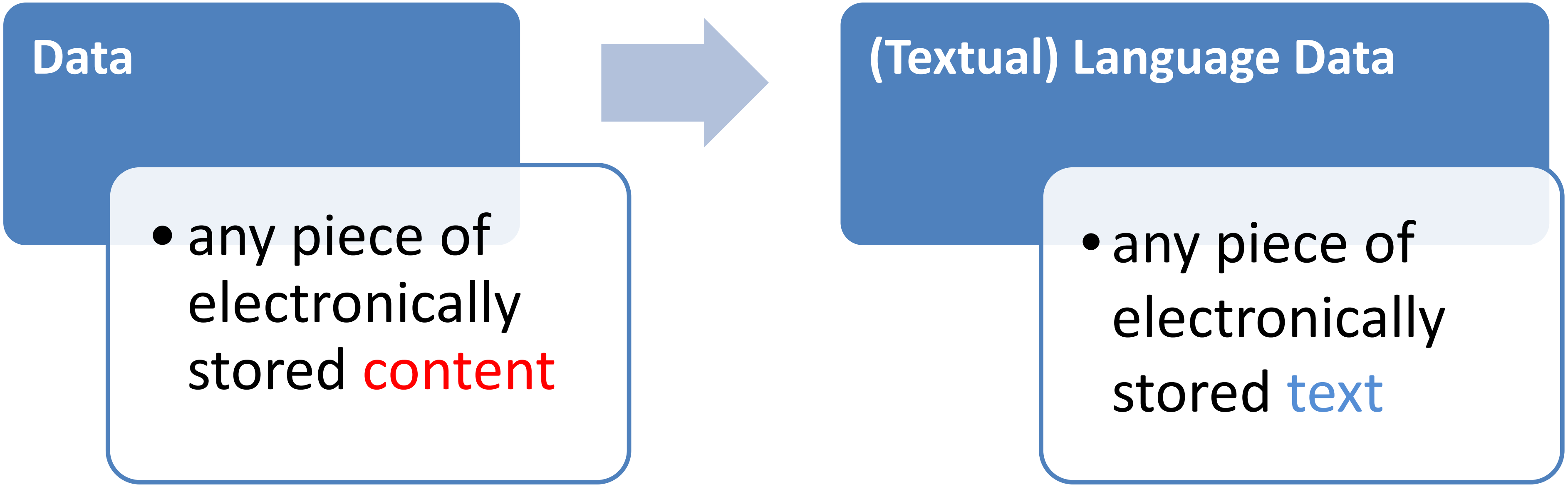
Preparing and sharing data via the ELRC-SHARE repository
ELRC on site Services and what happens next

Italian Workshop

Khalid Choukri, Hélène Mazo
ELRA/ELDA

The notion of language data



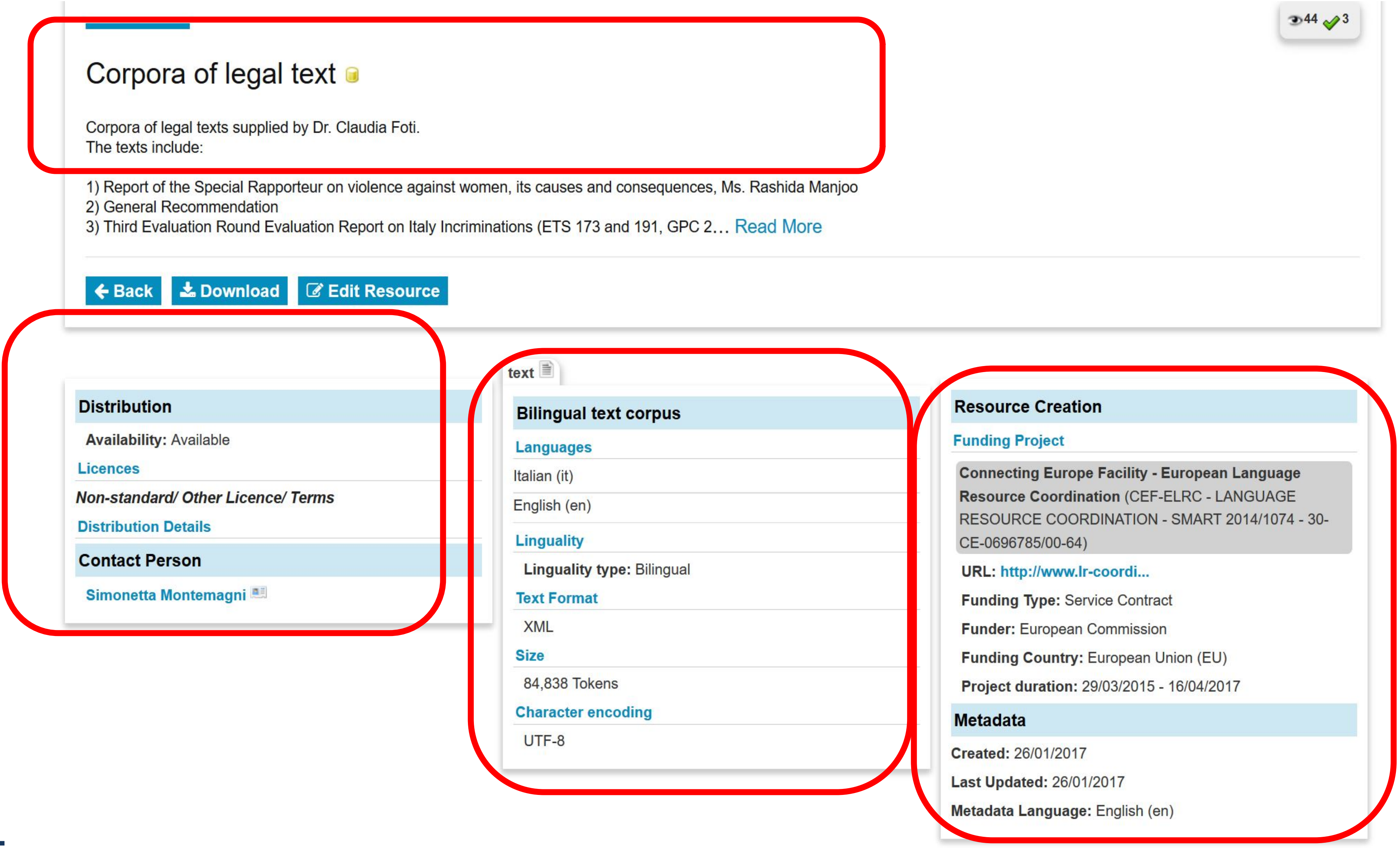
Data

- any piece of electronically stored content

(Textual) Language Data

- any piece of electronically stored text

The notion of data in the context of eTranslation



👁44 ✓3

Corpora of legal text

Corpora of legal texts supplied by Dr. Claudia Foti.
The texts include:

1) Report of the Special Rapporteur on violence against women, its causes and consequences, Ms. Rashida Manjoo
2) General Recommendation
3) Third Evaluation Round Evaluation Report on Italy Incriminations (ETS 173 and 191, GPC 2… Read More

Distribution

Availability: Available

Licences

Non-standard/ Other Licence/ Terms

Distribution Details

Contact Person

Simonetta Montemagni

text

Bilingual text corpus

Languages

Italian (it)

English (en)

Linguality

Linguality type: Bilingual

Text Format

XML

Size

84,838 Tokens

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: 26/01/2017

Last Updated: 26/01/2017

Metadata Language: English (en)

