



European Language
Resource Coordination
Connecting Europe Facility



Preparing and sharing data via the
ELRC-SHARE repository
ELRC on site Services and what happens next


Maltese Workshop
Khalid Choukri, H  l  ne Mazo
ELRA/ELDA




Connecting
Europe
Facility

Second ELRC Workshop, Malta, 2019/11/01

1



European Language
Resource Coordination
Connecting Europe Facility



The notion of language data

Data


- any piece of electronically stored **content**

(Textual) Language Data


- any piece of electronically stored **text**


Second ELRC Workshop, Malta, 2019/11/01

2



The notion of data

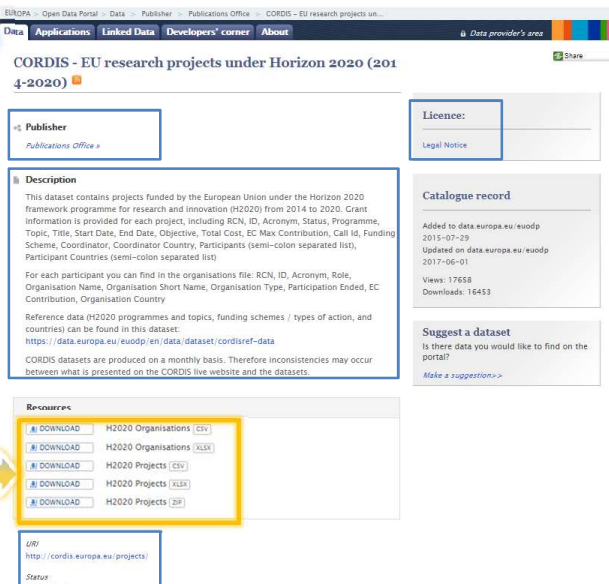




[Stemap](#) | [Legal notice](#) | [Contact](#) | [English vlv](#)


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)




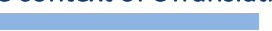
Second ELRC Workshop, Malta, 2019/11/01

3



The notion of data in the context of eTranslation





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

Multilingual subtitle data 2BDutch

Attribution details: NTU - Nederlandse Taalunie

The website www.2BDutch.nl presents videos with spoken Dutch containing English, German, Spanish, French and Portuguese subtitles. Student of Dutch can practice their listening skills and learn new Dutch words. These subtitles form the present corpus.

[Back](#) [Download](#)

Distribution

Availability: Available

Licences: [License Agreement between ELRC and NTU](#)

Non-restricted Other Licence Terms: [Conditions: Non Commercial Use](#)

Distribution Details: [Attribution Details: NTU - Nederlandse Taalunie](#)

Contact Person

Carole Thebaux ^{RI}

Multilingual text corpus

Languages: French (fr), English (en), Dutch: Flemish (nl), Portuguese (pt), German (de), Spanish, Castilian (es)

Linguality: Multilingual

Text Format: TEI-X

Size: 32,549 Translation Units

Character encoding: UTF-8

Resource Creation

Funding Project: [Connecting Europe Facility - European Language Resource Coordination \(CEF-ELRC - LANGUAGE RESOURCE COORDINATION - SMART 2014-1074 - 30-02-06/07/2015/0164\)](#)

URL: <http://www.fr-ccordn...>

Funding Type: Service Contract

Funder: European Commission

Funding Country: European Union (EU)

Project duration: 20/03/2015 - 16/04/2017

Metadata

Created: 12/04/2017

Last Updated: 12/04/2017

Metadata Language: English (en)

Metadata Creator: [Fraser Bowen](#) ^{RI}

Relations

Related Resource: [Multilingual subtitle data 2BDutch \(Processed\)](#)

Relation Type: has Converted version

People who looked at this resource also viewed the following:

- [Paralel Corbal Voces \(Greek - Spanish\)](#)
- [Paralel kids from Sweden Labour market agency](#)
- [Spanish-English website paralel corpus](#)
- [Spanish-Portuguese website paralel corpus](#)


Resources from the same project

Second ELRC Workshop, Malta, 2019/11/01

4

European Language Resource Coordination
Connecting Europe Facility

The notion of data in the context of eTranslation



```
File01_mt.txt
File01_en.txt
File02_mt.txt
File02_en.txt
File03_mt.txt
...
```

Trans.
Da

Malta ([/ˈmɒltə/](#),^[14] [/ˈmɒl.tə/](#) (listen)); Maltese: [ˈmɛltɛ], officially known as the **Republic of Malta** (Maltese: *Repubblika ta' Malta*), is a Southern European [island country](#) consisting of an [archipelago](#) in the [Mediterranean Sea](#).^[12] It lies 80 km (50 mi) south of [Italy](#), 284 km (176 mi) east of [Tunisia](#),^[13] and 333 km (207 mi) north of [Libya](#).^[14] With a population of about 475,000^[15] over an area of 316 km² (122 sq mi),^[16] Malta is the [world's tenth smallest](#)^{[15][16]} and [fifth most densely populated](#) sovereign country. Its capital is [Valletta](#), which is the smallest national capital in the [European Union](#) by area at 0.8 km². The official languages are [Maltese](#) and English, with Maltese officially recognised as the national language and the only [Semitic language](#) in the European Union.


