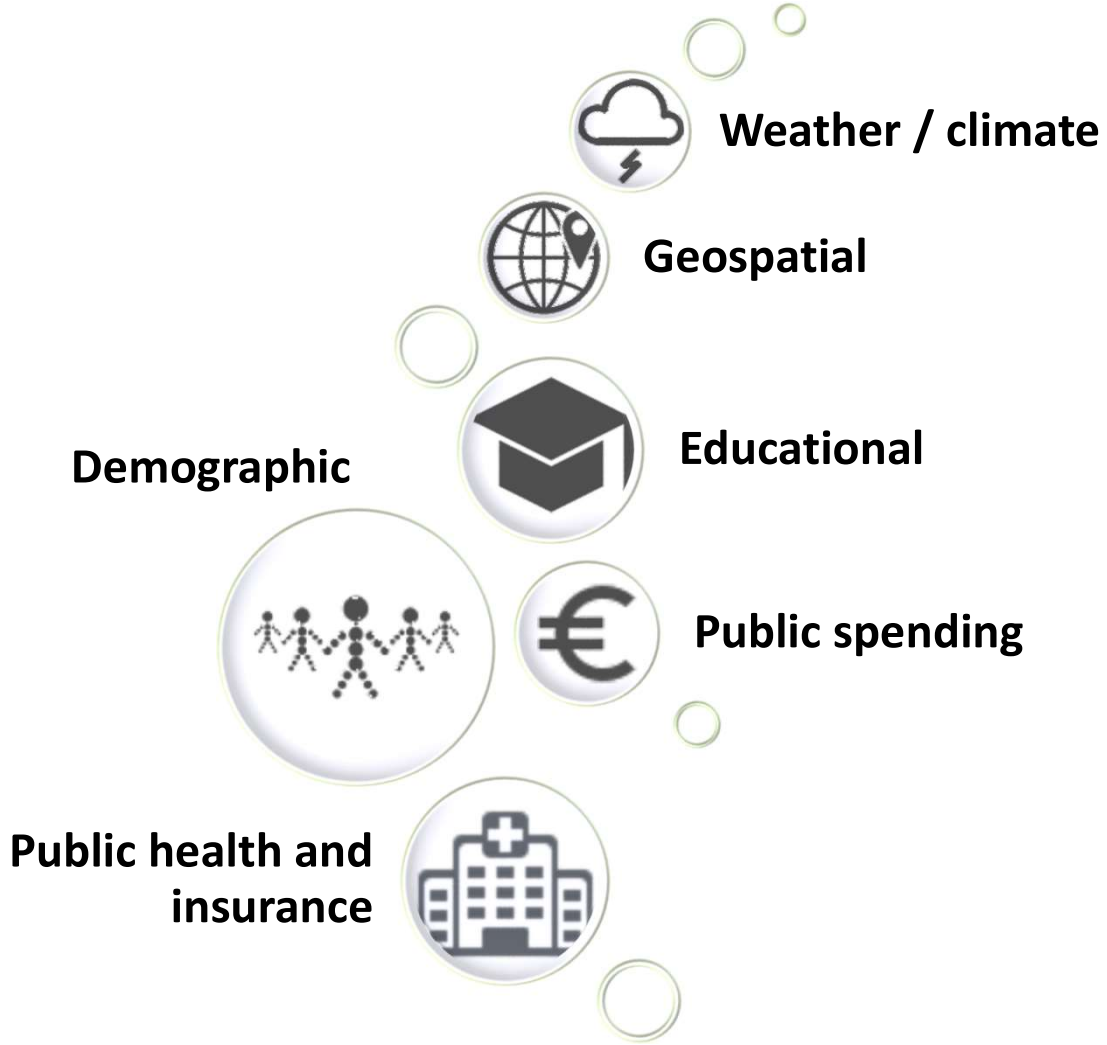
European Language Resource Coordination
Connecting Europe Facility

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

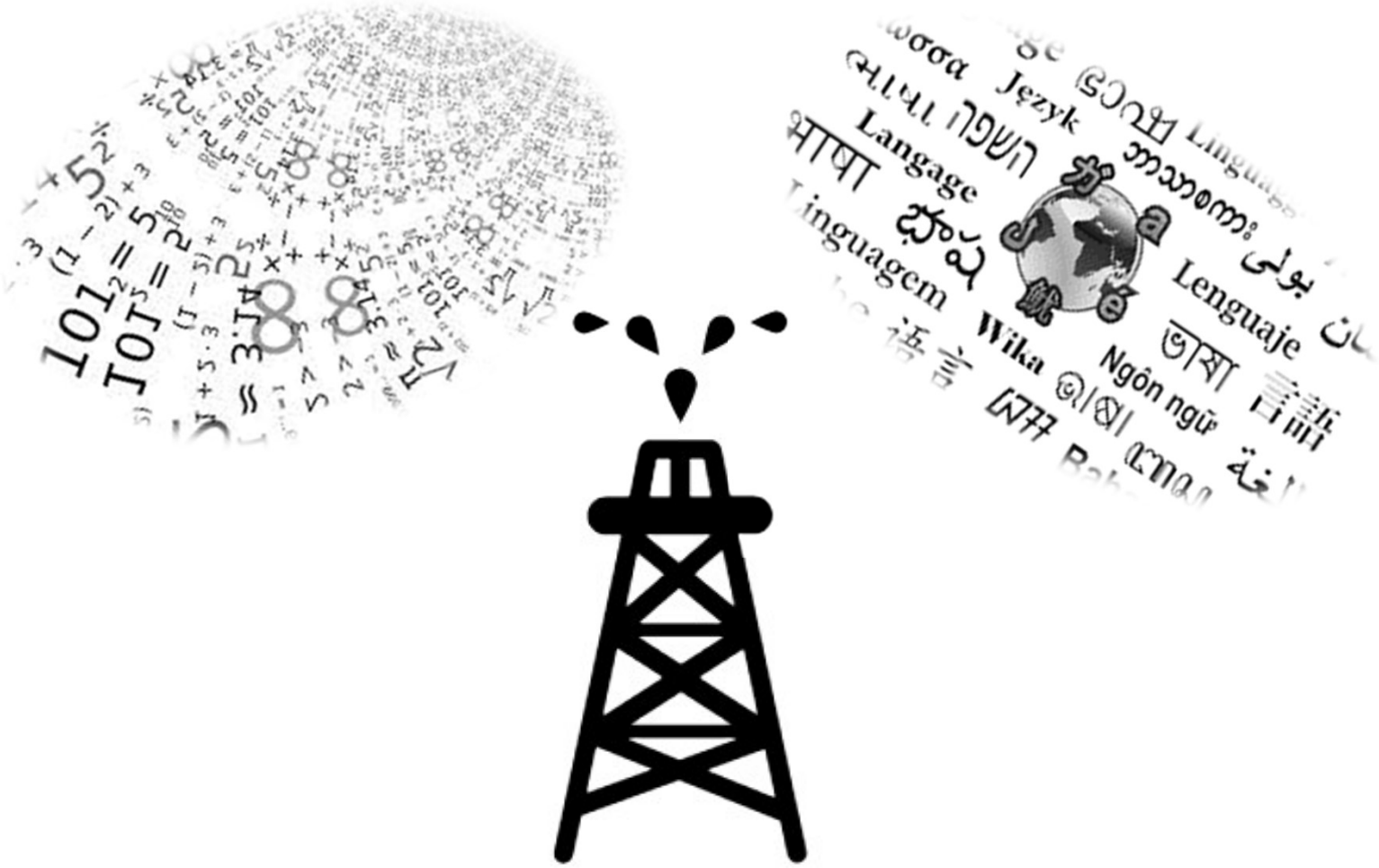


The notion of data

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 data

Basic concepts:

- **Data**: any piece of electronically stored content
- **Dataset (or resource):** the collection of one or many data files **grouped** according to certain **criteria**
- **Metadata:** *data about the data,* i.e. description of a dataset with properties (e.g. title, publisher, description of the content and URL)

