European Language Resource Coordination
Connecting Europe Facility

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

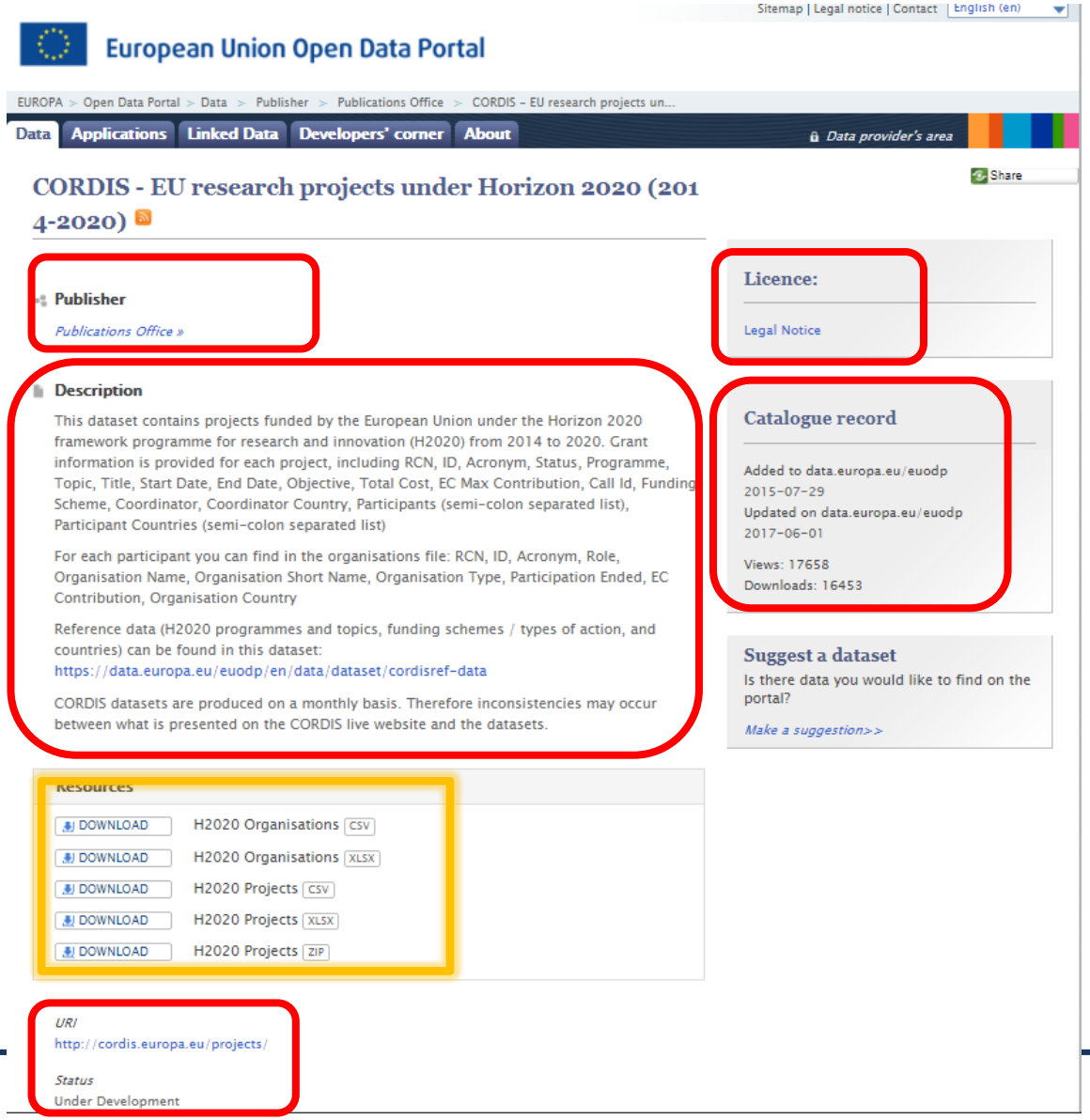
Thierry Declerck, Lilli Smal
Multilinguale Technologien,
Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI)

Der Datenbegriff

Der Datenbegriff

Grundlegende Konzepte:

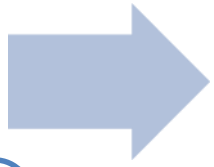
- **Daten**: jede Art von gespeicherten Inhalten
- **Datensatz (oder Ressource):** Die Sammlung einer oder mehrerer Datendateien, die nach bestimmten **Kriterien gruppiert** sind
- **Metadaten:** *Daten über die Daten*, d.h. Beschreibung eines Datensatzes mit Eigenschaften (z.B. Titel, Verlag, Beschreibung des Inhalts und URL)



Der Begriff “Sprachdaten”

