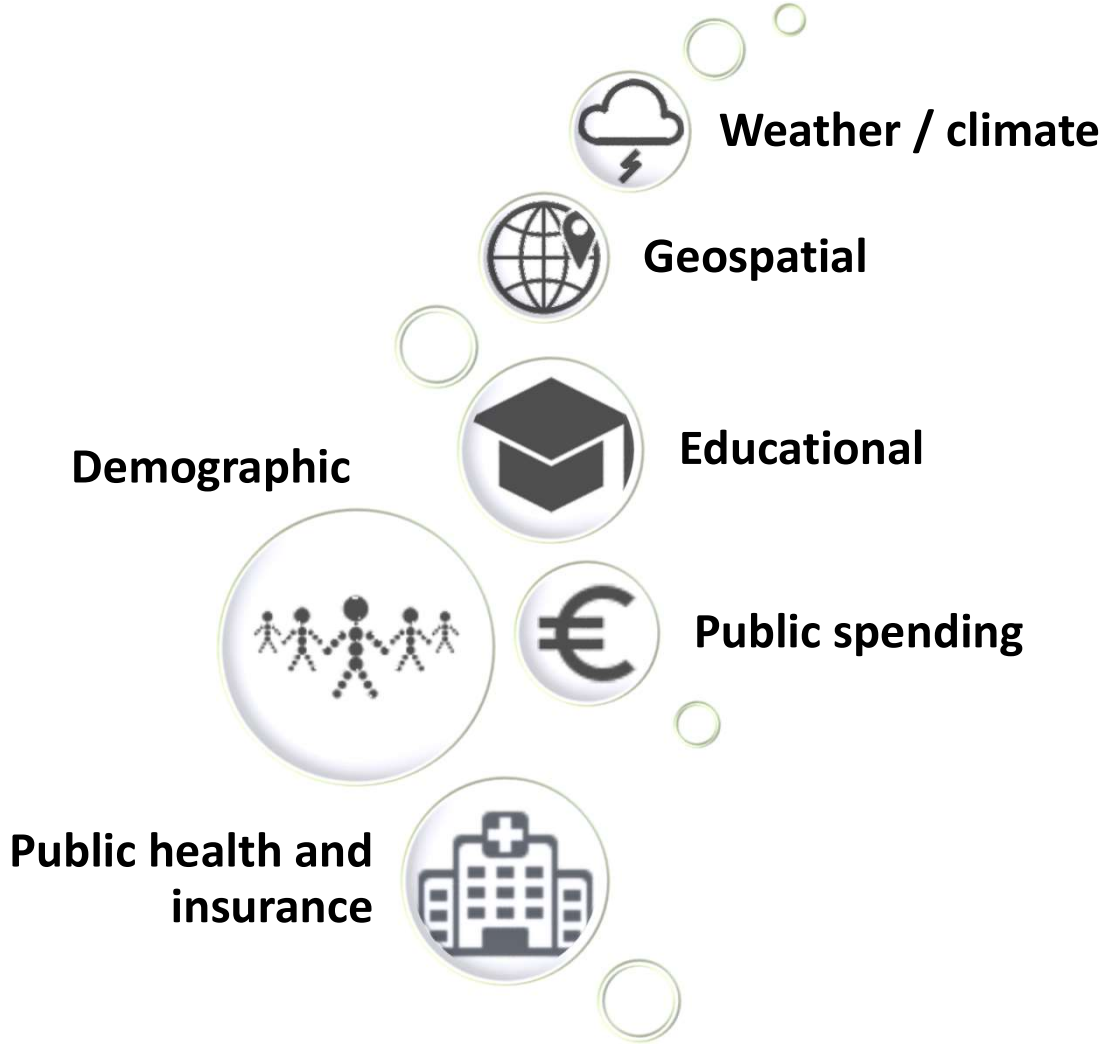
European Language Resource Coordination
Connecting Europe Facility

