txt
File01_en.txt
File02_it.txt
File02_en.txt
File03_it.txt
…
```

Trans.
Da

```
tuv xml:lang="nl"
changedate="20151214T133604Z">
    <seg>Deze gecoördineerde wet
stelt een regeling voor
verplichte verzekering voor
geneeskundige verzorging en
uitkeringen in; ze organiseert
die in twee onderscheiden takken
die betrekking hebben, de ene op
de geneeskundige verstrekkingen,
de andere op de uitkeringen
wegens arbeidsongeschiktheid […]
en op de
moederschapsverzekering.</seg>
```

```
 </tuv>
   <tuv xml:lang="fr"
changedate="20151214T133604Z">
    <seg>La présente loi
coordonnée institue un régime
d'assurance obligatoire soins
de santé et indemnités; elle
l'organise en deux secteurs
distincts relatifs, l'un aux
prestations de santé, l'autre
aux indemnités d'incapacité de
travail, […] et à l'assurance
maternité.</seg>
   </tuv>
```

The notion of data in the context of eTranslation

```
File01_nl.txt
File01_en.txt
File02_nl.txt
File02_en.txt
File03_nl.txt
…
```

Trans.
Da

```
Amsterdam is de hoofdstad van ons land.

Nu is het een grote en drukke stad.

En dit is 'De Dam',

het beroemde plein in Amsterdam.
```

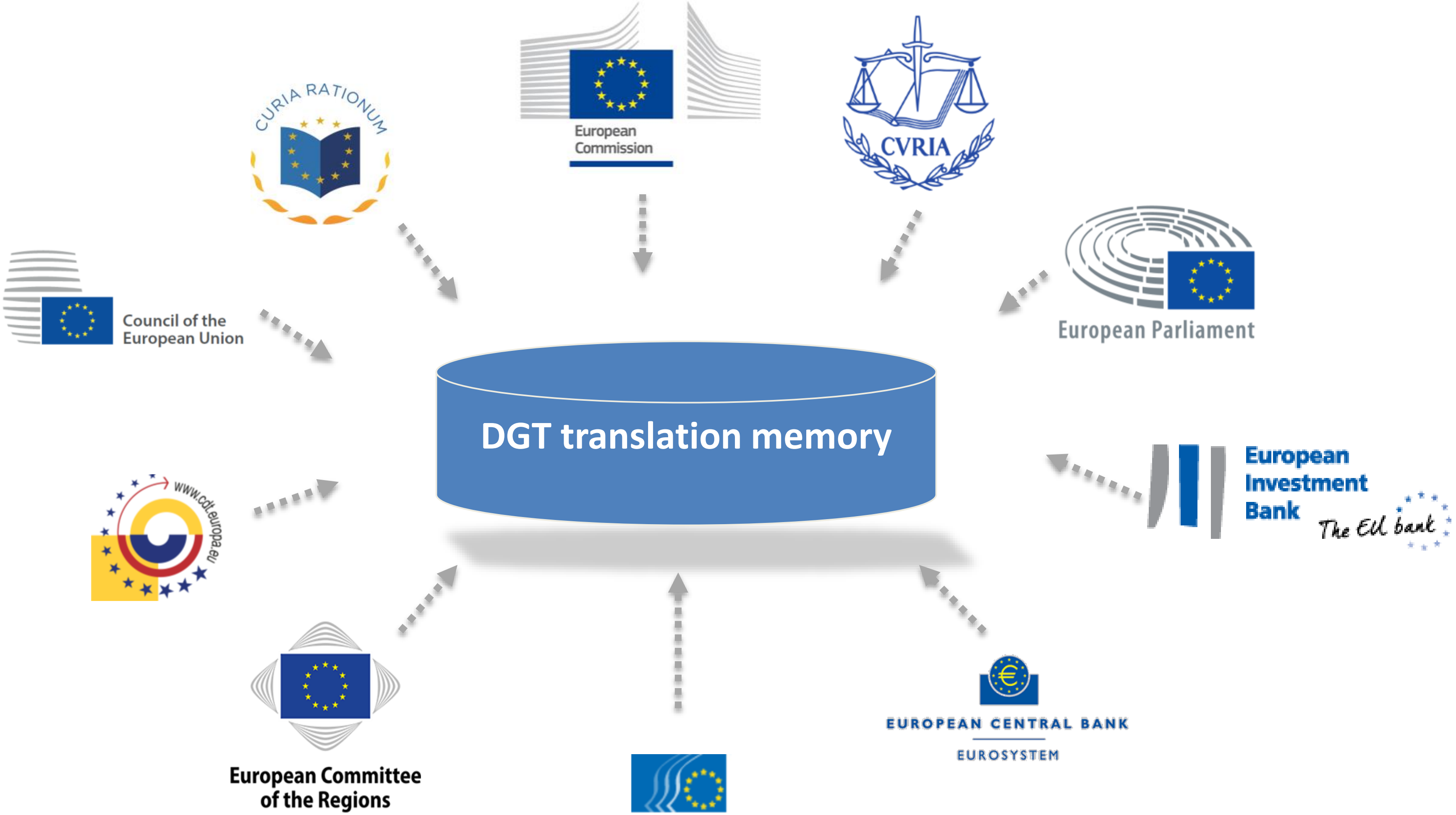
```
Amsterdam is the capital of
our country

It is a big and busy city

This is Dam square

It is the famous square of
Amsterdam
```

Data used by eTranslation



Such data are already available
BUT
they are not enough…

What does eTranslation need?

- Data residing in local public organisations, produced in-house or outsourced, e.g.
  - Reports
  - Communication
  - News
  - Web Content that is managed for several languages
  - Policies
  - Terminologies
  - Archives
  - Forms
  - FAQs

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)

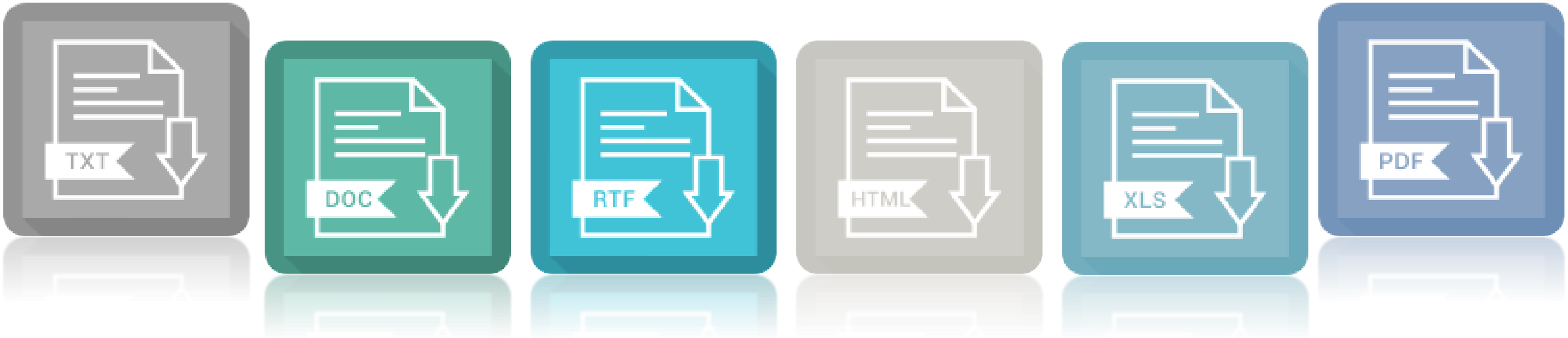
Dutch text

Ter toepassing van de bepalingen van deze gecoördineerde wet worden de landsbonden gemachtigd die het waren ter toepassing van het organiek koninklijk besluit van 22 september 1955 van de ziekte- en invaliditeitsverzekering

Translation in French

Sont agréées pour l'application des dispositions de la présente loi coordonnée les unions nationales qui l'étaient pour l'application de l'arrêté royal du 22 septembre 1955 organique de l'assurance maladie-invalidité.

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

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

Ter toepassing van de bepalingen van deze gecoördineerde wet worden de landsbonden gemachtigd die het waren ter toepassing van het organiek koninklijk besluit van 22 september 1955 van de ziekte- en invaliditeitsverzekering.

Sont agréées pour l'application des dispositions de la présente loi coordonnée les unions nationales qui l'étaient pour l'application de l'arrêté royal du 22 septembre 1955 organique de l'assurance maladie-invalidité.

De landsbonden waarborgen in hun statuten de bij deze wet bedoelde prestaties.

Les unions nationales garantissent, dans leurs statuts, les prestations prévues par la présente loi.

De machtiging van landsbonden die deze gecoördineerde wet of haar uitvoeringsbesluiten en – verordeningen niet naleven kan door de Koning, op advies of voorstel van het Algemeen comité van het Instituut, worden ingetrokken.

L'agréation peut être retirée par le Roi, sur avis ou sur proposition du Comité général de l'Institut, aux unions nationales qui n'observent pas la présente loi coordonnée ou ses arrêtés et règlements d'exécution.

Don’t merge the source and translated text into a single document

European Language Resource Coordination
Connecting Europe Facility

Preparing your data |2

Don'ts

Ter toepassing van de bepalingen van deze gecoördineerde wet worden de landsbonden gemachtigd die het waren ter toepassing van het organiek koninklijk besluit van 22 september 1955 van de ziekte- en invaliditeitsverzekering.

De landsbonden waarborgen in hun statuten de bij deze wet bedoelde prestaties.

