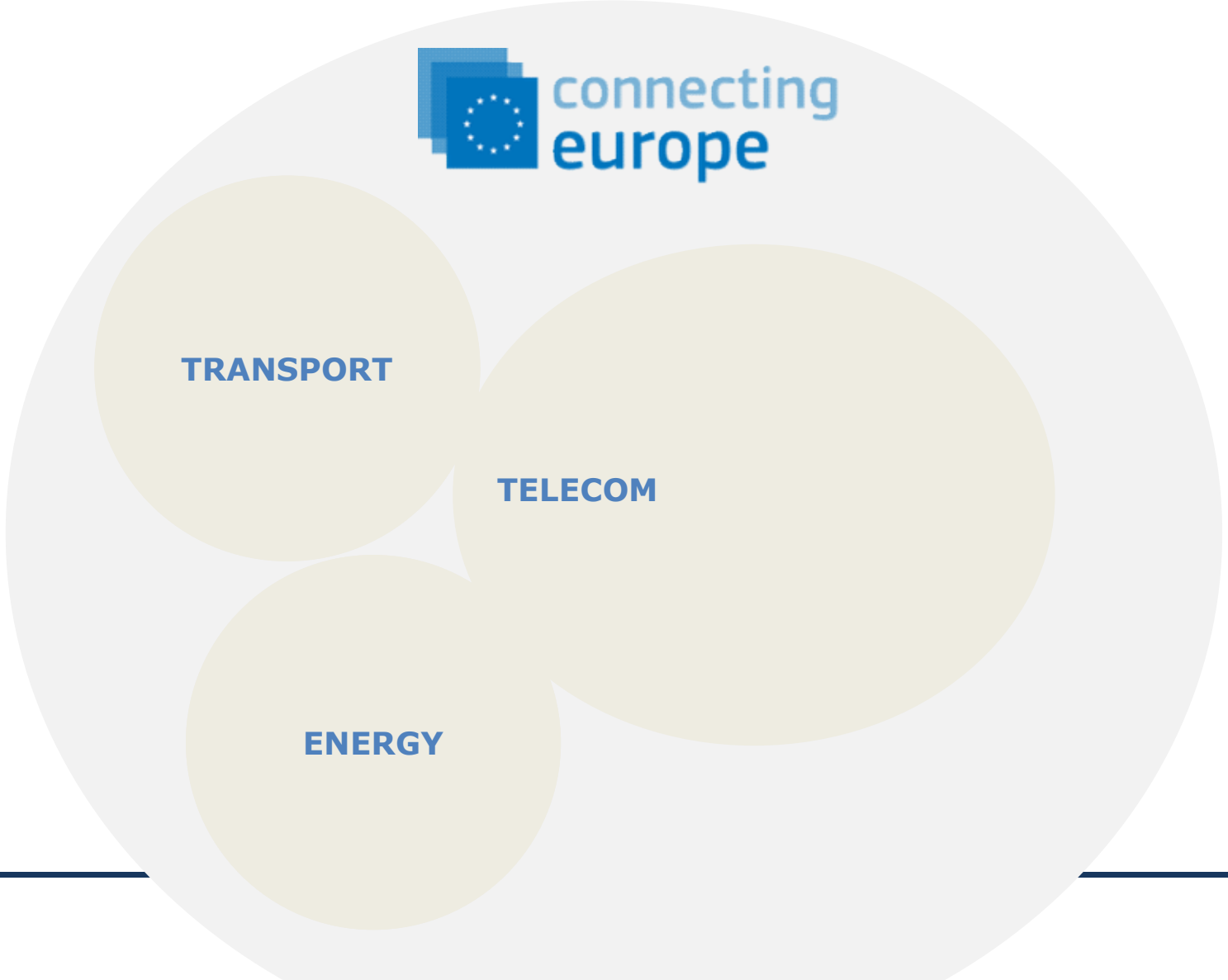
European Language Resource Coordination
Connecting Europe Facility

The European Language Resource Coordination (ELRC): why and how you can engage and benefit

Stelios Piperidis
Institute for Language and Speech Processing / Athena R.C.
ELRC

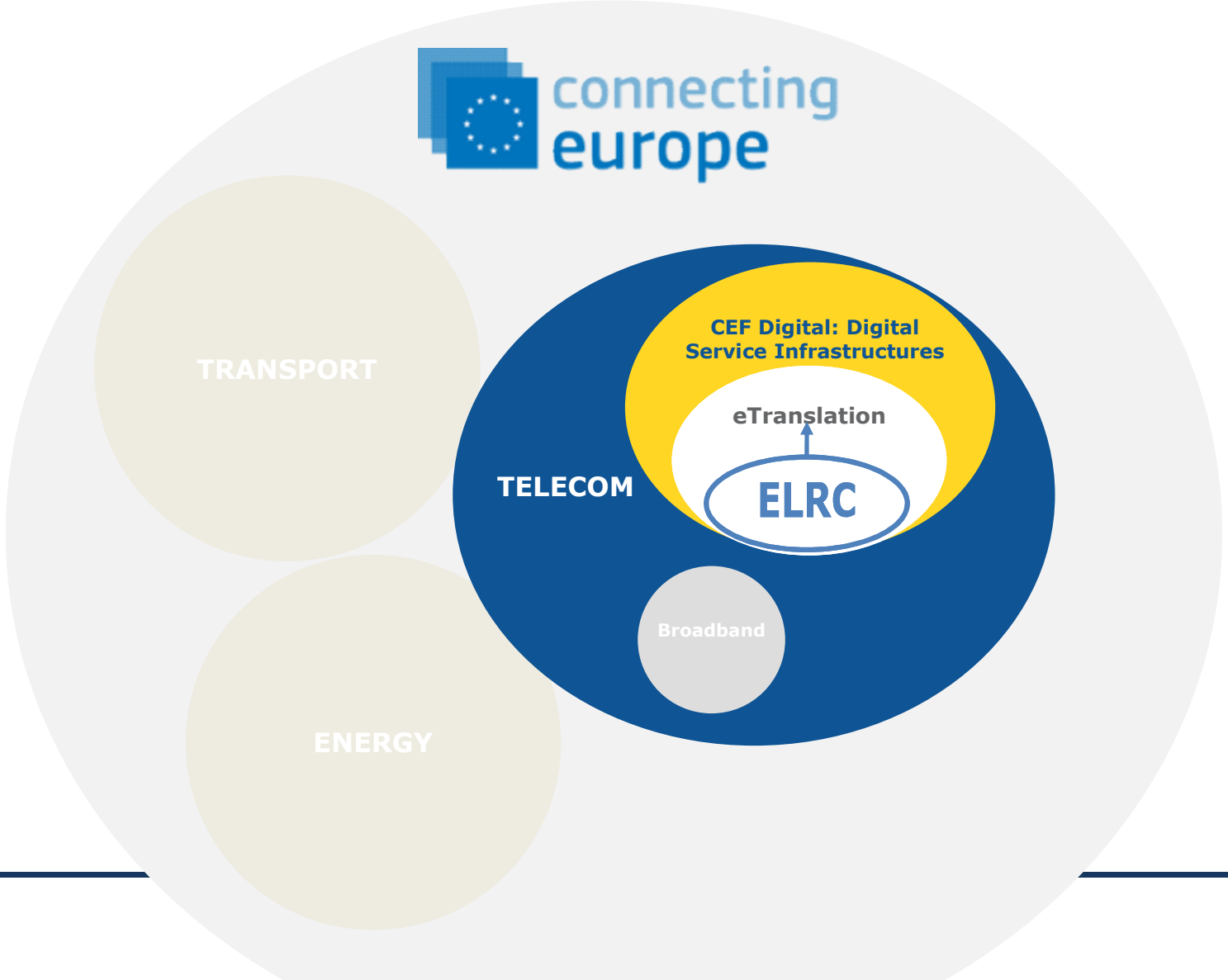
European Language Resource Coordination
Connecting Europe Facility

The ELRC context



European Language Resource Coordination
Connecting Europe Facility

The ELRC context



European Language Resource Coordination
Connecting Europe Facility

Heads & organisations behind the ELRC

The ELRC Consortium:

German Research Center for Artificial Intelligence

European Language Resource Coordination
Connecting Europe Facility

Heads & organisations behind the ELRC

30 ELRC Technological National Anchor Points

30 ELRC Public Services National Anchor Points

European Language Resource Coordination
Connecting Europe Facility

The ELRC anchor points in Hungary

Tamás Váradi
Research Institute for Linguistics, Hungarian Academy of Sciences

Zoltán Bódi
Institute for Hungarian Language Strategy

Janos Szenthe
Department of EU Law at the Ministry of Justice

European Language Resource Coordination
Connecting Europe Facility

The ELRC anchor points in Hungary

Tamás Váradi

Research Institute for Linguistics, Hungarian Academy of Sciences

Zoltán Bódi

Institute for Hungarian Language Strategy

Janos Szenthe

Department of EU Law at the Ministry of Justice

European Language Resource Coordination
Connecting Europe Facility

Why ELRC…?

- To facilitate multilingual communication and exchange of information in key public service scenarios
- How? With the right language resources!
- Examples of public services scenarios to which eTranslation may be applied:
  - Consumer Rights (ODR),
  - Social Security (EESSI),
  - Justice (eJustice),
  - Commerce (BRIS),
  - Open Data (EDP),
  - …

European Language Resource Coordination
Connecting Europe Facility

Language equality in the digital age

THE EUROPEAN PARLIAMENT VOTED IN FAVOR OF THE RESOLUTION ON **"LANGUAGE EQUALITY IN THE DIGITAL AGE"** ON 11. SEPTEMBER

European Language Resource Coordination
Connecting Europe Facility

Language equality in the digital age

"18. Calls on the Commission and the Member States to develop strategies and policy action to facilitate multilingualism in the digital market; requests, in this context, that the Commission and the Member States define the minimum language resources that all European languages should possess, such as data sets, lexicons, speech records, translation memories, annotated corpora and encyclopaedic content, in order to prevent digital extinction;"

European Language Resource Coordination
Connecting Europe Facility

Language equality in the digital age

"21. Calls on the Commission to make as a priority of language technology those Member States which are small in size and have their own language, in order to pay heed to the linguistic challenges that they face;"

European Language Resource Coordination
Connecting Europe Facility

Language equality in the digital age

Committee on Industry, Research and Energy (11.9.18): stresses "the role of the EU, the Member States, universities and other public institutions in contributing to the preservation of their languages in the digital world and in developing databases and translation technologies for all EU languages, including languages that are less widely spoken; calls for coordination between research and industry with a common objective of enhancing the digital possibilities for language translation and with open access to the data required for technological advancement".

European Language Resource Coordination
Connecting Europe Facility

Why ELRC…?

- Cracking the Language Barrier Federation (2016): "We recommend that Europe actively makes an effort to compete in the global landscape for research and development in language technology since we cannot expect third parties from other continents to solve our translation and knowledge management problems in a way that suits our specific communicative, societal and cultural needs. Language Technology made for Europe in Europe is the key. It will contribute to future European crossborder and cross-language communication, economic growth and social stability."

European Language Resource Coordination
Connecting Europe Facility

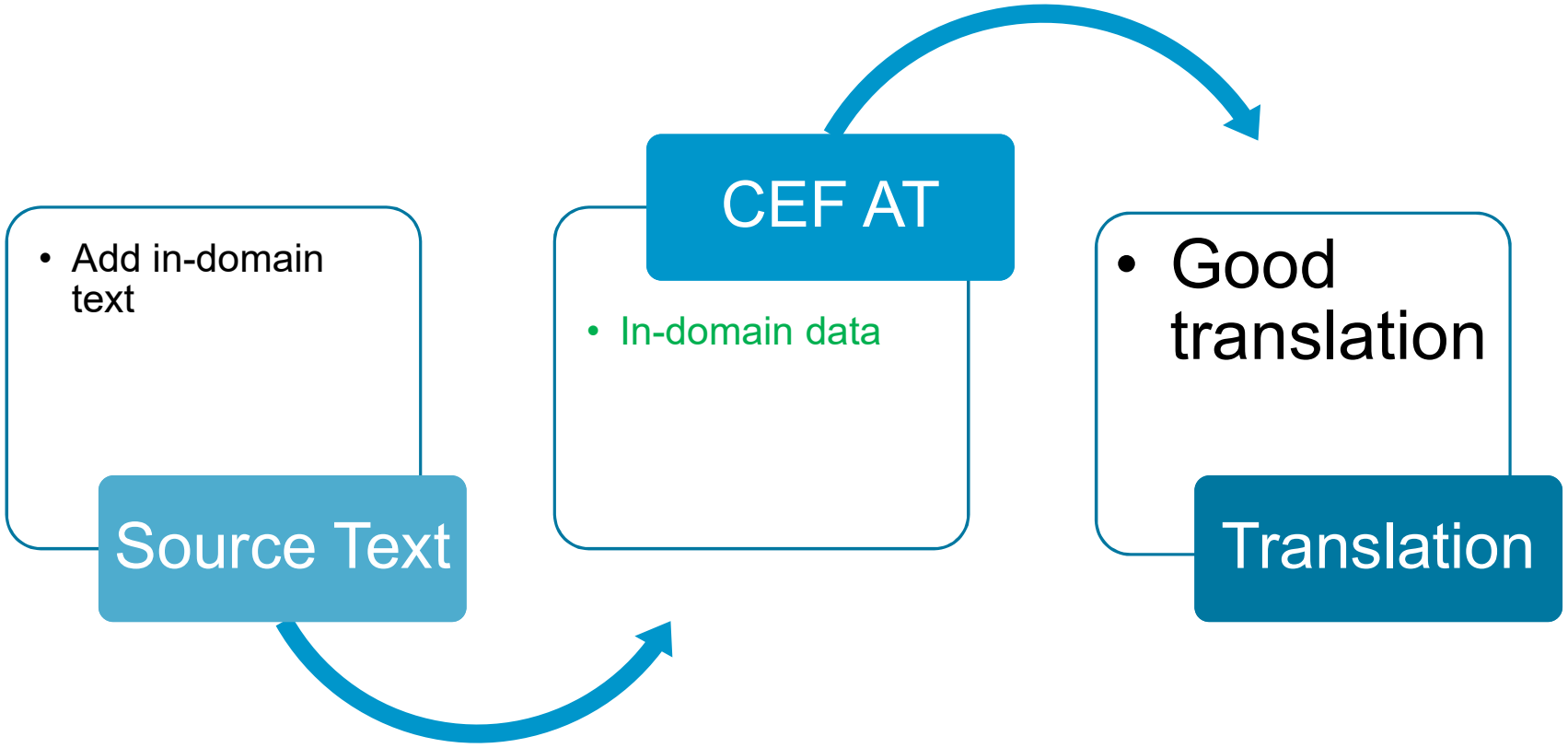
What does ELRC do?