Daten

- Jede Art von elektronisch gespeicherten Inhalten

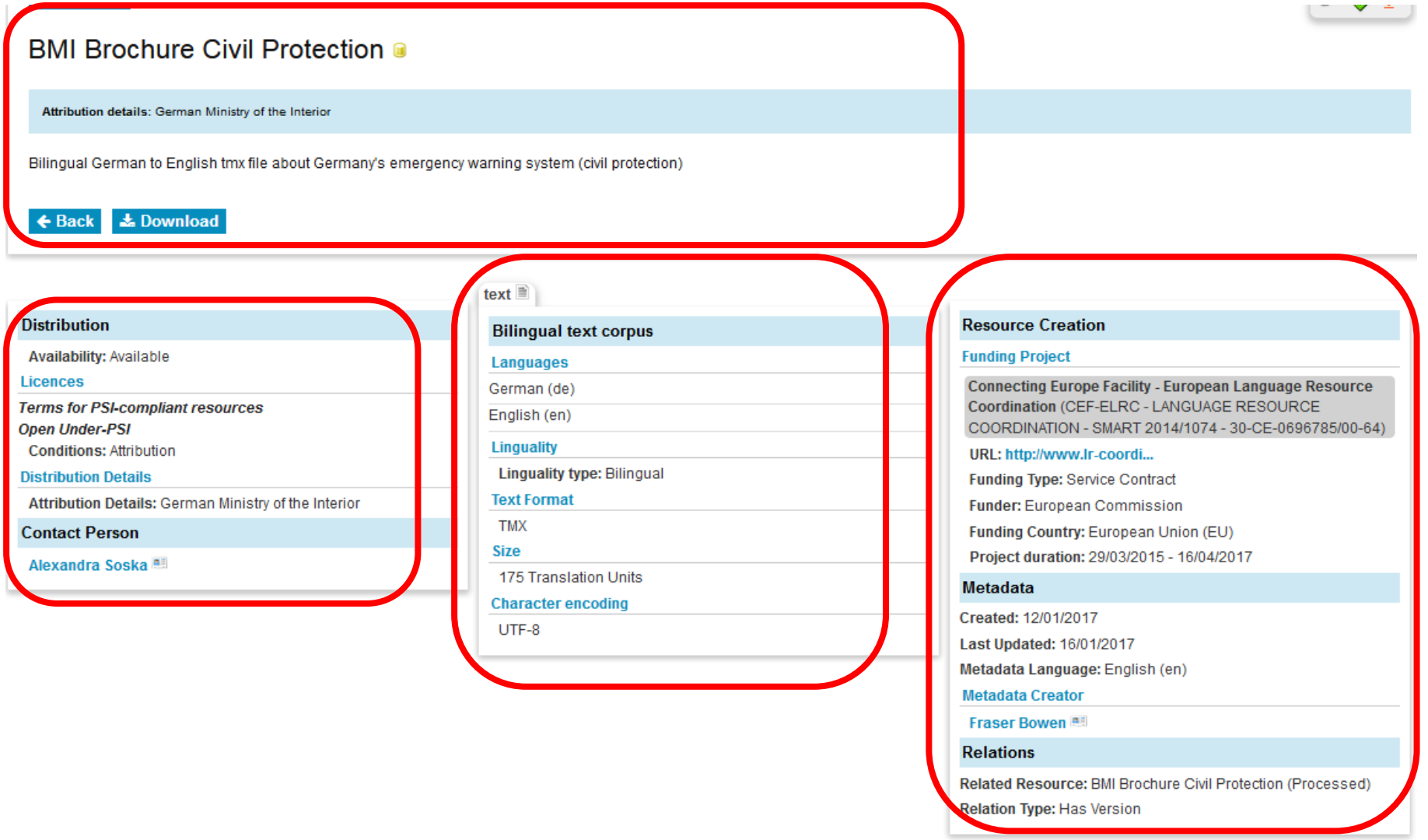


(Textuelle) Sprachdaten

- Jede Art von elektronisch gespeicherten Texten

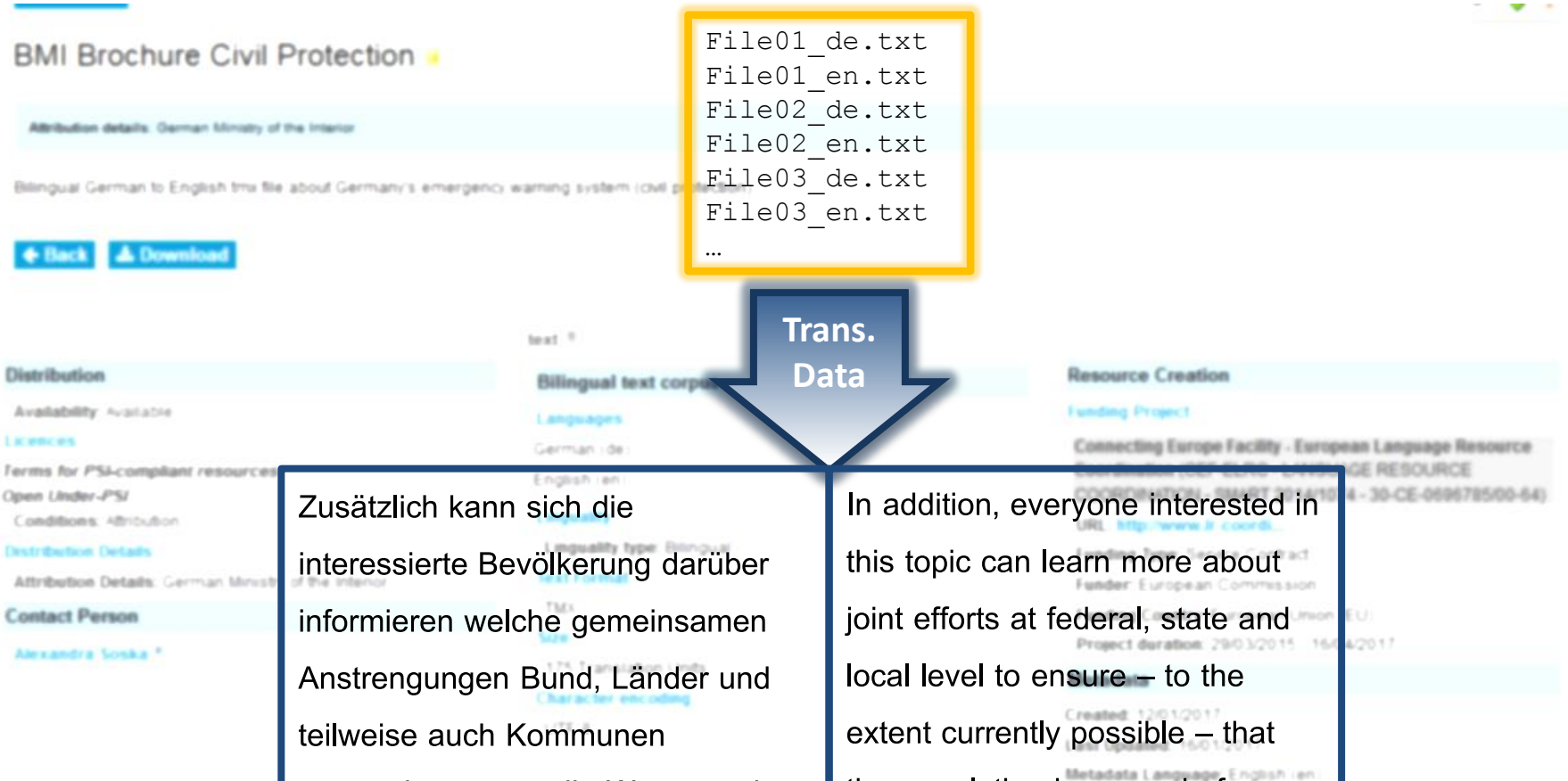
European Language Resource Coordination
Connecting Europe Facility

