

# Datenaufbereitung und -bereitstellung mit Hilfe des ELRC-SHARE-Repository

Thierry Declerck, Lilli Smal

Multilinguale Technologien,  
Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI)

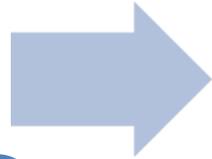






## Daten

- Jede Art von elektronisch gespeicherten **Inhalten**



## (Textuelle) Sprachdaten

- Jede Art von elektronisch gespeicherten **Texten**

## BMI Brochure Civil Protection

Attribution details: German Ministry of the Interior

Bilingual German to English tmx file about Germany's emergency warning system (civil protection)

[← Back](#) [Download](#)

### Distribution

Availability: Available

#### Licences

*Terms for PSI-compliant resources*

*Open Under-PSI*

Conditions: Attribution

#### Distribution Details

Attribution Details: German Ministry of the Interior

### Contact Person

Alexandra Soska 

text 

### Bilingual text corpus

#### Languages

German (de)

English (en)

#### Linguality

Linguality type: Bilingual

#### Text Format

TMX

#### Size

175 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-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: 12/01/2017

Last Updated: 16/01/2017

Metadata Language: English (en)

#### Metadata Creator

Fraser Bowen 

### Relations

Related Resource: BMI Brochure Civil Protection (Processed)

Relation Type: Has Version

## BMI Brochure Civil Protection

Attribution details: German Ministry of the Interior

Bilingual German to English brochure about Germany's emergency warning system (civil protection)

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

```
File01_de.txt
File01_en.txt
File02_de.txt
File02_en.txt
File03_de.txt
File03_en.txt
...
```

Trans.  
Data

Zusätzlich kann sich die interessierte Bevölkerung darüber informieren welche gemeinsamen Anstrengungen Bund, Länder und teilweise auch Kommunen unternehmen, um die Warnung der Bevölkerung in Bedrohungslagen im derzeit gegebenen Rahmen des Möglichen zu gewährleisten.

In addition, everyone interested in this topic can learn more about joint efforts at federal, state and local level to ensure – to the extent currently possible – that the population is warned of imminent threats.

BMI Brochure Civil Protection

germ-eng\_corpus.tmx

Trans.  
Data

Attribution details: German Ministry of the Interior

Bilingual German to English tmx file about Germany's emergency warning system (civil protection)