Malta, magħrufa uffiċjalment bħala **r-Repubblika ta' Malta**, hija repubblika konstituzzjonali żviluppata [Ewropew](#) fi hdan l-[Unjoni Ewropea](#).^{[15][16]} Il-pajjiż ikkonsisti minn [arċipelagu](#) ta' seba' gżejjer li jinsabu fil-[Bahar Mediterran](#), fl-[Ewropa t'Isfel](#), 93 km mill-kosta ta' [Sgallija](#) (l-[Italia](#)), 288 km fil-Lvant tat-[Tuneżija](#) u 300 km fit-Tramuntana tal-[Libja](#). Dawn il-gżejjer igawdu minn [klima Mediterranja](#).^{[12][16]} Matul l-[istorja](#), il-pożizzjoni ta' Malta fil-Bahar Mediterran kellha sinjifikat strateġiku.^[12] Minhabba dan il-fattur, numru ta' qawwiet fosthom il-[Feniċi](#), ir-[Rumani](#), l-[Gharab](#), l-[Isgallin](#), il-[Kavallieri ta' San Ġwann](#) u l-[Inglizi](#) ikoll hakmu lil Malta. Malta kisbet l-indipendenza mir-[Renju Unit](#) fl-1964 u għaxar snin wara, fl-1974, saret Repubblika però xorta baqgħet membru tal-[Commonwealth tan-Nazzjonijiet](#). Attwalment hi membru tal-[Unjoni Ewropea](#), wara s-sħubija fl-2004, u wkoll tan-[Nazzjonijiet Uniti](#). Il-Lingwa Maltija hija l-unika lingwa Semitika li hija lingwa formali tal-Unjoni Ewropea.

Second ELRC Workshop, Malta, 2019/11/01

5

European Language Resource Coordination
Connecting Europe Facility

The notion of data in the context of eTranslation



```
File01_it.txt
File01_en.txt
File02_it.txt
File02_en.txt
File03_it.txt
...
```

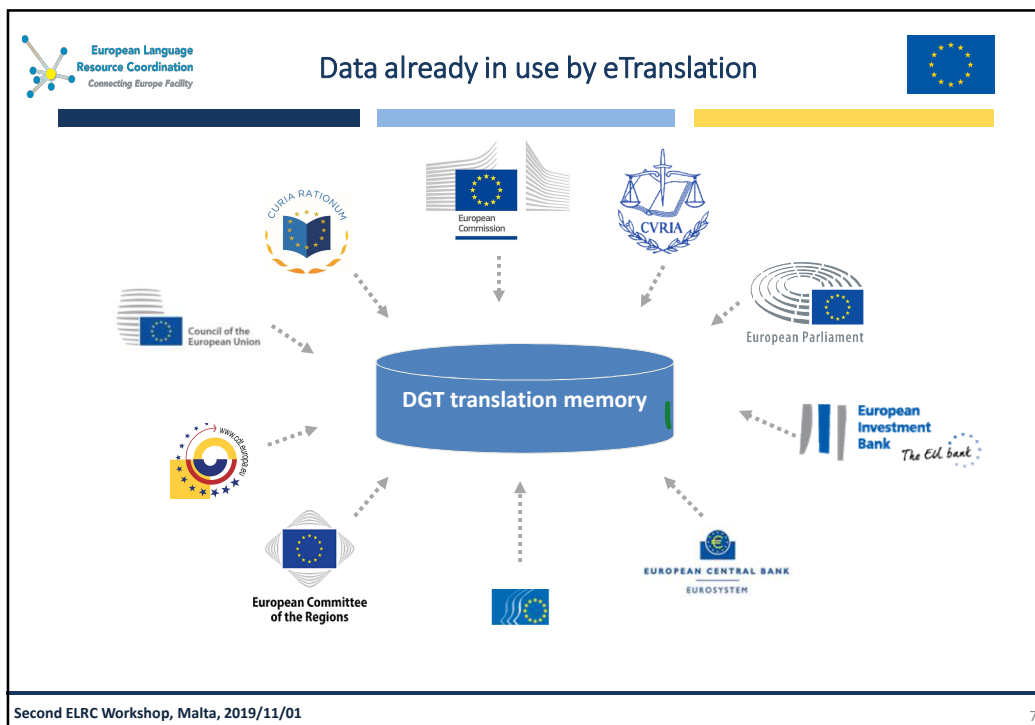
Trans.
Da

```
tuv xml:lang="nl"
changedate="20151214T133604Z">
  <seg>Deze gecoördineerde wet
  stelt een regeling voor
  verplichte verzekering voor
  geneeskundige verzorging en
  uitkeringen in; ze organiseert
  die in twee onderscheiden takken
  die betrekking hebben, de ene op
  de geneeskundige verstrekkingen,
  de andere op de uitkeringen
  wegens arbeidsongeschiktheid [...]
  en op de
  moederschapsverzekering.</seg>
```

```
</tuv>
<tuv xml:lang="fr"
  changedate="20151214T133604Z">
  <seg>La présente loi
  coordonnée institue un régime
  d'assurance obligatoire soins
  de santé et indemnités; elle
  l'organise en deux secteurs
  distincts relatifs, l'un aux
  prestations de santé, l'autre
  aux indemnités d'incapacité de
  travail, [...] et à l'assurance
  maternité.</seg>
</tuv>
```