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

CS

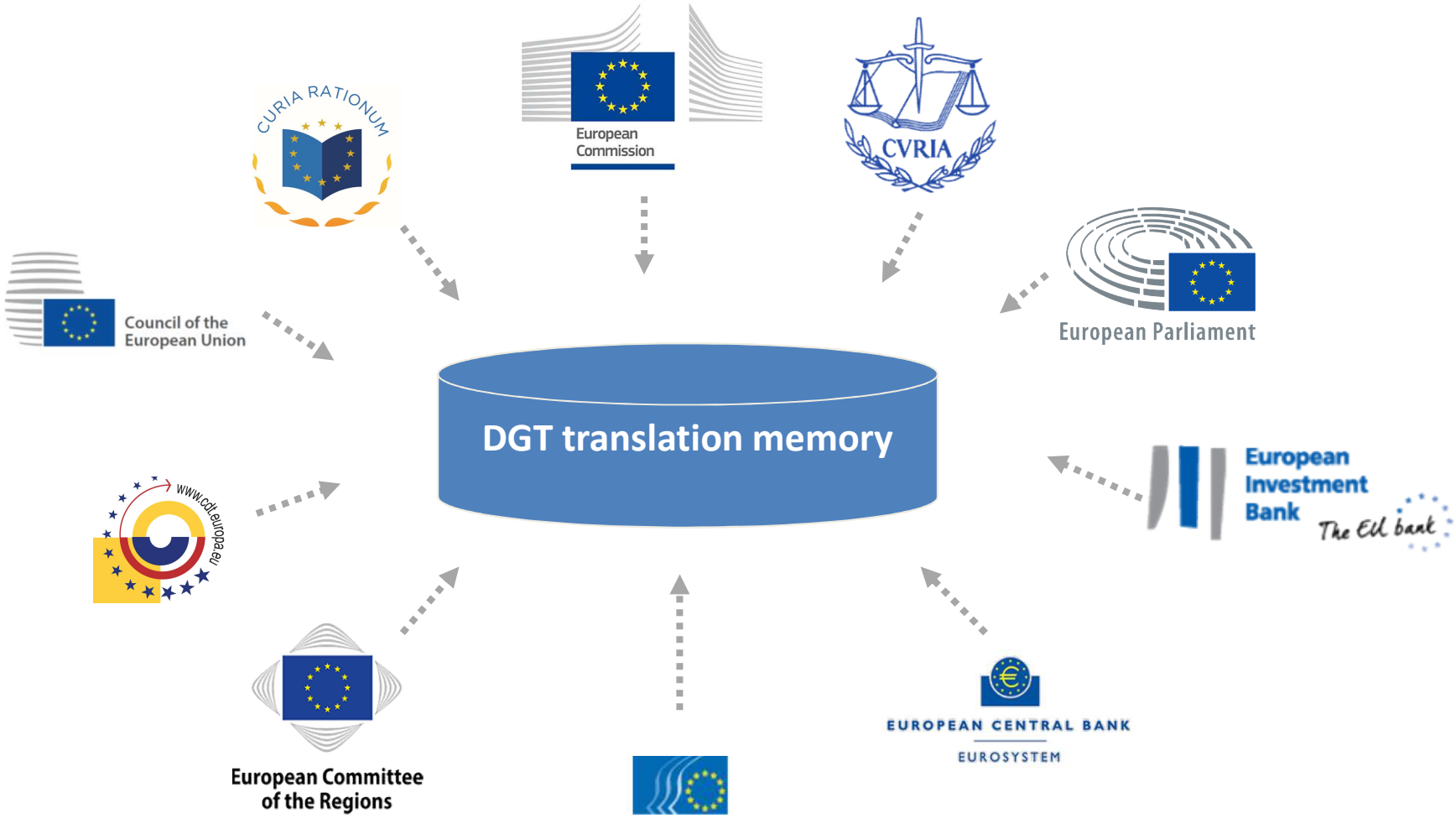
Celkové výdaje na dávky nemocenského pojištění v roce 2005, tedy včetně výdajů proplácených organizacím nad 25 zaměstnanců, činily 31 miliard 660 milionů korun. Nemocensky pojištění obyvatelé ČR v roce 2005 prostonali celkem 107 095 134 dnů, což je o 1 139 608 dnů více než v roce 2004.

EN

Total sickness benefit expenditures in 2005, including reimbursements to organisations with over 25 employees, were CZK 31,660 million. In 2005, Czech citizens affiliated to the sickness insurance scheme spent 107,095,134 days on sick leave, which is 1,139,608 days more than in 2004.

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)

Czech text

Úvodní ustanovení
(1) Rada zasedá podle potřeby, nejméně však čtyřikrát ročně.
(2) Zasedání Rady svolává její předseda / předsedkyně (dále jen „předseda“) prostřednictvím.
(3) Program zasedání navrhuje předsednictvo Rady a schvaluje její předseda v souladu s úkoly Rady nebo podle závěrů jejího předchozího zasedání.
(4) Rada je způsobilá jednat a přijímat závěry, pokud je přítomna nadpoloviční většina jejích členů nebo jejich stálých zástupců.
(5) Závěry se přijímají hlasováním.

Translation in English

Introductory provisions
(1) The Council shall meet as needed, however at least four times a year.
(2) Council sessions shall be called by its Chair via the Council Secretariat.
(3) The meeting programme is presented by the Council Board and approved by its Chair in accordance with the Council's tasks or the conclusions of its previous meeting.
(4) The Council is competent to act and adopt conclusions if a majority of its members or their permanent representatives is present.
(5) Conclusions shall be adopted by vote.

European Language Resource Coordination
Connecting Europe Facility

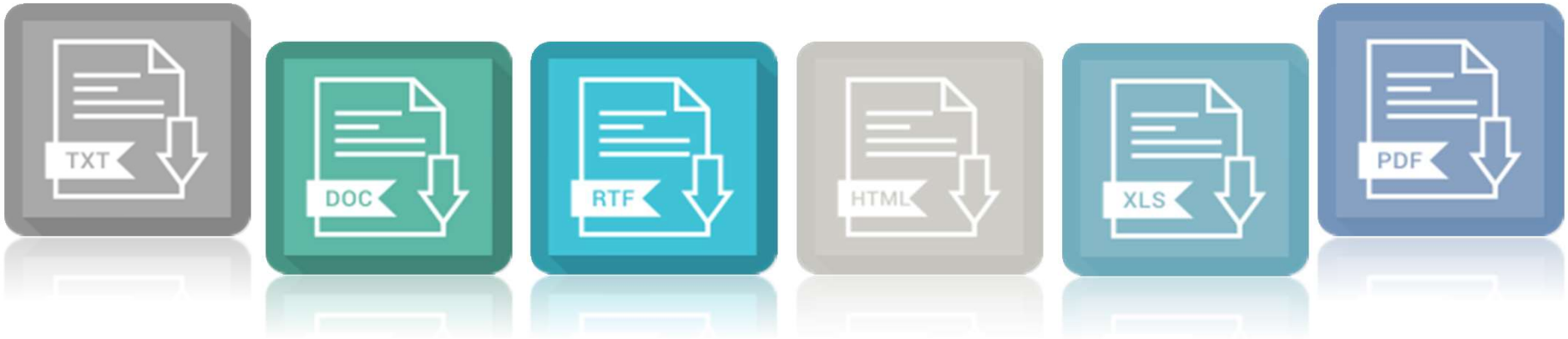
What data are useful for eTranslation as per type |2

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

Czech	English
Korporátní daň	Corporate tax
Korporátní obligace	Corporate bond
Koš (ve vztahu k derivátům)	Basket
Košová opce	Basket option
Kotace	Quote
Kotovací tabule	Quotation board
Kotovaná cena	Quoted price
Kotovaná společnost	Quoted company
Kotování	Quotation
Koupě při otevření	Buy on opening
…	…

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

European Language Resource Coordination
Connecting Europe Facility

File formats for parallel texts

Ročník 1993

SBÍRKA ZÁKONŮ
ČESKÉ REPUBLIKY

Částka 1

ÚSTAVNÍ ZÁKON
České národní rady

ÚSTAVA
ČESKÉ REPUBLIKY

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)

European Language Resource Coordination
Connecting Europe Facility

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.

Toto je výše uvedený odstavec v českém překladu. Toto je výše uvedený odstavec v českém překladu. Toto je výše uvedený odstavec v českém překladu. Toto je výše uvedený odstavec v českém překladu. Toto je výše uvedený odstavec v českém překladu. Toto je výše uvedený odstavec v českém překladu.

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. A second paragraph in English.

Toto je výše uvedený odstavec v českém překladu. Toto je výše uvedený odstavec v českém překladu. Toto je výše uvedený odstavec v českém překladu. Toto je výše uvedený odstavec v českém překladu.

