

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

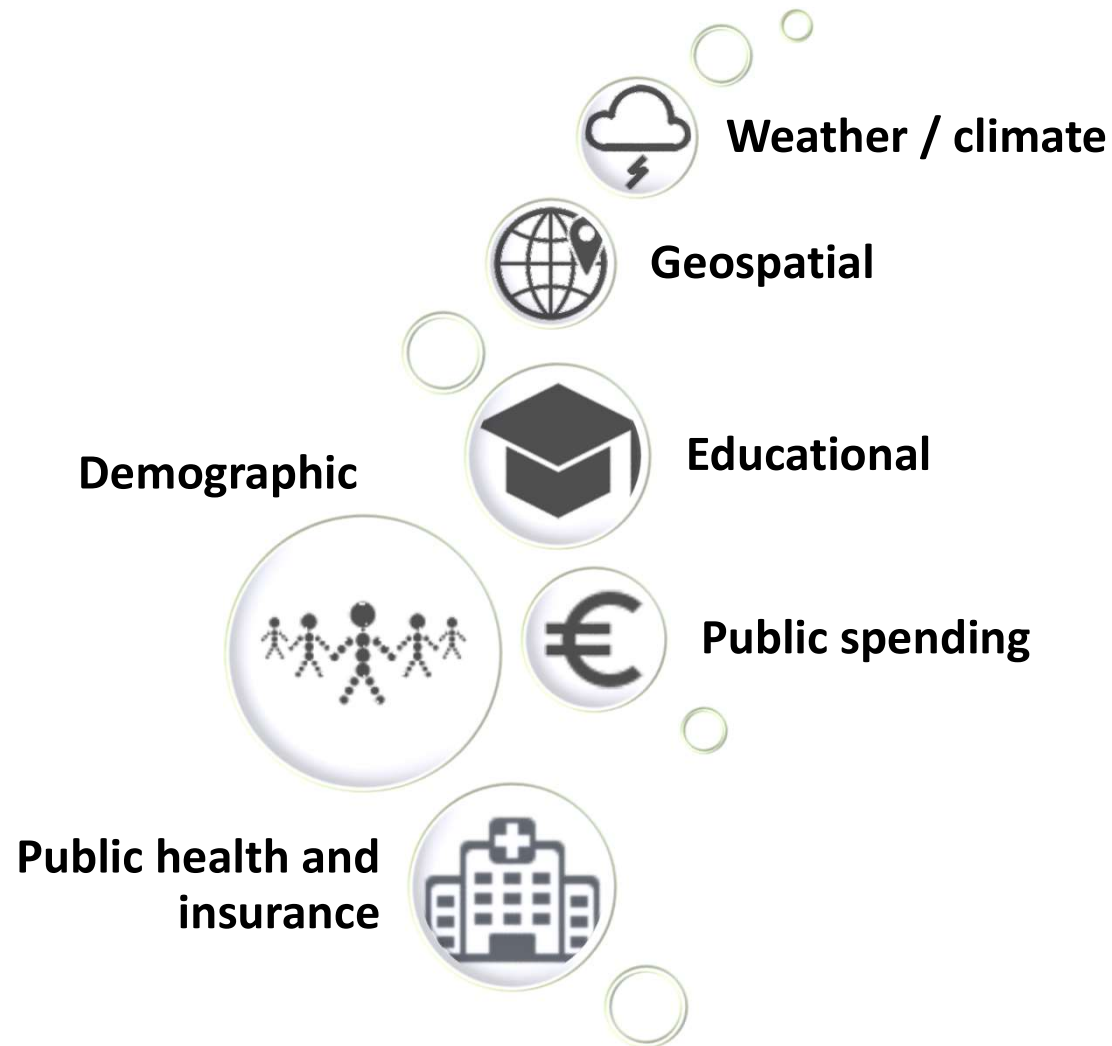
Maria Giagkou

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



Connecting  
Europe  
Facility