De machtiging van landsbonden die deze gecoördineerde wet of haar uitvoeringsbesluiten en – verordeningen niet naleven kan door de Koning, op advies of voorstel van het Algemeen comité van het Instituut, worden ingetrokken.

Sont agréées pour l'application des dispositions de la présente loi coordonnée les unions nationales qui l'étaient pour l'application de l'arrêté royal du 22 septembre 1955 organique de l'assurance maladie-invalidité.

Les unions nationales garantissent, dans leurs statuts, les prestations prévues par la présente loi.

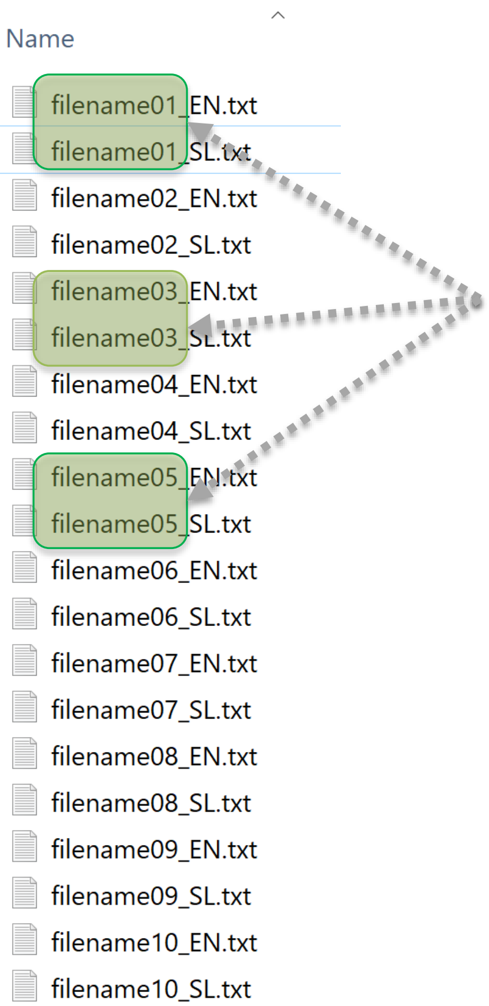
L'agréation peut être retirée par le Roi, sur avis ou sur proposition du Comité général de l'Institut, aux unions nationales qui n'observent pas la présente loi coordonnée ou ses arrêtés et règlements d'exécution.

Preparing your data |3

Don'ts

Ter toepassing van de bepalingen van deze gecoördineerde wet worden de landsbonden gemachtigd die het waren ter toepassing van het organiek koninklijk besluit van 22 september 1955 van de ziekte- en invaliditeitsverzekering.	Sont agréées pour l'application des dispositions de la présente loi coordonnée les unions nationales qui l'étaient pour l'application de l'arrêté royal du 22 septembre 1955 organique de l'assurance maladie-invalidité.
De landsbonden waarborgen in hun statuten de bij deze wet bedoelde prestaties.	Les unions nationales garantissent, dans leurs statuts, les prestations prévues par la présente loi.

Preparing your data |4



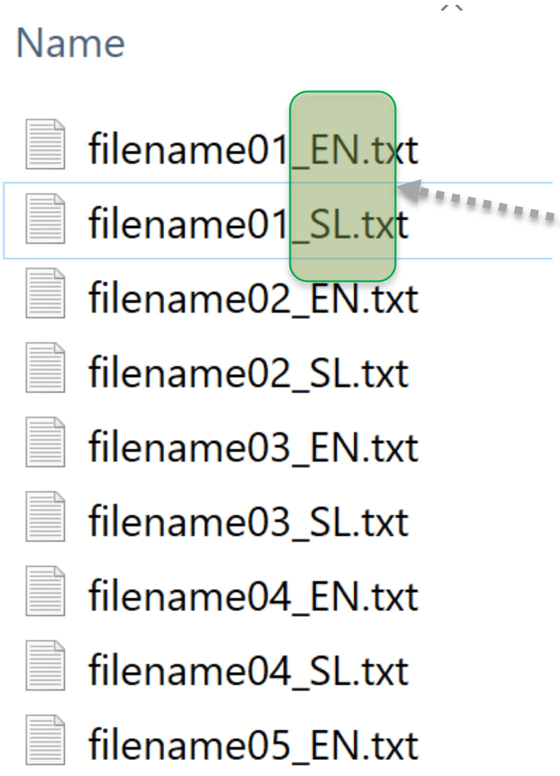
Name

filename01_EN.txt
filename01_SL.txt
filename02_EN.txt
filename02_SL.txt
filename03_EN.txt
filename03_SL.txt
filename04_EN.txt
filename04_SL.txt
filename05_EN.txt
filename05_SL.txt
filename06_EN.txt
filename06_SL.txt
filename07_EN.txt
filename07_SL.txt
filename08_EN.txt
filename08_SL.txt
filename09_EN.txt
filename09_SL.txt
filename10_EN.txt
filename10_SL.txt

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

Preparing your data |5

Do’s



Name

filename01_EN.txt

filename01_SL.txt

filename02_EN.txt

filename02_SL.txt

filename03_EN.txt

filename03_SL.txt

filename04_EN.txt

filename04_SL.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

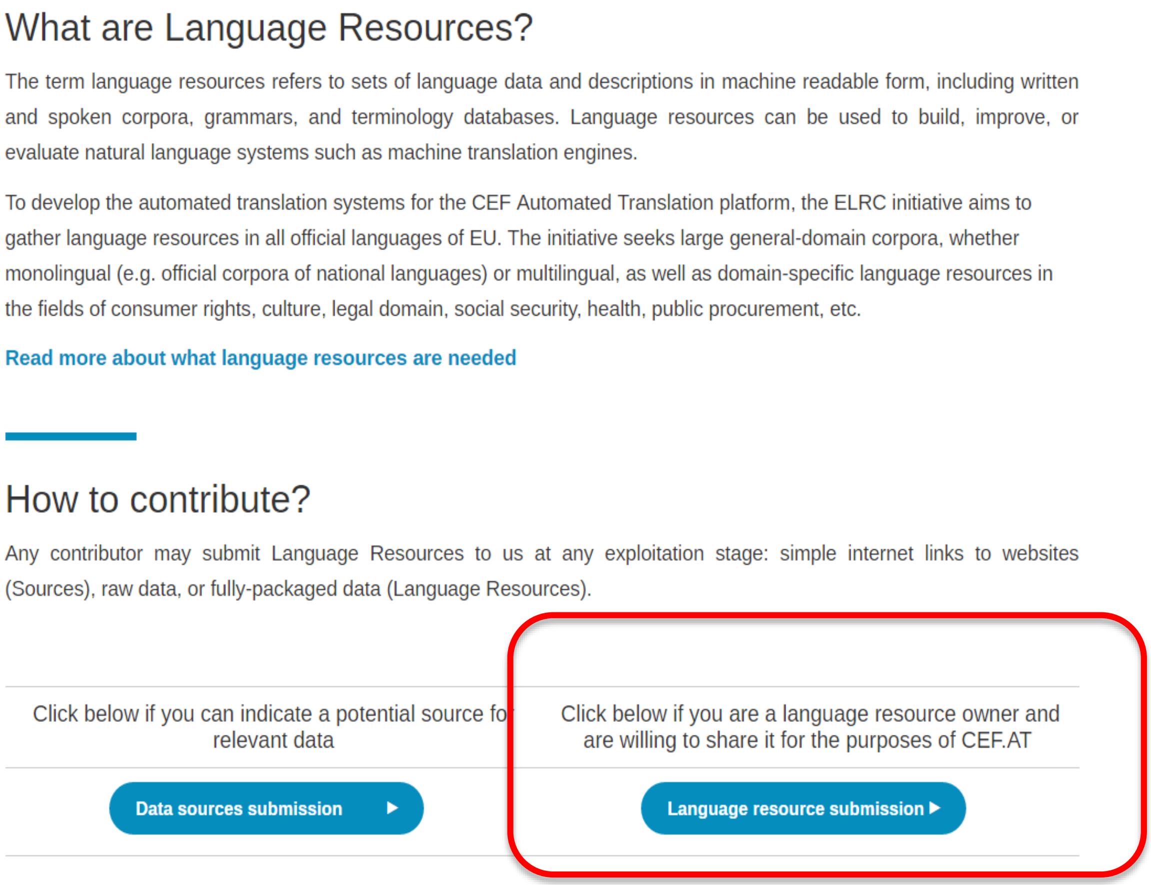
To ELRC-SHARE

- 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

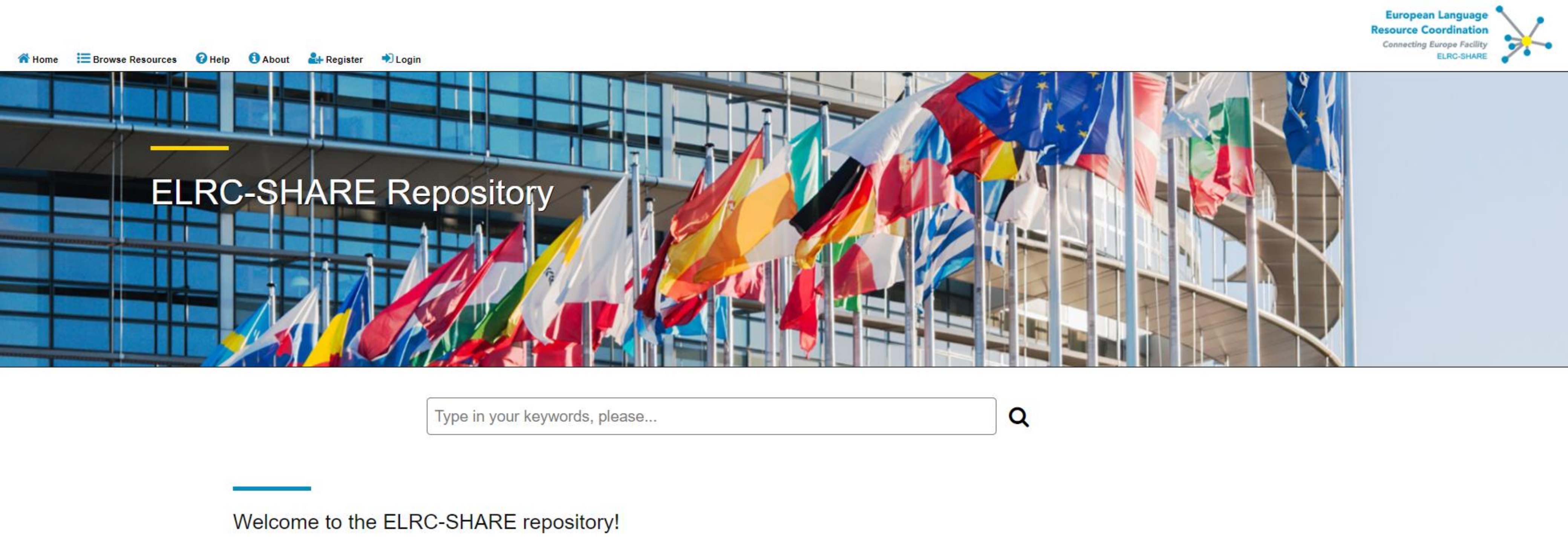
How to contribute?