Der Datenbegriff im Kontext von eTranslation



European Language Resource Coordination
Connecting Europe Facility

Der Datenbegriff im Kontext von eTranslation




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

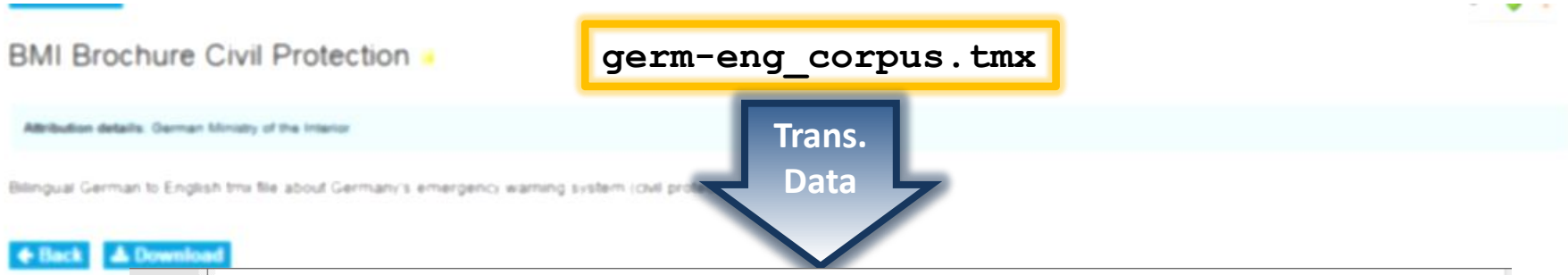
Trans. Data

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.

European Language Resource Coordination
Connecting Europe Facility

Der Datenbegriff im Kontext von eTranslation



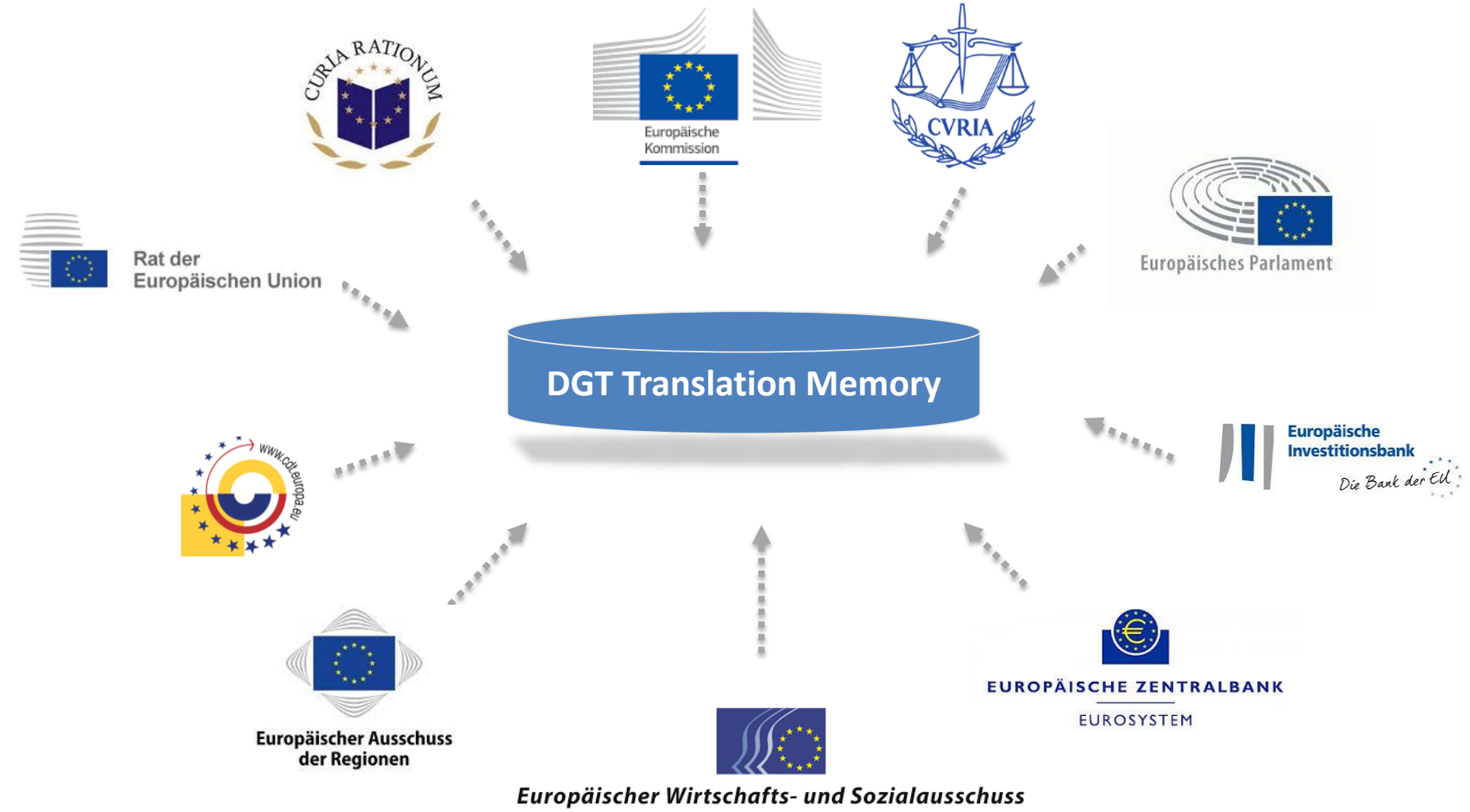
germ-eng_corpus.tmx

Trans. Data

```
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>
```

European Language Resource Coordination
Connecting Europe Facility

Von eTranslation verwendete Daten



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

Was wird von eTranslation zusätzlich benötigt?

- 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

Welche Daten sind für eTranslation nützlich? gemäß Typ |1

- 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 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 welchen 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.