Collect	• Language resources
Identify	• Needs of public services
Engage	• With the public sector in the identification of LR
Help	• With any technical or legal issues
Act	• Observatory for language resources across Europe

European Language Resource Coordination
Connecting Europe Facility

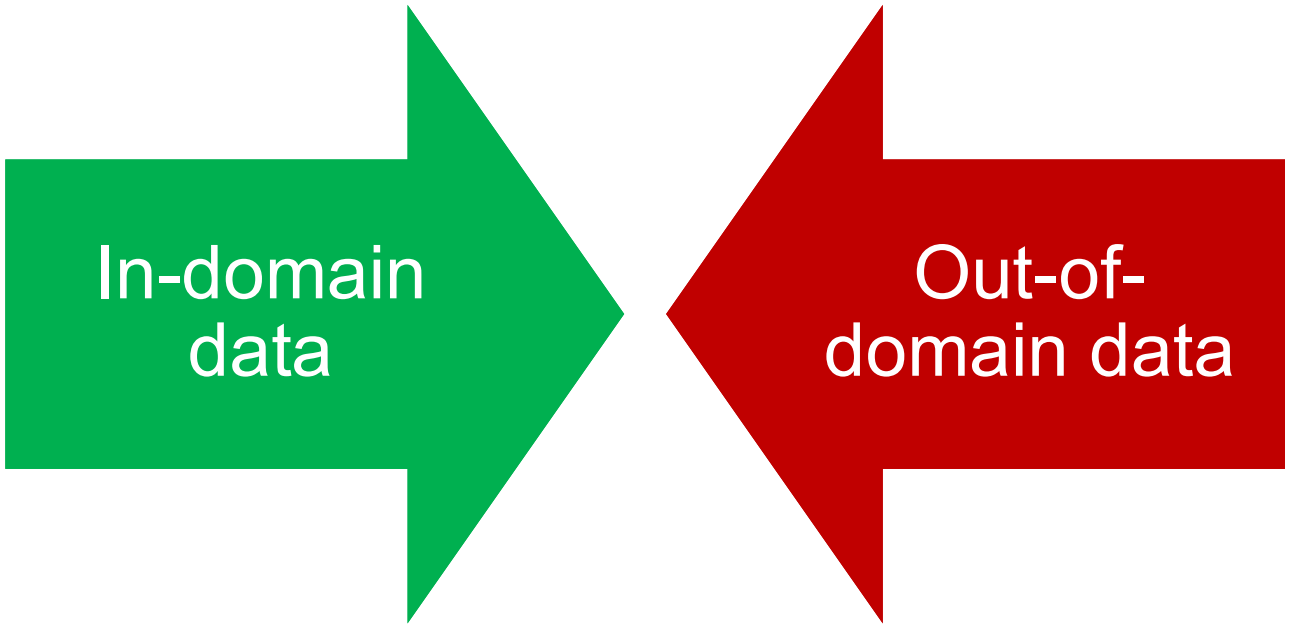
How to make it work…?

With the right language resources!



European Language Resource Coordination
Connecting Europe Facility

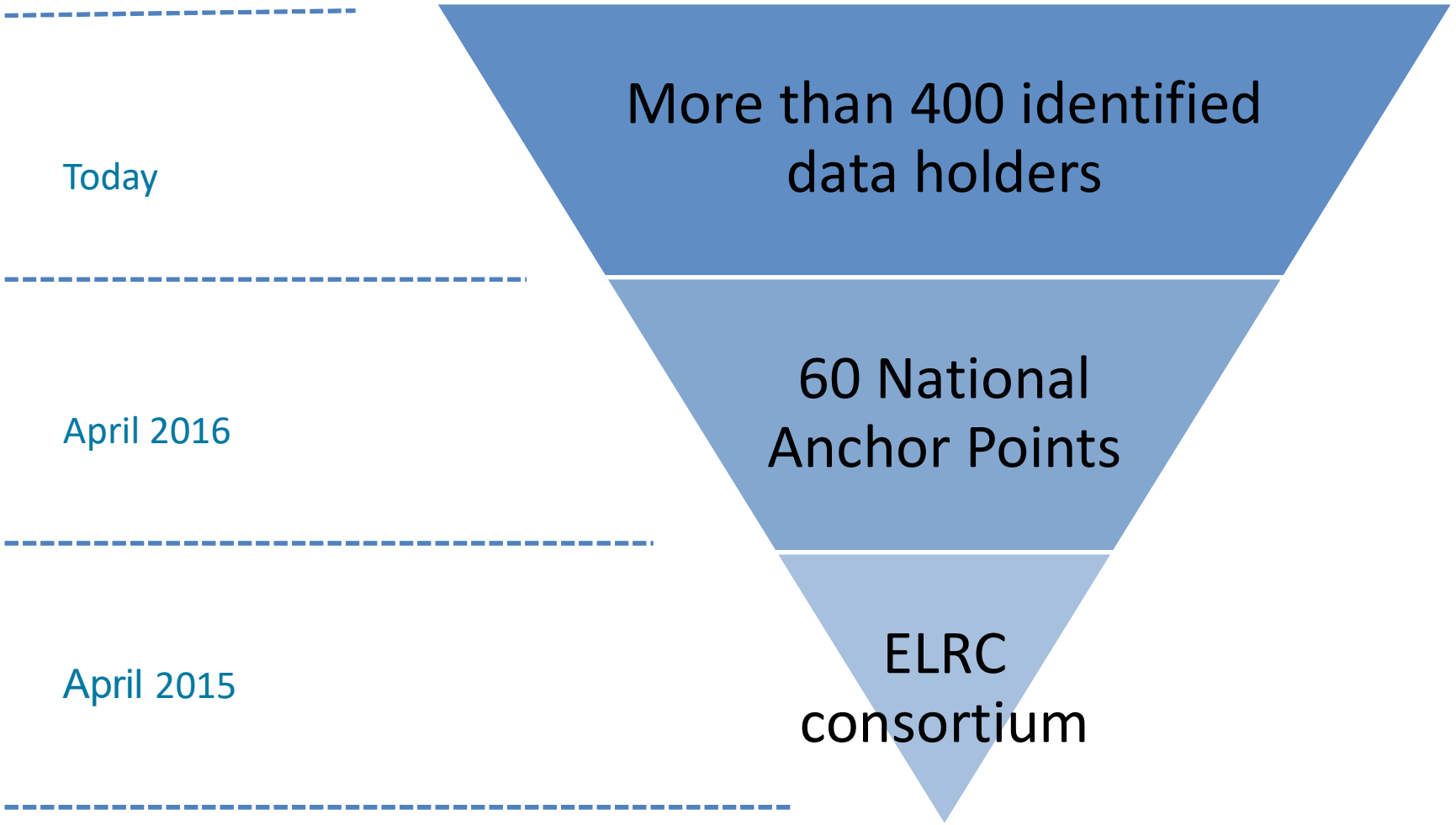
How to make it work…?



➔ **Your language resources make a difference!**

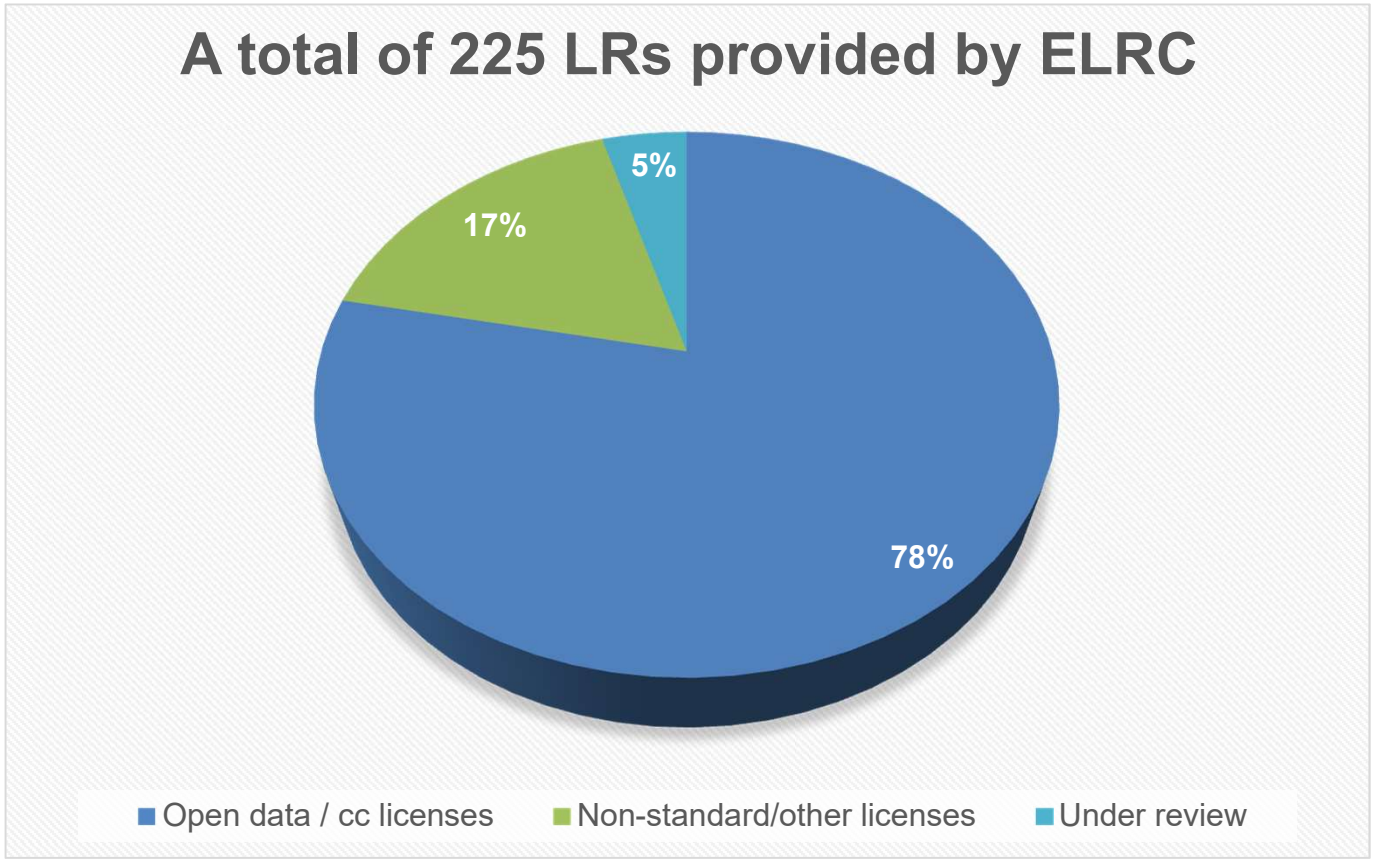
European Language Resource Coordination
Connecting Europe Facility

What has been achieved so far…? The ELRC network



European Language Resource Coordination
Connecting Europe Facility

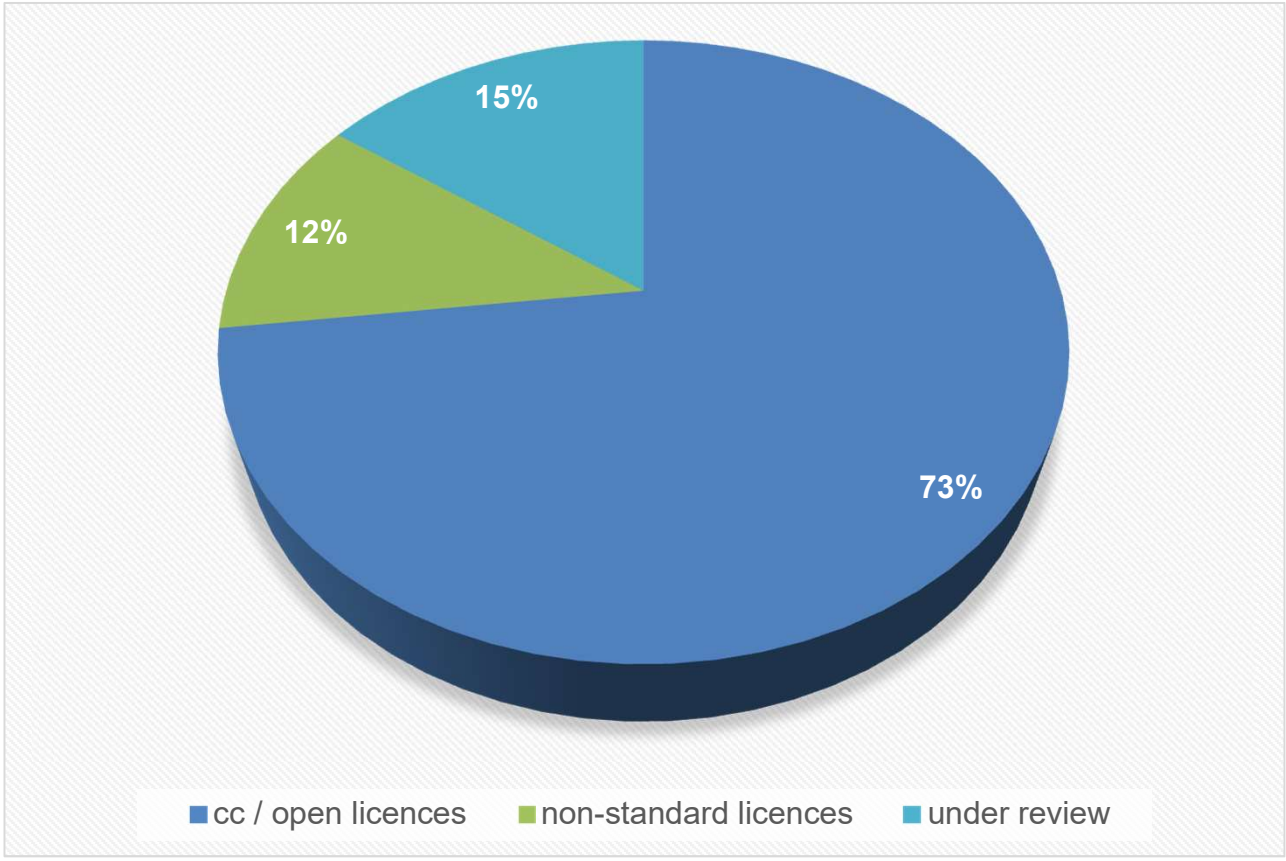
What has been achieved so far…? ELRC Data in April 2017