The notion of data
in the context of eTranslation

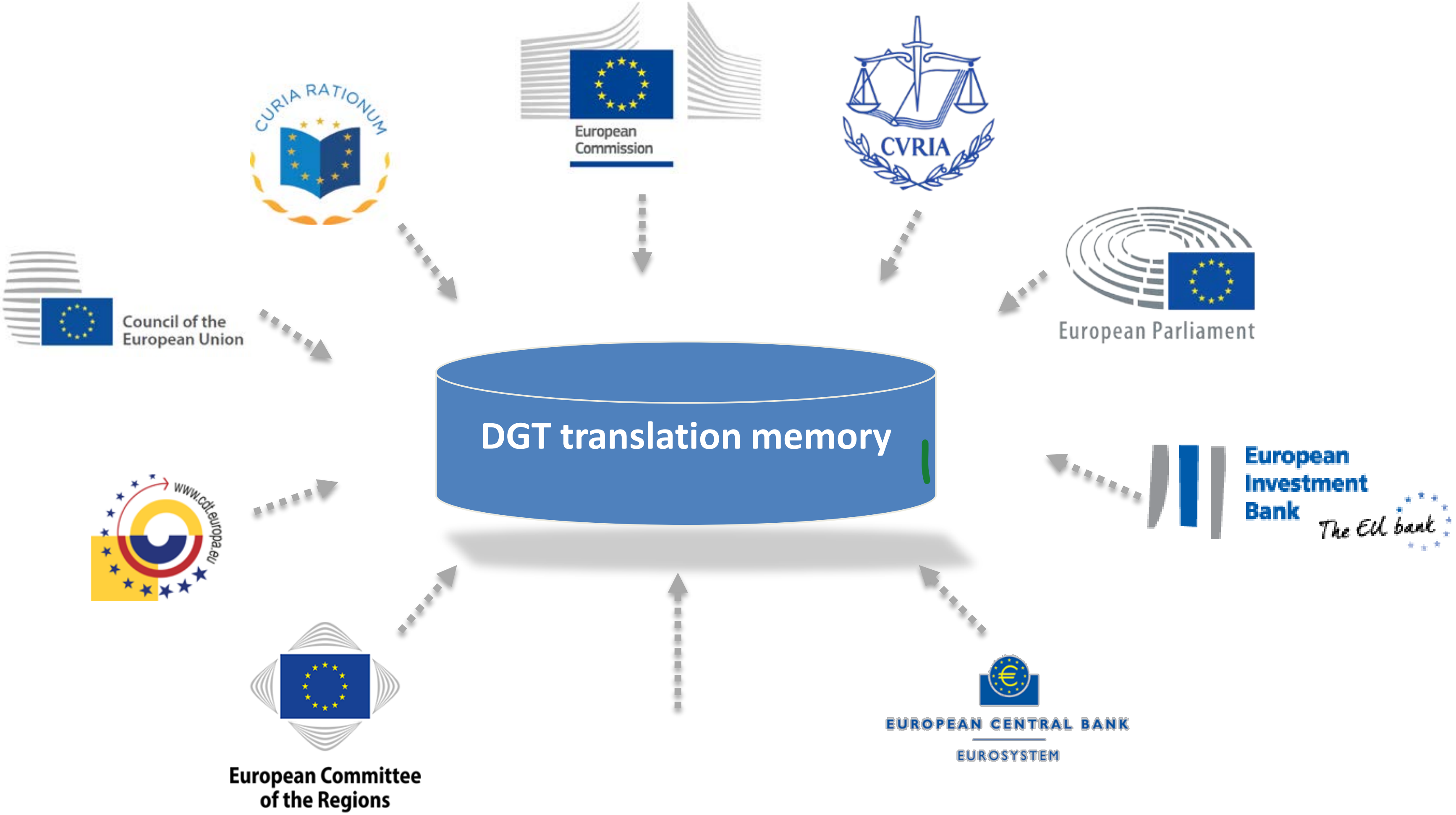
```
File01_it.txt
File01_en.txt
File02_it.txt
File02_en.txt
File03_it.txt
…
```

Trans.
Da

With reference to the procedures referred to in paragraph 16, letter b), of this Article, the contracting authorities shall in any event be required to publish on their institutional websites: the proposing structure; the object of the call; the list of the operators invited to tender; the awarded contractor; the amount of the contract; the timing of completion of the works, service or supply; the amount of the sums paid. By 31 January of each year, such information relevant to the previous year shall be published in summary tables, freely downloadable in a standard open digital format enabling analysis and re-processing, also for statistical purposes, of the data contained.

Con riferimento ai procedimenti di cui al comma 16, lettera b), del presente articolo, le stazioni appaltanti sono in ogni caso tenute a pubblicare nei propri siti web istituzionali: la struttura proponente; l'oggetto del bando; l'elenco degli operatori invitati a presentare offerte; l'aggiudicatario; l'importo di aggiudicazione; i tempi di completamento dell'opera, servizio o fornitura; l'importo delle somme liquidate. Entro il 31 gennaio di ogni anno, tali informazioni, relativamente all'anno precedente, sono pubblicate in tabelle riassuntive rese liberamente scaricabili in un formato digitale standard aperto che consenta di analizzare e rielaborare, anche a fini statistici, i dati informatici.

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)

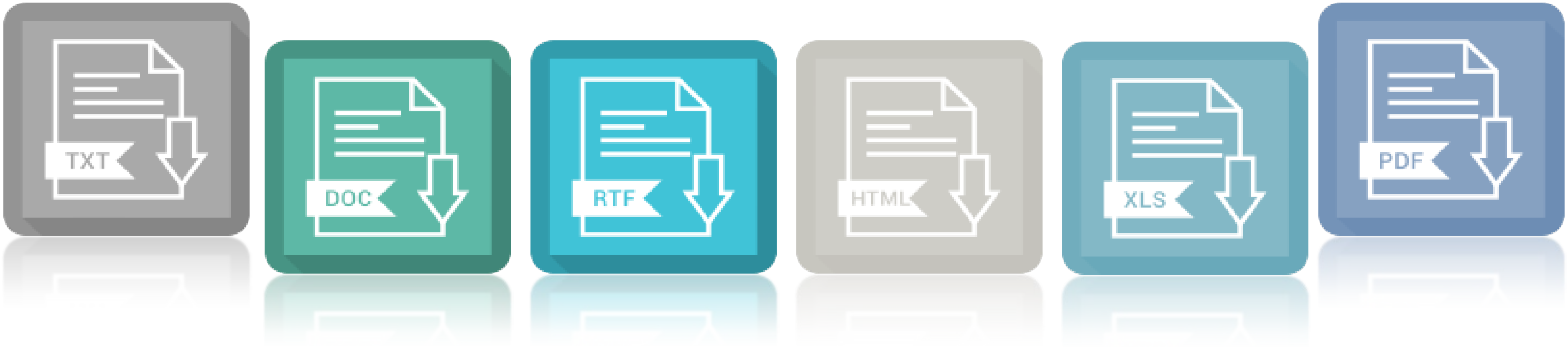
Portuguese text

entregue na câmara municipal juntamente com uma relação de todos os seus delegados com a indicação da assembleia ou secção de voto para que foram designados, nos prazos e para os efeitos legais.

Translation in French

Ce document doit être rempli par le parti politique et déposé à la mairie avec la liste de tous ses délégués, en mentionnant la section ou le bureau de vote auquel ils ont été affectés, dans les délais fixés et à toutes fins légales.

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

O site na Internet também tem uma seção Vítimas e Testemunhas que talvez você considere útil.

En la página web, también encontrará una sección para víctimas y testigos que le puede resultar útil.

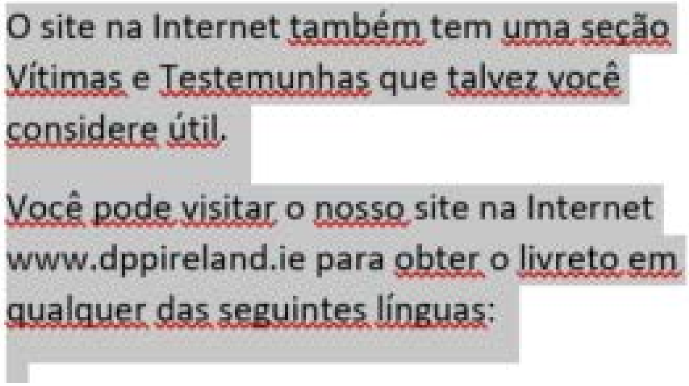
Você pode visitar o nosso site na Internet www.dppireland.ie para obter o livreto em qualquer das seguintes línguas:

Puede visitar nuestro sitio web www.dppireland.ie para descargar el folleto en cualquiera de los idiomas siguientes:

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

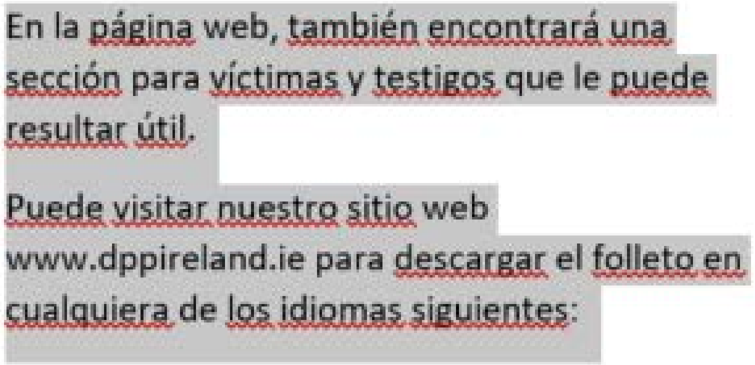
Preparing your data |2

Don’ts



O site na Internet também tem uma seção Vítimas e Testemunhas que talvez você considere útil.

Você pode visitar o nosso site na Internet www.dppireland.ie para obter o livreto em qualquer das seguintes línguas:



En la página web, también encontrará una sección para víctimas y testigos que le puede resultar útil.

