European Language Resource Coordination
Connecting Europe Facility

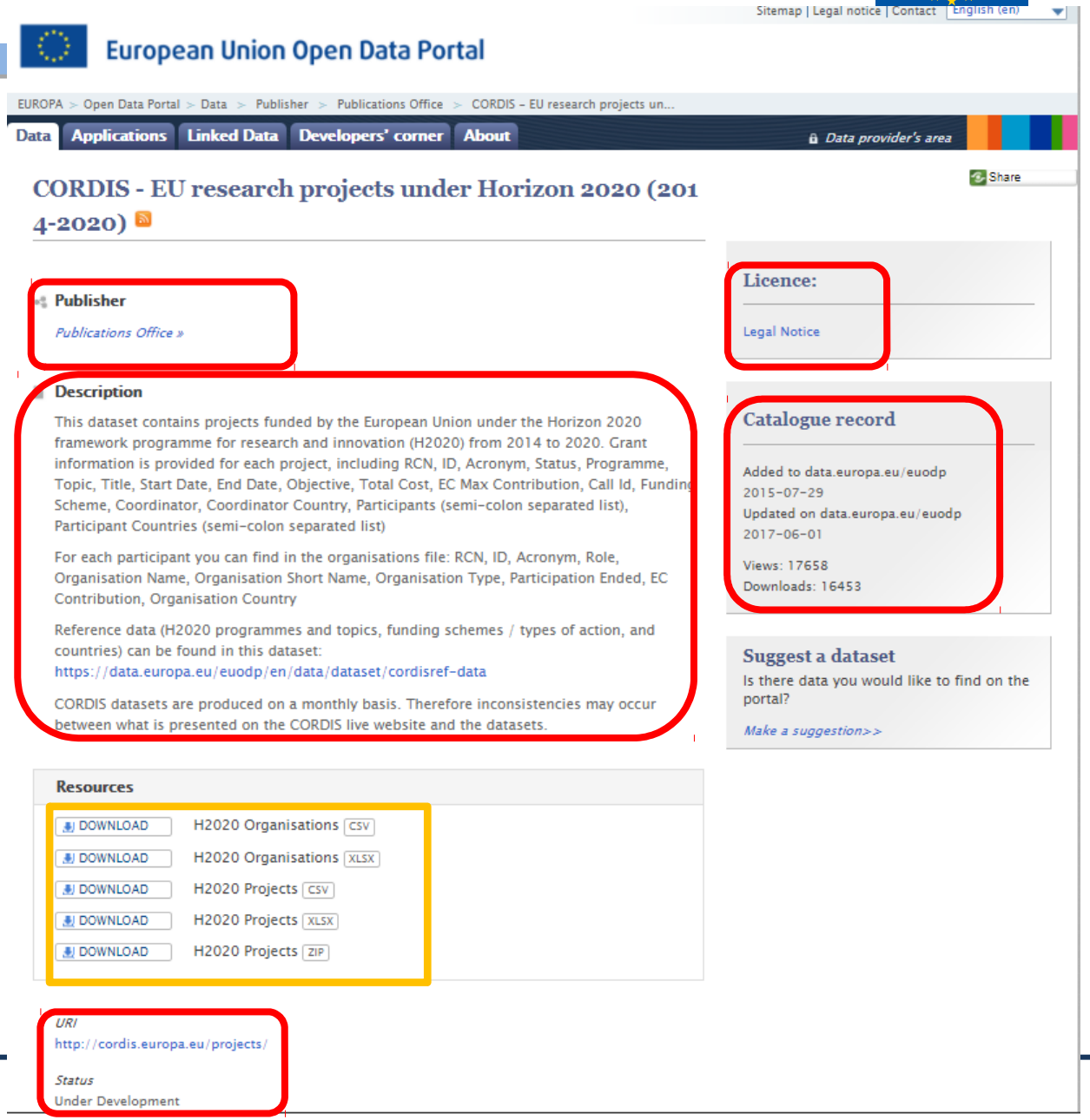
Cómo preparar y compartir datos: El repositorio ELRC y pasos a seguir

Victoria Arranz
ELDA

El concepto “datos”

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)



European Language Resource Coordination
Connecting Europe Facility

El concepto “datos lingüísticos”