European Language Resource Coordination
Connecting Europe Facility

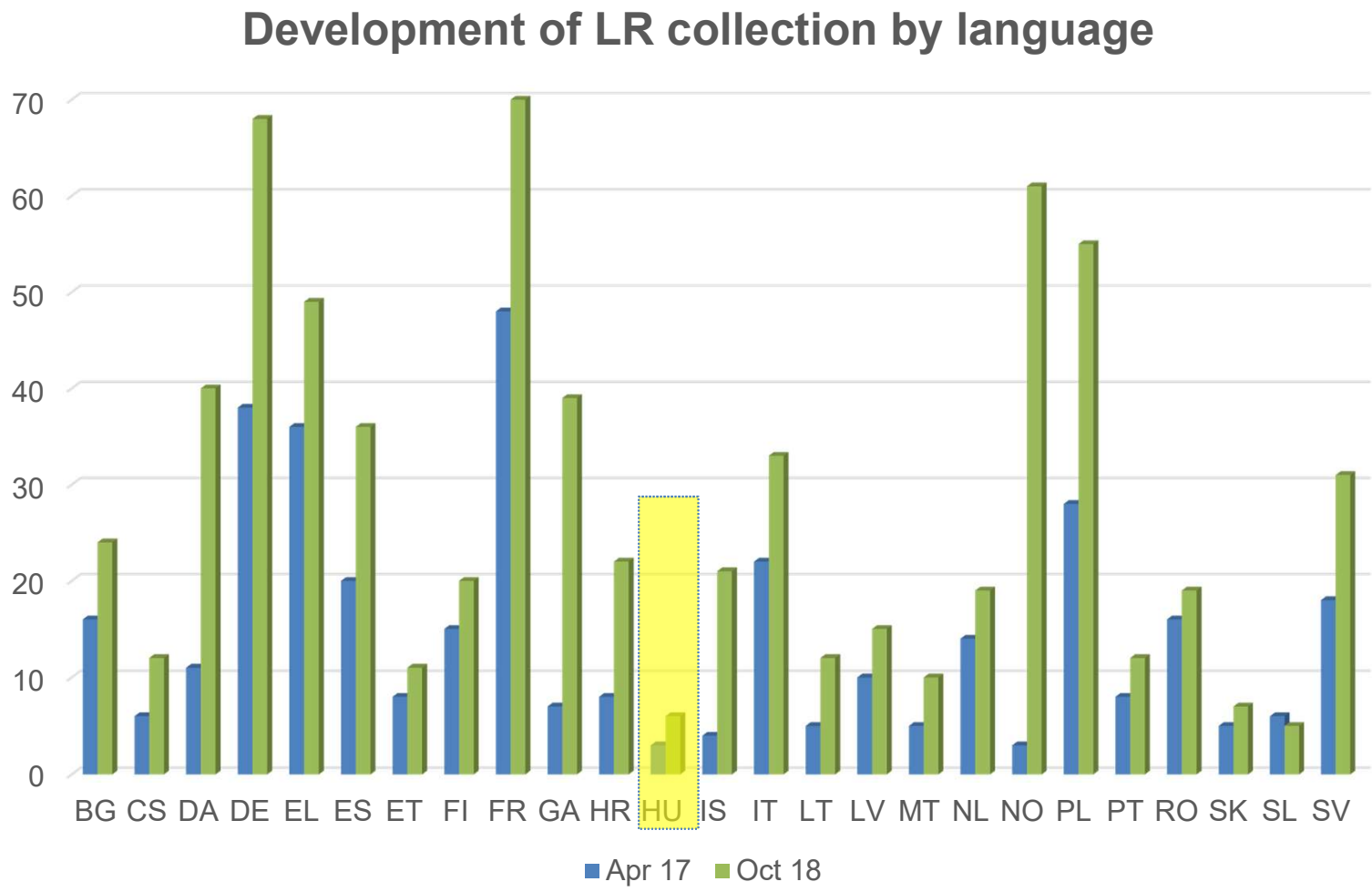
What has been achieved so far…? ELRC Data in October 2018

More than 500 LRs contributed through ELRC



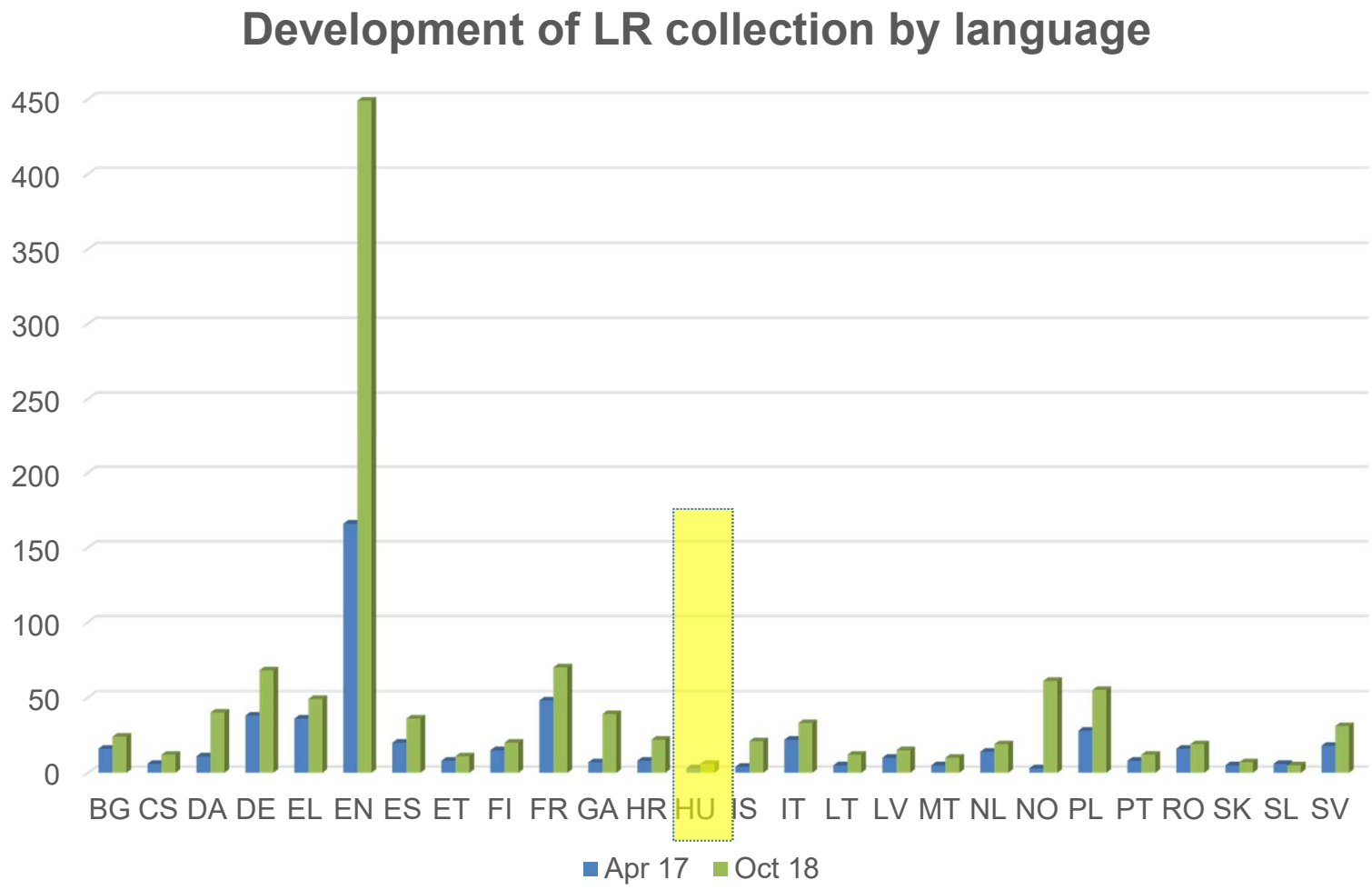
European Language Resource Coordination
Connecting Europe Facility

What has been achieved so far…?
ELRC Data by language



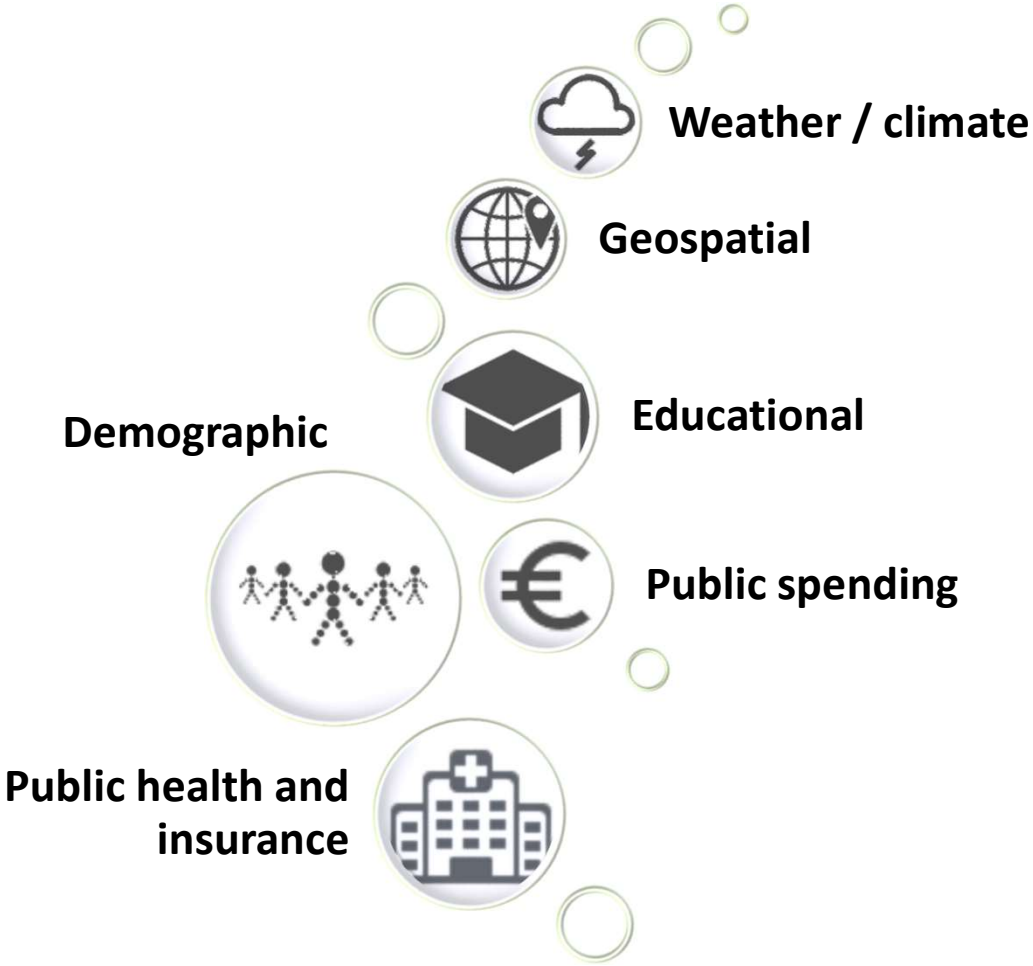
European Language Resource Coordination
Connecting Europe Facility

