

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

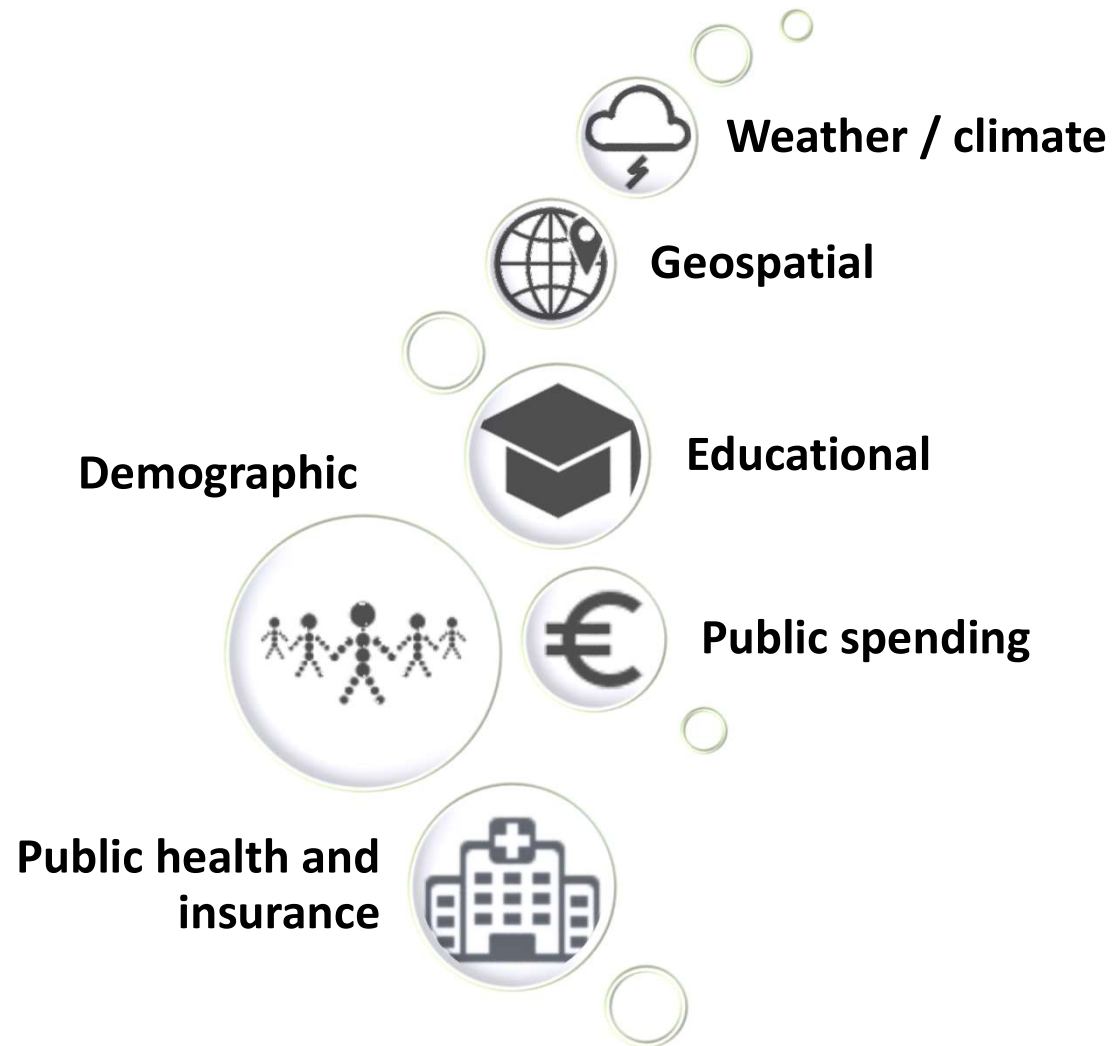
Maria Giagkou

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



Connecting
Europe
Facility

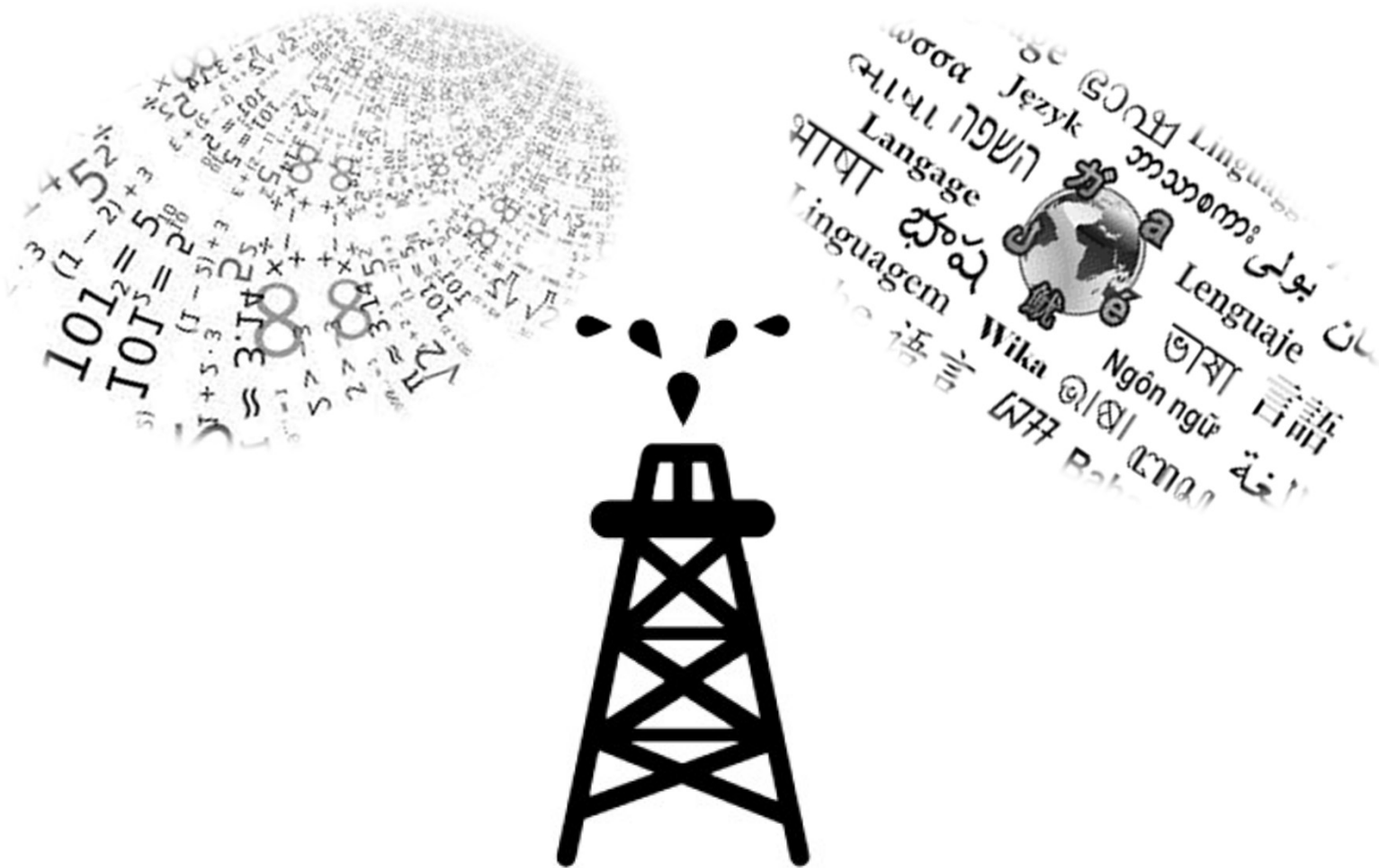
The notion of data



Data: the oil of the 21st century



The notion of data



The notion of data

Basic concepts:

- **Data:** any piece of electronically stored content
- **Dataset (or resource):** the collection of one or many data files **grouped** according to certain **criteria**
- **Metadata:** *data about the data*, i.e. description of a dataset with properties (e.g. title, publisher, description of the content and URL)



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

Data Applications Linked Data Developers' corner About

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

Publisher
Publications Office »

