

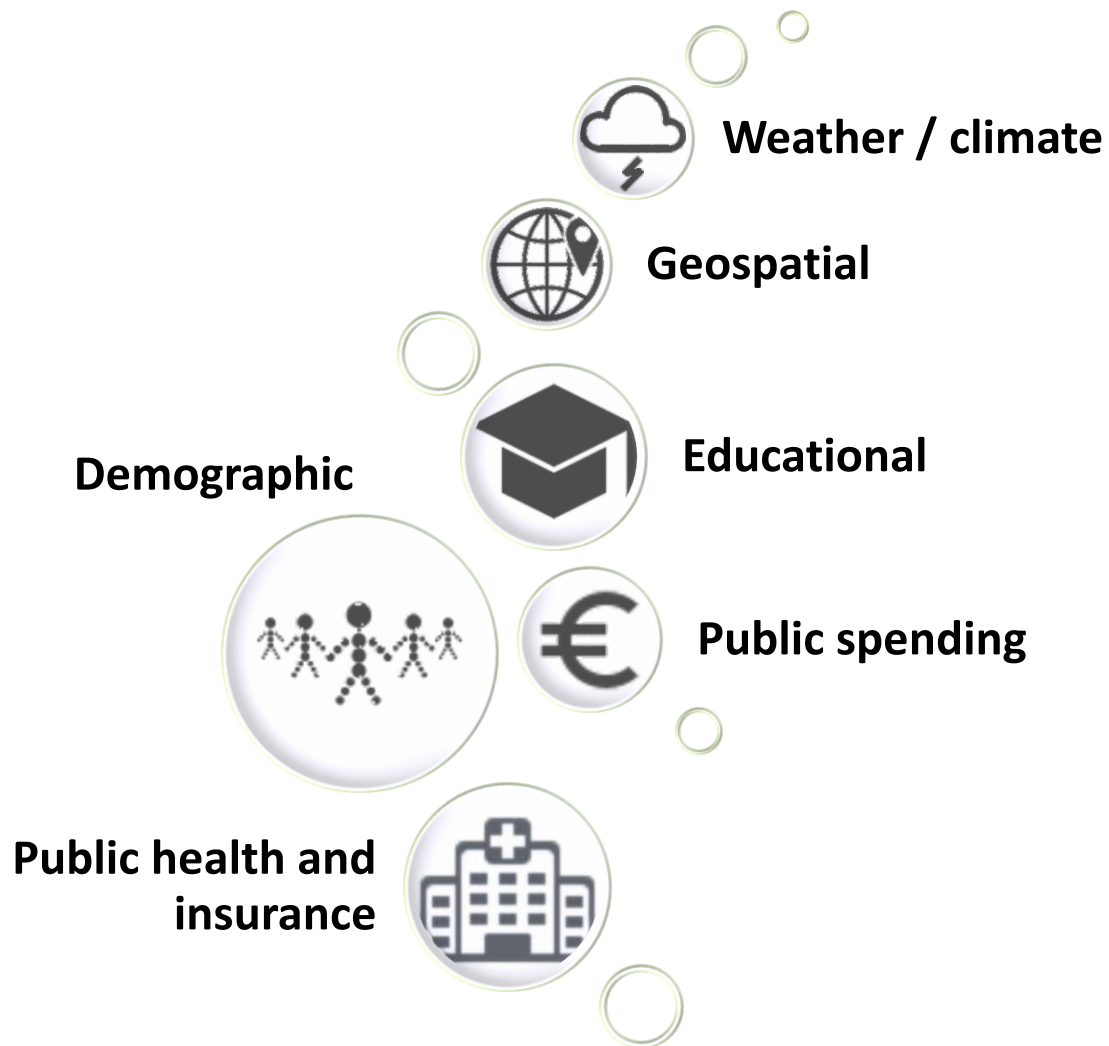
# ELRC Workshop in Ireland

## Preparing and sharing data with the ELRC repository

Dr Teresa Lynn  
Dublin City University



# 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)



EUROPA > Open Data Portal > Data > Publisher > Publications Office > CORDIS - EU research projects un...