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

A·sentence·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.¶

Toto·je·odstavec·vlevo,·přeložený·v·češtině.·Toto·je·odstavec·vlevo,·přeložený·v·češtině.·Toto·je·odstavec·vlevo,·přeložený·v·češtině.·Toto·je·odstavec·vlevo,·přeložený·v·češtině.·Toto·je·odstavec·vlevo,·přeložený·v·češtině.¶

Toto·je·věta·vlevo,·přeložená·v·češtině.·¶

¶

Jedná·se·o·druhý·odstavec,·přeložený·česky.·Jedná·se·o·druhý·odstavec,·přeložený·česky.·Jedná·se·o·druhý·odstavec,·přeložený·česky.·Jedná·se·o·druhý·odstavec,·přeložený·česky.¶

¶

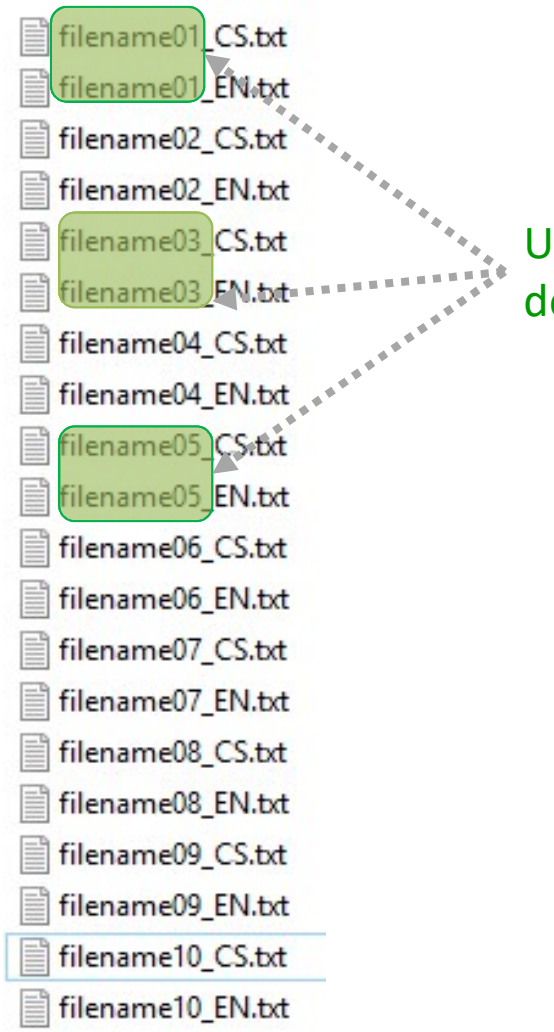
Preparing your data |3

Don'ts

English	čeština
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.	Toto je odstavec vlevo, přeložený v češtině. Toto je odstavec vlevo, přeložený v češtině. Toto je odstavec vlevo, přeložený v češtině. Toto je odstavec vlevo, přeložený v češtině. Toto je odstavec vlevo, přeložený v češtině.
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. A second paragraph in English.	Jedná se o druhý odstavec, přeložený česky. Jedná se o druhý odstavec, přeložený česky. Jedná se o druhý odstavec, přeložený česky. Jedná se o druhý odstavec, přeložený česky.

Preparing your data |4

Do's

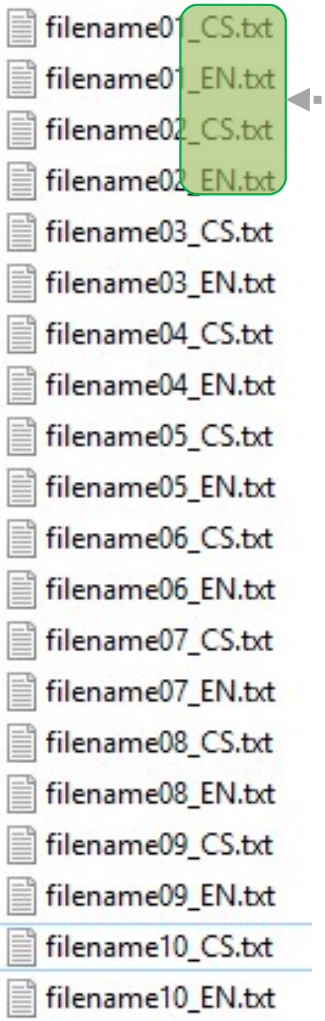


Use **identical filenames** for each document pair (source – translation)

European Language Resource Coordination
Connecting Europe Facility

Preparing your data |5

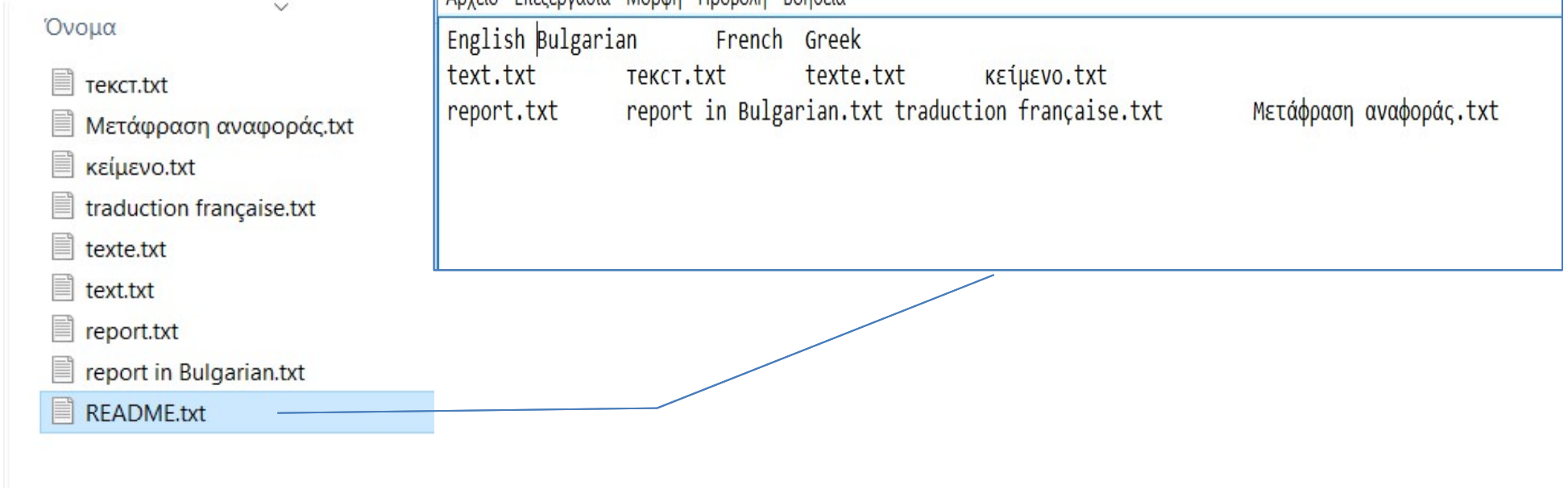
Do's



Include **language identifiers** in the filename

Preparing your data |5

Do's



European Language Resource Coordination
Connecting Europe Facility

Criteria for grouping your data into meaningful datasets

- 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 – Czech*
    - *Documents in English – Czech - French*
  - **Domain**: each collection ideally belongs to a single domain, e.g.
    - *Collection of texts in English – Czech in the culture domain*
    - *Social security documents in English – Czech - French*

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

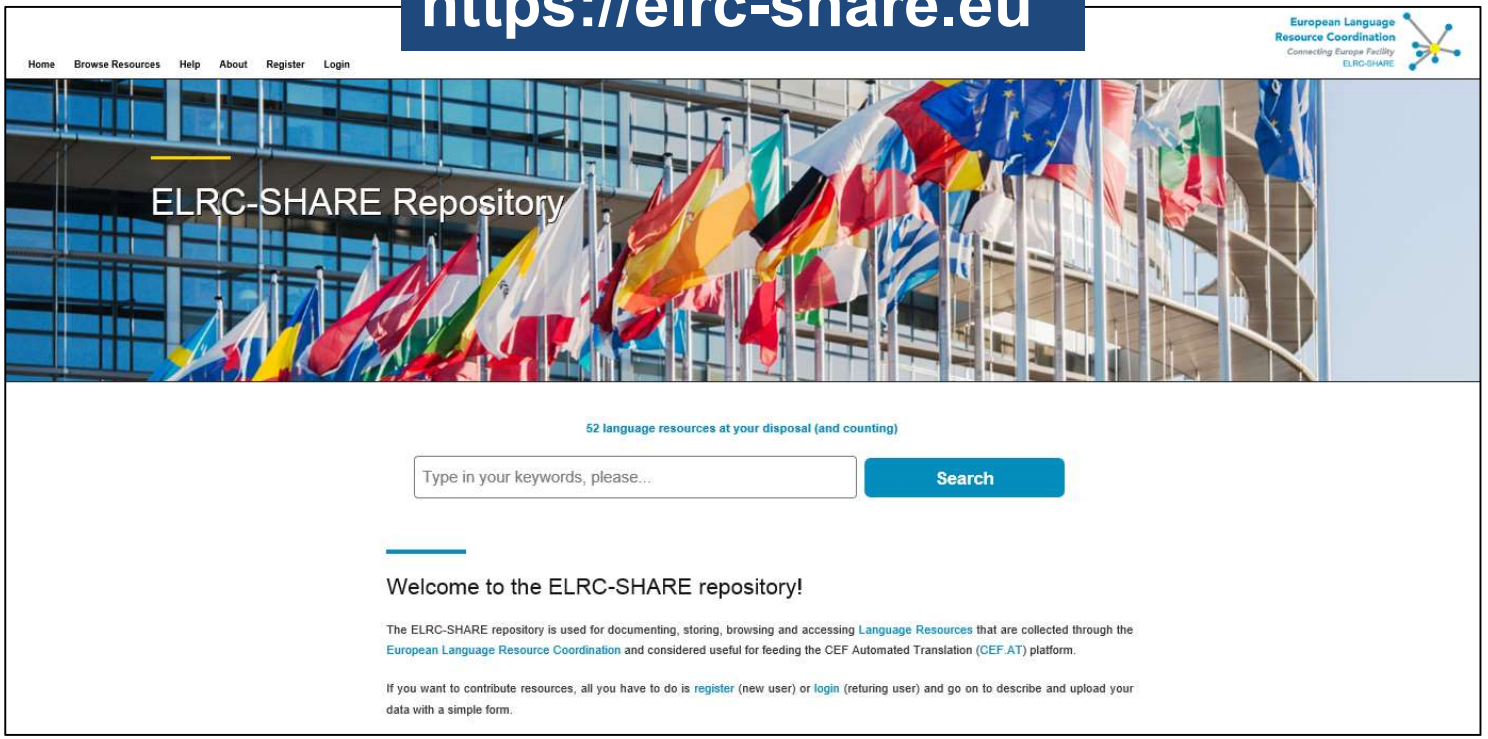
How to contribute your data to CEF eTranslation
A step-by-step guide

