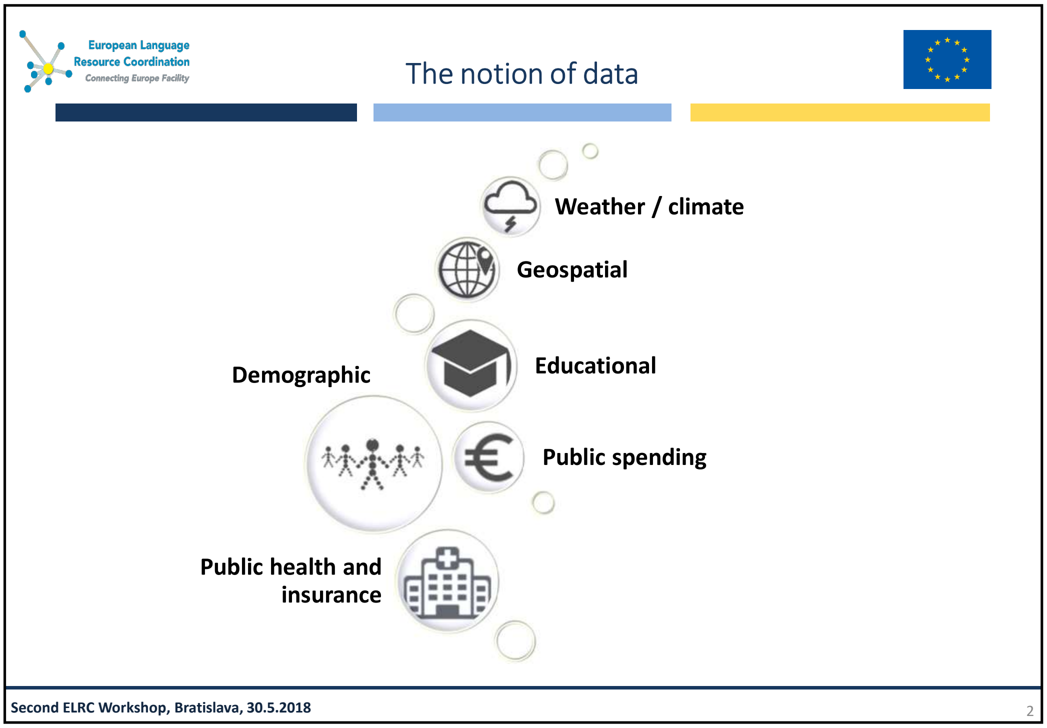
Preparing and sharing data with the ELRC-SHARE repository
and what happens next

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



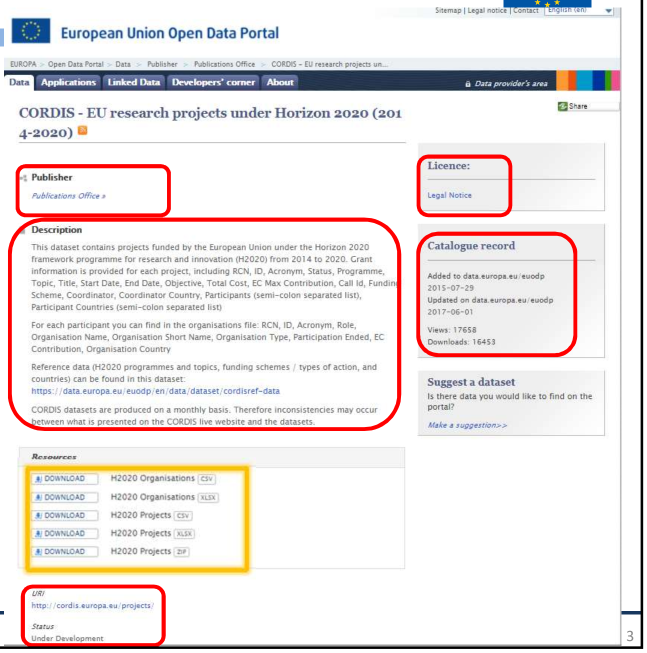
The notion of data

- Weather / climate
- Geospatial
- Educational
- Demographic
- Public spending
- Public health and insurance

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)



Sitemap | Legal notice | Contact | English (en)

European Union Open Data Portal

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

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

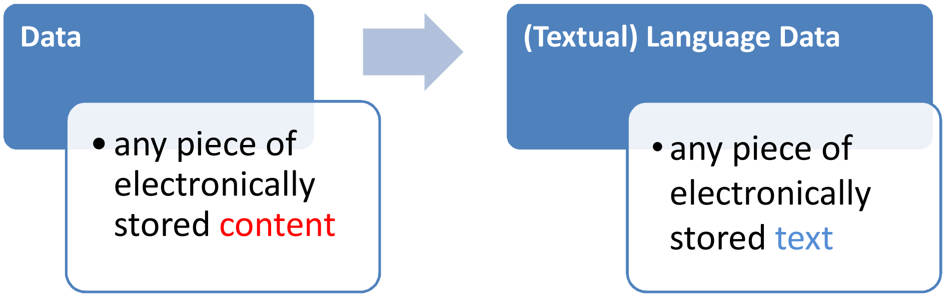
Views: 17658
Downloads: 16453

Suggest a dataset

Is there data you would like to find on the portal?

Make a suggestion>>

The notion of language data



Data

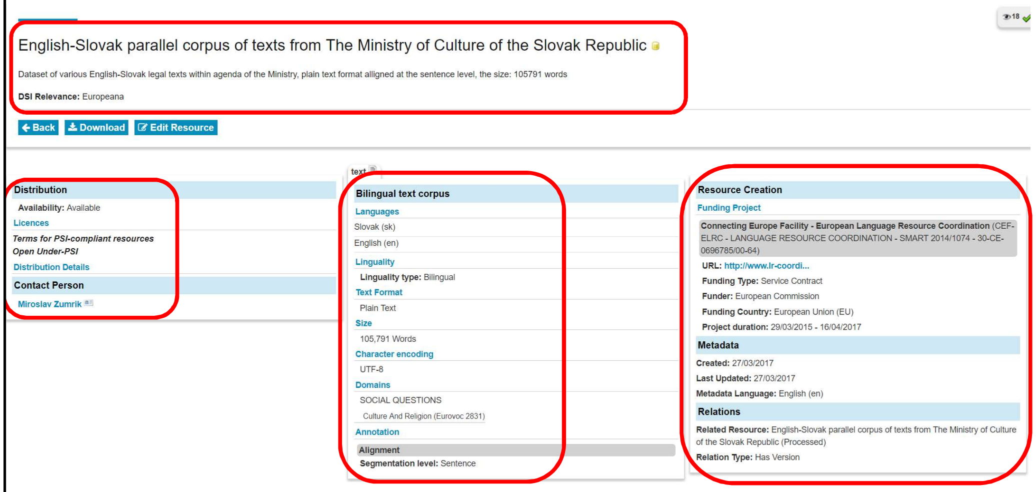
- any piece of electronically stored content

(Textual) Language Data

- any piece of electronically stored text

European Language Resource Coordination
Connecting Europe Facility

The notion of data in the context of eTranslation



English-Slovak parallel corpus of texts from The Ministry of Culture of the Slovak Republic

Dataset of various English-Slovak legal texts within agenda of the Ministry, plain text format alligned at the sentence level, the size: 105791 words

DSI Relevance: Europeana

Distribution

Availability: Available

Licences

Terms for PSI-compliant resources
Open Under-PSI

Distribution Details

Contact Person

Miroslav Zumrik

text

Bilingual text corpus

Languages

Slovak (sk)

English (en)

Linguality

Linguality type: Bilingual

Text Format

Plain Text

Size

105,791 Words

Character encoding

UTF-8

Domains

SOCIAL QUESTIONS

Culture And Religion (Eurovoc 2831)

Annotation

Alignment

Segmentation level: Sentence

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: 27/03/2017

Last Updated: 27/03/2017

Metadata Language: English (en)

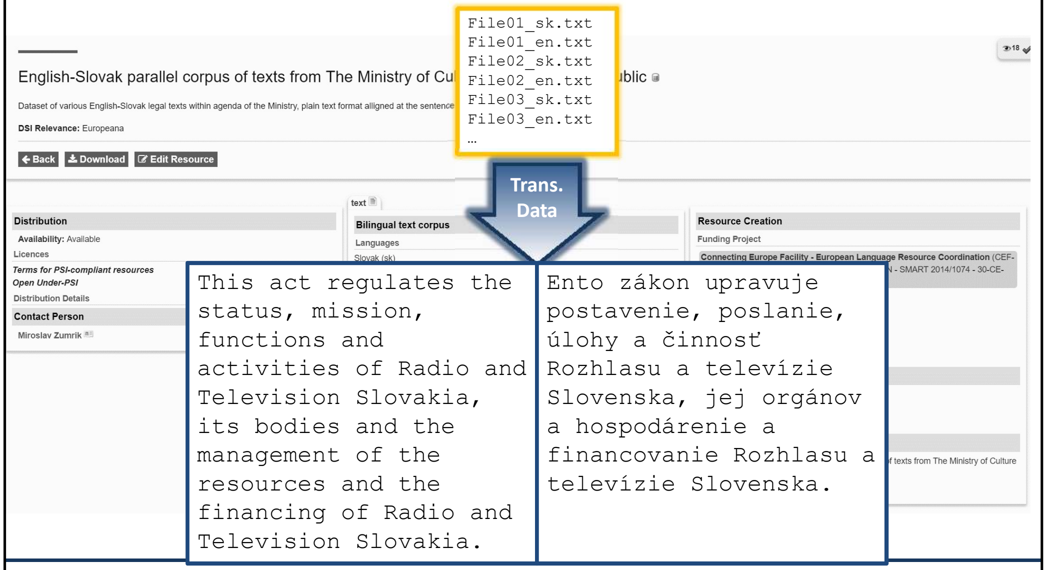
Relations

Related Resource: English-Slovak parallel corpus of texts from The Ministry of Culture of the Slovak Republic (Processed)

Relation Type: Has Version

European Language Resource Coordination
Connecting Europe Facility

The notion of data in the context of eTranslation



```
File01_sk.txt
File01_en.txt
File02_sk.txt
File02_en.txt
File03_sk.txt
File03_en.txt
...
```

Trans. Data

This act regulates the status, mission, functions and activities of Radio and Television Slovakia, its bodies and the management of the resources and the financing of Radio and Television Slovakia.	Ento zákon upravuje postavenie, poslanie, úlohy a činnosť Rozhlasu a televízie Slovenska, jej orgánov a hospodárenie a financovanie Rozhlasu a televízie Slovenska.

