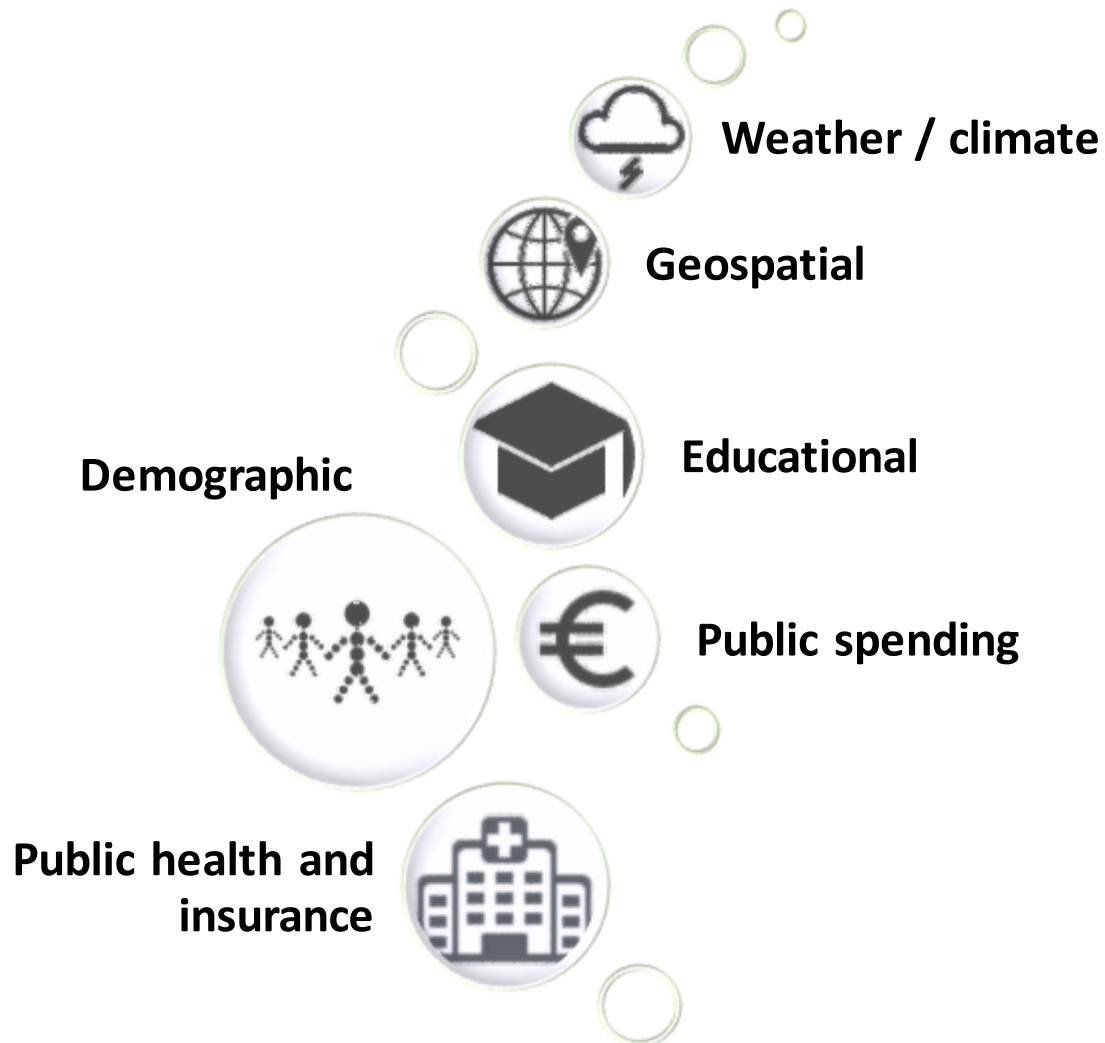


Preparing and sharing data with the ELRC repository and what happens next

Maria Giagkou

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

The notion of data



The notion of data



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?