Quelle: https://www.bmi.bund.de/DE/service/datenschutz/datenschutz_node.html

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 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

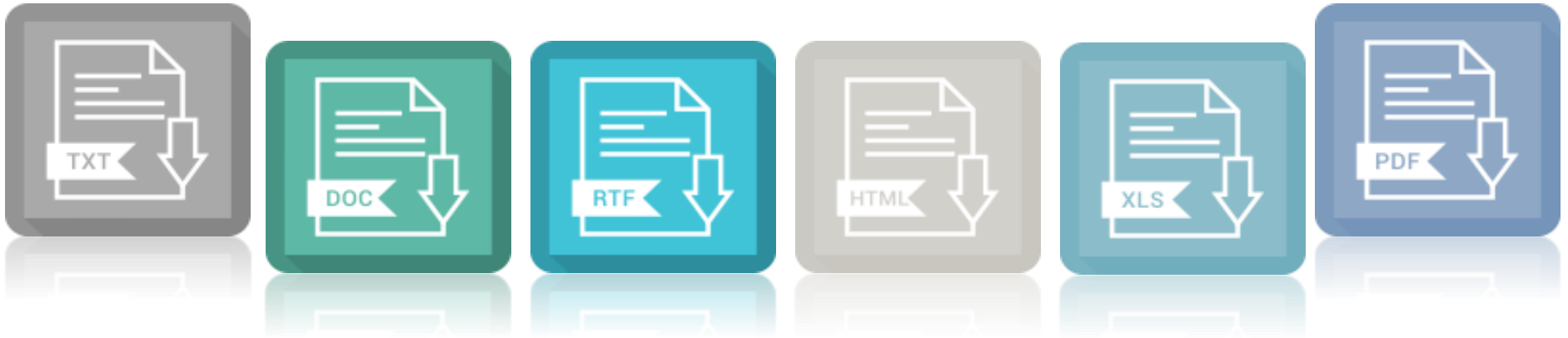
Welche Daten sind für eTranslation nützlich? gemäß Typ |2

- 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

Welche Daten sind für eTranslation nützlich? gemäß Format |1



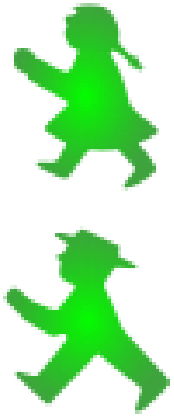
- Grundsätzlich ist jeder Text in maschinenlesbarer Form geeignet
- Einige Formate sind jedoch "MÜ-fähiger" als andere, da sie weniger manuelle oder automatische Vorverarbeitung benötigen
- Mehr Vorverarbeitung führt zu mehr Fehlern im Output und macht diese Formate somit weniger nützlich für eTranslation

Welche Daten sind für eTranslation nützlich? gemäß Format|2

- 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

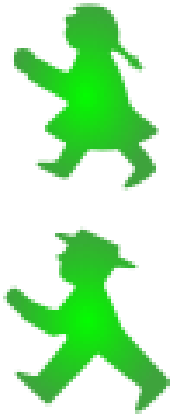
Aufbereitung Ihrer Daten |1



- 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)

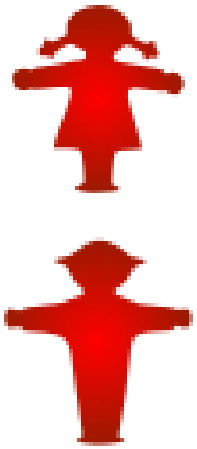
Aufbereitung Ihrer Daten |2



- 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

Aufbereitung Ihrer Daten |3



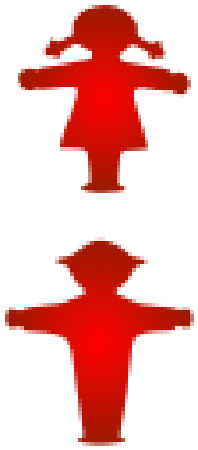
This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph 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. 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. Dies ist ein zweiter Absatz auf Deutsch. Dies ist ein zweiter Absatz auf Deutsch.

Aufbereitung Ihrer Daten |4



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

Kriterien für die Gruppierung Ihrer Daten

- 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*

Bevorzugte Domänen

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

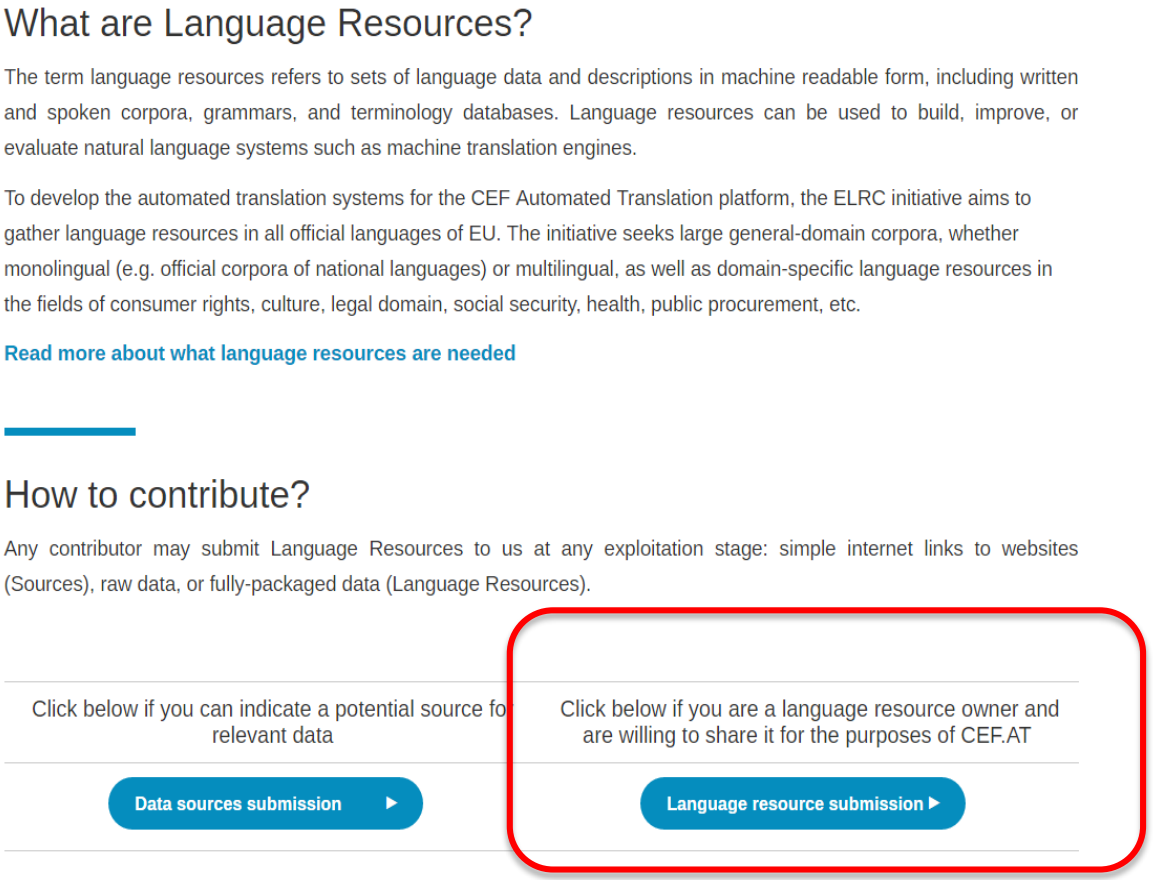
Wie Sie Ihre Daten im ELRC-SHARE Repository hochladen können - Eine Schritt-für-Schritt-Anleitung