English-Slovak parallel corpus of texts from The Ministry of Cul ublic

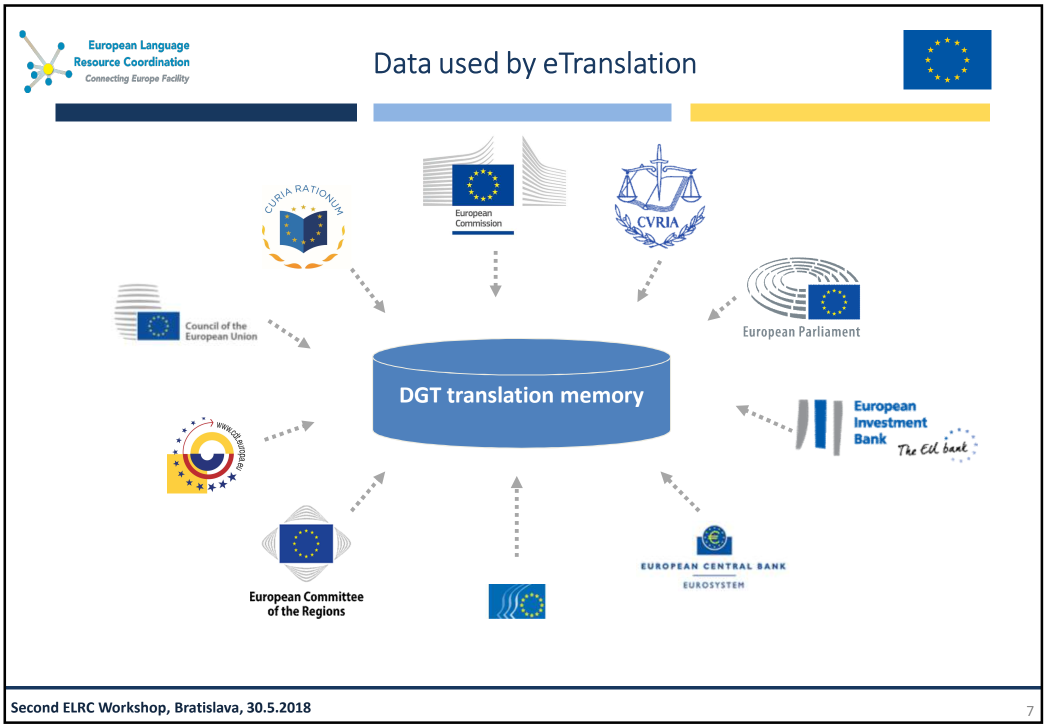
Dataset of various English-Slovak legal texts within agenda of the Ministry, plain text format aligned at the sentence

DSI Relevance: Europeana

Distribution
Availability: Available
Licences
Terms for PSI-compliant resources
Open Under-PSI
Distribution Details
Contact Person
Miroslav Zumrik

text
Bilingual text corpus
Languages
Slovak (sk)

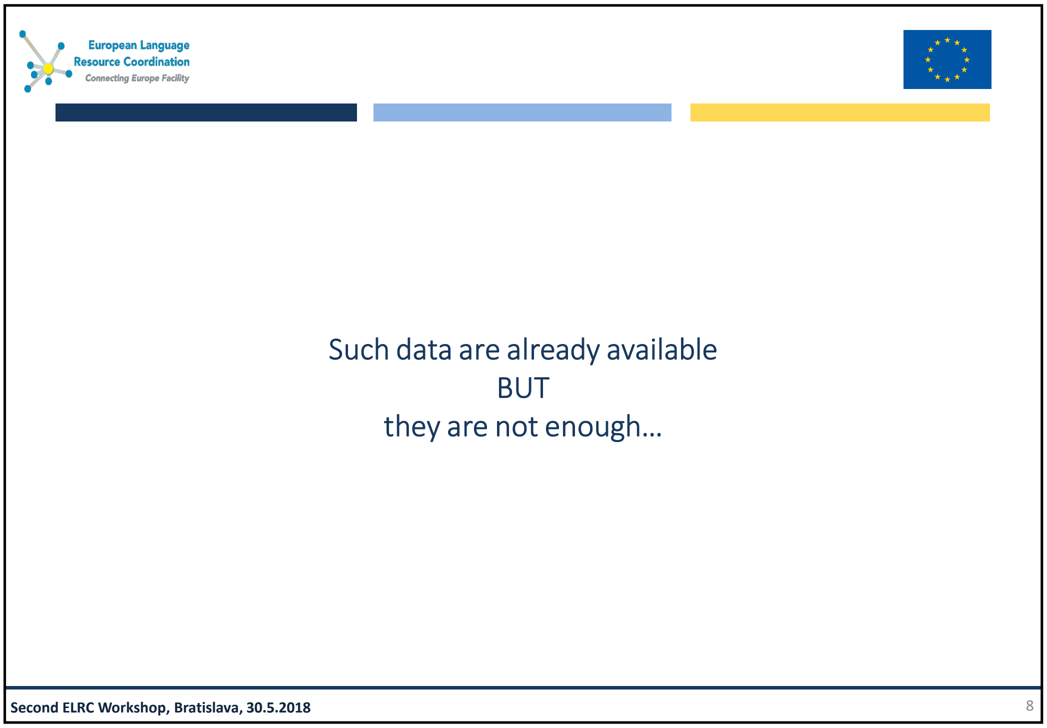
Resource Creation
Funding Project
Connecting Europe Facility - European Language Resource Coordination (CEF- N - SMART 2014/1074 - 30-CE-
f texts from The Ministry of Culture



European Language
Resource Coordination
Connecting Europe Facility

Data used by eTranslation

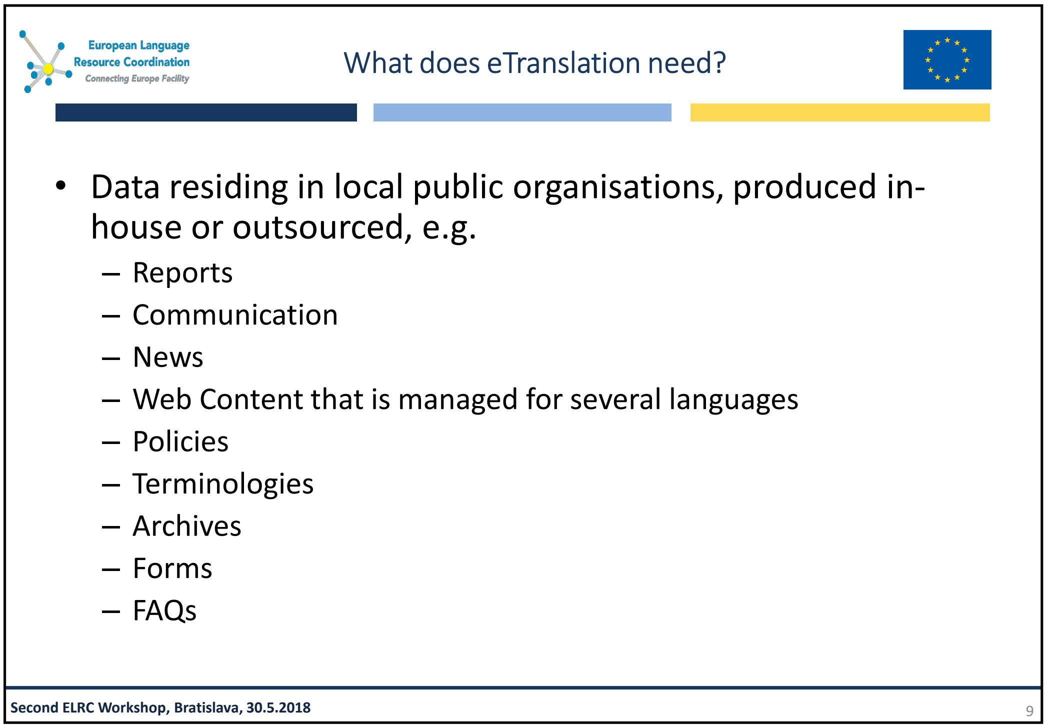
Second ELRC Workshop, Bratislava, 30.5.2018 7



European Language
Resource Coordination
Connecting Europe Facility

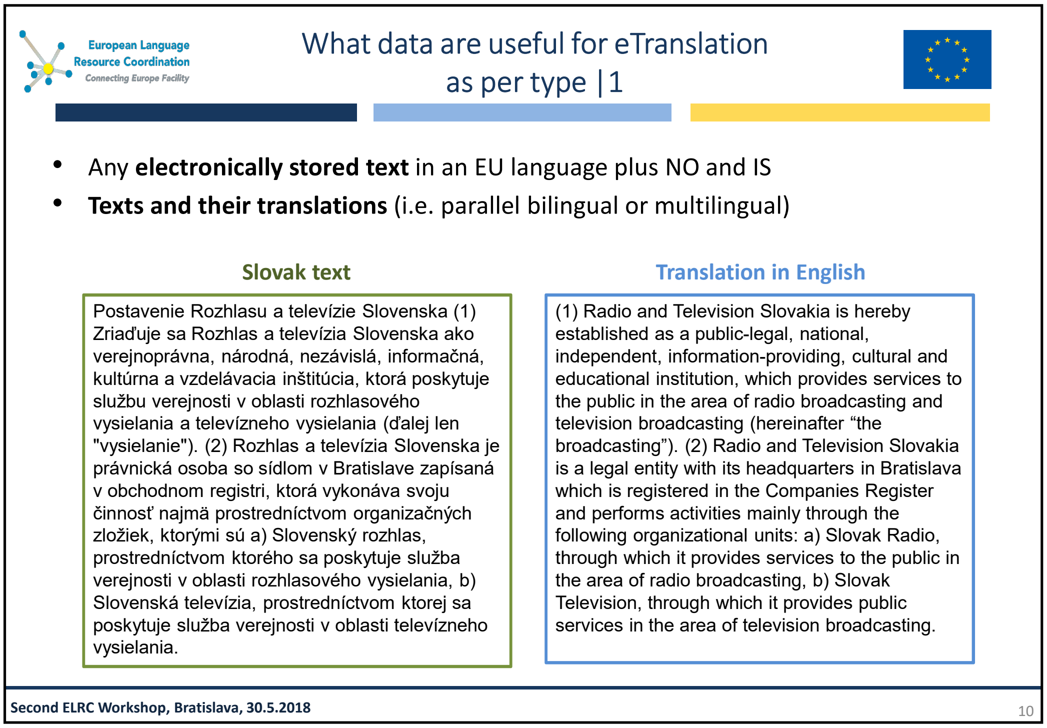
Such data are already available
BUT
they are not enough…

Second ELRC Workshop, Bratislava, 30.5.2018 8



What does eTranslation need?

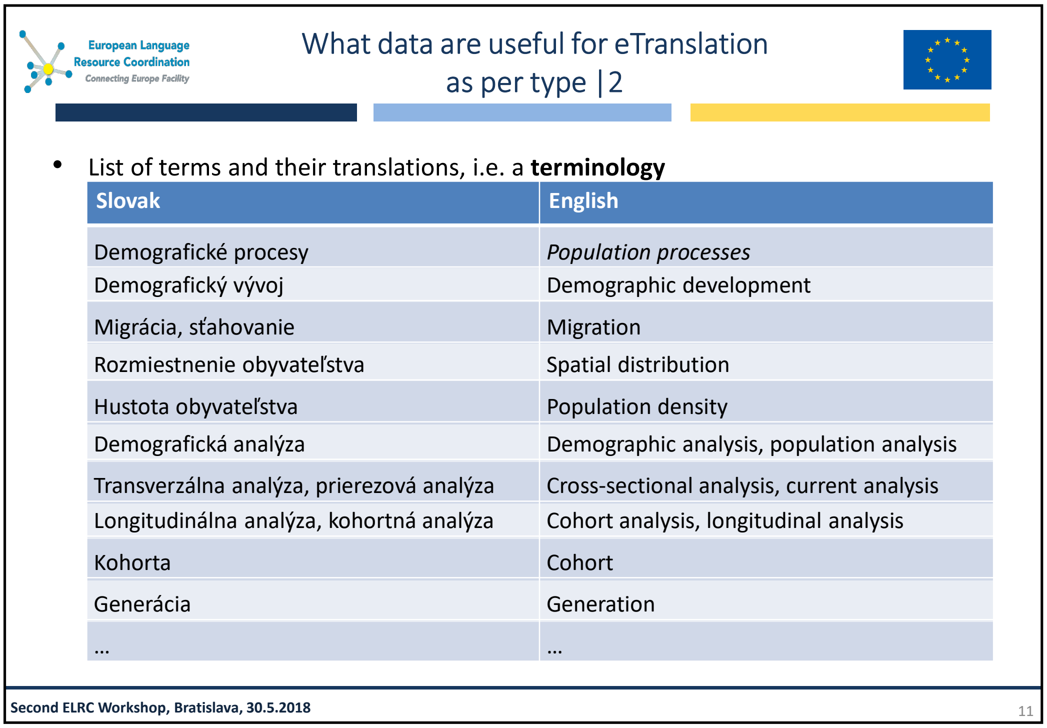
- Data residing in local public organisations, produced in-house or outsourced, e.g.
  - Reports
  - Communication
  - News
  - Web Content that is managed for several languages
  - Policies
  - Terminologies
  - Archives
  - Forms
  - FAQs