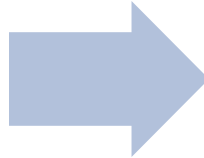
[Make a suggestion>>](#)

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)

Data

- any piece of electronically stored **content**



(Textual) Language Data

- any piece of electronically stored **text**

Secretariat-General parallel corpus SL-EN and EN-SL (part 1) 📄

English-Slovenian parallel corpus in TMX format from the Secretariat-General of the Government of the Republic of Slovenia in the legal domain

DSI Relevance: eJustice

[← Back](#) [Download](#)

Distribution

Availability: Available

Licences

[Terms for PSI-compliant resources](#)

[Open Under-PSI](#)

Distribution Details

Contact Person

[Simon Krek](#) 🇸🇮

text 📄

Bilingual text corpus

Languages

English (en)

Slovenian (sl)

Linguality

Linguality type: Bilingual

Multi-linguality type: Parallel

Text Format

TMX

Size

2,708,160 Words

55,184 Translation Units

Character encoding

UTF-16LE

Domains

LAW (Eurovoc 12)

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: 16/02/2017

Last Updated: 28/02/2017

Metadata Language: English (en)

Metadata Creator

[Kanella Pouli](#) 🇸🇮

[Miltos Deligiannis](#) 🇸🇮

Relations

Related Resource: Secretariat-General parallel corpus SL-EN and EN-SL (part 1) (Processed)

Relation Type: Has Version

Secretariat-General parallel corpus SL-EN and EN

English-Slovenian parallel corpus in TMX format from the Secretariat-General of the Government of the Republic of Slovenia

DSI Relevance: eJustice

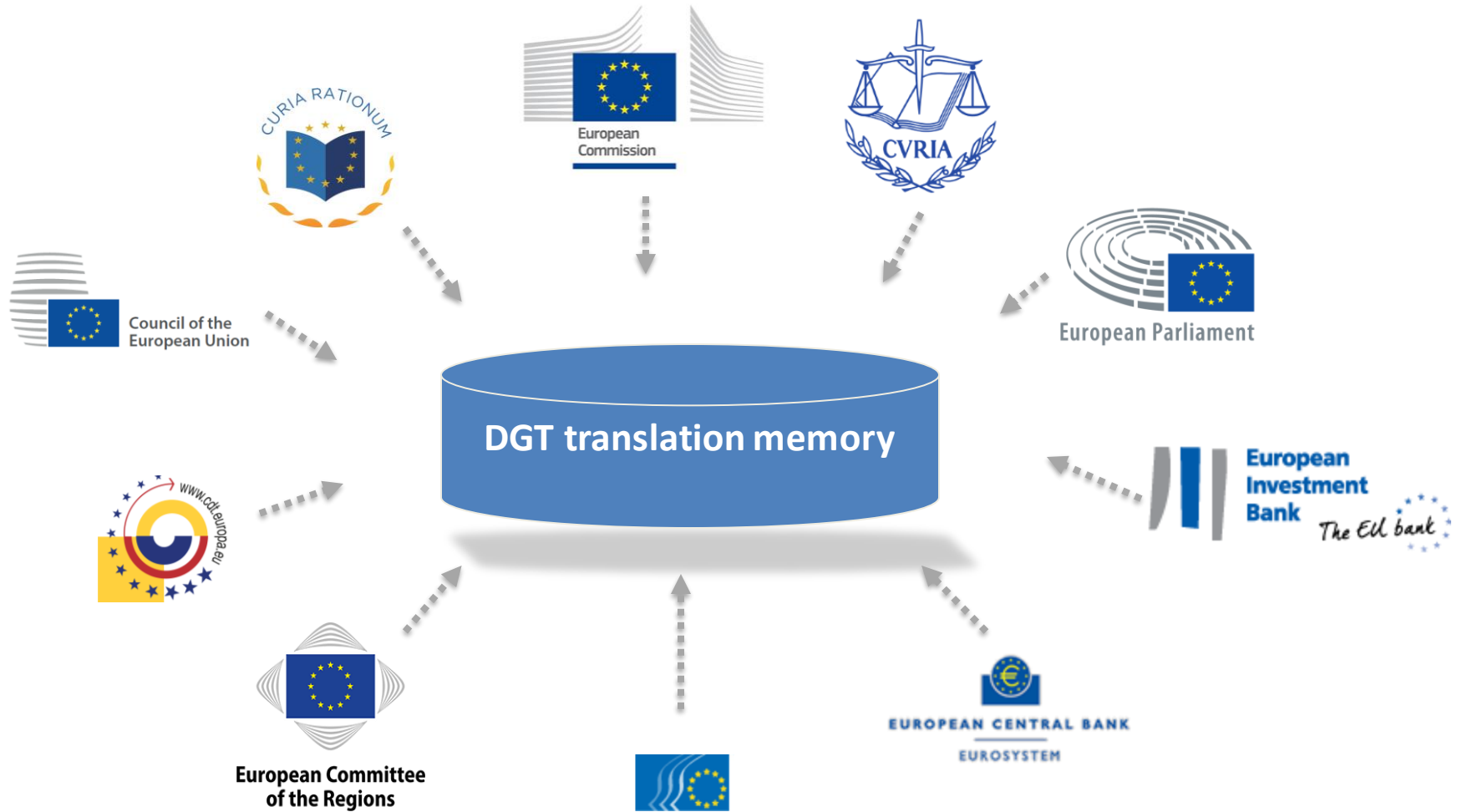
[← Back](#) [Download](#)