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

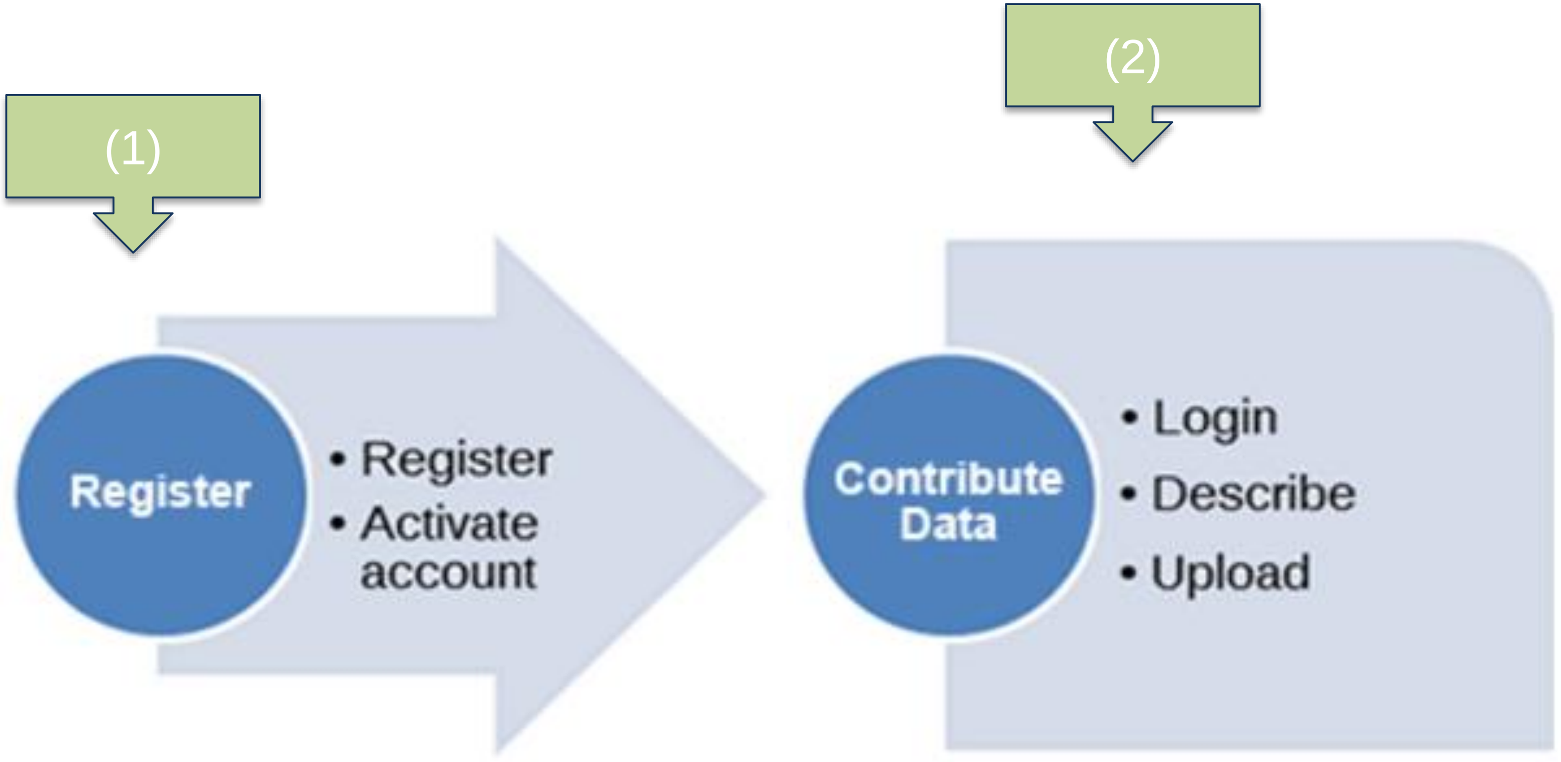
ELRC-SHARE repository



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 Contribute Data



(1)

Register

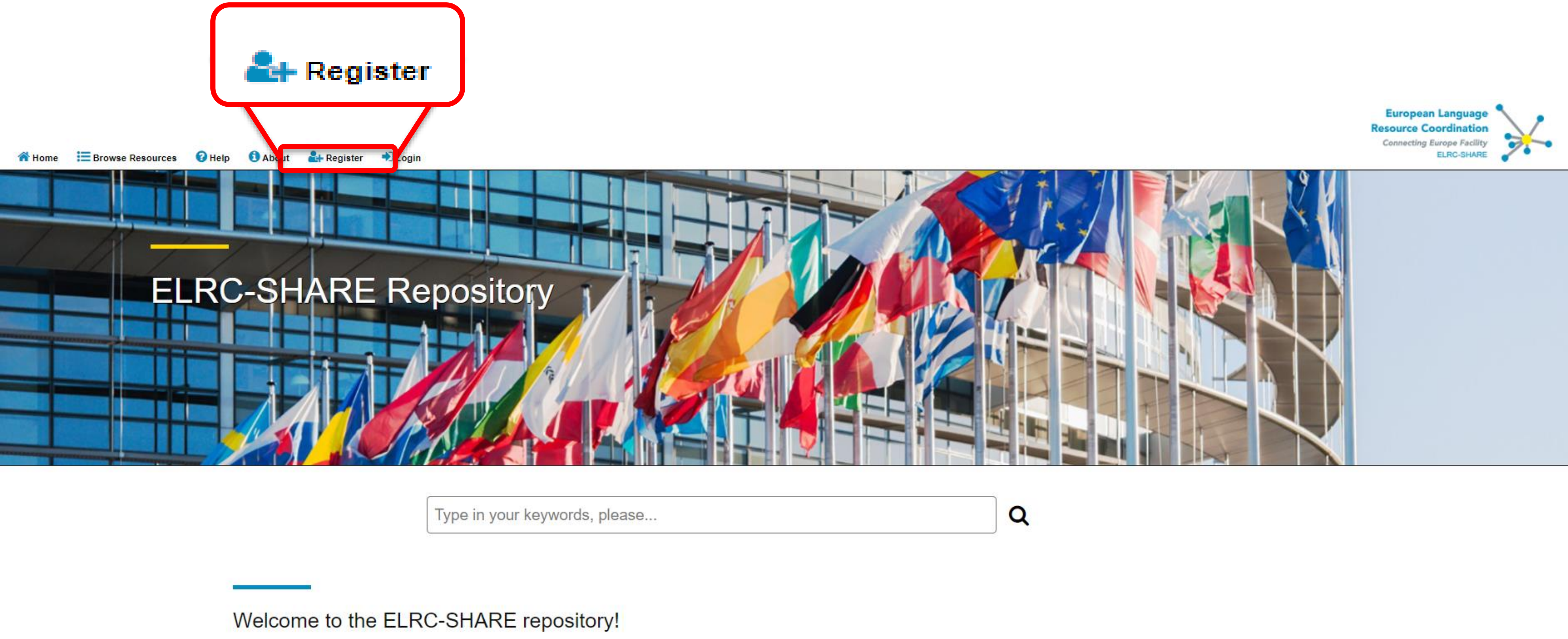
- Register
- Activate account

(2)

Contribute Data

- Login
- Describe
- Upload

How to Register (1/2)



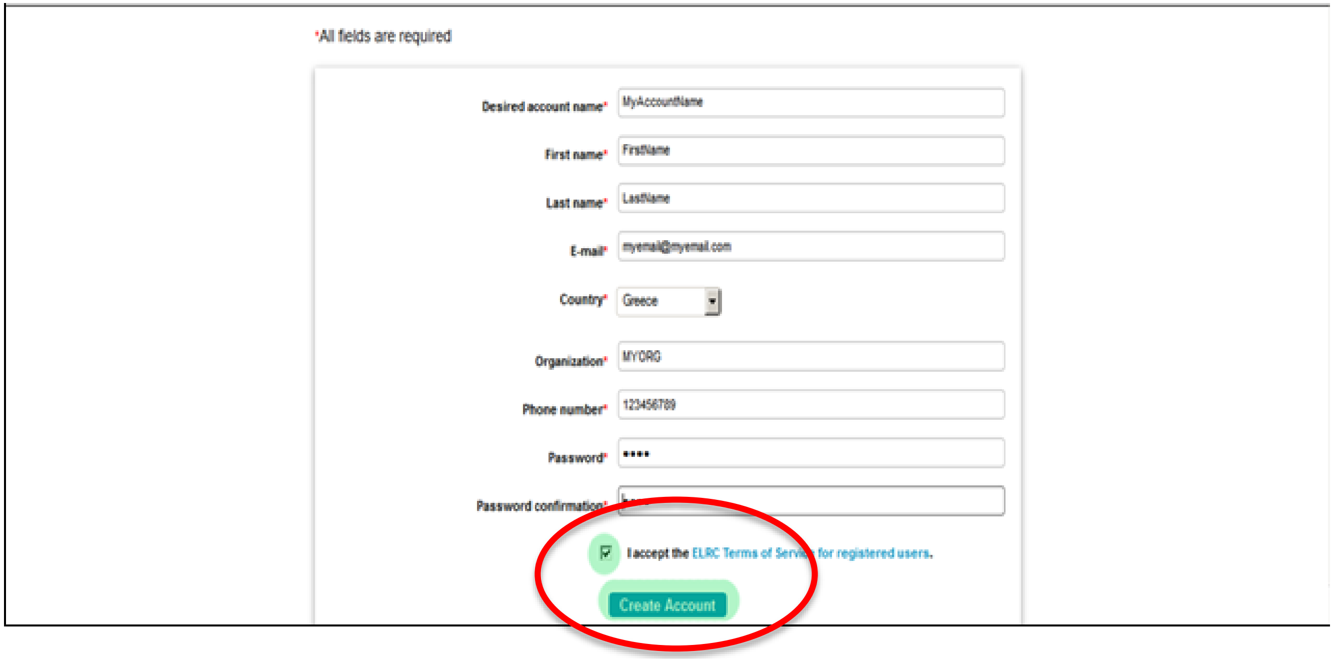
ELRC-SHARE Repository

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

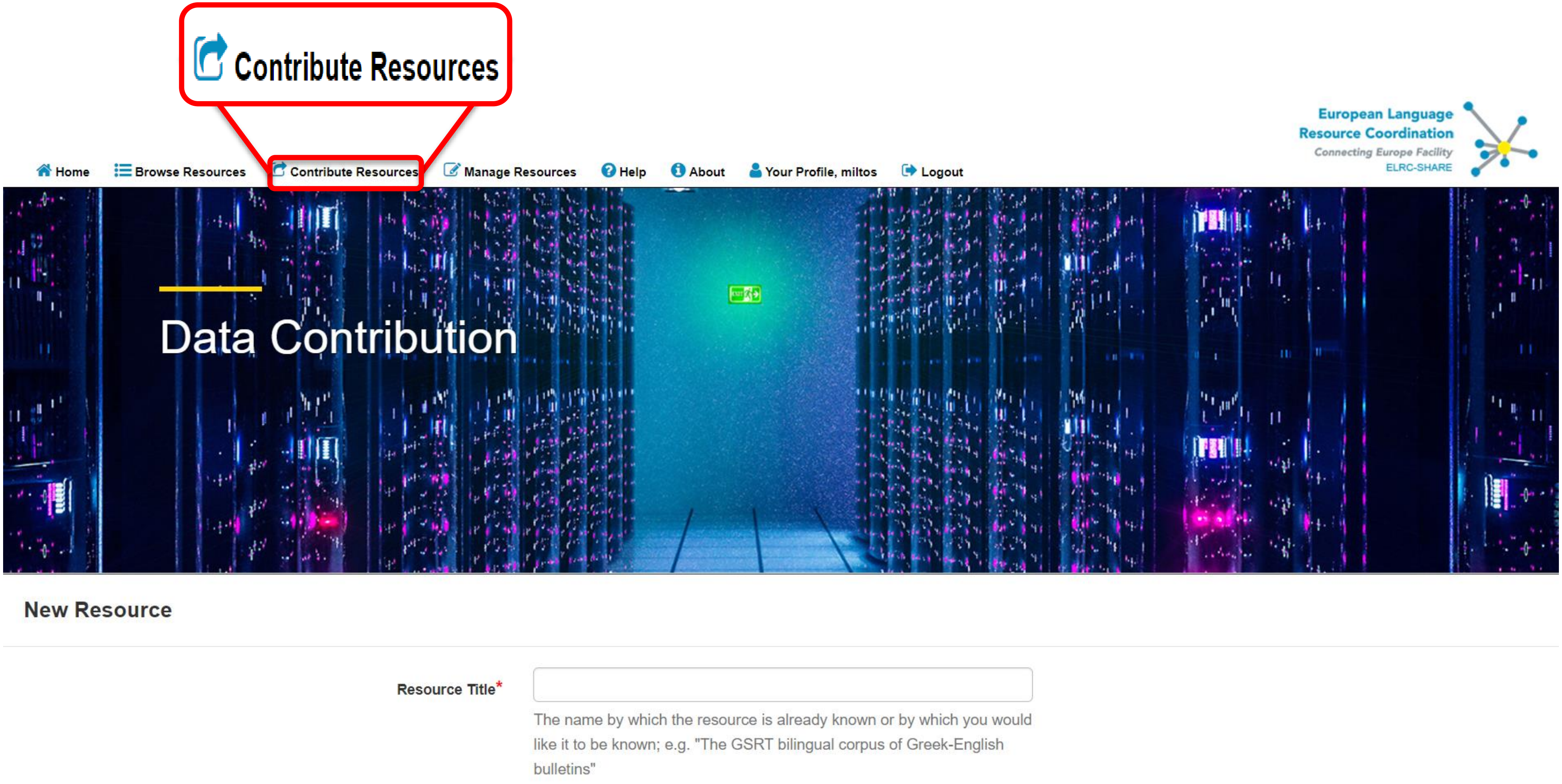


*All fields are required

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)



Data Contribution

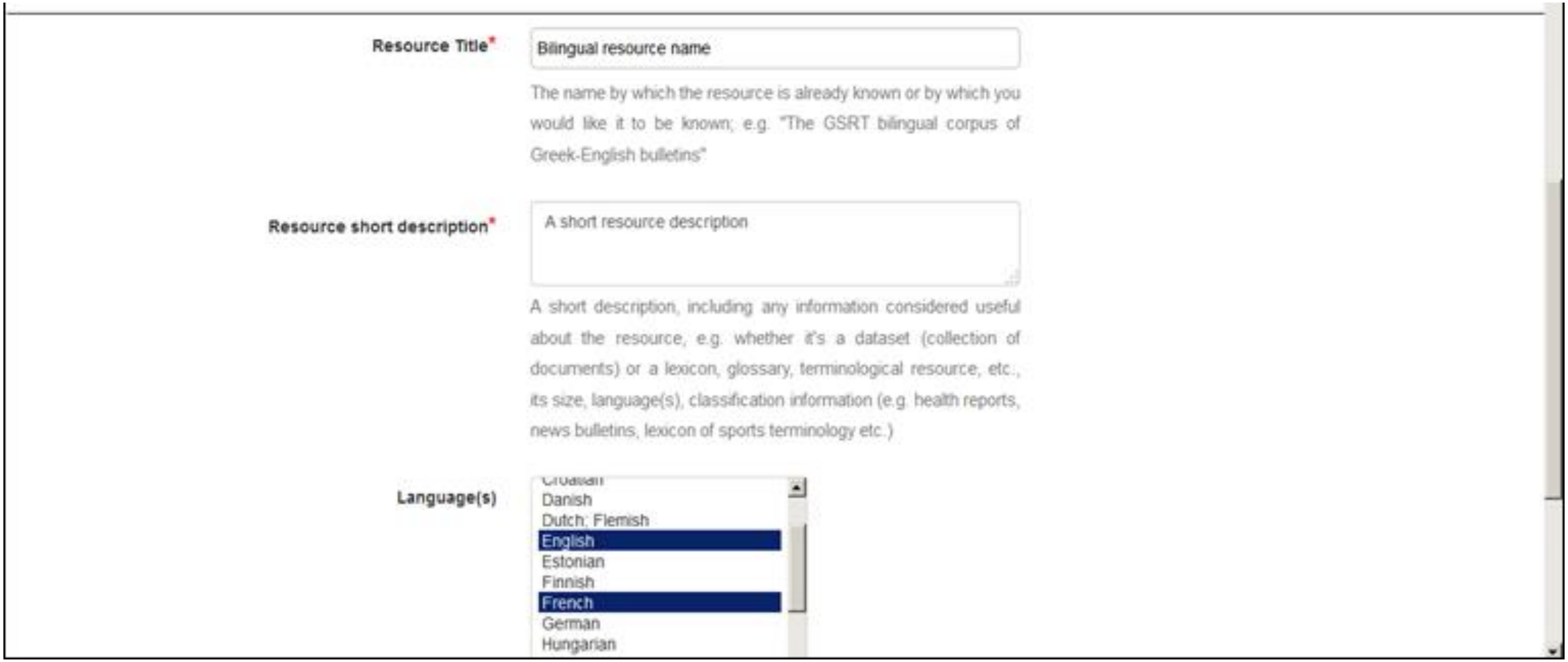
New Resource