What data are useful for eTranslation as per type |1

- Any **electronically stored text** in an EU language plus NO and IS
- **Texts and their translations** (i.e. parallel bilingual or multilingual)

Slovak text	Translation in English
Postavenie Rozhlasu a televízie Slovenska (1) Zriaďuje sa Rozhlas a televízia Slovenska ako verejnoprávna, národná, nezávislá, informačná, kultúrna a vzdelávacia inštitúcia, ktorá poskytuje službu verejnosti v oblasti rozhlasového vysielania a televízneho vysielania (ďalej len "vysielanie"). (2) Rozhlas a televízia Slovenska je právnická osoba so sídlom v Bratislave zapísaná v obchodnom registri, ktorá vykonáva svoju činnosť najmä prostredníctvom organizačných zložiek, ktorými sú a) Slovenský rozhlas, prostredníctvom ktorého sa poskytuje služba verejnosti v oblasti rozhlasového vysielania, b) Slovenská televízia, prostredníctvom ktorej sa poskytuje služba verejnosti v oblasti televízneho vysielania.	(1) Radio and Television Slovakia is hereby established as a public-legal, national, independent, information-providing, cultural and educational institution, which provides services to the public in the area of radio broadcasting and television broadcasting (hereinafter “the broadcasting”). (2) Radio and Television Slovakia is a legal entity with its headquarters in Bratislava which is registered in the Companies Register and performs activities mainly through the following organizational units: a) Slovak Radio, through which it provides services to the public in the area of radio broadcasting, b) Slovak Television, through which it provides public services in the area of television broadcasting.



European Language Resource Coordination
Connecting Europe Facility

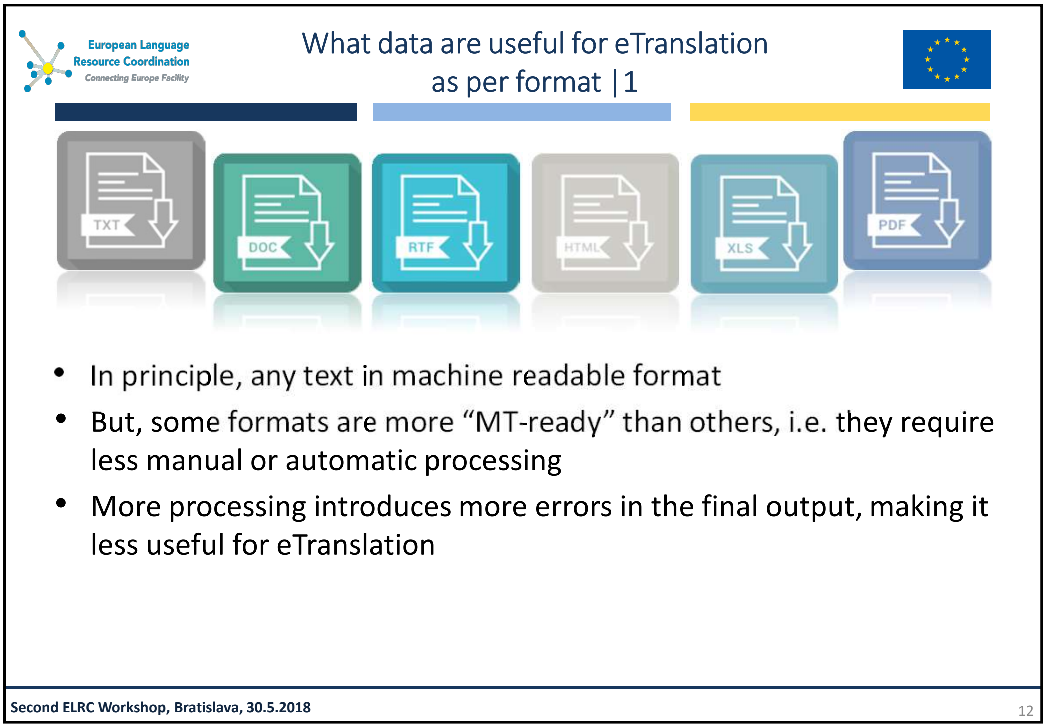
What data are useful for eTranslation as per type |2

- List of terms and their translations, i.e. a **terminology**

Slovak	English
Demografické procesy	*Population processes*
Demografický vývoj	Demographic development
Migrácia, sťahovanie	Migration
Rozmiestnenie obyvateľstva	Spatial distribution
Hustota obyvateľstva	Population density
Demografická analýza	Demographic analysis, population analysis
Transverzálna analýza, prierezová analýza	Cross-sectional analysis, current analysis
Longitudinálna analýza, kohortná analýza	Cohort analysis, longitudinal analysis
Kohorta	Cohort
Generácia	Generation
…	…

Second ELRC Workshop, Bratislava, 30.5.2018

11



European Language Resource Coordination
Connecting Europe Facility

What data are useful for eTranslation as per format |1

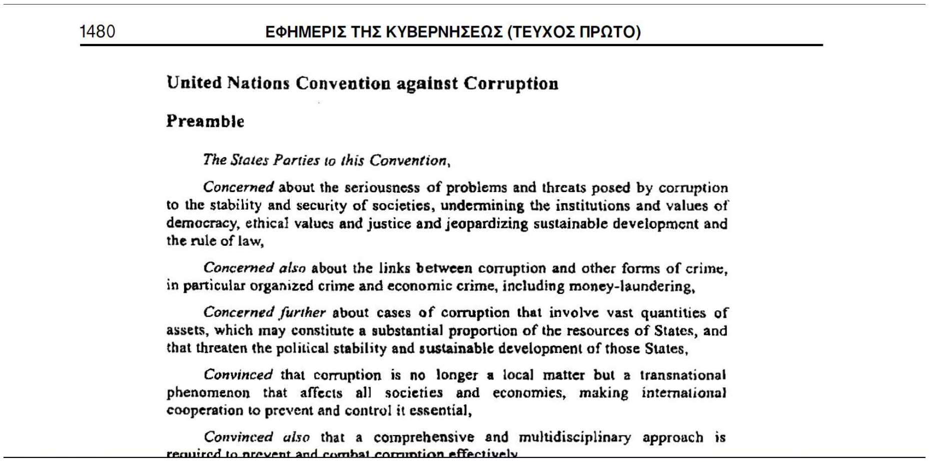
TXT
DOC
RTF
HTML
XLS
PDF

- 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

Second ELRC Workshop, Bratislava, 30.5.2018

12

File formats for parallel texts



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

What data are useful for eTranslation as per format |2

- The following formats are particularly useful (in descending order):
  - For bilingual/multilingual parallel texts
    1. Translation memories (.tmx)
    2. XML translation files (.xliff)
    3. Plain text (.txt, .csv)
    4. Spreadsheets (e.g. xlsx)
  - For terminologies
    1. TermBase eXchange (.tbx)
    2. Plain text (.txt, .csv)
    3. Spreadsheets (e.g. xlsx)
  - For monolingual texts
    1. Plain text (.txt, .csv)

File formats of parallel texts and their manipulation

Preparing your data |1

Don’ts

This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English.

Tento odsek je preložený v slovenčine. Tento odsek je preložený v slovenčine. Tento odsek je preložený v slovenčine. Tento odsek je preložený v slovenčine. Tento odsek je preložený v slovenčine. Tento odsek je preložený v slovenčine. Tento odsek je preložený v slovenčine.

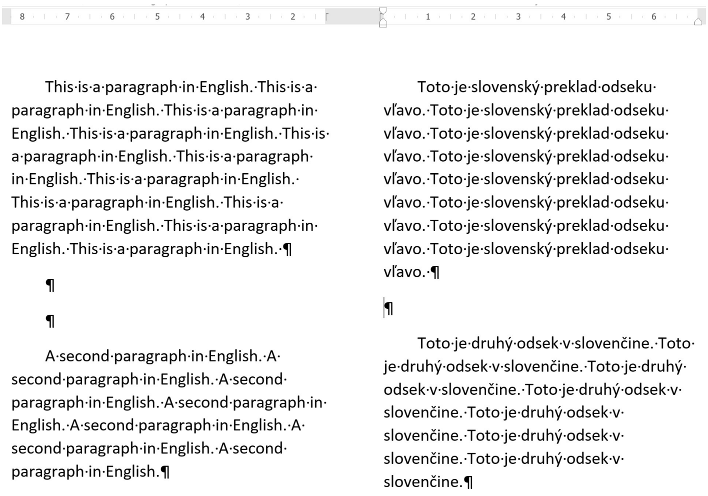
A second paragraph in English. A second paragraph in English. A second paragraph in English. A second paragraph in English. A second paragraph in English. A second paragraph in English. A second paragraph in English.

Toto je druhý odsek v slovenčine. Toto je druhý odsek v slovenčine. Toto je druhý odsek v slovenčine. Toto je druhý odsek v slovenčine. Toto je druhý odsek v slovenčine. Toto je druhý odsek v slovenčine. Toto je druhý odsek v slovenčine.

European Language Resource Coordination
Connecting Europe Facility

Preparing your data |2

Don'ts



This·is·a·paragraph·in·English.·This·is·a·paragraph·in·English.·This·is·a·paragraph·in·English.·This·is·a·paragraph·in·English.·This·is·a·paragraph·in·English.·This·is·a·paragraph·in·English.·This·is·a·paragraph·in·English.·This·is·a·paragraph·in·English.·This·is·a·paragraph·in·English.·This·is·a·paragraph·in·English.·This·is·a·paragraph·in·English.·¶

¶

¶

A·second·paragraph·in·English.·A·second·paragraph·in·English.·A·second·paragraph·in·English.·A·second·paragraph·in·English.·A·second·paragraph·in·English.·A·second·paragraph·in·English.·A·second·paragraph·in·English.¶

Toto·je·slovenský·preklad·odseku·vľavo.·Toto·je·slovenský·preklad·odseku·vľavo.·Toto·je·slovenský·preklad·odseku·vľavo.·Toto·je·slovenský·preklad·odseku·vľavo.·Toto·je·slovenský·preklad·odseku·vľavo.·Toto·je·slovenský·preklad·odseku·vľavo.·Toto·je·slovenský·preklad·odseku·vľavo.·Toto·je·slovenský·preklad·odseku·vľavo.·¶

