

# Cómo preparar y compartir datos: El repositorio ELRC y pasos a seguir

**Victoria Arranz**  
**ELDA**



## 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?

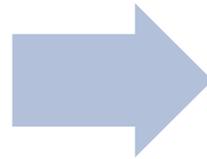
[Make a suggestion>>](#)

## Nociones básicas:

- **Datos:** Cualquier elemento de contenidos almacenado electrónicamente
- **Conjunto de datos (o recurso):** la colección de uno o más ficheros de datos **agrupados** de acuerdo a unos **criterios** particulares.
- **Metadatos:** *datos sobre los datos:* descripción de las características de un conjunto de datos (título, publicador, descripción del contenido, URL)

## Datos

- cualquier elemento de **contenido** almacenado electrónicamente



## Datos lingüísticos (textuales)

- cualquier elemento de **texto** almacenado electrónicamente

## 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

### 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](#)

text

### 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)

#### Metadata Creator

[Kanella Pouli](#)

[Maciej Ogrodniczuk](#)

# El concepto de “datos” en el contexto de eTranslation



Central Statistical Office Dataset

File01\_pl.txt  
File01\_en.txt  
File02\_pl.txt  
File02\_en.txt  
File03\_pl.txt  
File03\_en.txt

Trans  
Data

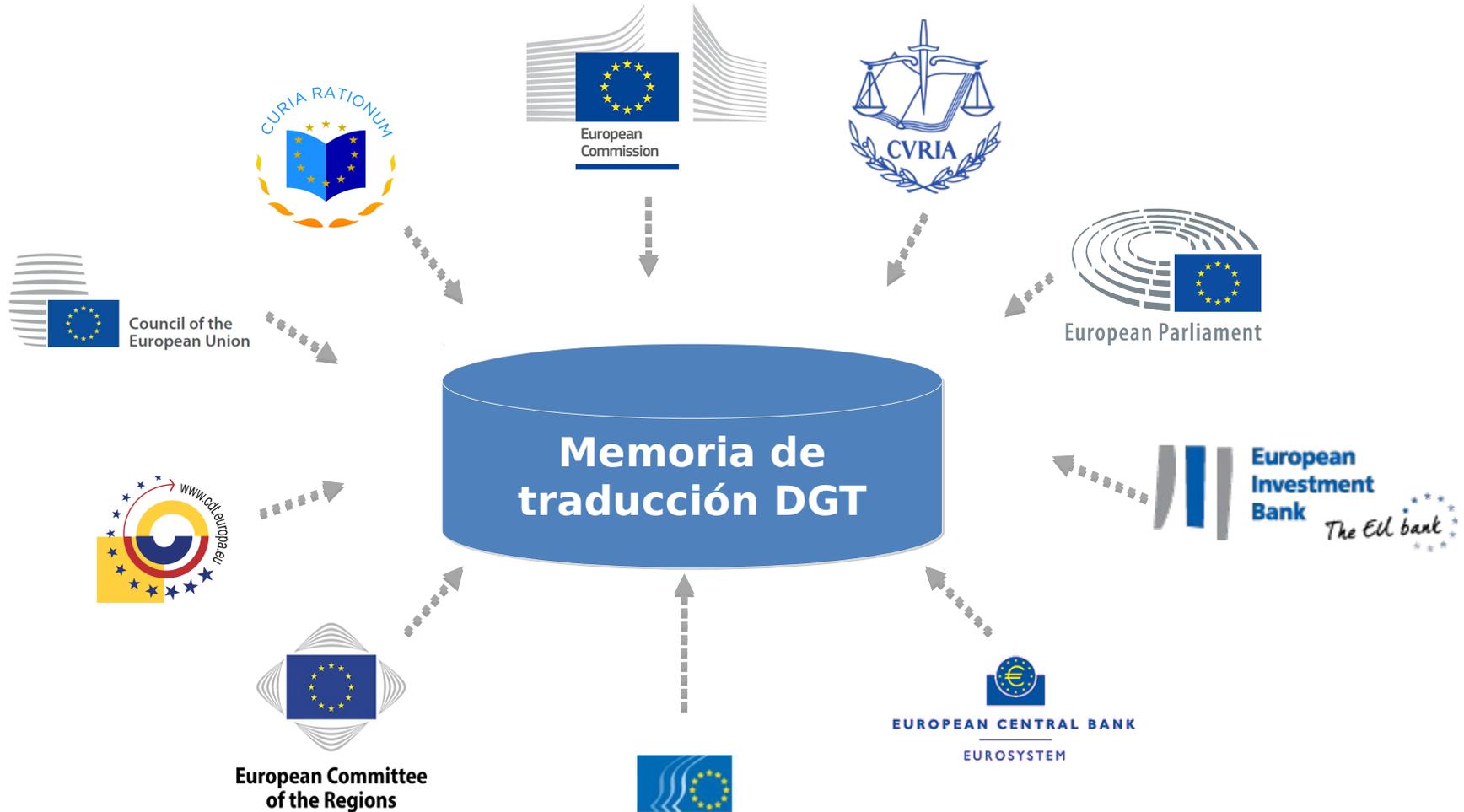
Distribution  
Availability  
Licence  
CC-BY 4.0  
Conditions: Attribution