What has been achieved so far…?
ELRC Data by language



European Language Resource Coordination
Connecting Europe Facility

The notion of data



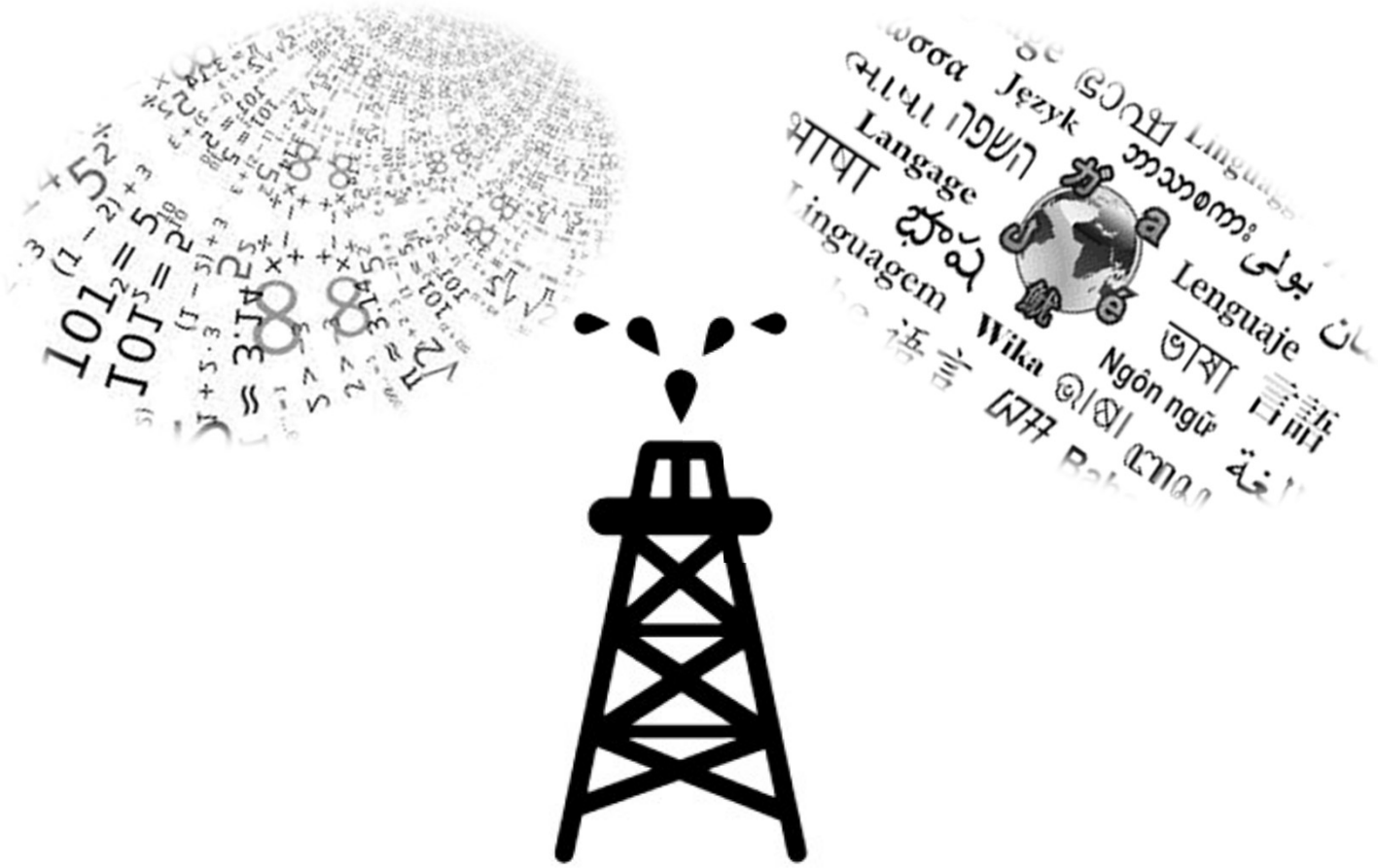
European Language Resource Coordination
Connecting Europe Facility

The notion of data

Data: the oil of the 21st century

European Language Resource Coordination
Connecting Europe Facility

The notion of data

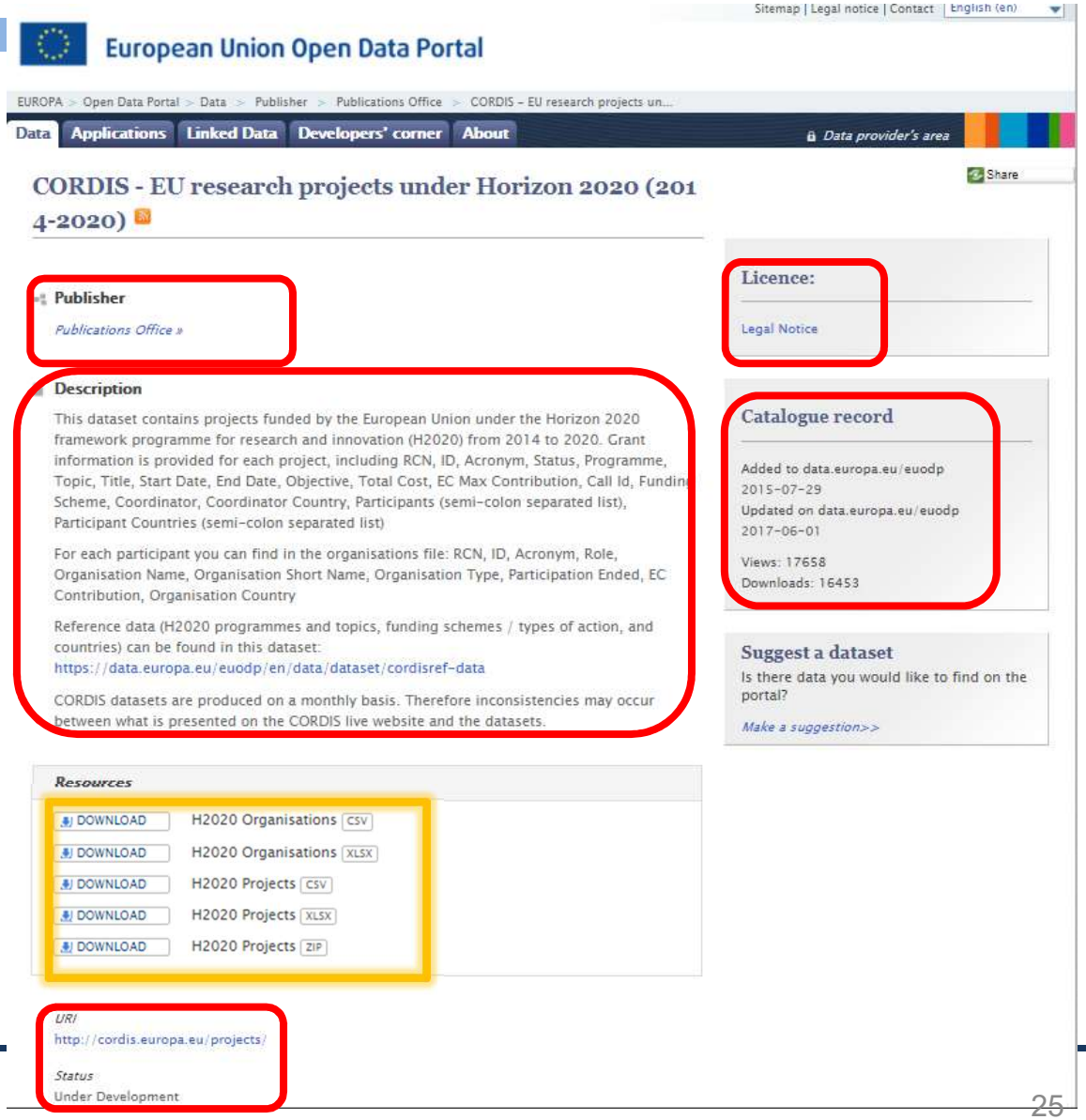


European Language Resource Coordination
Connecting Europe Facility

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)



European Language Resource Coordination
Connecting Europe Facility

Language data (resources) for eTranslation through ELRC

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

HU

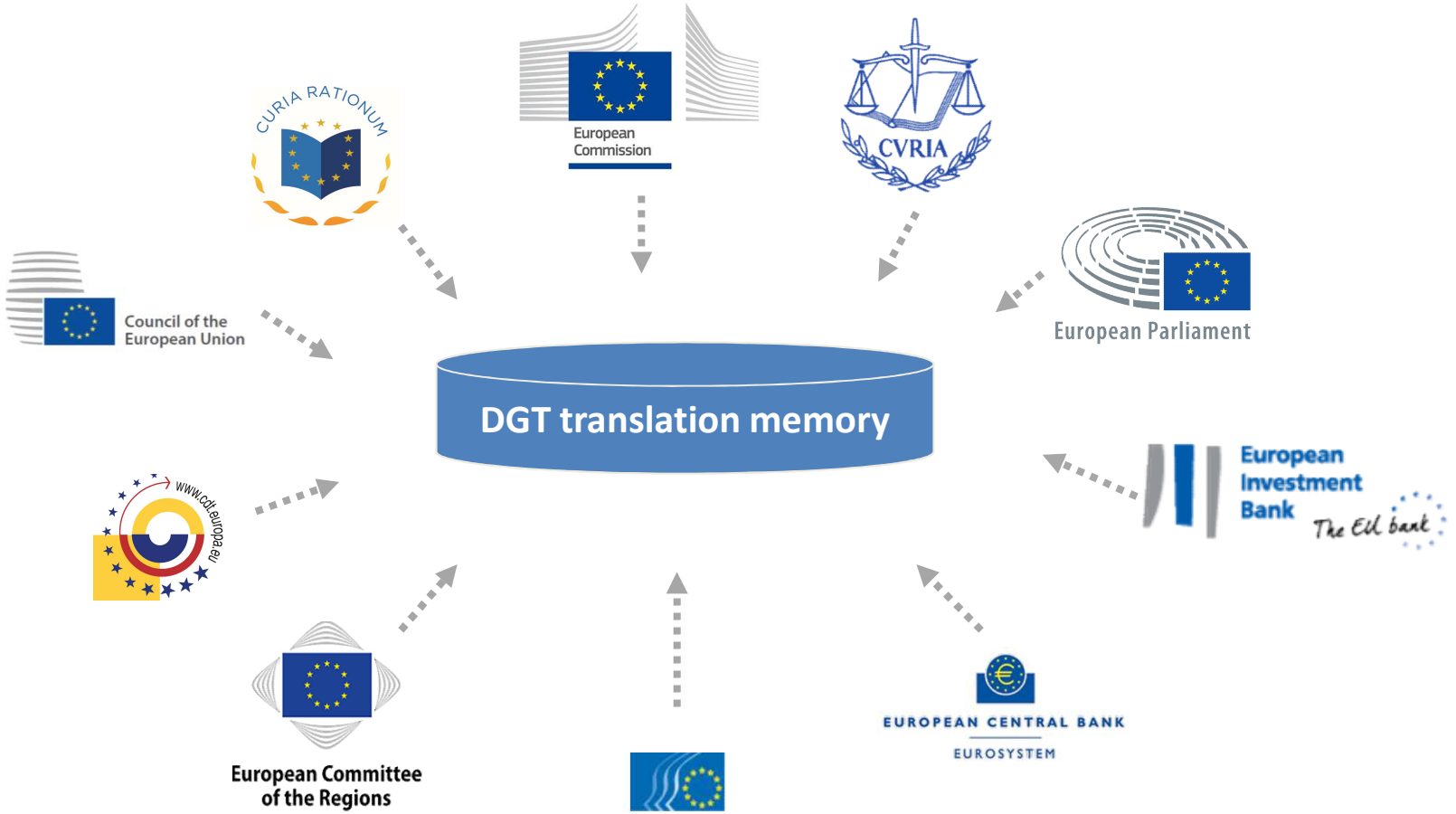
Az emberi jogokra vonatkozó nemzetközi törvény abban az alapvetésben gyökerezik, hogy minden egyes embernek joga van az Emberi Jogok Egyetemes Nyilatkozatában lefektetett minden joghoz és szabadsághoz. Az emberi jogok szexuális orientációtól vagy a nemi azonosságtól függetlenül minden egyes emberre egyformán vonatkoznak.

EN

International human rights law is grounded on the premise that all individuals are entitled to the same rights and freedoms as set out in the Universal Declaration of Human Rights. Human rights shall apply equally to every human being, regardless of sexual orientation or gender identity.

European Language Resource Coordination
Connecting Europe Facility

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)

Hungarian text

A feladatok harmadik csoportját az Országgyűlés saját felkészülése képezi, amelynek során a Parlament felkészül az elnökség alatt megrendezendő parlamenti bizottsági találkozókra. Ennek irányítására létrejött egy Koordinációs csoport, amelynek vezetője az Országgyűlés Külügyi Hivatalának vezetője, tagjai az EU Főoszály vezetője, a Protokoll Főosztály vezetője, valamint a Főtitkárság, a Sajtóiroda, a Gazdasági Főigazgatóság és az Országgyűlés Kutatószolgálatának képviselői, illetve az Országgyűlés brüsszeli állandó irodájának vezetője.

Translation in English

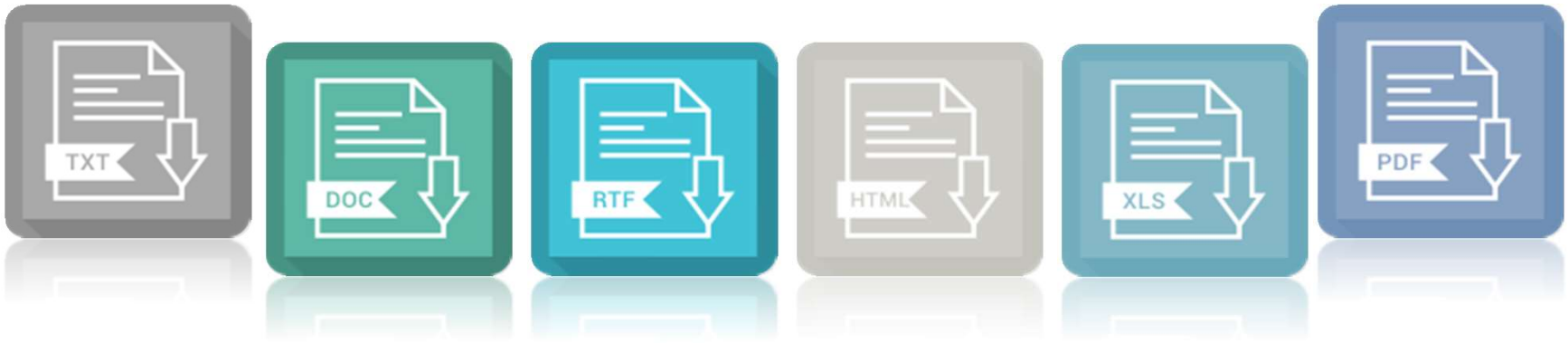
The third group of the preparatory tasks is the National Assembly's own preparation, where the Parliament prepares for the meetings of the parliamentary committees to be held during its presidency. To coordinate these meetings, a Coordination Group was established that is led by the head of the Office for Foreign Relations of the National Assembly, its members are the head of the EU Department, the head of the Protocol Department and the representatives of the General Secretariat, the Press Office, the Economic Directorate and the Research Service of the National Assembly, as well as the Permanent Representative of the National Assembly to the EU in Brussels.