Back Download

Distribution

Availability

Licences

Terms for PSI-co

Open Under-PSI

Conditions

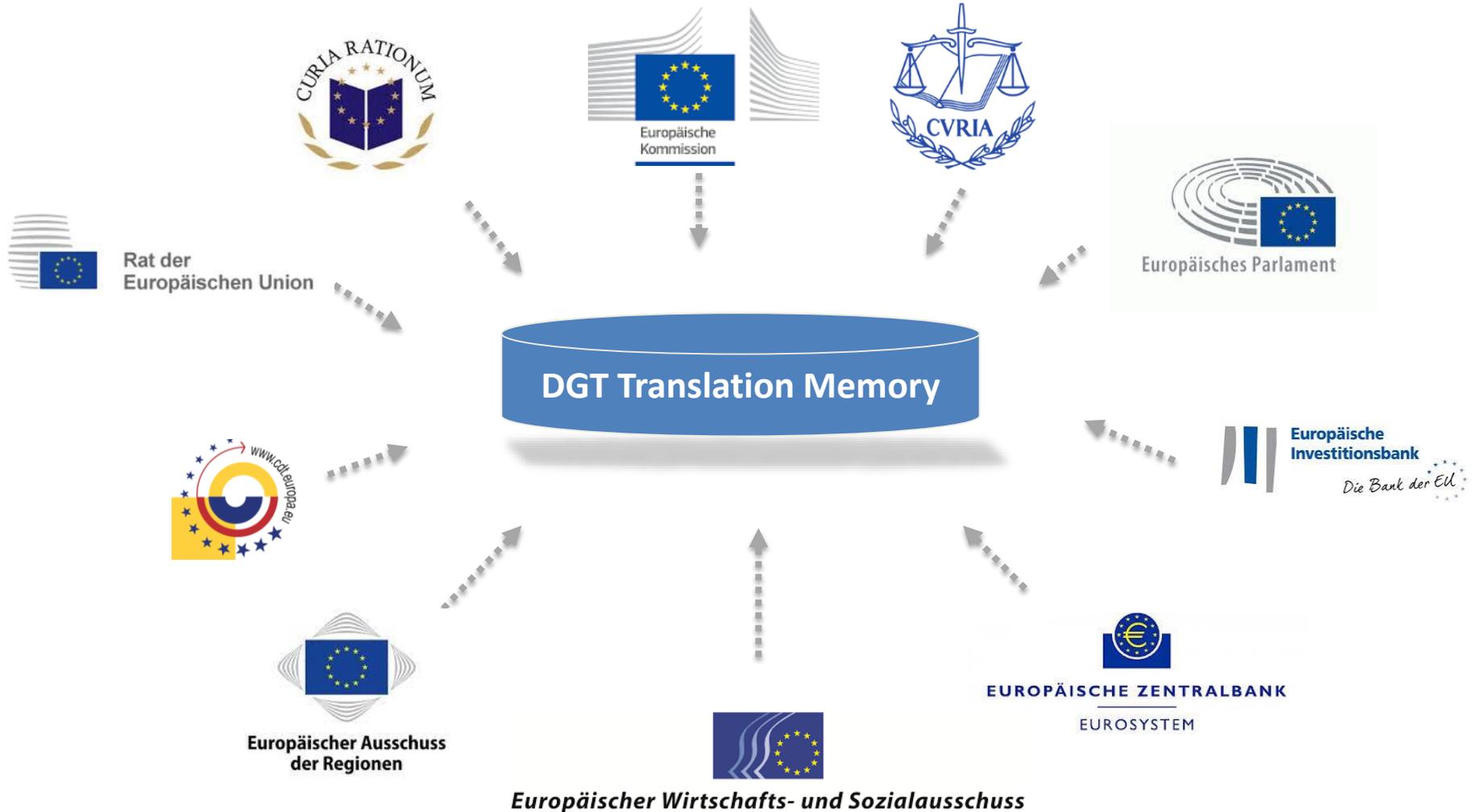
Distribution Data

Attribution Data

Contact Person

Alexandra Sork

```
66 <tu>
67   <tuv xml:lang="de-DE">
68     <seg>Dieses Informationsblatt stellt die grundsätzlichen Fähigkeiten von MoWaS
        vor.</seg>
69   </tuv>
70   <tuv xml:lang="en-GB">
71     <seg>This brochure presents the basic features of MoWaS,</seg>
72   </tuv>
73 </tu>
74 <tu>
75   <tuv xml:lang="de-DE">
76     <seg>Zusätzlich kann sich die interessierte Bevölkerung darüber informieren
        welche gemeinsamen Anstrengungen Bund, Länder und teilweise auch Kommunen
        unternehmen, um die Warnung der Bevölkerung in Bedrohungslagen im derzeit
        gegebenen Rahmen des Möglichen zu gewährleisten.</seg>
77   </tuv>
78   <tuv xml:lang="en-GB">
79     <seg>In addition, everyone interested in this topic can learn more about joint
        efforts at federal, state and local level to ensure - to the extent currently
        possible - that the population is warned of imminent threats.</seg>
80   </tuv>
81 </tu>
```



Diese Daten sind bereits verfügbar,  
ABER  
sie reichen nicht aus .....

- Sprachdaten, die von lokalen öffentlichen Institutionen generiert oder ausgelagert wurden, wie z.B.:
  - Berichte
  - Informationsmaterialien für den externen Gebrauch
  - Presseinformationen
  - Mehrsprachige Web-Inhalte
  - Richtlinien
  - Terminologien
  - Archive
  - Formulare
  - FAQs

- Jeder **elektronisch gespeicherte Text** in einer EU-Sprache (plus norwegisch und isländisch)
- **Texte und deren Übersetzungen** (zwei- oder mehrsprachige parallele Texte)

## Deutscher Text

Das Bundesministerium des Innern, für Bau und Heimat (BMI) betreibt als nichtrechtsfähige Behörde der Bundesrepublik Deutschland unter der Domain [www.bmi.bund.de](http://www.bmi.bund.de) eine Internetseite, auf der es die Öffentlichkeit über seine Tätigkeit informiert und der Öffentlichkeit niedrigschwellig Informationen zur Verfügung stellt.

Personenbezogene Daten werden von uns nur im notwendigen Umfang verarbeitet. Welche Daten zu welchem Zweck und auf welcher Grundlage benötigt und verarbeitet werden, richtet sich maßgeblich nach der Art der Leistung, die von Ihnen in Anspruch genommen wird, beziehungsweise ist abhängig davon, für welchen Zweck diese benötigt werden.

## Translation in English

As an agency of the Federal Republic of Germany without legal capacity, the Federal Ministry of the Interior, Building and Community (BMI) operates a website at the domain [www.bmi.bund.de](http://www.bmi.bund.de) where it informs the public of its activities and makes information easily available to the public.

We process personal data only to the extent necessary. Which data are needed and processed for what purposes and on what basis depends on the type of service you choose and for what purpose the data are needed.

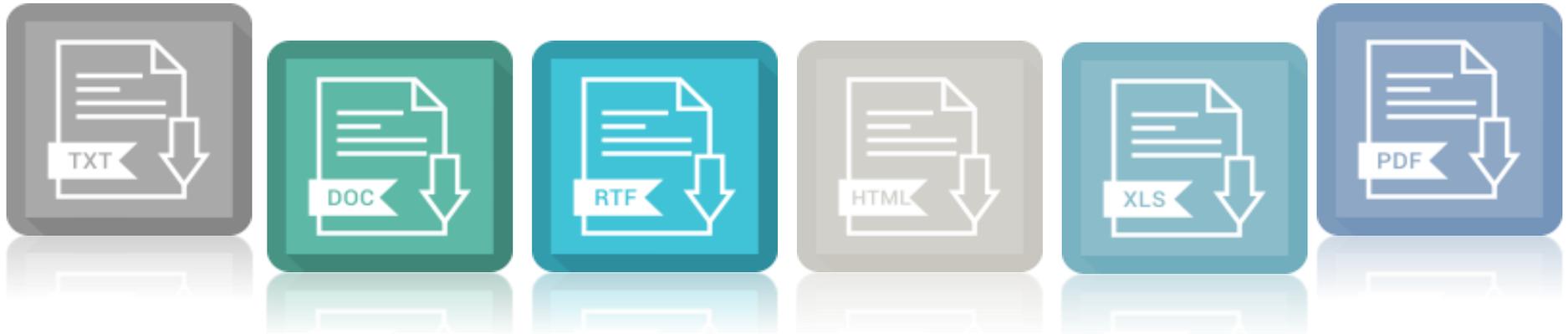
Source: <https://www.bmi.bund.de/EN/service/privacy-policy/privacy-policy-node.html>