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

European Language Resource Coordination
Connecting Europe Facility

The notion of data

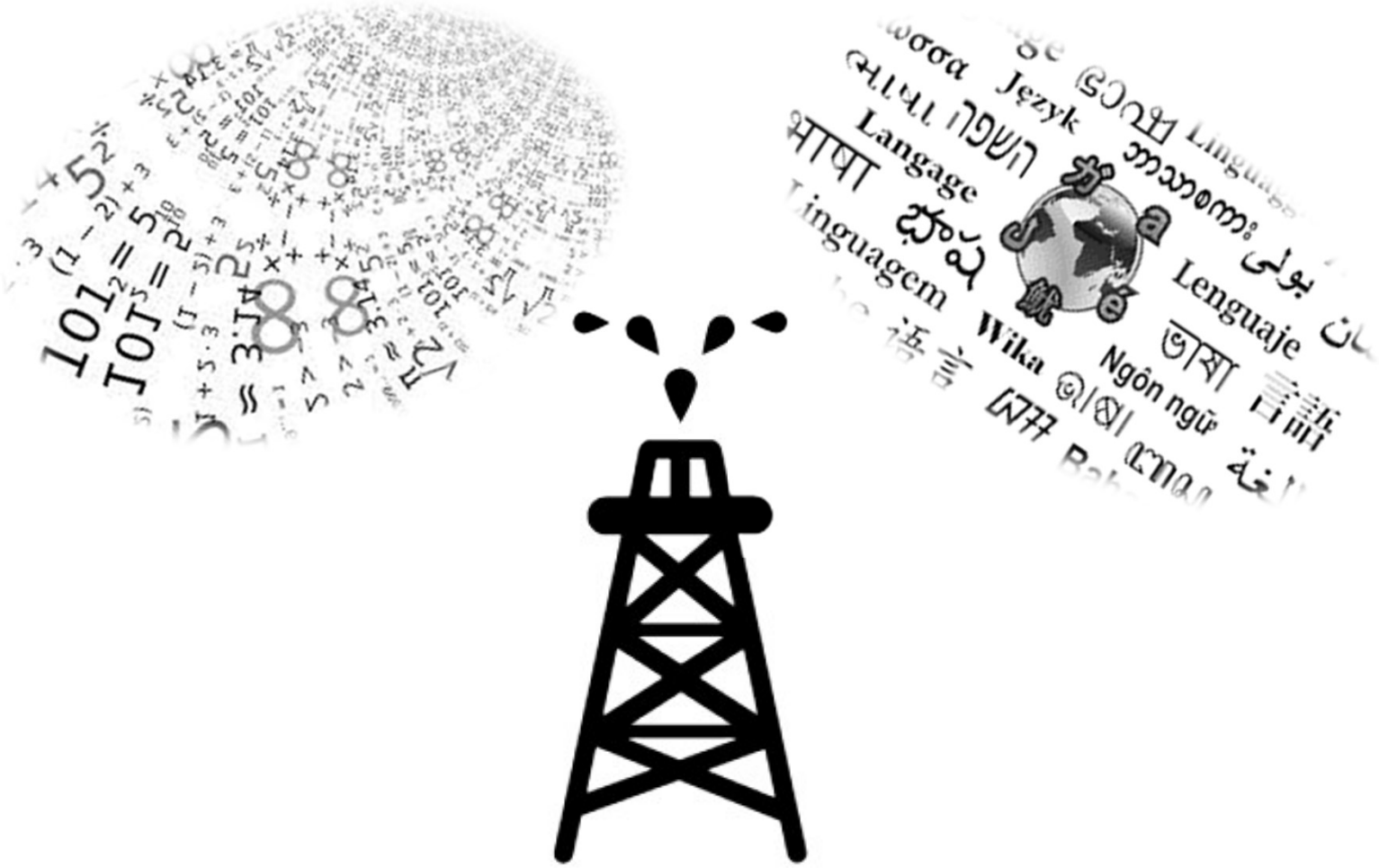


The notion of data

Data: the oil of the 21st century

European Language Resource Coordination
Connecting Europe Facility

The notion of data



European Language Resource Coordination
Connecting Europe Facility

The notion of language data

Data

- any piece of electronically stored content

(Textual) Language Data

- any piece of electronically stored text

European Language Resource Coordination
Connecting Europe Facility

Language data for eTranslation

DE

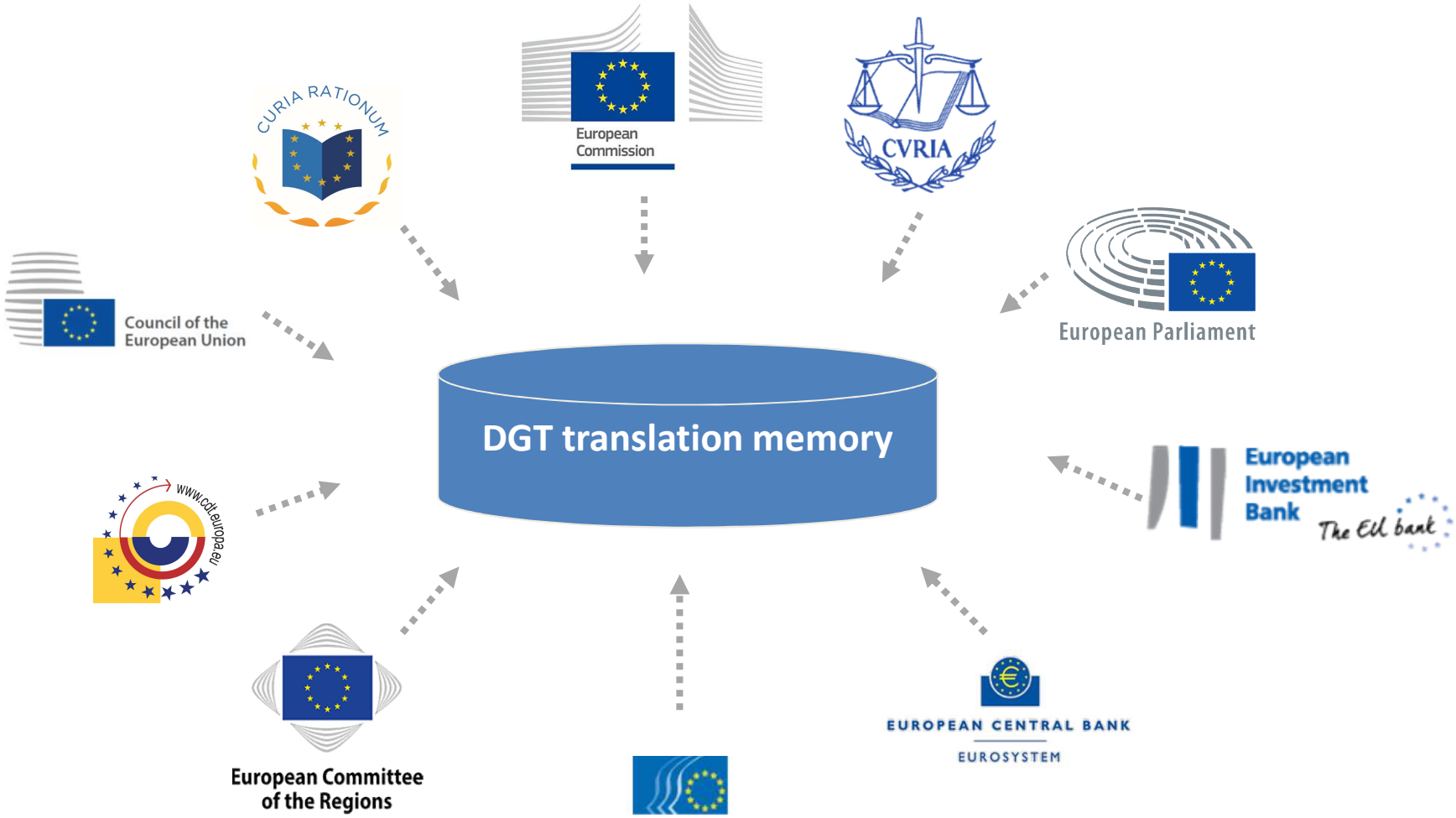
Dies gilt nicht, wenn die genannten Personen auf Grund dieser Tätigkeit den Bestimmungen der gesetzlichen Pensionsversicherung oder Bestimmungen, die in ihren Grundsätzen jenen der gesetzlichen Pensionsversicherung entsprechen, unterliegen.

EN

This shall not apply if the aforementioned persons are – on account of the respective activity carried out by them – subject to the provisions of the statutory pension insurance or other provisions which in principle conform to those of the statutory pension insurance.

European Language Resource Coordination
Connecting Europe Facility

Data used by eTranslation



Such data are already available
BUT
they are not enough…

European Language Resource Coordination
Connecting Europe Facility

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)

German text	Translation in English
Beschluß der Provisorischen Nationalversammlung vom 30. Oktober 1918. Präambel/Promulgationsklausel 1.Jede Zensur ist als dem Grundrecht der Staatsbürger widersprechend als rechtsungültig aufgehoben. 2.Die Einstellung von Druckschriften und die Erlassung eines Postverbotes gegen solche findet nicht mehr statt. Auf Grund des § 7 des Beschlusses der Provisorischen Nationalversammlung vom 30. Oktober 1918 über die grundlegenden Einrichtungen der Staatsgewalt wird beurkundet, daß der obenstehende Beschluß von der Provisorischen Nationalversammlung am 30. Oktober 1918 gefaßt worden ist.	Resolution of the Provisional National Assembly of 30 October 1918. Preamble/Clause of Promulgation 1.All censorship is abolished as illegal because contradictory to the basic rights of the citizen. 2.Stops on publications and the issue of a postal distribution veto on such cease forthwith. Pursuant to Para. 7 of the Resolution of the National Assembly of 30 October 1918 about the fundamental establishment of the government is documented that the above-mentioned resolution has been taken by the Provisional National Assembly on 30 October 1918.

European Language Resource Coordination
Connecting Europe Facility

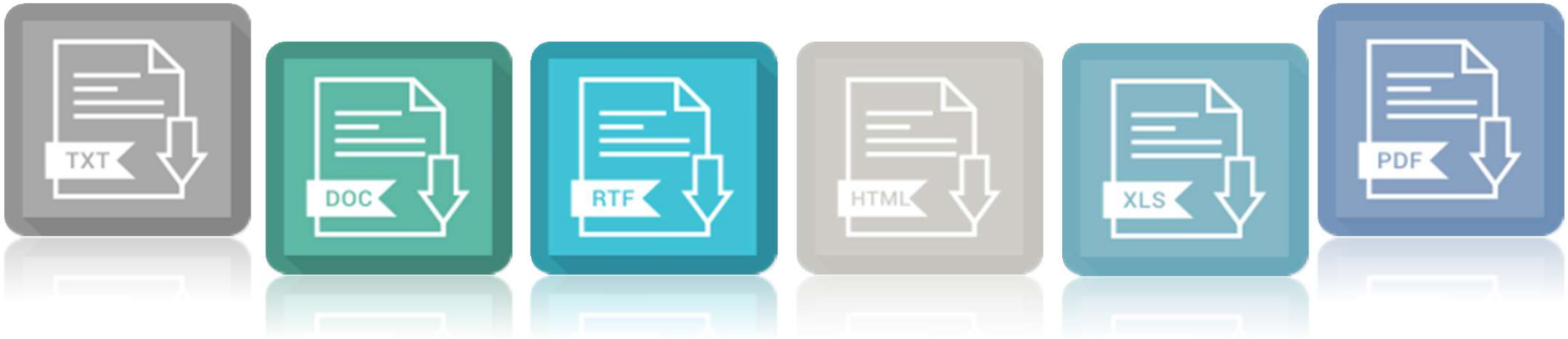
What data are useful for eTranslation as per type |2

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

German	English
abfragen (Datenbank)	to query (a database)
abgängig *(z.B. eine Person ist abgängig)*	missing *(e.g. a person is missing)*
AbgängigeR (VermissteR)	missing person
Abhängigkeit von Suchtgiften und psychotropen Substanzen	addiction to narcotic drugs and psychotropic substances
ablaufen *(z.B. das Dokument läuft am … ab; abgelaufenes Reisedokument)* (vgl. gültig sein)	to expire *(e.g. the document expires on …; document that has expired)*
abmelden (vgl. anmelden, ummelden – Kfz)	to deregister (a vehicle)
sich abmelden (vgl. sich anmelden – MeldeG)	to deregister
sich (von einem Computersystem) abmelden (vgl. sich anmelden)	to log off
Abflug	departure
Absatz *(z.B. § 3 Abs 2 BDG)*	paragraph *(e.g. sec 3 para 2, Civil Servants Act)*
…	…

European Language Resource Coordination
Connecting Europe Facility

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

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

European Language Resource Coordination
Connecting Europe Facility

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

European Language Resource Coordination
Connecting Europe Facility

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.