What data are useful for eTranslation as per type |2

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

English	Hungarian
accident insurance	balesetbiztosítás
health insurance	betegbiztosítás, egészségbiztosítás
household insurance	háztartási biztosítás
insurance broker	biztosítási ügynök
medical insurance	betegbiztosítás
sickness insurance	betegbiztosítás
unemployment insurance	munkanélküliségi biztosítás
health insurance card	egészségbiztosítási kártya
nursing care insurance	gondozási biztosítás
…	…

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

A pokoli komédia még egyre tartott a börzén. A halálra ítélt papírok, a bondavári gyártelep s a bondavári vasút részvényei egyik kézből a másikba repültek. Most már a komikumig vitték a tragédiát, s kezdett humor vegyülni a szerencsétlenségbe. Ez a szó: "Bondavár" csak arra való volt, hogy derültséget idézzen elő a börzeemberek között. Aki az utolsó részvényén túladhatott nagy veszteséggel, nevetett azon, aki megszerezte azt. Kezdték a részvényeket megfoghatlan becsű tárgyakért cserébe kínálgatni. Ráadásul egy új esernyőre egy ócska esernyőért. Használták előfizetési ekvivalensül olyan lapokért, amiket valakinek a szerkesztő a nyakára köt erővel. Ajándékozgatták jótékony célokra.

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.

Ez a fenti bekezdés magyar fordítása. Ez a fenti bekezdés magyar fordítása. Ez a fenti bekezdés magyar fordítása. Ez a fenti bekezdés magyar fordítása. Ez a fenti bekezdés magyar fordítása. Ez a fenti bekezdés magyar fordítása. Ez a fenti bekezdés magyar fordítása. Ez a fenti bekezdés magyar fordítása. Ez a fenti bekezdés magyar fordítása.

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.

Ez a második bekezdés magyar fordítása. Ez a második bekezdés magyar fordítása. Ez a második bekezdés magyar fordítása. Ez a második bekezdés magyar fordítása. Ez a második bekezdés magyar fordítása. Ez a második bekezdés magyar fordítása. Ez a második bekezdés magyar fordítása. Ez a második bekezdés magyar fordítása.

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.·¶

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.¶

¶

Ez·a·bal·oldali·bekezdés·magyar·fordítása.·Ez·a·bal·oldali·bekezdés·magyar·fordítása.·Ez·a·bal·oldali·bekezdés·magyar·fordítása.·Ez·a·bal·oldali·bekezdés·magyar·fordítása.·Ez·a·bal·oldali·bekezdés·magyar·fordítása.·Ez·a·bal·oldali·bekezdés·magyar·fordítása.¶

¶

Ez·a·mondat·magyar·fordítása·balra.·¶

¶

Ez·a·bal·oldali·bekezdés·magyar·fordítása.·Ez·a·bal·oldali·bekezdés·magyar·fordítása.·Ez·a·bal·oldali·bekezdés·magyar·fordítása.·Ez·a·bal·oldali·bekezdés·magyar·fordítása.·Ez·a·bal·oldali·bekezdés·magyar·fordítása.·Ez·a·bal·oldali·bekezdés·magyar·fordítása.¶

European Language Resource Coordination
Connecting Europe Facility

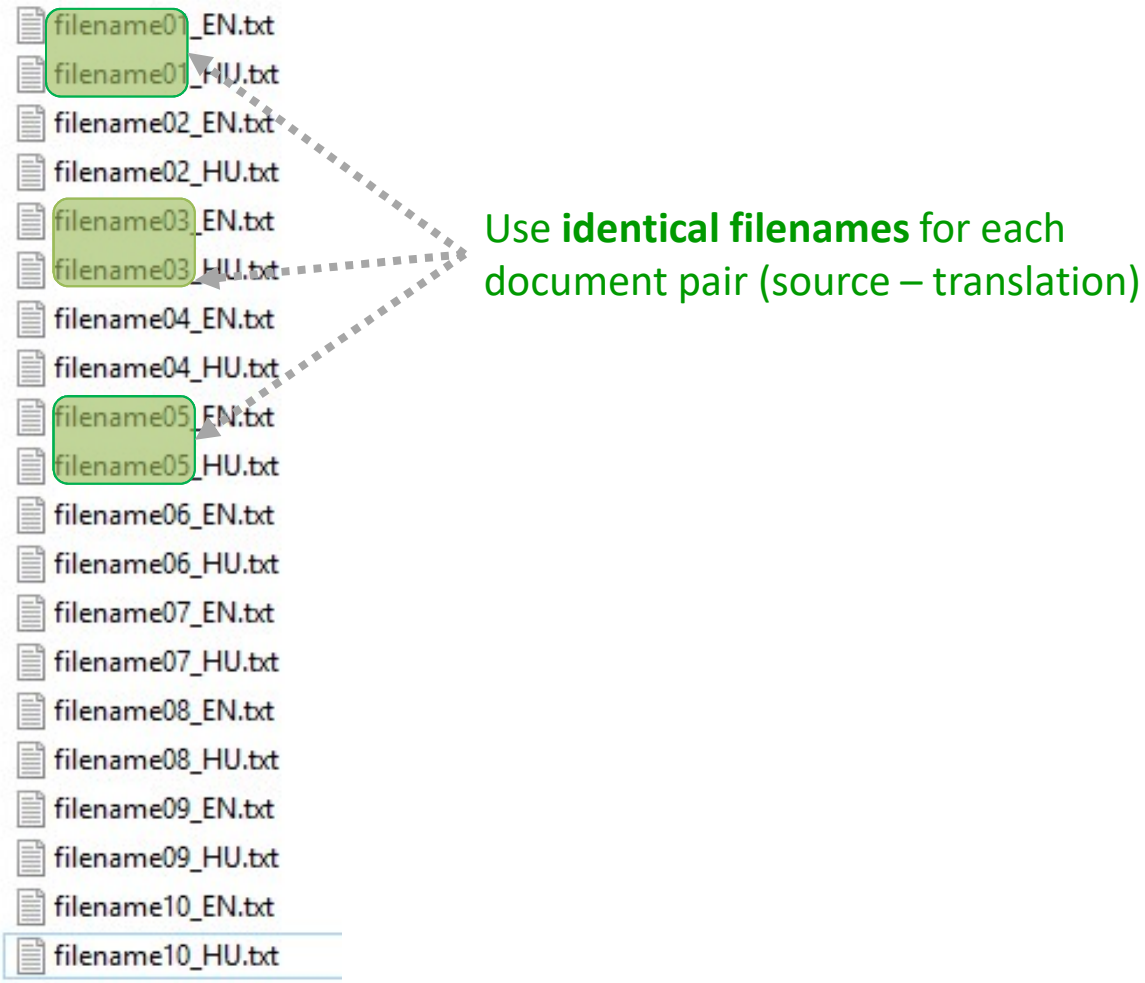
Preparing your data |3

Don’ts

English	Magyar
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.	Ez a bal oldali bekezdés magyar fordítása. Ez a bal oldali bekezdés magyar fordítása. Ez a bal oldali bekezdés magyar fordítása. Ez a bal oldali bekezdés magyar fordítása. Ez a bal oldali bekezdés magyar fordítása. Ez a bal oldali bekezdés magyar fordítása.
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.	Ez a bal oldali bekezdés magyar fordítása. Ez a bal oldali bekezdés magyar fordítása. Ez a bal oldali bekezdés magyar fordítása. Ez a bal oldali bekezdés magyar fordítása. Ez a bal oldali bekezdés magyar fordítása. Ez a bal oldali bekezdés magyar fordítása.