Data Applications Linked Data Developers' corner About

CORDIS - EU research projects under Horizon 2020 (2014-2020)

**Publisher**  
Publications Office »

**Description**  
This dataset contains projects funded by the European Union under the Horizon 2020 framework programme for research and innovation (H2020) from 2014 to 2020. Grant information is provided for each project, including RCN, ID, Acronym, Status, Programme, Topic, Title, Start Date, End Date, Objective, Total Cost, EC Max Contribution, Call Id, Funding Scheme, Coordinator, Coordinator Country, Participants (semi-colon separated list), Participant Countries (semi-colon separated list)  
For each participant you can find in the organisations file: RCN, ID, Acronym, Role, Organisation Name, Organisation Short Name, Organisation Type, Participation Ended, EC Contribution, Organisation Country  
Reference data (H2020 programmes and topics, funding schemes / types of action, and countries) can be found in this dataset:  
<https://data.europa.eu/euodp/en/data/dataset/cordisref-data>  
CORDIS datasets are produced on a monthly basis. Therefore inconsistencies may occur between what is presented on the CORDIS live website and the datasets.

**Resources**

DOWNLOAD	H2020 Organisations	CSV
DOWNLOAD	H2020 Organisations	XLSX
DOWNLOAD	H2020 Projects	CSV
DOWNLOAD	H2020 Projects	XLSX
DOWNLOAD	H2020 Projects	ZIP

**URI**  
<http://cordis.europa.eu/projects/>

**Status**  
Under Development

**Licence:**

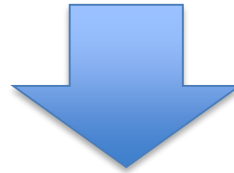
**Legal Notice**

**Catalogue record**

Added to data.europa.eu/euodp  
2015-07-29  
Updated on data.europa.eu/euodp  
2017-06-01  
Views: 17658  
Downloads: 16453

**Suggest a dataset**  
Is there data you would like to find on the portal?  
[Make a suggestion>>](#)

- **Data:** any piece of electronically stored **content**



- (Textual) **Language Data:** any piece of electronically stored **text**
- Public Administration **Language Data** also Open Data



## Central Statistical Office Dataset

Two Polish-English publications of the Polish Central Statistical Office in the XLIFF format  
1. "Statistical Yearbook of the Republic of Poland 2015" is the main summary publication of the Central Statistical Office, including a comprehensive set of statistical data describing the condition of the... [Read More](#)

Appropriateness for DSI: Open Data Portal

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



- Data files within the dataset:
- File01\_pl.txt
  - File01\_en.txt
  - File02\_pl.txt
  - File02\_en.txt
  - File03\_pl.txt
  - File03\_en.txt
  - ...



**Distribution**

**Availability**  
Available

**Licence**  
CC-BY 4.0  
Conditions: Attribution  
Attribution Details: Central Statistical Office Dataset was created for the European Language Resources Coordination Action (ELRC) (http://lr-coordination.eu) by Ogrodniczuk Maciej, Institute of Computer Science, Polish Academy of Sciences, with primary data copyrighted by the Central Statistical Office of Poland (http://stat.gov.pl/en) and is licensed under "CC-BY 4.0" (https://creativecommons.org/licenses/by/4.0/)

Allows Uses Besides DGT

**Contact Person**  
Maciej Ogrodniczuk

**Bilingual text corpus**

**Languages**  
Polish (pl)  
English (en)

**Linguality**  
Linguality type: Bilingual  
Multi-linguality type: Parallel

**Text Format**  
XML

**Size**  
1,532 Translation Units

**Character encoding**  
UTF-8

**Domains**  
SOCIAL QUESTIONS (Demography And Population)  
Conforms to EUROVOC  
ENVIRONMENT (Natural Environment)  
Conforms to EUROVOC  
SOCIAL QUESTIONS (Social Framework)  
Conforms to EUROVOC

**Annotation**

**Metadata**

Created: Sept. 18, 2016  
Last Updated: Dec. 15, 2016  
Metadata Language: English (en)

Zapewnienie równości szans dla kobiet i mężczyzn oraz pełnoprawnego uczestnictwa w życiu społeczeństwa jest jednym z podstawowych praw człowieka.

Ensuring equality of chances for men and women, as well as full participation in the social life is one of the basic human rights.