File01_sl.txt
File01_en.txt
File02_sl.txt
File02_en.txt
File03_sl.txt
File03_en.txt
...

Trans.
Data

posodobitev
infrastrukture,
rehabilitacija,
izboljšanje in
varovanje okolja ter
boljša
pripravljenost za
ukrepanje ob
naravnih in drugih
nesrečah

Modernisation of
infrastructure,
rehabilitation,
improvement and
protection of the
environment, and
improved capacity
to act in cases of
natural and other
disasters



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

- Data residing in local public organisations, produced in-house or outsourced, e.g.
 - Reports
 - Communication
 - News
 - Web Content that is managed for several languages
 - 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)

Slovenian text

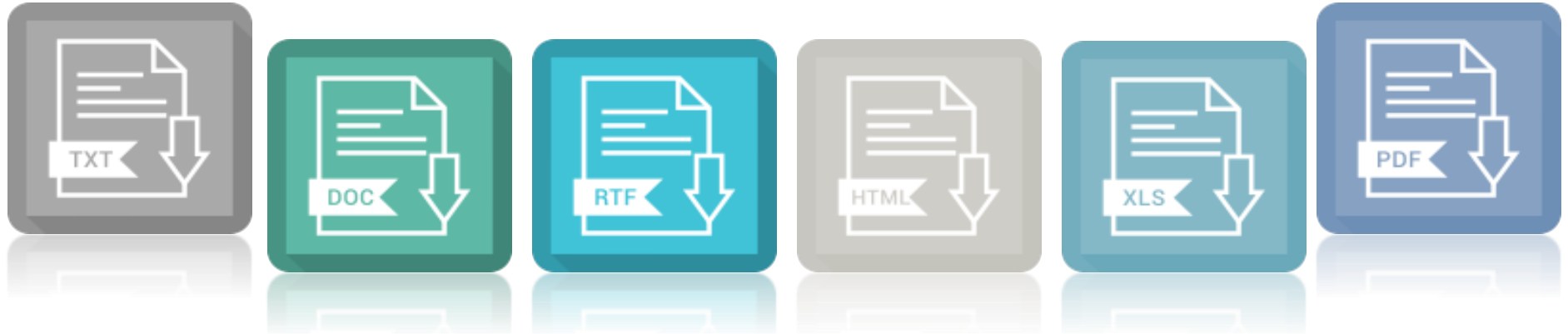
Pobudo za izvedbo konkretnega programa ali projekta da skupni odbor iz 6. člena tega sporazuma ali posamezni nosilci in izvajalci konkretne dejavnosti ali projekta razvojnega sodelovanja. Projekti in dejavnosti, ki se financirajo v okviru programa razvojne pomoči, so skladni s politiko Evropske unije, vključno s politiko o varovanju okolja, enakosti med spoloma, prometu, čezevropskih omrežjih (TREN), konkurenci in javnih naročilih. Skupni odbor sestavlja največ pet članov iz vsake pogodbenice.