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

Resources

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

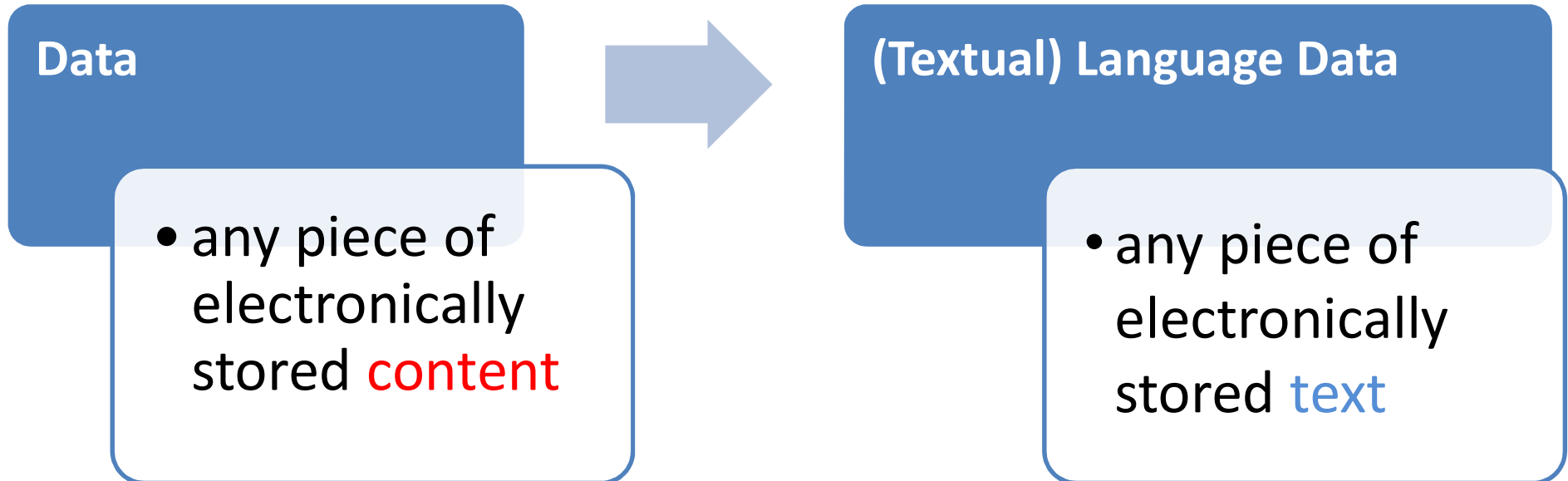
URI
<http://cordis.europa.eu/projects/>

Status
Under Development

Licence:
Legal Notice

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

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





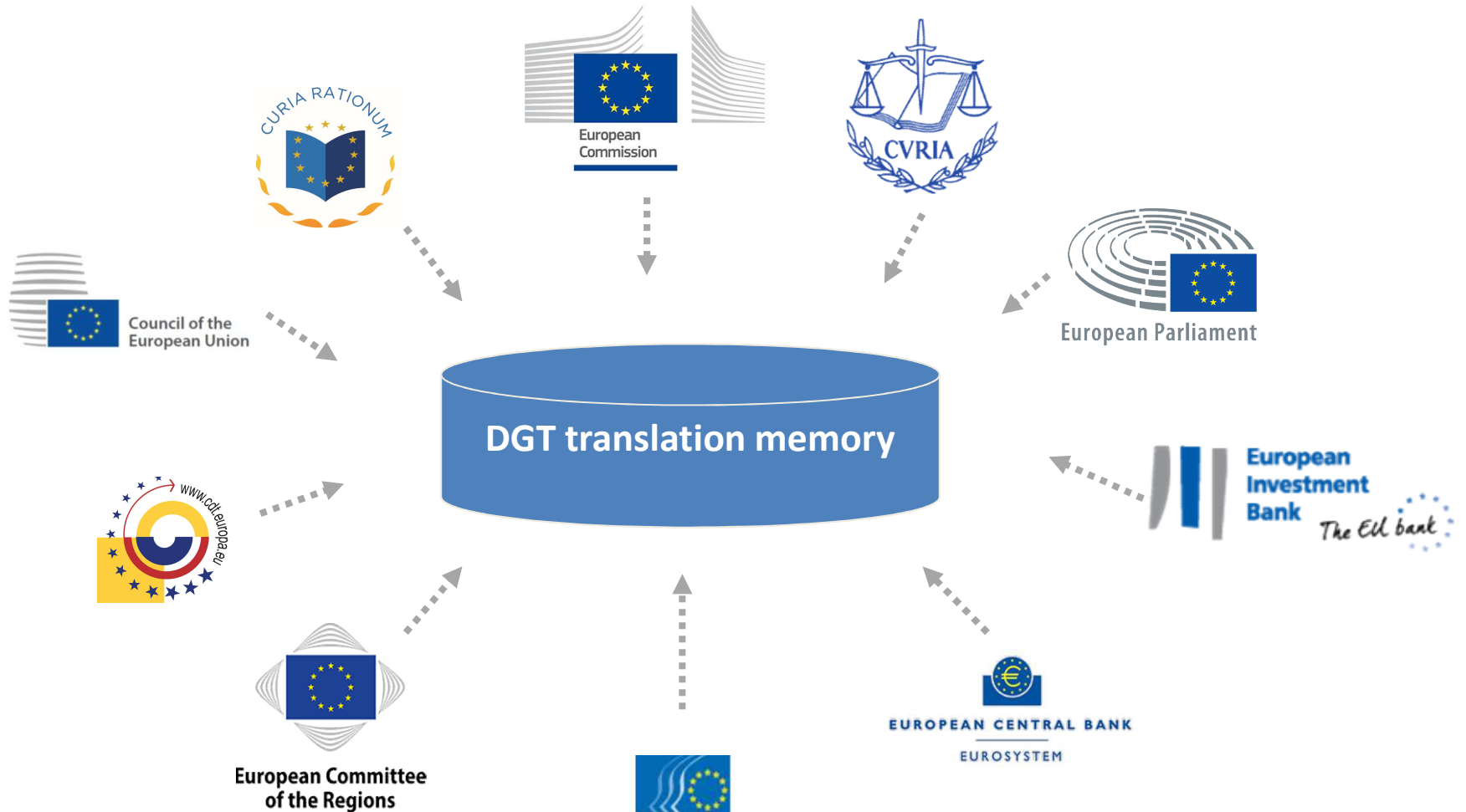
CS

Celkové výdaje na dávky nemocenského pojištění v roce 2005, tedy včetně výdajů proplácených organizacím nad 25 zaměstnanců, činily 31 miliard 660 milionů korun. Nemocensky pojištění obyvatelé ČR v roce 2005 prostonali celkem 107 095 134 dnů, což je o 1 139 608 dnů více než v roce 2004.

EN

Total sickness benefit expenditures in 2005, including reimbursements to organisations with over 25 employees, were CZK 31,660 million. In 2005, Czech citizens affiliated to the sickness insurance scheme spent 107,095,134 days on sick leave, which is 1,139,608 days more than in 2004.

Data used by eTranslation



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



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