Puede visitar nuestro sitio web www.dppireland.ie para descargar el folleto en cualquiera de los idiomas siguientes:

Preparing your data |3

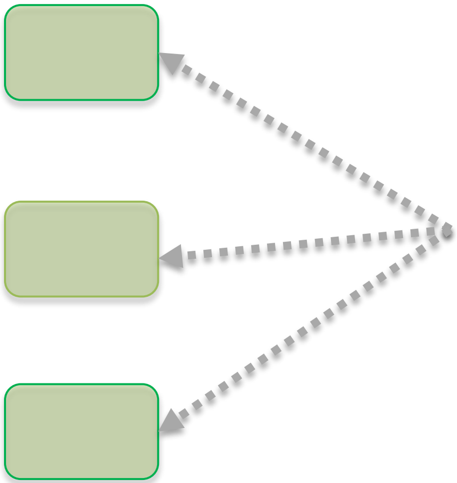
Don’ts

O site na Internet também tem uma seção Vítimas e Testemunhas que talvez você considere útil.
Você pode visitar o nosso site na Internet www.dppireland.ie para obter o livreto em qualquer das seguintes línguas:

En la página web, también encontrará una sección para víctimas y testigos que le puede resultar útil.
Puede visitar nuestro sitio web www.dppireland.ie para descargar el folleto en cualquiera de los idiomas siguientes:

Preparing your data |4

Do’s



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

Preparing your data |5

Do's

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

European Language Resource Coordination
Connecting Europe Facility

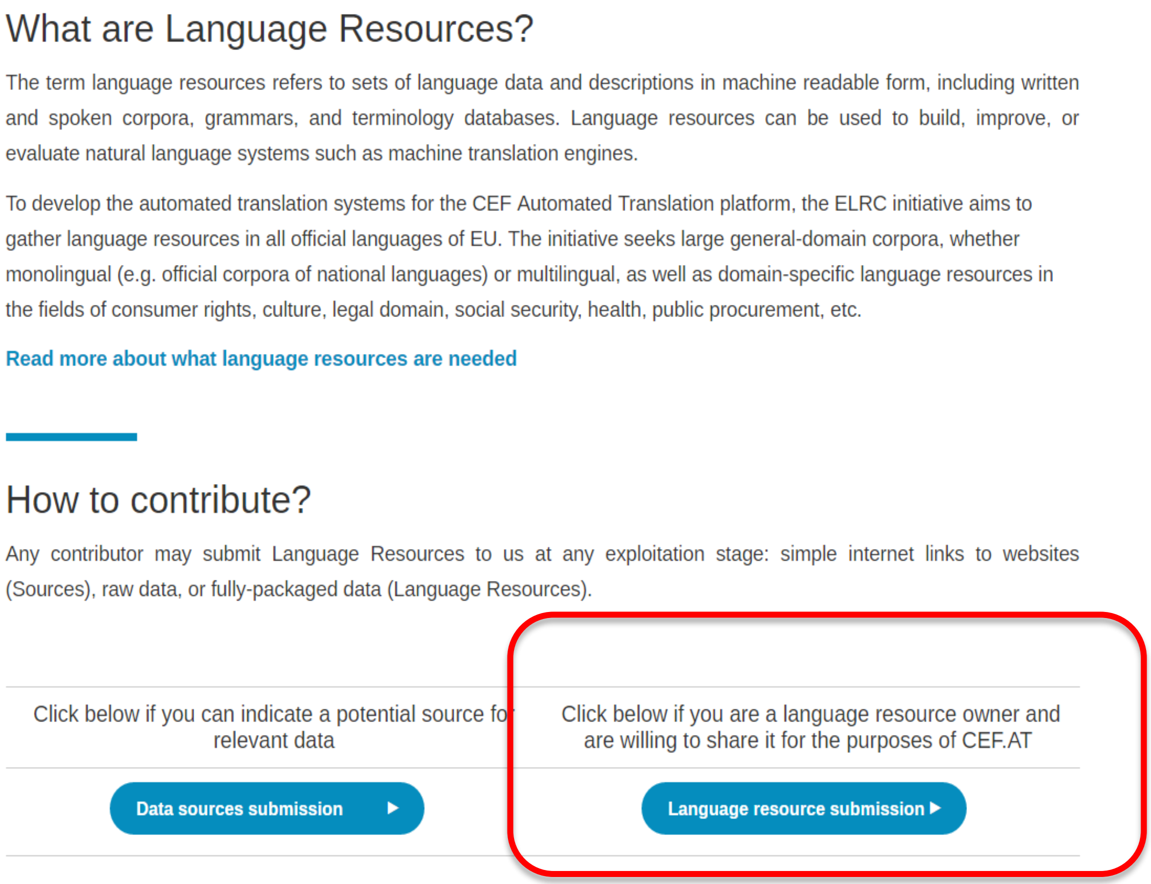
To ELRC-SHARE

- 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

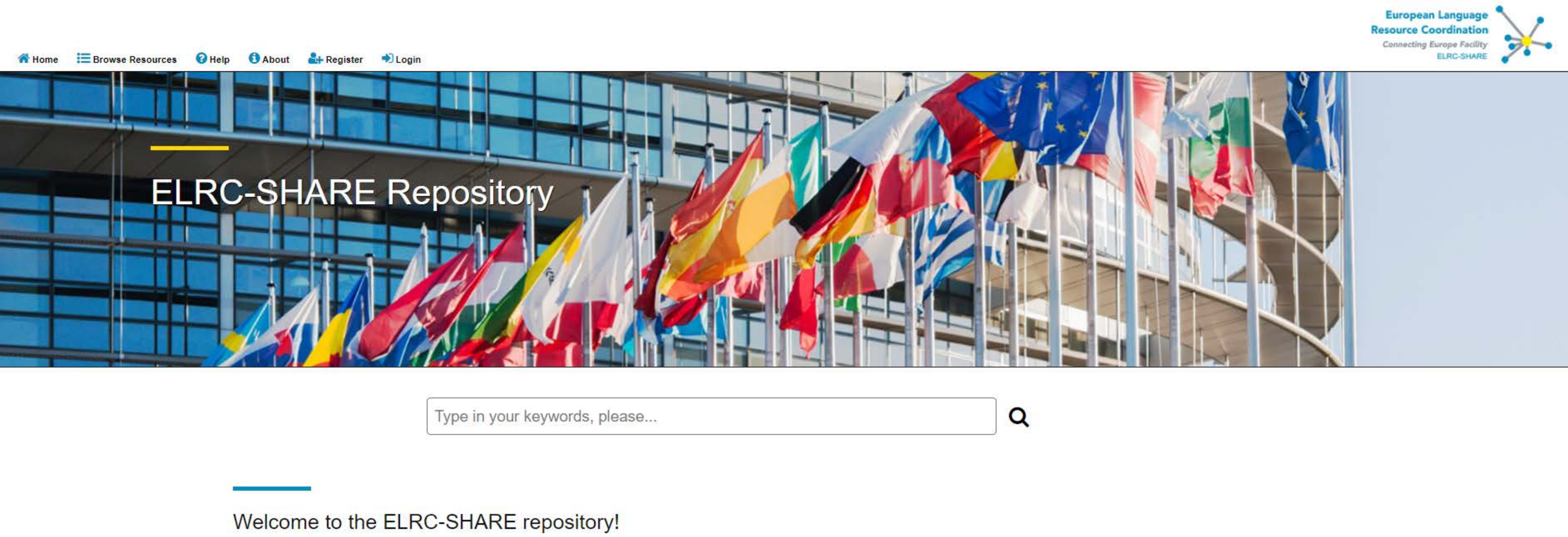
How to contribute?

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 for 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

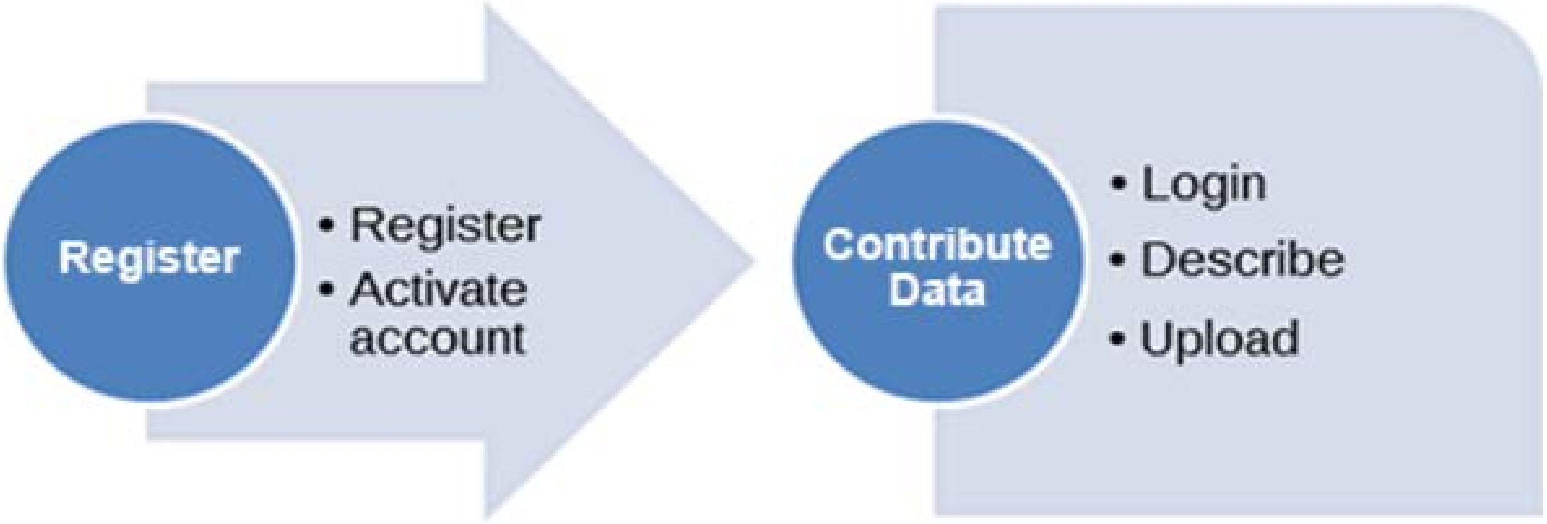


ELRC-SHARE Repository

Type in your keywords, please...

Welcome to the ELRC-SHARE repository!