# The notion of data

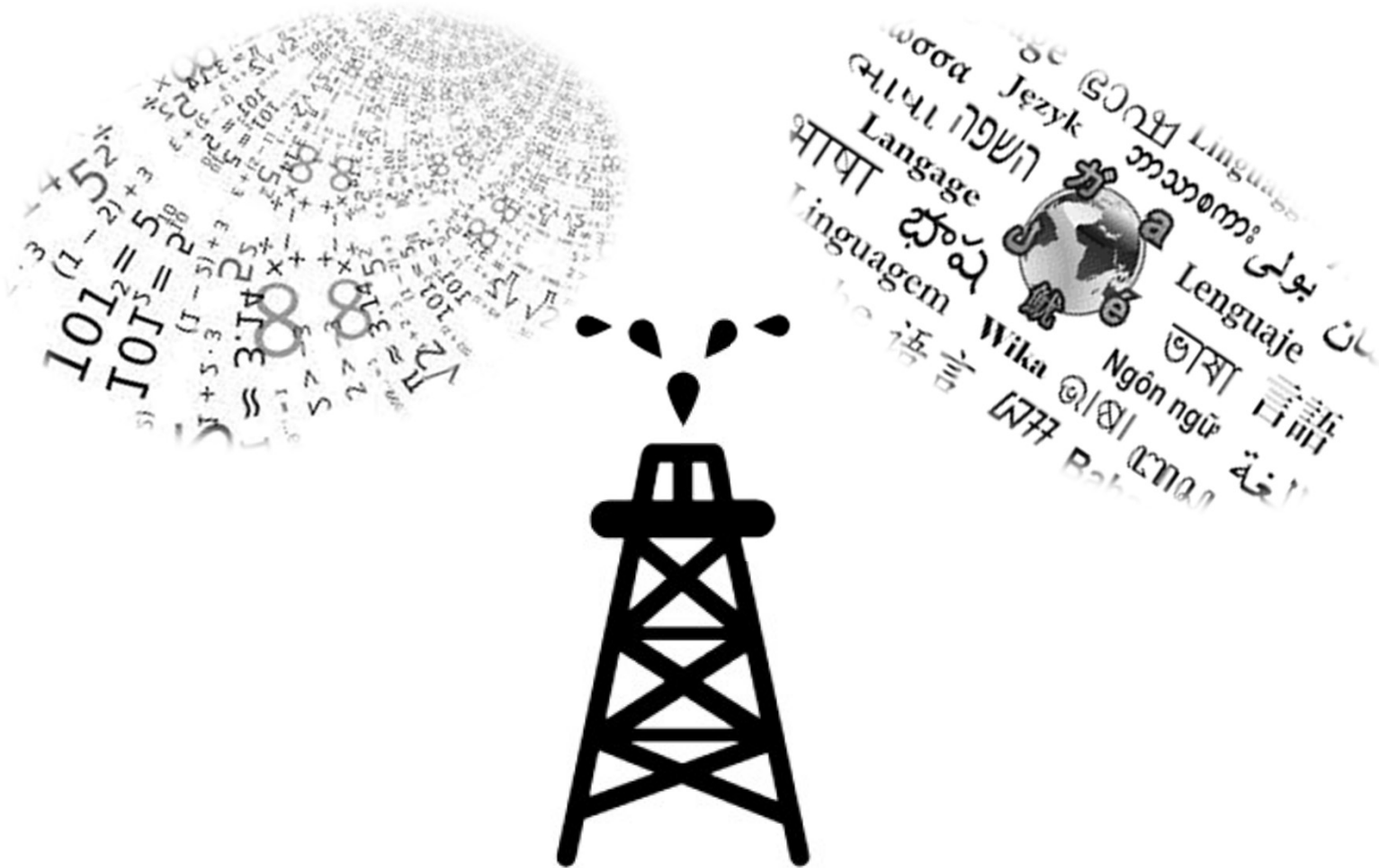


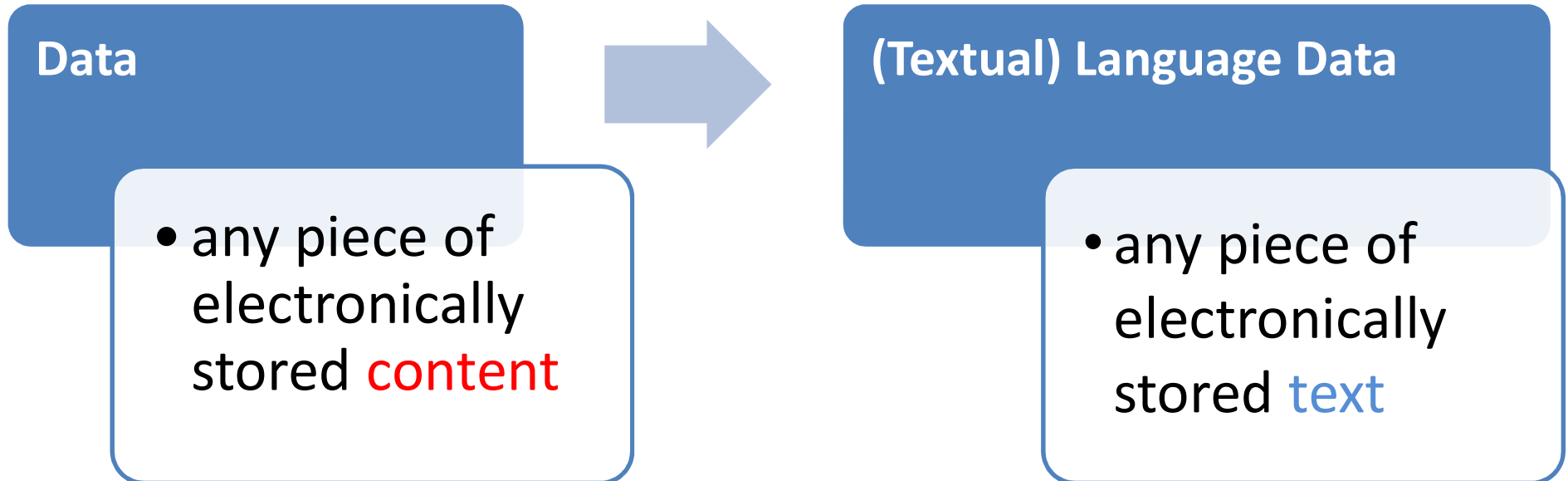


Data: the oil of the 21<sup>st</sup> century



# The notion of data





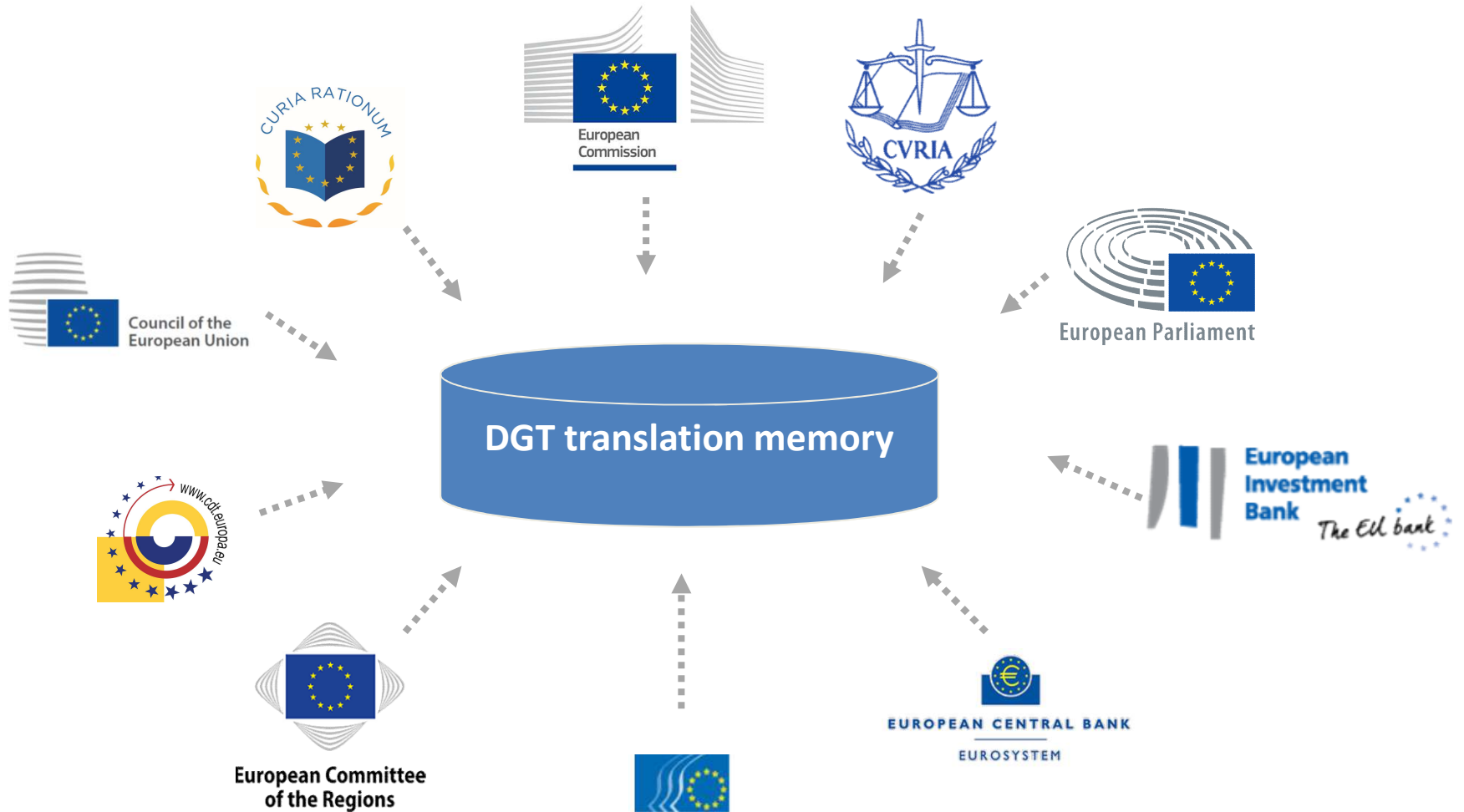
DE

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.

# Data used by eTranslation



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)

## German text

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.

