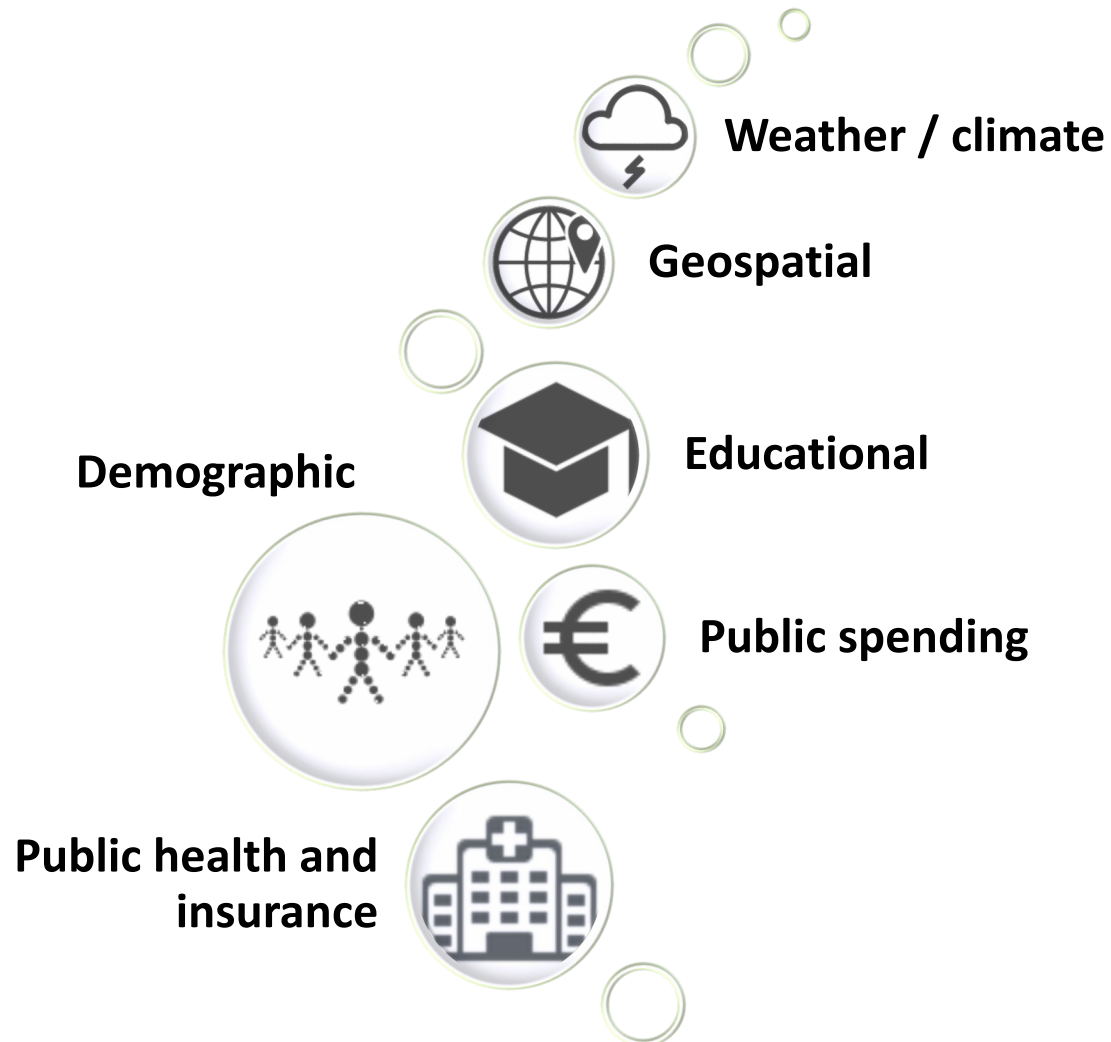


# Preparing and sharing data with the ELRC-SHARE repository and what happens next

Maria Giagkou

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

# The notion of data

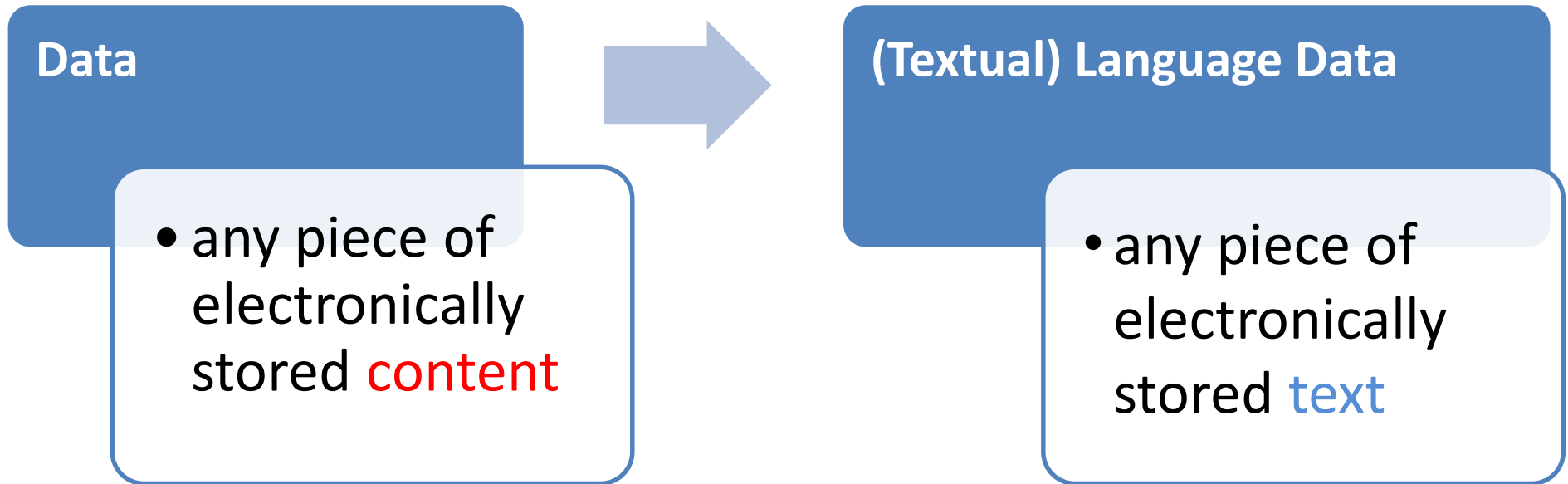


# The notion of data

Data: the oil of the 21<sup>st</sup> century







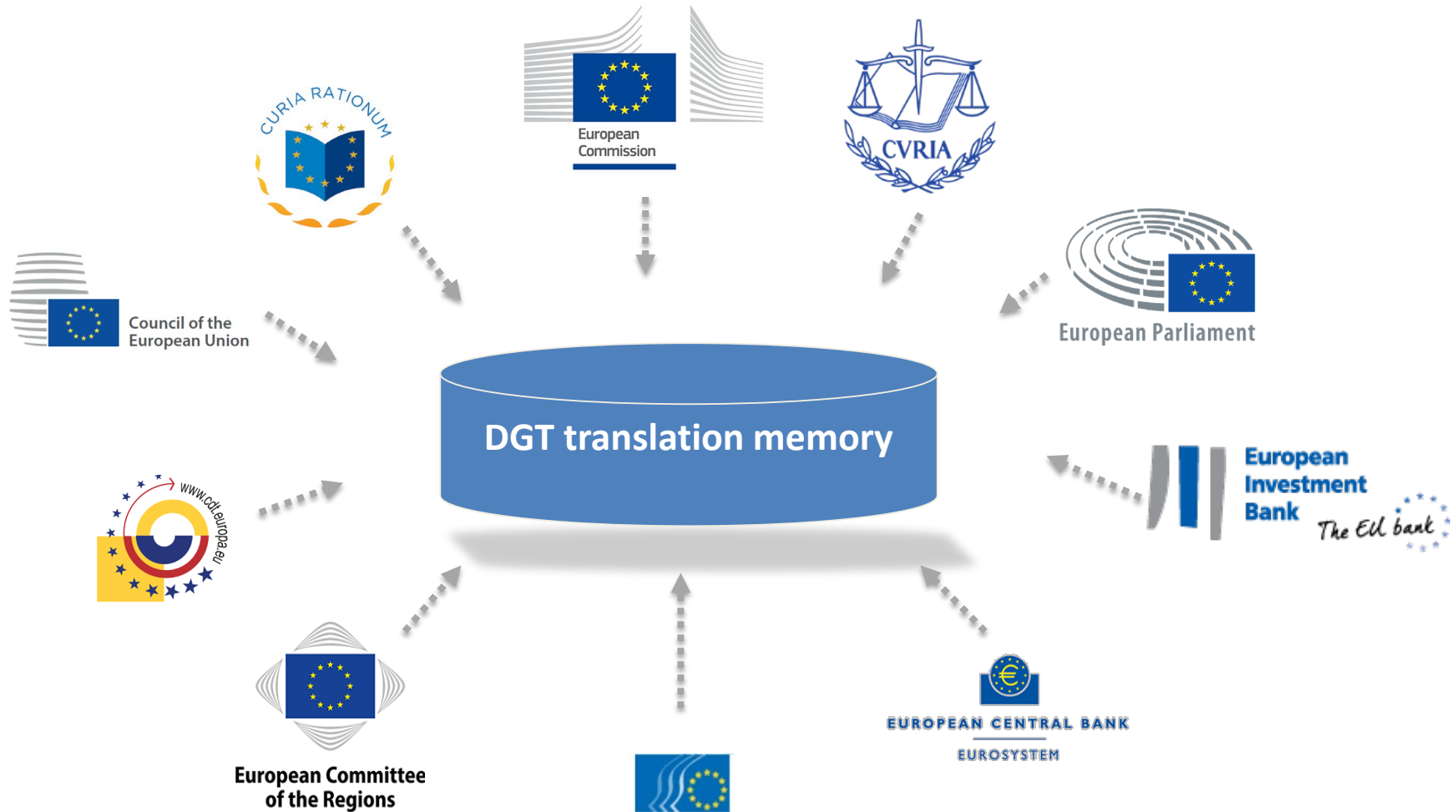
RO

Legea penală prevede pedepsele aplicabile și măsurile educative ce se pot lua față de persoanele care au săvârșit infracțiuni, precum și măsurile de siguranță ce se pot lua față de persoanele care au comis fapte prevăzute de legea penală.



EN

Criminal law establishes applicable penalties and educational measures that can be ruled against persons who committed offenses, as well as security measures that can be ruled against persons who committed actions covered by criminal law.



Such data are already available  
BUT  
they are not enough...