Second ELRC Workshop, Malta, 2019/11/01

6





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

The slide features the same header and decorative bar as the previous slide. The main content is centered text. The footer contains the text "Second ELRC Workshop, Malta, 2019/11/01" on the left and the number "8" on the right.





What does eTranslation need?






What data are useful for eTranslation as per type | 1







Dutch text



Ter toepassing van de bepalingen van deze gecoördineerde wet worden de landsbonden gemachtigd die het waren ter toepassing van het organiek koninklijk besluit van 22 september 1955 van de ziekte- en invaliditeitsverzekering




Translation in French




Sont agréées pour l'application des dispositions de la présente loi coordonnée les unions nationales qui l'étaient pour l'application de l'arrêté royal du 22 septembre 1955 organique de l'assurance maladie-invalidité.


Second ELRC Workshop, Malta, 2019/11/01

9




What data are useful for eTranslation as per type | 1







Dutch text



Ter toepassing van de bepalingen van deze gecoördineerde wet worden de landsbonden gemachtigd die het waren ter toepassing van het organiek koninklijk besluit van 22 september 1955 van de ziekte- en invaliditeitsverzekering




Translation in French




Sont agréées pour l'application des dispositions de la présente loi coordonnée les unions nationales qui l'étaient pour l'application de l'arrêté royal du 22 septembre 1955 organique de l'assurance maladie-invalidité.







Second ELRC Workshop, Malta, 2019/11/01

10




What data are useful for eTranslation as per format | 1











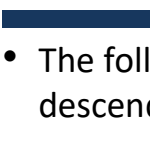
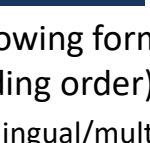
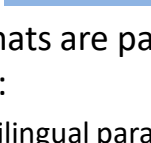
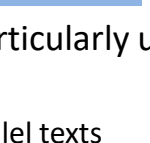
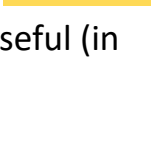

- In principle, any text in machine readable format
- But, some formats are more “MT-ready” than others, i.e. they require less manual or automatic processing
- More processing introduces more errors in the final output, making it less useful for eTranslation

Second ELRC Workshop, Malta, 2019/11/01
11





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)

Second ELRC Workshop, Malta, 2019/11/01
12


File formats of parallel texts and their manipulation

Second ELRC Workshop, Malta, 2019/11/01
13

Preparing your data | 1

Don'ts



Ter toepassing van de bepalingen van deze gecoördineerde wet worden de landsbonden gemachtigd die het waren ter toepassing van het organiek koninklijk besluit van 22 september 1955 van de ziekte- en invaliditeitsverzekering.

Sont agréées pour l'application des dispositions de la présente loi coordonnée les unions nationales qui l'étaient pour l'application de l'arrêté royal du 22 septembre 1955 organique de l'assurance maladie-invalidité.

De landsbonden waarborgen in hun statuten de bij deze wet bedoelde prestaties.


Les unions nationales garantissent, dans leurs statuts, les prestations prévues par la présente loi.

De machtiging van landsbonden die deze gecoördineerde wet of haar uitvoeringsbesluiten en – verordeningen niet naleven kan door de Koning, op advies of voorstel van het Algemeen comité van het Instituut, worden ingetrokken.


L'agrégation peut être retirée par le Roi, sur avis ou sur proposition du Comité général de l'Institut, aux unions nationales qui n'observent pas la présente loi coordonnée ou ses arrêtés et règlements d'exécution.


Don't merge the source and translated text into a single document

Second ELRC Workshop, Malta, 2019/11/01
14



Preparing your data | 2



Don'ts 

Ter toepassing van de bepalingen van deze gecoördineerde wet worden de landsbonden gemachtigd die het waren ter toepassing van het organiek koninklijk besluit van 22 september 1955 van de ziekte- en invaliditeitsverzekering.

De landsbonden waarborgen in hun statuten de bij deze wet bedoelde prestaties.


De machtiging van landsbonden die deze gecoördineerde wet of haar uitvoeringsbesluiten en – verordeningen niet naleven kan door de Koning, op advies of voorstel van het Algemeen comité van het Instituut, worden ingetrokken.

Sont agréées pour l'application des dispositions de la présente loi coordonnée les unions nationales qui l'étaient pour l'application de l'arrêté royal du 22 septembre 1955 organique de l'assurance maladie-invalidité.


Les unions nationales garantissent, dans leurs statuts, les prestations prévues par la présente loi.