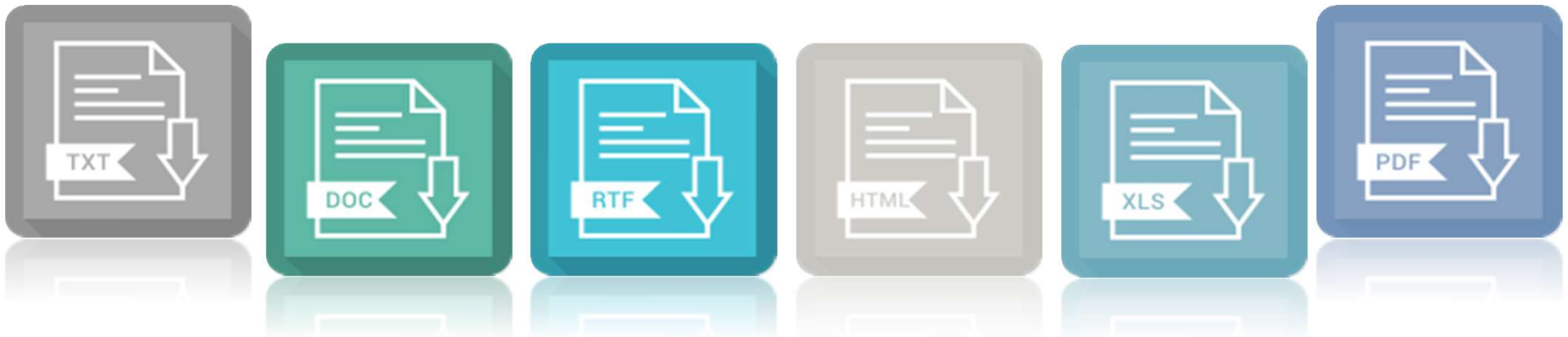
## Translation in English

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.

- 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)
...	...

## 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



1480

ΕΦΗΜΕΡΙΣ ΤΗΣ ΚΥΒΕΡΝΗΣΕΩΣ (ΤΕΥΧΟΣ ΠΡΩΤΟ)

## **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



- 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

**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. 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. 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. 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.



### 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-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.¶

¶






Don'ts



English	Deutsche
<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.</p>	<p>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.</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.</p>	<p>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.</p>

 (Ctrl) ▾



Name

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

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



Name

- filename01\_DE.txt
- filename01\_EN.txt
- filename02\_DE.txt
- filename02\_EN.txt
- filename03\_DE.txt
- filename03\_EN.txt
- filename04\_DE.txt
- filename04\_EN.txt
- filename05\_DE.txt
- filename05\_EN.txt

Include **language identifiers** in the filename



- 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*



- 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 ▶

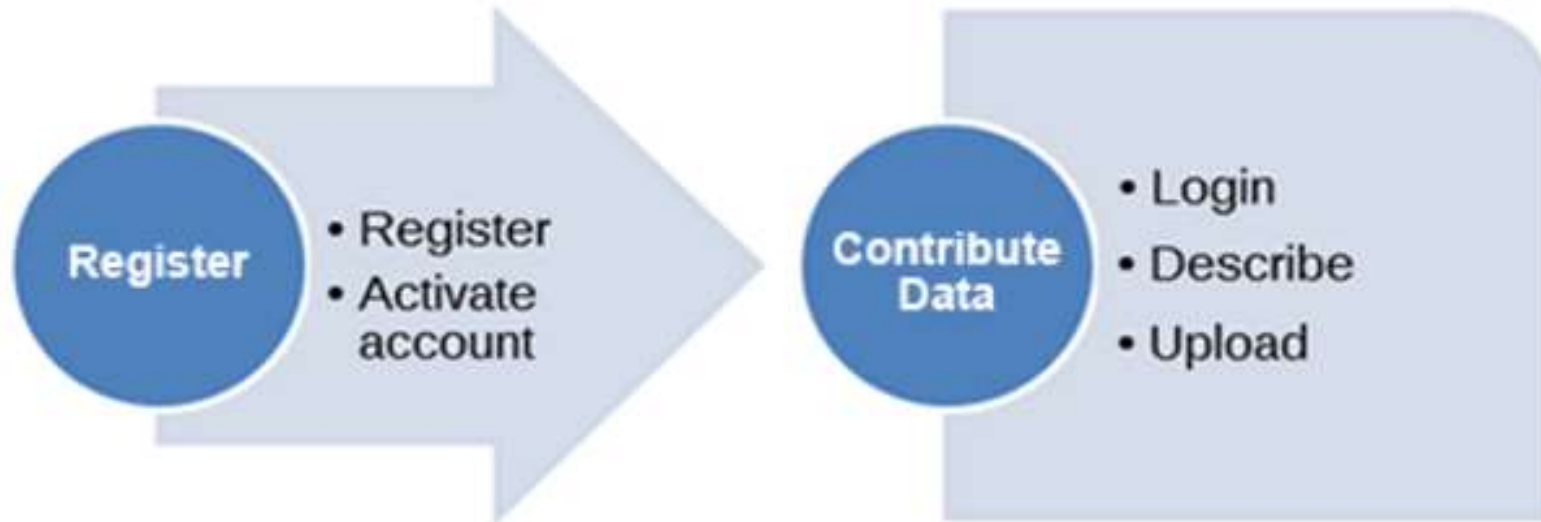


## ELRC-SHARE Repository



Welcome to the ELRC-SHARE repository!





# How to Register (1/2)



 Register

## ELRC-SHARE Repository

Type in your keywords, please...