- Any **electronically stored text** in an EU language plus NO and IS
- **Texts and their translations** (i.e. parallel bilingual or multilingual)

## Romanian text

Legalitatea sancțiunilor de drept penal

- (1) Legea penală prevede pedepsele aplicabile și măsurile educative ce se pot lua față de persoanele care au săvârșit infracțiuni, precum și măsurile de siguranță ce se pot lua față de persoanele care au comis fapte prevăzute de legea penală.
- (2) Nu se poate aplica o pedeapsă ori nu se poate lua o măsură educativă sau o măsură de siguranță dacă aceasta nu era prevăzută de legea penală la data când fapta a fost săvârșită.
- (3) Nicio pedeapsă nu poate fi stabilită și aplicată în afara limitelor generale ale acesteia.

## Translation in English

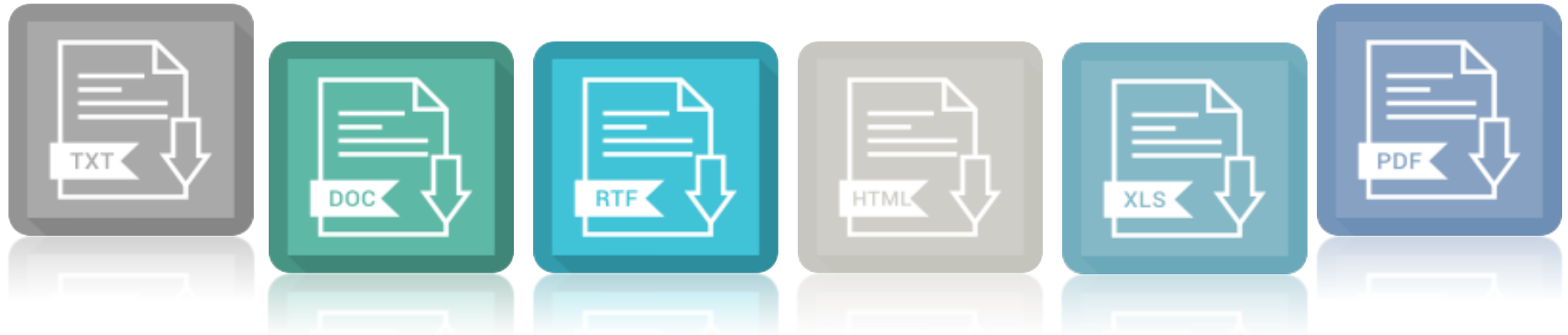
Lawfulness of criminal penalty

- (1) Criminal law establishes applicable penalties and educational measures that can be ruled against persons who committed offenses, as well as security measures that can be ruled against persons who committed actions covered by criminal law.
- (2) No penalty, educational or security measure can be ruled that was not stipulated in criminal law at the date when the violation was committed.
- (3) (3) No penalty can be ruled and enforced outside the law's general limits.

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

| Romanian            | English                   |
|---------------------|---------------------------|
| Risc                | risk                      |
| acceptarea riscului | risk acceptance           |
| risc acceptabil     | acceptable risk           |
| risc de incendiu    | fire risk                 |
| risc de inundație   | flood risk                |
| risc rezidual       | residual risk             |
| analiza riscului    | risk analysis             |
| modificarea climei  | climate change            |
| eroziune            | erosion                   |
| degradarea mediului | environmental degradation |
| ...                 | ...                       |

# 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



- 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





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. ¶

¶

Aceasta este traducerea în română a paragrafului din stânga. Aceasta este traducerea în română a paragrafului din stânga. Aceasta este traducerea în română a paragrafului din stânga. Aceasta este traducerea în română a paragrafului din stânga. ¶

¶

Aceasta este traducerea românească a sentinței la stânga. ¶

Aceasta este traducerea în română a paragrafului din stânga. Aceasta este traducerea în română a paragrafului din stânga. Aceasta este traducerea în română a paragrafului din stânga. Aceasta este traducerea în română a paragrafului din stânga. ¶





| English  | Română  |
|--|---|
| <p>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. This is a paragraph in English.</p> | <p>Aceasta este traducerea în română a paragrafului din stânga. Aceasta este traducerea în română a paragrafului din stânga. Aceasta este traducerea în română a paragrafului din stânga. Aceasta este traducerea în română a paragrafului din stânga. Aceasta este traducerea în română a paragrafului din stânga.</p> |
| <p>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.</p>   | <p>Aceasta este traducerea în română a paragrafului din stânga. Aceasta este traducerea în română a paragrafului din stânga. Aceasta este traducerea în română a paragrafului din stânga. Aceasta este traducerea în română a paragrafului din stânga. Aceasta este traducerea în română a paragrafului din stânga.</p> |



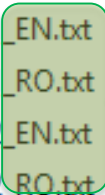
- filename01\_EN.txt
- filename01\_RO.txt
- filename02\_EN.txt
- filename02\_RO.txt
- filename03\_EN.txt
- filename03\_RO.txt
- filename04\_EN.txt
- filename04\_RO.txt
- filename05\_EN.txt
- filename05\_RO.txt
- filename06\_EN.txt
- filename06\_RO.txt
- filename07\_EN.txt
- filename07\_RO.txt
- filename08\_EN.txt
- filename08\_RO.txt
- filename09\_EN.txt
- filename09\_RO.txt
- filename10\_EN.txt
- filename10\_RO.txt