- Liste von Termen und deren Übersetzungen, d.h. eine Terminologie.

English	Deutsch
account	Konto
account allocation	Buchung
accountancy	Buchführung(Tätigkeit); Buchhaltung(Abteilung); Rechnungswesen(Funktionsbereich)
accrua	antizipative Abgrenzung
across-the-board	pauschal
capability to pay	Zahlungsfähigkeit
capacity costs	fixe Kosten, Kapazitätskosten
capacity volume variance	Beschäftigungsabweichung
...	...

Quelle: [https://www.fh-dortmund.de/de/fb/9/personen/lehr/schdie/medien/Fachwoerterbuch\\_Controlling\\_Englisch-Deutsch.pdf](https://www.fh-dortmund.de/de/fb/9/personen/lehr/schdie/medien/Fachwoerterbuch_Controlling_Englisch-Deutsch.pdf)

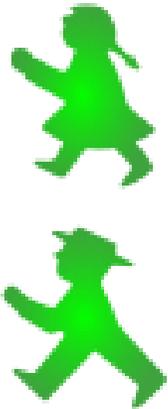
# Welche Daten sind für eTranslation nützlich? gemäß Format | 1



- Grundsätzlich ist jeder Text in maschinenlesbarer Form geeignet
- Einige Formate sind jedoch "MÜ-fähiger" als andere, da sie weniger manuelle oder automatische Vorverarbeitung benötigen
- Mehr Vorverarbeitung führt zu mehr Fehlern im Output und macht diese Formate somit weniger nützlich für eTranslation

- Die folgenden Formate sind besonders nützlich (in absteigender Reihenfolge):
  - Für zweisprachige/mehrsprachige parallele Texte
    1. Übersetzungsspeicher/Translation Memories (.tmx)
    2. XML-Übersetzungsdateien (.xliff)
    3. Klartext (.txt, .csv)
    4. Tabellenkalkulationen (zB. Xlsx)
  - Für Terminologien
    1. TermBase eXchange (.tbx)
    2. Klartext (.txt, .csv)
    3. Tabellenkalkulationen (zB. Xlsx)
  - Für einsprachige Texte
    1. Klartext (.txt, .csv)

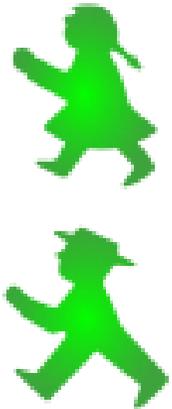
## (Gewünschte) Dateiformate von parallelen Texten



-  Dateinamen01\_DE.txt
-  Dateinamen01\_EN.txt
-  Dateinamen02\_DE.txt
-  Dateinamen02\_EN.txt
-  Dateinamen03\_DE.txt
-  Dateinamen03\_EN.txt
-  Dateinamen04\_DE.txt
-  Dateinamen04\_EN.txt
-  Dateinamen05\_DE.txt
-  Dateinamen05\_EN.txt

...

Verwendung identischer Dateinamen für  
jedes Dokumentenpaar (Quelle -  
Übersetzung)



-  Dateinamen01\_DE.txt
-  Dateinamen01\_EN.txt
-  Dateinamen02\_DE.txt
-  Dateinamen02\_EN.txt
-  Dateinamen03\_DE.txt
-  Dateinamen03\_EN.txt
-  Dateinamen04\_DE.txt
-  Dateinamen04\_EN.txt
-  Dateinamen05\_DE.txt
-  Dateinamen05\_EN.txt

...

**Sprachbezeichnungen in  
den Dateinamen  
aufnehmen**



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.



Dies ist ein Absatz auf Deutsch. Dies ist ein Absatz auf Deutsch. Dies ist ein Absatz auf  
Deutsch. Dies ist ein Absatz auf Deutsch. Dies ist ein Absatz auf Deutsch. Dies ist ein Absatz  
auf Deutsch. Dies ist ein Absatz auf Deutsch. Dies ist ein Absatz auf Deutsch. Dies ist ein  
Absatz auf Deutsch. Dies ist ein Absatz auf Deutsch. Dies ist ein Absatz auf Deutsch. Dies ist  
ein Absatz auf Deutsch.