L'agrément peut être retirée par le Roi, sur avis ou sur proposition du Comité général de l'Institut, aux unions nationales qui n'observent pas la présente loi coordonnée ou ses arrêtés et règlements d'exécution.

Second ELRC Workshop, Malta, 2019/11/01
15



Preparing your data | 3



Don'ts 


Ter toepassing van de bepalingen van deze gecoördineerde wet worden de landsbonden gemachtigd die het waren ter toepassing van het organiek koninklijk besluit van 22 september 1955 van de ziekte- en invaliditeitsverzekering.

De landsbonden waarborgen in hun statuten de bij deze wet bedoelde prestaties.


Sont agréées pour l'application des dispositions de la présente loi coordonnée les unions nationales qui l'étaient pour l'application de l'arrêté royal du 22 septembre 1955 organique de l'assurance maladie-invalidité.


Les unions nationales garantissent, dans leurs statuts, les prestations prévues par la présente loi.

Second ELRC Workshop, Malta, 2019/11/01
16



Preparing your data | 4





Do's


Name

- filename01_EN.txt
- filename01_SL.txt
- filename02_EN.txt
- filename02_SL.txt
- filename03_EN.txt
- filename03_SL.txt
- filename04_EN.txt
- filename04_SL.txt
- filename05_EN.txt
- filename05_SL.txt
- filename06_EN.txt
- filename06_SL.txt
- filename07_EN.txt
- filename07_SL.txt
- filename08_EN.txt
- filename08_SL.txt
- filename09_EN.txt
- filename09_SL.txt
- filename10_EN.txt
- filename10_SL.txt


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


Second ELRC Workshop, Malta, 2019/11/01

17



Preparing your data | 5





Do's


Name

- filename01_EN.txt
- filename01_SL.txt
- filename02_EN.txt
- filename02_SL.txt
- filename03_EN.txt
- filename03_SL.txt
- filename04_EN.txt
- filename04_SL.txt
- filename05_EN.txt


Include language identifiers in the filename


Second ELRC Workshop, Malta, 2019/11/01

18



Preparing your data | 5




Do's


Όνομα

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


Αρχείο επεξεργασίας μεταφράσεων

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

Second ELRC Workshop, Malta, 2019/11/01
19




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*

Second ELRC Workshop, Malta, 2019/11/01
20



Examples of 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

Second ELRC Workshop, Malta, 2019/11/01

21





How to contribute your data to CEF eTranslation


A step-by-step guide

Second ELRC Workshop, Malta, 2019/11/01

22

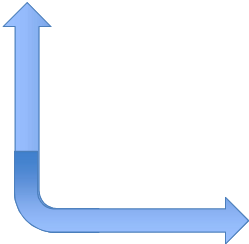


To ELRC-SHARE



- 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 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 CEFAT.

Language resource submission ▶

Second ELRC Workshop, Malta, 2019/11/01

23



ELRC-SHARE repository





Home
Browse Resources
Help
About
Register
Login

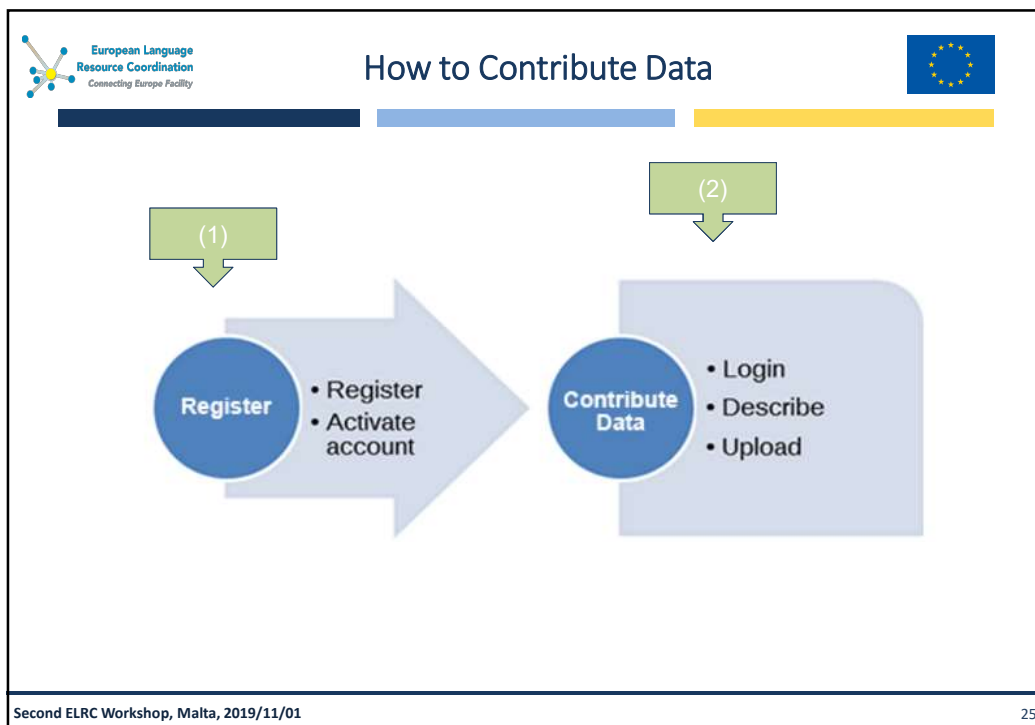
ELRC-SHARE Repository

Q

Welcome to the ELRC-SHARE repository!