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

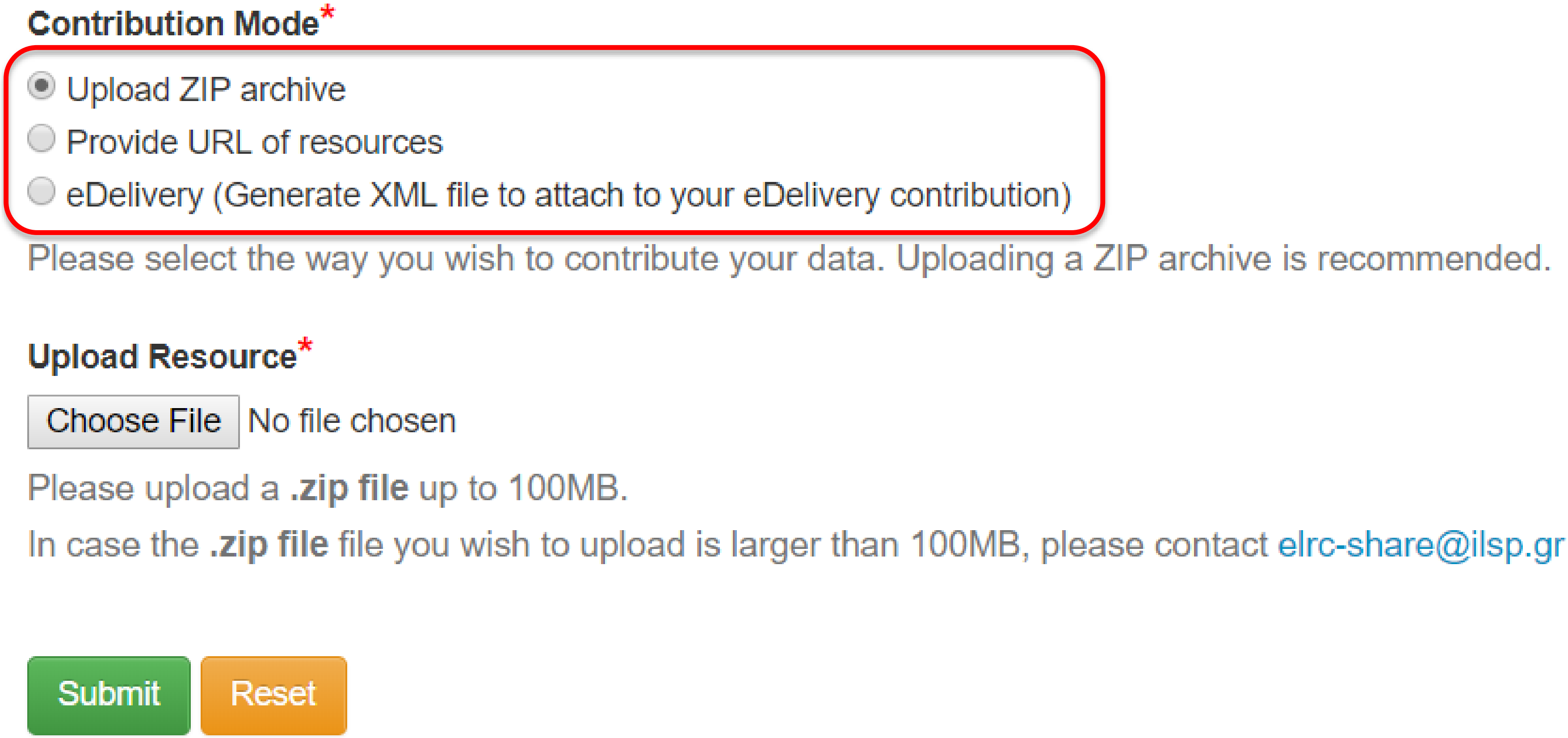
A short description, including any information considered useful about the resource, e.g. whether it's a dataset (collection of documents) or a lexicon, glossary, terminological resource, etc., its size, language(s), classification information (e.g. health reports, news bulletins, lexicon of sports terminology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

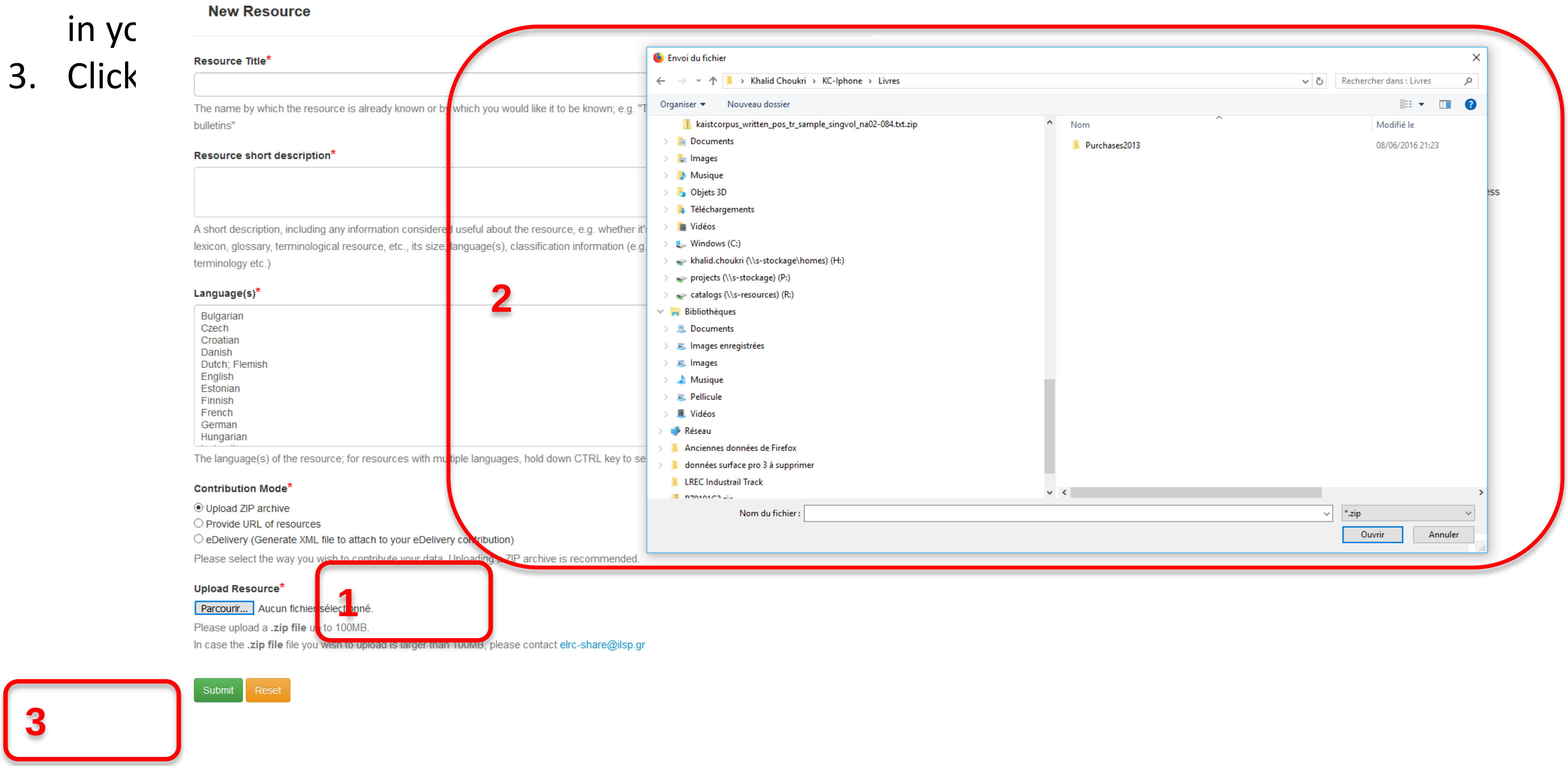
Choose File No file chosen

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 yc
3. Click



New Resource

Resource Title*

The name by which the resource is already known or b which you would like it to be known; e.g. "T bulletins"

Resource short description*

A short description, including any information considere useful about the resource, e.g. whether it' lexicon, glossary, terminological resource, etc., its size language(s), classification information (e.g. terminology etc.)

Language(s)*

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

The language(s) of the resource; for resources with mu tiple languages, hold down CTRL key to se

Contribution Mode*

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

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

Upload Resource*

Parcourir... Aucun fichier sélectionné.