Bilingual  
Language  
Polish (pl)  
English (en)  
Linguality  
Linguality type

Metadata  
Created: Sept. 18, 2016  
Last Updated: Dec. 15, 2016  
Metadata Language: English (en)  
Metadata Creator  
Kanela Pouli #5  
Maciej Orzadniczuk #5

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.



**Estos datos ya están disponibles  
PERO  
no son suficientes ...**

- Datos de las organizaciones públicas locales, producidas por ellas mismas o externalizados, por ejemplo:
  - Informes
  - Comunicaciones
  - Noticias
  - Contenidos Web que tengan versiones en varios idiomas
  - Leyes
  - Terminologías
  - Archivos
  - Formularios
  - Documentos de Preguntas Frecuentes



- En principio, cualquier texto almacenado electrónicamente en cualquiera de las lenguas de la UE, o noruego e islandés (las lenguas CEF)
- Lo ideal son textos y sus traducciones en una o más de las lenguas CEF (textos paralelos bilingües o multilingües)

## Texto español

Informe del Presidente de la Oficina de Contratación Pública sobre el funcionamiento del sistema de contratación pública en 2010.  
El informe sobre el funcionamiento del sistema de contratación pública abarca el período comprendido entre el 1 de enero y el 31 de diciembre de 2011 y se ha preparado sobre la base de la información obtenida en documentos oficiales y publicaciones, así como de otros documentos, informes y análisis sobre contratación pública de la Oficina de Contratación Pública disponibles.

## Traducción al inglés

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.

- También son útiles listas de términos y sus traducciones, es decir, una **terminología**

	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>



- En principio, cualquier texto en un formato legible por máquina
- Pero, algunos formatos son más fáciles de procesar que otros
- Cuanto más difíciles de procesar, más posibilidad de introducir errores, que son un problema para eTranslation

## **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*

---

- Los siguientes formatos son especialmente interesantes (en orden descendiente):
  - Para textos paralelos (documento y su traducción)
    1. Memorias de traducción (.tmx)
    2. Ficheros de traducción en XML (.xliff)
    3. Texto sin formato (.txt, .csv)
    4. Hojas de cálculo (.xlsx)
  - Para terminologías
    1. TermBase eXchange (.tbx)
    2. Texto sin formato (.txt, .csv)
    3. Hojas de cálculo (.xlsx)
  - Para textos monolingües
    1. Texto sin formato (.txt, .csv)

# Formato de ficheros de textos paralelos y su manipulación

# Preparar los datos|1

NO



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.

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

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.