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



- filename01\_EN.txt
- filename01\_RO.txt
- filename02\_EN.txt
- filename02\_RO.txt
- filename03\_EN.txt
- filename03\_RO.txt
- filename04\_EN.txt
- filename04\_RO.txt
- filename05\_EN.txt
- filename05\_RO.txt
- filename06\_EN.txt
- filename06\_RO.txt
- filename07\_EN.txt
- filename07\_RO.txt
- filename08\_EN.txt
- filename08\_RO.txt
- filename09\_EN.txt
- filename09\_RO.txt
- filename10\_EN.txt
- filename10\_RO.txt

Include **language identifiers** in the filename



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



- 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

- 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

Data sources submission ▶

Click below if you are a language resource owner and are willing to share it for the purposes of CEF.AT

Language resource submission ▶

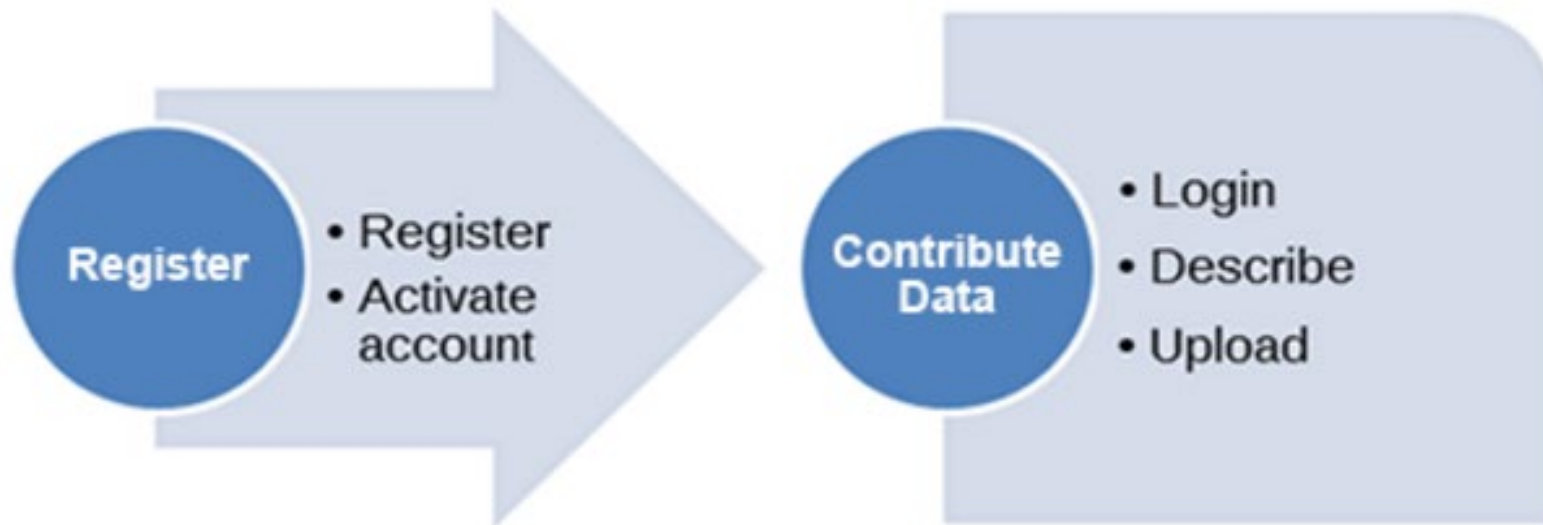


## ELRC-SHARE Repository



Welcome to the ELRC-SHARE repository!







 Register

## ELRC-SHARE Repository



Welcome to the ELRC-SHARE repository!



- 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.

Create Account



## Data Contribution

### New Resource

Resource Title\*

The name by which the resource is already known or by which you would like it to be known; e.g. "The GSRT bilingual corpus of Greek-English bulletins"



- Fill in the details of the dataset

**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"

**Resource short description\***

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

**Language(s)**

- Croatian
- Danish
- Dutch; Flemish
- English
- Estonian
- Finnish
- French
- German
- Hungarian

- Three modes for contributing your data

## Contribution Mode\*

- Upload ZIP archive
- Provide URL of resources
- eDelivery (Generate XML file to attach to your eDelivery contribution)