European Language Resource Coordination
Connecting Europe Facility

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.·¶

¶

¶

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

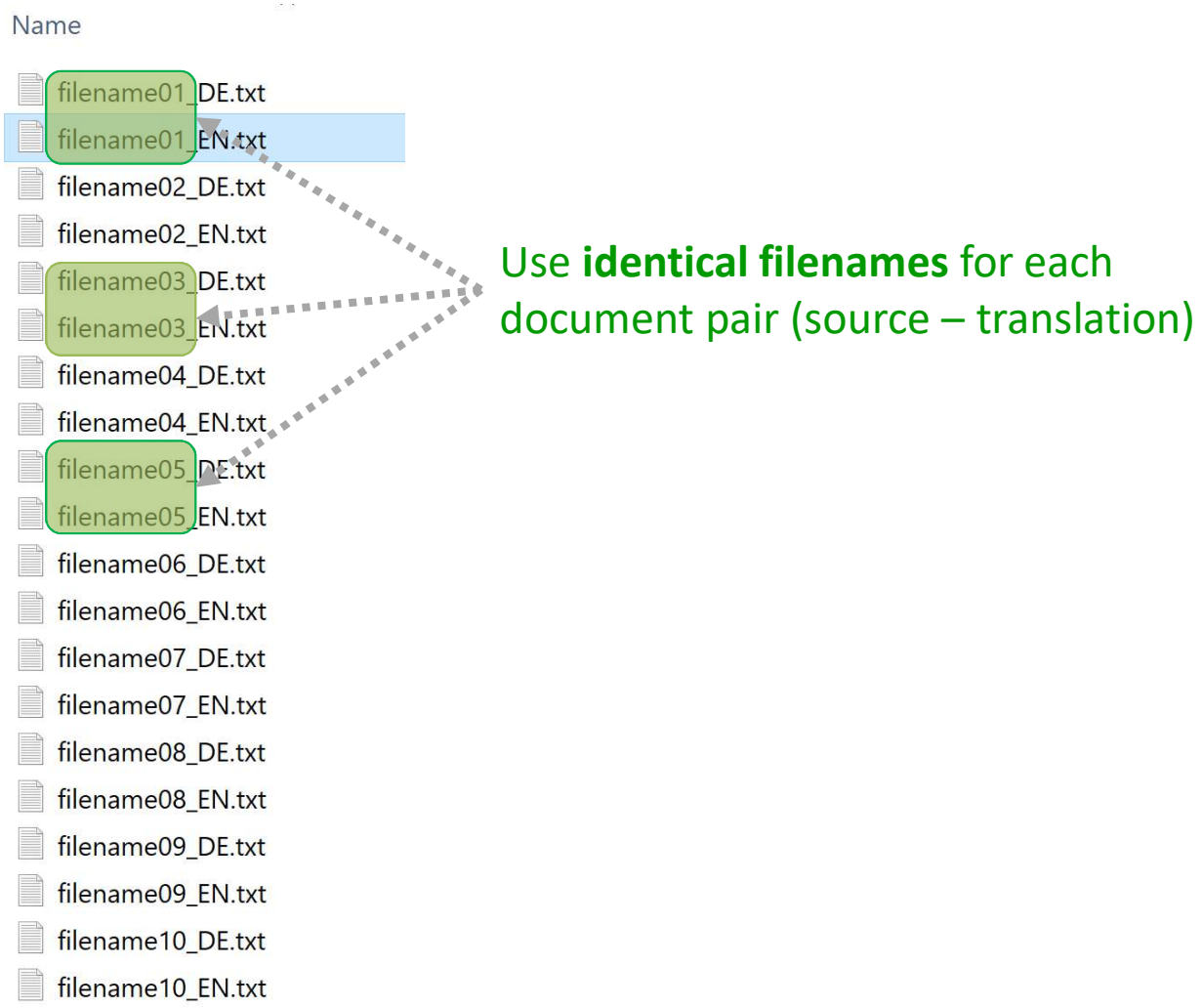
Don'ts

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.	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.

(Ctrl)

Preparing your data |4

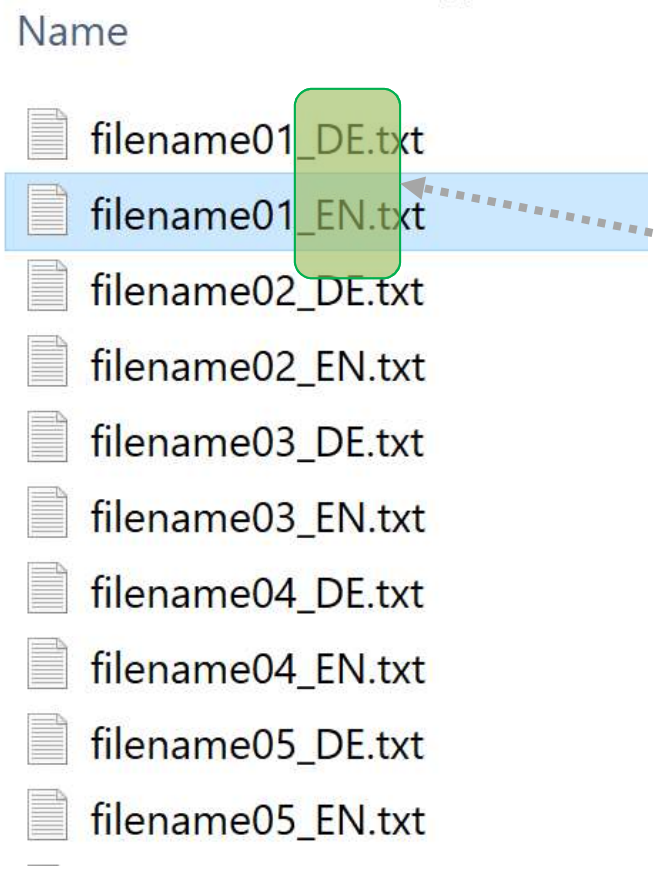
Do's



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

Preparing your data |5

Do's



Include **language identifiers** in the filename

European Language Resource Coordination
Connecting Europe Facility

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*

European Language Resource Coordination
Connecting Europe Facility

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

European Language Resource Coordination
Connecting Europe Facility

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

European Language Resource Coordination
Connecting Europe Facility

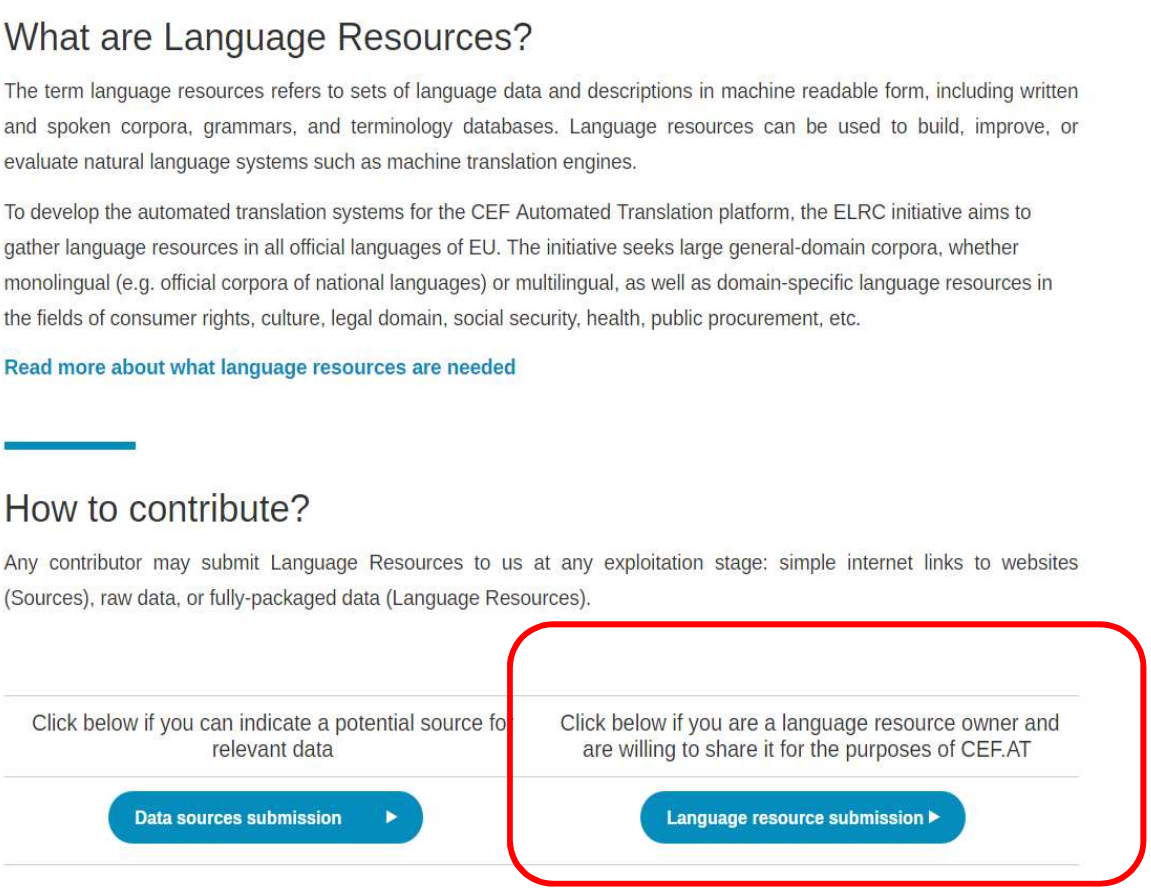
To ELRC-SHARE

- At the ELRC portal click on the “Language resource submission” button

Or

- Type in the url address:

elrc-share.eu

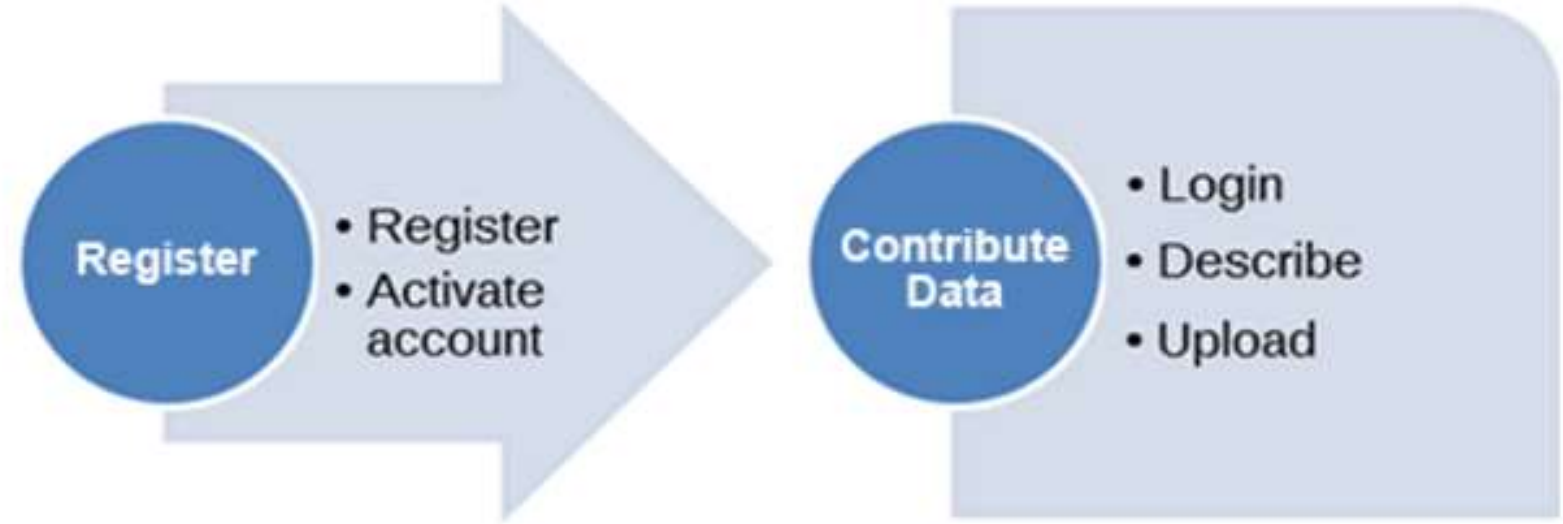


European Language Resource Coordination
Connecting Europe Facility

ELRC-SHARE repository

European Language Resource Coordination
Connecting Europe Facility

How to Contribute Data

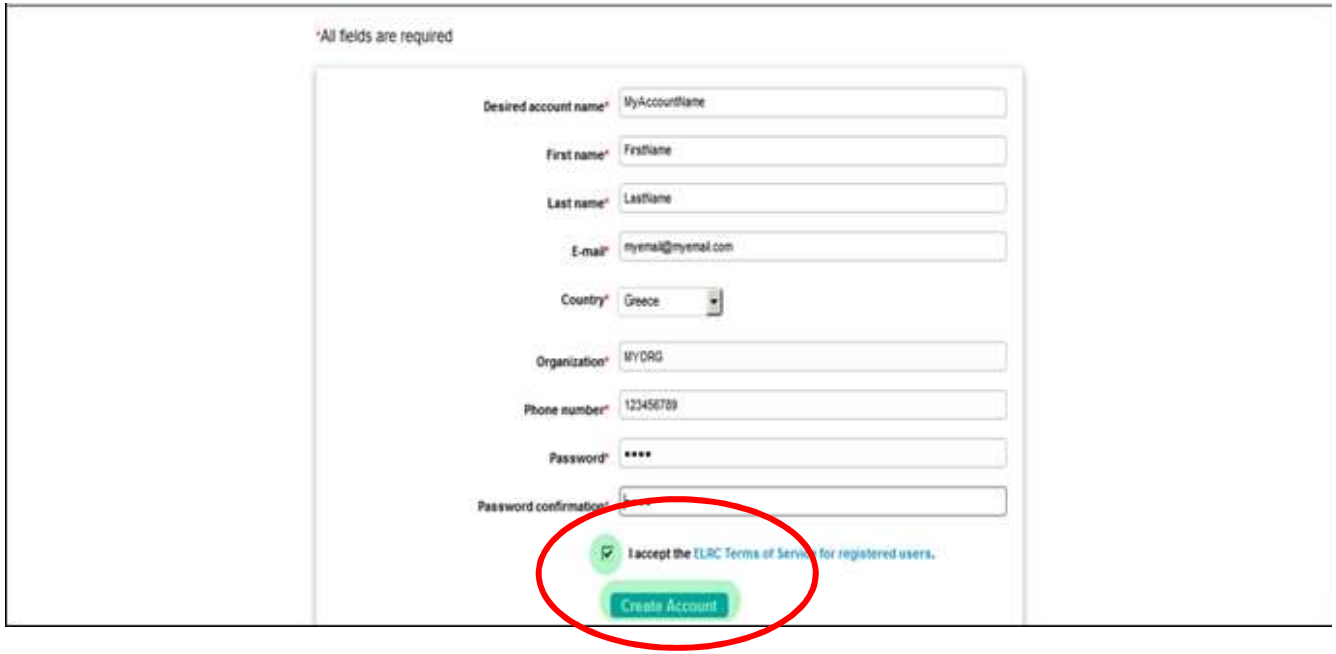


How to Register (1/2)

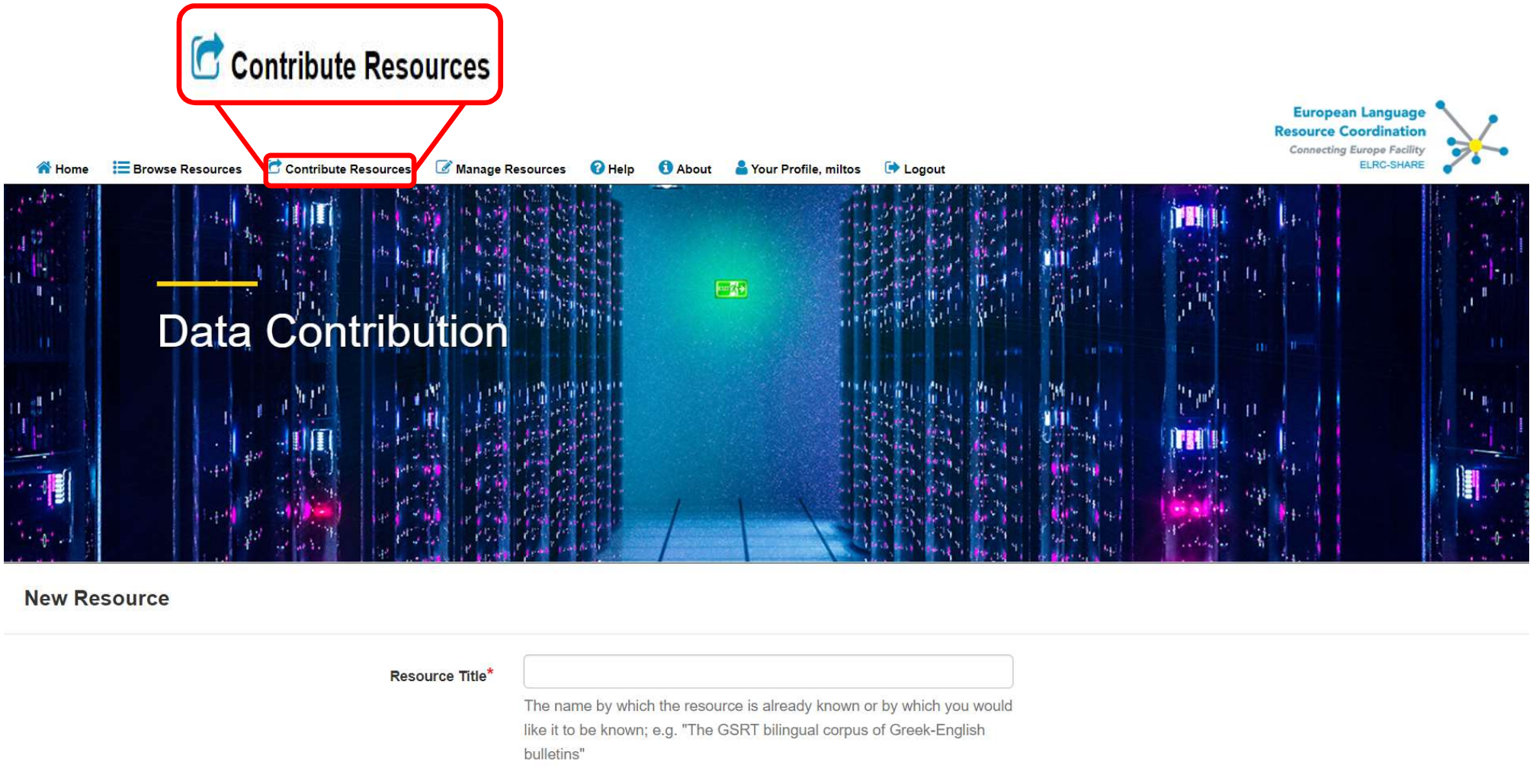
European Language Resource Coordination
Connecting Europe Facility