Second ELRC Workshop, Malta, 2019/11/01

24



European Language Resource Coordination
Connecting Europe Faculty

How to Register (1/2)

Register


Home Browse Resources Help About Register Login

ELRC-SHARE Repository


Type in your keywords, please...

Welcome to the ELRC-SHARE repository!

Second ELRC Workshop, Malta, 2019/11/01 26



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:

First name:

Last name:

E-mail:

Country:

Organization:

Phone number:

Password:


Password confirmation:

I accept the ELRC Terms of Service for registered users.


Create Account

Second ELRC Workshop, Malta, 2019/11/01

27



How to Contribute Data (1/6)



Contribute Resources

Home
Browse Resources
Contribute Resources
Manage Resources
Help
About
Your Profile, mltoos
Logout

Data Contribution


New Resource

Resource Title*


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

Second ELRC Workshop, Malta, 2019/11/01

28



How to Contribute Data (2/6)



- Fill in the details of the dataset

Resource Title*


Resource short description*

Language(s)


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"

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.)

Second ELRC Workshop, Malta, 2019/11/01
29



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 you wish to upload is larger than 100MB, please contact elrc-share@ilsp.gr

Second ELRC Workshop, Malta, 2019/11/01
30



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
 - ZipltFree
- Windows embedded compression functionality


Second ELRC Workshop, Malta, 2019/11/01

31

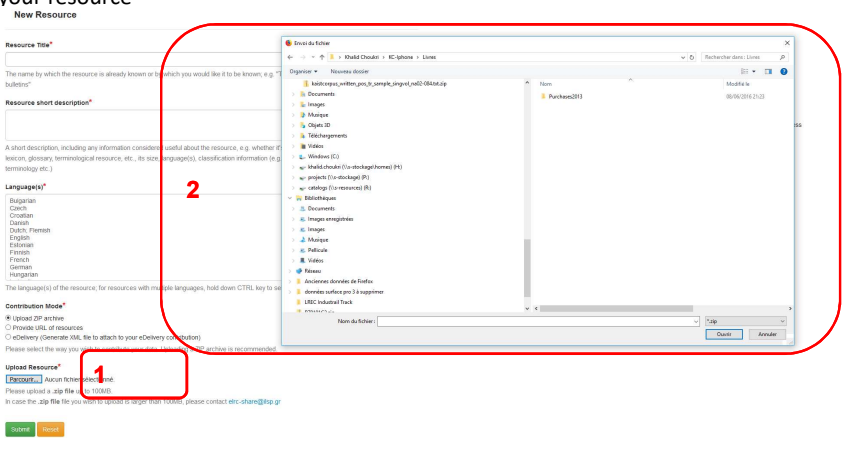


European Language
Resource Coordination
Connecting Europe Facility

How to Contribute Data (4/6)




1. Click on Choose file
2. Locate your resource in your computer
3. Click




Second ELRC Workshop, Malta, 2019/11/01

32



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*


Please provide a URL containing the files you wish to contribute

Second ELRC Workshop, Malta, 2019/11/01


33




How to Contribute Data (6/6)



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



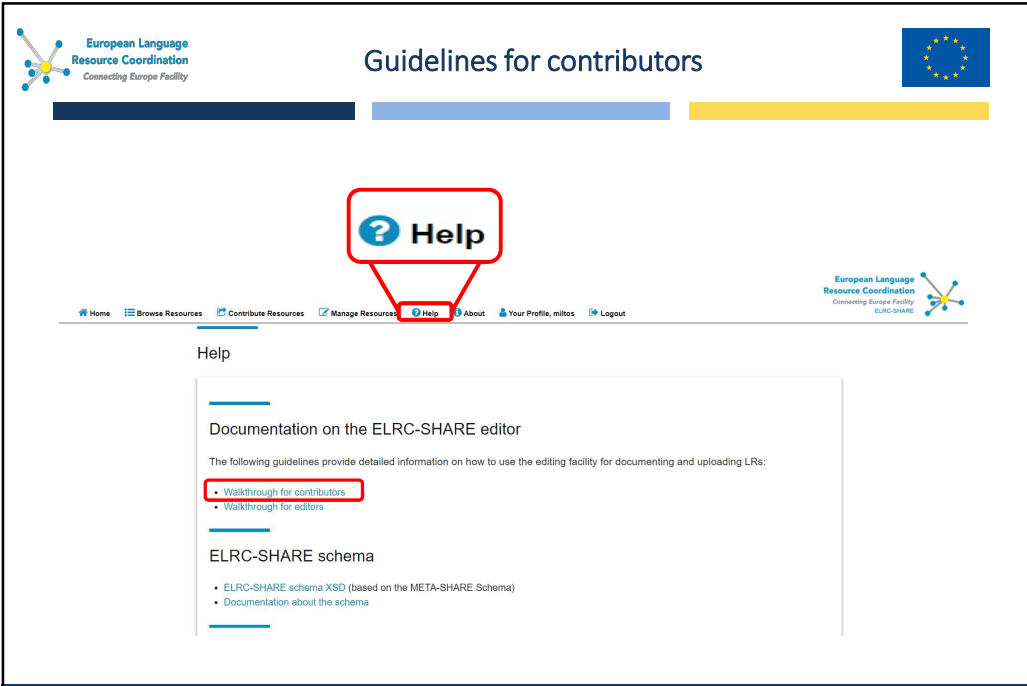
Home
Browse Resources
Contribute Resources
Manage Resources
Help
About
Your Profile, milton
Logout