Preparing your data |4

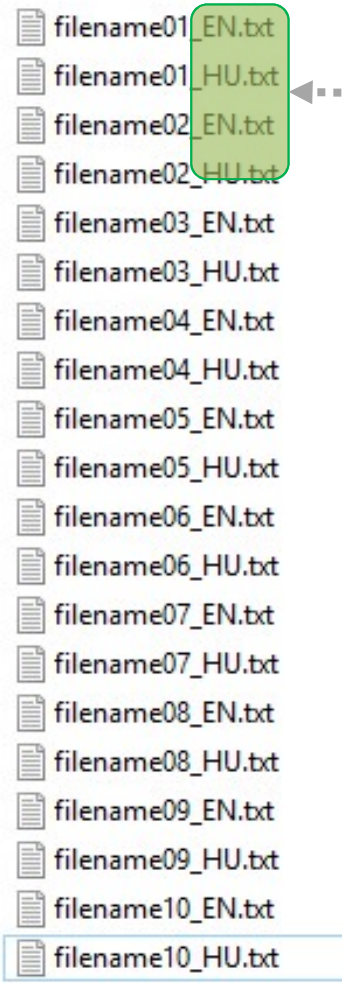
Do's



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

Preparing your data |5

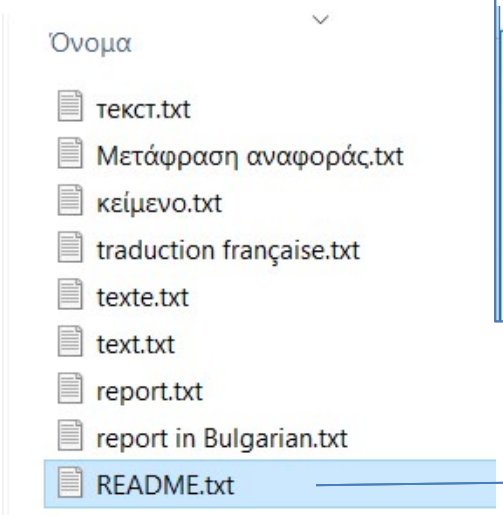
Do's

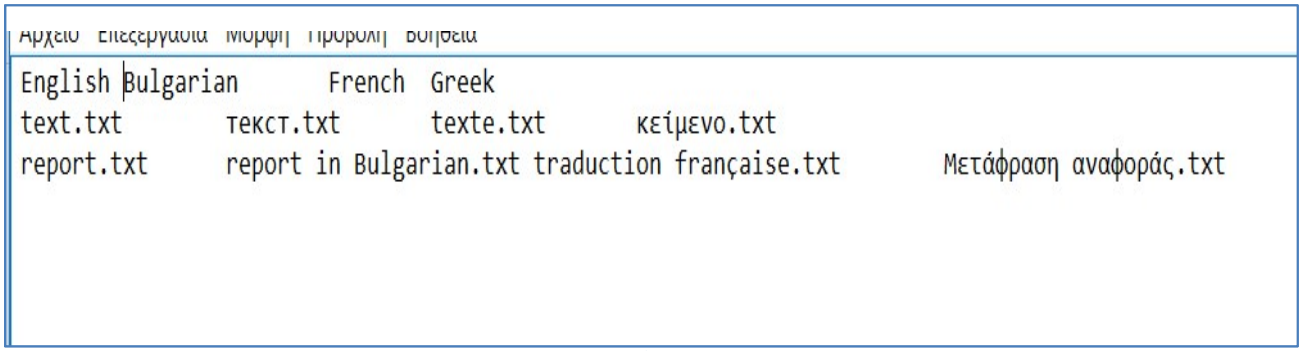


Include **language identifiers** in the filename

Preparing your data |5

Do's





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

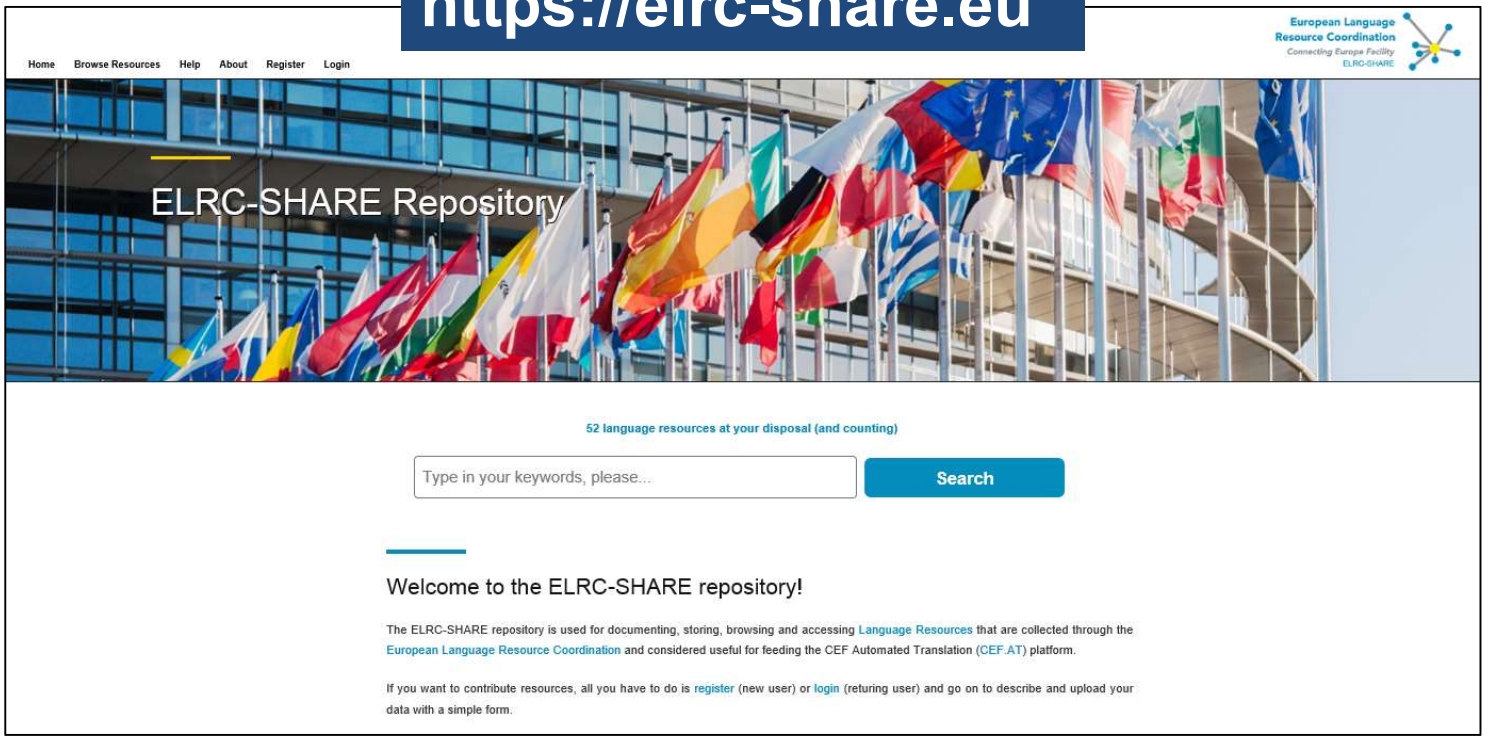
European Language Resource Coordination
Connecting Europe Facility

To ELRC-SHARE

The ELRC-SHARE Repository

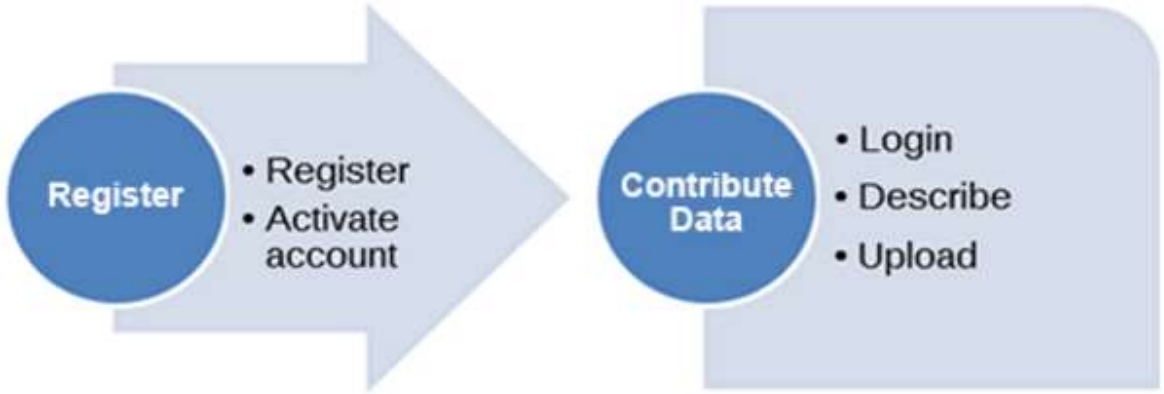
- Access to, sharing and contribution of language resources
- Access to tools and services catalogue (upcoming)

https://elrc-share.eu



European Language Resource Coordination
Connecting Europe Facility

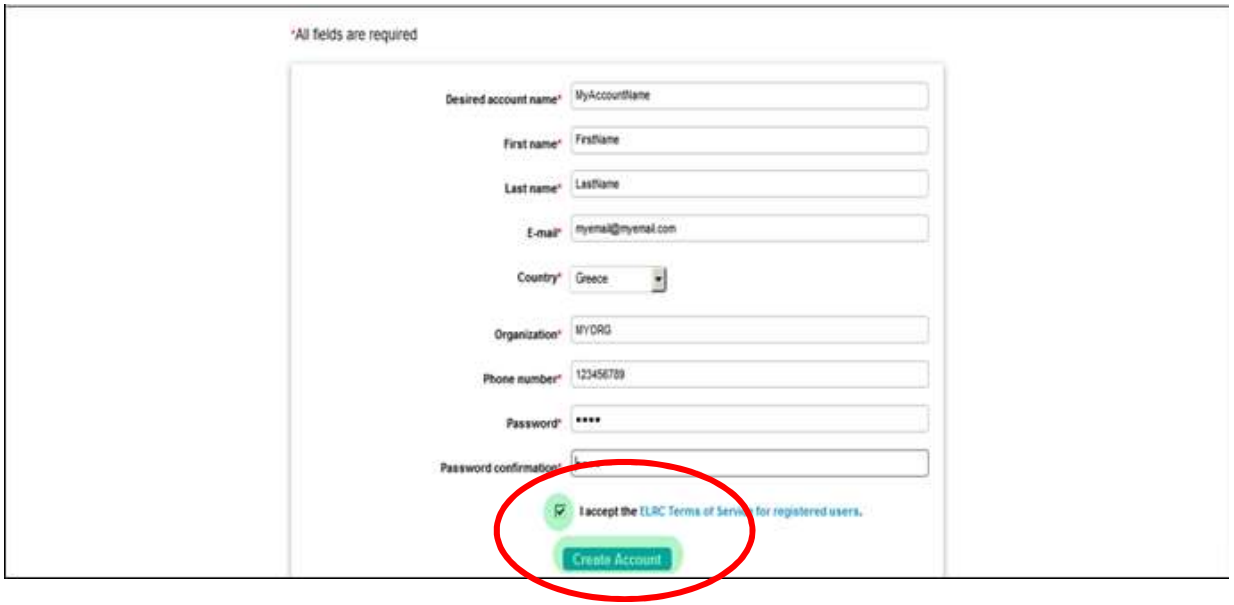
How to Contribute Data



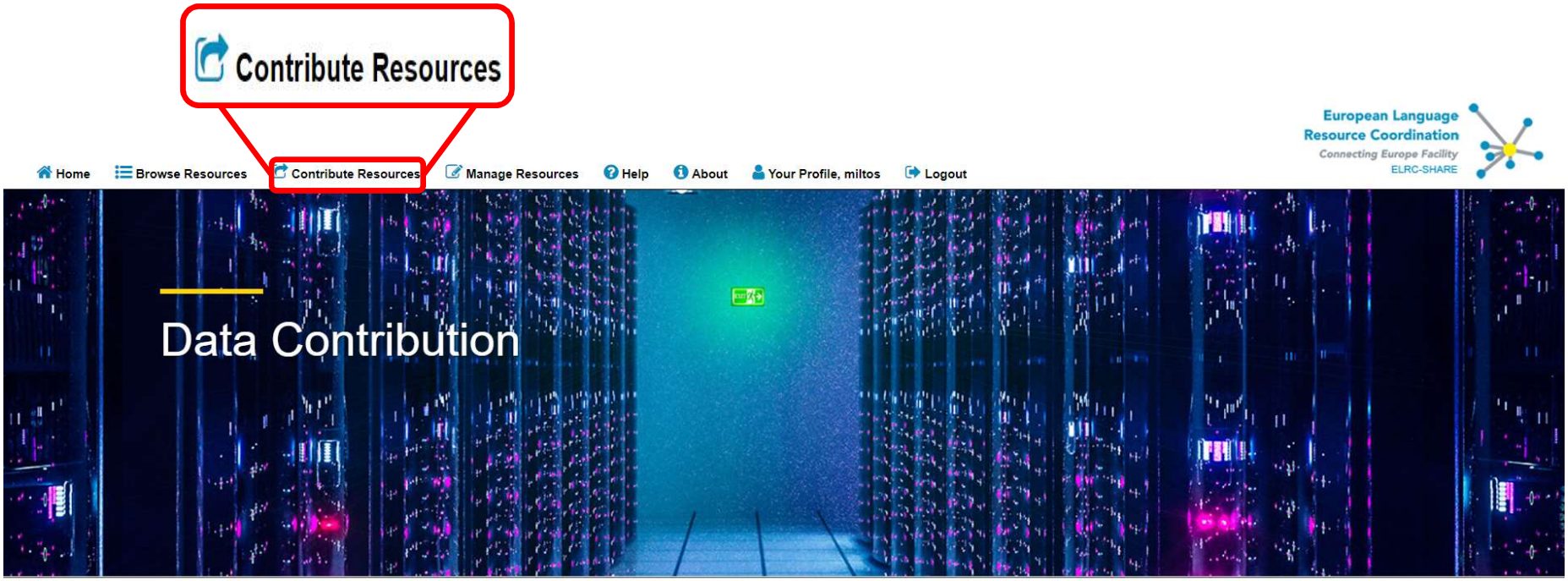
How to Register (1/2)

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)



New Resource

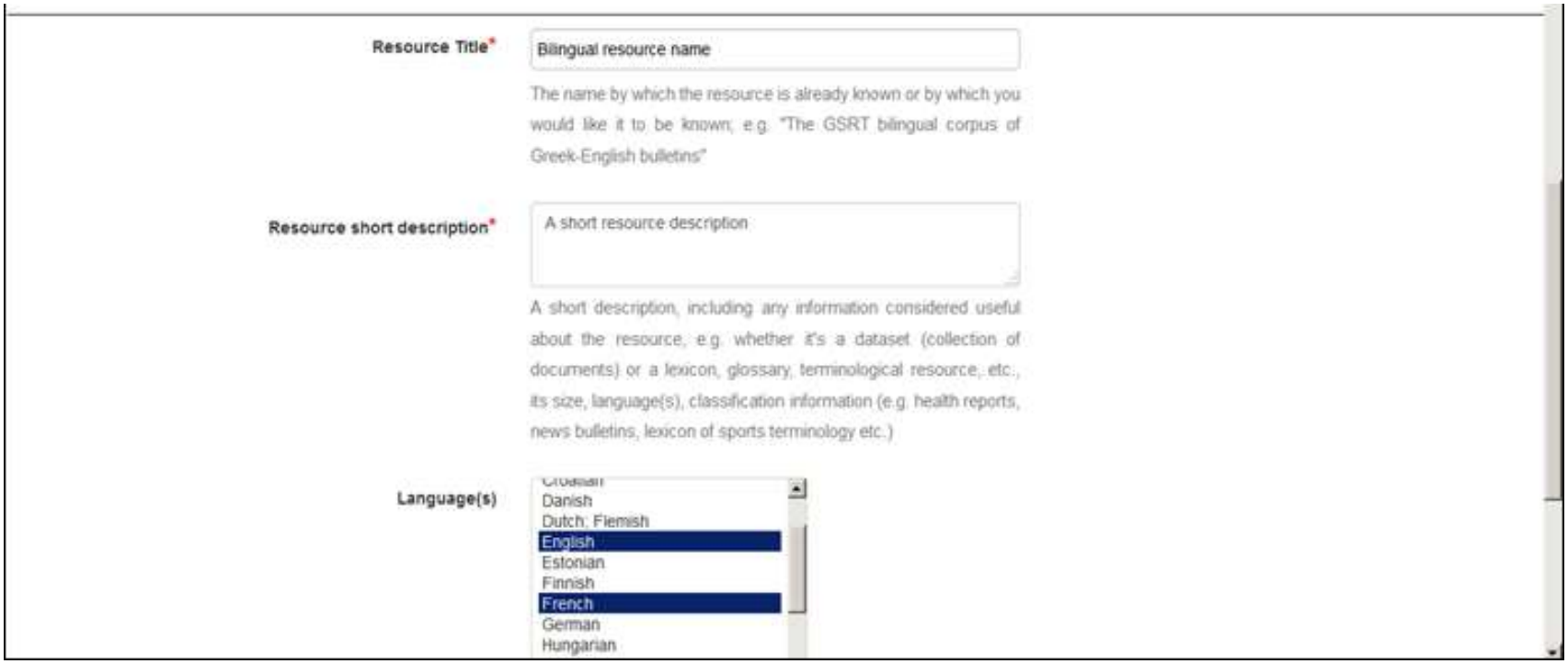
Resource Title*

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"