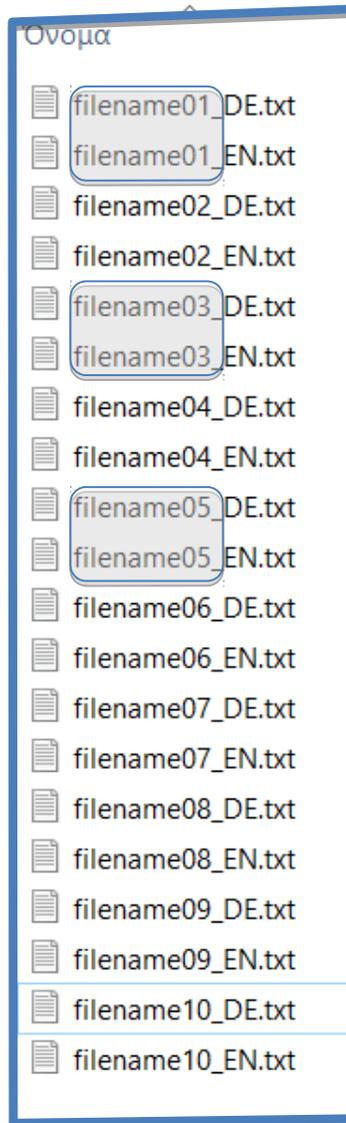


NO

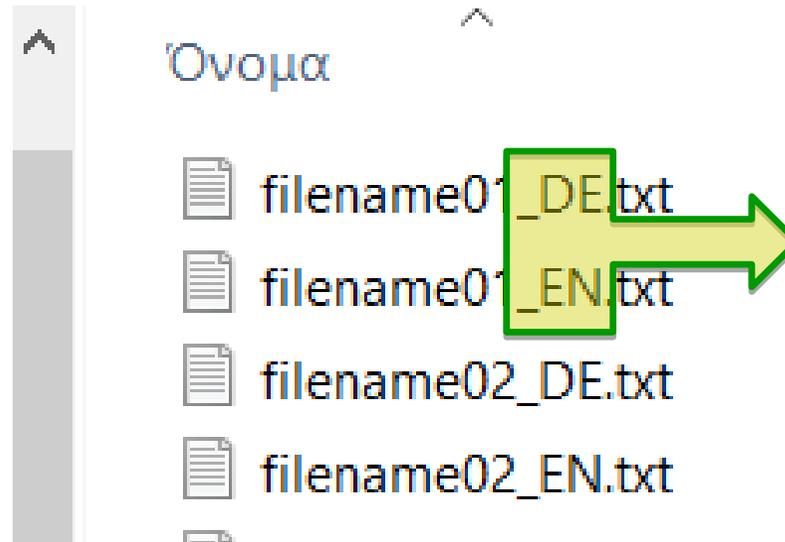


English	Greek
<p>This is the content of an English document. This is the content of an English document. This is the content of an English document. This is the content of an English document.</p> <p>This is the content of an English document.</p> <p>This is the content of an English document. This is the content of an English document.</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>

# Preparar los datos|4



**Utilizar el mismo  
nombre para ambos  
ficheros**



Incluir  
identificadores  
de idioma en el  
nombre del  
fichero



- Recuerde: un conjunto de datos es una colección de datos **agrupada de acuerdo a unos criterios específicos**
- Para mejorar y adaptar eTranslation hay dos criterios críticos:
  - **Idioma(s)**: cada colección se define por el idioma o pares de idiomas de sus datos, por ejemplo
    - *Colección de textos en inglés - alemán*
    - *Documentos en inglés - noruego - francés*
  - **Dominio**: cada colección pertenece idealmente a un único dominio, por ejemplo
    - *Colección de textos en inglés - alemán del dominio: cultura*
    - *Documentos de Seguridad Social en inglés - noruego - francés*

- Administrativo / reglamentos
- Temas relevantes para los servicios digitales (DSI) CEF

CEF DSI	Dominio
Online Dispute Resolution	Derechos de los consumidores
Electronic Exchange of Social Security Information	Seguridad Social
eProcurement	Compra pública, contratos, acuerdos ...
European e-Justice Portal	Justicia, Leyes
eHealth	Salud, Medicina
Business Registers Interconnection System	Mercantil, empresarial
Safer Internet	Seguridad en internet
Cybersecurity	Ciberseguridad
Public Open Data	Datos abiertos públicos
Europeana	Cultura

# **Cómo aportar datos a CEF eTranslation**

## **Una guía paso a paso**

[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!

 Contribute Resources

[Home](#) [Browse Resources](#) [Contribute Resources](#) [Manage Resources](#) [Help](#) [About](#) [Your Profile, milto](#) [Logout](#)