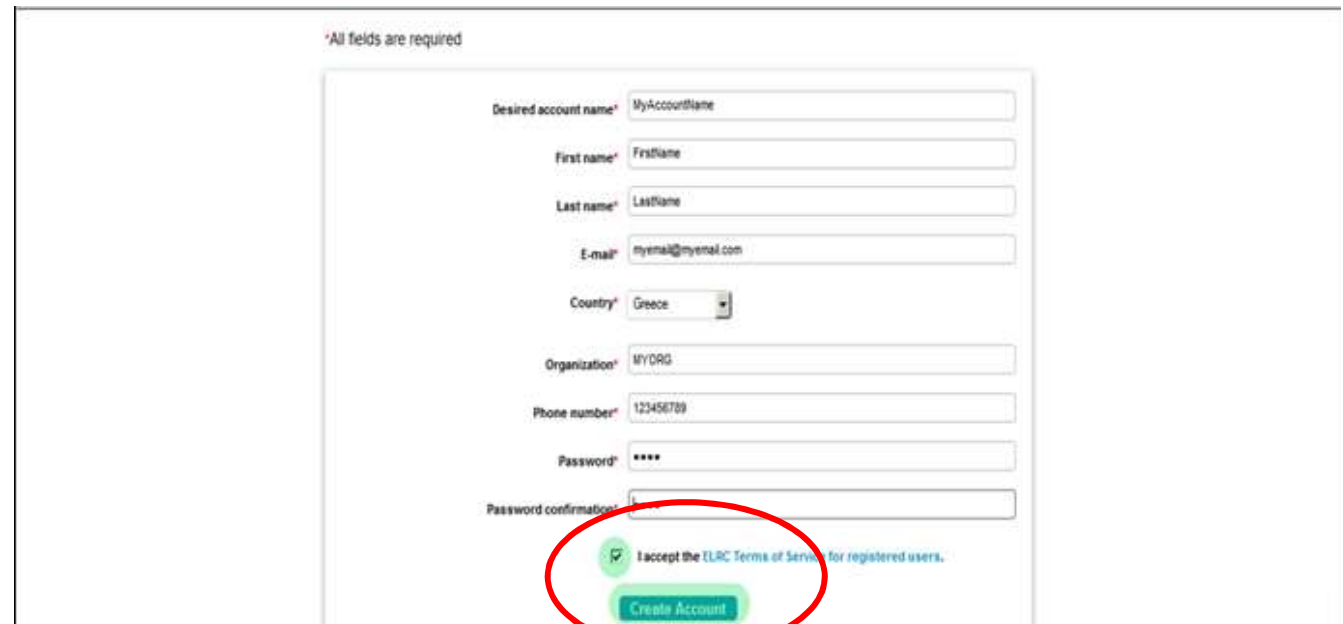


Welcome to the ELRC-SHARE repository!



- 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



The screenshot shows a registration form with the following fields: Desired account name\* (MyAccountName), First name\* (Firstname), Last name\* (Lastname), E-mail\* (myemail@myemail.com), Country\* (Greece), Organization\* (MYORG), Phone number\* (123456789), Password\* (\*\*\*\*), and Password confirmation\*. A red circle highlights the 'I accept the ELRC Terms of Service for registered users.' checkbox and the 'Create Account' button.

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



## 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



The screenshot shows a web form with three main sections:

- Resource Title\***: A text input field containing "Bilingual resource name". Below it is a descriptive paragraph: "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 larger text area containing "A short resource description:". Below it is a descriptive paragraph: "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)**: A dropdown menu showing a list of languages: Croatian, Danish, Dutch, Flemish, English, Estonian, Finnish, French, German, and Hungarian. The "English" and "French" options are highlighted in blue.

- 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.

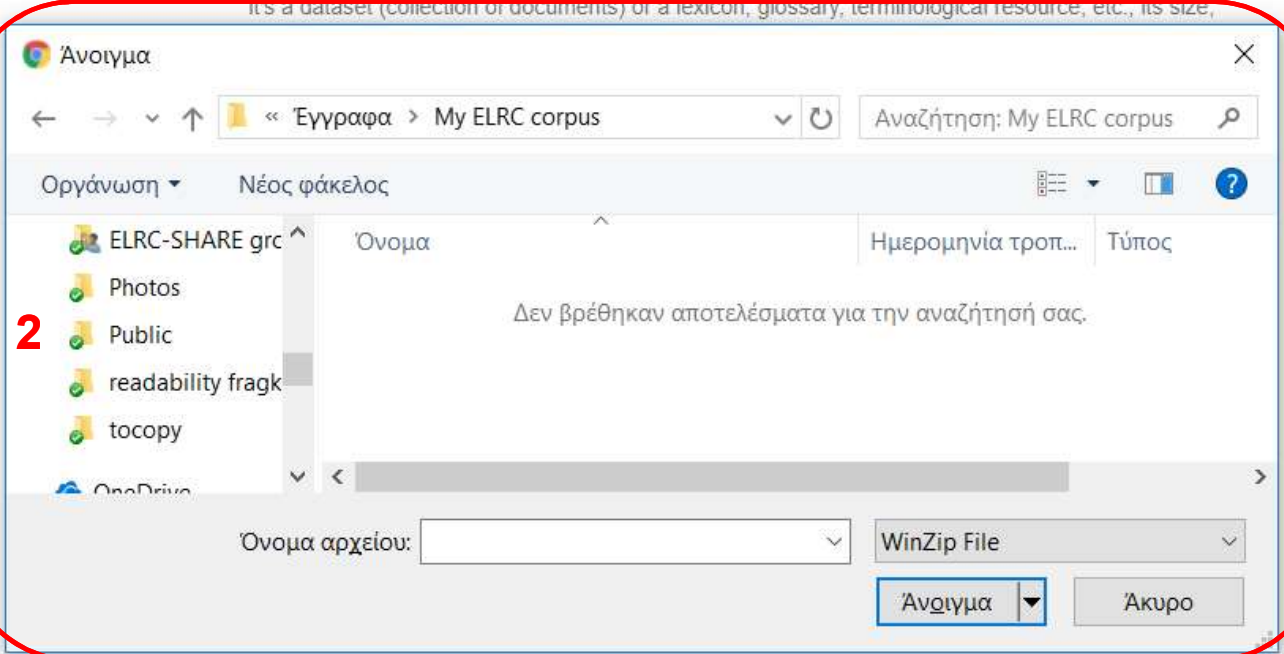
In case the **.zip file** file you wish to upload is larger than 100MB, please contact [elrc-share@ilsp.gr](mailto:elrc-share@ilsp.gr)