How to Contribute Data



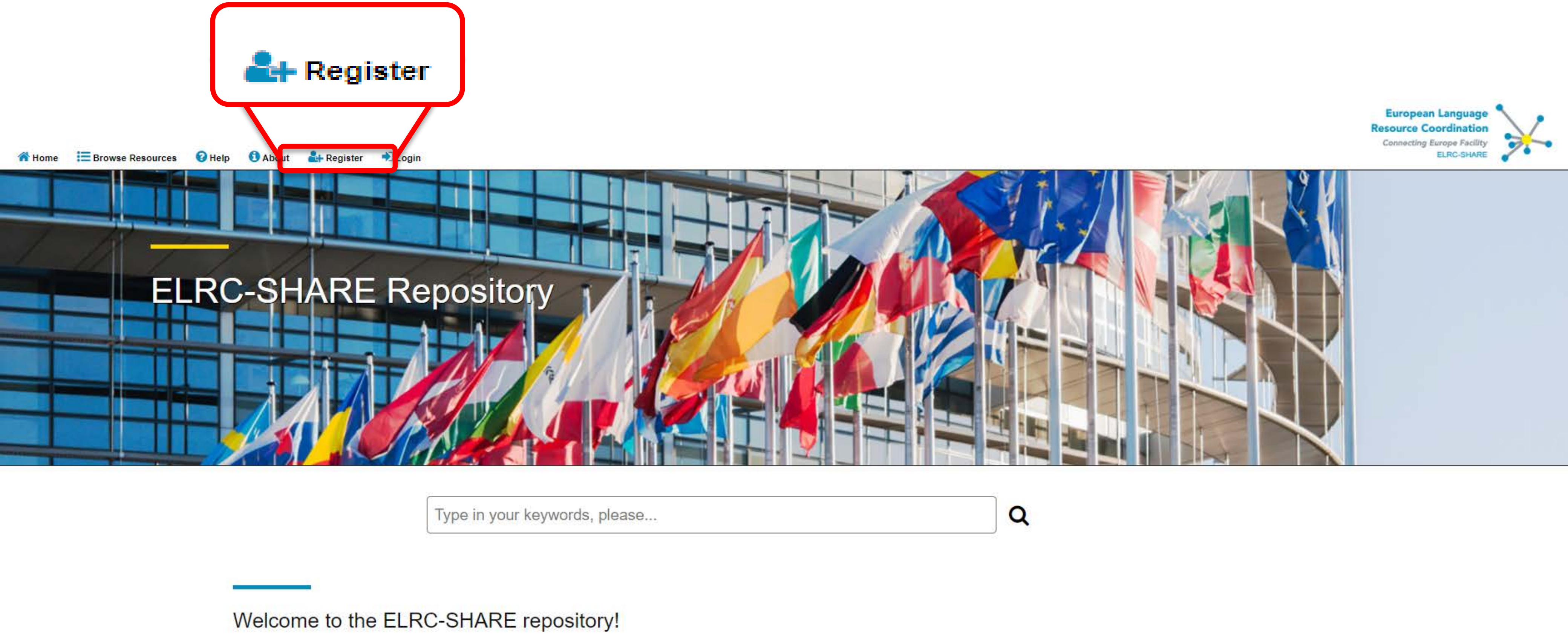
Register

- Register
- Activate account

Contribute Data

- Login
- Describe
- Upload

How to Register (1/2)


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

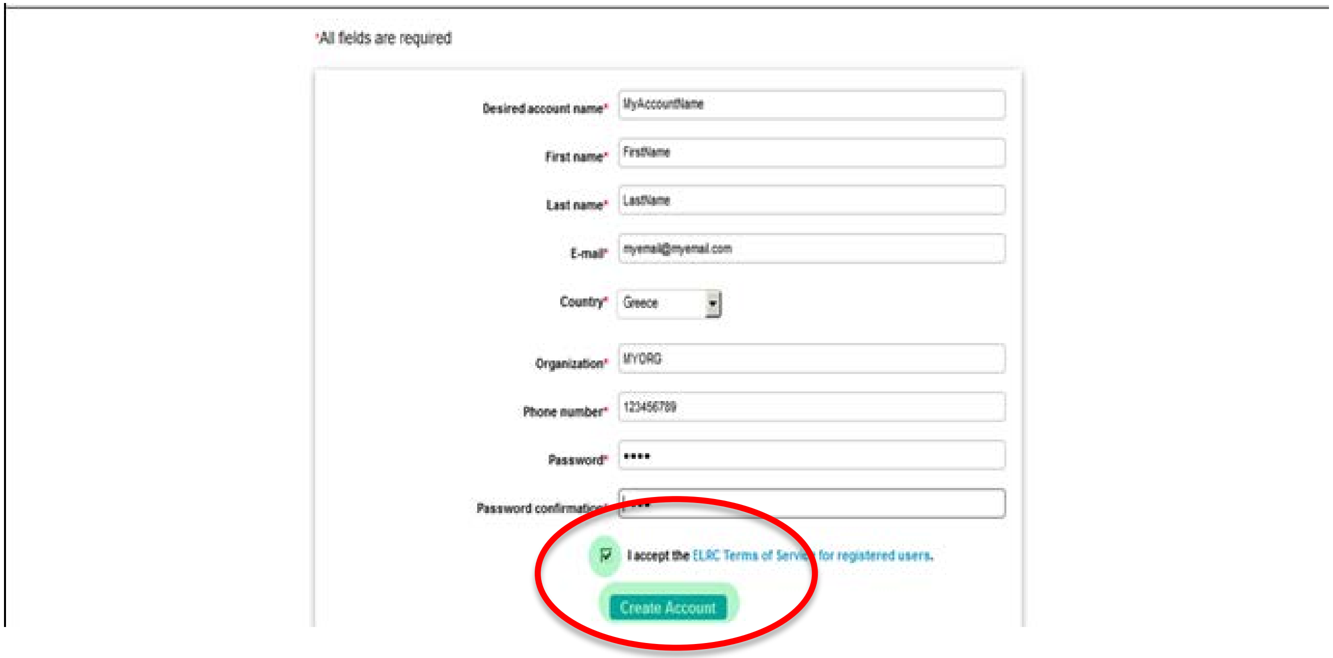
ELRC-SHARE Repository

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

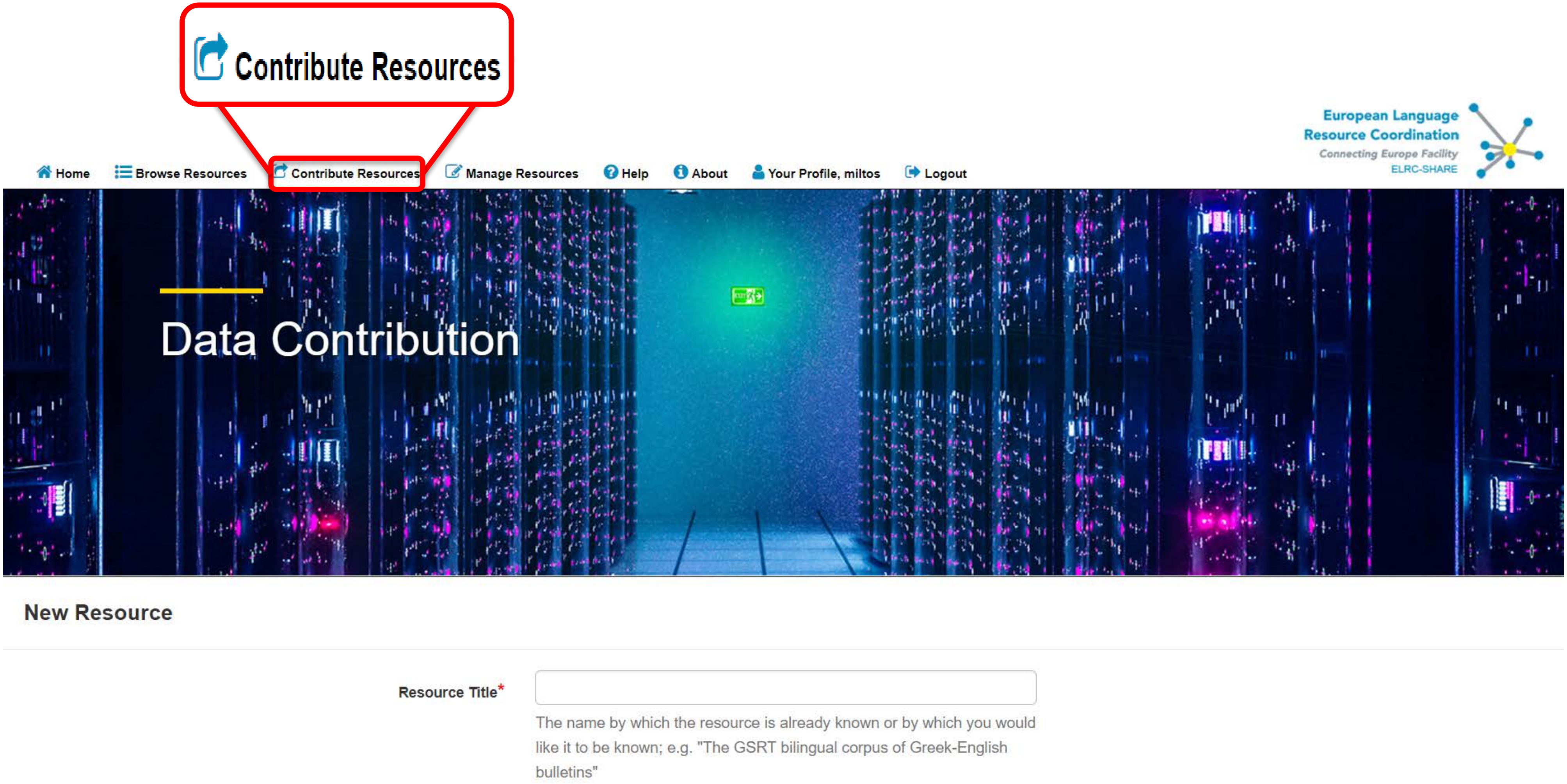


*All fields are required

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)