Translation in English

The initiative for the execution of a specific programme or project shall be given either by the Joint Committee referred to in Article 6 hereof or by individual holders or contractors of a specific development cooperation activity or project. Projects and activities financed under the development assistance programme shall comply with European Union policies, including those concerning environmental protection, gender equality, transport, Trans European Networks (TREN), competition, as well as public procurement. The Joint Committee shall be composed of no more than five members from each Party.

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

English	Polish
aircraft commander	dowódca statku powietrznego
aircraft configuration	konfiguracja statku powietrznego
aircraft flight manual	instrukcja użytkowania w locie statku powietrznego
aircraft hangar	hangar dla statków powietrznych
airframe	płatowiec
airworthiness certificate	świadcstwo zdatości do lotu
automatic flight control system	automatyczny układ sterowania lotem
European Technical Standard Order	Europejska Norma Techniczna
...	



- 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 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.

To jest polskie tłumaczenie poprzedniego akapitu. To jest polskie tłumaczenie poprzedniego akapitu. To jest polskie tłumaczenie poprzedniego akapitu. To jest polskie tłumaczenie poprzedniego akapitu. To jest polskie tłumaczenie poprzedniego akapitu. To jest polskie tłumaczenie poprzedniego akapitu. To jest polskie tłumaczenie poprzedniego akapitu.

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.

To jest polskie tłumaczenie drugiego akapitu. To jest polskie tłumaczenie drugiego akapitu. To jest polskie tłumaczenie drugiego akapitu. To jest polskie tłumaczenie drugiego akapitu. To jest polskie tłumaczenie drugiego akapitu.



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. ¶

¶

To jest polskie tłumaczenie poprzedniego akapitu. To jest polskie tłumaczenie poprzedniego akapitu. To jest polskie tłumaczenie poprzedniego akapitu. To jest polskie tłumaczenie poprzedniego akapitu. To jest polskie tłumaczenie poprzedniego akapitu. To jest polskie tłumaczenie poprzedniego akapitu. To jest polskie tłumaczenie poprzedniego akapitu. ¶

¶

To jest polskie tłumaczenie drugiego akapitu. To jest polskie tłumaczenie drugiego akapitu. To jest polskie tłumaczenie drugiego akapitu. To jest polskie tłumaczenie drugiego akapitu. ¶



English	Polskie
<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.</p>	<p>To jest polskie tłumaczenie poprzedniego akapitu. To jest polskie tłumaczenie poprzedniego akapitu. To jest polskie tłumaczenie poprzedniego akapitu. To jest polskie tłumaczenie poprzedniego akapitu. To jest polskie tłumaczenie poprzedniego akapitu. To jest polskie tłumaczenie poprzedniego akapitu. To jest polskie tłumaczenie poprzedniego akapitu.</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.</p>	<p>To jest polskie tłumaczenie drugiego akapitu. To jest polskie tłumaczenie drugiego akapitu. To jest polskie tłumaczenie drugiego akapitu. To jest polskie tłumaczenie drugiego akapitu.</p>



Do's

- Όνομα
- filename01_EN.txt
 - filename01_PL.txt
 - filename02_EN.txt
 - filename02_PL.txt
 - filename03_EN.txt
 - filename03_PL.txt
 - filename04_EN.txt
 - filename04_PL.txt
 - filename05_EN.txt
 - filename05_PL.txt
 - filename06_EN.txt
 - filename06_PL.txt
 - filename07_EN.txt
 - filename07_PL.txt
 - filename08_EN.txt
 - filename08_PL.txt
 - filename09_EN.txt
 - filename09_PL.txt
 - filename10_EN.txt
 - filename10_PL.txt

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



Do's

- filename01_EN.txt
- filename01_PL.txt
- filename02_EN.txt
- filename02_PL.txt
- filename03_EN.txt
- filename03_PL.txt
- filename04_EN.txt
- filename04_PL.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
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



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



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)