To ELRC-SHARE

The ELRC-SHARE Repository

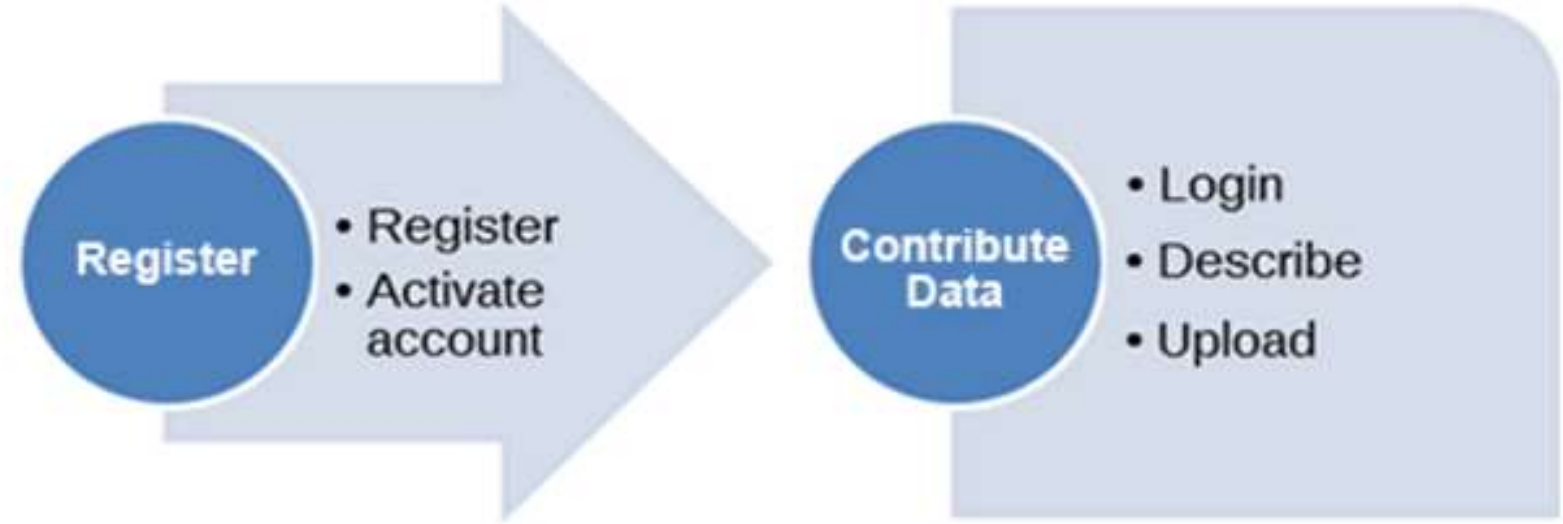
- Access to, sharing and contribution of language resources
- Access to tools and services catalogue (upcoming)

https://elrc-share.eu



European Language Resource Coordination
Connecting Europe Facility

How to Contribute Data

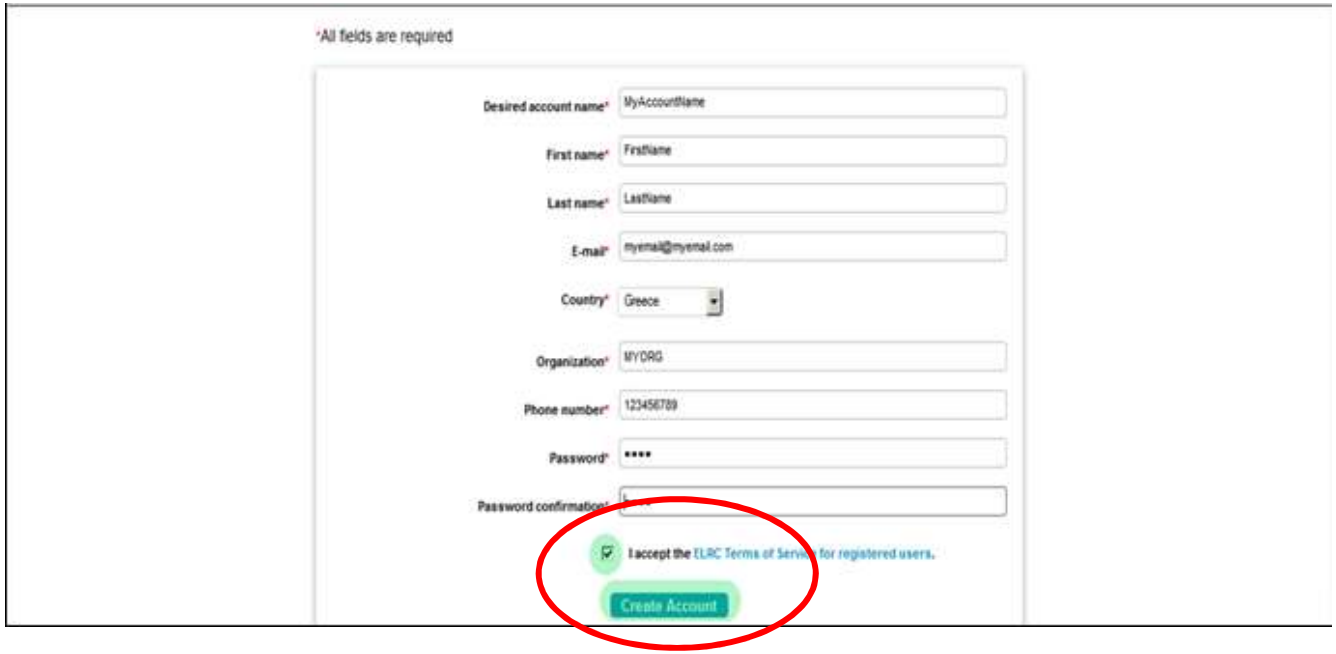


How to Register (1/2)

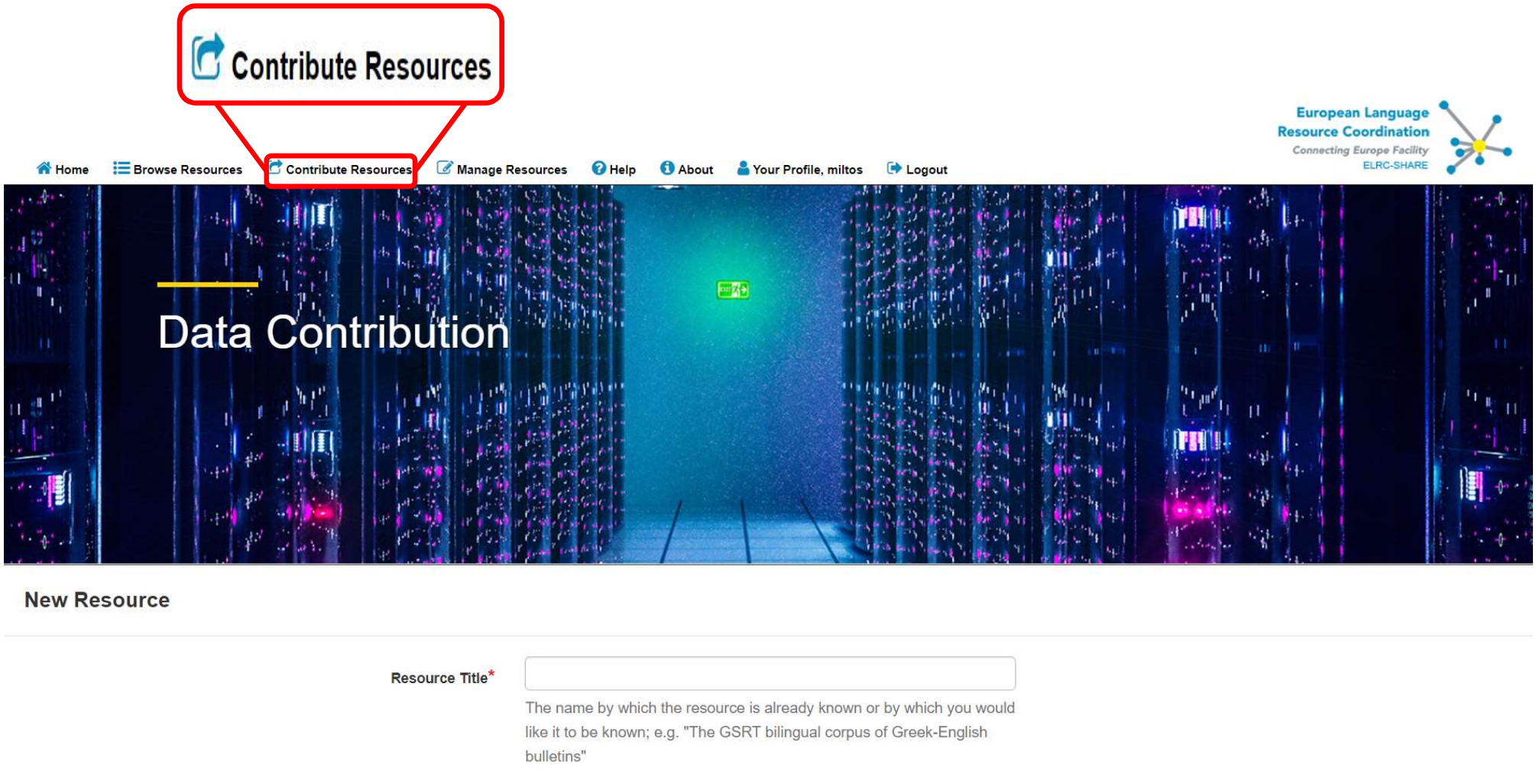
European Language Resource Coordination
Connecting Europe Facility