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

Dies ist ein zweiter Absatz auf Deutsch. Dies ist ein zweiter Absatz auf Deutsch. Dies ist ein  
zweiter Absatz auf Deutsch. Dies ist ein zweiter Absatz auf Deutsch. Dies ist ein zweiter  
Absatz auf Deutsch. Dies ist ein zweiter Absatz auf Deutsch. Dies ist ein zweiter Absatz auf  
Deutsch. Dies ist ein zweiter Absatz auf Deutsch. Dies ist ein zweiter Absatz auf Deutsch. Dies  
ist ein zweiter Absatz auf Deutsch.



- Zur Erinnerung: Ein Datensatz ist eine Sammlung von Daten, die nach **bestimmten Kriterien gruppiert** sind.
- Für die Weiterentwicklung und Anpassung des eTranslation Übersetzungsservice sind zwei Kriterien entscheidend:
  - **Sprachen(n)**: Jede Sammlung wird durch die Sprache oder Sprachpaare ihrer Daten definiert, z.B.
    - *Textsammlung in Englisch – Deutsch*
    - *Dokumente in Englisch – Norwegisch - Finnisch*
  - **Domäne**: Jede Sammlung gehört idealerweise zu einer einzelnen Domäne, z.B.
    - *Textsammlung in Englisch – Deutsch im Kulturbereich*
    - *Dokumente der Sozialversicherung in Englisch – Norwegisch - Finnisch*

CEF DSI	Domäne
Online-Streitbeilegung (ODR)	Verbraucherrechte, Verbraucherbeschwerde
Elektronischer Austausch von Sozialversicherungsdaten (EESSI)	Sozialversicherung, Sozialversicherungssystem
Elektronischen Auftragsvergabe (eProcurement)	Vergabe öffentlicher Aufträge, vertragliche Vereinbarungen
Europäisches Justizportal (eJustice)	Justiz, Recht
eHealth	Gesundheit, Medizin
System zur Verknüpfung von Unternehmensregistern (Business Registers Interconnection System – BRIS)	Handel, Geschäftsleben, Unternehmen, Markt
Sichere Nutzung des Internets (Safer Internet)	
Cybersicherheit	
Public Open Data	
Europeana	Kultur

# Wie Sie Ihre Daten im ELRC-SHARE Repository hochladen können - Eine Schritt-für-Schritt-Anleitung

- Klicken Sie auf der ELRC-Webseite unter “Resources” auf die Schaltfläche “Language resource submission”

Oder

- Geben Sie die URL-Adresse ein: [elrc-share.eu](https://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

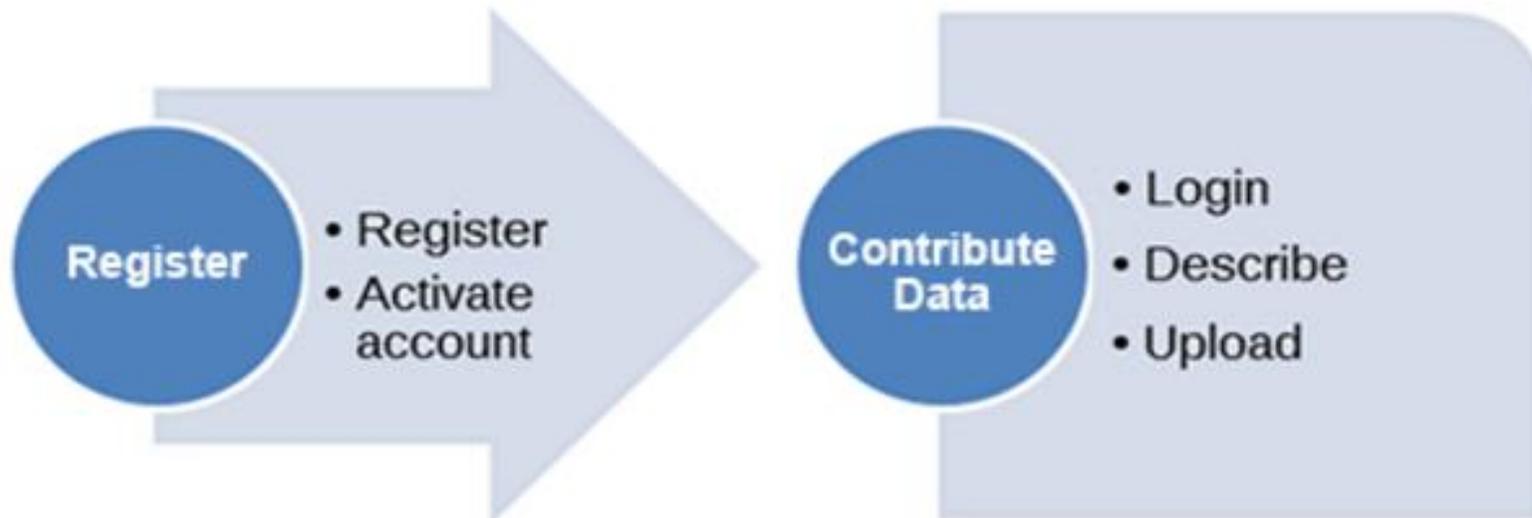
Language resource submission ▶

## ELRC-SHARE Repository



Welcome to the ELRC-SHARE repository!