Datos

- cualquier elemento de contenido almacenado electrónicamente

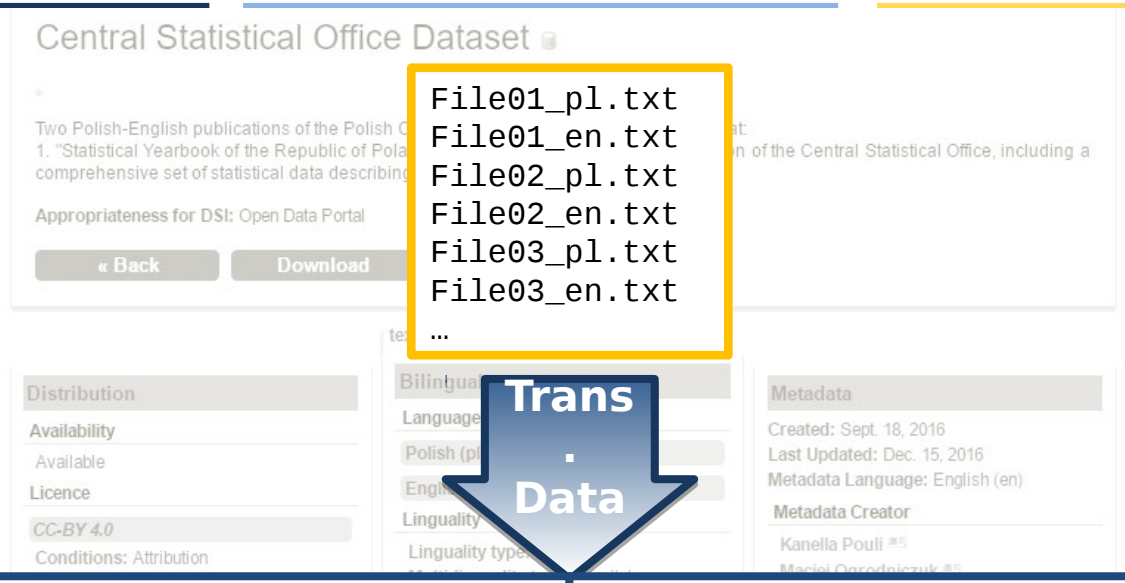
Datos lingüísticos (textuales)

- cualquier elemento de texto almacenado electrónicamente

European Language Resource Coordination
Connecting Europe Facility

El concepto de “datos” en el contexto de eTranslation

El concepto de “datos” en el contexto de eTranslation




```
File01_pl.txt
File01_en.txt
File02_pl.txt
File02_en.txt
File03_pl.txt
File03_en.txt
…
```