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



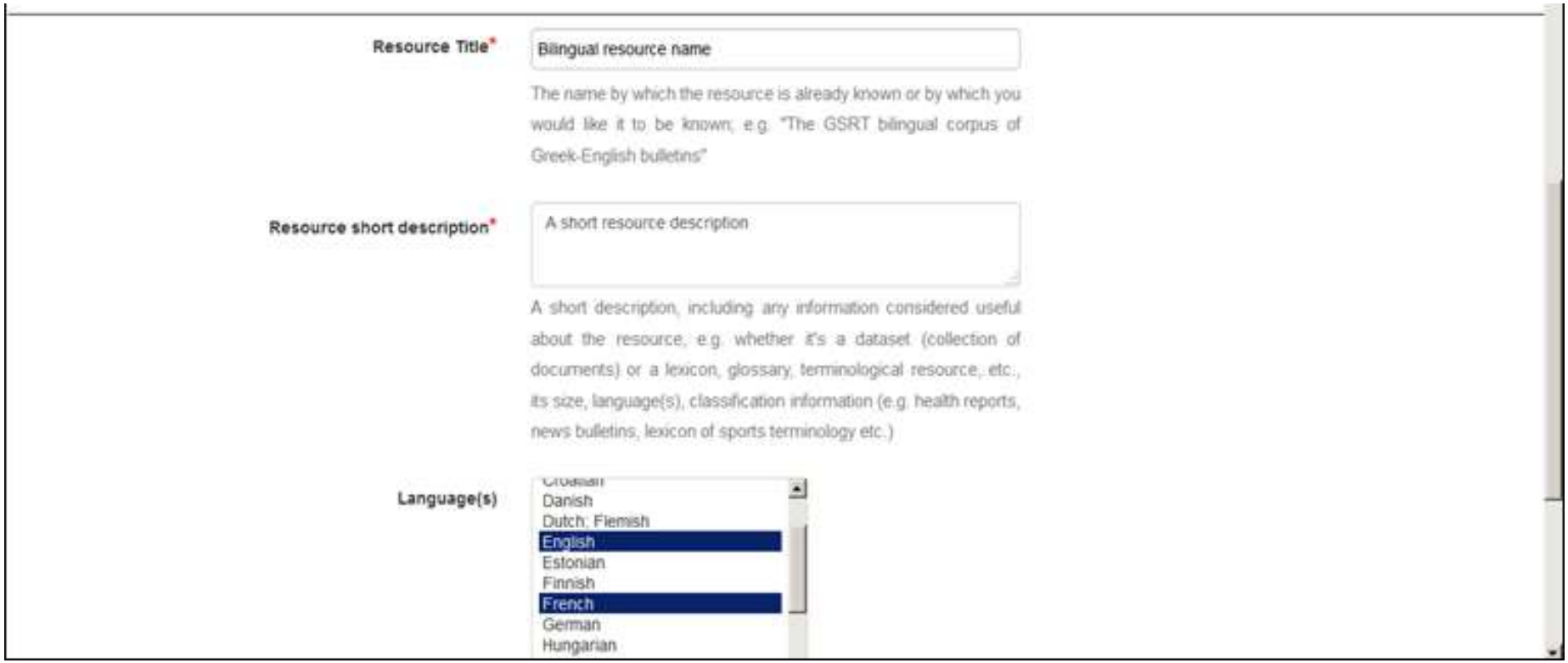
How to Contribute Data (1/6)



European Language Resource Coordination
Connecting Europe Facility

How to Contribute Data (2/6)

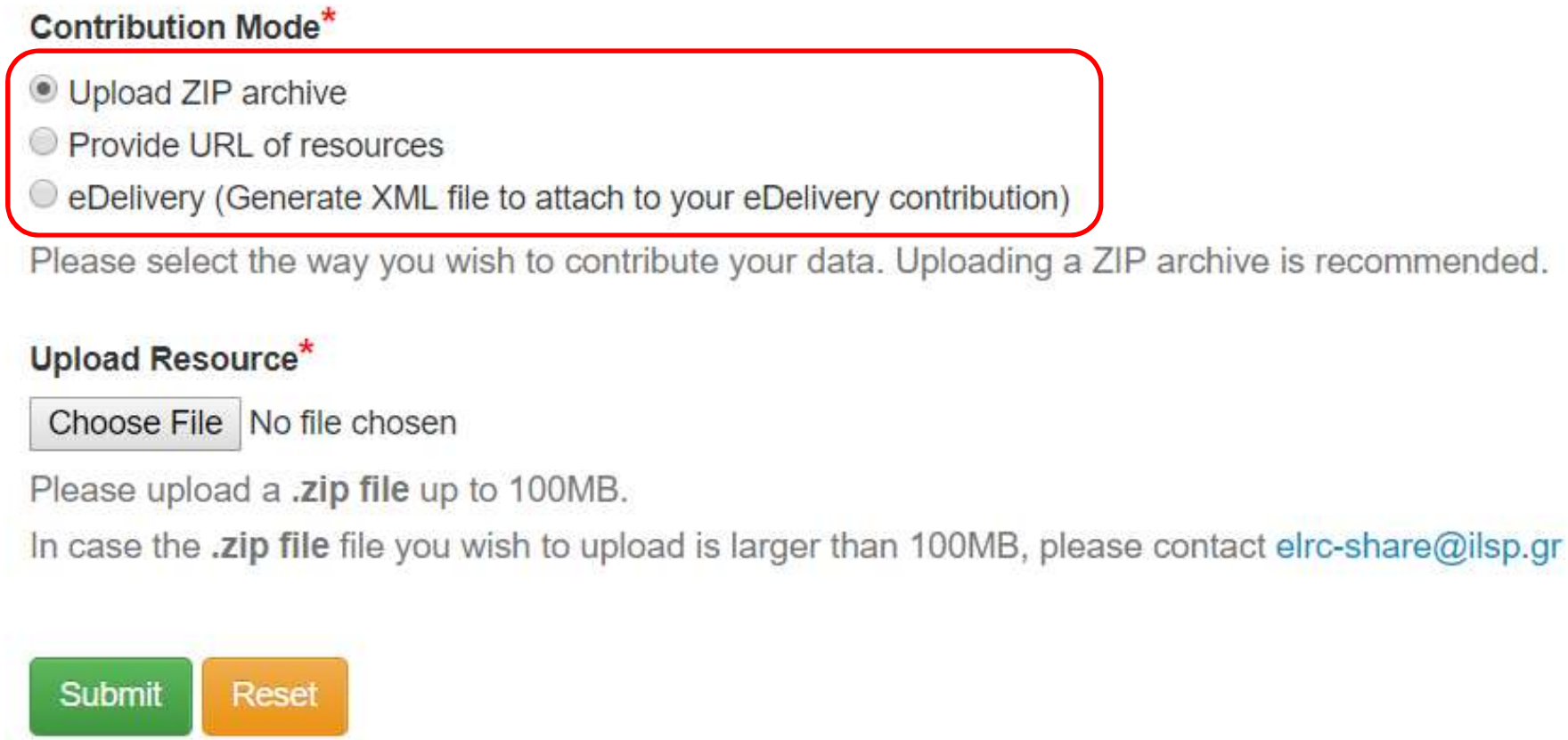
- Fill in the details of the dataset



European Language Resource Coordination
Connecting Europe Facility

How to Contribute Data (3/6)