Czech text

Úvodní ustanovení

- (1) Rada zasedá podle potřeby, nejméně však čtyřikrát ročně.
- (2) Zasedání Rady svolává její předseda / předsedkyně (dále jen „předseda“) prostřednictvím.
- (3) Program zasedání navrhuje předsednictvo Rady a schvaluje její předseda v souladu s úkoly Rady nebo podle závěrů jejího předchozího zasedání.
- (4) Rada je způsobilá jednat a přijímat závěry, pokud je přítomna nadpoloviční většina jejích členů nebo jejich stálých zástupců.
- (5) Závěry se přijímají hlasováním.

Translation in English

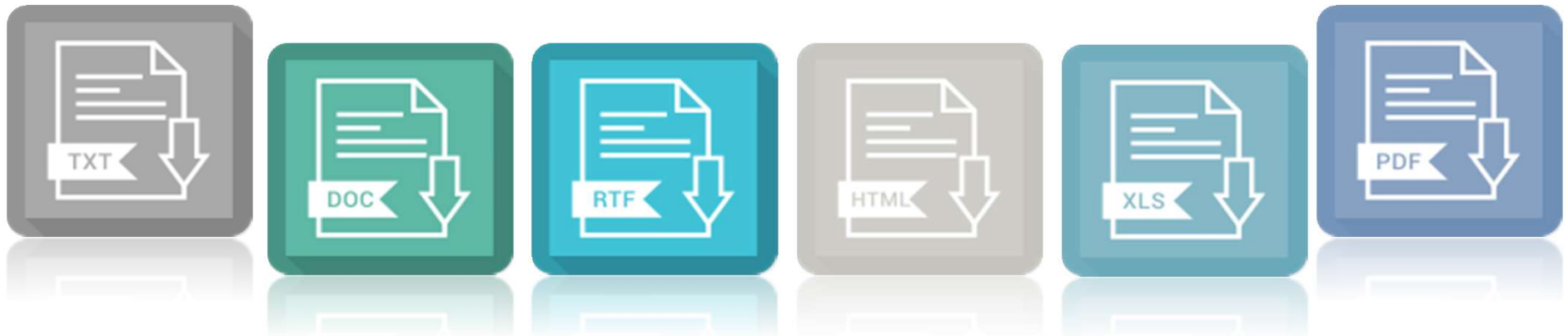
Introductory provisions

- (1) The Council shall meet as needed, however at least four times a year.
- (2) Council sessions shall be called by its Chair via the Council Secretariat.
- (3) The meeting programme is presented by the Council Board and approved by its Chair in accordance with the Council's tasks or the conclusions of its previous meeting.
- (4) The Council is competent to act and adopt conclusions if a majority of its members or their permanent representatives is present.
- (5) Conclusions shall be adopted by vote.