Data Contribution

Second ELRC Workshop, Malta, 2019/11/01

34



European Language Resource Coordination
Connecting Europe Facility

Guidelines for contributors

Home Browse Resources Contribute Resources Manage Resources **Help** About Your Profile, roles Logout

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

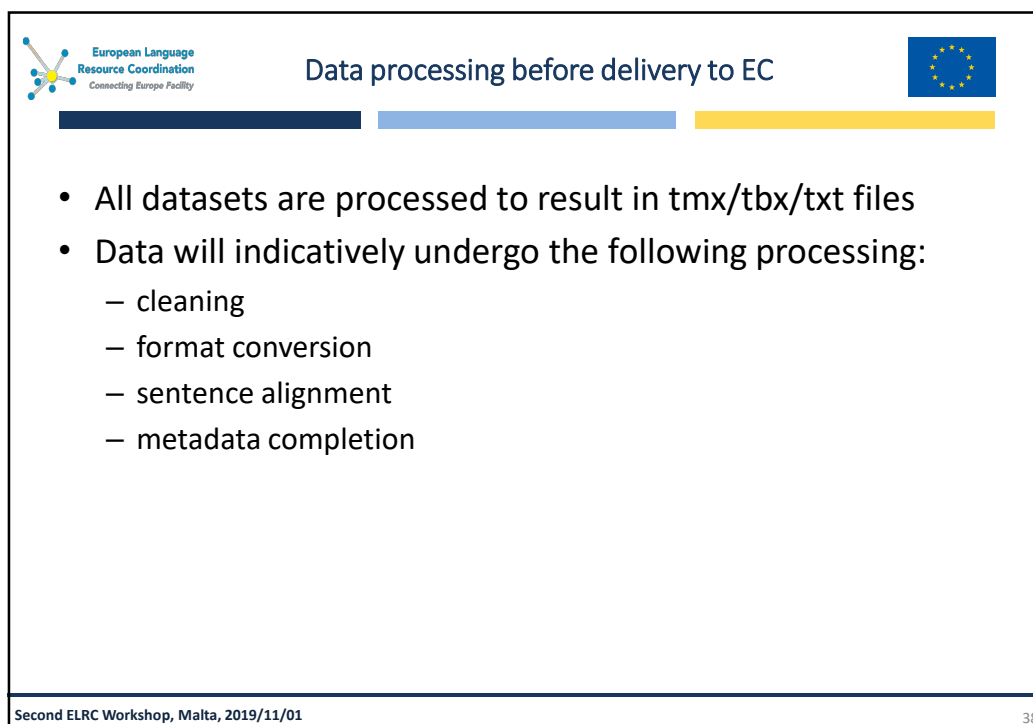
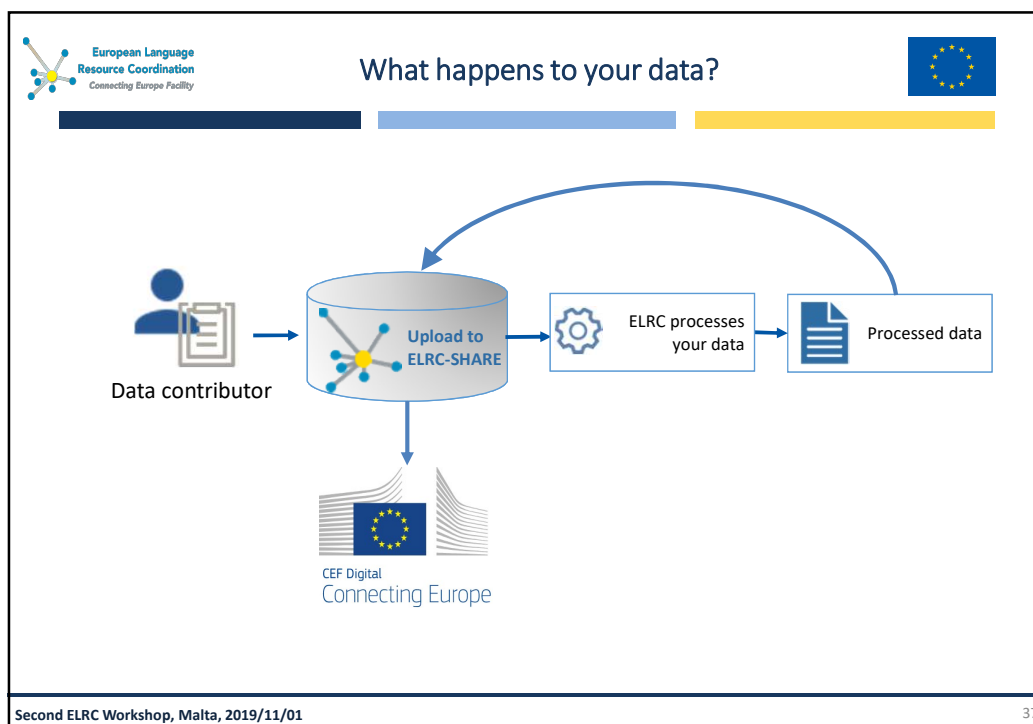
Second ELRC Workshop, Malta, 2019/11/01 35