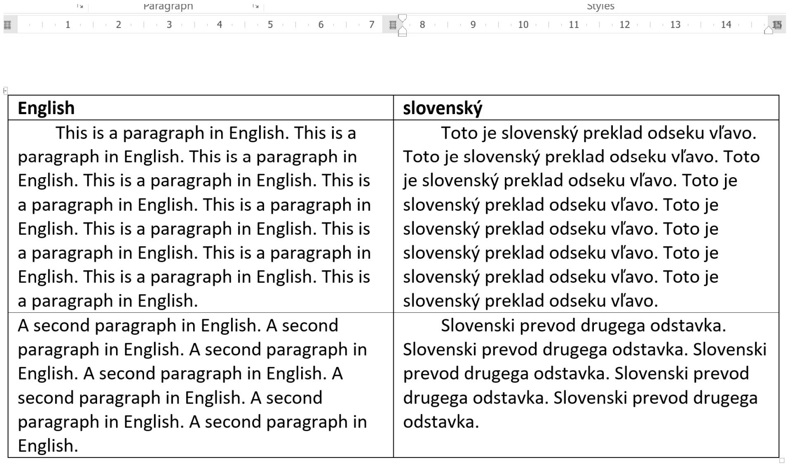
¶

Toto·je·druhý·odsek·v·slovenčine.·Toto·je·druhý·odsek·v·slovenčine.·Toto·je·druhý·odsek·v·slovenčine.·Toto·je·druhý·odsek·v·slovenčine.·Toto·je·druhý·odsek·v·slovenčine.·Toto·je·druhý·odsek·v·slovenčine.·Toto·je·druhý·odsek·v·slovenčine.¶

European Language Resource Coordination
Connecting Europe Facility

Preparing your data |3

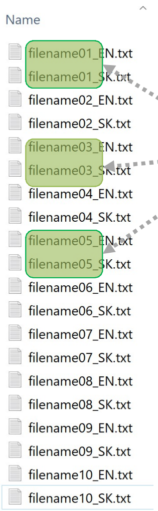
Don'ts



English	slovenský
This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English.	Toto je slovenský preklad odseku vľavo. Toto je slovenský preklad odseku vľavo. Toto je slovenský preklad odseku vľavo. Toto je slovenský preklad odseku vľavo. Toto je slovenský preklad odseku vľavo. Toto je slovenský preklad odseku vľavo. Toto je slovenský preklad odseku vľavo. Toto je slovenský preklad odseku vľavo.
A second paragraph in English. A second paragraph in English. A second paragraph in English. A second paragraph in English. A second paragraph in English. A second paragraph in English. A second paragraph in English.	Slovenski prevod drugega odstavka. Slovenski prevod drugega odstavka. Slovenski prevod drugega odstavka. Slovenski prevod drugega odstavka. Slovenski prevod drugega odstavka.

Preparing your data |4

Do's



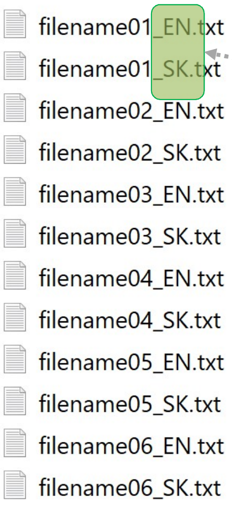
Name

filename01_EN.txt
filename01_SK.txt
filename02_EN.txt
filename02_SK.txt
filename03_EN.txt
filename03_SK.txt
filename04_EN.txt
filename04_SK.txt
filename05_EN.txt
filename05_SK.txt
filename06_EN.txt
filename06_SK.txt
filename07_EN.txt
filename07_SK.txt
filename08_EN.txt
filename08_SK.txt
filename09_EN.txt
filename09_SK.txt
filename10_EN.txt
filename10_SK.txt

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

Preparing your data |5

Do's

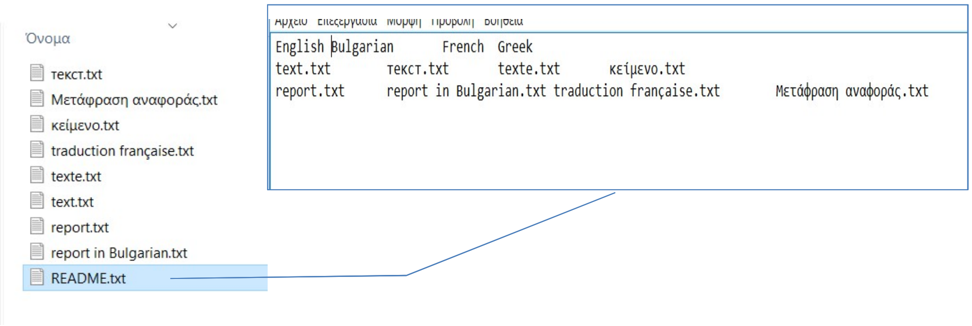


filename01_EN.txt
filename01_SK.txt
filename02_EN.txt
filename02_SK.txt
filename03_EN.txt
filename03_SK.txt
filename04_EN.txt
filename04_SK.txt
filename05_EN.txt
filename05_SK.txt
filename06_EN.txt
filename06_SK.txt

Include **language identifiers** in the filename

Preparing your data |5

Do's



- текст.txt
- Μετάφραση αναφοράς.txt
- κείμενο.txt
- traduction française.txt
- texte.txt
- text.txt
- report.txt
- report in Bulgarian.txt
- README.txt

```
English Bulgarian       French  Greek
text.txt        текст.txt       texte.txt       κείμενο.txt
report.txt      report in Bulgarian.txt traduction française.txt        Μετάφραση αναφοράς.txt
```

Criteria for grouping your data

- 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*

Preferred domains

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

CEF DSI	Domain
Online Dispute Resolution	Consumers’ rights, complaints
Electronic Exchange of Social Security Information	Social security, insurance
eProcurement	Public procurement, contractual agreements
European e-Justice Portal	Justice, Law
eHealth	Health, Medicine
Business Registers Interconnection System	Business, market
Safer Internet	
Cybersecurity	
Public Open Data	
Europeana	Culture

How to contribute your data to CEF eTranslation
A step-by-step guide

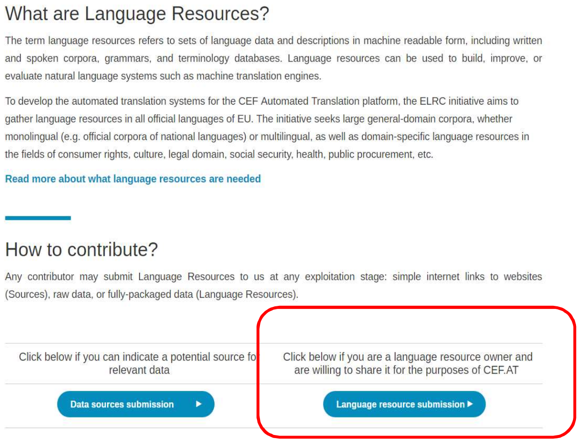
To ELRC-SHARE

- 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

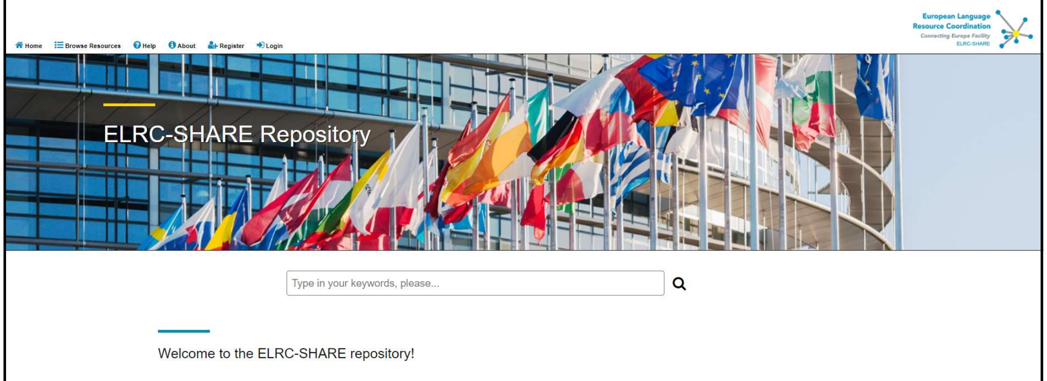
How to contribute?

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 fo relevant data

Click below if you are a language resource owner and are willing to share it for the purposes of CEF.AT

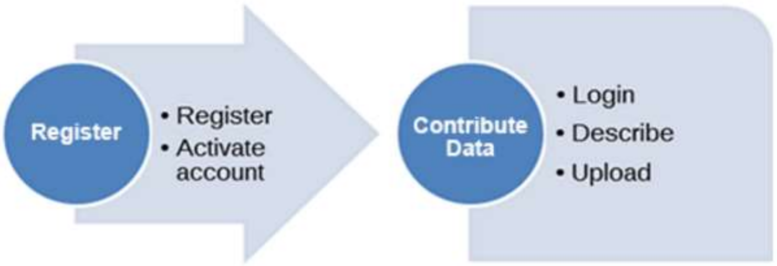
ELRC-SHARE repository



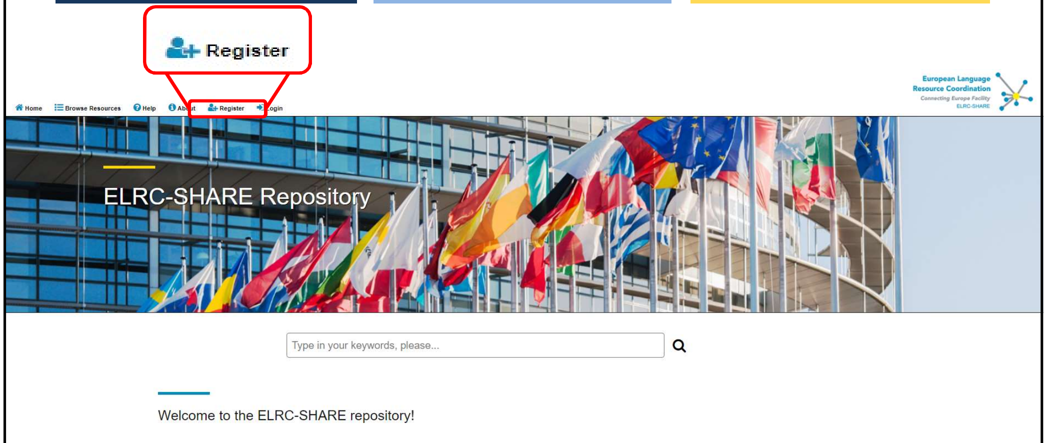
Home
Browse Resources
Help
About
Register
Login
European Language Resource Coordination
Connecting Europe Facility
ELRC-SHARE
ELRC-SHARE Repository
Type in your keywords, please...

Welcome to the ELRC-SHARE repository!

How to Contribute Data



How to Register (1/2)

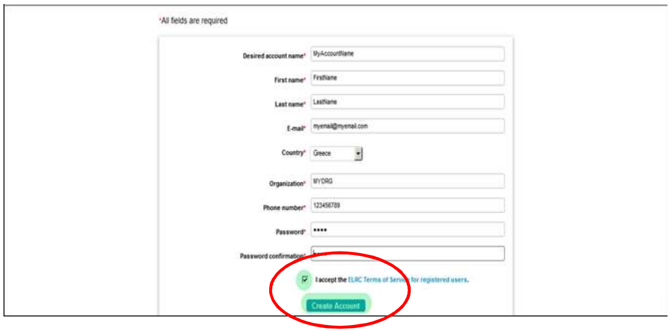


Welcome to the ELRC-SHARE repository!