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

## Upload Resource\*

Choose File No file chosen

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

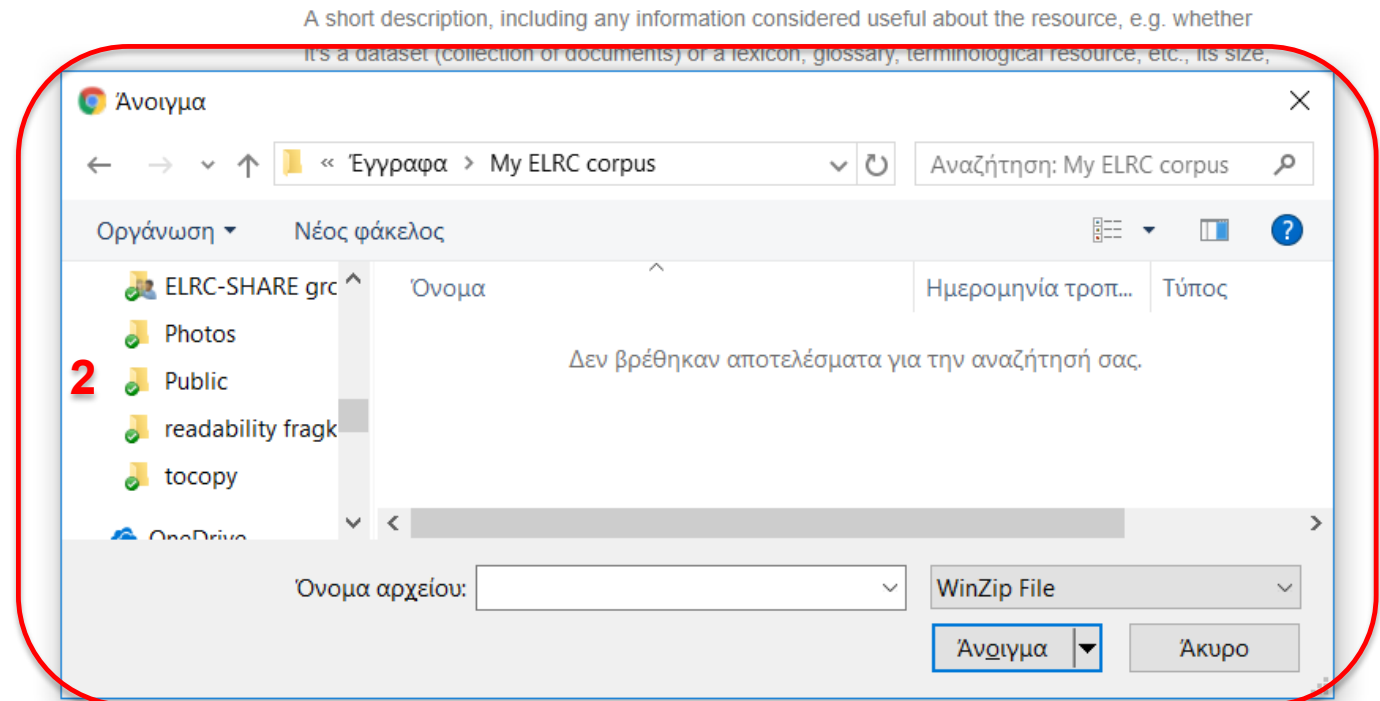
In case the **.zip file** you wish to upload is larger than 100MB, please contact [elrc-share@ilsp.gr](mailto:elrc-share@ilsp.gr)

Submit

Reset

## How to Contribute Data (4/6)

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



Upload Resources

**1** Choose File No file chosen

Please upload a .zip file up to 100MB.

In case the .zip file you wish to upload is larger than 100MB, please contact [elrc-share@ilsp.gr](mailto:elrc-share@ilsp.gr)

**3**

Submit

Reset



- 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

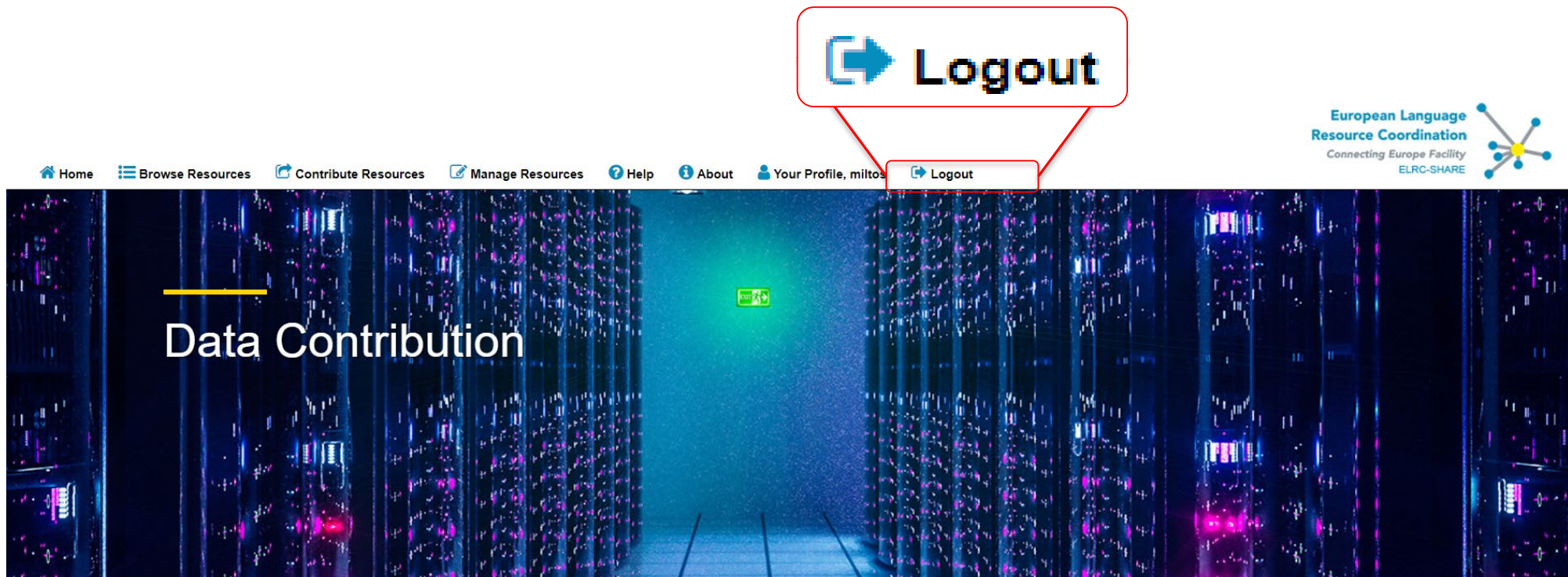
Please select the way you wish to contribute your data. Uploading a ZIP archive is recommended.