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



- Data residing in local public organisations, produced in-house or outsourced:
  - Reports
  - Communication
  - News
  - Web Content
  - Policies
  - Terminologies
  - Archives
  - Forms
  - FAQs



- In principle, any **electronically stored text** in any of the EU languages, plus Norwegian and Icelandic (i.e. the CEF languages)
- Ideally, **texts and their translations** in one or more of the CEF languages (i.e. parallel bilingual or multilingual)

## Polish text

Sprawozdanie Prezesa Urzędu Zamówień Publicznych o funkcjonowaniu systemu zamówień publicznych w 2010 roku.  
Sprawozdanie z funkcjonowania systemu zamówień publicznych obejmuje okres od 1 stycznia do 31 grudnia 2011 r. i zostało opracowane na podstawie informacji zawartych w oficjalnych dokumentach i publikacjach urzędowych jak również innych dokumentach, raportach i analizach z zakresu zamówień publicznych, będących w dyspozycji Urzędu Zamówień Publicznych.

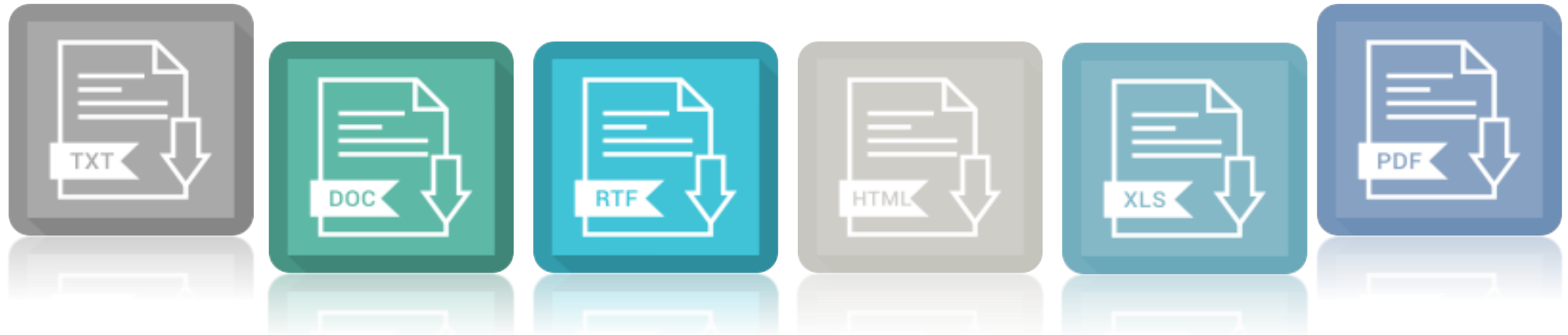
## Translation in English

Report of the President of Public Procurement Office on functioning of public procurement system in 2010.  
The report on the functioning of the public procurement system covers the period from 1 January to 31 December 2011 and it has been prepared on the basis of information obtained from official documents and publications as well as other documents, reports and analyses regarding public procurement which were available to the Public Procurement Office.

- Can also be a list of specific terms and their translations, i.e. a **terminology**

	A	B
1	Slovak	English
2	Demografia	<i>Demography</i>
3	Populácia	<i>Population</i>
4	Obyvateľstvo	<i>Inhabitants</i>
5	Demografická štruktúra	<i>Demographic structure</i>
6	Demografická reprodukcia	<i>Demographic reproduction</i>
7	Demografické správanie	<i>Demographic behaviour</i>
8	Demografické udalosti, demografické javy	<i>Demographic events</i>
9	Demografické procesy	<i>Population processes</i>
10	Demografický vývoj, populačný vývoj	<i>Demographic development, population development</i>
11	Pohyb obyvateľstva	<i>Population change</i>
12	Prirodzený pohyb	<i>Natural changes of population</i>
13	Migrácia, sťahovanie	<i>Migration</i>
14	Rozmiestnenie obyvateľstva	<i>Spatial distribution</i>
15	Hustota obyvateľstva	<i>Population density</i>
16	Demografická analýza	<i>Demographic analysis, population analysis</i>
17	Transverzálna analýza, prierezová analýza	<i>Cross-sectional analysis, current analysis</i>
18	Longitudinálna analýza, kohortná analýza	<i>Cohort analysis, longitudinal analysis</i>
19	Kohorta	<i>Cohort</i>
20	Generácia	<i>Generation</i>
21	Rodinný stav	<i>Marital status</i>
22	Slobodní	<i>Single persons</i>
23	Ženatí, vydaté	<i>Married persons</i>
24	Rozvedení, rozvedené	<i>Divorced persons</i>
25	Vdovci, vdovy, ovdovení	<i>Widowed persons</i>
26	Tabuľky života	<i>Life tables</i>
27	Skrátené tabuľky života	<i>Abridged life tables</i>
28	Podrobné tabuľky života	<i>Complete life tables</i>