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

Czech	English
Korporátní daň	Corporate tax
Korporátní obligace	Corporate bond
Koš (ve vztahu k derivátům)	Basket
Košová opce	Basket option
Kotace	Quote
Kotovací tabule	Quotation board
Kotovaná cena	Quoted price
Kotovaná společnost	Quoted company
Kotování	Quotation
Koupě při otevření	Buy on opening
...	...

What data are useful for eTranslation as per format | 1



- In principle, any text in machine readable format
- But, some formats are more “MT-ready” than others, i.e. they require less manual or automatic processing
- More processing introduces more errors in the final output, making it less useful for eTranslation





- 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



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. ¶

A sentence 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 odstavec vlevo, přeložený v češtině. Toto je odstavec vlevo, přeložený v češtině. Toto je odstavec vlevo, přeložený v češtině. Toto je odstavec vlevo, přeložený v češtině. Toto je odstavec vlevo, přeložený v češtině. Toto je odstavec vlevo, přeložený v češtině. ¶

Toto je věta vlevo, přeložená v češtině. ¶

¶

Jedná se o druhý odstavec, přeložený česky. Jedná se o druhý odstavec, přeložený česky. Jedná se o druhý odstavec, přeložený česky. Jedná se o druhý odstavec, přeložený česky. Jedná se o druhý odstavec, přeložený česky. ¶

¶



Don'ts



English	čeština
<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. 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>Toto je odstavec vlevo, přeložený v češtině. Toto je odstavec vlevo, přeložený v češtině. Toto je odstavec vlevo, přeložený v češtině. Toto je odstavec vlevo, přeložený v češtině. Toto je odstavec vlevo, přeložený v češtině. Toto je odstavec vlevo, přeložený v češtině. Toto je odstavec vlevo, přeložený v češtině.</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. A second paragraph in English. A second paragraph in English. A second paragraph in English.</p>	<p>Jedná se o druhý odstavec, přeložený česky. Jedná se o druhý odstavec, přeložený česky. Jedná se o druhý odstavec, přeložený česky. Jedná se o druhý odstavec, přeložený česky. Jedná se o druhý odstavec, přeložený česky.</p>



- filename01_CS.txt
- filename01_EN.txt
- filename02_CS.txt
- filename02_EN.txt
- filename03_CS.txt
- filename03_EN.txt
- filename04_CS.txt
- filename04_EN.txt
- filename05_CS.txt
- filename05_EN.txt
- filename06_CS.txt
- filename06_EN.txt
- filename07_CS.txt
- filename07_EN.txt
- filename08_CS.txt
- filename08_EN.txt
- filename09_CS.txt
- filename09_EN.txt
- filename10_CS.txt
- filename10_EN.txt

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



- filename01_CS.txt
- filename01_EN.txt
- filename02_CS.txt
- filename02_EN.txt
- filename03_CS.txt
- filename03_EN.txt
- filename04_CS.txt
- filename04_EN.txt
- filename05_CS.txt
- filename05_EN.txt
- filename06_CS.txt
- filename06_EN.txt
- filename07_CS.txt
- filename07_EN.txt
- filename08_CS.txt
- filename08_EN.txt
- filename09_CS.txt
- filename09_EN.txt
- filename10_CS.txt
- filename10_EN.txt

Include **language identifiers** in the filename



Όνομα	Αρχείο	Επεξεργασία	Μορφή	Προβολή	Βοήθεια
τεκστ.txt	English	Bulgarian	French	Greek	
Μετάφραση αναφοράς.txt	text.txt	текст.txt	texte.txt	κείμενο.txt	
κείμενο.txt	report.txt	report in Bulgarian.txt	traduction française.txt	Μετάφραση αναφοράς.txt	
traduction française.txt					
texte.txt					
text.txt					
report.txt					
report in Bulgarian.txt					
README.txt					



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



- 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

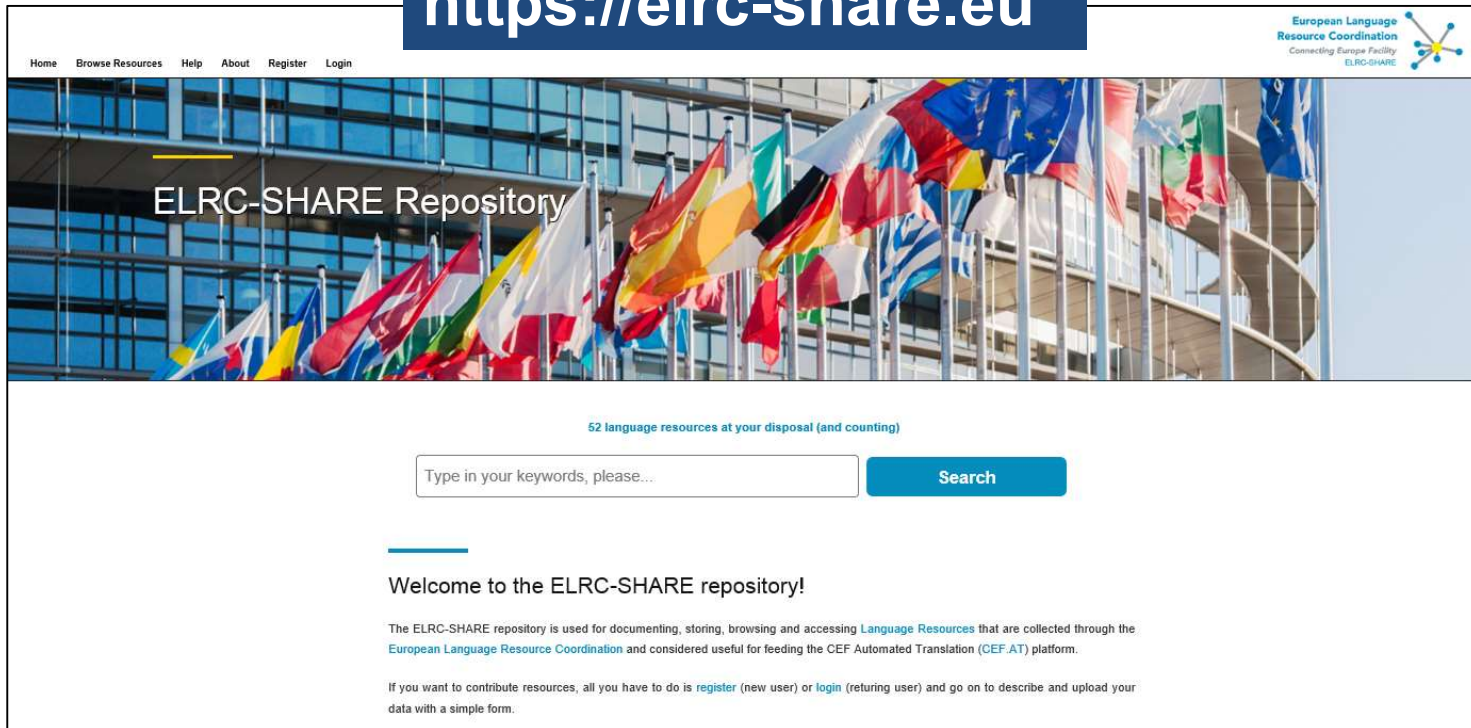
How to contribute your data to CEF eTranslation