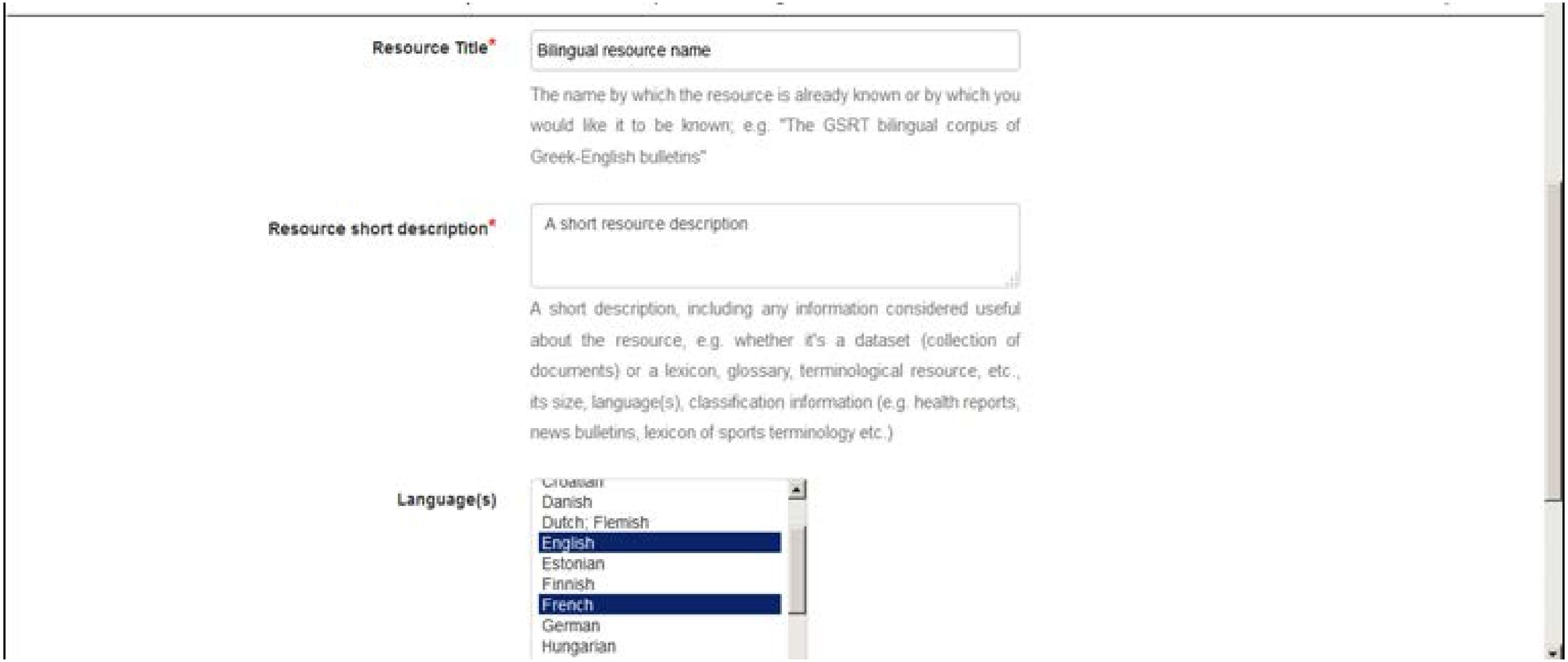
New Resource

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)

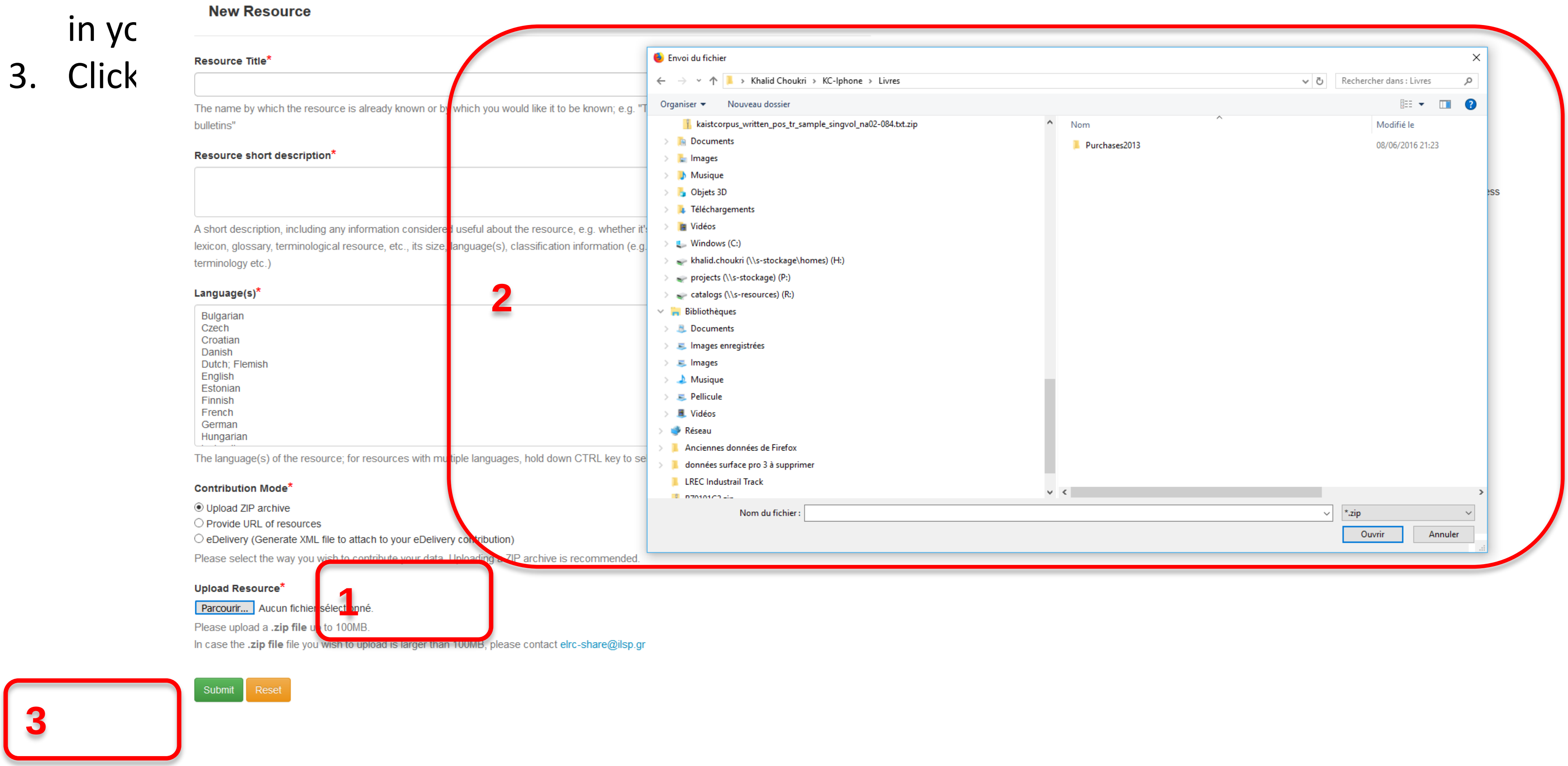
How to Contribute Data (3/6)

- Three modes for contributing your data

European Language Resource Coordination
Connecting Europe Facility

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 by which you would like it to be known; e.g. "T bulletins"

Resource short description*

A short description, including any information considered useful about the resource, e.g. whether it's 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 multiple 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 contribution)

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** up 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

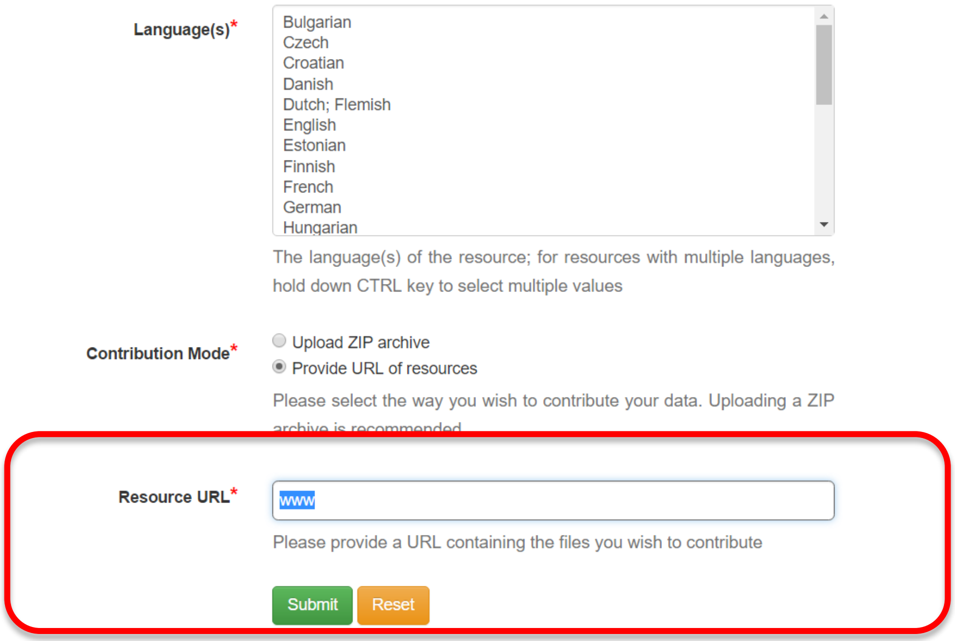
kaistcorpus_written_pos_tr_sample_singvol_na02-084.txt.zip
Documents
Images
Musique
Objets 3D
Téléchargements
Vidéos
Windows (C:)
khalid.choukri (\\s-stockage\homes) (H:)
projects (\\s-stockage) (P:)
catalogs (\\s-resources) (R:)
Bibliothèques
Documents
Images enregistrées
Images
Musique
Pellicule
Vidéos
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