European Language Resource Coordination
Connecting Europe Facility

How to Register (2/2)

- 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



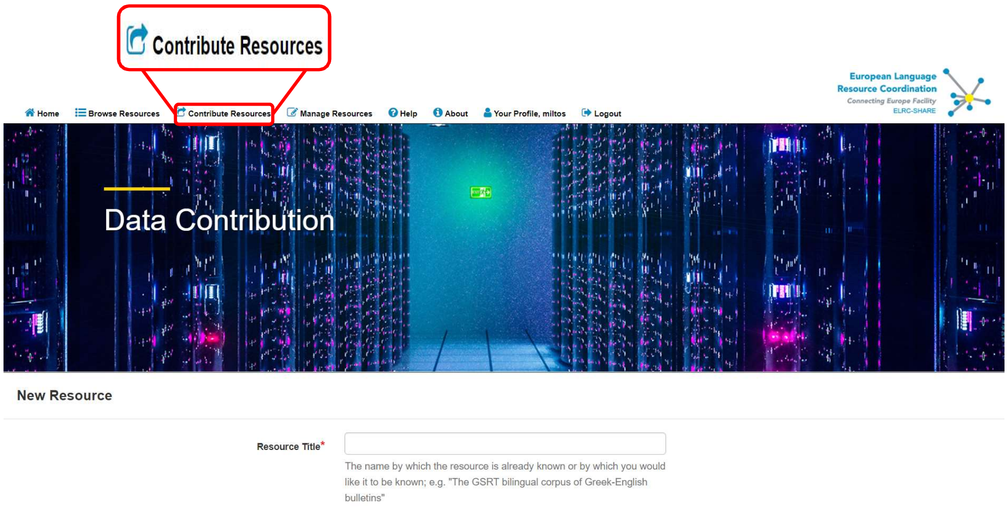
*All fields are required

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.

European Language Resource Coordination
Connecting Europe Facility

How to Contribute Data (1/6)



New Resource

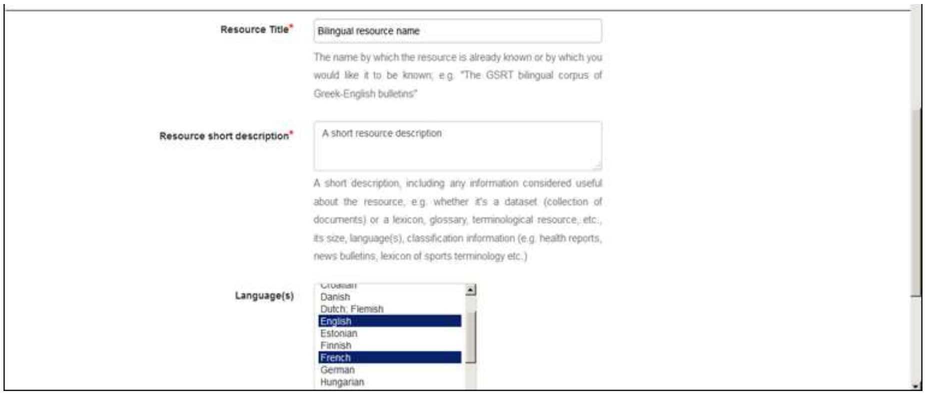
Resource Title*

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

How to Contribute Data (2/6)

- Fill in the details of the dataset



Resource Title*

Bilingual resource name

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 resource 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.)

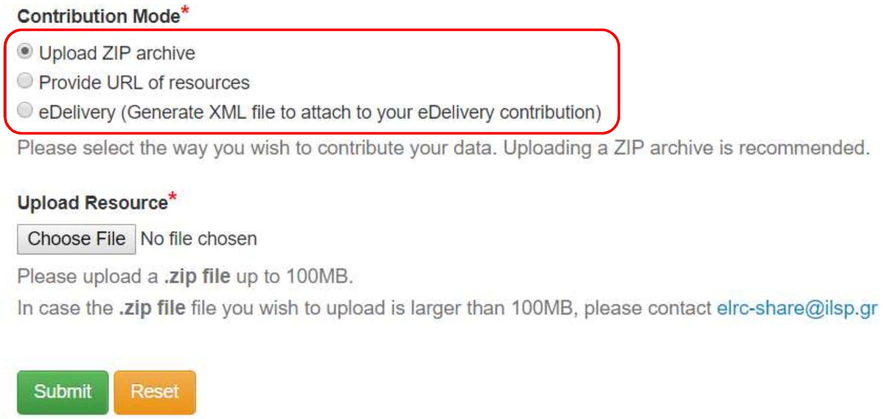
Language(s)

Croatian
Danish
Dutch; Flemish
English
Estonian
Finnish
French
German
Hungarian

European Language Resource Coordination
Connecting Europe Facility

How to Contribute Data (3/6)

- Three modes for contributing your data



Contribution Mode*

● Upload ZIP archive
○ Provide URL of resources
○ eDelivery (Generate XML file to attach to your eDelivery contribution)

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

Upload Resource*

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

European Language Resource Coordination
Connecting Europe Facility

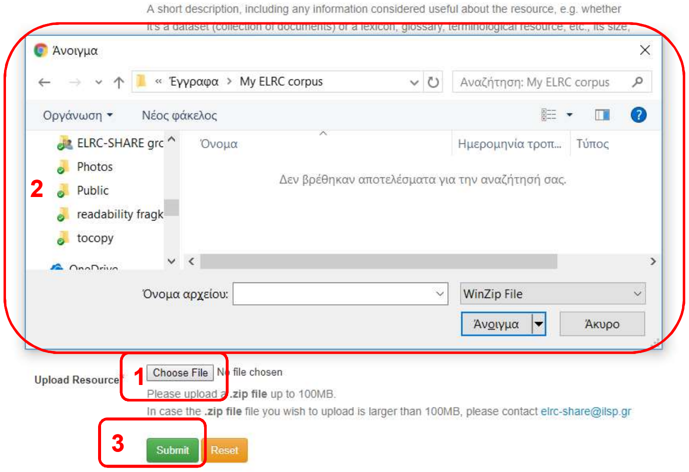
How to create a .zip file

- Free file compression tools (indicative):
  - 7zip
  - PeaZip
  - Hamster Free Zip Archiver
  - Universal Extractor
  - ZipItFree
- Windows embedded compression functionality

European Language Resource Coordination
Connecting Europe Facility

How to Contribute Data (4/6)

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



A short description, including any information considered useful about the resource, e.g. whether it's a dataset (collection of documents) or a lexicon, glossary, terminological resource, etc., its size,

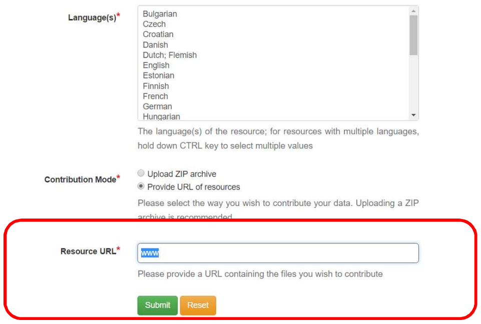
Upload Resource* Choose File No file chosen

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

How to Contribute Data (5/6)

- 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

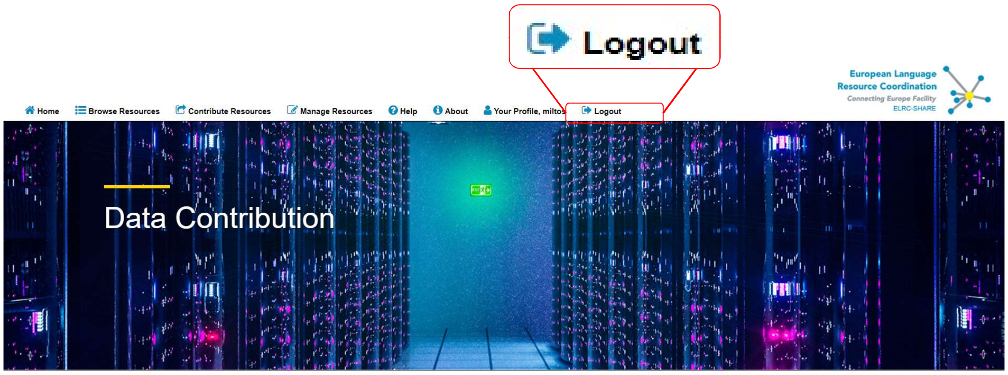
Please select the way you wish to contribute your data. Uploading a ZIP archive is recommended

Resource URL* www

Please provide a URL containing the files you wish to contribute

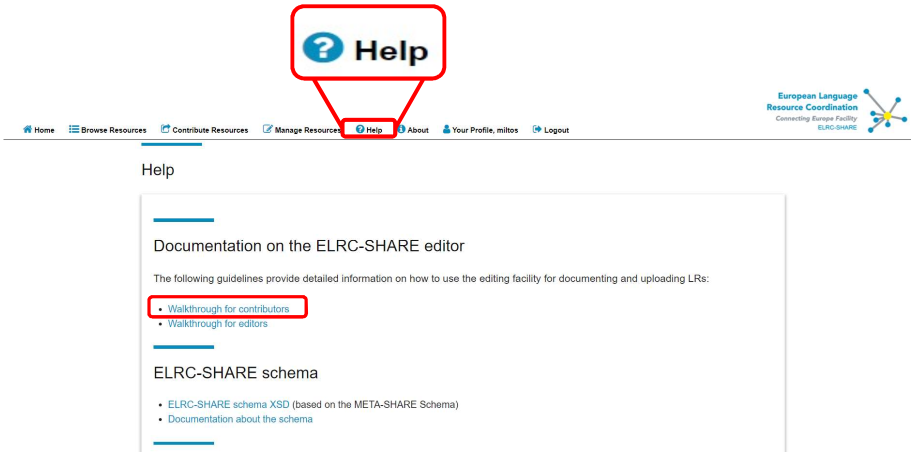
How to Contribute Data (6/6)

- Repeat the process if you want to contribute another resource, or log out



European Language Resource Coordination
Connecting Europe Facility

Guidelines for contributors



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

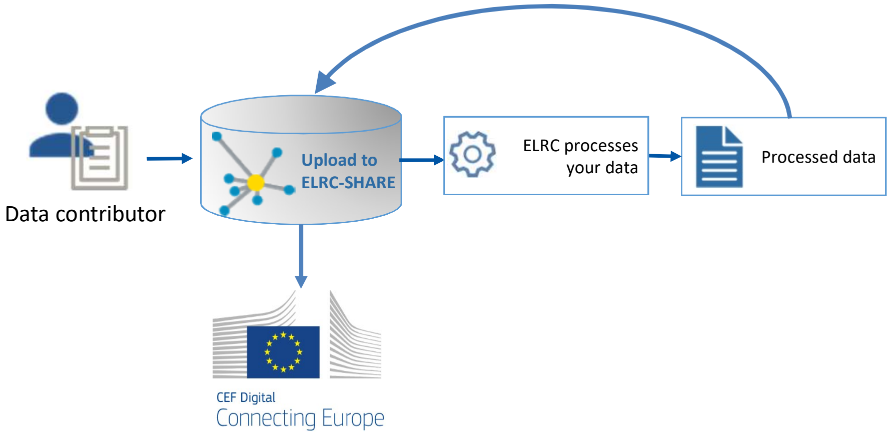
- ELRC-SHARE schema XSD (based on the META-SHARE Schema)
- Documentation about the schema

European Language Resource Coordination
Connecting Europe Facility

What happens next?