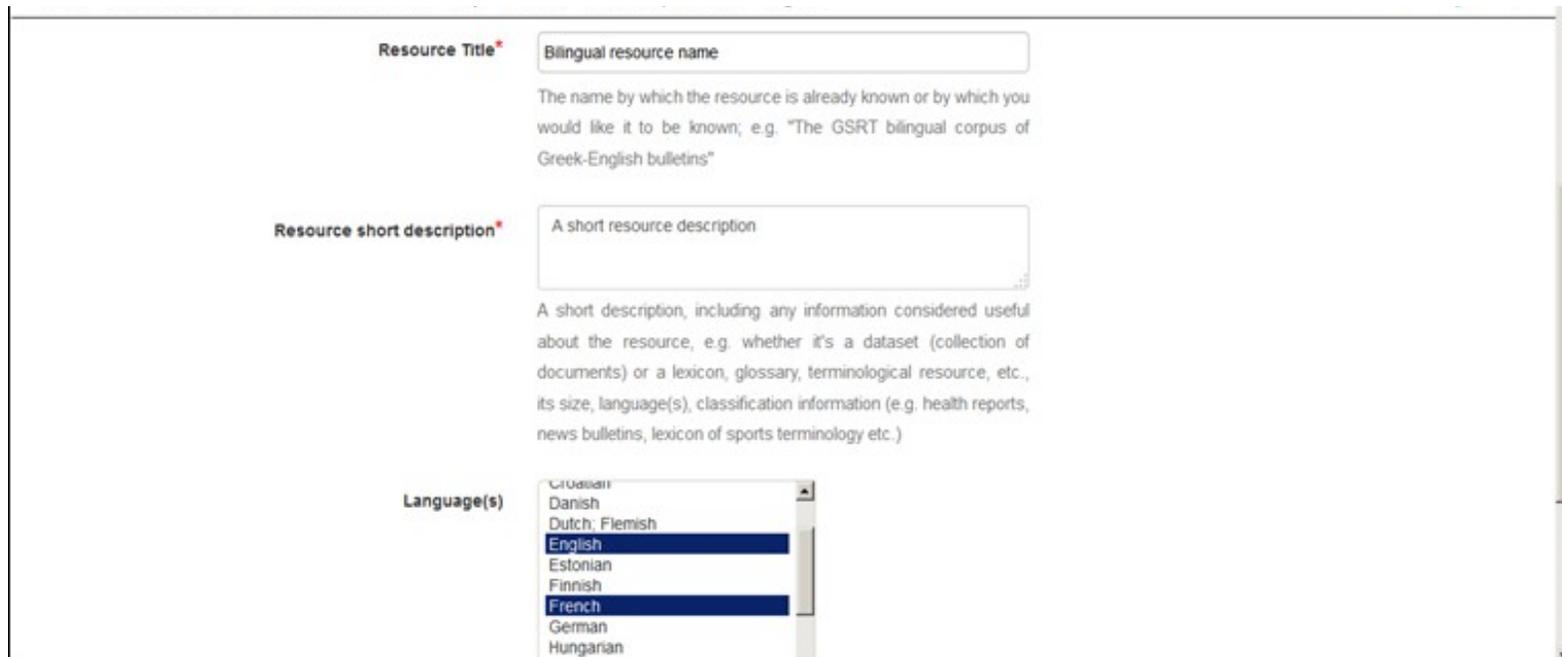
## Data Contribution

### 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"

- Hay que rellenar los metadatos que describen el conjunto de datos



The screenshot shows a 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 text input field 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 with a scroll bar. The visible options are: "Croatian", "Danish", "Dutch, Flemish", "English", "Estonian", "Finnish", "French", "German", and "Hungarian". "English" and "French" are highlighted in blue.



- Hay dos maneras de introducir los datos

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)



- Se pueden utilizar los programas de compresión más conocidos :
  - 7zip
  - PeaZip
  - Hamster Free Zip Archiver
  - Universal Extractor
  - ZipltFree
- o la función de compresión de Windows



- También puede indicar una url (lista de carpetas)

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

- En la próxima versión de ELRC-SHARE se incorporará una funcionalidad adicional para transferir de forma segura datos sensibles basada en un desarrollo del componente CEF eDelivery