So finden Sie das ELRC-SHARE Repository

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

Oder

- Geben Sie die URL-Adresse ein: **elrc-share.eu**

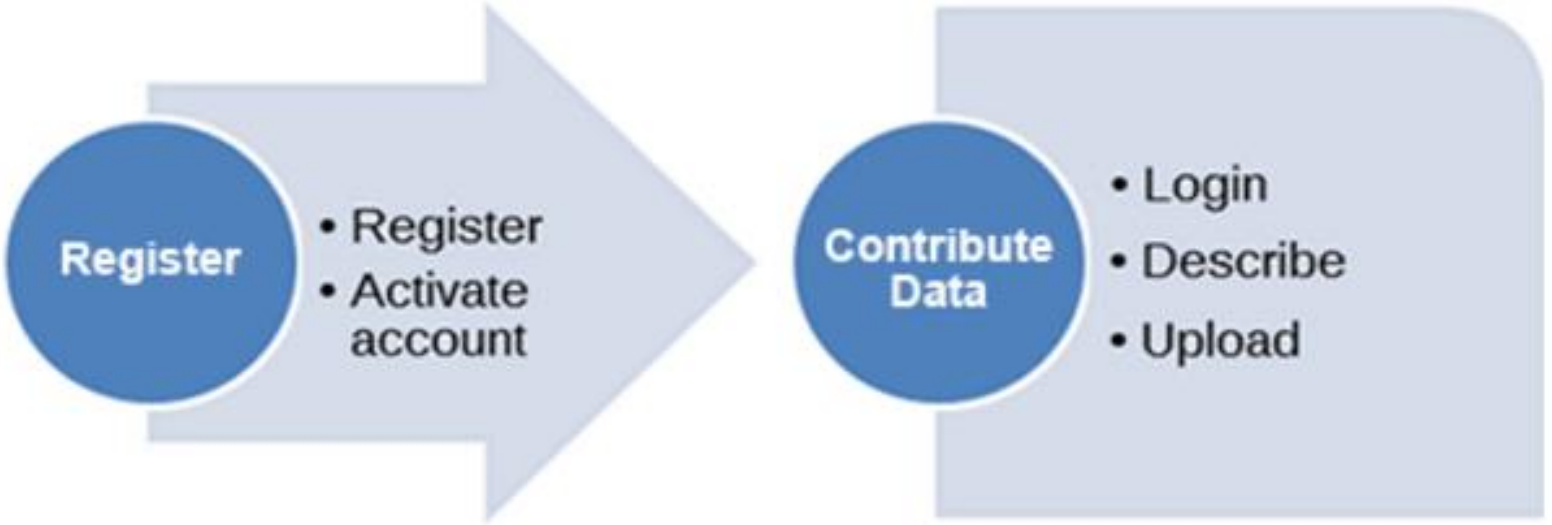


European Language Resource Coordination
Connecting Europe Facility

Das ELRC-SHARE Repository

European Language Resource Coordination
Connecting Europe Facility

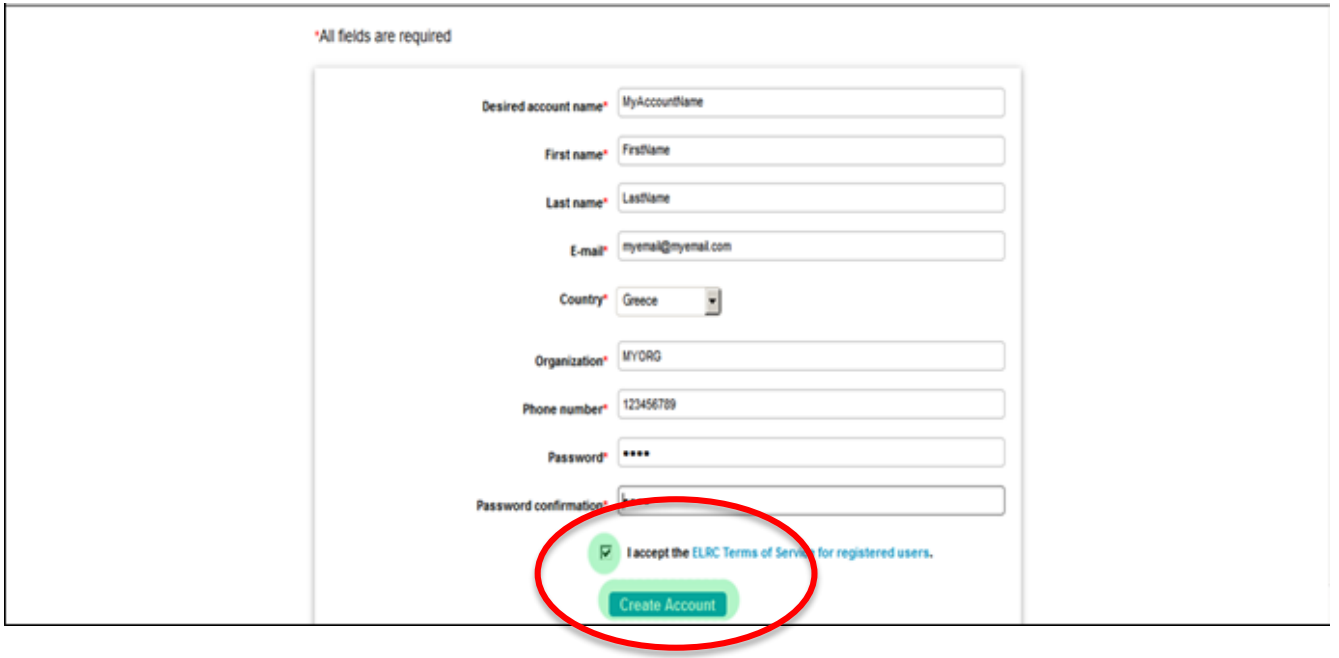
Wie kann man Daten beisteuern?