European Language Resource Coordination
Connecting Europe Facility

What happens next?


Second ELRC Workshop, Malta, 2019/11/01 36




 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, Malta, 2019/11/01 39

 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, Malta, 2019/11/01 40


On-site assistance


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




Second ELRC Workshop, Malta, 2019/11/01 41


On-site assistance



**We will (help) 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, Malta, 2019/11/01 42




Language processing services on-site






Data extraction

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



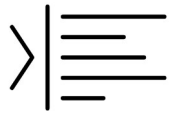
Anonymisation

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!


Second ELRC Workshop, Malta, 2019/11/01

43




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!

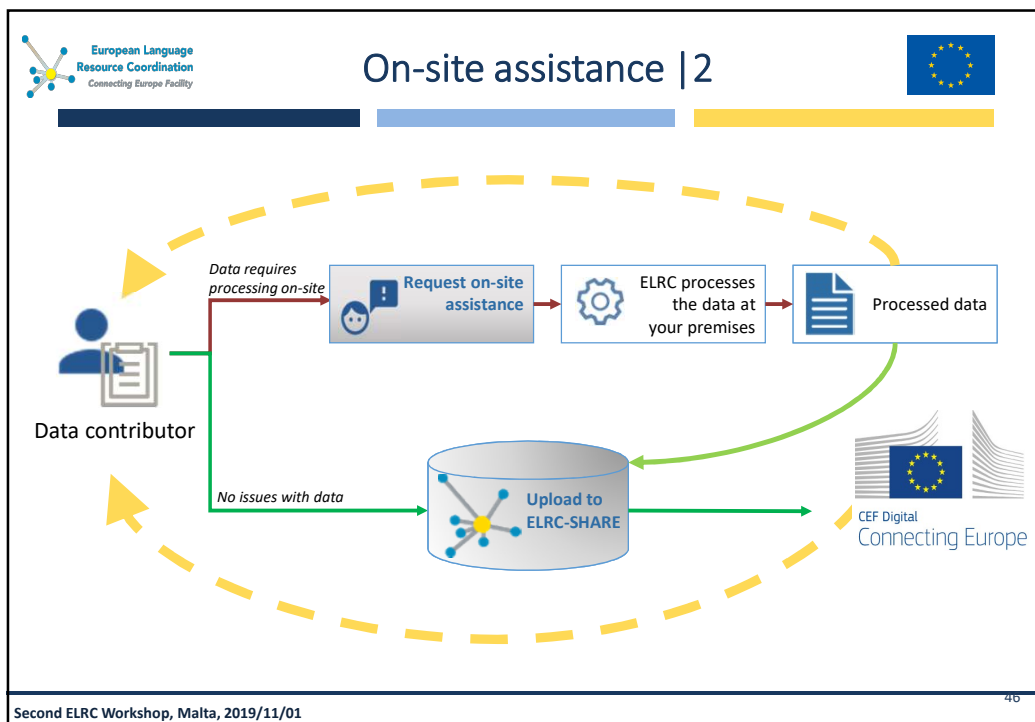
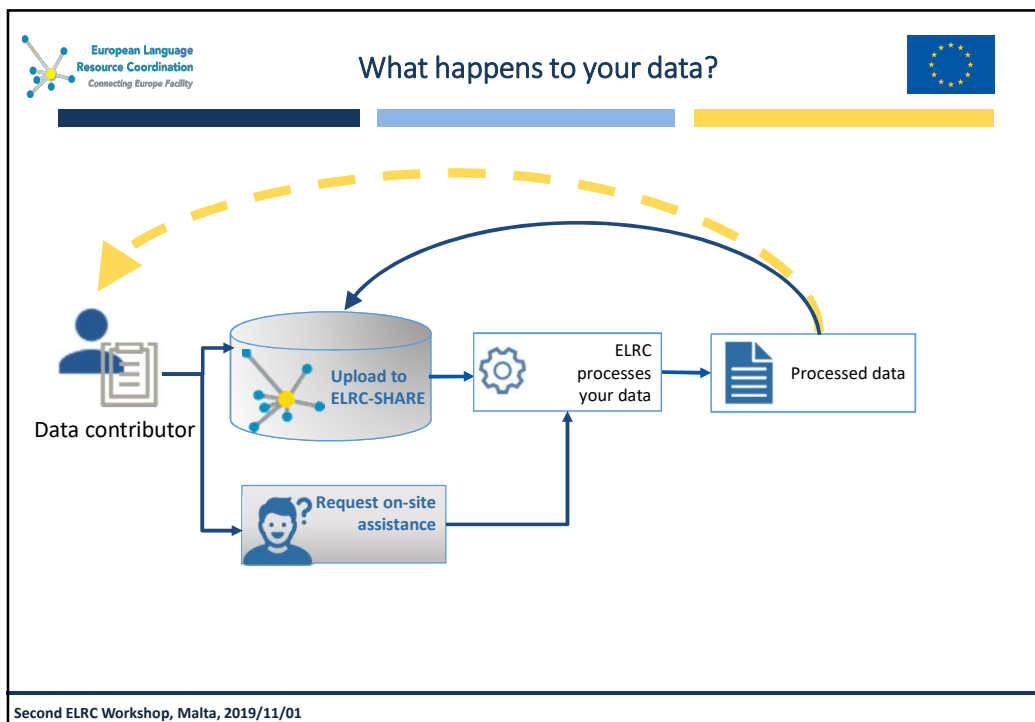