# 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*

1480

ΕΦΗΜΕΡΙΣ ΤΗΣ ΚΥΒΕΡΝΗΣΕΩΣ (ΤΕΥΧΟΣ ΠΡΩΤΟ)

## United Nations Convention against Corruption

### Preamble

*The States Parties to this Convention*

*Concerned about the serious and threats posed by corruption to the stability and security of societies, the institutions and values of democracy, ethical values and justice, and the goal of sustainable development and the rule of law,*

*Concerned also about the link 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 parallel texts
    1. Translation memories (.tmx)
    2. XML translation files (.xliff)
    3. Plain text (.txt, .csv)
    4. Spreadsheets (e.g. xlsx)
    5. Word Docs
  - For terminologies
    1. TermBase eXchange (.tbx)
    2. Plain text (.txt, .csv)
    3. Spreadsheets (e.g. xlsx)
  - For monolingual texts
    1. Plain text (.txt, .csv)
    2. Word Docs

# File formats of parallel texts and their manipulation



Don'ts



This is an English paragraph. This is an English paragraph. This is an English paragraph. This is an English paragraph. This is an English paragraph. This is an English paragraph. This is an English paragraph. This is an English paragraph.

Αυτή είναι η Ελληνική μετάφραση της παραπάνω παραγράφου. Αυτή είναι η Ελληνική μετάφραση της παραπάνω παραγράφου. Αυτή είναι η Ελληνική μετάφραση της παραπάνω παραγράφου. Αυτή είναι η Ελληνική μετάφραση της παραπάνω παραγράφου. Αυτή είναι η Ελληνική μετάφραση της παραπάνω παραγράφου. Αυτή είναι η Ελληνική μετάφραση της παραπάνω παραγράφου.

This is another English paragraph. This is an English paragraph. This is an English paragraph. This is an English paragraph. This is an English paragraph. This is an English paragraph. This is an English paragraph. This is an English paragraph.

Αυτή είναι η Ελληνική μετάφραση της παραπάνω παραγράφου. Αυτή είναι η Ελληνική μετάφραση της παραπάνω παραγράφου. Αυτή είναι η Ελληνική μετάφραση της παραπάνω παραγράφου. Αυτή είναι η Ελληνική μετάφραση της παραπάνω παραγράφου. Αυτή είναι η Ελληνική μετάφραση της παραπάνω παραγράφου.

This is an English sentence.

Αυτή είναι η μετάφραση της παραπάνω πρότασης.





## Don'ts



This-is-an-English-paragraph.·This-is-an-English-paragraph.·This-is-an-English-paragraph.·This-is-an-English-paragraph.·This-is-an-English-paragraph.·This-is-an-English-paragraph.→ ¶

¶

¶

This-is-another-English-paragraph.·This-is-an-English-paragraph.·This-is-an-English-paragraph.·This-is-an-English-paragraph.·This-is-an-English-paragraph.·This-is-an-English-paragraph.¶

¶

.....

This-is-an-English-sentence.¶

¶

Αυτή·είναι·η·Ελληνική·μετάφραση·της·παραπάνω·παραγράφου.·Αυτή·είναι·η·Ελληνική·μετάφραση·της·παραπάνω·παραγράφου.·Αυτή·είναι·η·Ελληνική·μετάφραση·της·παραπάνω·παραγράφου.·Αυτή·είναι·η·Ελληνική·μετάφραση·της·παραπάνω·παραγράφου.·Αυτή·είναι·η·Ελληνική·μετάφραση·της·παραπάνω·παραγράφου.¶

Αυτή·είναι·η·Ελληνική·μετάφραση·της·παραπάνω·παραγράφου.·Αυτή·είναι·η·Ελληνική·μετάφραση·της·παραπάνω·παραγράφου.·Αυτή·είναι·η·Ελληνική·μετάφραση·της·παραπάνω·παραγράφου.·Αυτή·είναι·η·Ελληνική·μετάφραση·της·παραπάνω·παραγράφου.·Αυτή·είναι·η·Ελληνική·μετάφραση·της·παραπάνω·παραγράφου.¶