European Language Resource Coordination
Connecting Europe Facility

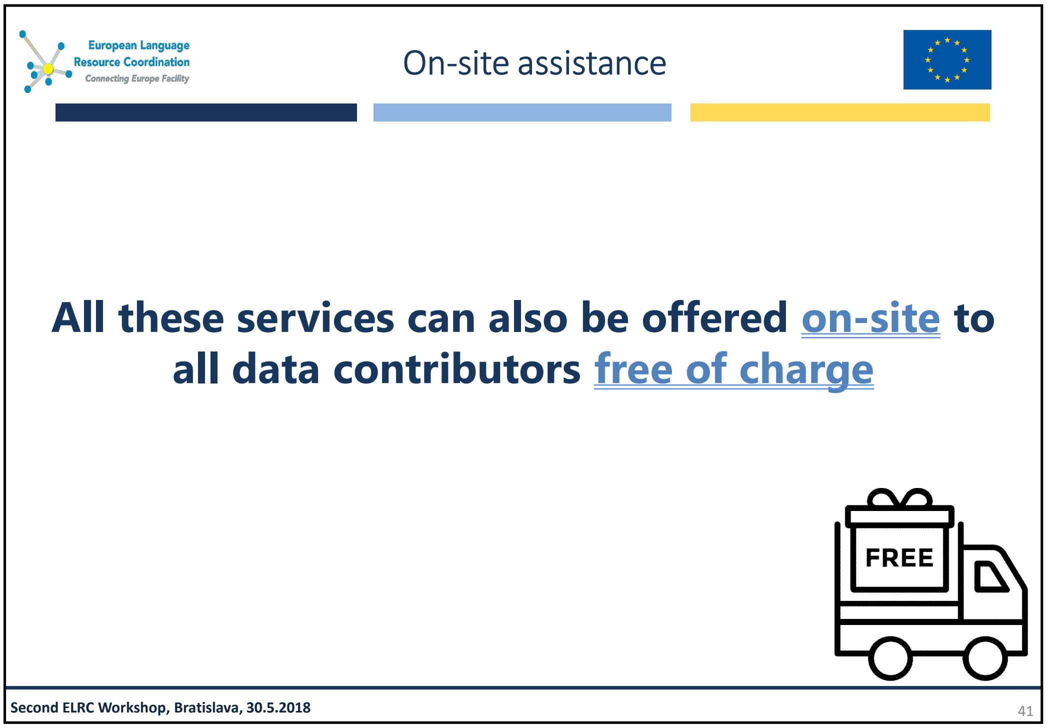
What happens to your data?



European Language Resource Coordination
Connecting Europe Facility

Data processing before delivery to EC

- 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



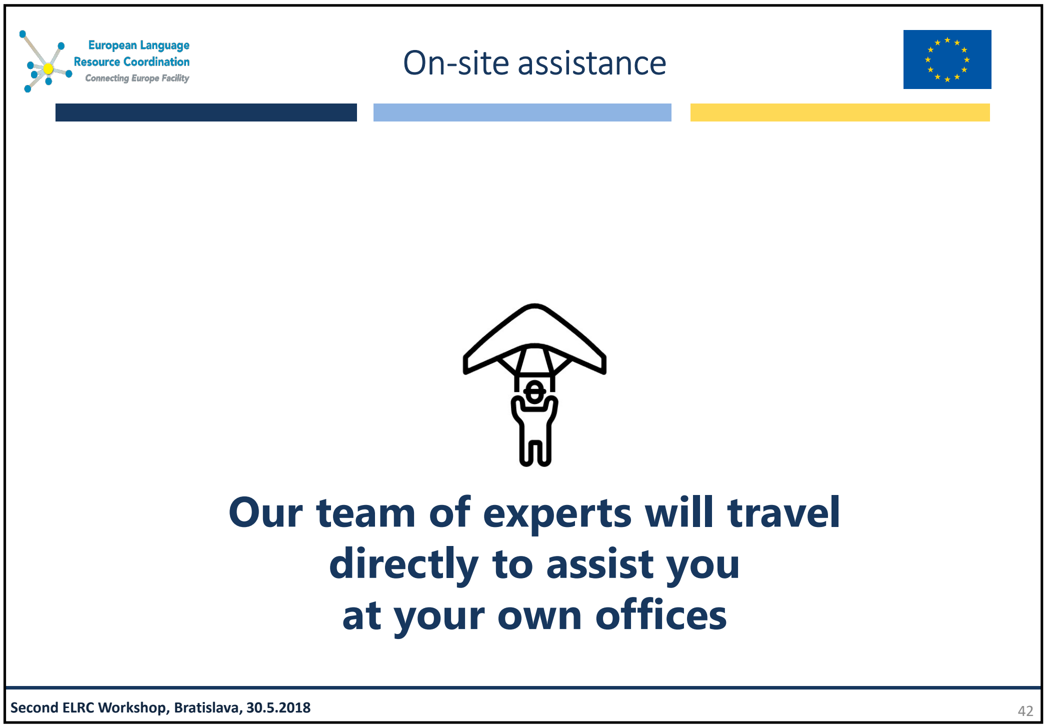
European Language Resource Coordination
Connecting Europe Facility

On-site assistance

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

Second ELRC Workshop, Bratislava, 30.5.2018

41



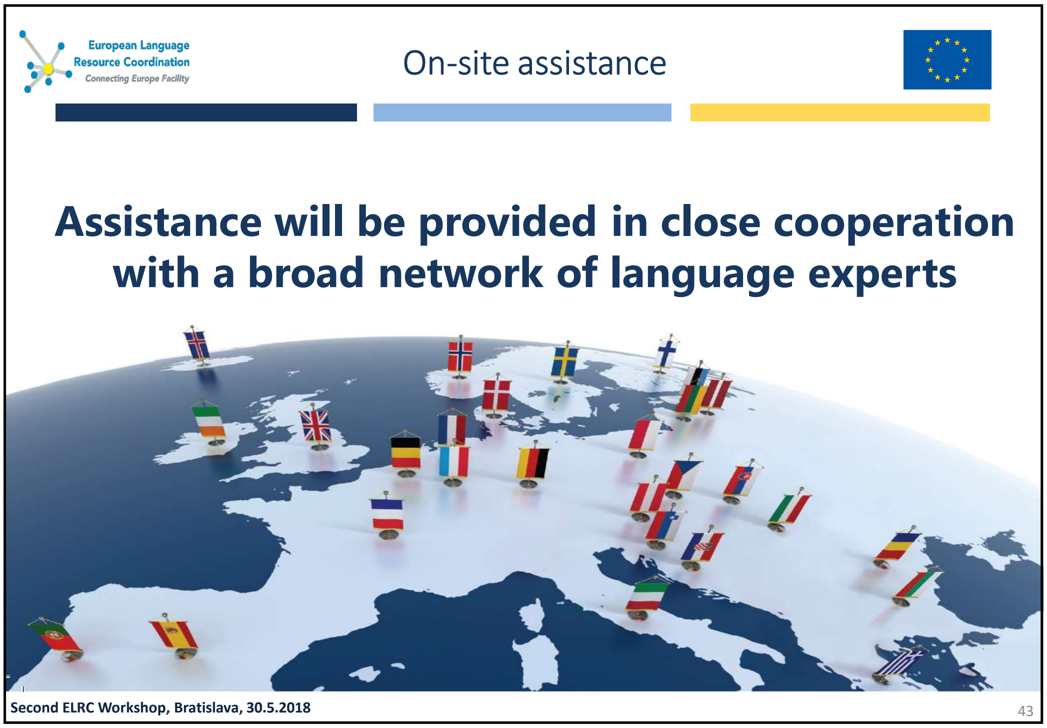
European Language Resource Coordination
Connecting Europe Facility

On-site assistance

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

Second ELRC Workshop, Bratislava, 30.5.2018

42



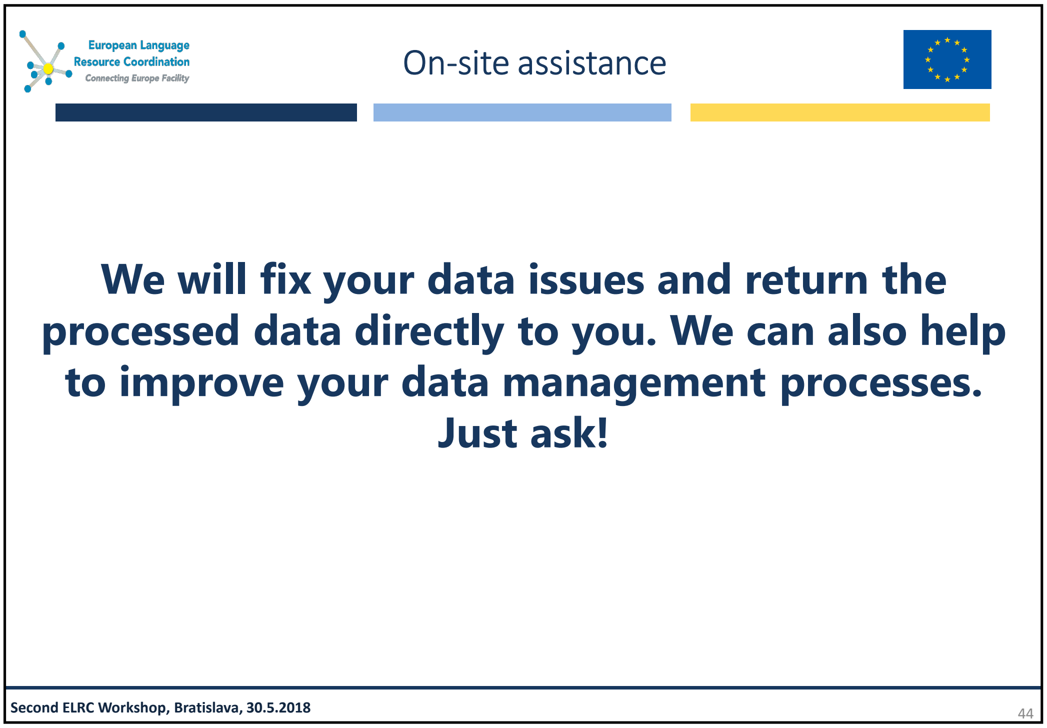
European Language Resource Coordination
Connecting Europe Facility

On-site assistance

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

Second ELRC Workshop, Bratislava, 30.5.2018

43



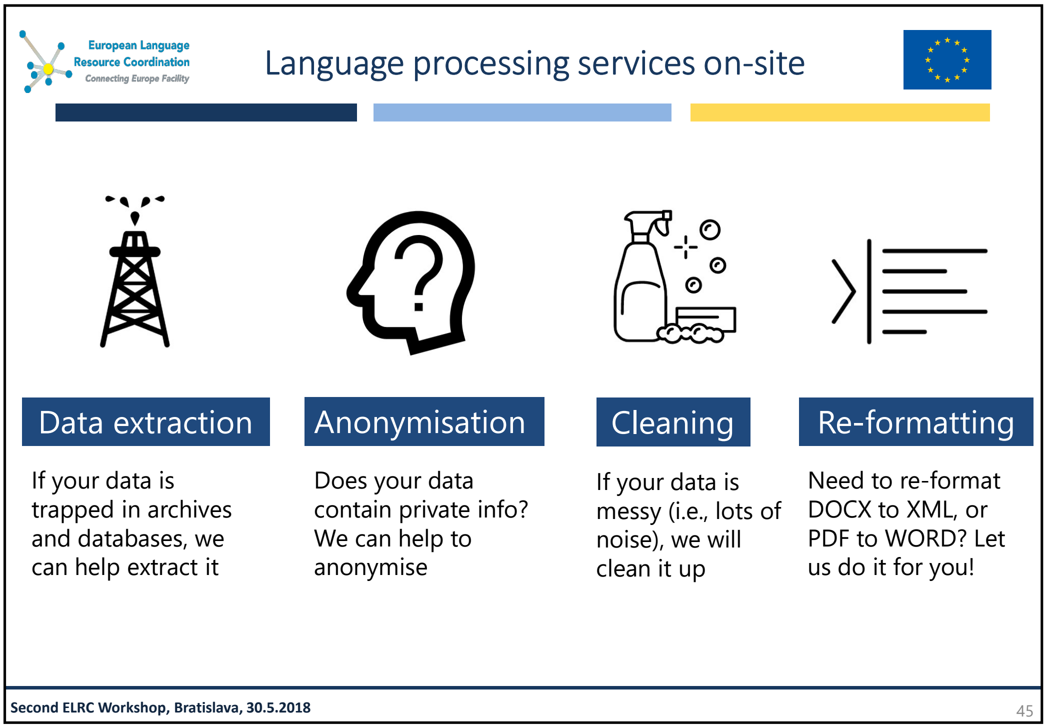
European Language Resource Coordination
Connecting Europe Facility

On-site assistance

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!