Submit

Reset

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

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,



Upload Resource **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@lsp.gr](mailto:elrc-share@lsp.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

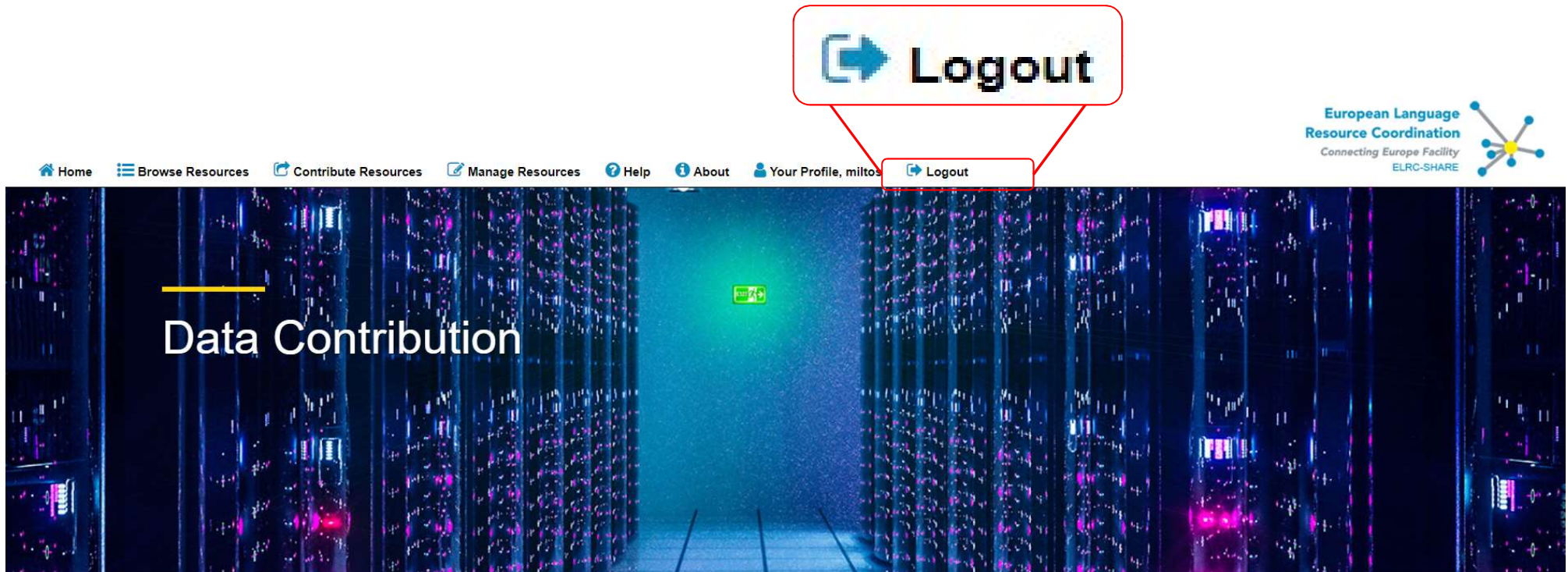
Please select the way you wish to contribute your data. Uploading a ZIP archive is recommended.

**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 shows 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 box, and a callout box with a red border and a blue arrow icon points to it with the text 'Logout'. Below the navigation bar is a large banner image of a server room with the text 'Data Contribution' overlaid in white. The ELRC-SHARE logo is visible in the top right corner of the banner area.