Metadata

Metadata are crucial! We can organise and validate metadata for your team


Second ELRC Workshop, Malta, 2019/11/01

44






European Language
Resource Coordination
Connecting Europe Facility




How to request services and help

Second ELRC Workshop, Malta, 2019/11/01 47



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](#).

Ir-coordination.eu/request-onsite-assistance

First name *

Last name *

Institution *

Country *

Email *

Types of assistance required *

Legal assistance
 Data processing
 Anonymisation
 Other

Description of assistance required

Second ELRC Workshop, Malta, 2019/11/01 48



European Language Resource Coordination
Connecting Europe Facility

ELRC Helpdesk

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

Helpdesk for Language Resources

Helpdesk for Language Resources

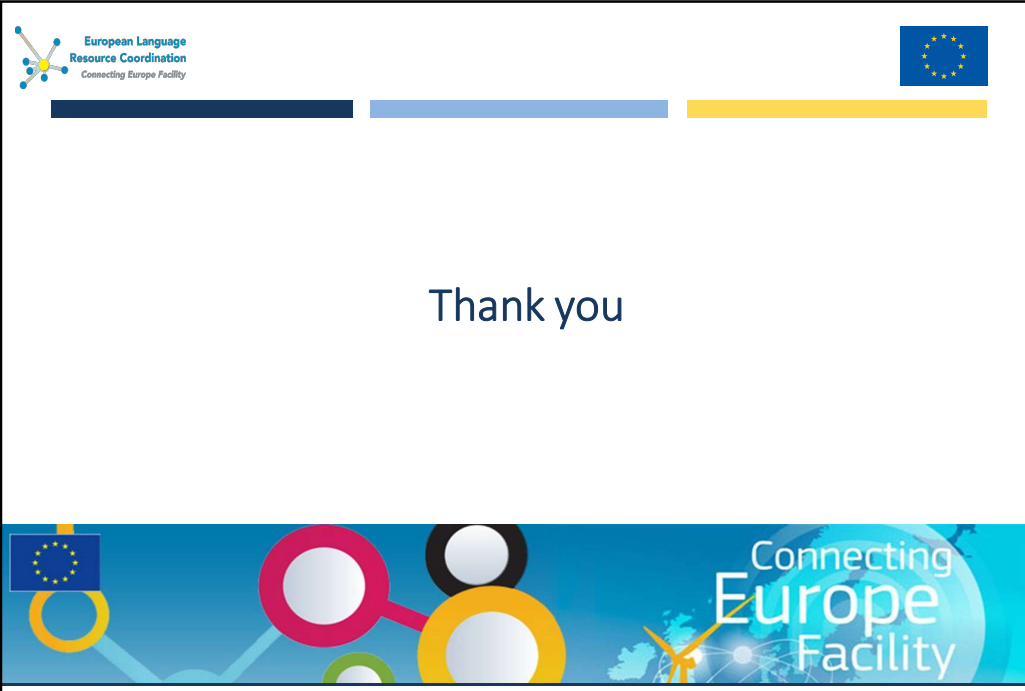
We are happy to answer any questions on the technical or legal aspects related to the use, production, collection, processing, and sharing of language resources.

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-coordination.eu

lr-coordination.eu/helpdesk

Second ELRC Workshop, Malta, 2019/11/01 49



European Language Resource Coordination
Connecting Europe Facility

Thank you

Connecting Europe Facility

Second ELRC Workshop, Malta, 2019/11/01 50