- Three modes for contributing your data



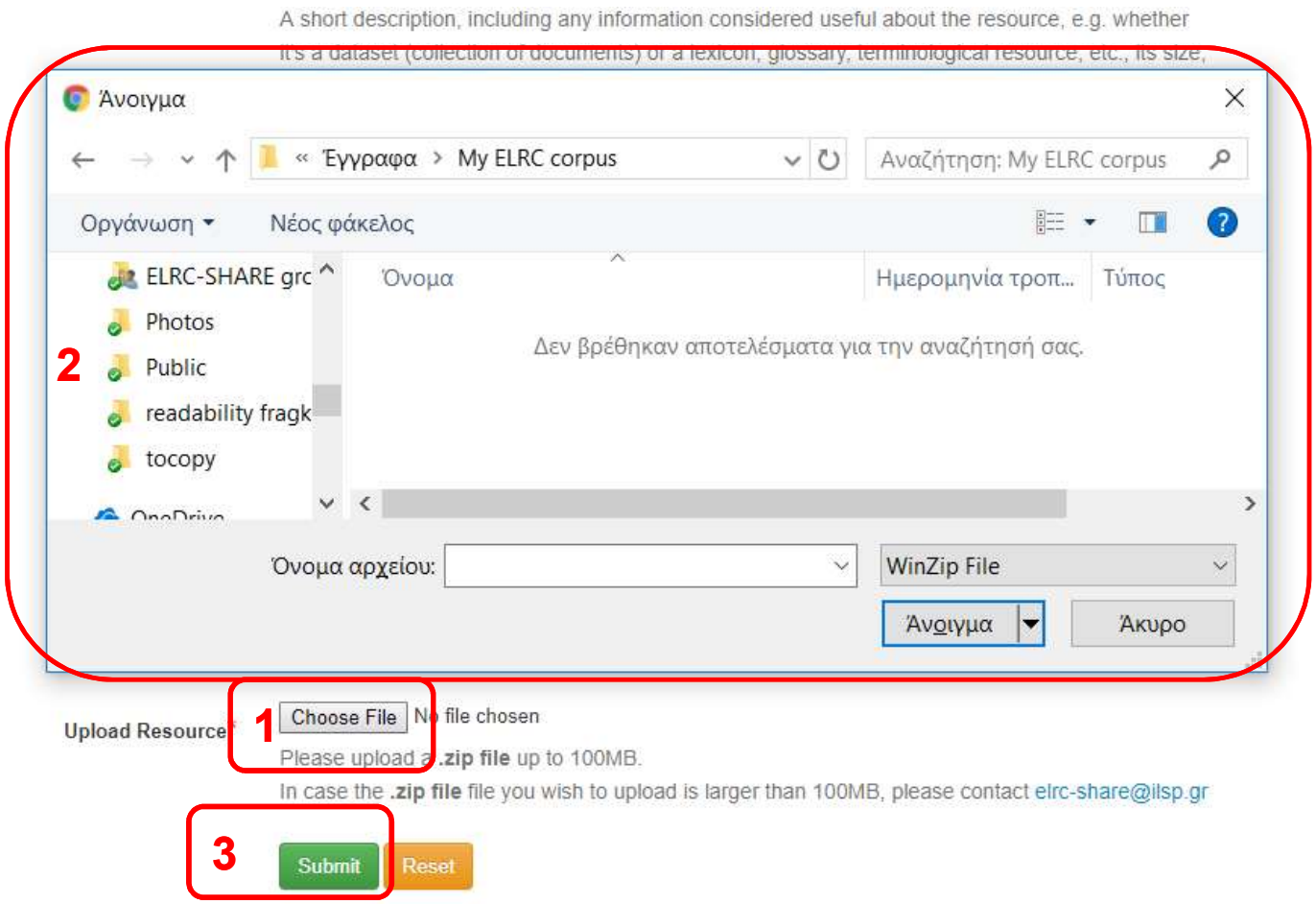
European Language Resource Coordination
Connecting Europe Facility

How to create a .zip file

- Free file compression tools (indicative):
  - 7zip
  - PeaZip
  - Hamster Free Zip Archiver
  - Universal Extractor
  - ZipItFree
- Windows embedded compression functionality

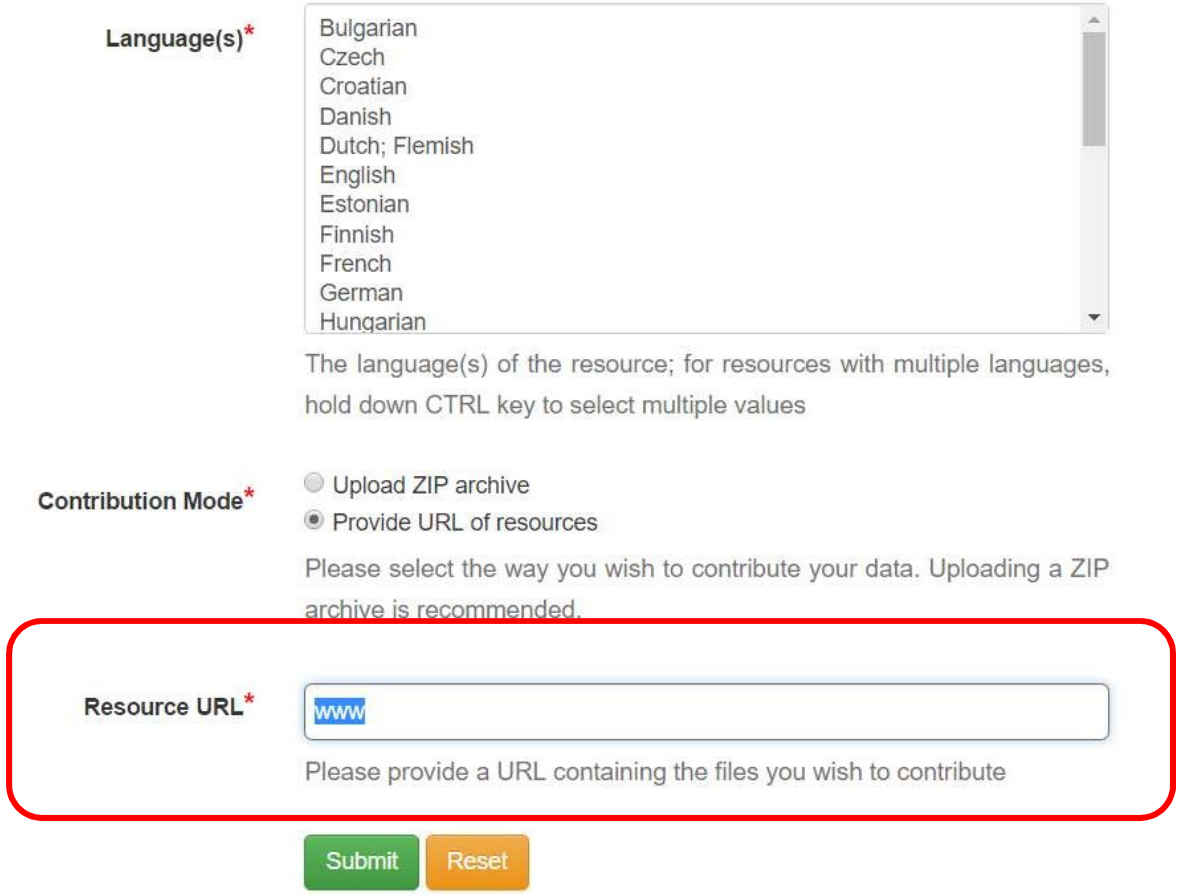
How to Contribute Data (4/6)

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



How to Contribute Data (5/6)

- Alternatively indicate a url (directory listing)