A step-by-step guide

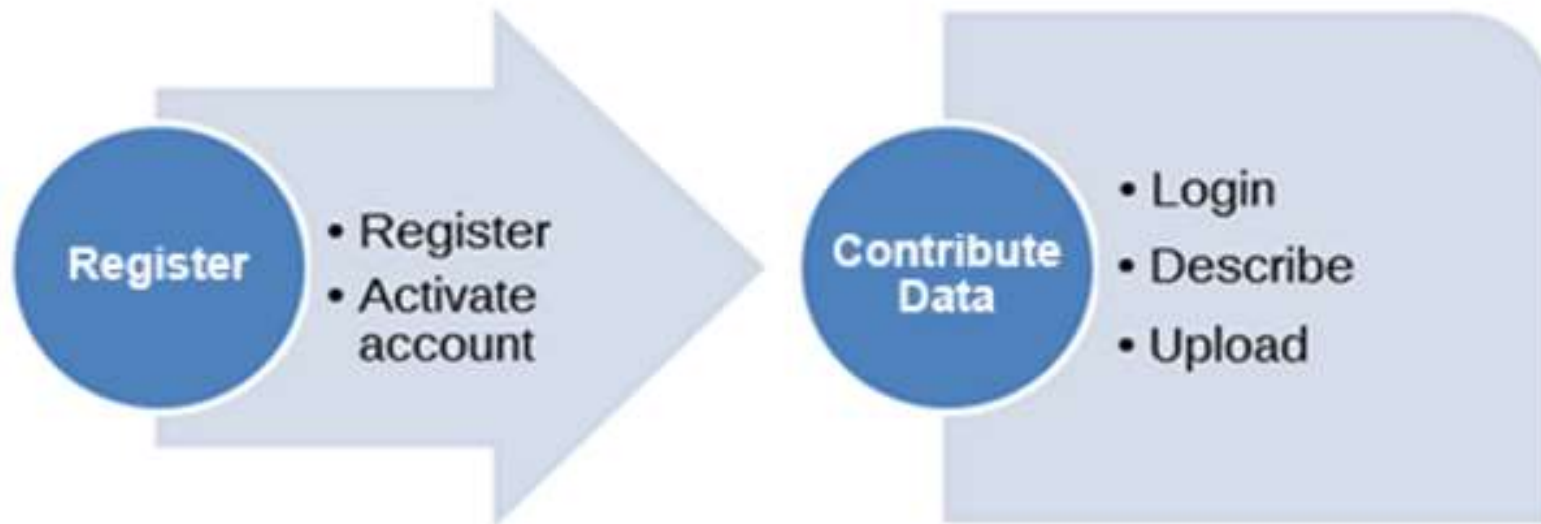
The ELRC-SHARE Repository

- Access to, sharing and contribution of language resources
- Access to tools and services catalogue (upcoming)

<https://elrc-share.eu>



The screenshot shows the homepage of the ELRC-SHARE Repository. At the top, there is a navigation menu with links for Home, Browse Resources, Help, About, Register, and Login. The main header features the text "ELRC-SHARE Repository" overlaid on a background image of various European national flags. Below the header, a blue box indicates "52 language resources at your disposal (and counting)". A search bar with the placeholder text "Type in your keywords, please..." and a blue "Search" button is provided. The main content area includes a welcome message: "Welcome to the ELRC-SHARE repository!" followed by a paragraph explaining the repository's purpose: "The ELRC-SHARE repository is used for documenting, storing, browsing and accessing Language Resources that are collected through the European Language Resource Coordination and considered useful for feeding the CEF Automated Translation (CEF.AT) platform." A final paragraph states: "If you want to contribute resources, all you have to do is register (new user) or login (returning user) and go on to describe and upload your data with a simple form."