Registrieren (1/2)

Registrieren (2/2)

- 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.



Wie kann man Daten beisteuern? (1/6)

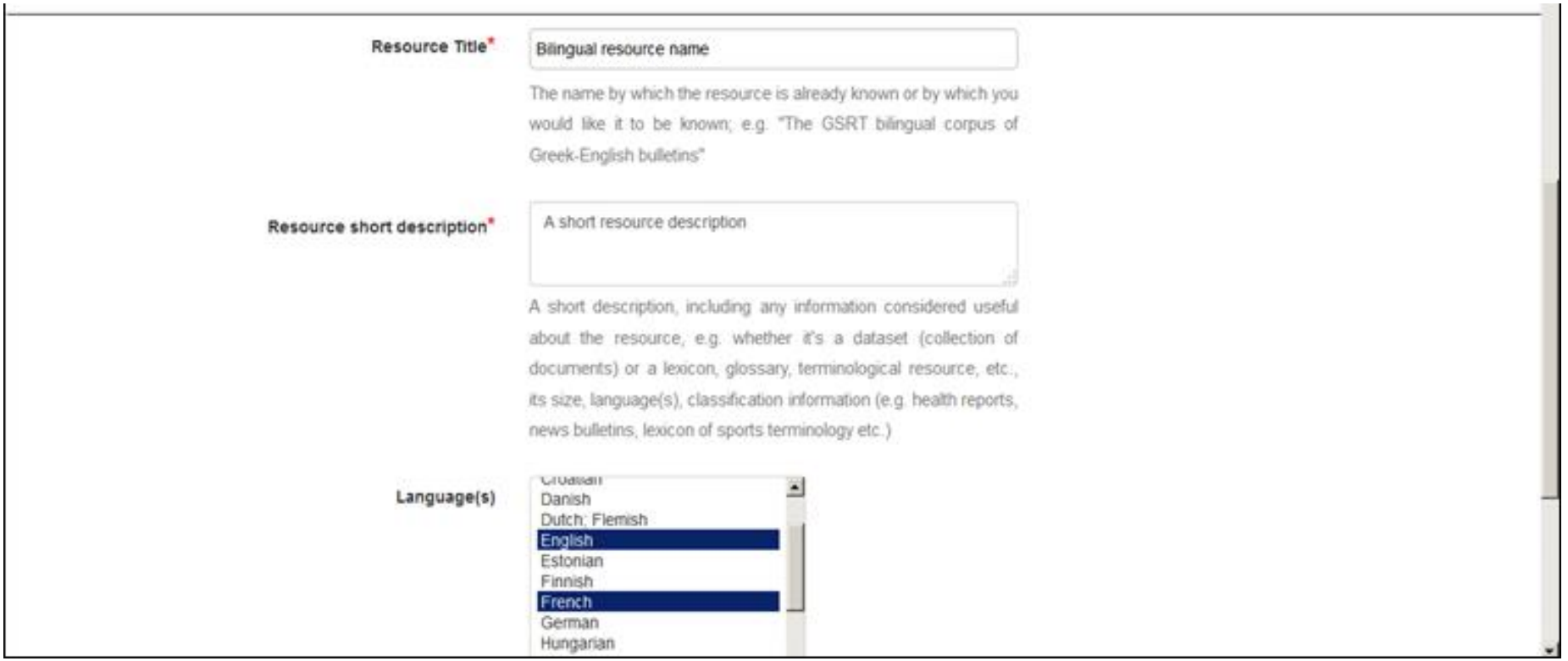
New Resource

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"

Wie kann man Daten beisteuern? (2/6)

- Beschreiben Sie Ihre Sprachdaten.



Wie kann man Daten beisteuern? (3/6)