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

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

Submit Reset

Envoi du fichier

Khalid Choukri › KC-Iphone › Livres

Rechercher dans : Livres

Organiser ▾ Nouveau dossier

kaistcorpus_written_pos_tr_sample_singvol_na02-084.txt.zip
Documents
Images
Musique
Objets 3D
Téléchargements
Vidéos
Windows (C:)
khalid.choukri (\\s-stockage\homes) (H:)
projects (\\s-stockage) (P:)
catalogs (\\s-resources) (R:)
Bibliothèques
Documents
Images enregistrées
Images
Musique
Pellicule
Vidéos
Réseau
Anciennes données de Firefox
données surface pro 3 à supprimer
LREC Industrail Track

Nom	Modifié le
Purchases2013	08/06/2016 21:23

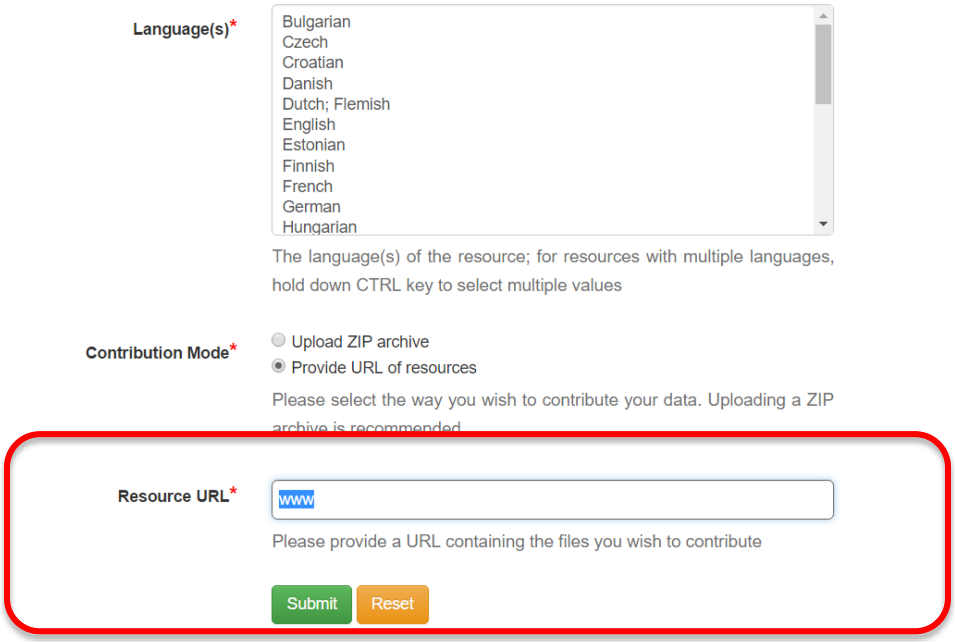
Nom du fichier :

*.zip

Ouvrir Annuler

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

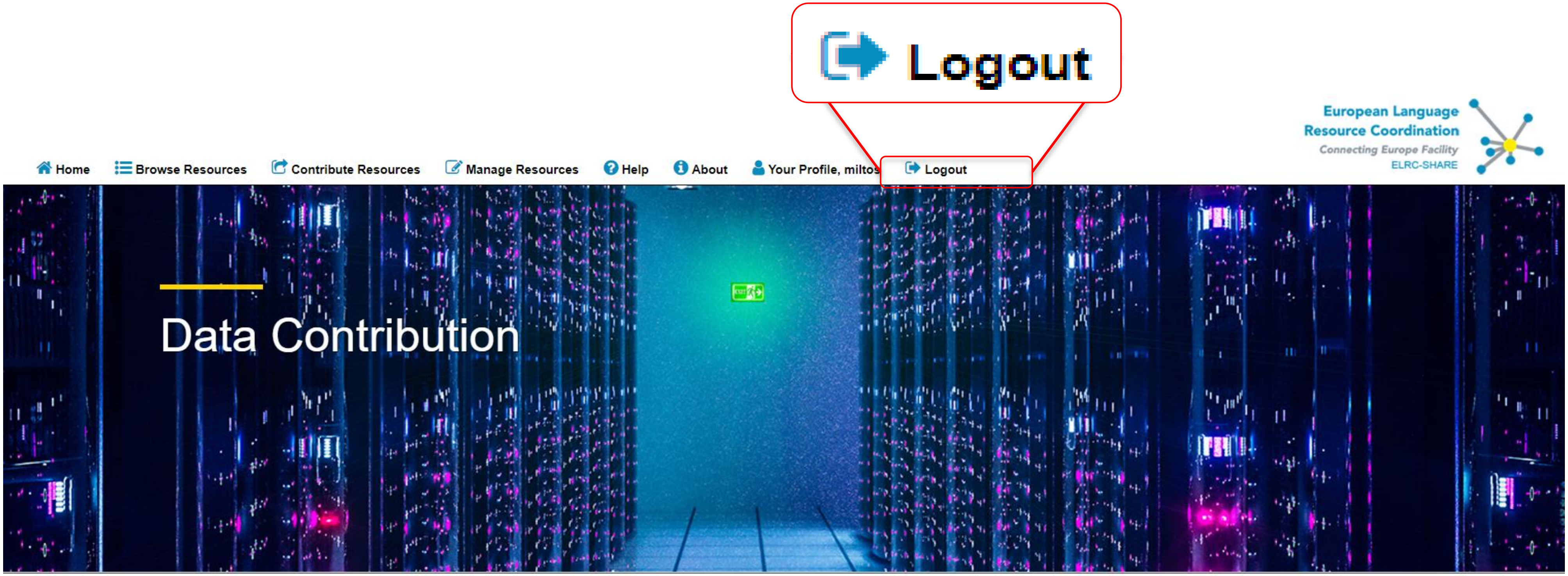
Please select the way you wish to contribute your data. Uploading a ZIP archive is recommended

Resource URL* www

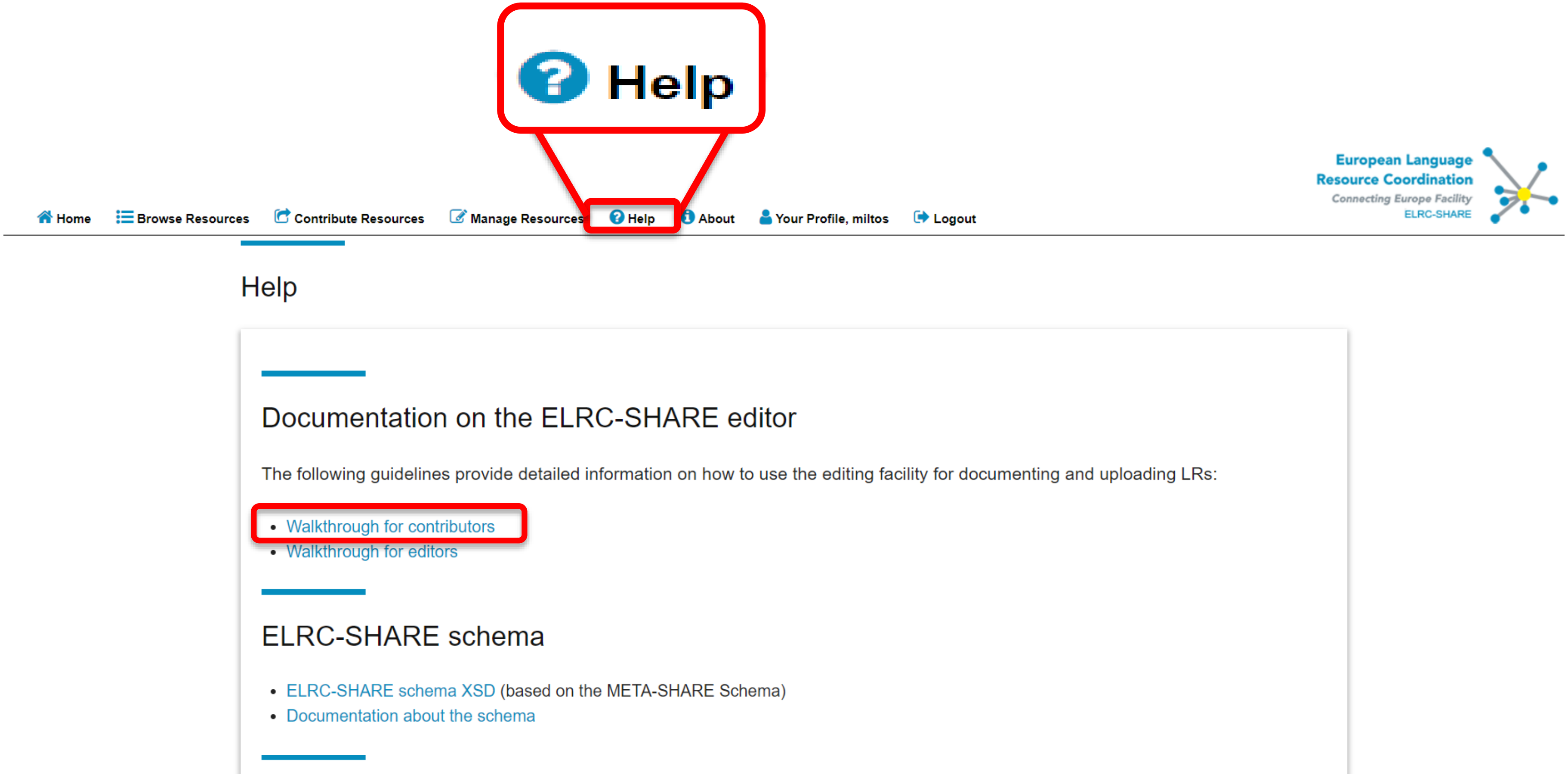
Please provide a URL containing the files you wish to contribute