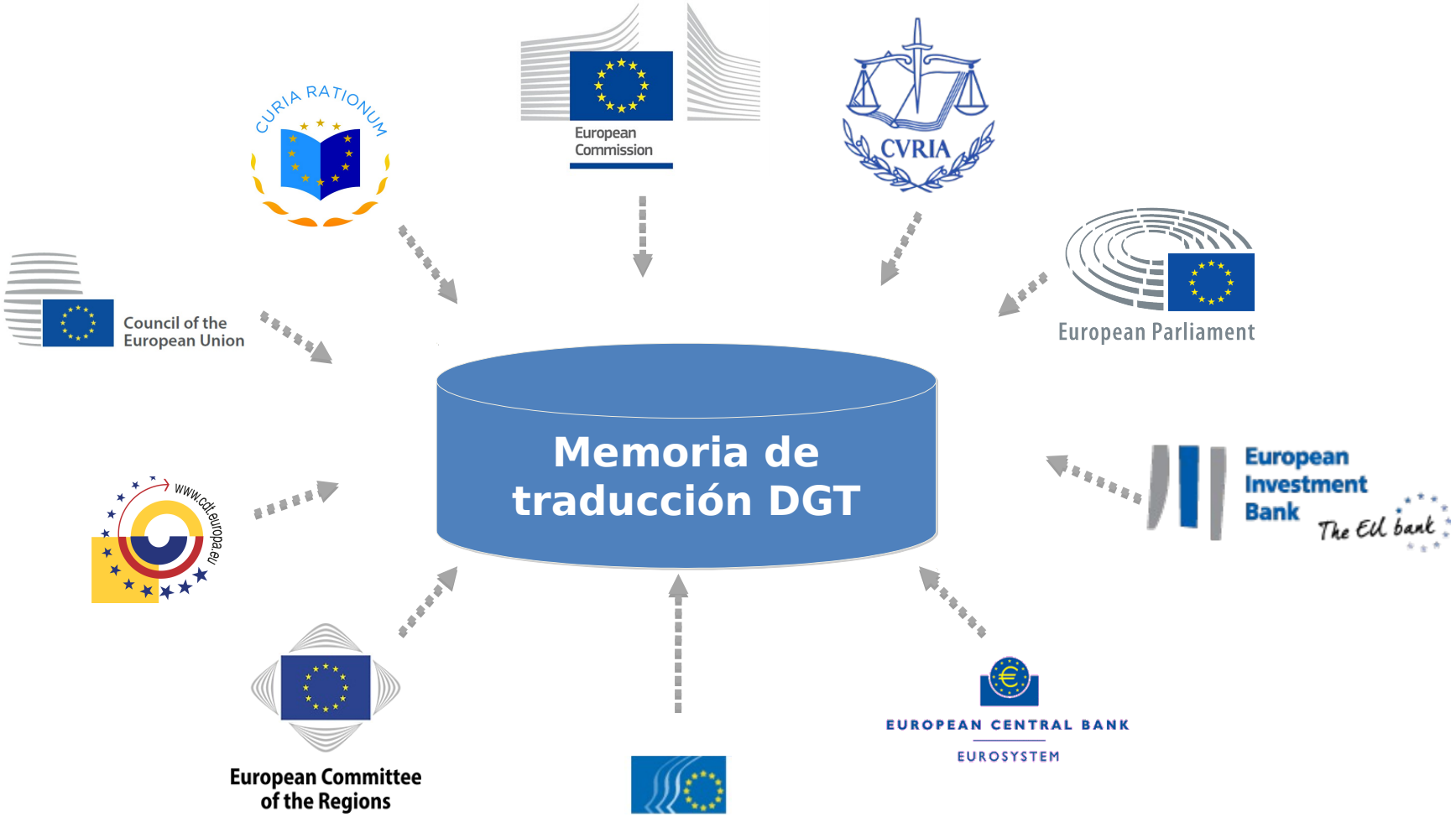
Trans . Data

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.

European Language Resource Coordination
Connecting Europe Facility

Datos utilizados por eTranslation



Estos datos ya están disponibles
PERO
no son suficientes …

European Language Resource Coordination
Connecting Europe Facility

¿Qué más necesita eTranslation?

- 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

Tipos de datos útiles para eTranslation|1

- 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	Traducción al inglés
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.	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.

Tipos de datos útiles para eTranslation |2

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

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

European Language Resource Coordination
Connecting Europe Facility

Formatos de datos útiles para eTranslation|1

- 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

Formato de ficheros para datos paralelos

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

European Language Resource Coordination
Connecting Europe Facility

Formatos de datos útiles para eTranslation|2

- 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)

European Language Resource Coordination
Connecting Europe Facility

Formato de ficheros de textos paralelos y su manipulación

Preparar los datos|1

NO

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 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.

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.

European Language Resource Coordination
Connecting Europe Facility

Preparar los datos|2

NO

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·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·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.¶

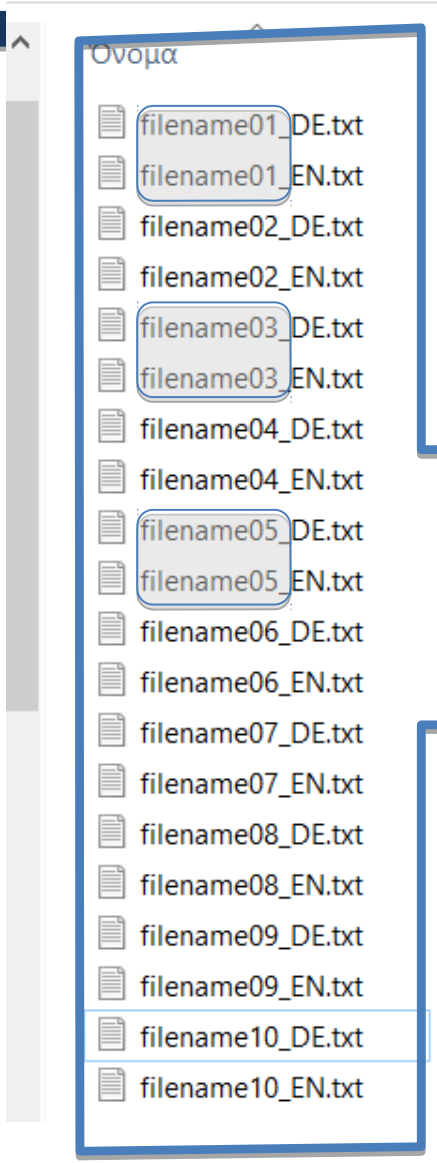
European Language Resource Coordination
Connecting Europe Facility