## Help

### Documentation on the ELRC-SHARE editor

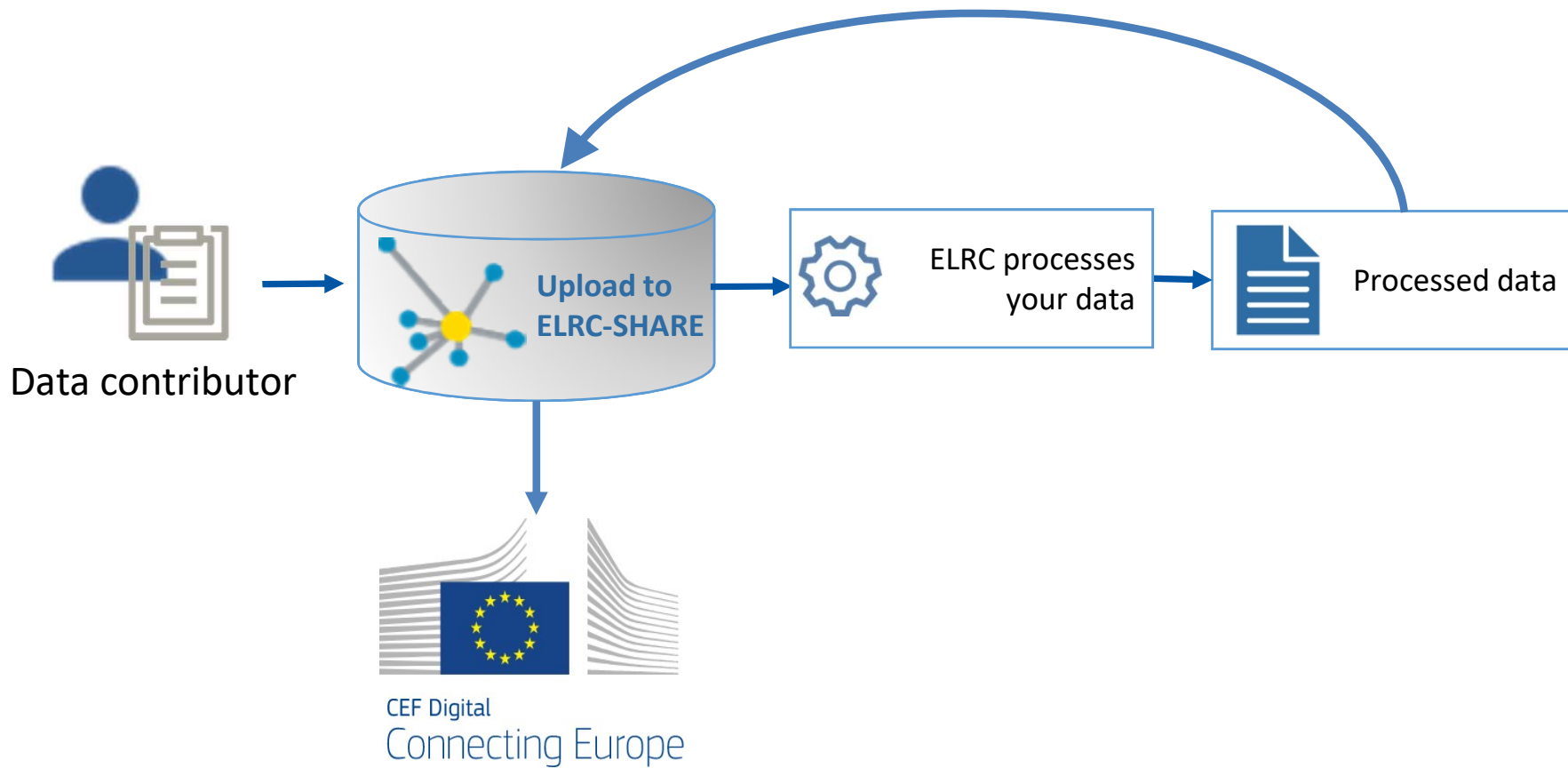
The following guidelines provide detailed information on how to use the editing facility for documenting and uploading LR:

- [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?





## Data extraction

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



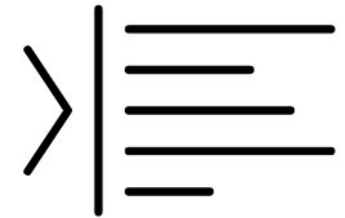
## Anonymisation

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



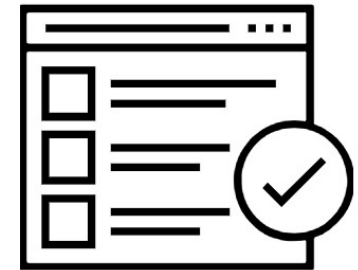
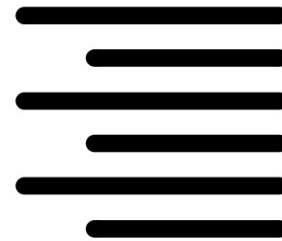
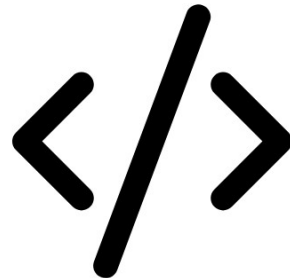
## Cleaning

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\_de.txt  
File01\_en.doc  
File02\_de.pdf  
File02\_en.txt  
File03\_de.doc  
File03\_en.doc  
...