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



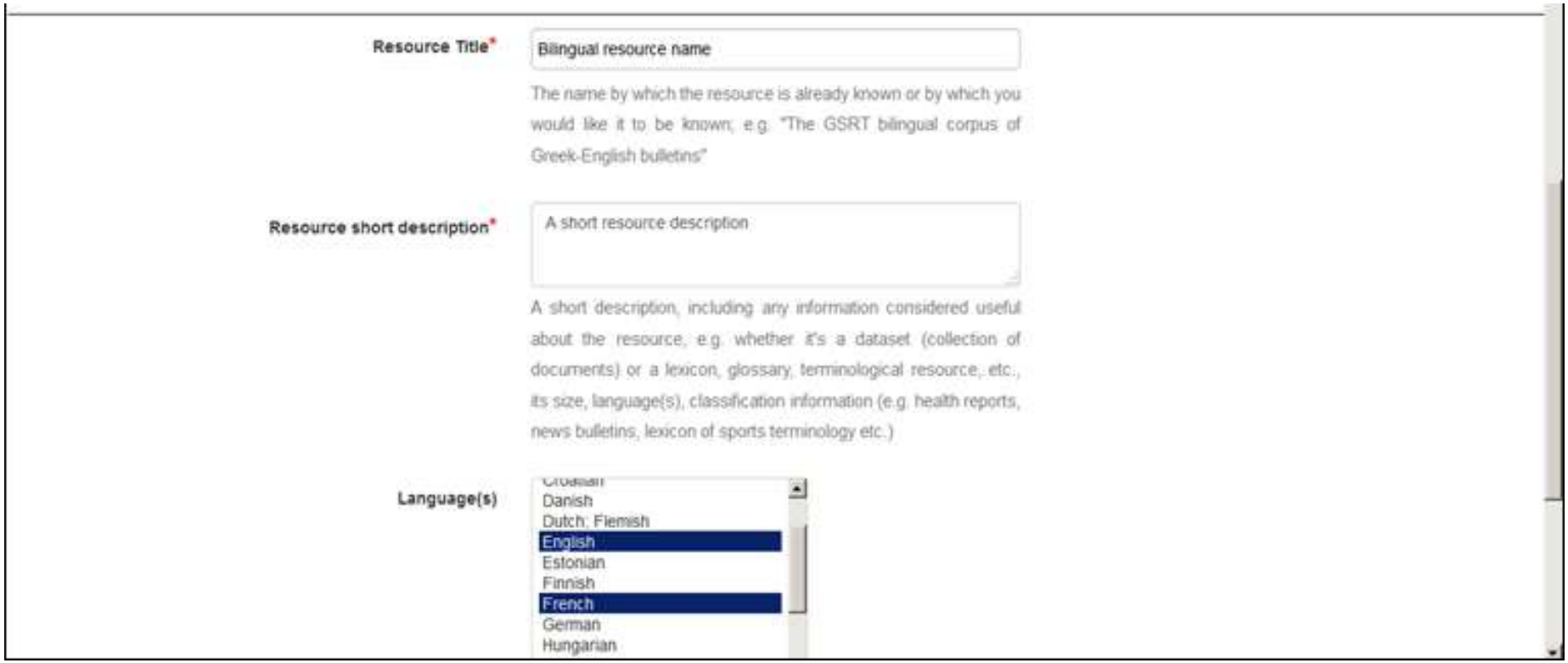
How to Contribute Data (1/6)



European Language Resource Coordination
Connecting Europe Facility

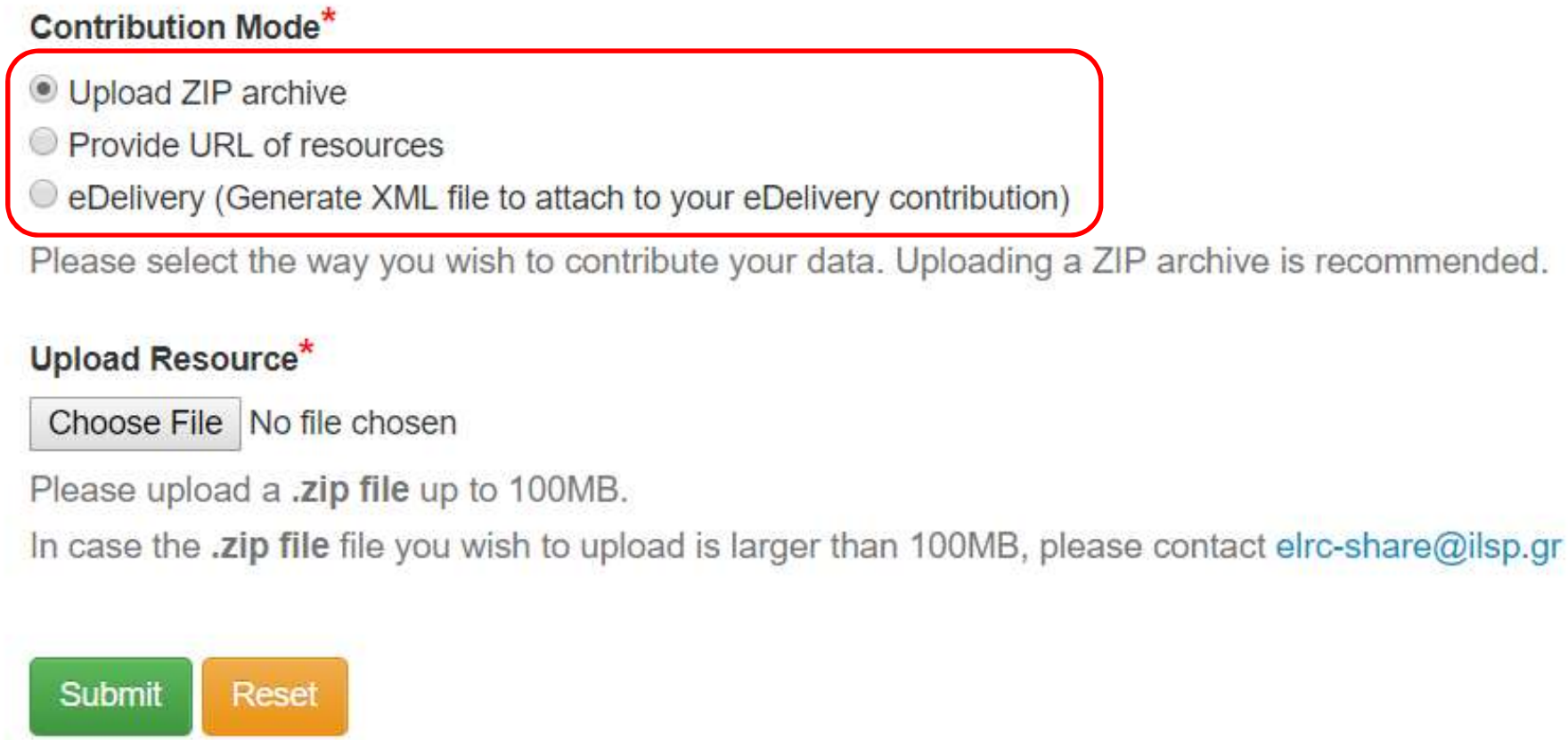
How to Contribute Data (2/6)