Preparar los datos|3

NO

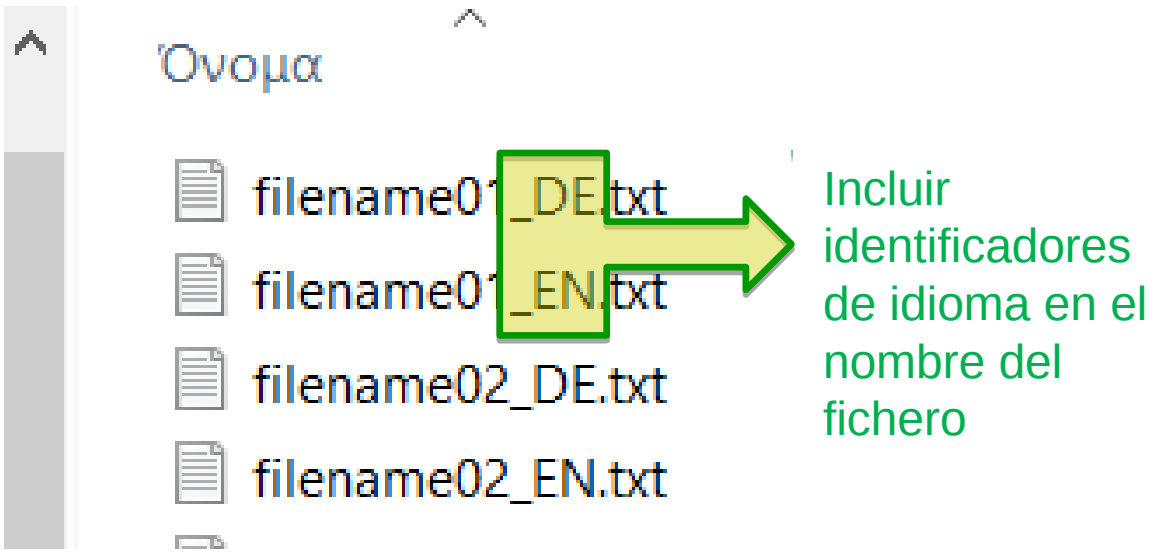
English	Greek
This is the content of an English document. This is the content of an English document. This is the content of an English document. This is the content of an English document. This is the content of an English document. This is the content of an English document. This is the content of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

Preparar los datos|4



Utilizar el mismo nombre para ambos ficheros

Preparar los datos|5



Criterios para agrupar los datos

- 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*

European Language Resource Coordination
Connecting Europe Facility

Prioridades en cuanto a dominios

- 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

European Language Resource Coordination
Connecting Europe Facility

Cómo aportar datos a CEF eTranslation
Una guía paso a paso

European Language Resource Coordination
Connecting Europe Facility

En ELRC-SHARE

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 fo relevant data	Click below if you are a language resource owner and are willing to share it for the purposes of CEF.AT
Data sources submission ▶	**Language resource submission ▶**

European Language Resource Coordination
Connecting Europe Facility

Repositorio ELRC-SHARE

European Language Resource Coordination
Connecting Europe Facility

Cómo aportar recursos (1/4)