How to Register (1/2)



 Register

[Home](#) [Browse Resources](#) [Help](#) [About](#) [Register](#) [Login](#)

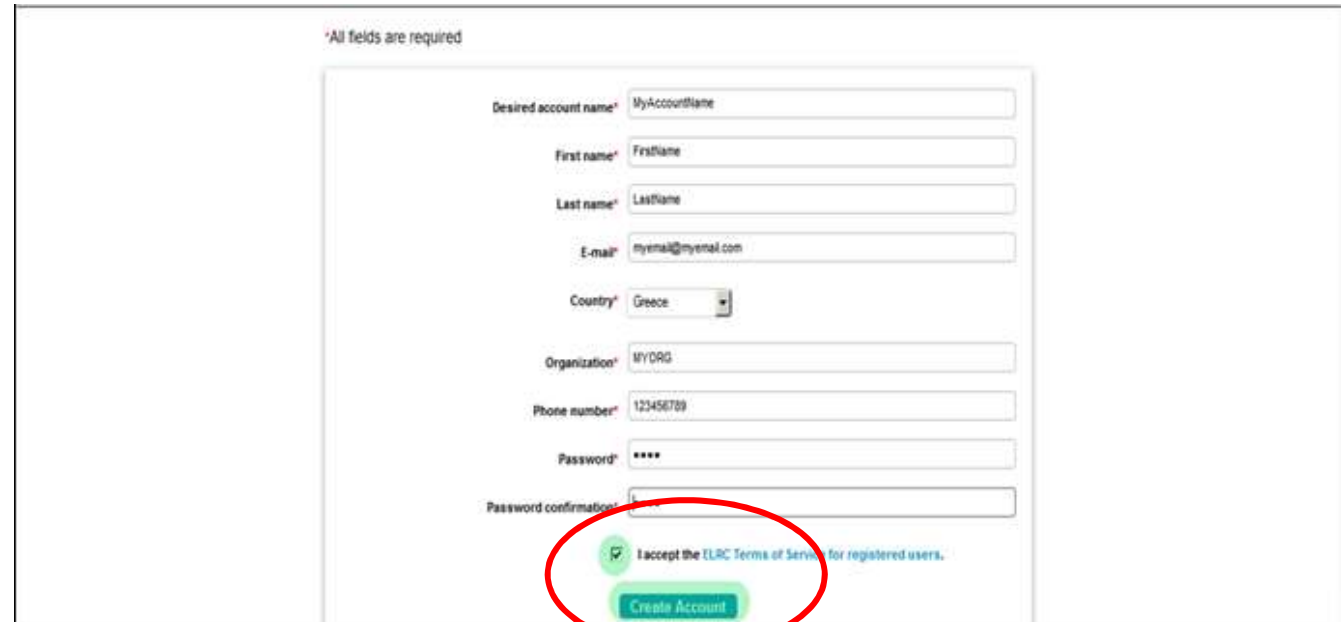
ELRC-SHARE Repository



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



The screenshot shows a registration form with the following fields: Desired account name* (MyAccountName), First name* (Firstname), Last name* (Lastname), E-mail* (myemail@myemail.com), Country* (Greece), Organization* (MYORG), Phone number* (123456789), Password* (****), and Password confirmation*. A red circle highlights the 'I accept the ELRC Terms of Service for registered users.' checkbox and the 'Create Account' button.

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



The screenshot shows a web 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 area 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 (highlighted in blue), Estonian, Finnish, French (highlighted in blue), German, and Hungarian.

- 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*

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

Submit

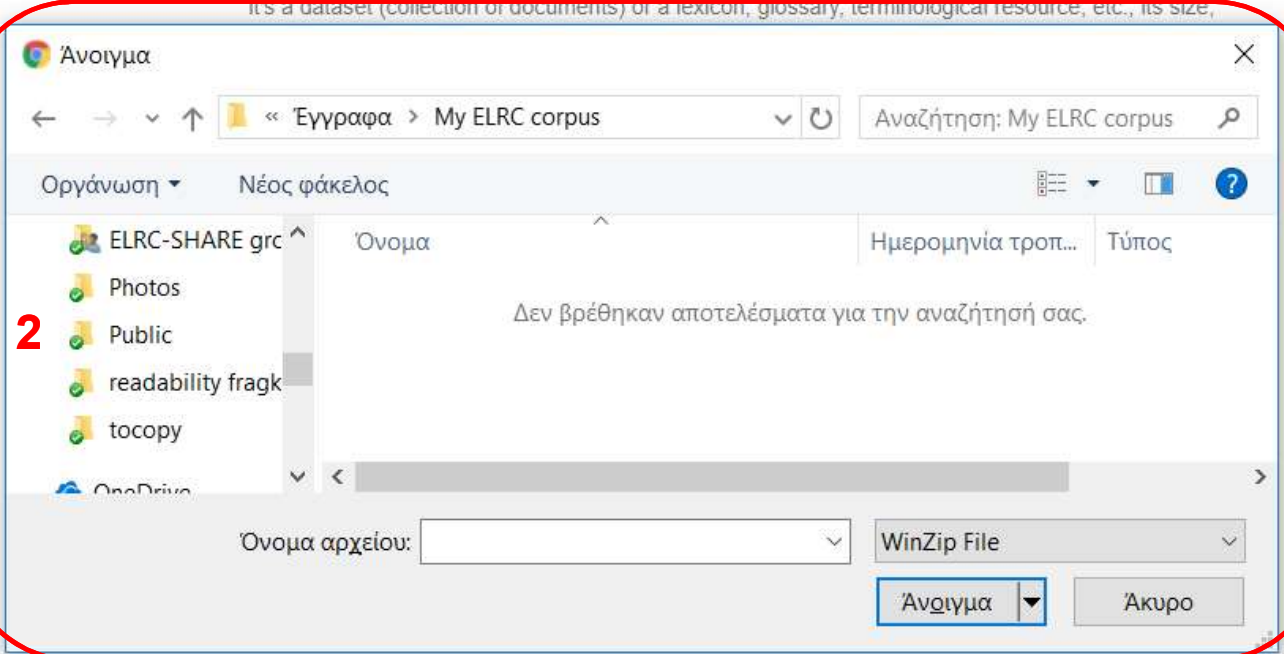
Reset



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

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,



2

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@lsp.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*

Please provide a URL containing the files you wish to contribute



Contribute Your Data Through eDelivery

If you wish to share your data through [eDelivery](#), you can use the ELRC-SHARE CEF eDelivery Access Point. For more information click [here](#).
In such a case, please fill in the form on the left and choose eDelivery in the Contribution mode.