Αυτή·είναι·η·μετάφραση·της·παραπάνω·πρότασης.¶



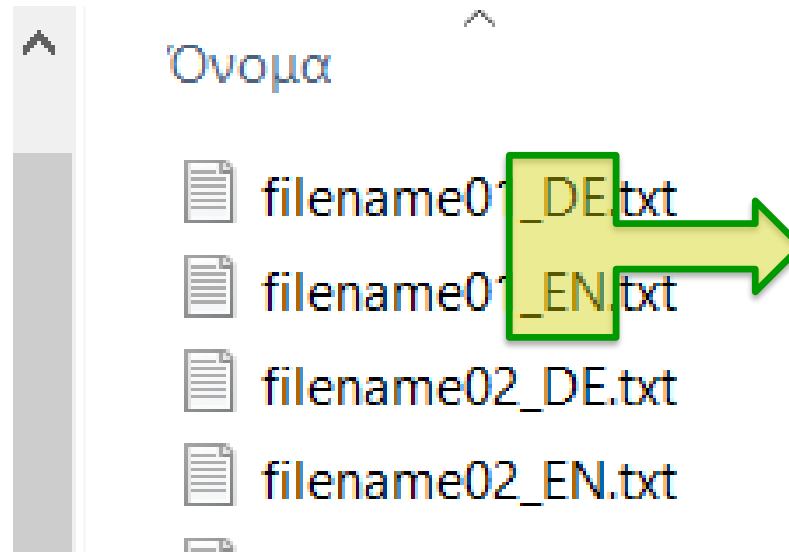
English	Greek
<p>This is the content of an English document. This is the content of an English document. This is the content of an English document. This is the content of an English document.</p> <p>This is the content of an English document.</p> <p>This is the content of an English document. This is the content of an English document. This is the content of an English document. This is the content of an English document. This is the content of an English document. This is the content of an English document. This is the content of an English document. This is the content of an English document.</p>	<p>Αυτή είναι η Ελληνική μετάφραση του κειμένου. Αυτή είναι η Ελληνική μετάφραση του κειμένου. Αυτή είναι η Ελληνική μετάφραση του κειμένου.</p> <p>Αυτή είναι η Ελληνική μετάφραση του κειμένου. Αυτή είναι η Ελληνική μετάφραση του κειμένου.</p> <p>Αυτή είναι η Ελληνική μετάφραση του κειμένου.</p> <p>Αυτή είναι η Ελληνική μετάφραση του κειμένου. Αυτή είναι η Ελληνική μετάφραση του κειμένου. Αυτή είναι η Ελληνική μετάφραση του κειμένου. Αυτή είναι η Ελληνική μετάφραση του κειμένου.</p>



- Όνομα
- 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 an identical filename for each document pair



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

- Administrative/regulatory domain and
- Topics relevant to the CEF DSIs

CEF DSI	Domain
Online Dispute Resolution	Consumers' rights
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](http://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 ▶