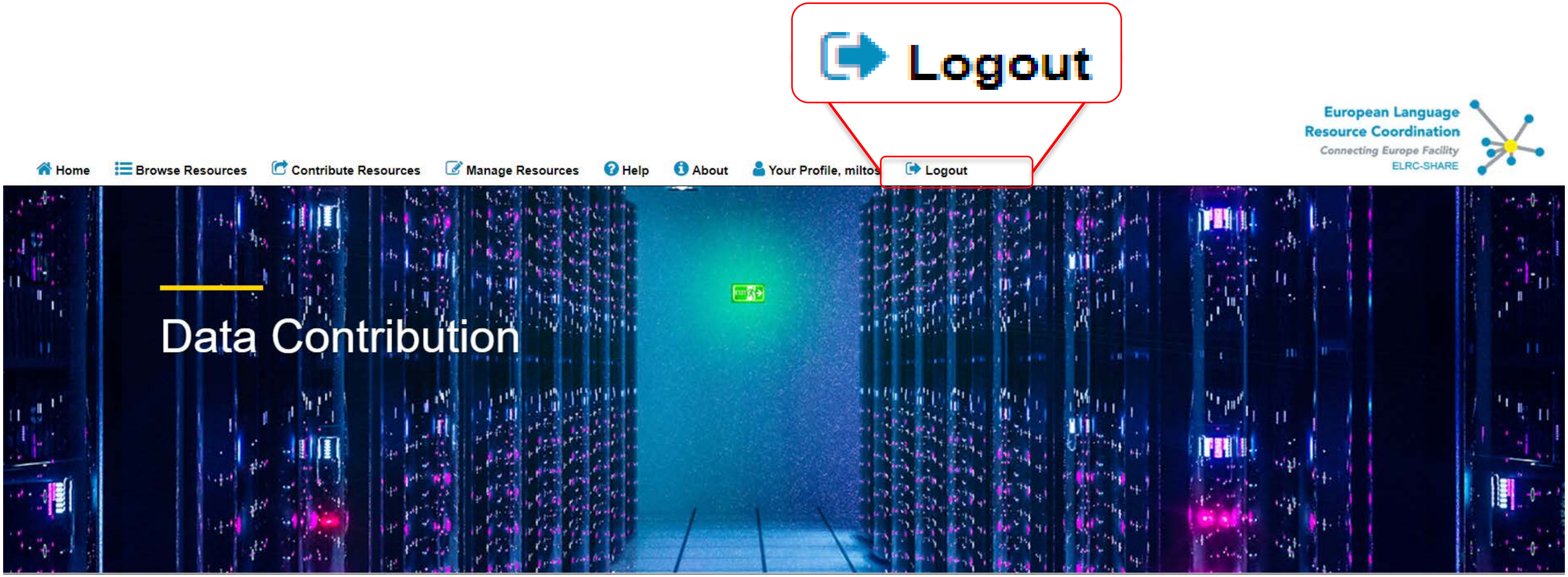
Please select the way you wish to contribute your data. Uploading a ZIP archive is recommended

Resource URL* www

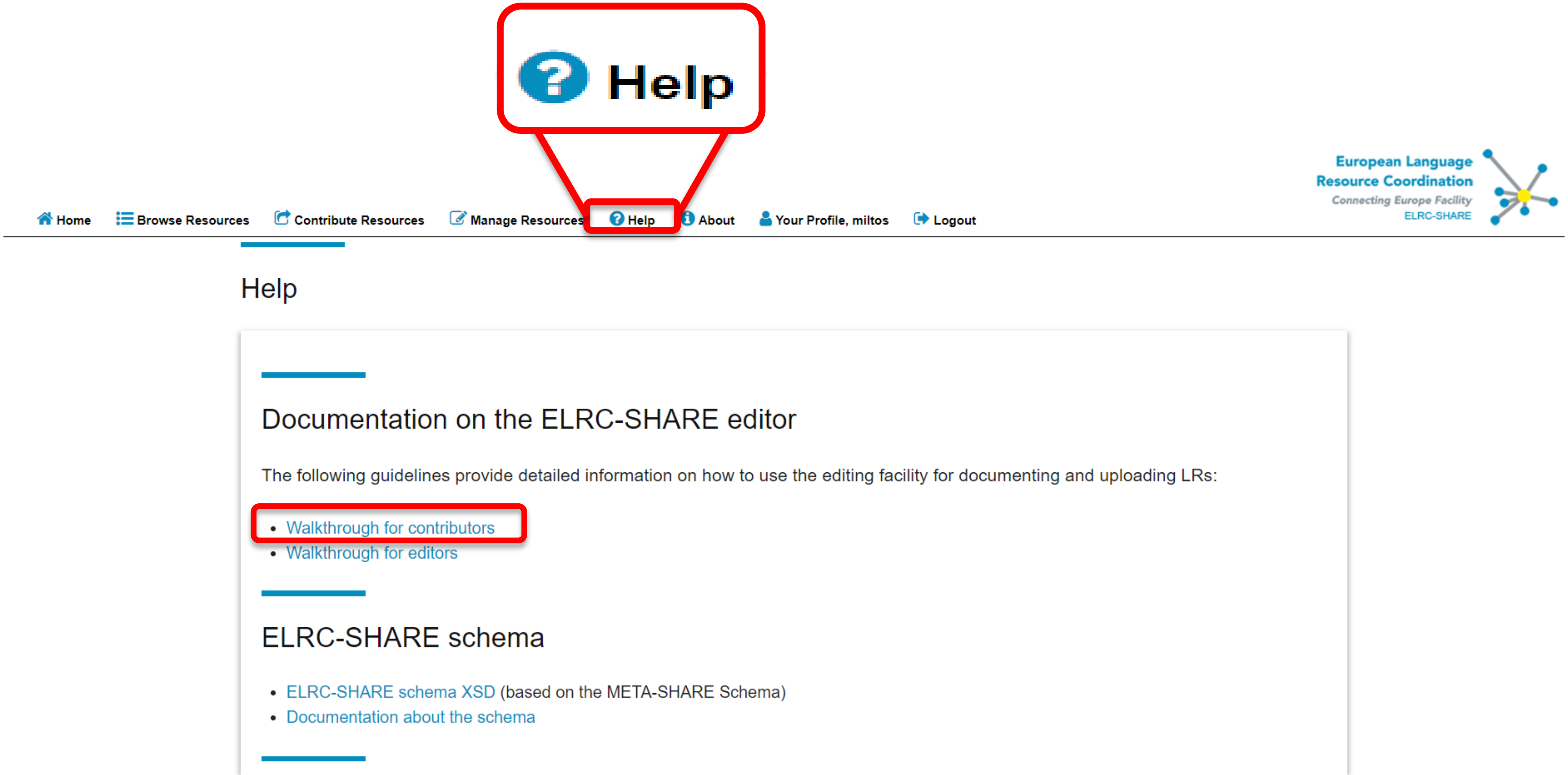
Please provide a URL containing the files you wish to contribute

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:

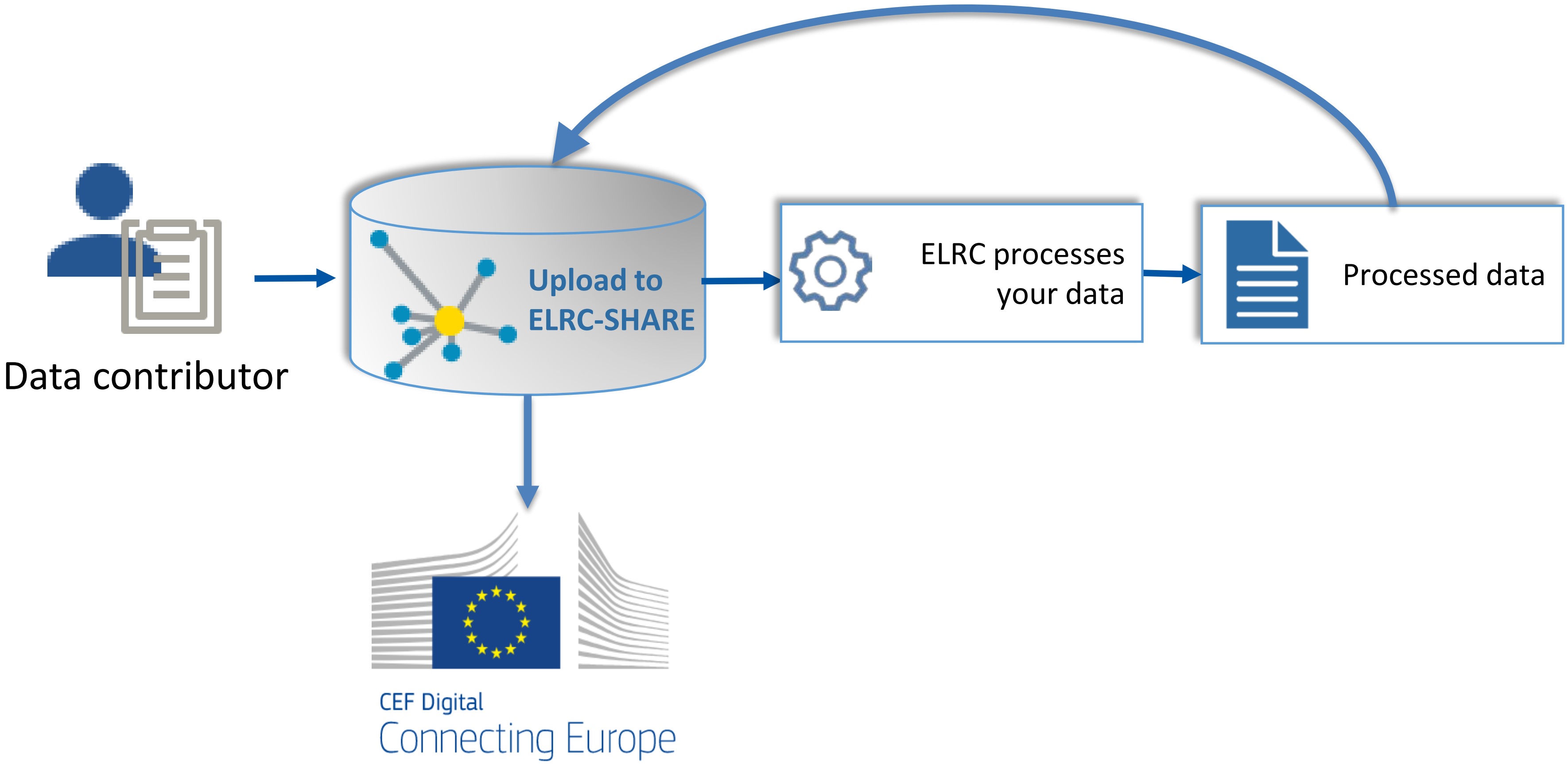
- Walkthrough for contributors
- Walkthrough for editors

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!

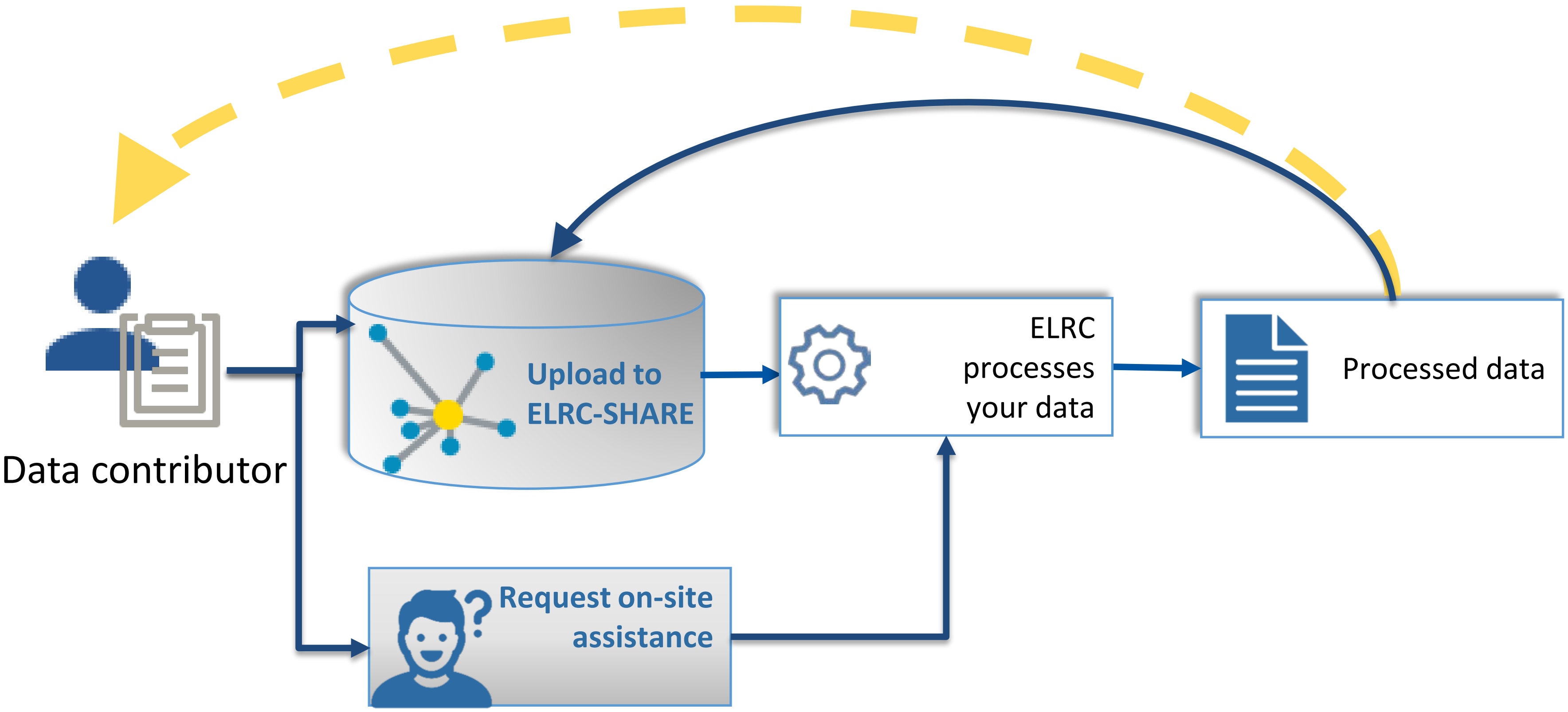
Alignment

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

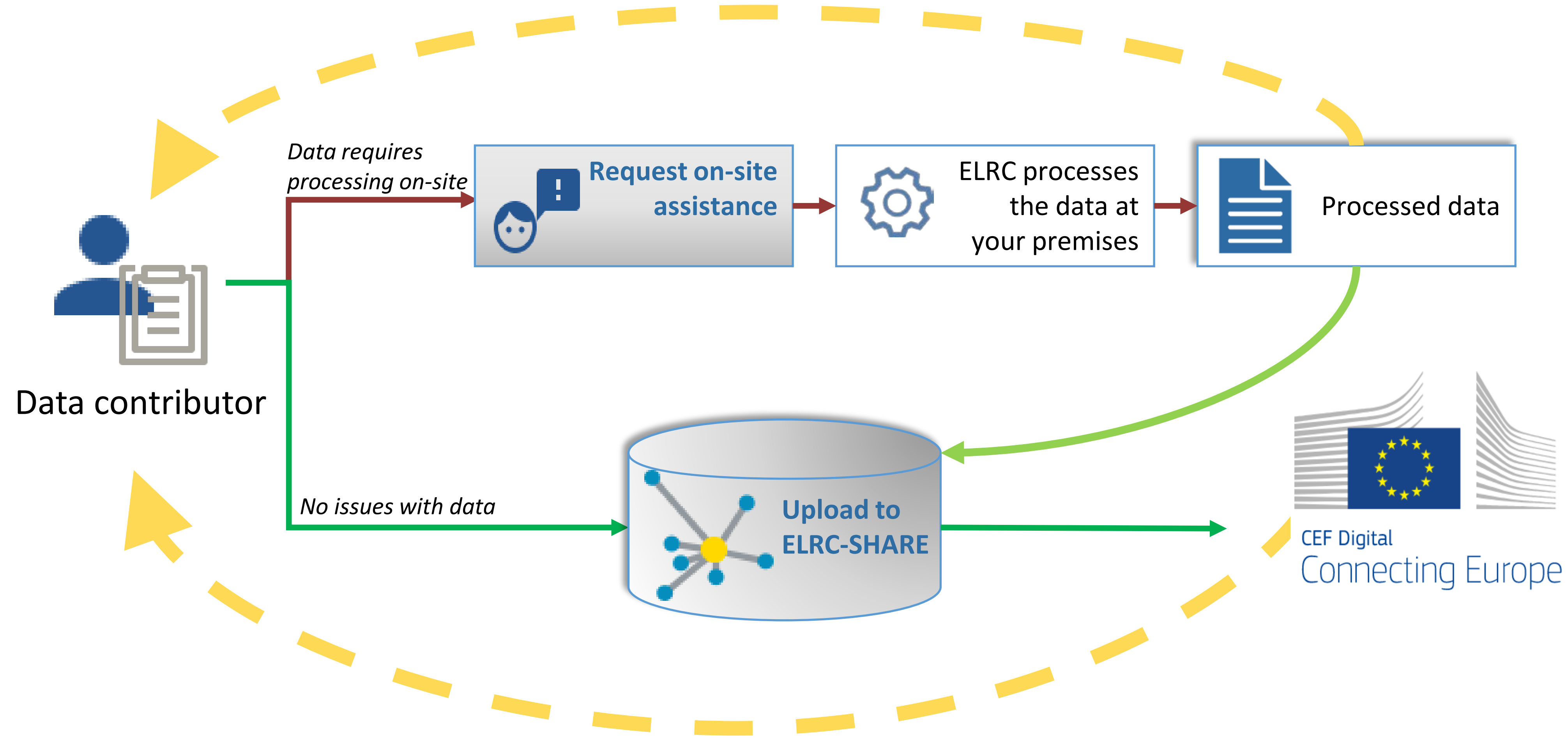
Metadata

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

What happens to your data?