## Help

### Documentation on the ELRC-SHARE editor

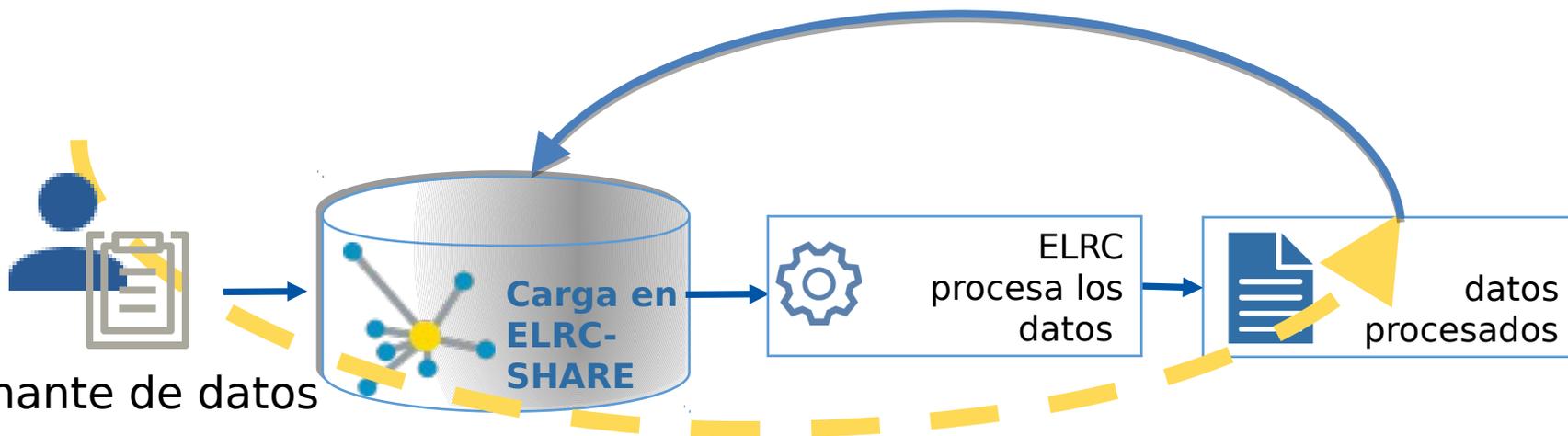
The following guidelines provide detailed information on how to use the editing facility for documenting and uploading LR:

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

### ELRC-SHARE schema

- [ELRC-SHARE schema XSD](#) (based on the META-SHARE Schema)
- [Documentation about the schema](#)

**¿Y después?**



- Todos los conjuntos de datos son procesados para convertirlos en ficheros tmx/tbx/txt
- De forma general, los datos sufrirán el siguiente procesamiento:
  - limpieza
  - conversión de formato
  - alineación de segmentos
  - anotación completa de metadatos



También pueden prestarse todos estos servicios  
[in situ y de forma gratuita](#)





**Nuestro equipo de expertos  
viajará directamente para  
asistirle en sus oficinas**

**Nosotros resolveremos los problemas que hayan surgido con los datos y entregaremos los datos procesados directamente al donante. También podemos ayudarle a mejorar sus procesos de gestión de datos.  
¡Pregúntenos!**



## Extracción de datos

Si sus datos están atrapados en bases de datos o archivos, podemos ayudarle a extraerlos



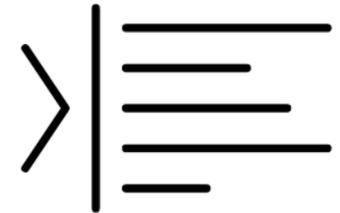
## Anonimización

¿Contienen sus recursos datos privados? Podemos ayudarle a eliminarlos



## Limpieza

Si sus datos están mezclados con código, los limpiaremos



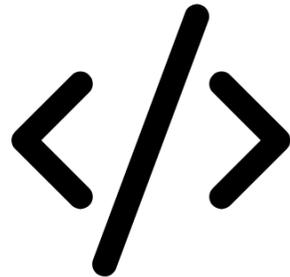
## Reformateo

Hay que reformatear los datos, de DOCX a XML, o de PDF a TXT? ¡Lo haremos por usted!



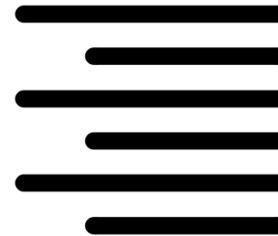
## Conversión de datos

Si los datos no están en los formatos adecuados, le ayudaremos a convertirlos



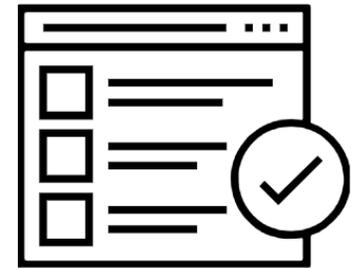
## Borrado de etiquetas

¿Contienen sus datos etiquetas innecesarias?  
Podemos hacer que desaparezcan



## Alineación

¿Las traducciones no están alineadas?  
Nuestras herramientas lo harán



## Metadatos

¡Los metadatos son cruciales! Podemos organizar y validar los metadatos necesarios

# Cómo solicitar asistencia y servicios



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

[Ir-coordination.eu/request-onsite-assistance](http://Ir-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-cooridantion.eu">help@lr-cooridantion.eu</a>

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

# ¡Póngase en contacto con nosotros!



**¡Gracias por su interés!**





- 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