- 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

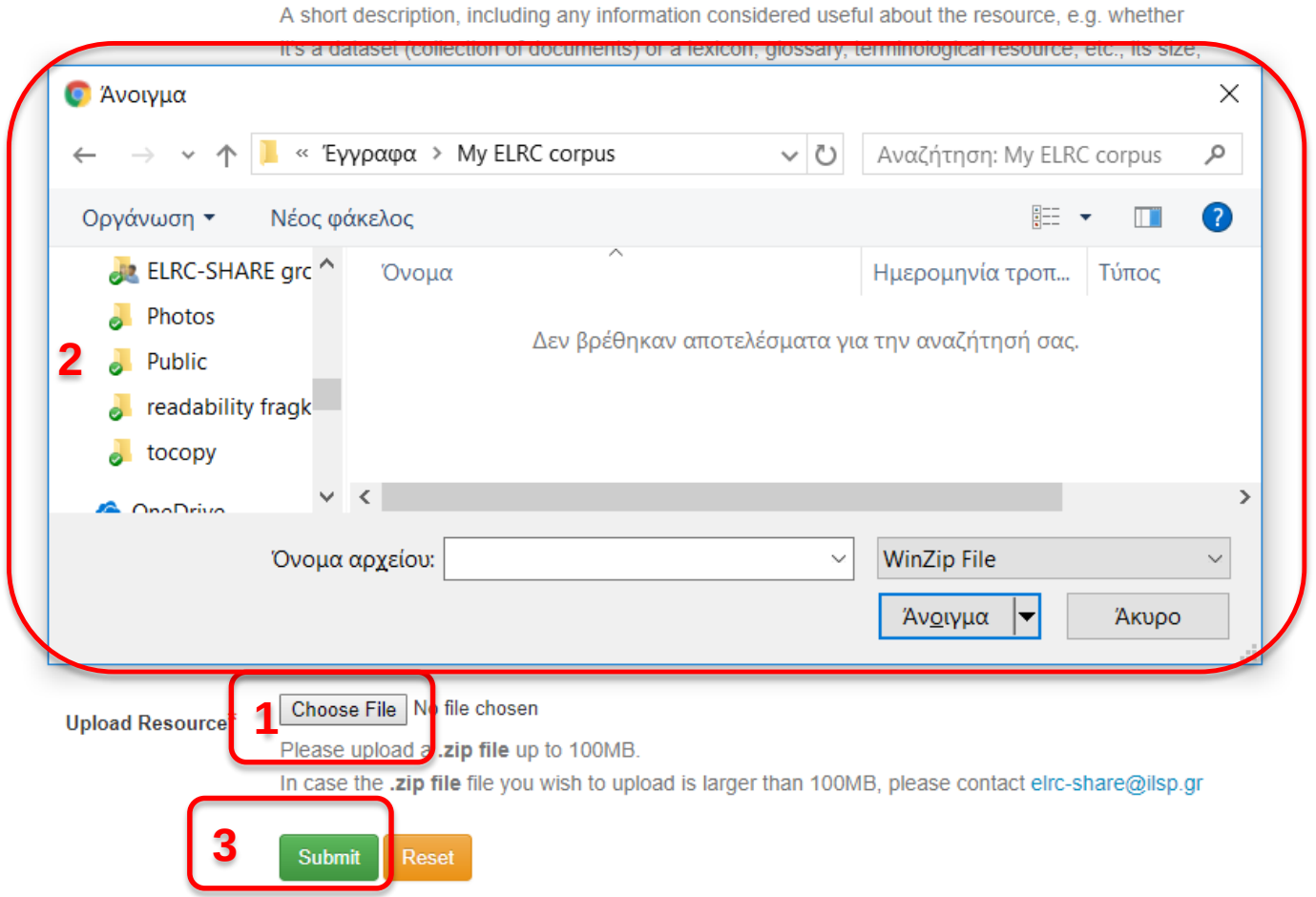
Please upload a **.zip file** up to 100MB.

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

Submit | Reset

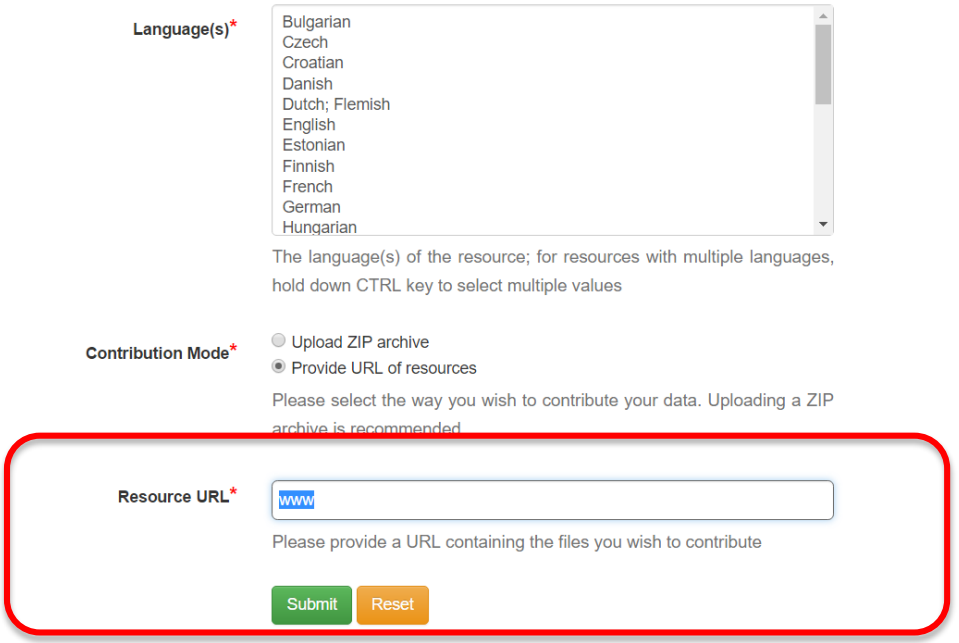
Wie kann man Daten beisteuern? (4/6)

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



Wie kann man Daten beisteuern? (5/6)