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:

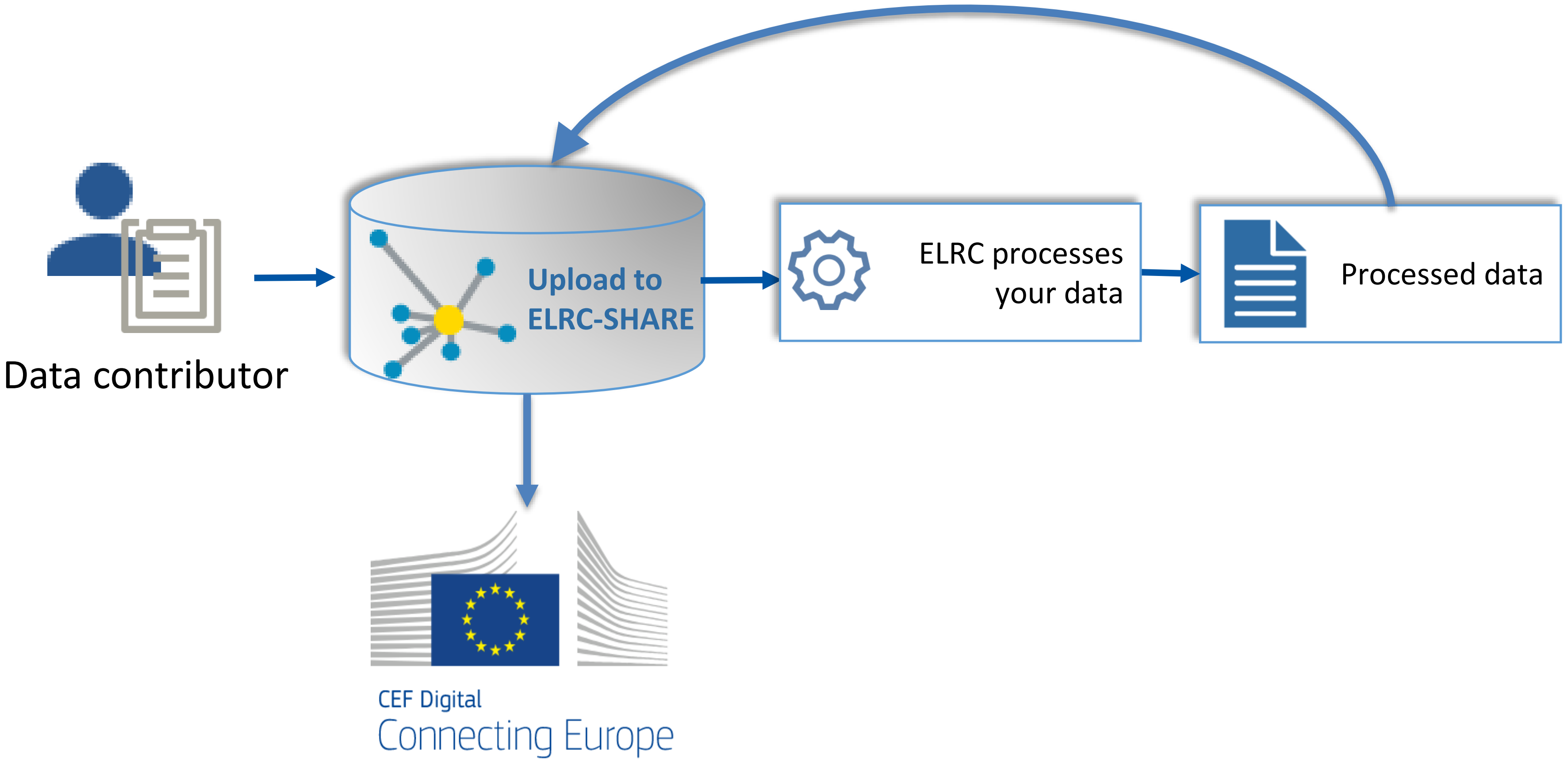
- Walkthrough for contributors
- Walkthrough for editors

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?



Data processing before delivery to EC

- All datasets are processed to result in tmx/tbx/txt files
- Data will indicatively undergo the following processing:
  - cleaning
  - format conversion
  - sentence alignment
  - metadata completion

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

Assistance will be provided in close cooperation with a broad network of language experts

On-site assistance

We will (help) fix your data issues and return the processed data directly to you.

We can also help to improve your data management processes. Just ask!

Language processing services on-site

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!

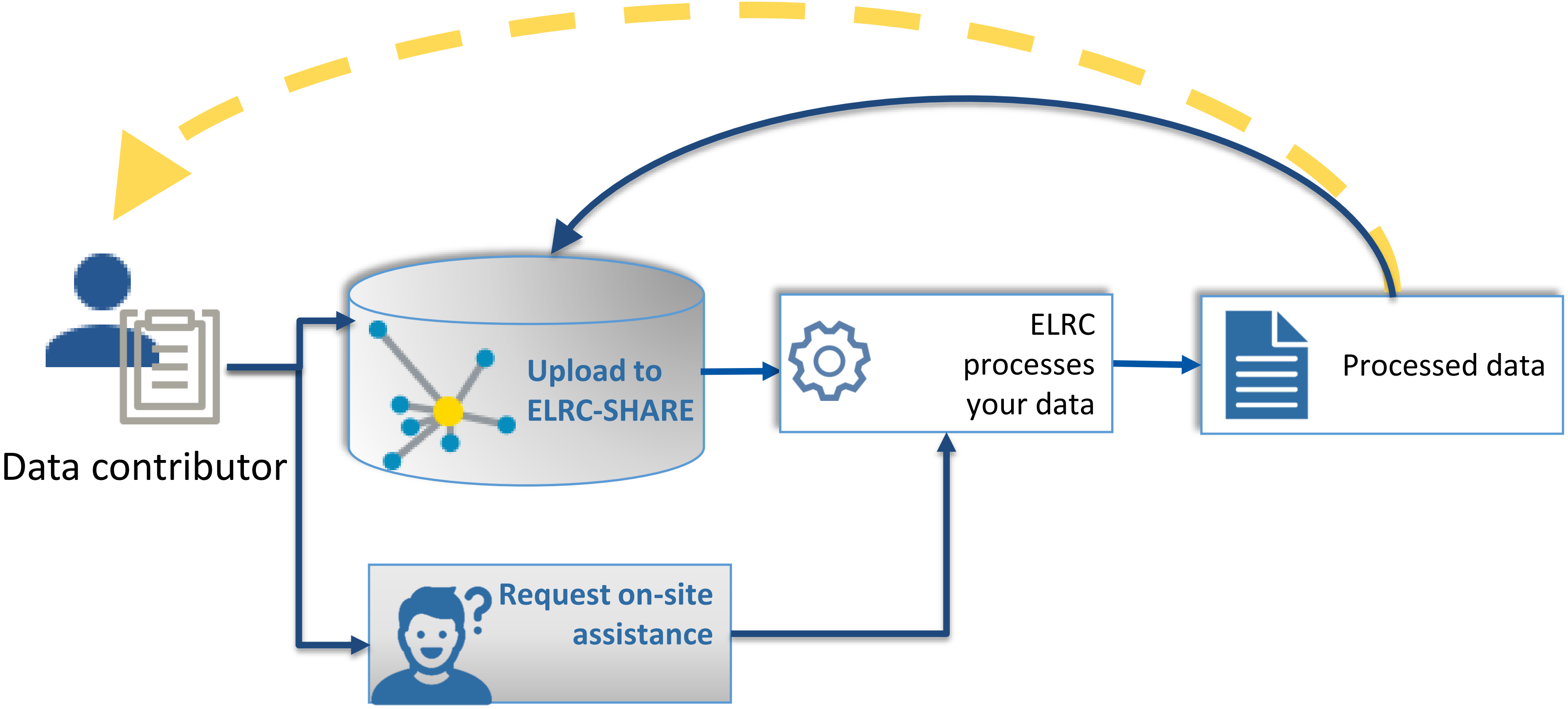
Alignment

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

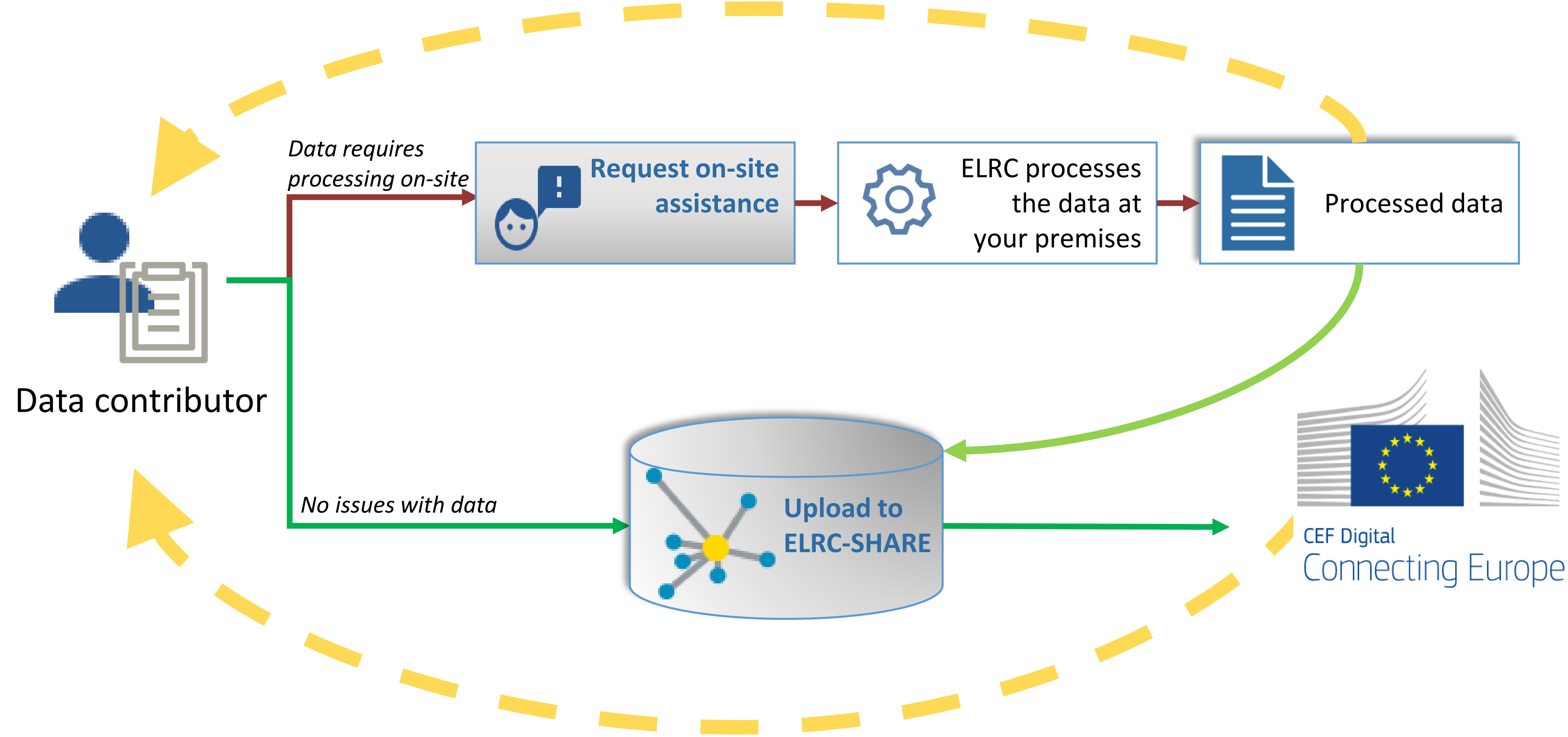
Metadata

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

What happens to your data?