New Resource

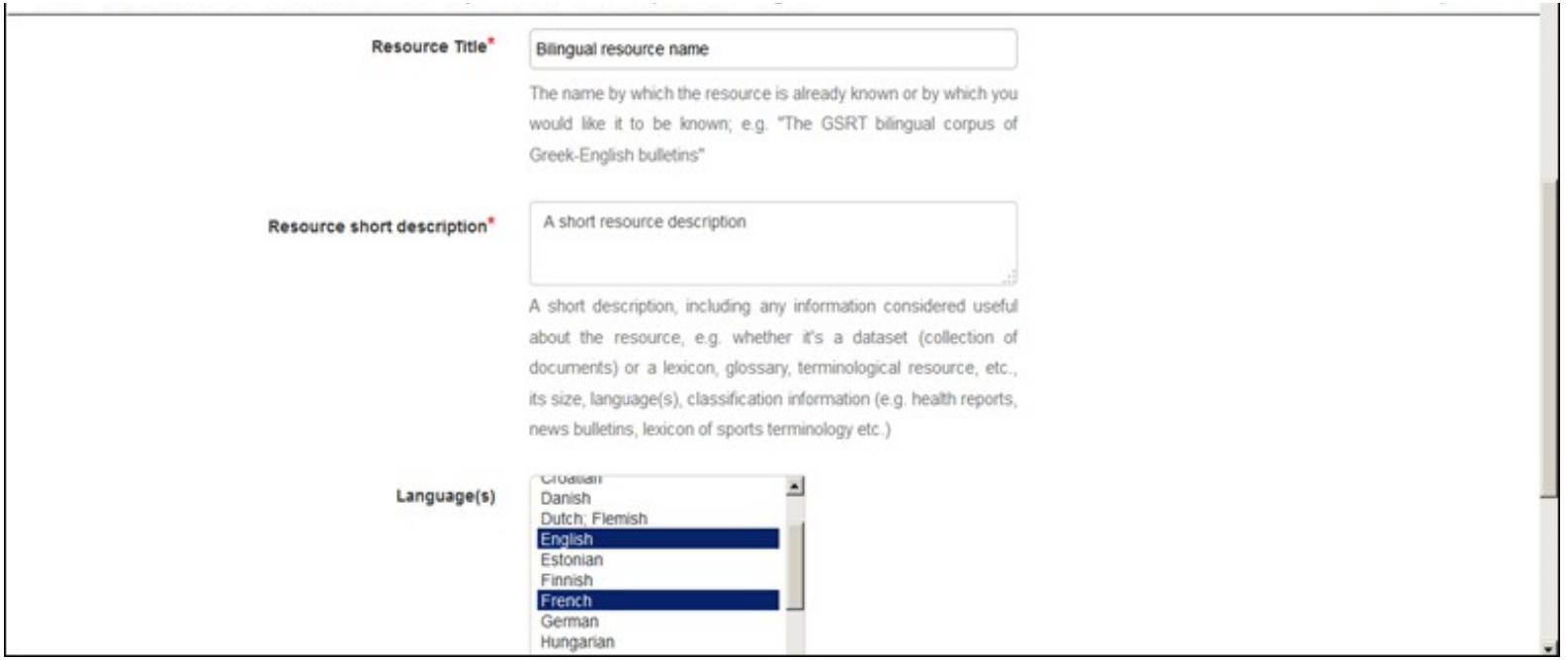
Resource Title*

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

European Language Resource Coordination
Connecting Europe Facility

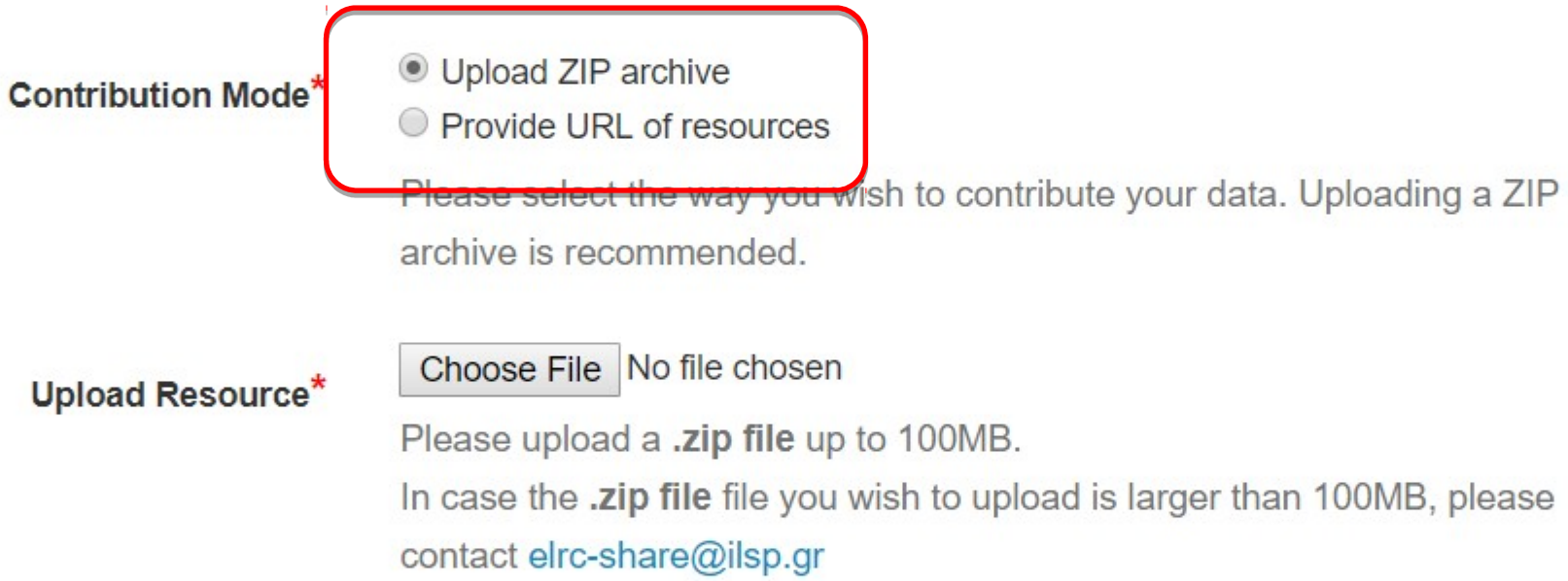
Cómo aportar recursos (2/4)