**Resource URL\***

Please provide a URL containing the files you wish to contribute



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



The screenshot displays the top navigation bar of the ELRC-SHARE website. The navigation items are: Home, Browse Resources, Contribute Resources, Manage Resources, Help, About, Your Profile, milto, and Logout. The 'Logout' button is highlighted with a red callout box that contains a blue arrow icon pointing right and the text 'Logout'. Below the navigation bar is a large image of a server room with the text 'Data Contribution' overlaid in white.



## Help

### Documentation on the ELRC-SHARE editor

The following guidelines provide detailed information on how to use the editing facility for documenting and uploading LR:

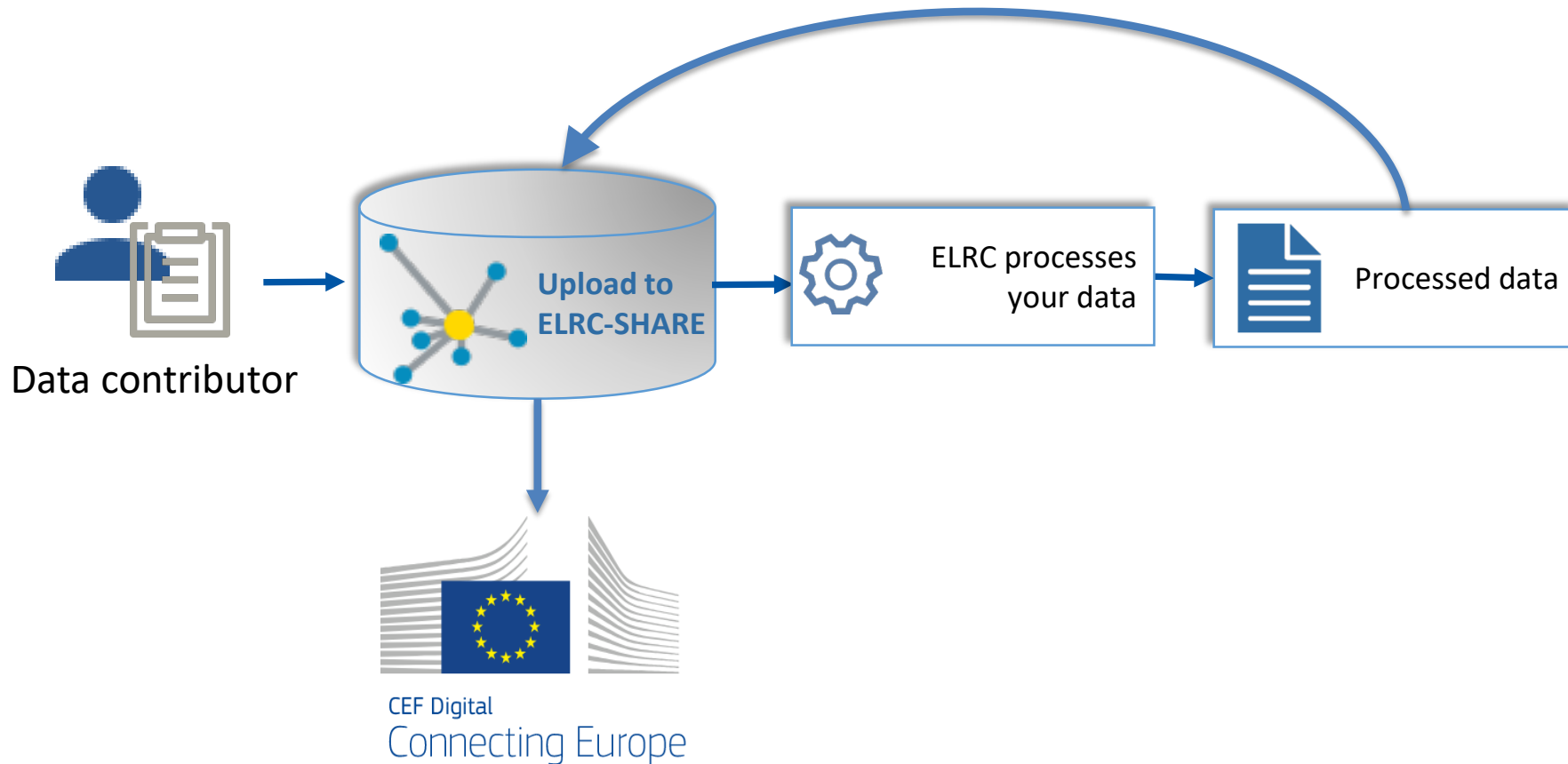
- [Walkthrough for contributors](#)
- [Walkthrough for editors](#)

### 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 extraction

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



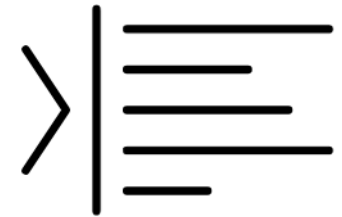
## Anonymisation

Does your data contain private info? We can help to anonymise



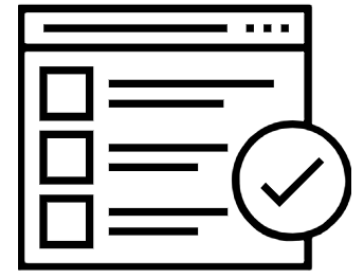
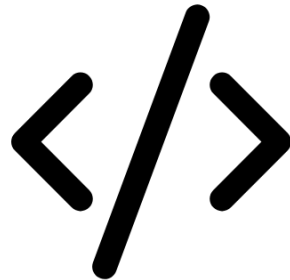
## Cleaning

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!