- Crudean
- 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

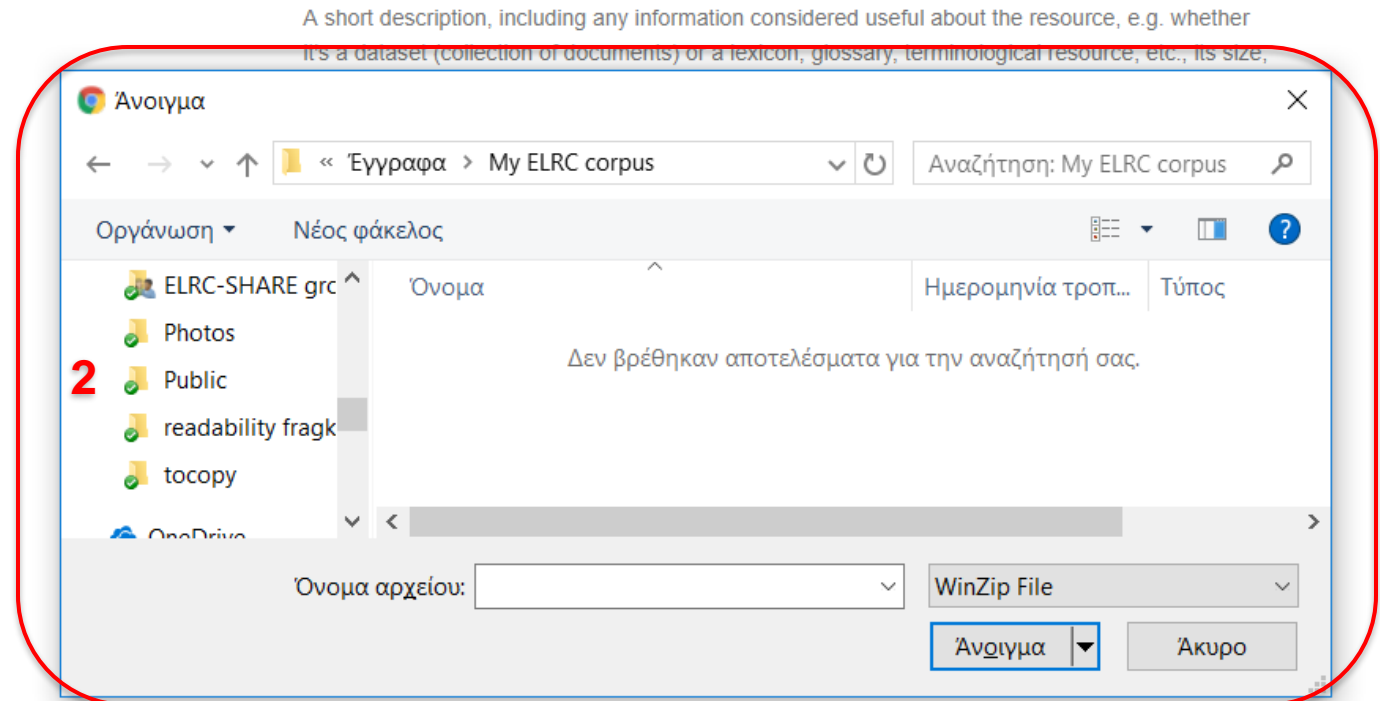
Please upload a **.zip file** up to 100MB.