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



Cómo aportar recursos (3/4)

- Hay dos maneras de introducir los datos

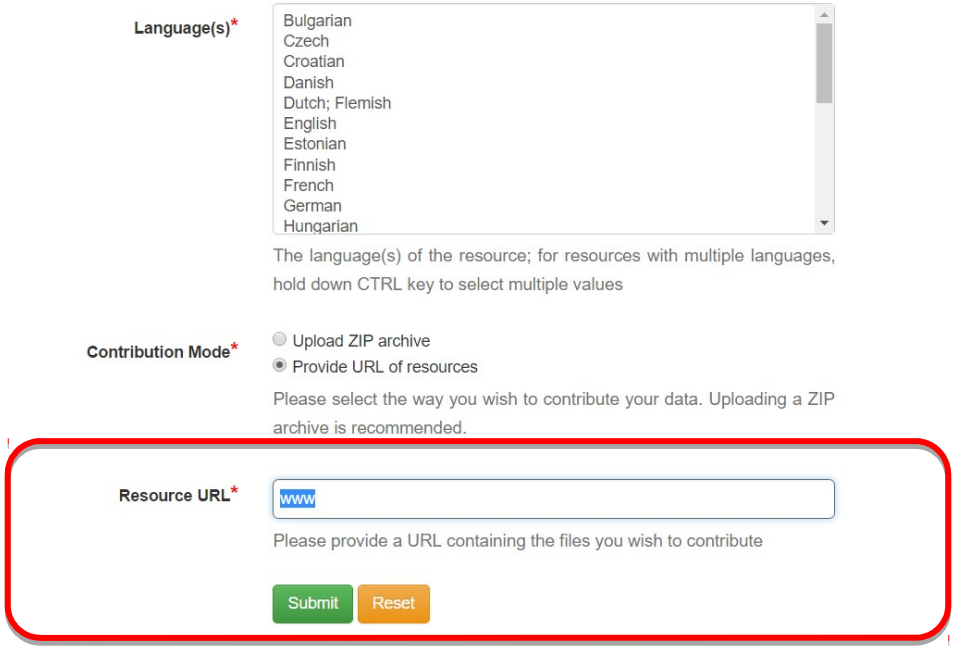


Cómo crear un fichero .zip

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

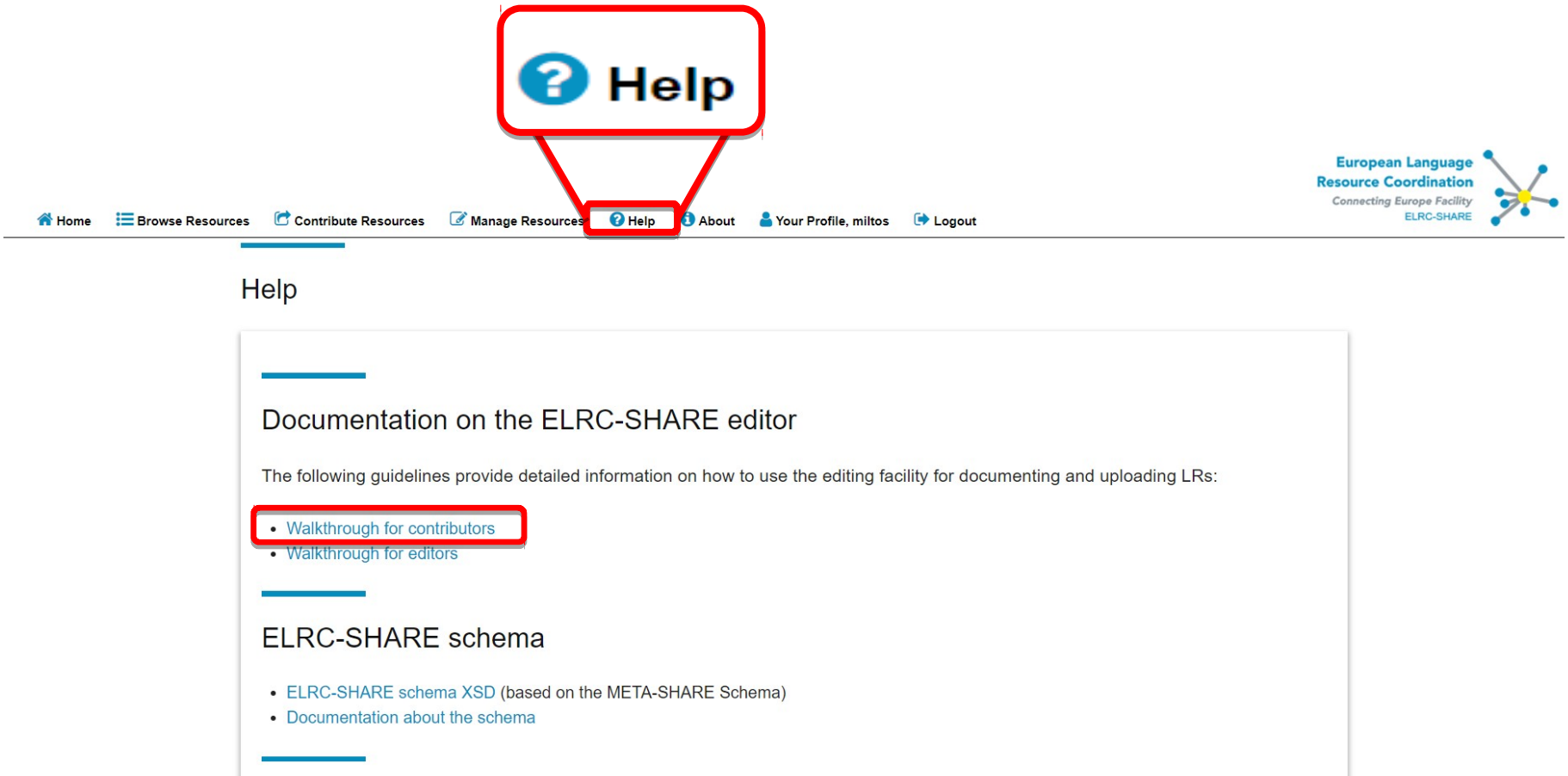
Cómo aportar recursos (4/4)