- Fill in the details of the dataset



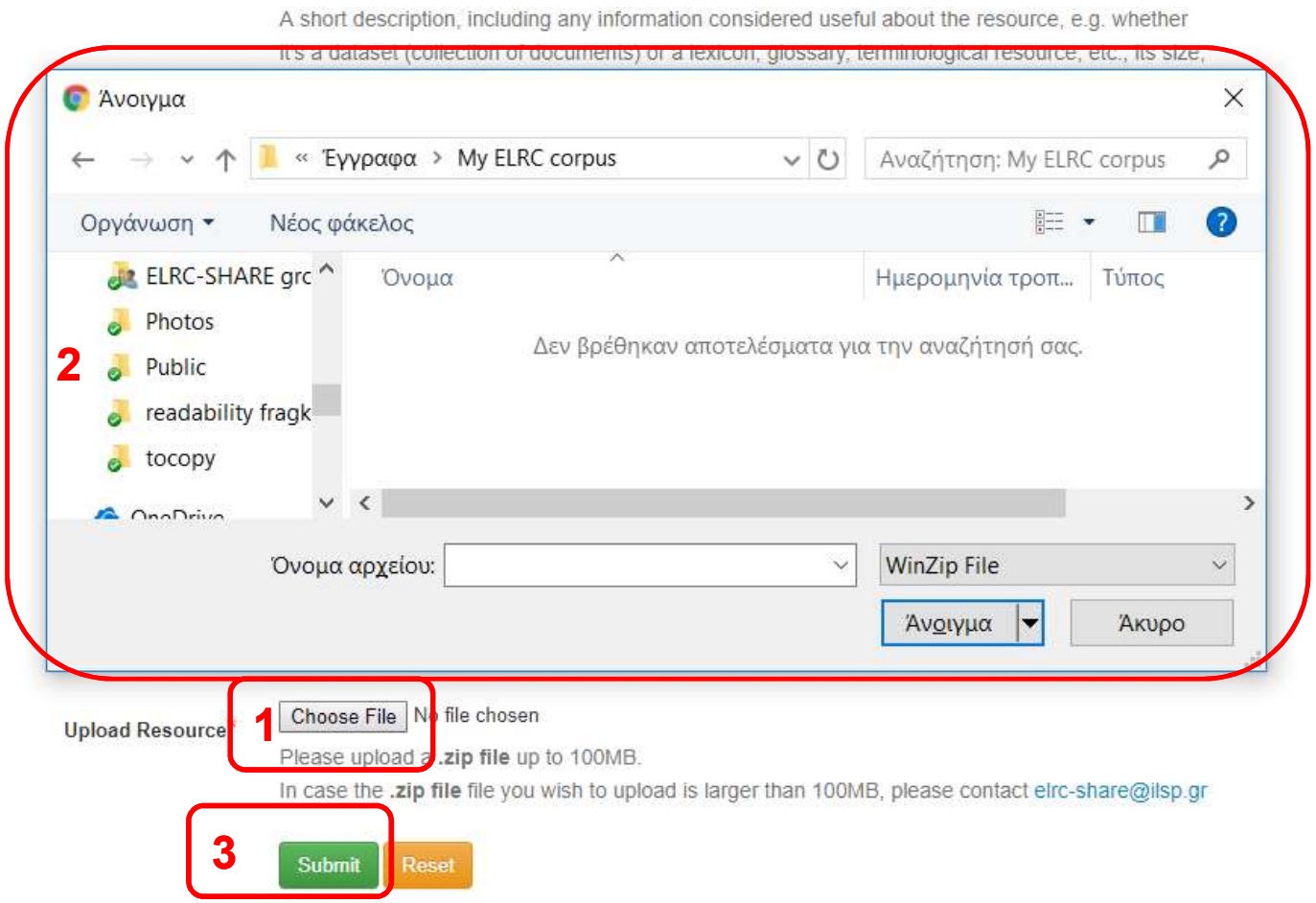
How to Contribute Data (3/6)

- Three modes for contributing your data



How to Contribute Data (4/6)

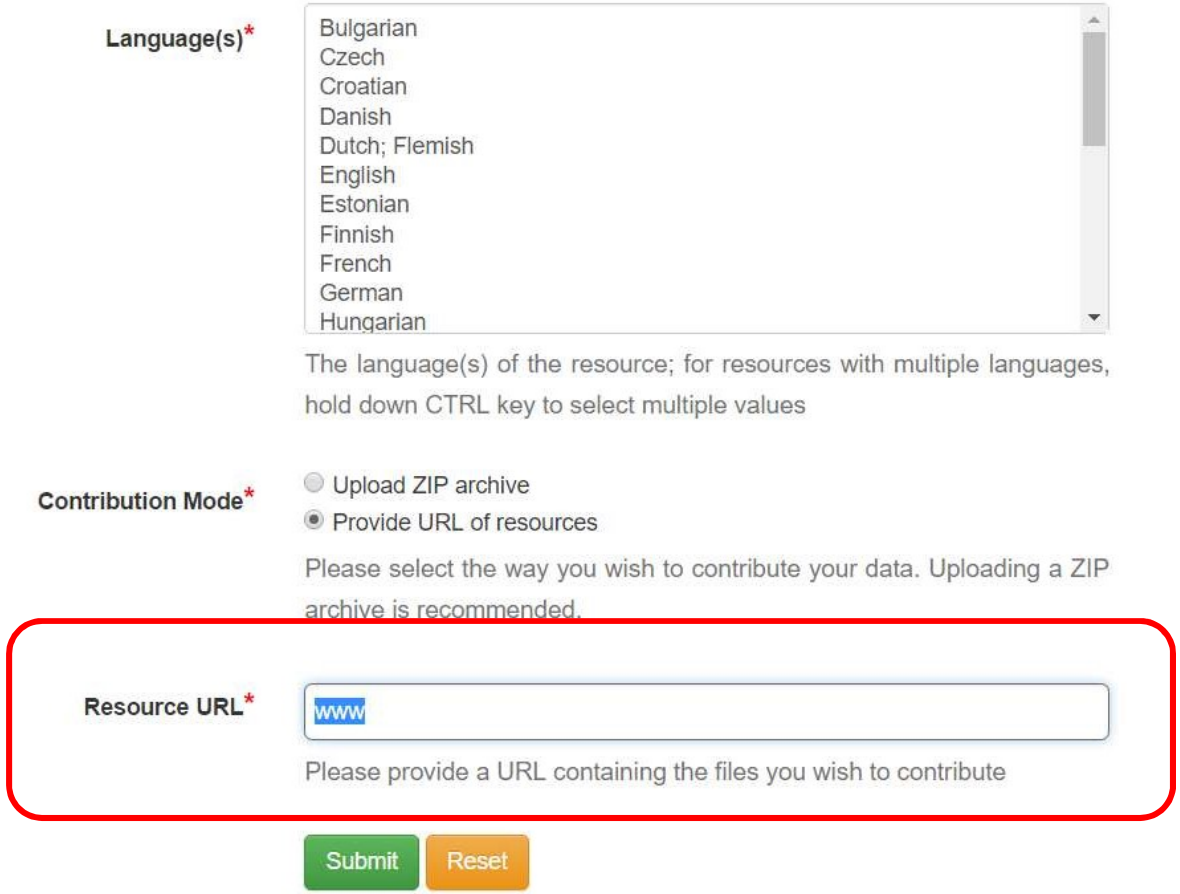
1. Click on Choose file
2. Locate your resource in your hard disk
3. Click on Submit



European Language Resource Coordination
Connecting Europe Facility

How to Contribute Data (5/6)