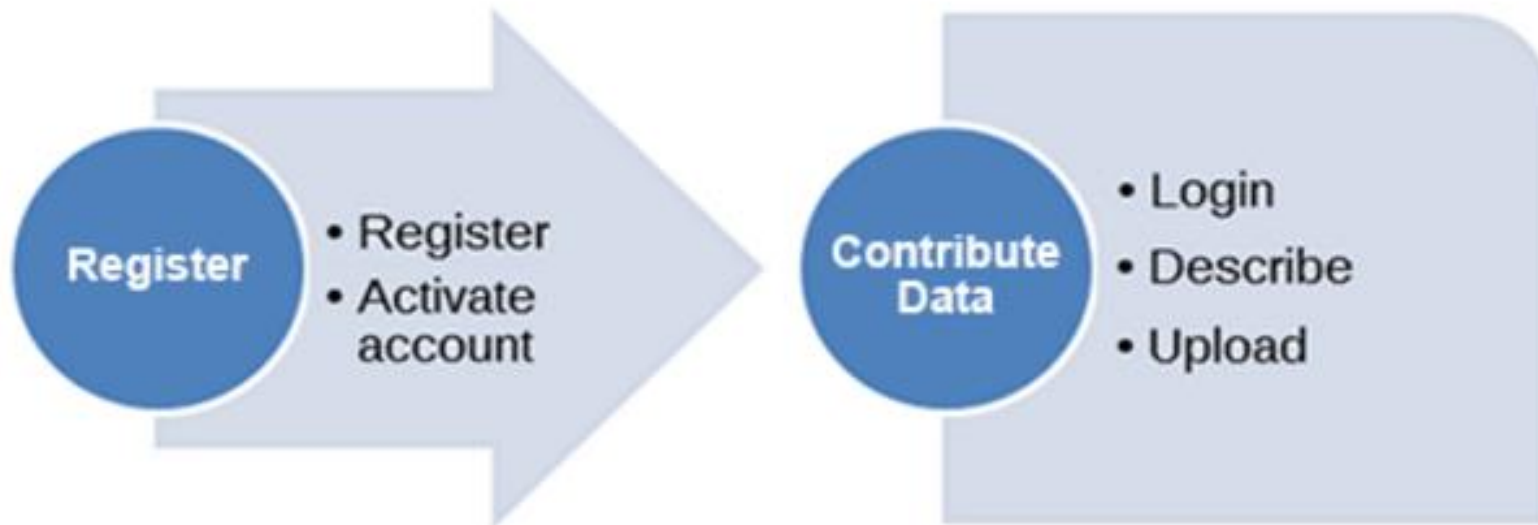




## ELRC-SHARE Repository



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

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)**

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



- Two modes for contributing your data

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.

Upload Resource\*

Choose File No file chosen

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

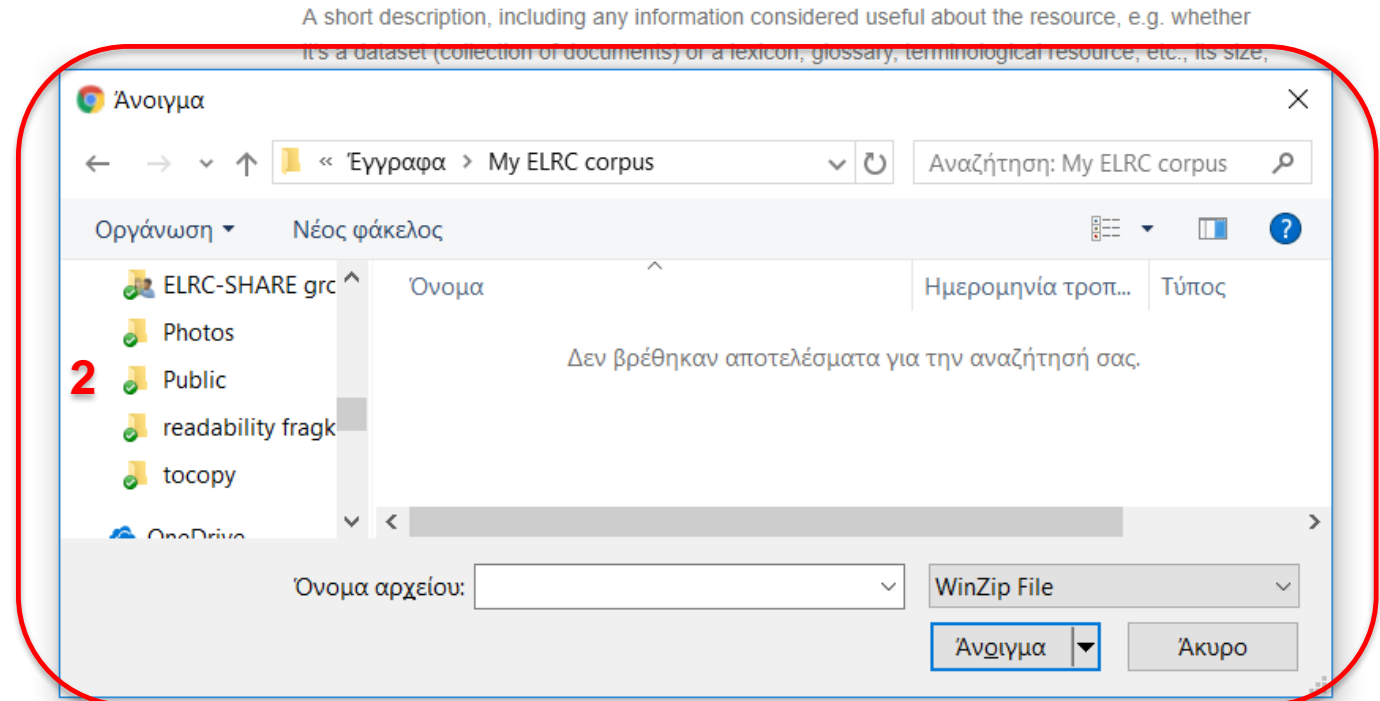
In case the **.zip file** file you wish to upload is larger than 100MB, please contact [elrc-share@ilsp.gr](mailto:elrc-share@ilsp.gr)