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

- 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

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*

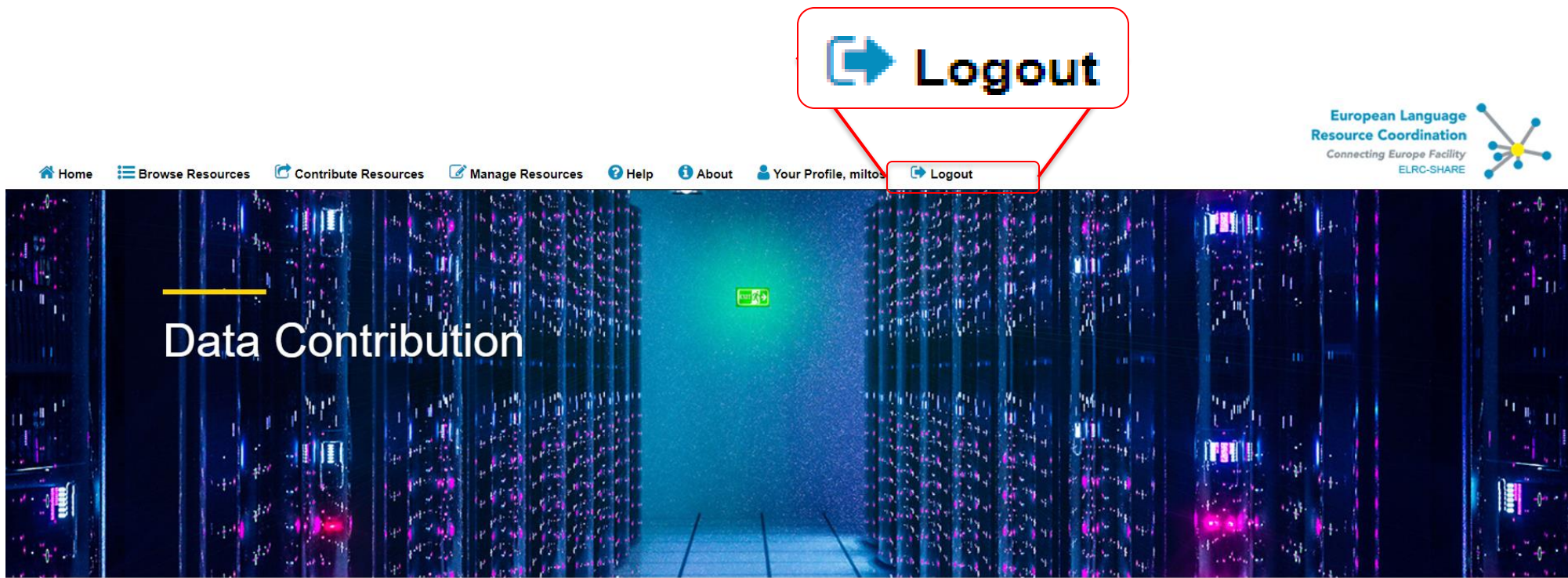
www

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 that contains a blue arrow icon pointing right 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 LRs:

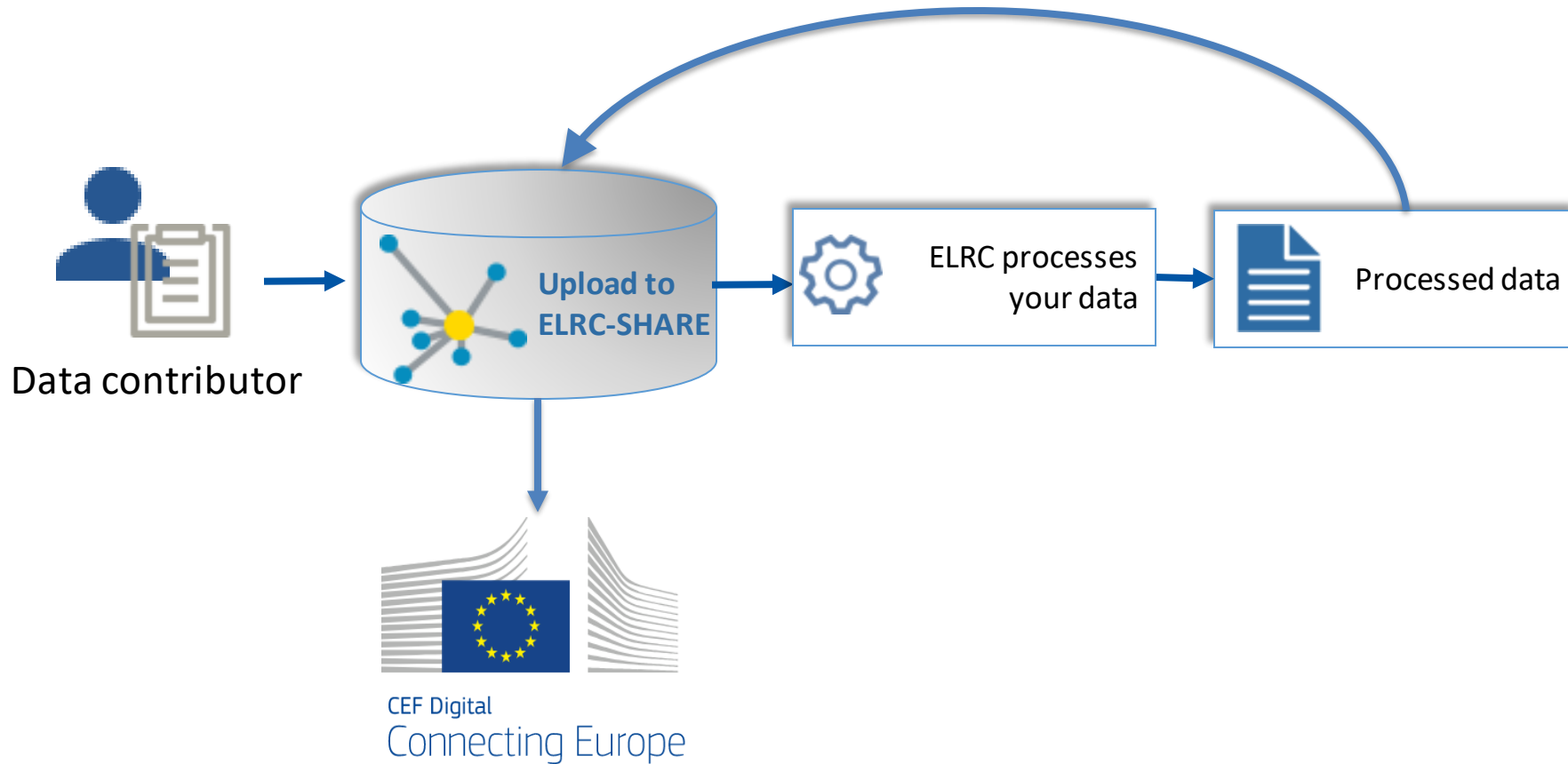
- [Walkthrough for contributors](#)
- [Walkthrough for editors](#)