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:

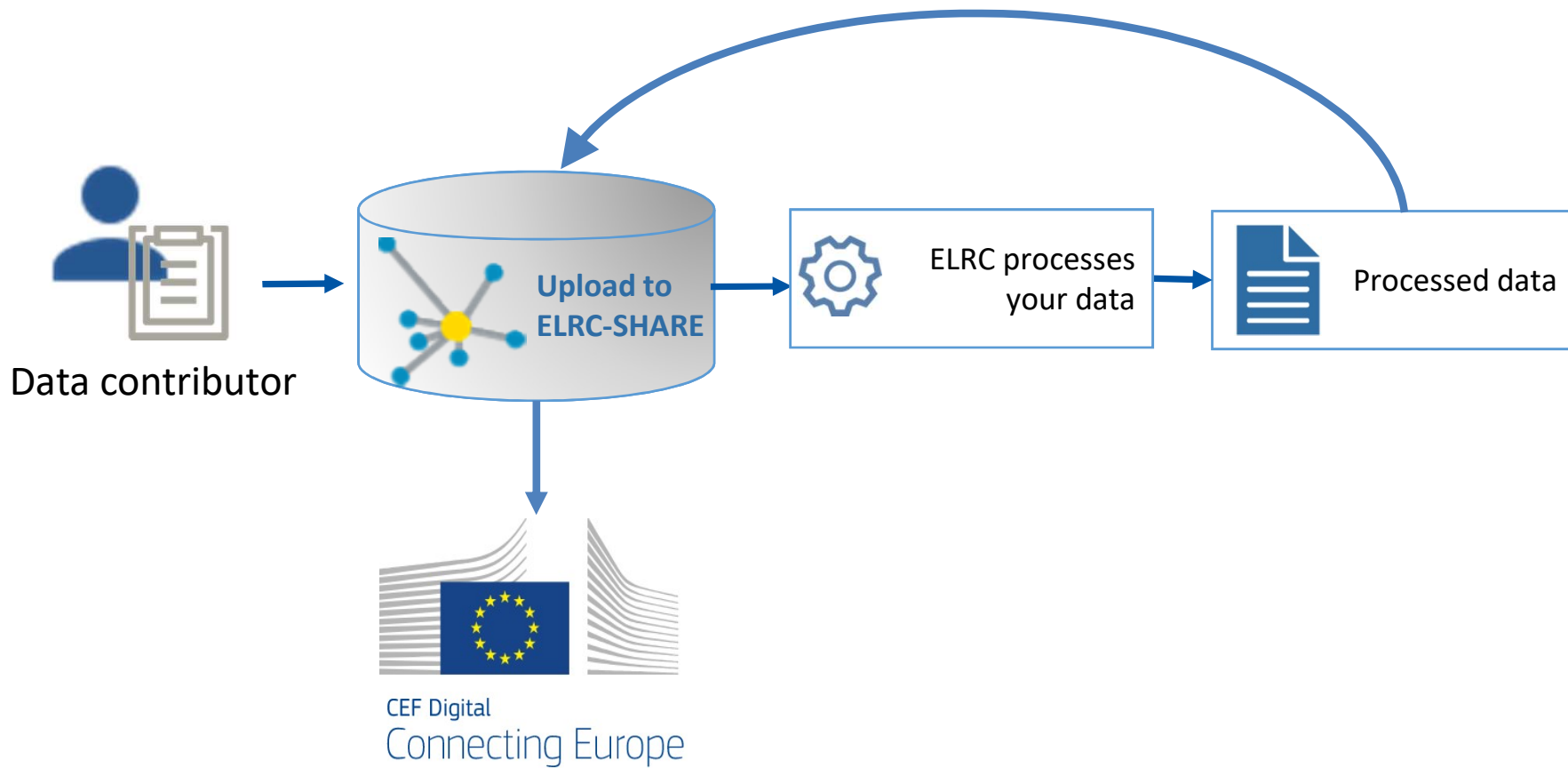
- [Walkthrough for contributors](#)
- [Walkthrough for editors](#)

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?





Data extraction

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



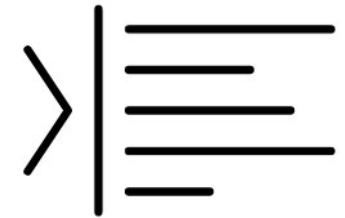
Anonymisation

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



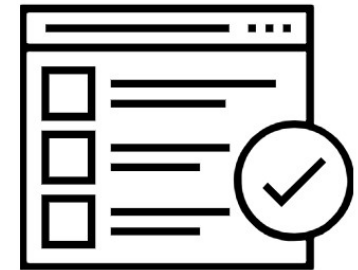
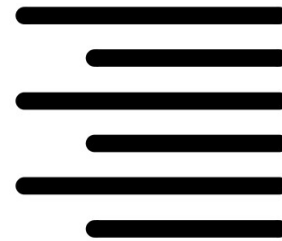
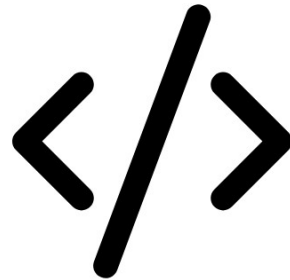
Cleaning

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 has happened to your data?

File01_cs.txt
File01_en.doc
File02_cs.pdf
File02_en.txt
File03_cs.doc
File03_en.doc
...