- Free file compression tools (indicative):
  - 7zip
  - PeaZip
  - Hamster Free Zip Archiver
  - Universal Extractor
  - ZipltFree
- 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



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@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\*

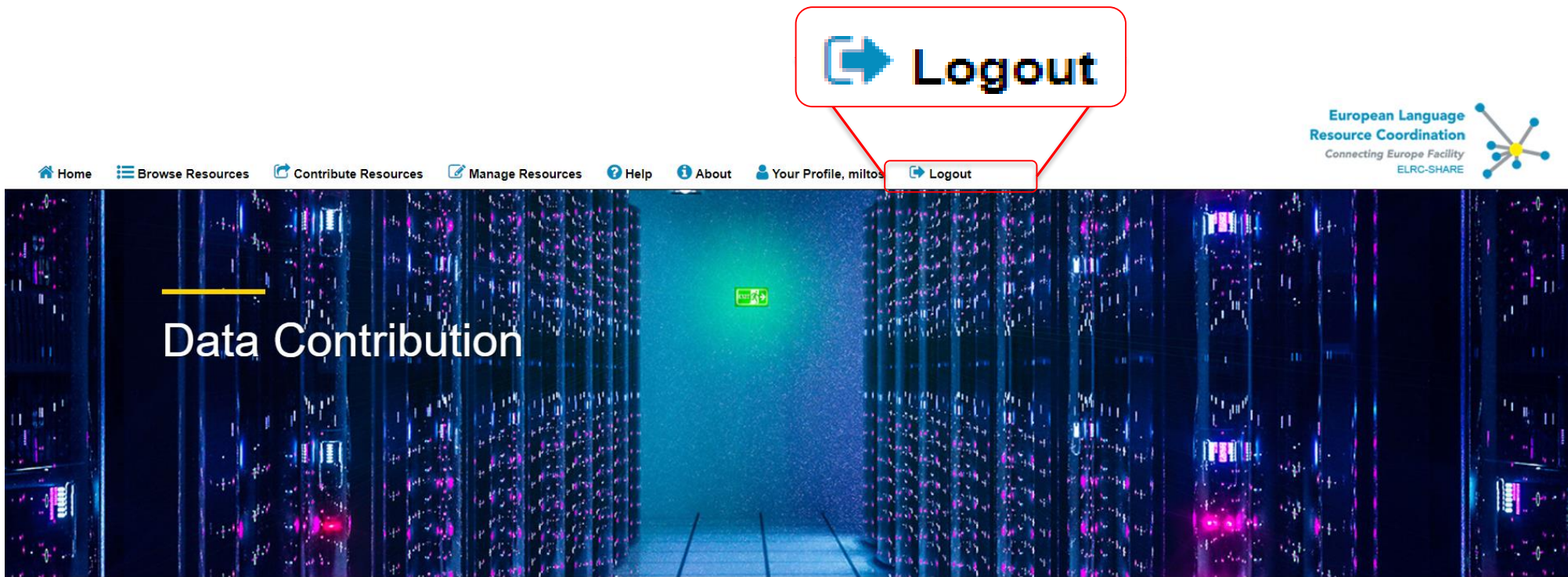
www

Please provide a URL containing the files you wish to contribute

Submit Reset

- Additional functionality for securely transferring sensitive data based on a deployment of the eDelivery CEF building block coming soon in the next version of ELRC-SHARE

- 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 callout box containing a blue arrow icon and the text 'Logout'. The background of the page is 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](#)



## Helpdesk for Language Resources

Telephone\* +33 970 440 522

Secretariat Support +49 681 857 7552 85

Skype ELRC Helpdesk

E-mail [help@lr-coordination.eu](mailto:help@lr-coordination.eu)

and/or request onsite assistance

(<http://www.lr-coordination.eu/request-onsite-assistance>)

Thank you for your attention!