# Wie kann man Daten beisteuern?





 Register

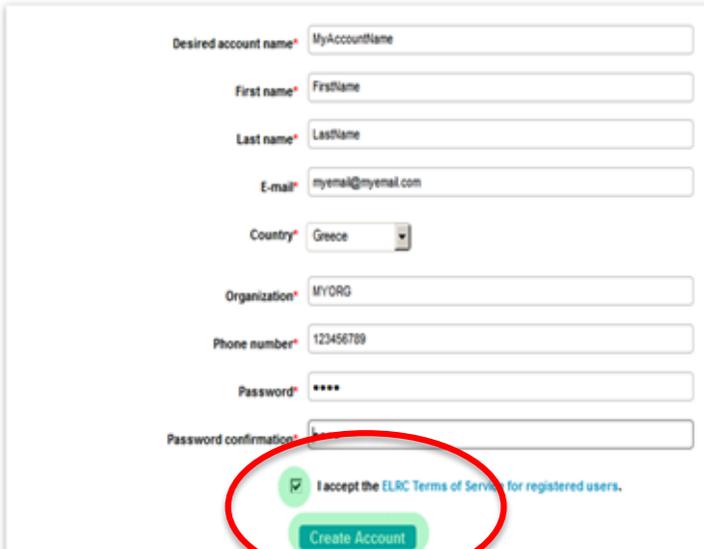
## ELRC-SHARE Repository



Welcome to the ELRC-SHARE repository!

- Geben Sie die erforderlichen Informationen ein
- Lesen Sie die *Allgemeinen Geschäftsbedingungen* und klicken Sie auf *Akzeptieren*, wenn Sie damit einverstanden sind.
- Klicken Sie auf die Schaltfläche *Create Account*
- Aktivieren Sie Ihr Konto gemäß den Ihnen per E-Mail zugesandten Richtlinien.

\*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



## 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"

- Beschreiben Sie Ihre Sprachdaten.

**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

- Drei Modi für die Bereitstellung Ihrer Daten

## 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

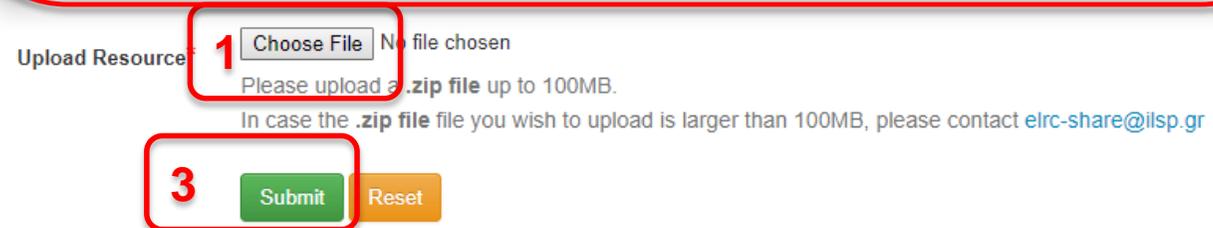
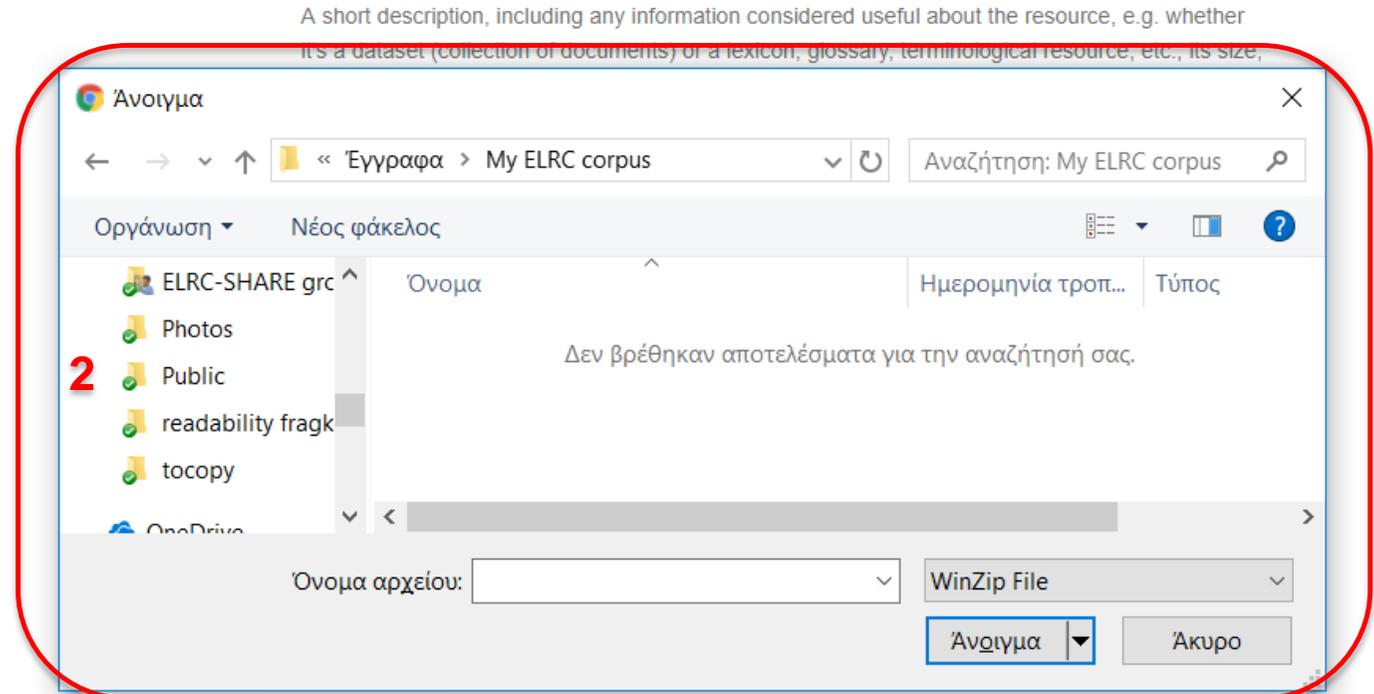
Please upload a **.zip file** up to 100MB.