European Language Resource Coordination
Connecting Europe Facility

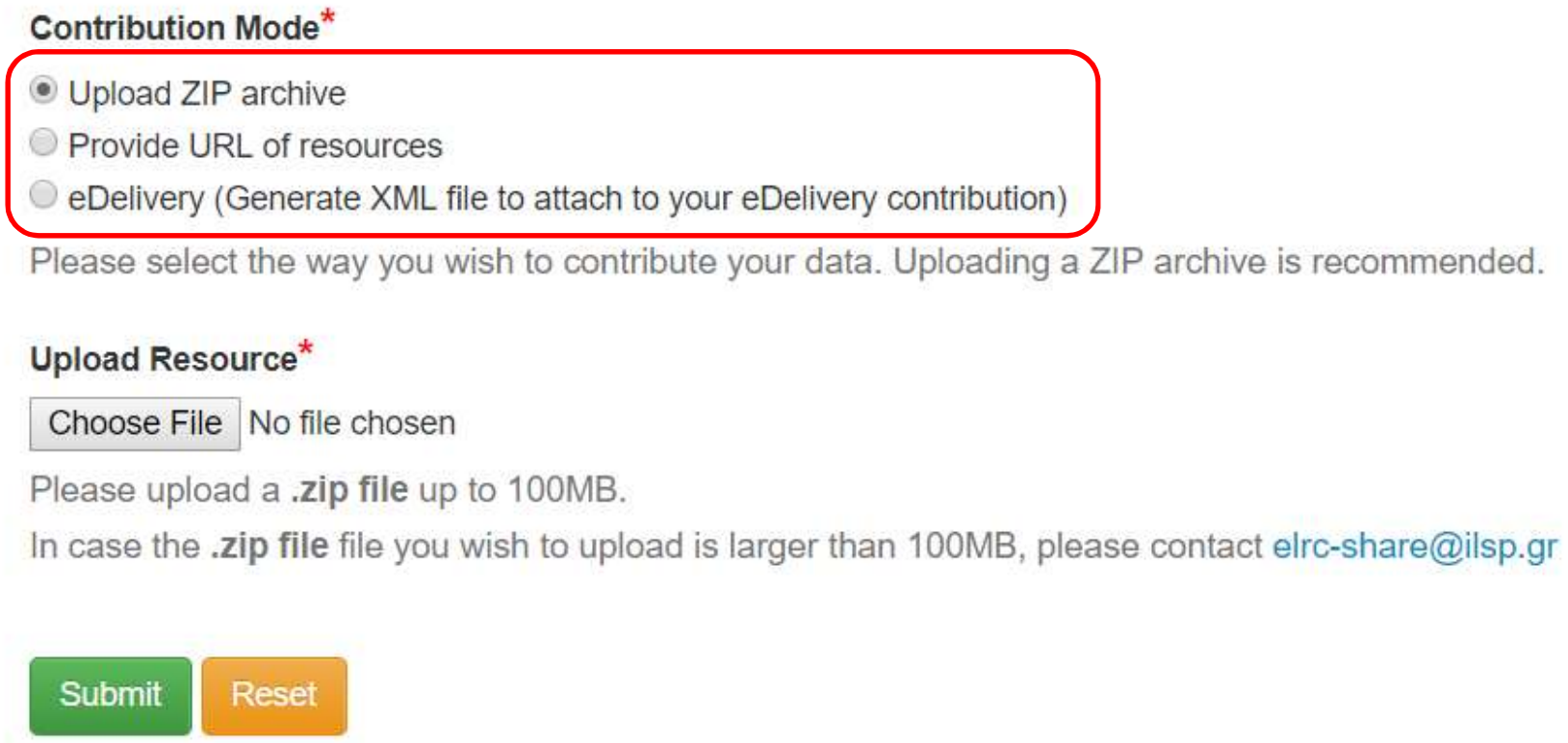
How to Contribute Data (2/6)

- Fill in the details of the dataset



How to Contribute Data (3/6)

- Three modes for contributing your data



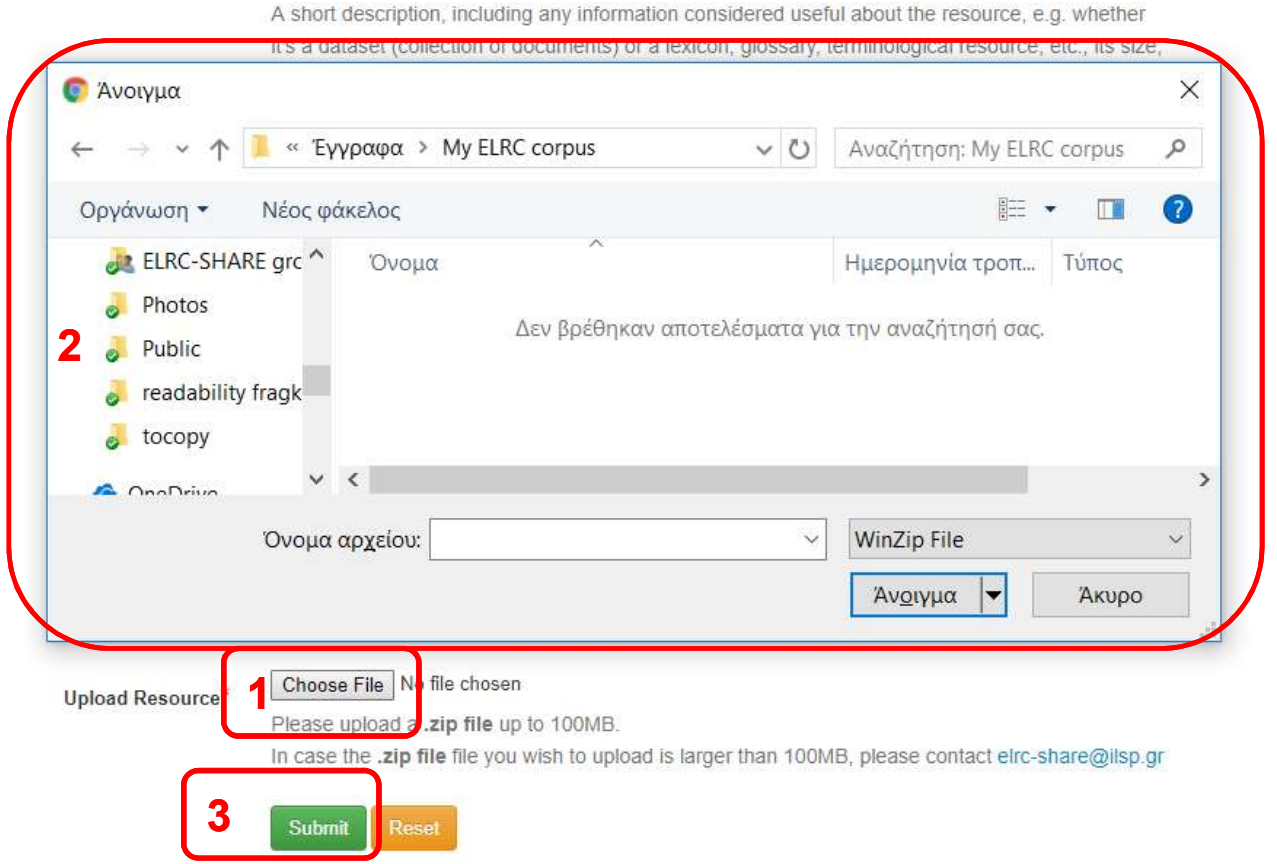
European Language Resource Coordination
Connecting Europe Facility

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)

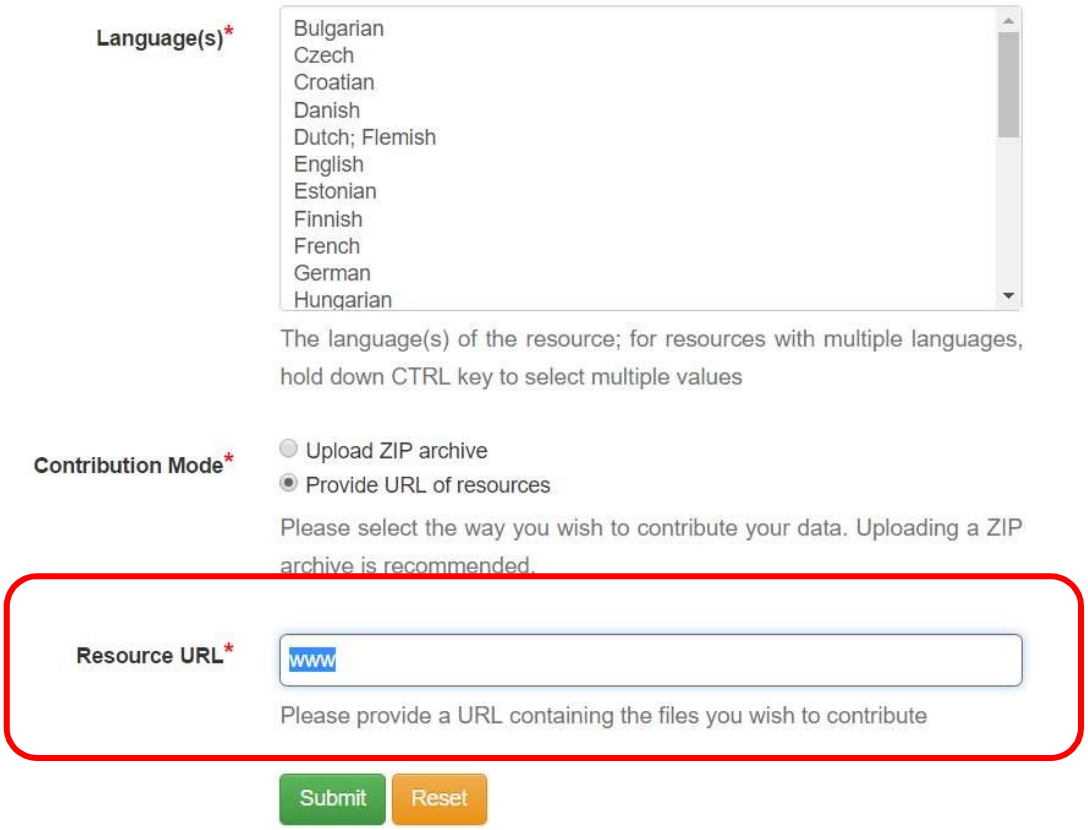
1. Click on Choose file
2. Locate your resource in your hard disk
3. Click on Submit



European Language Resource Coordination
Connecting Europe Facility

How to Contribute Data (5/6)

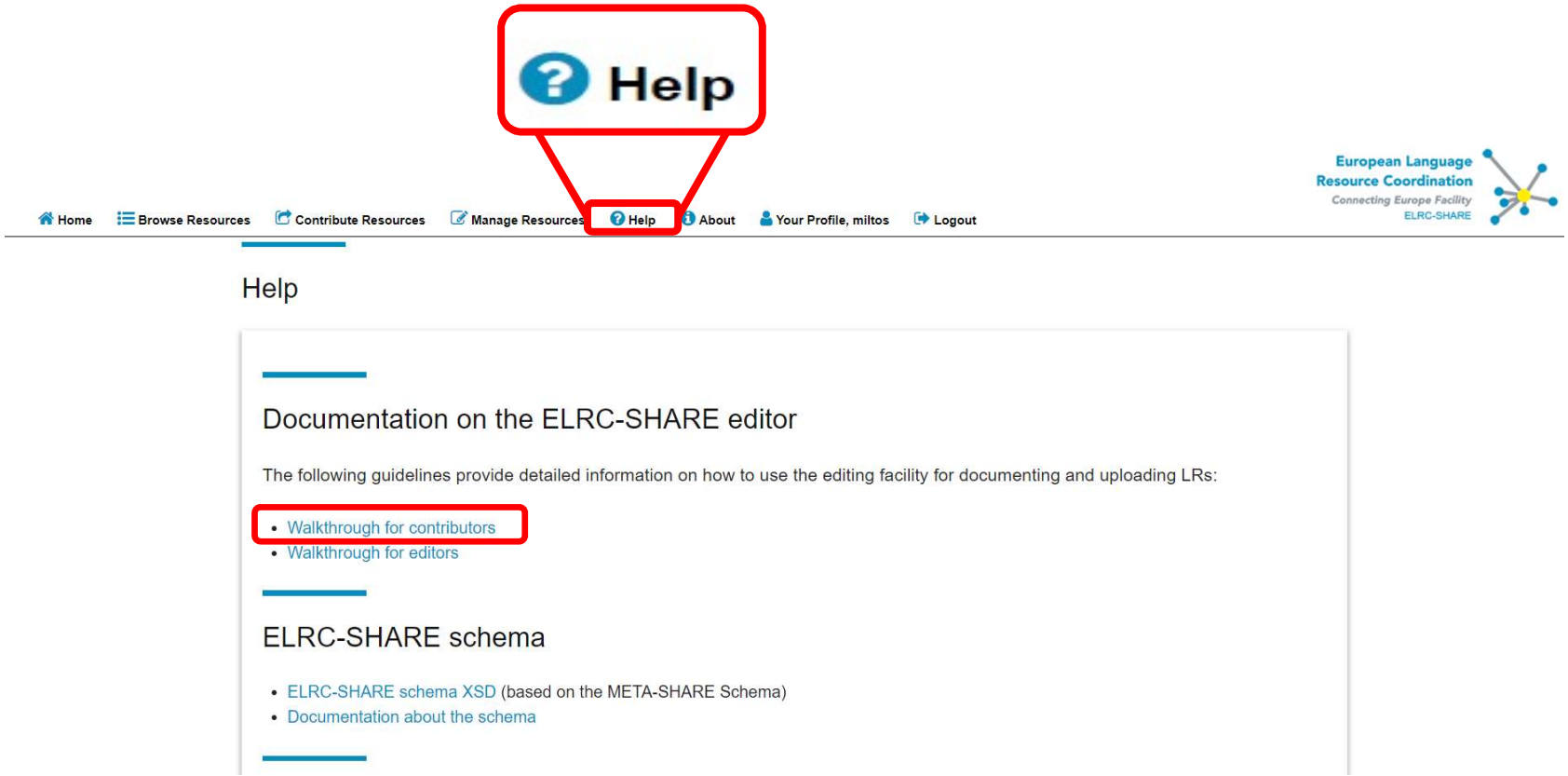
- Alternatively indicate a url (directory listing)



How to Contribute Data (6/6)