ELRC-SHARE schema

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

What happens next?

What happens to your data?

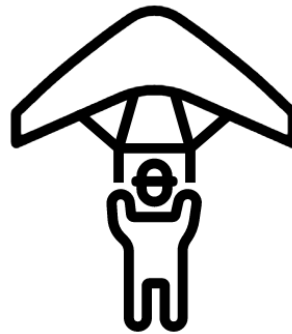


- All datasets are processed to result in tmx/tbx/txt files
- Data will indicatively undergo the following processing:
 - cleaning
 - format conversion
 - sentence alignment
 - metadata completion



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**

**Assistance will be provided in close cooperation
with a broad network of language experts**



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!



Data extraction

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



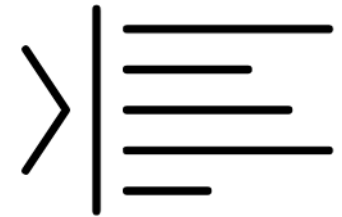
Anonymisation

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



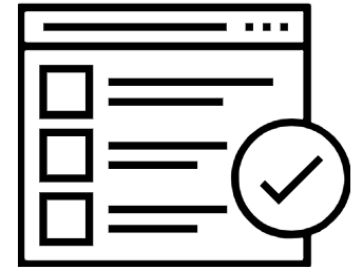
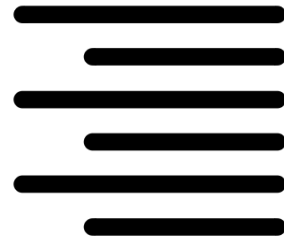
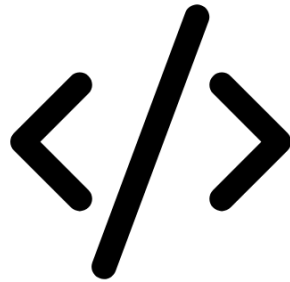
Cleaning

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!