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

Submit

Reset

1. Klicken Sie auf “Choose file”
2. Suchen Sie Ihre Ressource auf Ihrer Festplatte.
3. Klicken Sie auf “Submit”



- Alternativ können Sie auch eine URL angeben (Verzeichnisauflistung)

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.

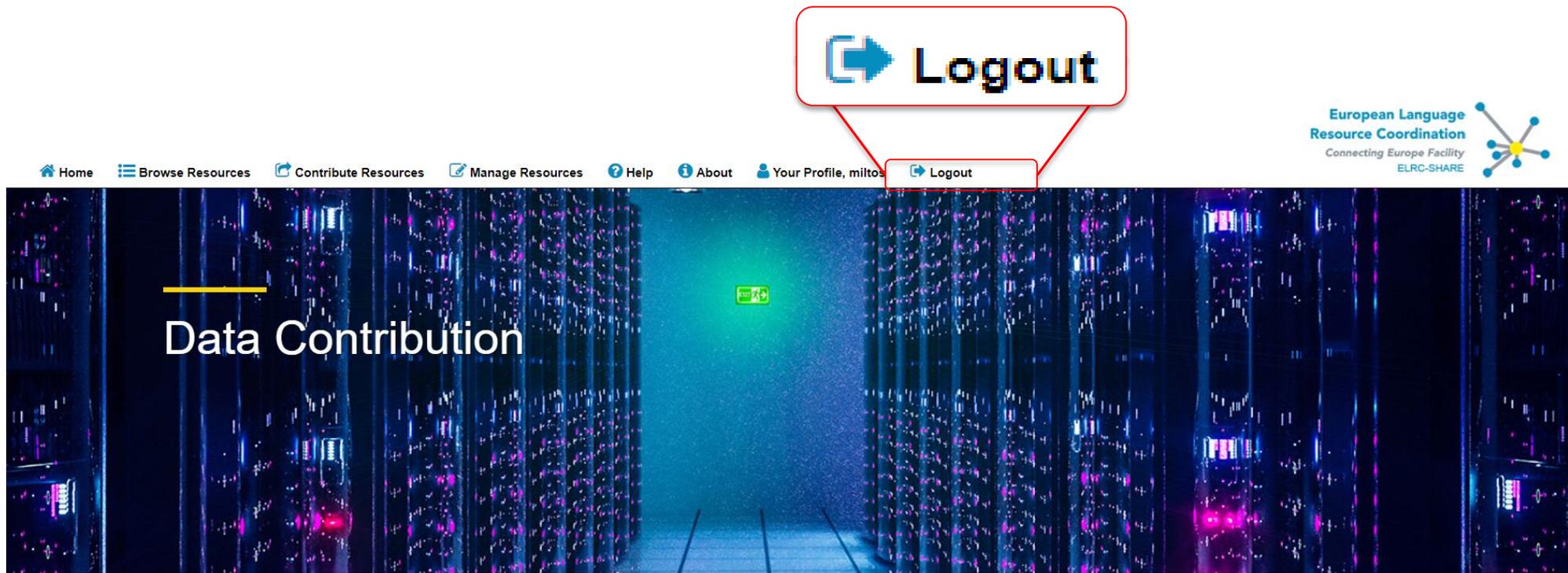
Resource URL\*

www

Please provide a URL containing the files you wish to contribute

Submit Reset

- Wiederholen Sie den Vorgang, wenn Sie eine andere Ressource hochladen wollen, oder loggen Sie sich aus.



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'. Below the navigation bar is a large image of a server room with the text 'Data Contribution' overlaid in white.