- También puede indicar una url (lista de carpetas)



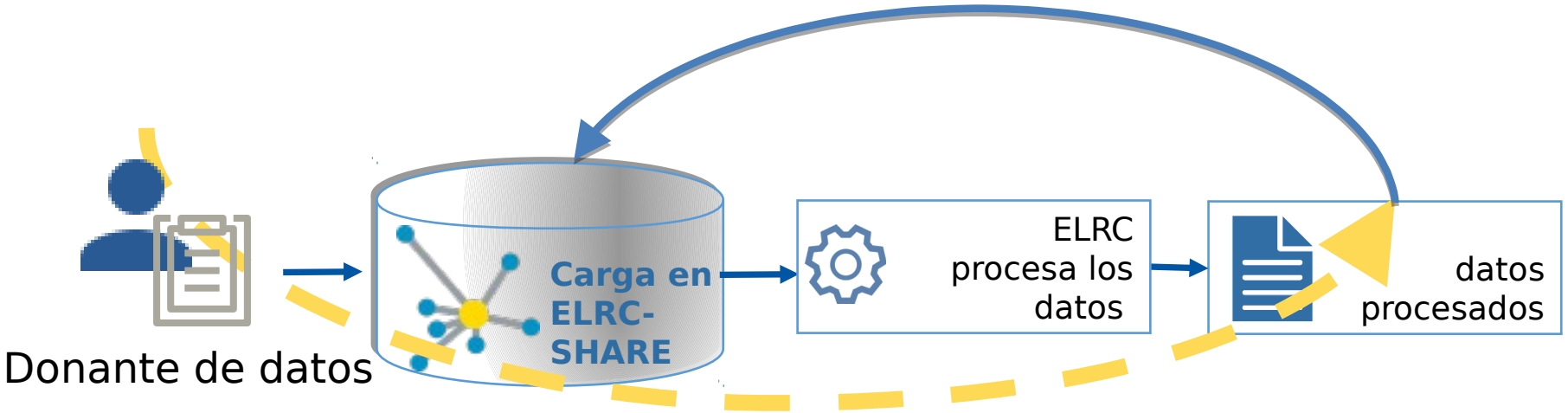
- En la próxima version de ELRC-SHARE se incorporará una funcionalidad adicional para transferir de forma segura datos sensibles basada en un desarrollo del componente CEF eDelivery

Guías para los usuarios



¿Y después?