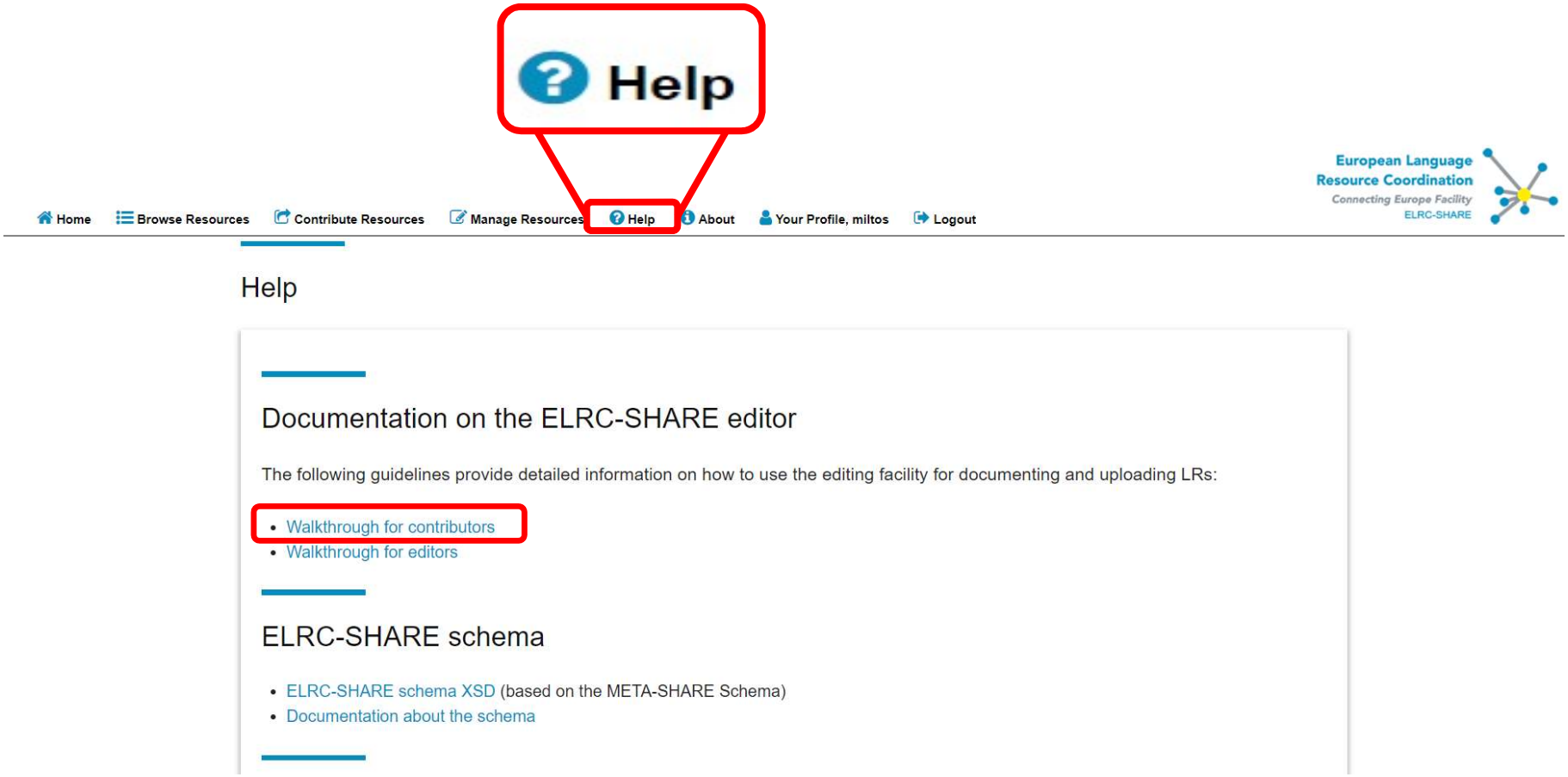
European Language Resource Coordination
Connecting Europe Facility

How to Contribute Data (6/6)

Contribute Your Data Through eDelivery

If you wish to share your data through eDelivery , you can use the ELRC-SHARE CEF eDelivery Access Point. For more information click here.
In such a case, please fill in the form on the left and choose eDelivery in the Contribution mode.

Guidelines for contributors

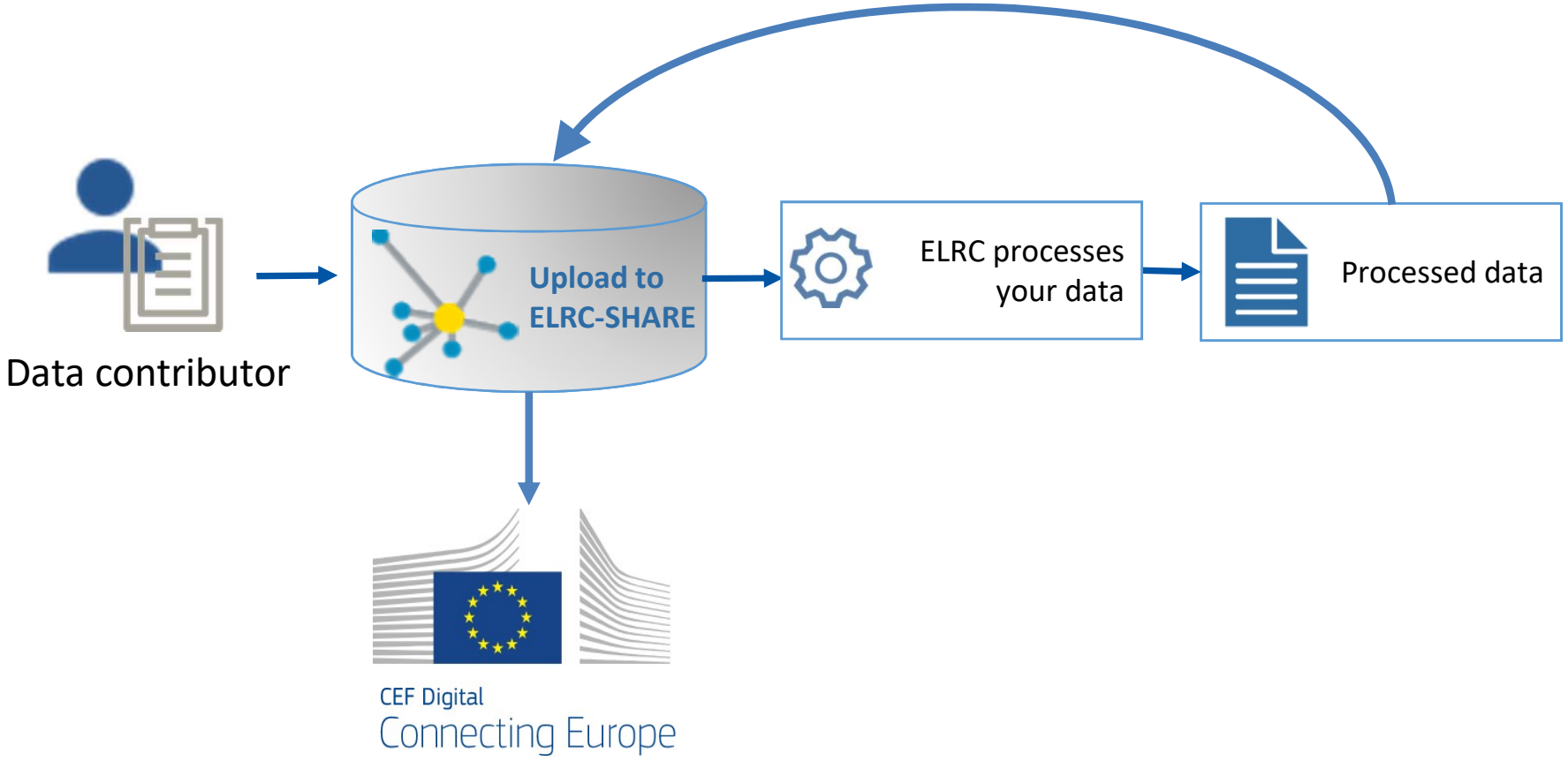


European Language Resource Coordination
Connecting Europe Facility

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

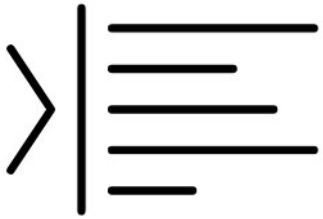
If your data is trapped in archives and databases, we can help extract it

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

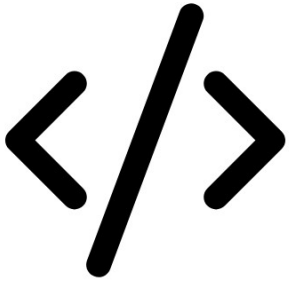
Need to re-format DOCX to XML, or PDF to WORD? Let us do it for you!

European Language Resource Coordination
Connecting Europe Facility

Language processing services

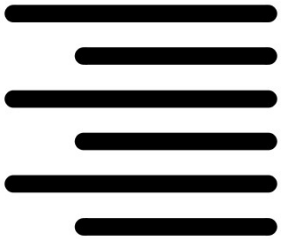
Data conversion

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



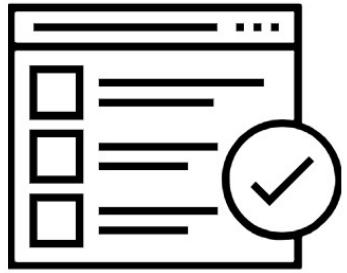
Tag removal

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?