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

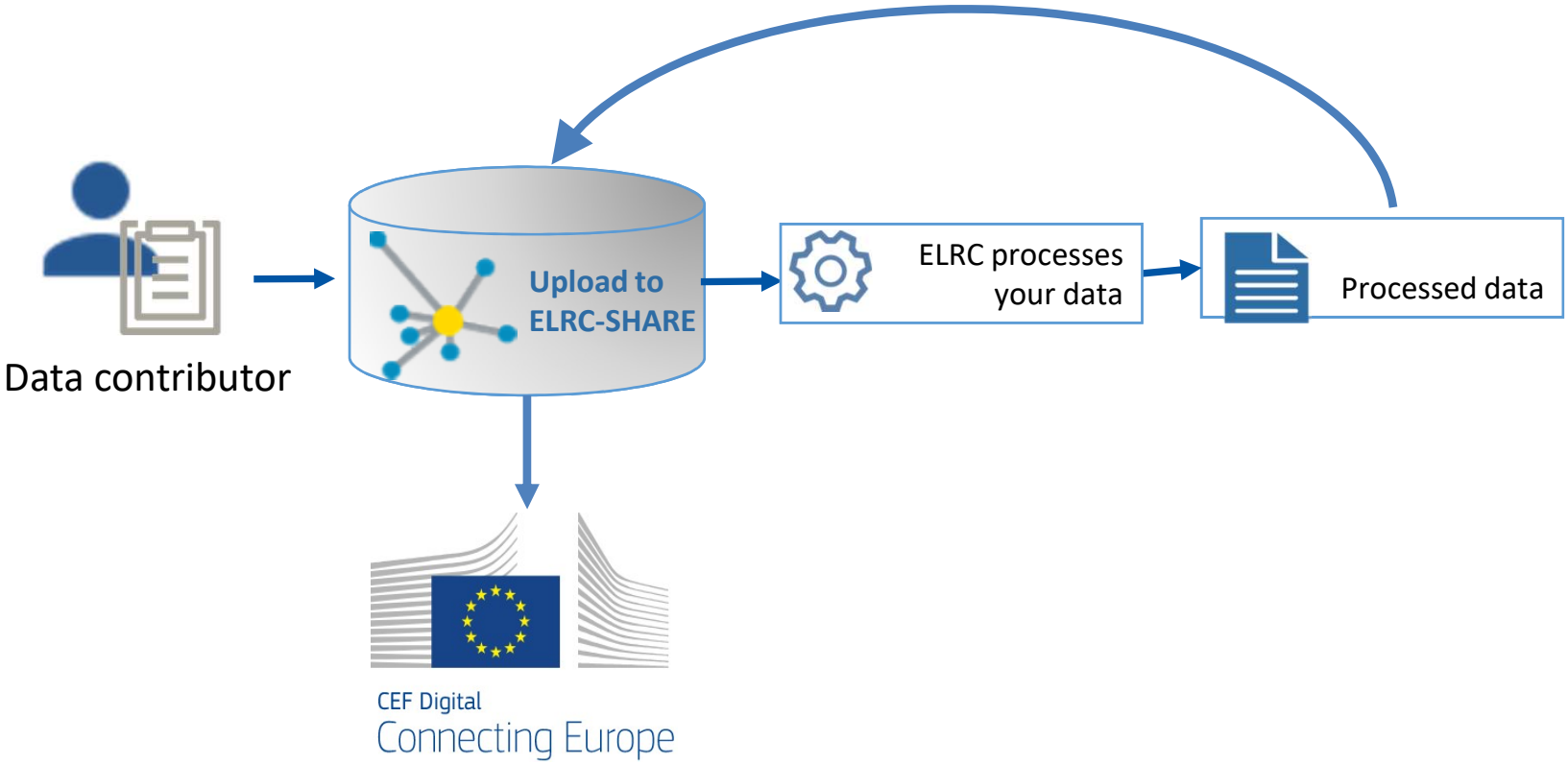
Guidelines for contributors



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

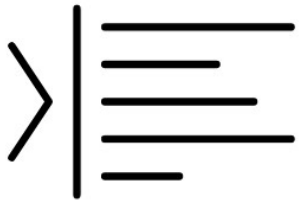
If your data is trapped in archives and databases, we can help extract it

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

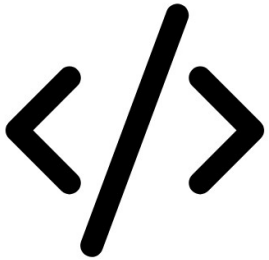
Need to re-format DOCX to XML, or PDF to WORD? Let us do it for you!

European Language Resource Coordination
Connecting Europe Facility

Language processing services

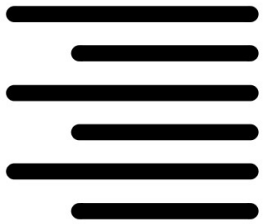
Data conversion

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



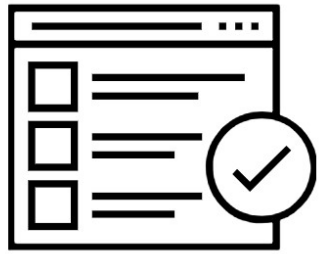
Tag removal

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?

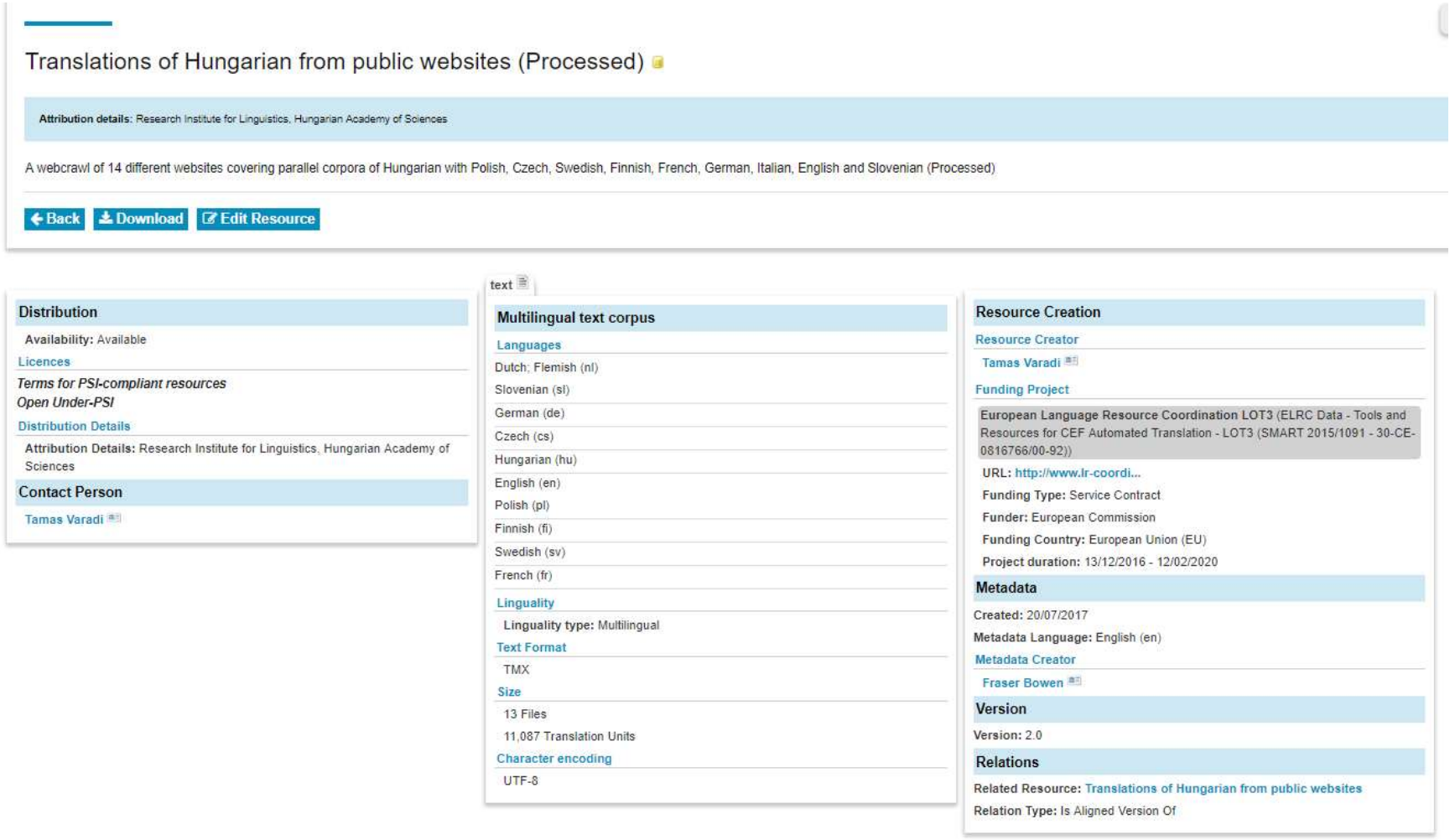
File01_hu.txt
File01_en.doc
File02_hu.pdf
File02_en.txt
File03_hu.doc
File03_en.doc
…

After processing

```
<tu tuid="3931">
 <tuv xml:lang="en">
  <seg>Weights: they express the typical
percentages the groups of different goods or services
represent within the total expenditures of
households.</seg> </tuv>
 <tuv xml:lang="hu">
  <seg>Súlyarány: kifejezi, hogy az egyes termékek és
szolgáltatások csoportjai a lakosság összes kiadásán
belül milyen jellemző százalékarányt
képviselnek.</seg></tuv></tu>
<tu tuid="3932">
 <tuv xml:lang="en">
  <seg>Weights are revised annually, based on household
consumption in the year two years before the reference
year.</seg>
 </tuv>
 <tuv xml:lang="hu">
  <seg>A súlyok felülvizsgálata évente történik, a
tárgyévet két évvel megelőző év lakossági fogyasztása
alapján.</seg></tuv></tu>
```

European Language Resource Coordination
Connecting Europe Facility

How your dataset is described



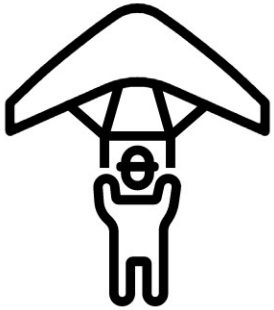
European Language Resource Coordination
Connecting Europe Facility

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

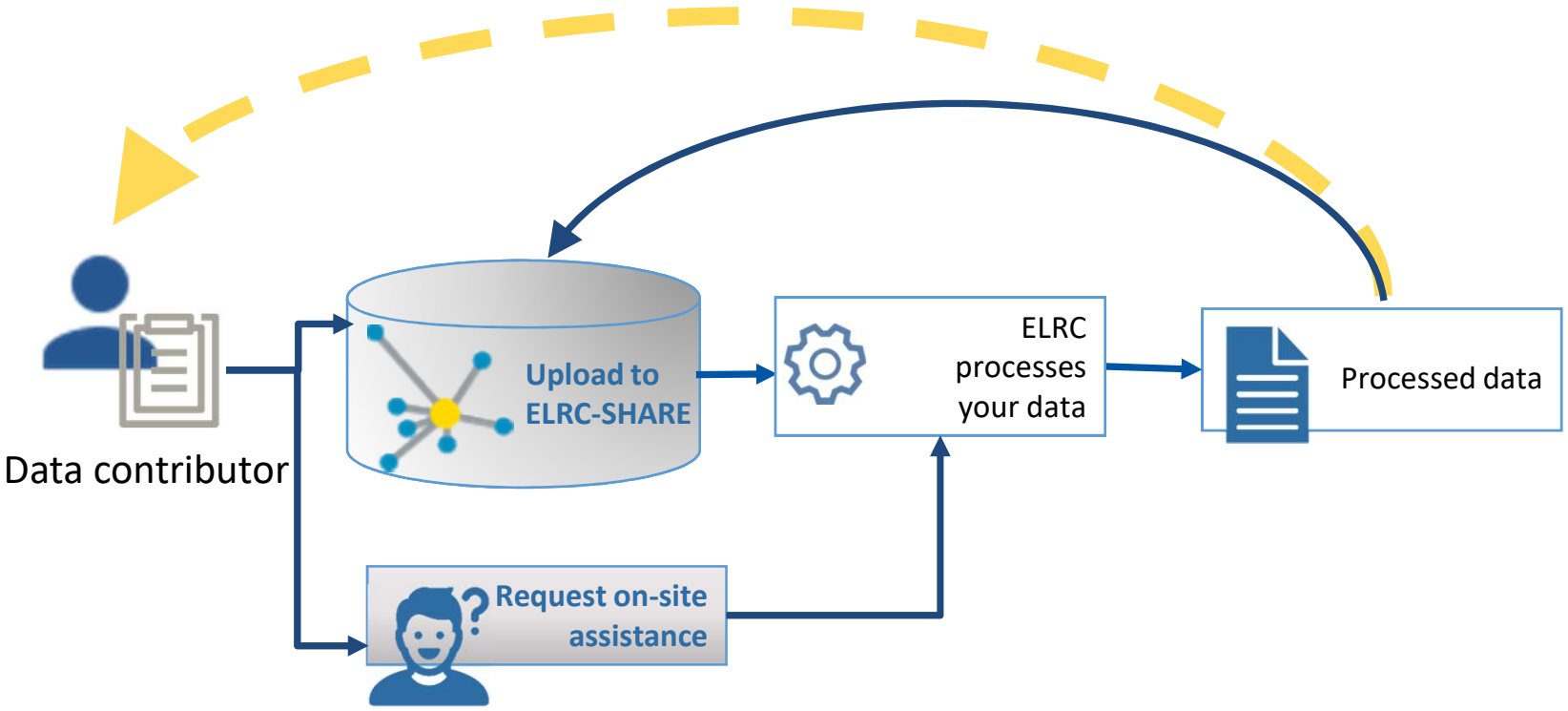
European Language Resource Coordination
Connecting Europe Facility

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?



European Language Resource Coordination
Connecting Europe Facility

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

Köszönöm!

Email: info@lr-coordination.eu

Website: www.lr-coordination.eu

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