After
proces
sing

```
<tu tuid="818">
  <tuv xml:lang="cs">
    <seg>Celkové výdaje na dávky nemocenského
    pojištění v roce 2005, tedy včetně výdajů proplácených
    organizacím nad 25 zaměstnanců, činily 31 miliard 660
    milionů korun.</seg> </tuv>
  <tuv xml:lang="en">
    <seg>Total sickness benefit expenditures in
    2005, including reimbursements to organisations with
    over 25 employees, were CZK 31,660
    million.</seg></tuv></tu>
  <tu tuid="819">
    <tuv xml:lang="cs">
      <seg>Nemocensky pojištění obyvatelé ČR v roce
      2005 prostonali celkem 107 095 134 dnů, což je o 1 139
      608 dnů více než v roce 2004.</seg></tuv>
    <tuv xml:lang="en">
      <seg>In 2005, Czech citizens affiliated to the
      sickness insurance scheme spent 107,095,134 days on sick
      leave, which is 1,139,608 days more than in
      2004.</seg></tuv></tu>
```

Electronic Exchange of Social Security Information documents in Czech-English (Processed)

40 Word docs containing documents about EESSI translated into Czech. (Processed)

DSI Relevance: ElectronicExchangeOfSocialSecurityInformation

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

Distribution

Availability: Available

[Licences](#)

[Terms for PSI-compliant resources](#)

[Open Under-PSI](#)

[Distribution Details](#)

Contact Person

[Marie Cernikova](#) 

text 

Bilingual text corpus

Languages

English (en)

Czech (cs)

Linguality

Linguality type: Bilingual

Text Format

TMX

Size

17,356 Translation Units

Character encoding

UTF-8

Resource Creation

Funding Project

European Language Resource Coordination LOT3 (ELRC Data - Tools and Resources for CEF Automated Translation - LOT3 (SMART 2015/1091 - 30-CE-0816766/00-92))

URL: <http://www.lr-coordi...>

Funding Type: Service Contract

Funder: European Commission

Funding Country: European Union (EU)

Project duration: 13/12/2016 - 12/02/2020

Metadata

Created: 20/07/2017

Metadata Language: English (en)

Version

Version: 2.0

Relations

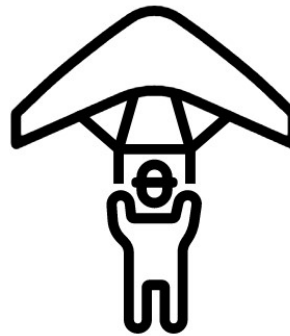
Related Resource: [Electronic Exchange of Social Security Information documents in Czech-English](#)

Relation Type: Is Aligned Version Of



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



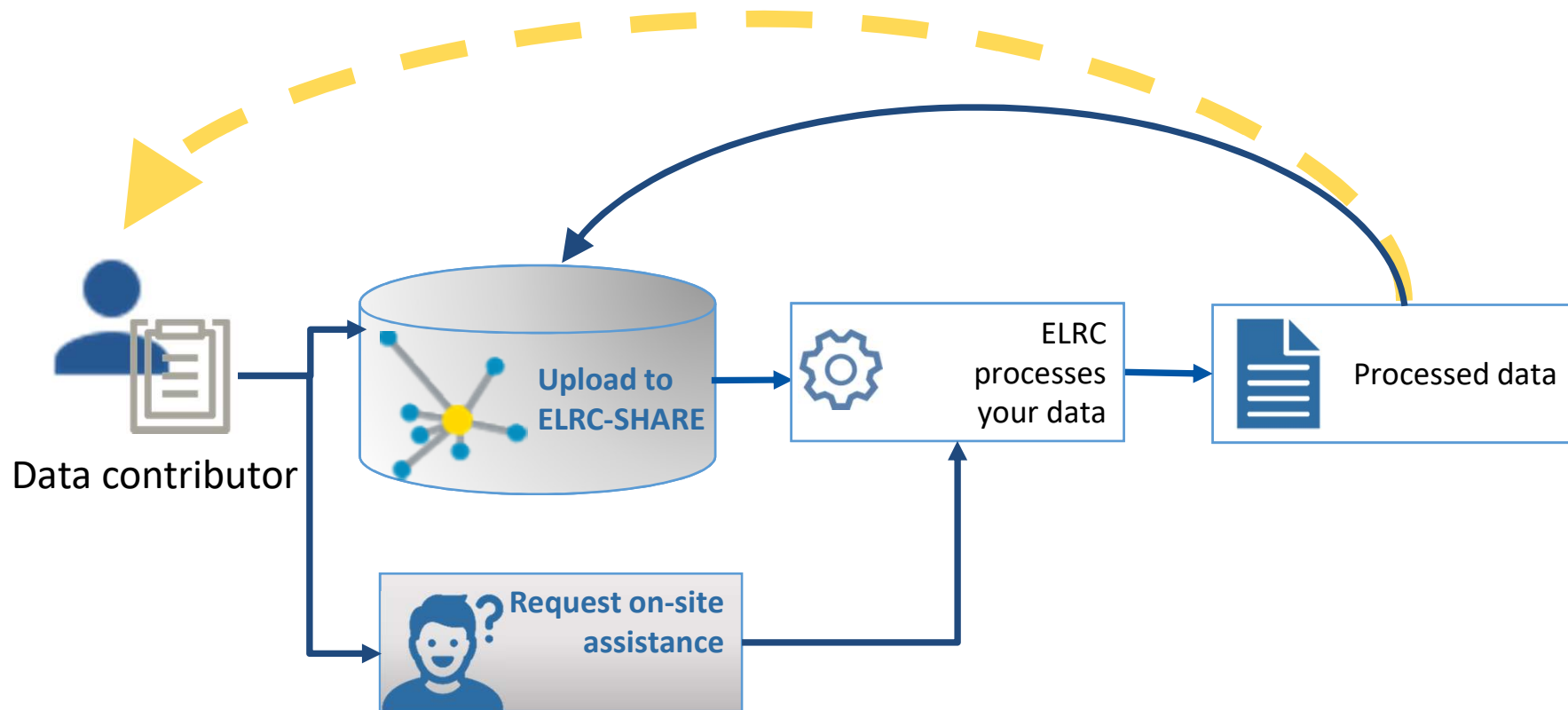


**Our team of experts will travel
directly to assist you
at your own offices**



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!

What happens to your data?



How to request services and help



www.lr-coordination.eu/request-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

Submit

www.lr-coordination.eu/helpdesk

[Home](#) [Discover](#) [Resources](#) [Services](#) [Events](#) [Anchor Points](#) [News](#) [Helpdesk](#)



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-coordination.eu

Děkuji!



- 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