After  
process

```
<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>
  </tu>

```

## Documents concerning Federal Constitutional Law in Austria 🇺🇸

34 ✓ 5

Attribution details: Language Institute, Austrian Armed Forces

Alignment documents concerning Austrian Federal Constitutional Law

DSI Relevance: eJustice

← Back   Download   Edit Resource

### Distribution

Availability: Available

#### Licences

Terms for PSI-compliant resources

Open Under-PSI

Conditions: Attribution

#### Distribution Details

Attribution Details: Language Institute, Austrian Armed Forces

### Contact Person

felix funda

### Bilingual text corpus

#### Languages

English (en)

German (de)

#### Linguality

Linguality type: Bilingual

Multi-linguality type: Parallel (Aligned data)

#### Text Format

TM format of the SDL alignment tool

#### Size

633 Kb

#### 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: 20/05/2016

Last Updated: 14/06/2016

Metadata Language: English (en)

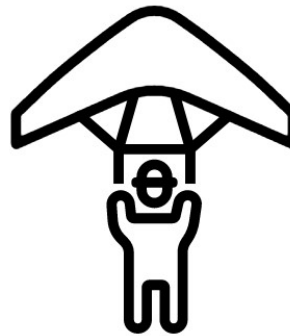
Metadata Creator





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



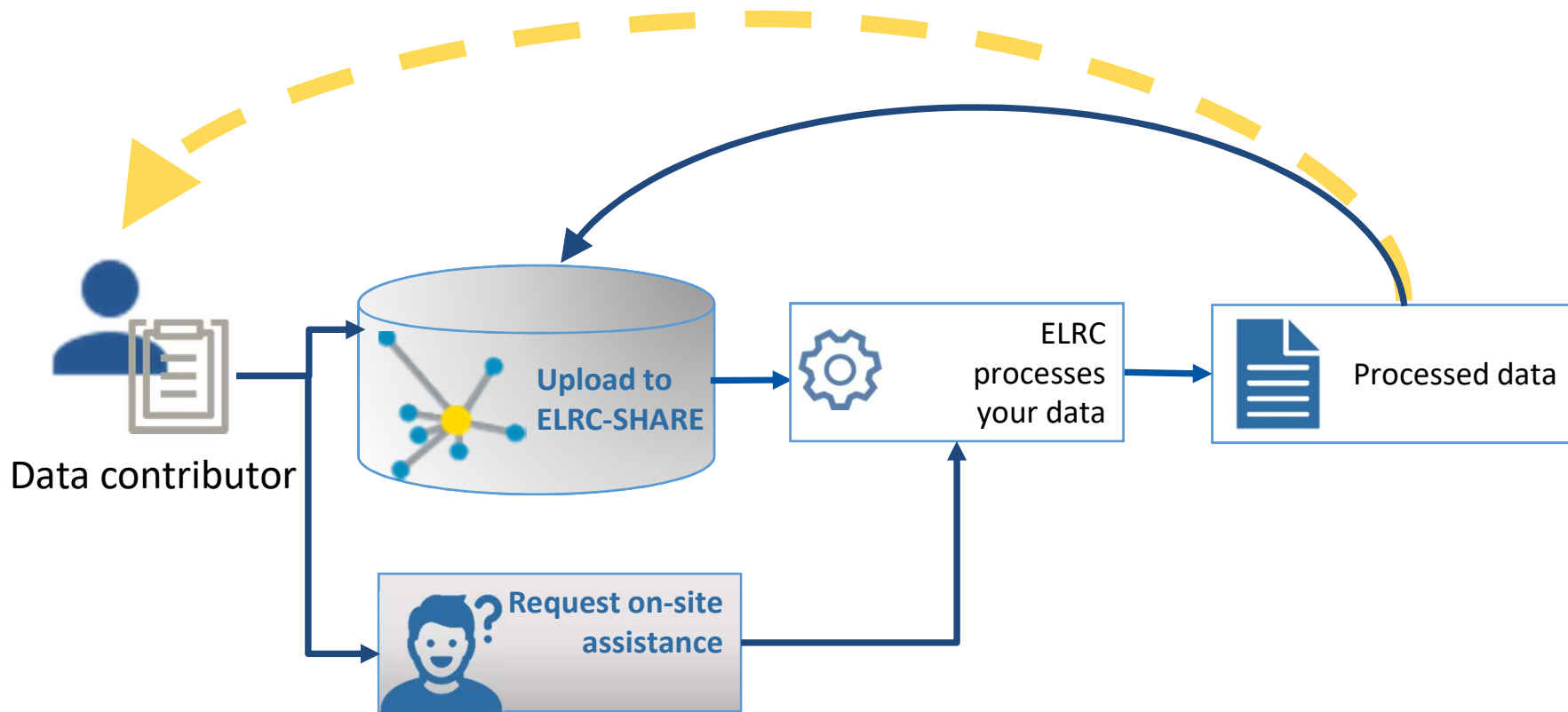


**Our team of experts will travel  
directly to assist you  
at your own offices**



**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



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](http://lr-coordination.eu/request-onsite-assistance)



## Helpdesk for Language Resources

### Helpdesk for Language Resources

We are happy to answer any questions on the technical or legal aspects related to the use, production, collection, processing, and sharing of language resources.

Please feel free to contact us through one of the following channels:

Telephone*	+33 970 440 522
Secretariat Support	+49 681 857 7552 85
Skype	ELRC Helpdesk
E-mail	<a href="mailto:help@lr-coordination.eu">help@lr-coordination.eu</a>

[lr-coordination.eu/helpdesk](https://lr-coordination.eu/helpdesk)

# ELRC consortium – come talk to us!





Danke für Ihre Aufmerksamkeit!



- 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