## Help

### Documentation on the ELRC-SHARE editor

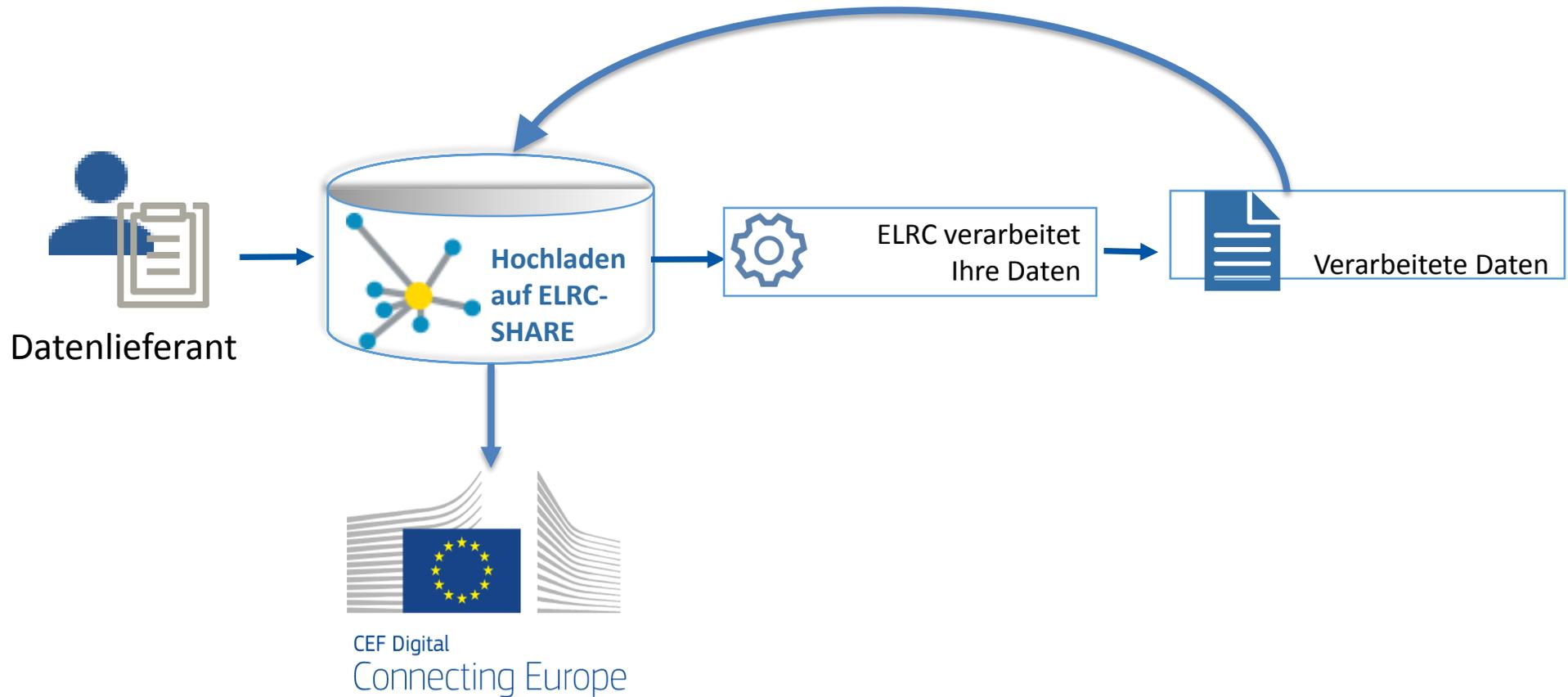
The following guidelines provide detailed information on how to use the editing facility for documenting and uploading LR:

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

### ELRC-SHARE schema

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

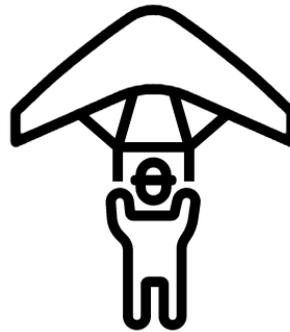
Was passiert als nächstes?



- Alle Datensätze werden so verarbeitet, dass sie zu tmx/tbx/txt-Dateien umgewandelt werden.
- Die Daten werden der folgenden Verarbeitung unterzogen:
  - Reinigung
  - Formatumwandlung
  - Alignierung von Sätzen
  - Vervollständigung der Metadaten

Diese Dienstleistungen werden auch direkt vor Ort und kostenlos für alle Datenspender angeboten.





Unser Expertenteam kann Sie bei Bedarf direkt vor Ort unterstützen, um Hilfe bei der Datenspende zu leisten

Die Unterstützung erfolgt in enger Zusammenarbeit mit einem breiten Netzwerk von Sprachexperten.



Wir verarbeiten Ihre Sprachdaten und geben die bereinigten Daten wieder an Sie zurück. Wir können auch helfen, Ihre Datenmanagementprozesse zu optimieren.  
Sprechen Sie uns einfach an!



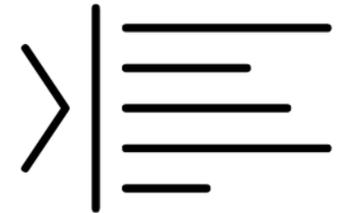
Wenn Ihre Daten in Archiven und Datenbanken “versteckt” sind, können wir Ihnen helfen, sie zu extrahieren.