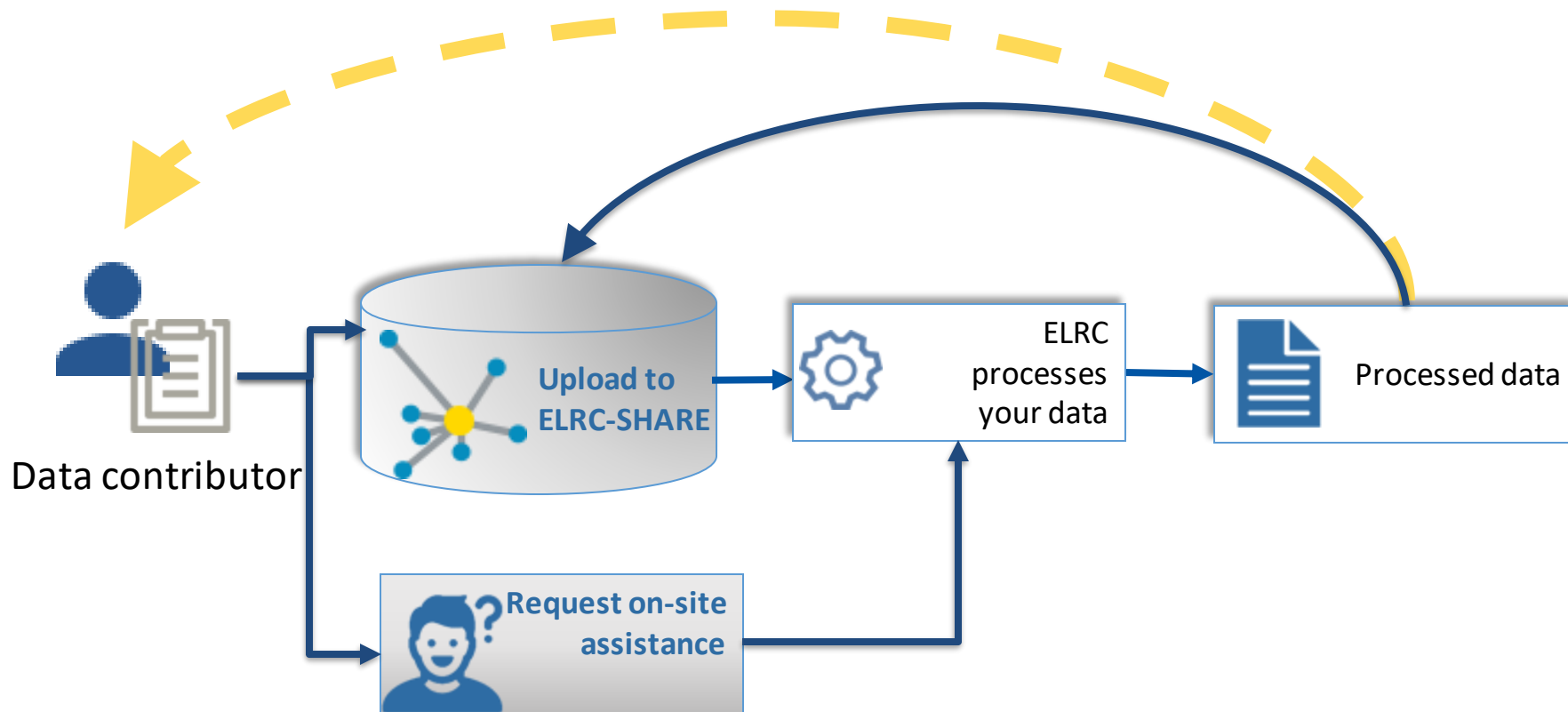
Alignment

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 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



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	help@lr-cooridantion.eu

lr-coordination.eu/helpdesk



Thank you for your attention!





- 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

Case studies (2015-2016)



Problem: Data provider didn't store translations as related documents, therefore source/target translation weren't paired

Solution: ELRC helped crawl a local system to find, related, and pair source/target translations





Problem: In some Spanish governmental departments, archives were only available in PDF

Solution: ELRC helped provide good converters to get usable documents



Problem: Data owner needed help with anonymization, as databases contained personal info. Another need: cleaning up 'junk' data (URLs, numbers, fragments)

Solution: ELRC helped provide anonymization services and data cleaning



Problem: Data donor found that legal acts in EN, ET, RU couldn't be aligned on a document level (no common machine-readable cross-language ID)

Solution: ELRC helped provide alignment services for documents