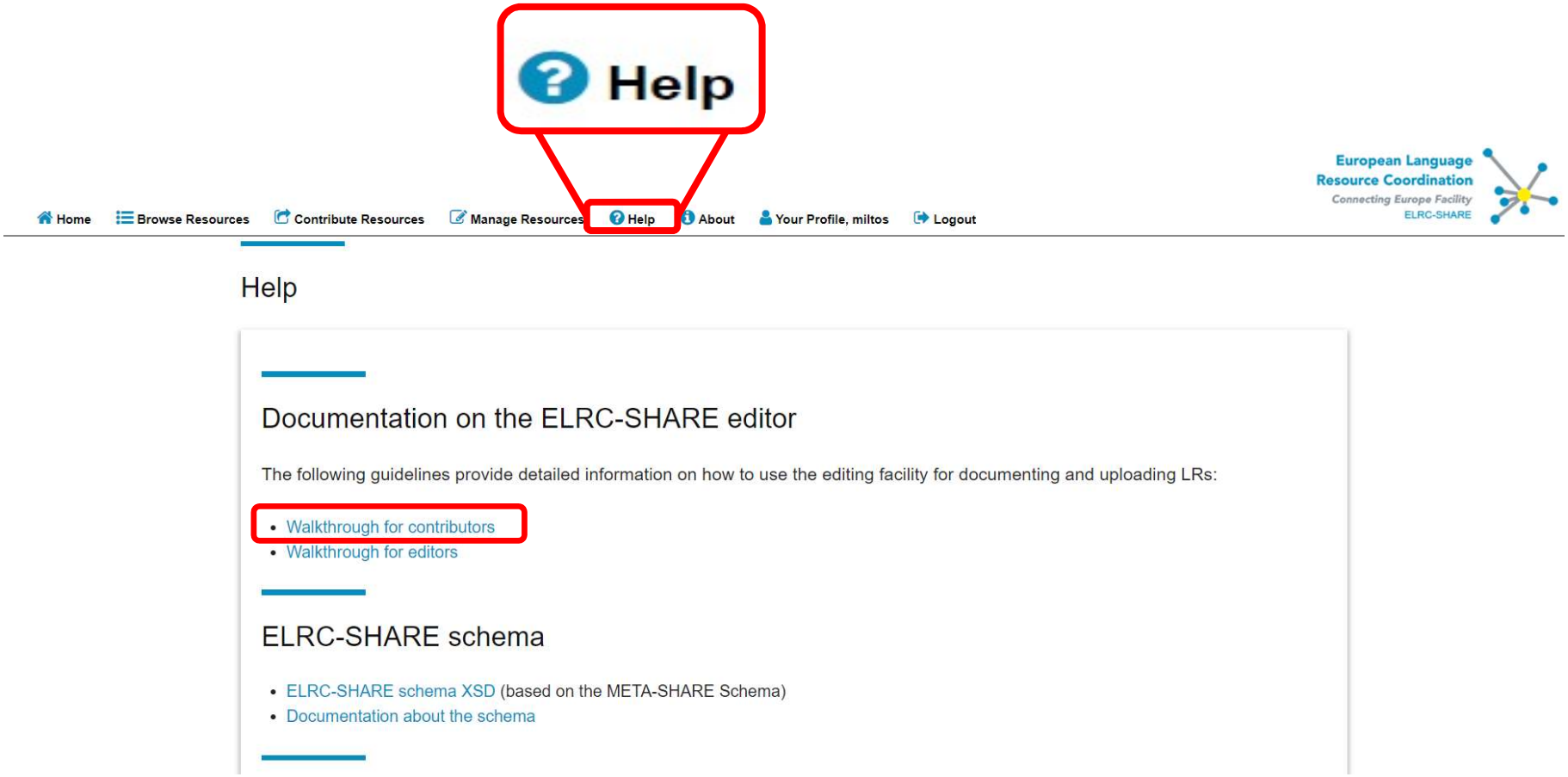
- Alternatively indicate a url (directory listing)



How to Contribute Data (6/6)

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

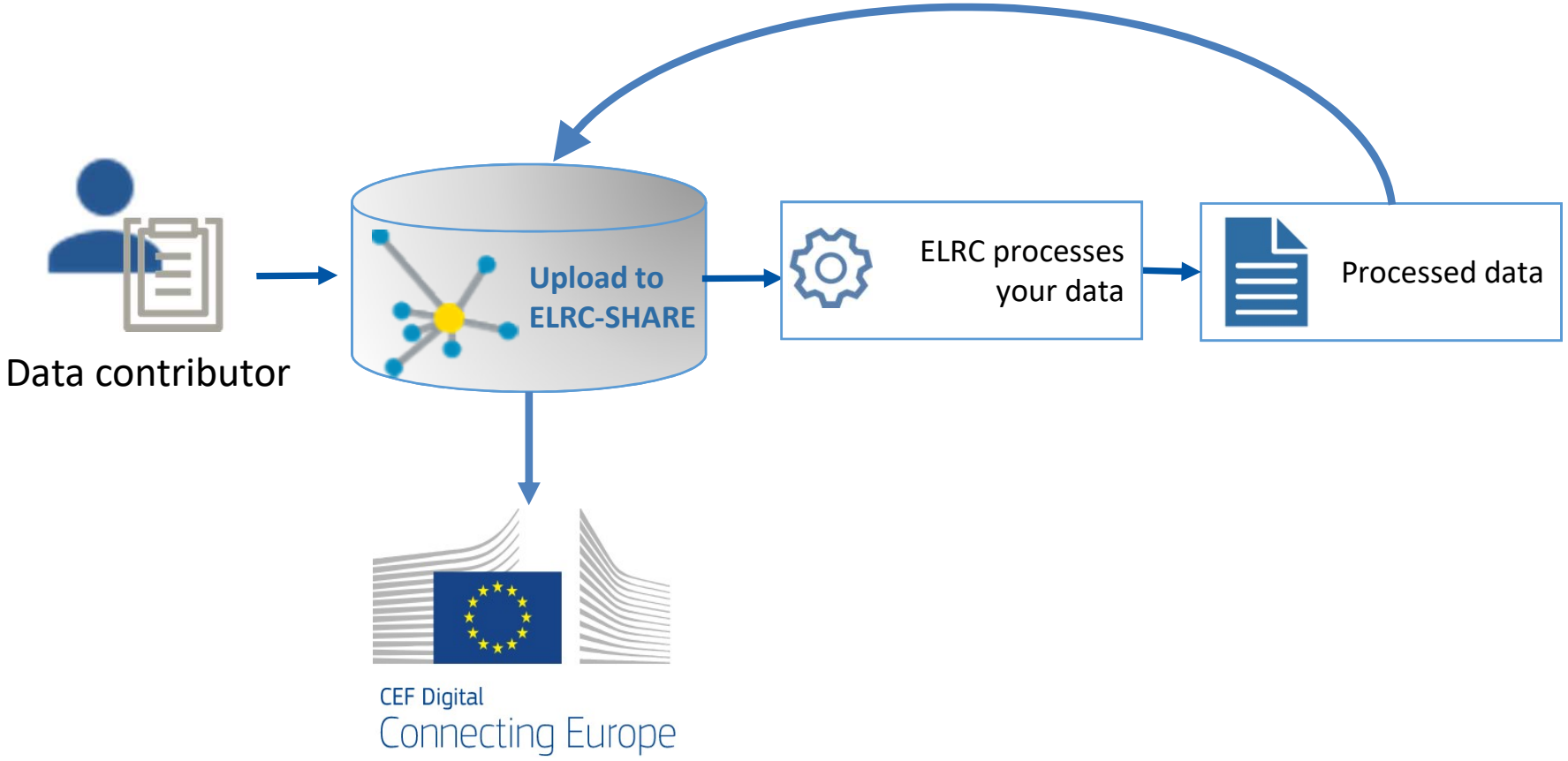
Guidelines for contributors



What happens next?

European Language Resource Coordination
Connecting Europe Facility

What happens to your data?



European Language Resource Coordination
Connecting Europe Facility

Language processing services

Data extraction

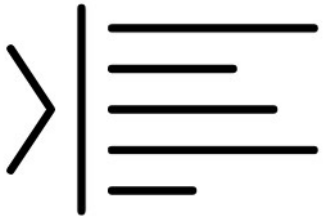
If your data is trapped in archives and databases, we can help extract it

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

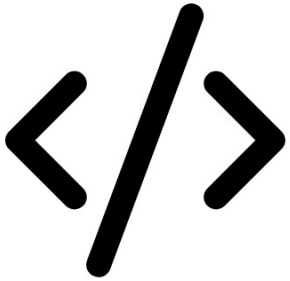
Need to re-format DOCX to XML, or PDF to WORD? Let us do it for you!