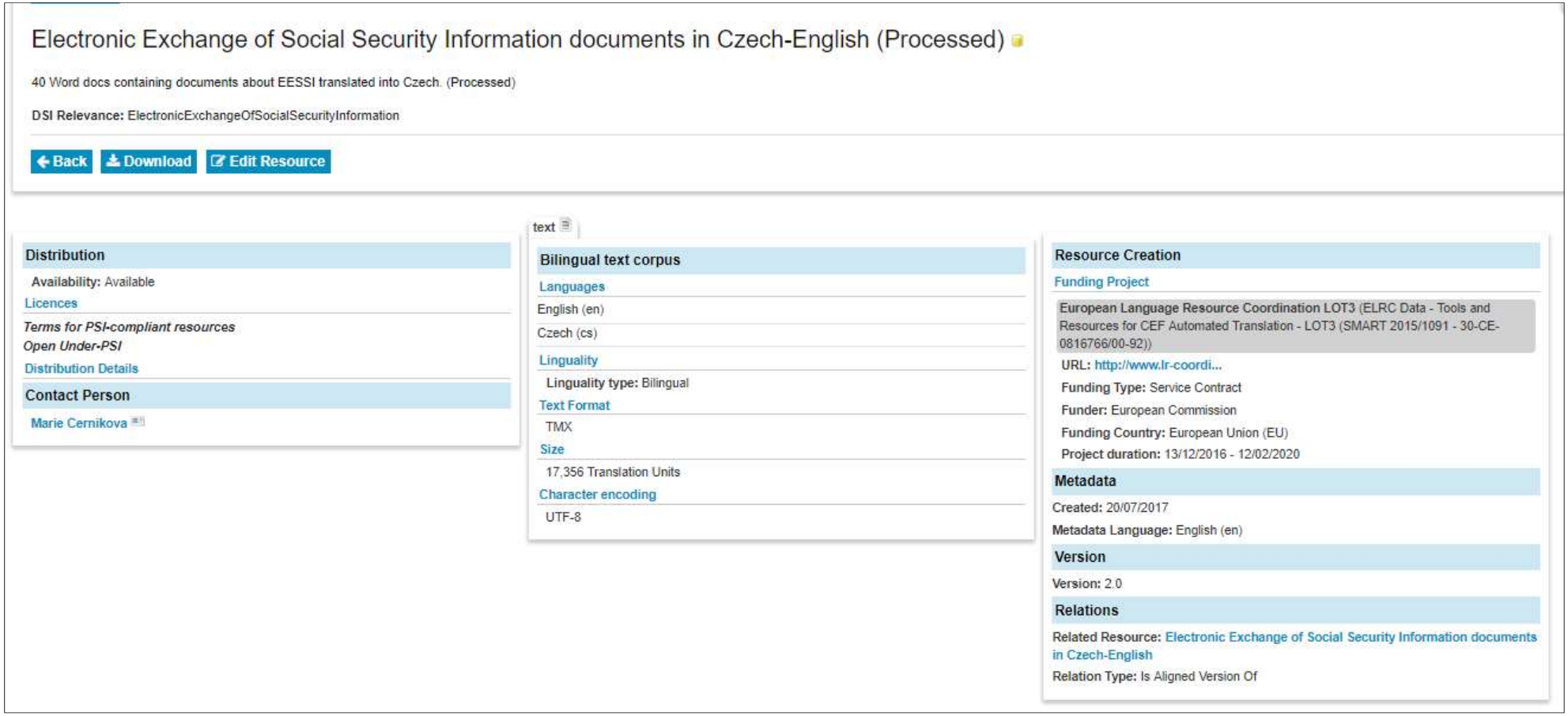
File01_cs.txt
File01_en.doc
File02_cs.pdf
File02_en.txt
File03_cs.doc
File03_en.doc
…

After processing

```
<tu tuid="818">
      <tuv xml:lang="cs">
        <seg>Celkové výdaje na dávky nemocenského
pojištění v roce 2005, tedy včetně výdajů proplácených
organizacím nad 25 zaměstnanců, činily 31 miliard 660
milionů korun.</seg> </tuv>
      <tuv xml:lang="en">
        <seg>Total sickness benefit expenditures in
2005, including reimbursements to organisations with
over 25 employees, were CZK 31,660
million.</seg></tuv></tu>
    <tu tuid="819">
      <tuv xml:lang="cs">
        <seg>Nemocensky pojištění obyvatelé ČR v roce
2005 prostonali celkem 107 095 134 dnů, což je o 1 139
608 dnů více než v roce 2004.</seg></tuv>
      <tuv xml:lang="en">
        <seg>In 2005, Czech citizens affiliated to the
sickness insurance scheme spent 107,095,134 days on sick
leave, which is 1,139,608 days more than in
2004.</seg></tuv></tu>
```

European Language Resource Coordination
Connecting Europe Facility

How your dataset is described



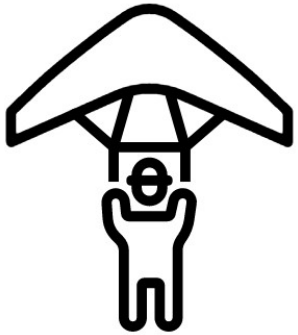
European Language Resource Coordination
Connecting Europe Facility

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

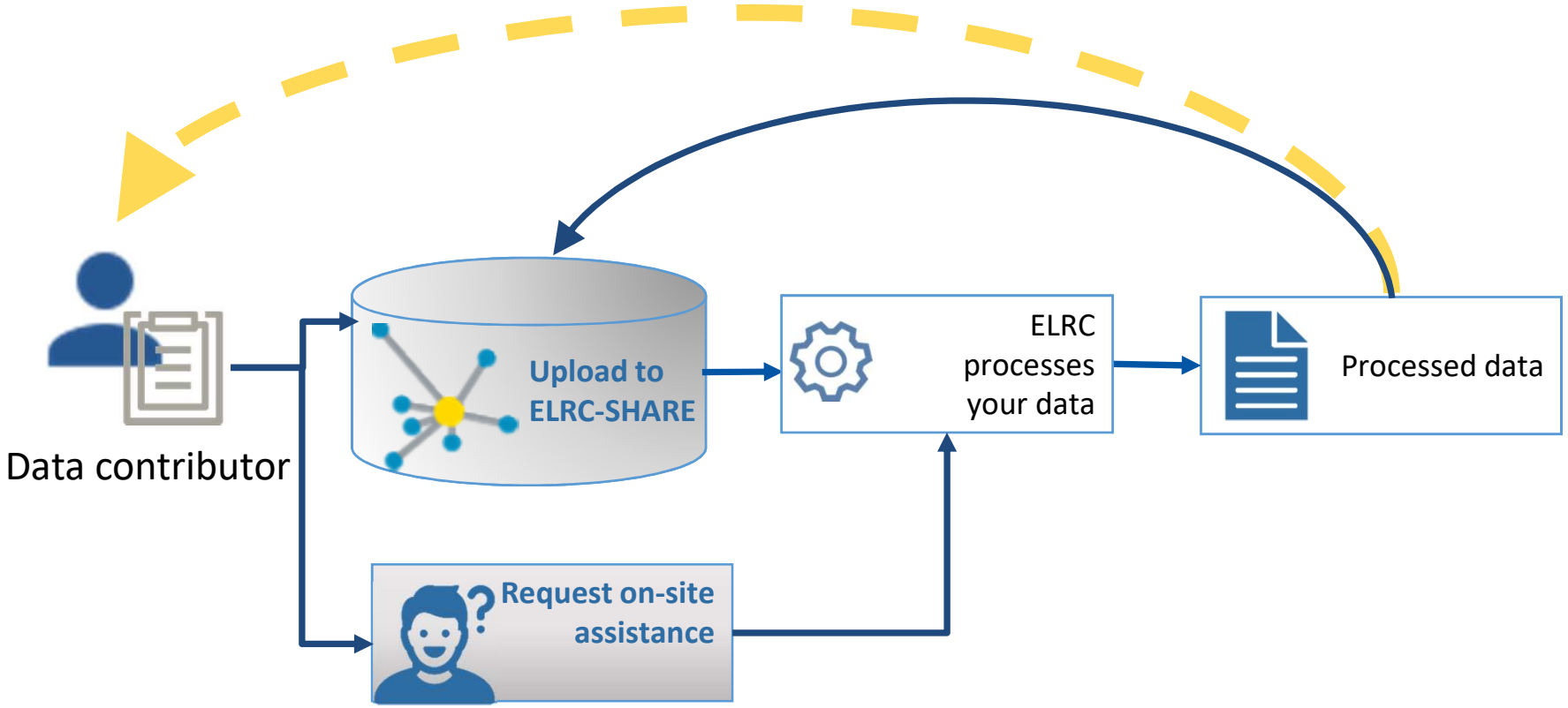
European Language Resource Coordination
Connecting Europe Facility

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

www.lr-coordination.eu/request-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

European Language Resource Coordination
Connecting Europe Facility

ELRC Helpdesk

www.lr-coordination.eu/helpdesk

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

European Language Resource Coordination
Connecting Europe Facility

Děkuji!

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

Case studies (2015-2016)

European Language Resource Coordination
Connecting Europe Facility

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

European Language Resource Coordination
Connecting Europe Facility

Spain

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

Solution: ELRC helped provide good converters to get usable documents

European Language Resource Coordination
Connecting Europe Facility

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