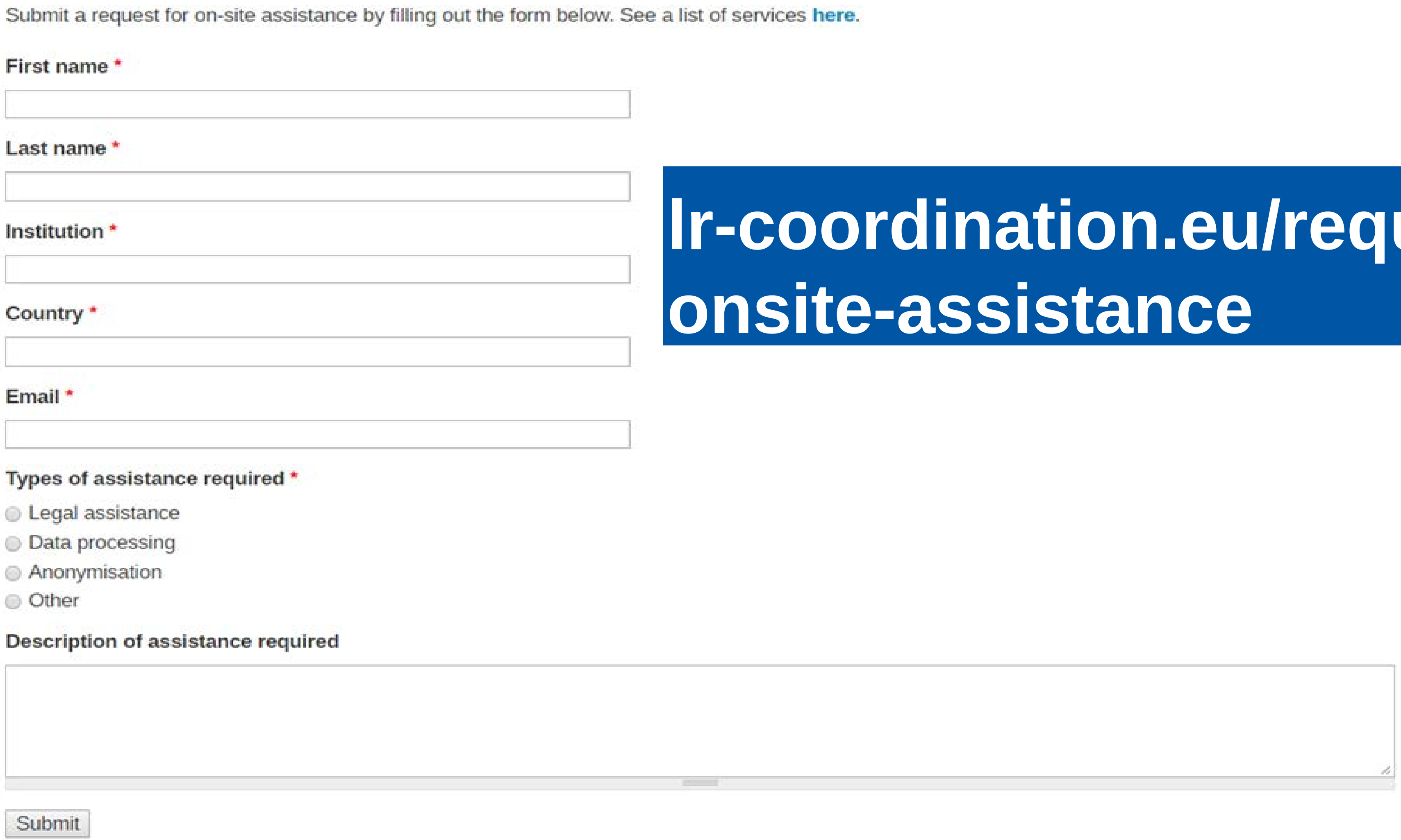


On-site assistance |2


Data contributor
Data requires processing on-site
Request on-site assistance
ELRC processes the data at your premises
Processed data
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 *

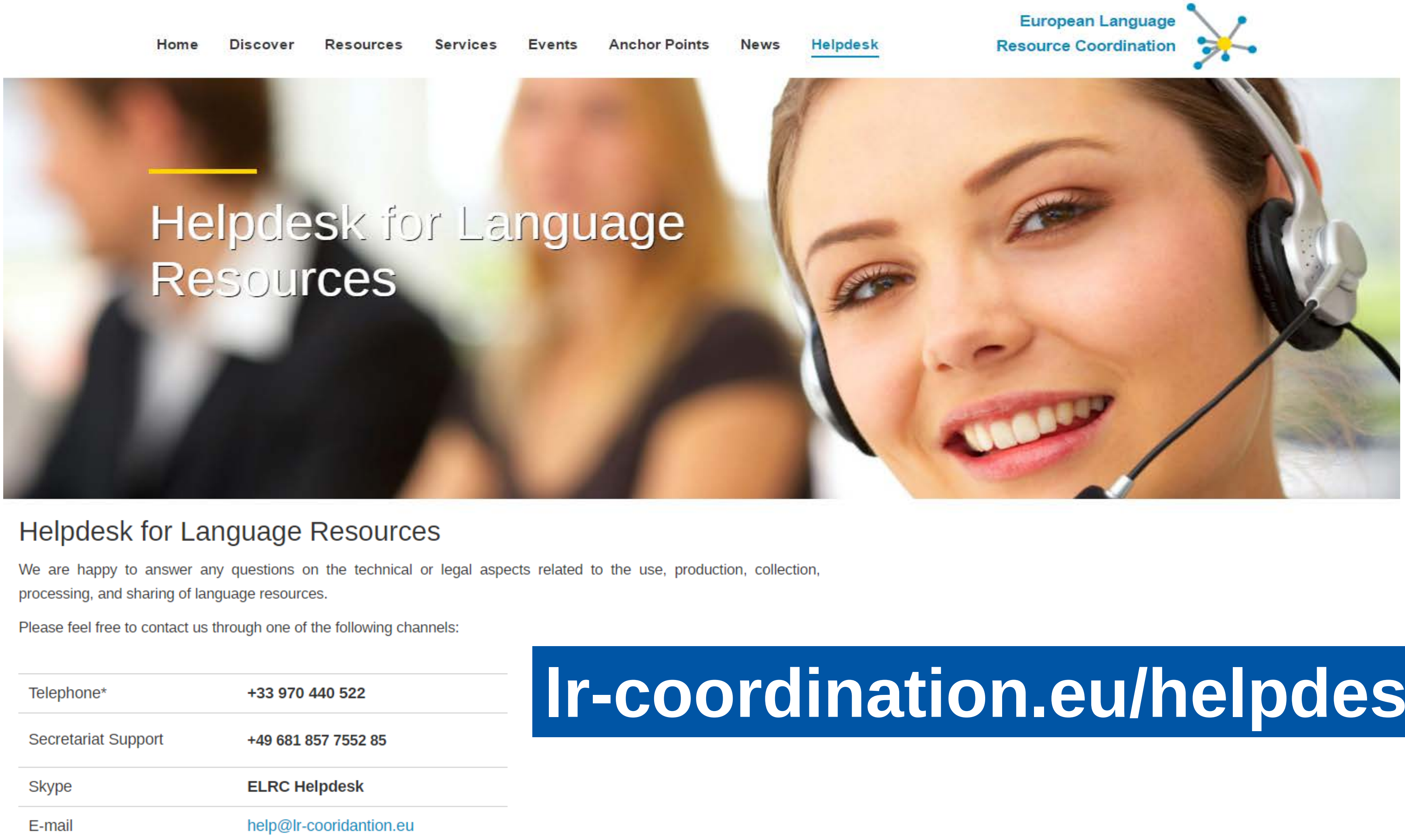
Types of assistance required *

○ Legal assistance
○ Data processing
○ Anonymisation
○ Other

Description of assistance required

lr-coordination.eu/request-onsite-assistance

ELRC Helpdesk



Helpdesk for Language Resources

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

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

Case studies (2015-2016)

Spain

Problem: Data provider didn’t store translations as related documents, therefore source/target translation weren’t paired

Solution: ELRC helped crawl a local system to find, related, and pair source/target translations

Spain

Problem: In some Spanish governmental departments, archives were only available in PDF

Solution: ELRC helped provide good converters to get usable documents

Germany

Problem: Data owner needed help with anonymization, as databases contained personal info. Another need: cleaning up ‘junk’ data (URLs, numbers, fragments)

Solution: ELRC helped provide anonymization services and data cleaning

Estonia

Problem: Data donor found that legal acts in EN, ET, RU couldn’t be aligned on a document level (no common machine-readable cross-language ID)

Solution: ELRC helped provide alignment services for documents