European Language Resource Coordination
Connecting Europe Facility

Language processing services

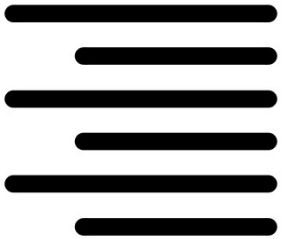
Data conversion

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



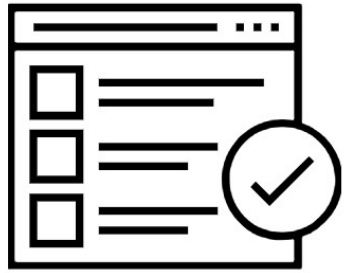
Tag removal

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

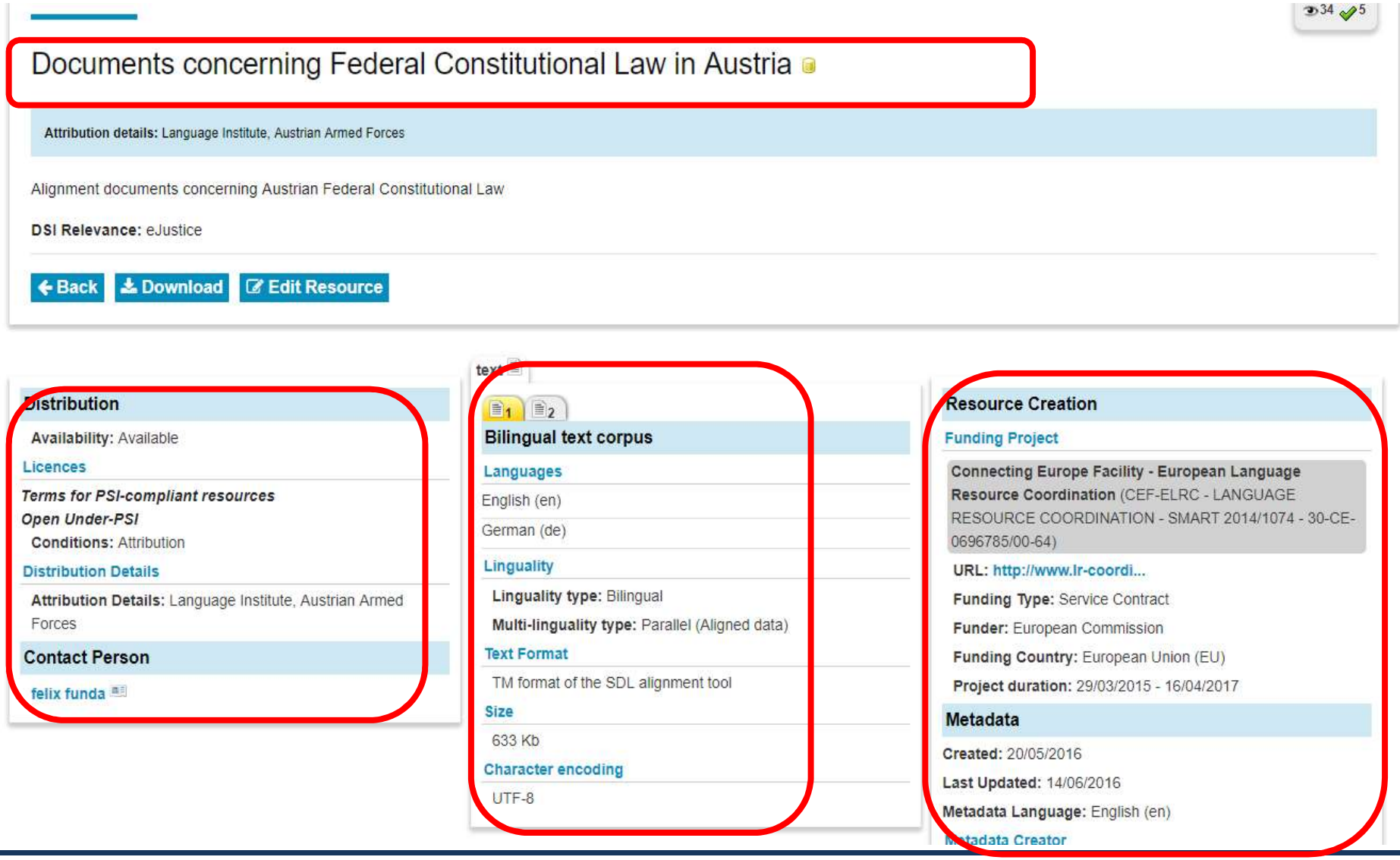
What has happened to your data?

File01_de.txt
File01_en.doc
File02_de.pdf
File02_en.txt
File03_de.doc
File03_en.doc
…


```
<body>
    <tu>
      <tuv xml:lang="de-DE">
        <seg>Präambel/Promulgationsklausel</seg>
      </tuv>
      <tuv xml:lang="en-GB">
        <seg>Preamble/Clause of Promulgation</seg>
      </tuv>
    </tu>
    <tu>
      <tuv xml:lang="de-DE">
        <seg>1.Jede Zensur ist als dem Grundrecht der
Staatsbürger widersprechend als rechtsungültig
aufgehoben.</seg>
      </tuv>
      <tuv xml:lang="en-GB">
        <seg>1.All censorship is abolished as illegal
because contradictory to the basic rights of the
citizen.</seg>
      </tuv>
```

European Language Resource Coordination
Connecting Europe Facility

The notion of data in the context of eTranslation



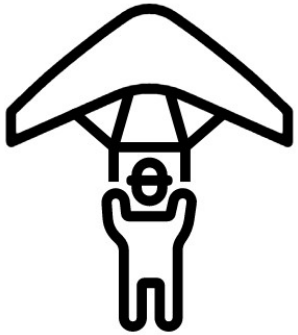
European Language Resource Coordination
Connecting Europe Facility

On-site assistance

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

European Language Resource Coordination
Connecting Europe Facility

On-site assistance



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

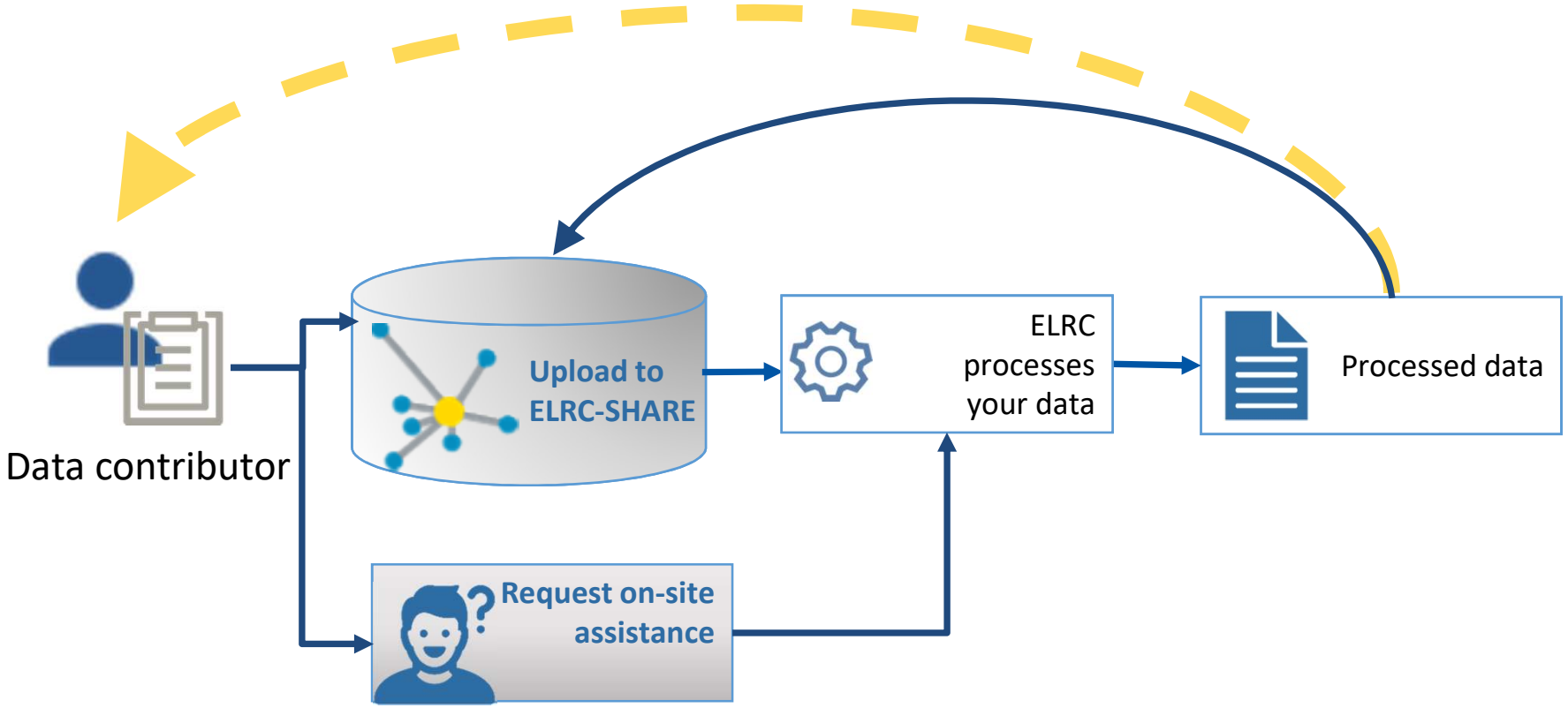
European Language Resource Coordination
Connecting Europe Facility

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!

European Language Resource Coordination
Connecting Europe Facility

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

European Language Resource Coordination
Connecting Europe Facility

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

European Language Resource Coordination
Connecting Europe Facility

ELRC consortium – come talk to us!

European Language Resource Coordination
Connecting Europe Facility

Danke für Ihre Aufmerksamkeit!

European Language Resource Coordination
Connecting Europe Facility

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