## 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\_ro.txt  
File01\_en.doc  
File02\_ro.pdf  
File02\_en.txt  
File03\_ro.doc  
File03\_en.doc  
...

After  
processing

```
<tu tuid="269">
  <tuv xml:lang="en">
    <seg>Ancillary penalties, additional penalties,
and security measures in case of multiple offenses</seg>
  </tuv>
  <tuv xml:lang="ro">
    <seg>Pedepsele complementare, pedepsele
accesorii și măsurile de siguranță în caz de pluralitate
de infracțiuni</seg>
  </tuv>
</tu>
<tu tuid="270">
  <tuv xml:lang="en">
    <seg>(1) If one of the committed violations
carries a ancillary penalty, such penalty shall be
awarded alongside the main penalty.</seg>
  </tuv>
  <tuv xml:lang="ro">
    <seg>(1) Dacă pentru una dintre infracțiunile
săvârșite s-a stabilit și o pedeapsă complementară,
aceasta se aplică alături de pedeapsa principală.</seg>
```

## General Romanian-English bilingual corpus

**Attribution details:** General Romanian-English bilingual corpus was created for the European Language Resources Coordination Action (ELRC) (<http://lr-coordination.eu/>) by Tufis Dan, Institutul de Cercetari pentru Inteligenta Artificiala "Mihai Draganescu", Academia Romana ([www.racai.ro/](http://www.racai.ro/)) with primary data copyrighted by Wikipedia (<https://en.wikipedia.org>) and is licensed under "CC-BY-SA 3.0" (<https://creativecommons.org/licenses/by-sa/3.0/>).

Romanian – English corpus built from a Wikipedia dump.

DSI Relevance: Europeana

[← Back](#) [Download](#) [Edit Resource](#)

### Distribution

Availability: Available

#### Licences

CC-BY-SA-3.0

Conditions: Attribution, Share Alike

#### Distribution Details

**Attribution Details:** General Romanian-English bilingual corpus was created for the European Language Resources Coordination Action (ELRC) (<http://lr-coordination.eu/>) by Tufis Dan, Institutul de Cercetari pentru Inteligenta Artificiala "Mihai Draganescu", Academia Romana ([www.racai.ro/](http://www.racai.ro/)) with primary data copyrighted by Wikipedia (<https://en.wikipedia.org>) and is licensed under "CC-BY-SA 3.0" (<https://creativecommons.org/licenses/by-sa/3.0/>).

#### IPR Holders

[Wikipedia, the free encyclopedia](#) 

### Contact Person

[Dan Tufis](#) 

text 

### Bilingual text corpus

#### Languages

English (en) (2,671,991 Words)

Romanian; Moldavian; Moldovan (ro) (2,729,213 Words)

#### Linguality

Linguality type: Bilingual

Multi-linguality type: Parallel

#### Text Format

Plain Text

#### Size

5,622,357 Words

#### 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: 18/11/2016

Last Updated: 16/12/2016

Metadata Language: English (en)

#### Metadata Creator

[Kanella Pouli](#) 

[Dan Tufis](#) 