¿Qué se hará con los datos aportados?



European Language Resource Coordination
Connecting Europe Facility

Procesamiento de los datos antes de entregárselos a la CE

- 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

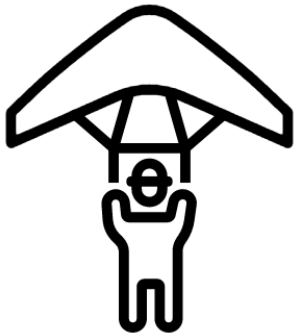
European Language Resource Coordination
Connecting Europe Facility

Asistencia in situ

También pueden prestarse todos estos servicios
in situ y de forma gratuita

European Language Resource Coordination
Connecting Europe Facility

Asistencia in situ



Nuestro equipo de expertos viajará directamente para asistirle en sus oficinas

European Language Resource Coordination
Connecting Europe Facility

Asistencia in situ

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

European Language Resource Coordination
Connecting Europe Facility

Servicios in situ de procesamiento lingüístico

Extracción de datos

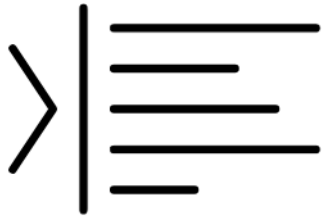
Si sus datos están atrapados en bases de datos o archivos, podemos ayudarle a extraerlos

Anonimización

¿Contienen sus recursos datos privados? Podemos ayudarle a eliminarlos

Limpieza

Si sus datos están mezclados con código, los limpiaremos



Reformateo

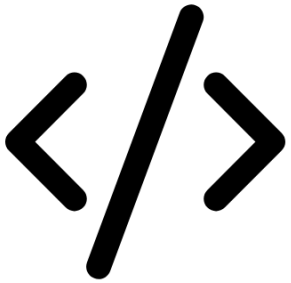
Hay que reformatear los datos, de DOCX a XML, o de PDF a TXT? ¡Lo haremos por usted!

European Language Resource Coordination
Connecting Europe Facility

Servicios de procesamiento lingüístico

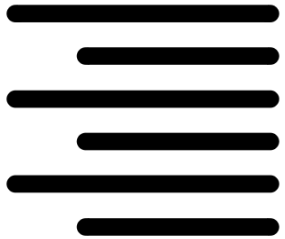
Conversión de datos

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



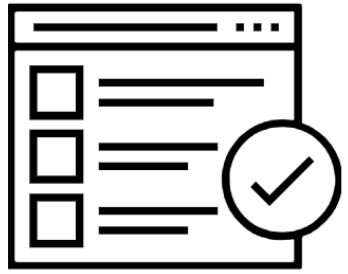
Borrado de etiquetas

¿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

European Language Resource Coordination
Connecting Europe Facility

Cómo solicitar asistencia y servicios

European Language Resource Coordination
Connecting Europe Facility

ELRC asistencia in situ

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

European Language Resource Coordination
Connecting Europe Facility

ELRC Helpdesk

Home Discover Resources Services Events Anchor Points News Helpdesk

European Language Resource Coordination

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	help@lr-cooridantion.eu

lr-coordination.eu/helpdesk

European Language Resource Coordination
Connecting Europe Facility

¡Póngase en contacto con nosotros!

European Language Resource Coordination
Connecting Europe Facility

¡Gracias por su interés!

European Language Resource Coordination
Connecting Europe Facility

Icons used in this presentation

- By Michael Mellon, GB, , CC-BY 3.0 US
- By Joana Pereira, BR, CC-BY 3.0 US
- By Becca O'Shea, NZ, CC-BY 3.0 US
- By Creative Stall, Basic licence www.iconfinder.com
- By Creative Stall, PK, CC-BY 3.0 US
- By Arthur Shlain, IL, CC-BY 3.0 US
- By Shmidt Sergey, US, CC-BY 3.0 US
- By Gregor Cresnar, CC-BY 3.0 US
- By anbileru adaleru, CC-BY 3.0 US
- By Vectors Market, CC-BY 3.0 US