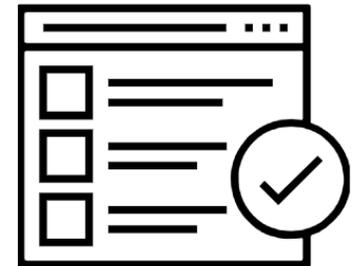
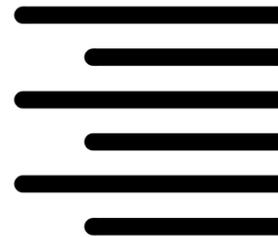
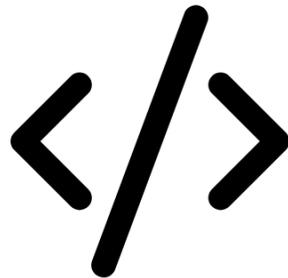
Enthalten Ihre Daten private Informationen? Wir können bei der Anonymisierung helfen.



Wenn Ihre Daten “unsauber” sind, werden wir sie bereinigen.



Müssen Sie DOCX in XML oder PDF in WORD umformatieren? Wir übernehmen das gerne für Sie!

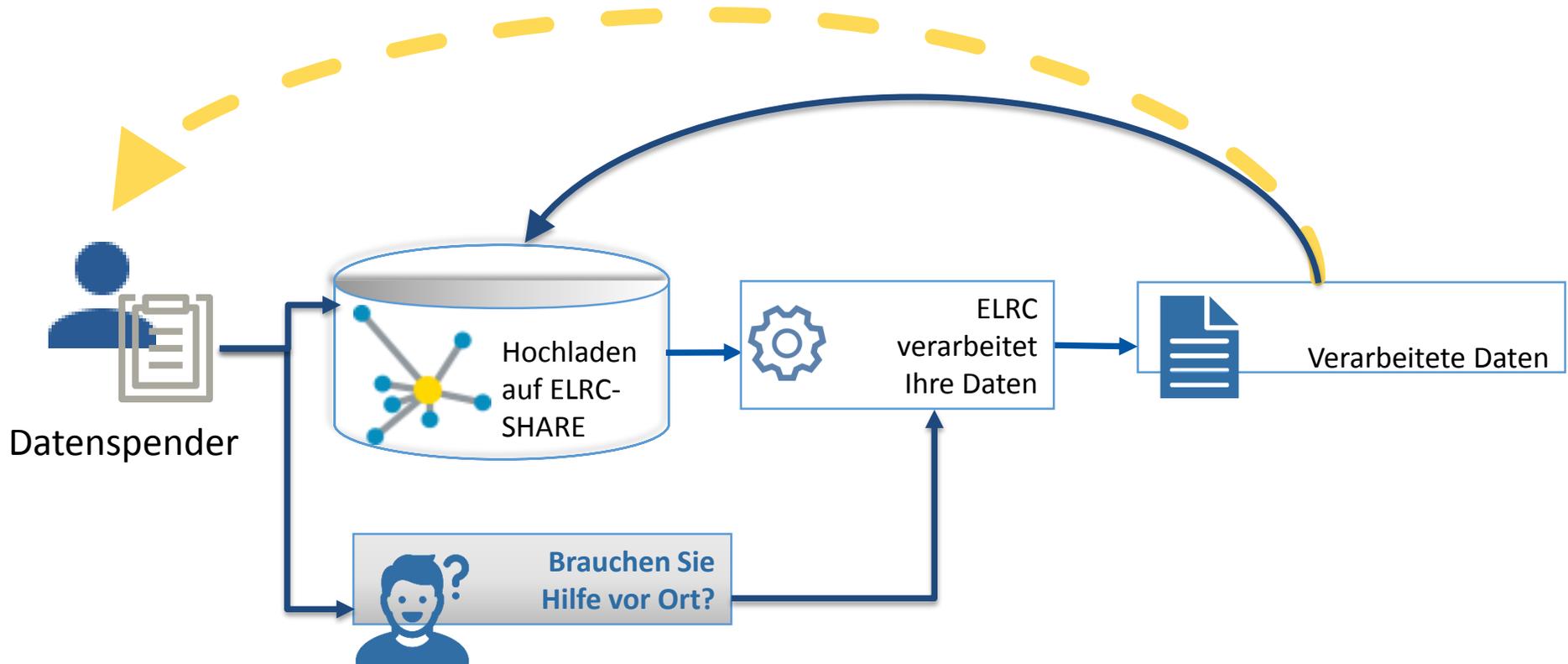


Wenn Ihre Daten nicht in den richtigen Formate konvertiert sind, können wir Ihnen helfen, sie zu konvertieren.

Enthalten Ihre Daten nicht benötigte Tags? Wir können Ihnen helfen, sie zu entfernen!

Sind die Übersetzungen nicht aligniert? Wir erledigen das für Sie mit unseren Werkzeugen!

Metadaten sind entscheidend! Wir können Metadaten für Ihr Team organisieren und validieren



*Einige Daten können auch aufs Open Data Portal hochgeladen werden, und somit anderartige Verwendung finden!*

Wie können Sie Hilfe vor Ort anfordern?

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



## 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	<a href="mailto:help@lr-coordantion.eu">help@lr-coordantion.eu</a>



Vielen Dank für Ihre Aufmerksamkeit!



- 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](http://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
- <https://de.wikipedia.org/wiki/Datei:Ampelmann.svg> (public domain)
- <https://upload.wikimedia.org/wikipedia/commons/0/0b/Ampelfrau.svg> (public domain)