Second ELRC Workshop, Bratislava, 30.5.2018

44



European Language Resource Coordination
Connecting Europe Facility

Language processing services on-site

Data extraction

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

Anonymisation

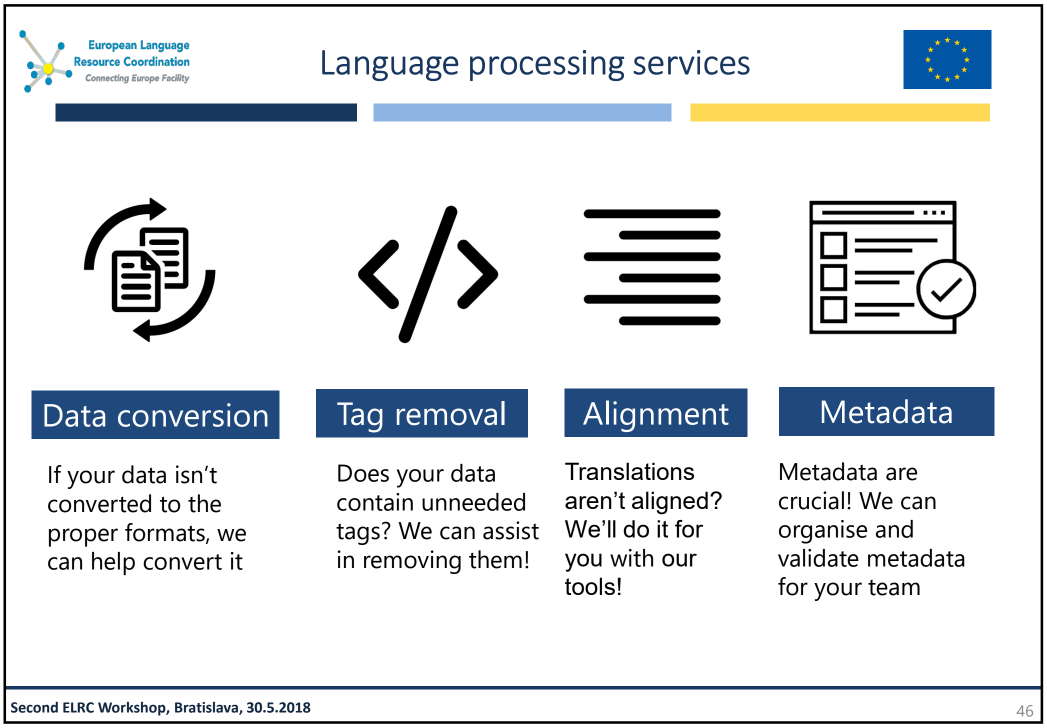
Does your data contain private info? We can help to anonymise

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!



European Language Resource Coordination
Connecting Europe Facility

Language processing services

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!

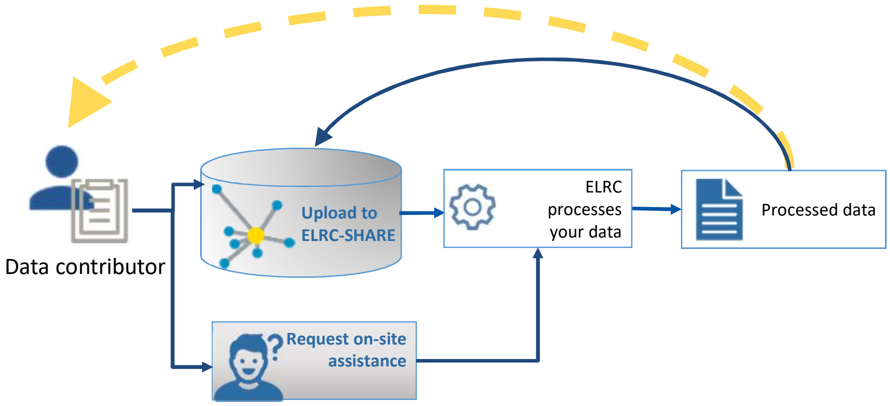
Alignment

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

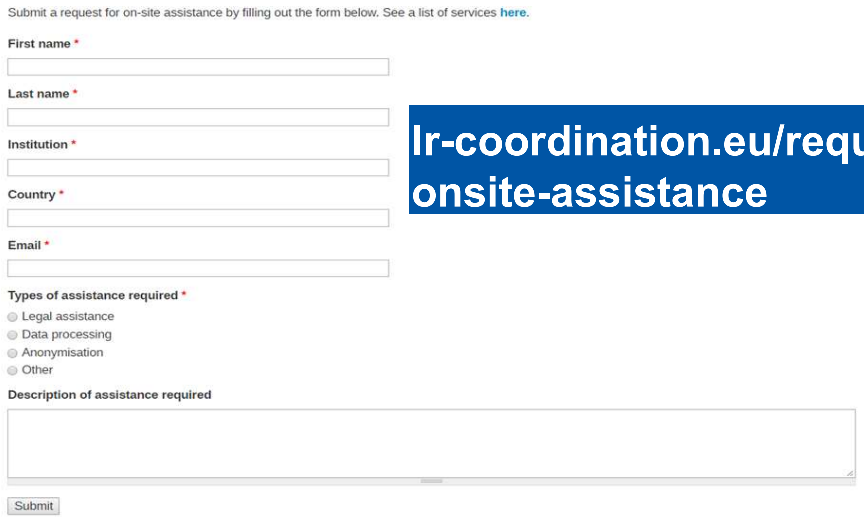
What happens to your data?



How to request services and help

European Language Resource Coordination
Connecting Europe Facility

ELRC onsite assistance



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 *

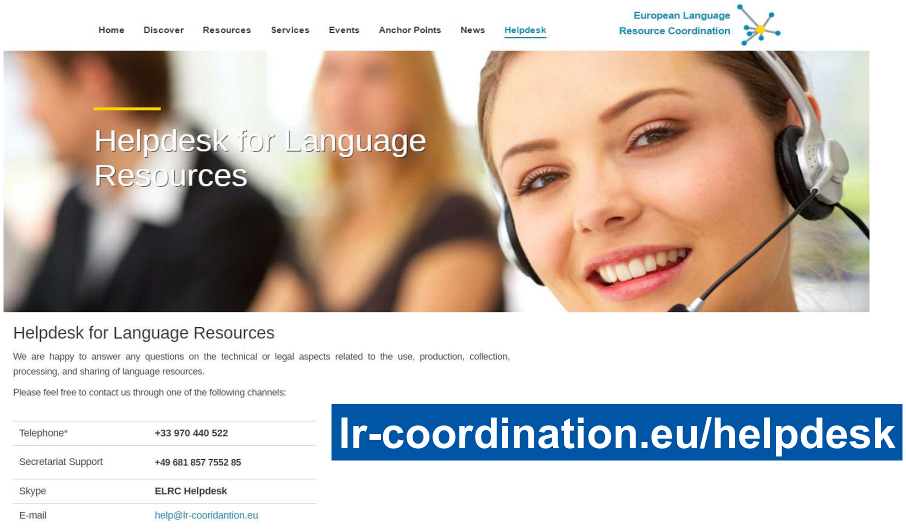
○ Legal assistance
○ Data processing
○ Anonymisation
○ Other

Description of assistance required

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

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:

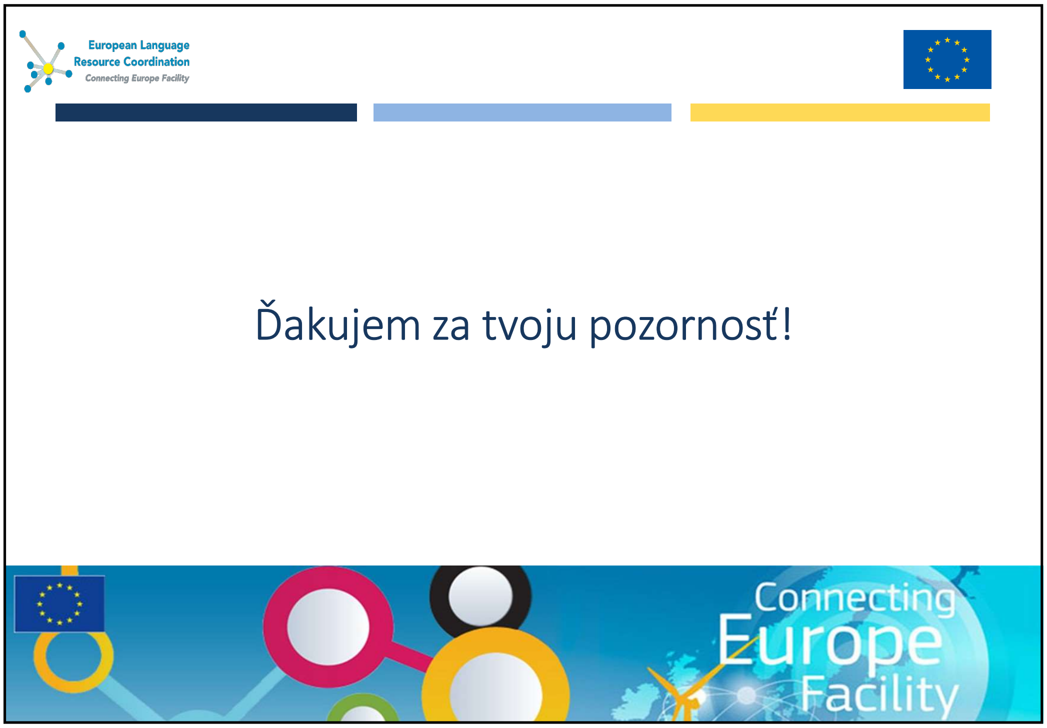
Telephone*	**+33 970 440 522**
Secretariat Support	**+49 681 857 7552 85**
Skype	**ELRC Helpdesk**
E-mail	help@lr-cooridantion.eu

lr-coordination.eu/helpdesk

ELRC consortium – come talk to us!