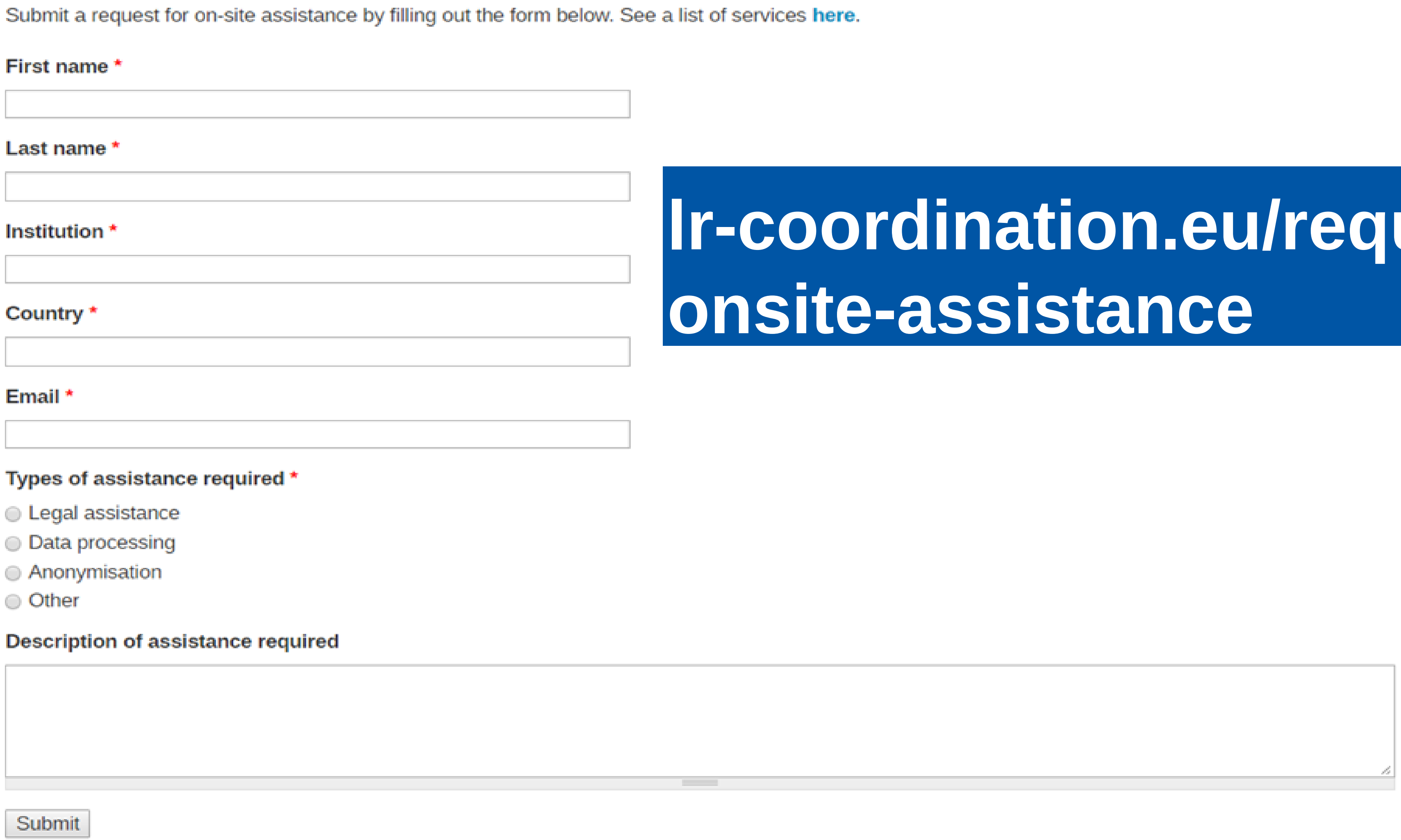


On-site assistance |2


Data requires processing on-site
Request on-site assistance
ELRC processes the data at your premises
Processed data
Data contributor
No issues with data
Upload to ELRC-SHARE
CEF Digital
Connecting Europe

How to request services and help

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 *

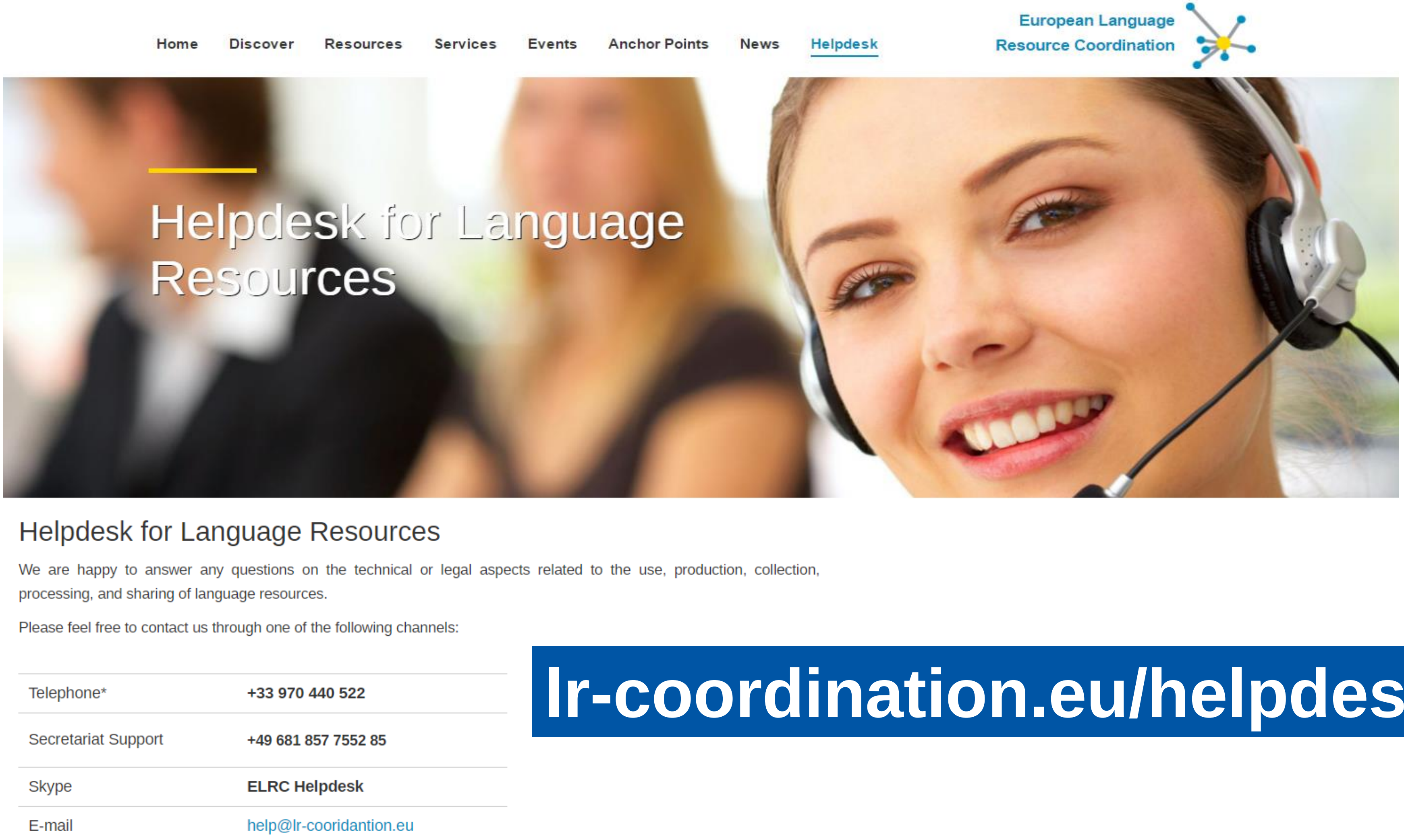
Types of assistance required *

○ Legal assistance
○ Data processing
○ Anonymisation
○ Other

Description of assistance required

lr-coordination.eu/request-onsite-assistance

ELRC Helpdesk



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

Thank you