#### Version

Version: 1.0





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

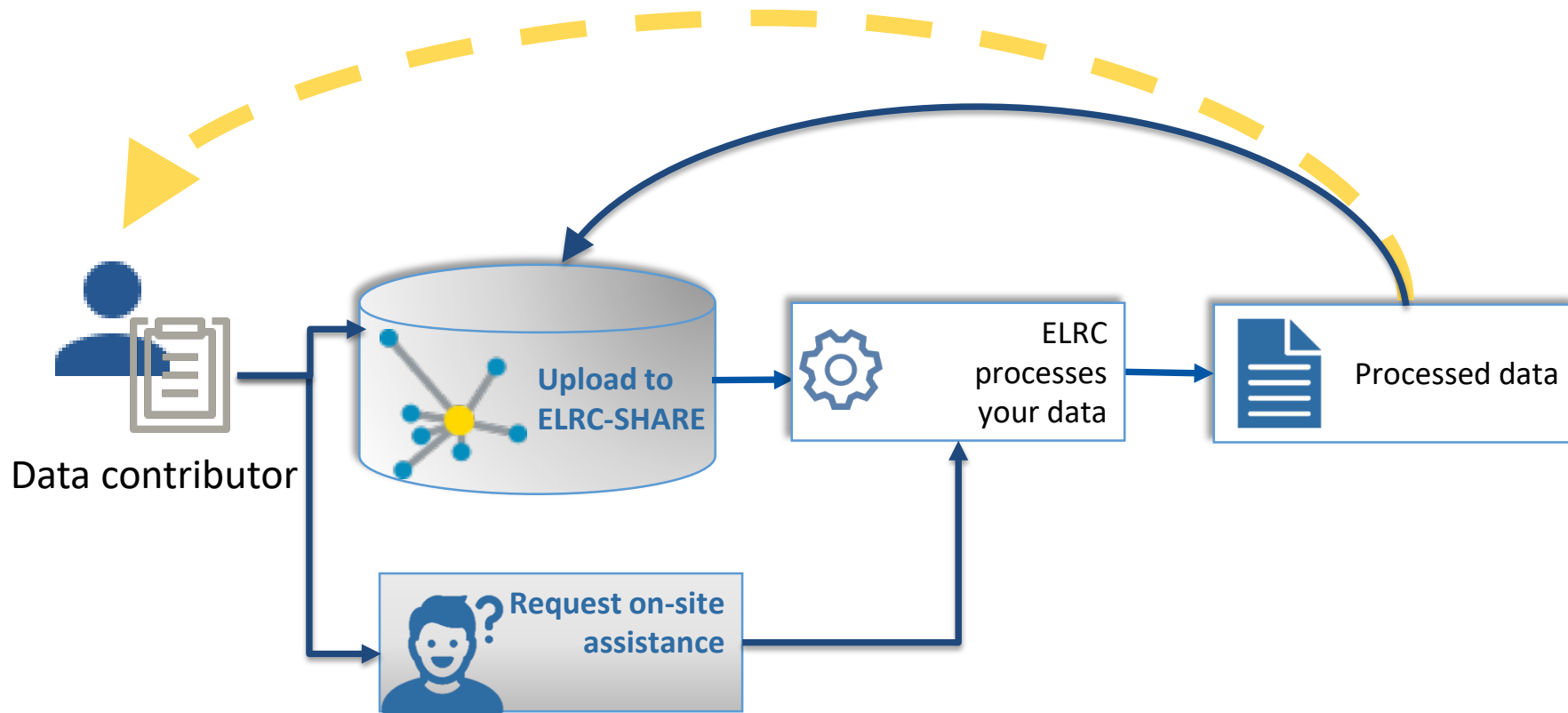




**Our team of experts will travel  
directly to assist you  
at your own offices**

**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!**

# What happens to your data?



# How to request services and help

[www.lr-coordination.eu/request-onsite-assistance](http://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 \***

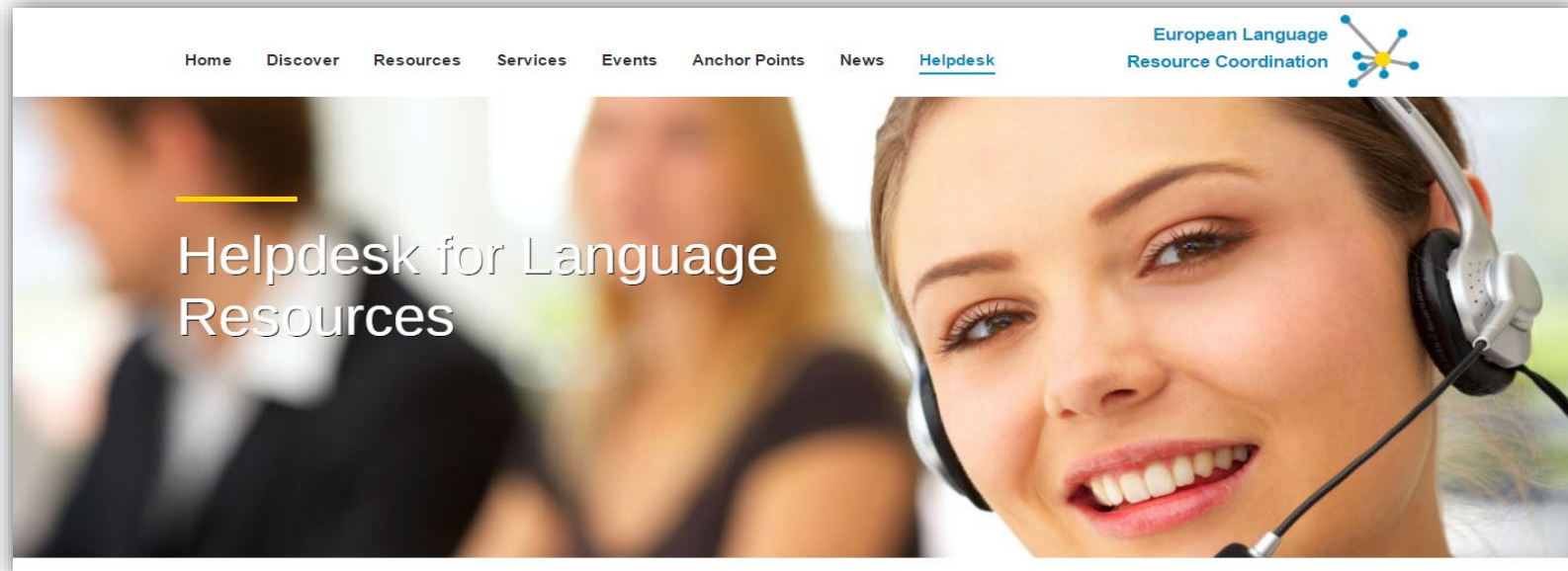
**Types of assistance required \***

- Legal assistance
- Data processing
- Anonymisation
- Other

**Description of assistance required**

Submit

[www.lr-coordination.eu/helpdesk](http://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              | <a href="mailto:help@lr-cooridantion.eu">help@lr-cooridantion.eu</a> |

Mulțumesc!





- 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](http://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