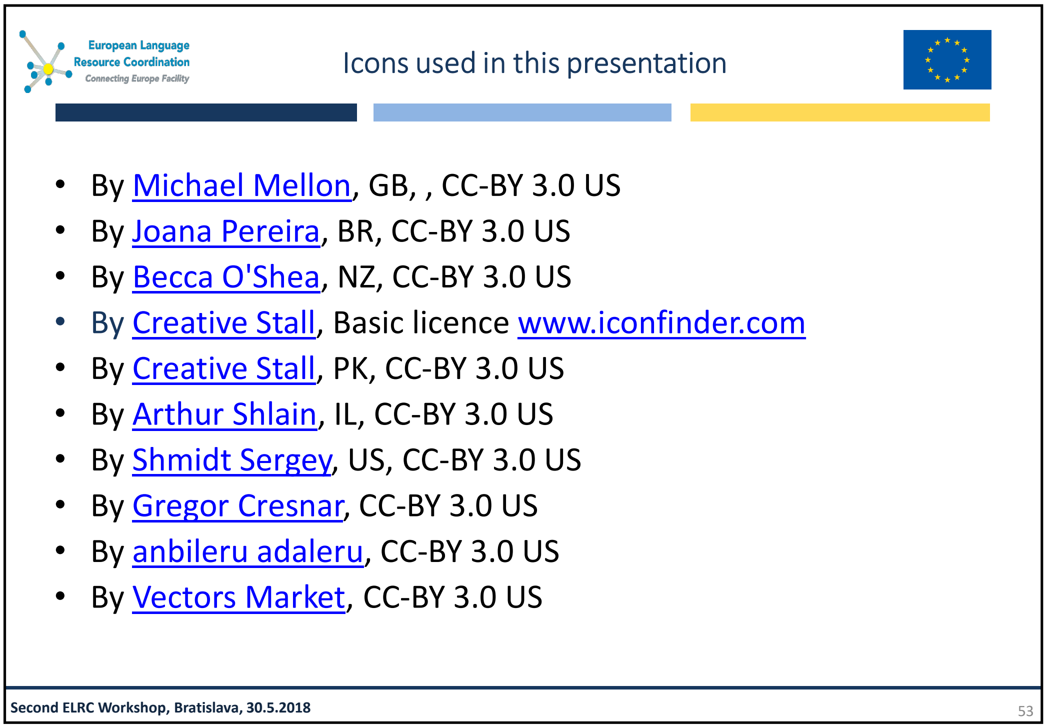
Second ELRC Workshop, Bratislava, 30.5.2018

51



Ďakujem za tvoju pozornosť!

Connecting
Europe
Facility



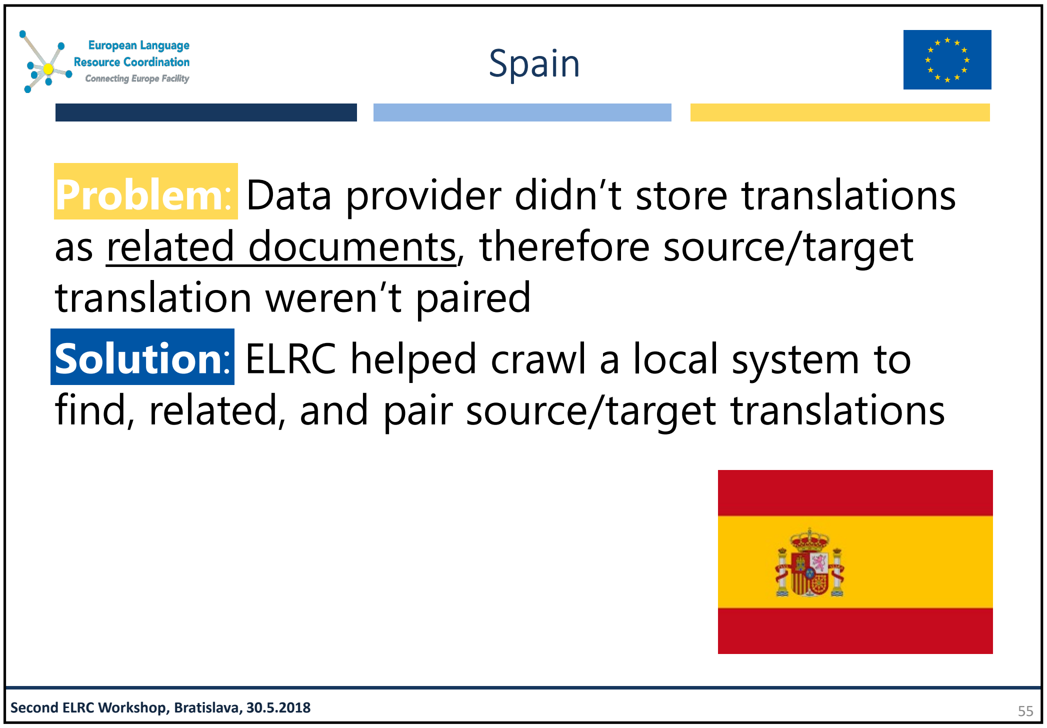
European Language Resource Coordination
Connecting Europe Facility

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

European Language Resource Coordination
Connecting Europe Facility

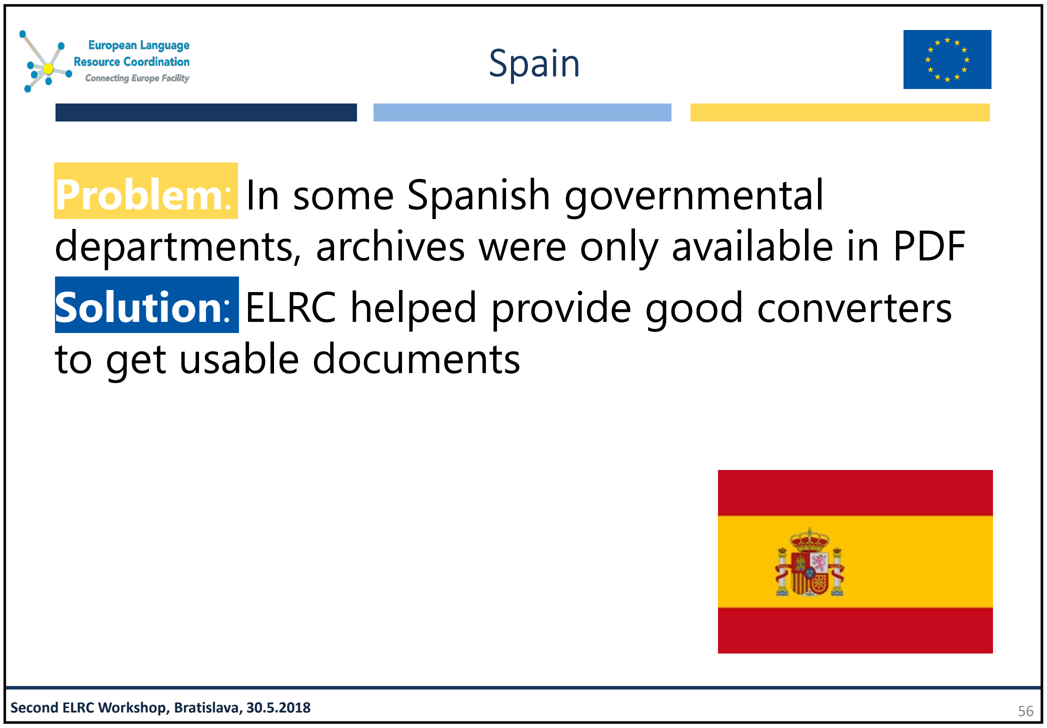
Case studies (2015-2016)



Spain

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

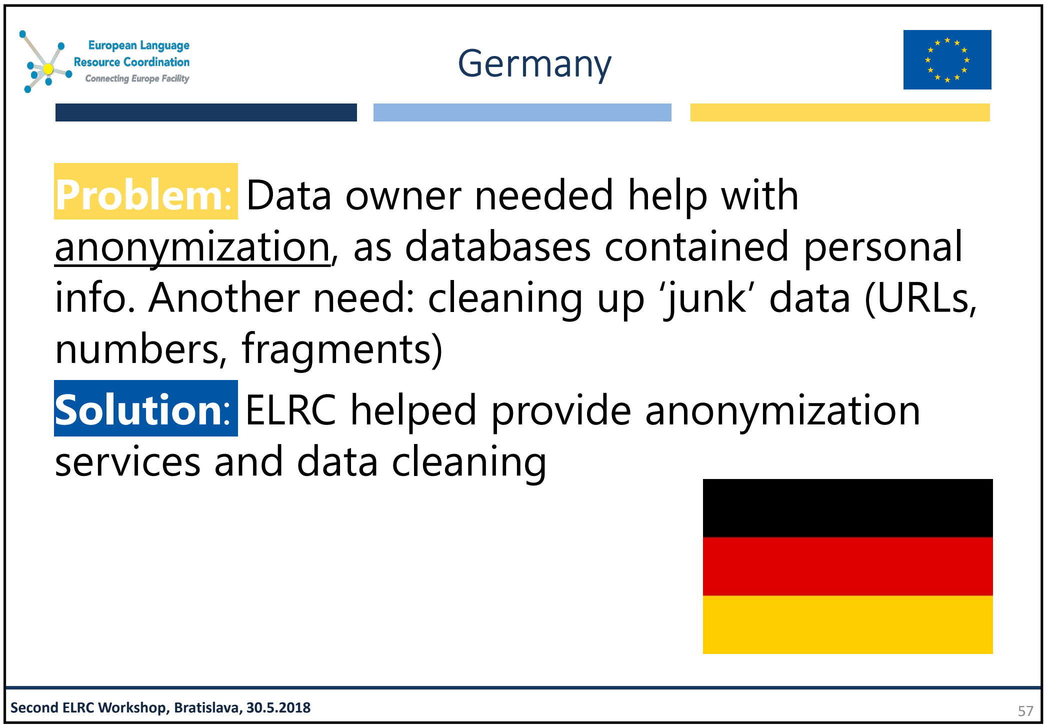
Solution: ELRC helped crawl a local system to find, related, and pair source/target translations



Spain

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

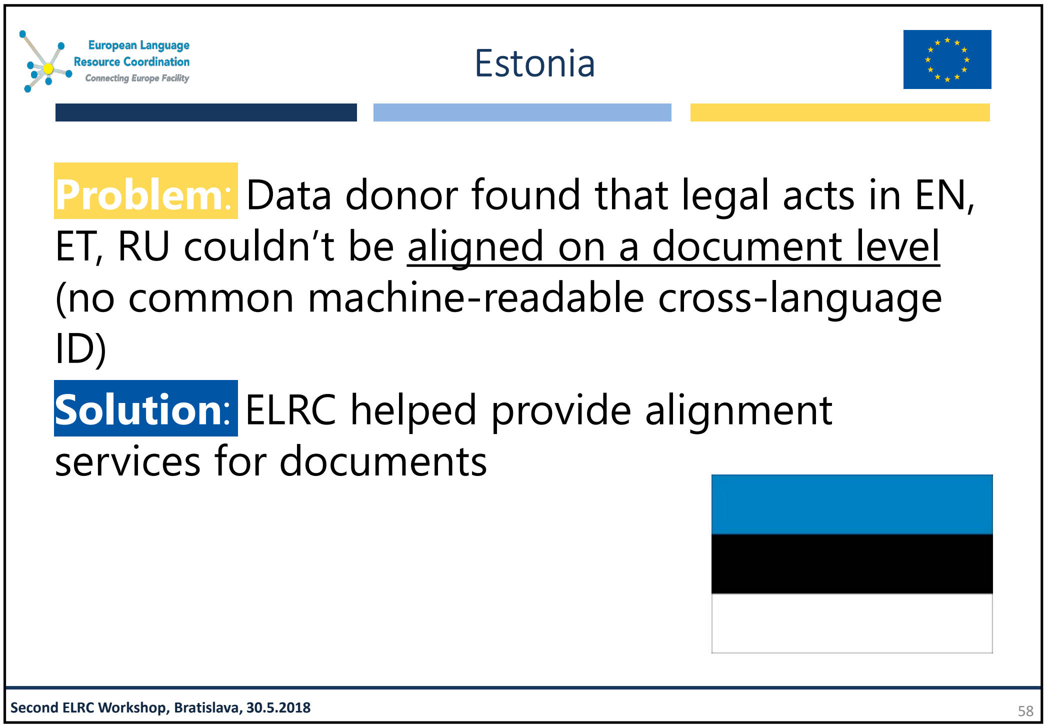
Solution: ELRC helped provide good converters to get usable documents



Germany

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



Estonia

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