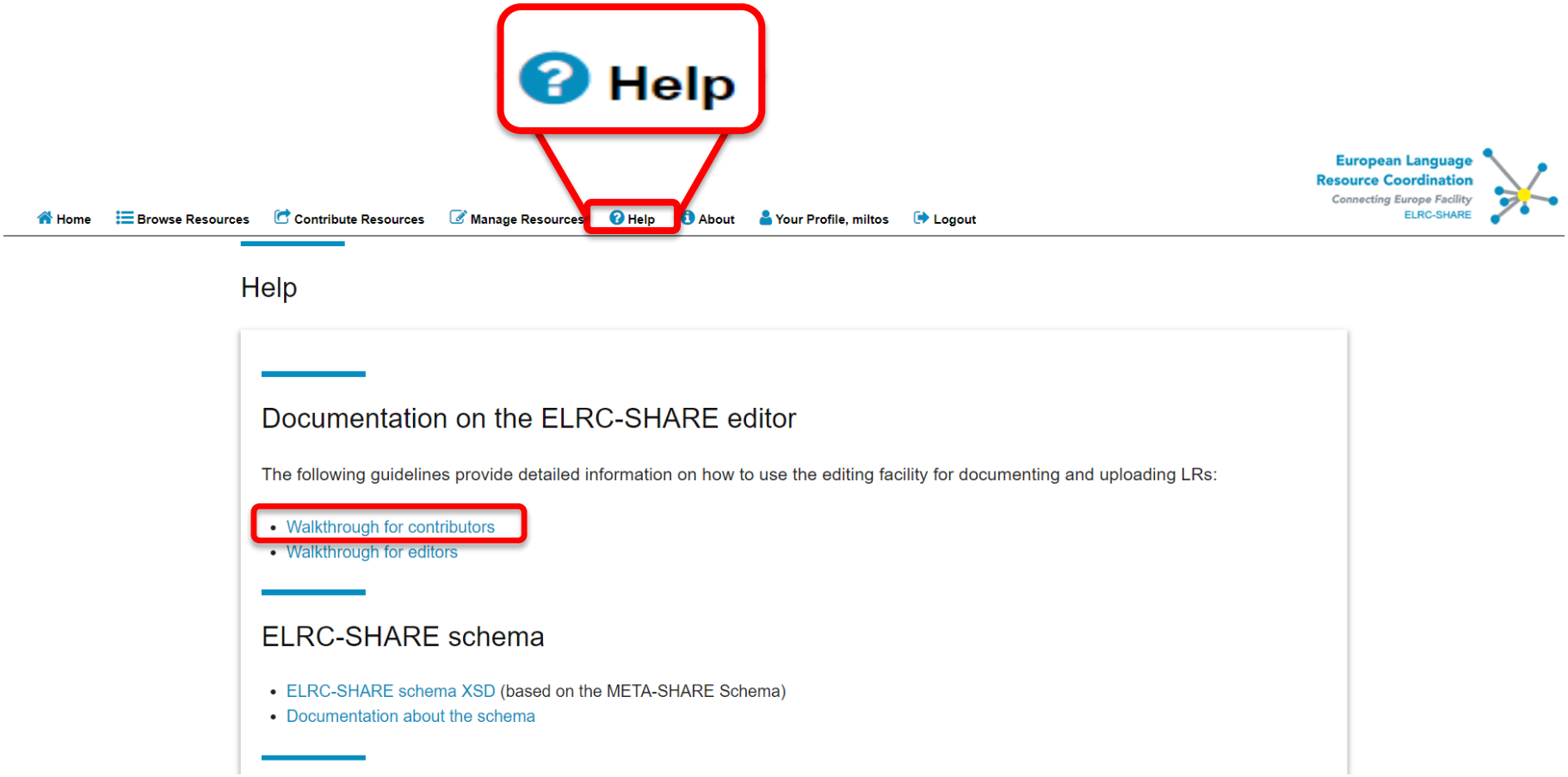
- Alternativ können Sie auch eine URL angeben (Verzeichnisauflistung)



Wie kann man Daten beisteuern? (6/6)

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

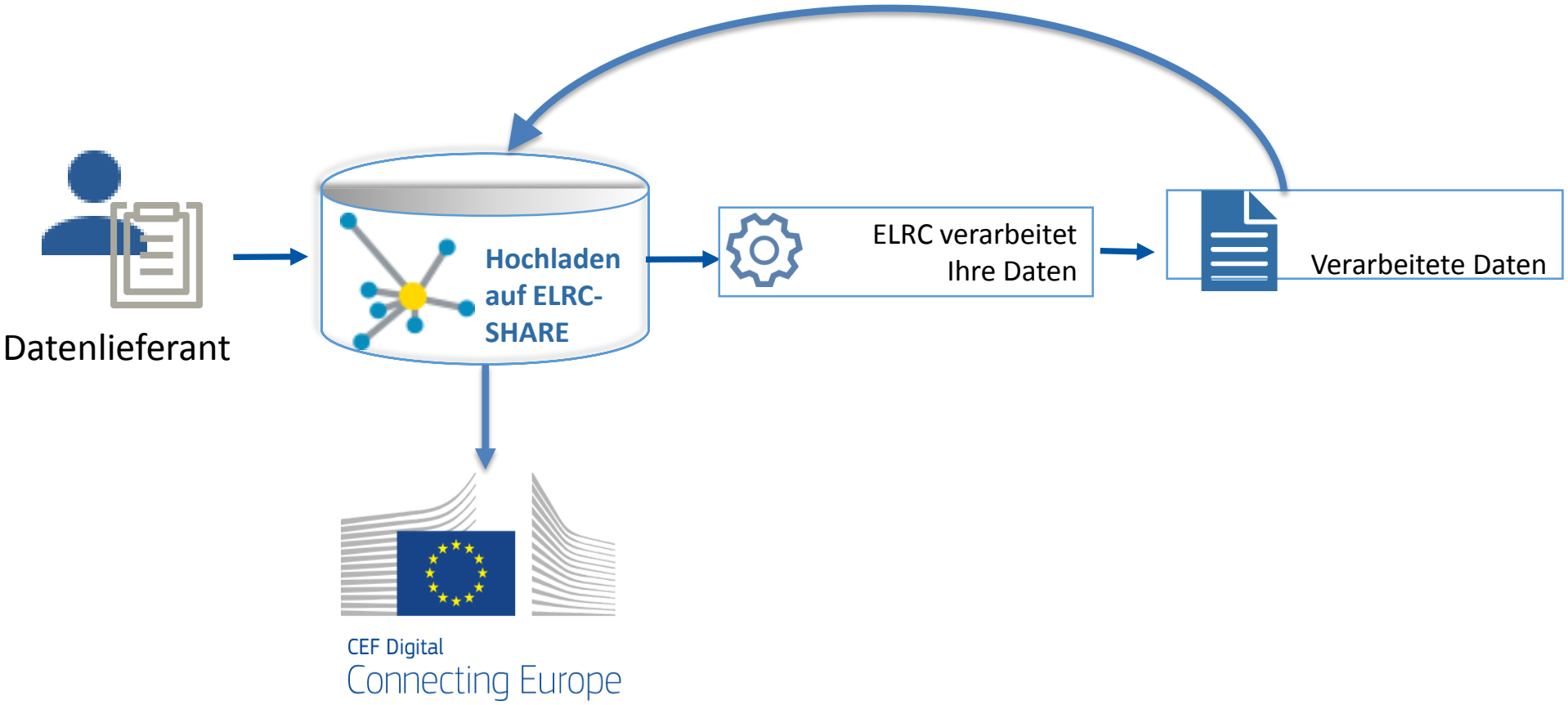
Richtlinien für Beitragende



Was passiert als nächstes?

European Language Resource Coordination
Connecting Europe Facility

Was passiert mit Ihren Daten?



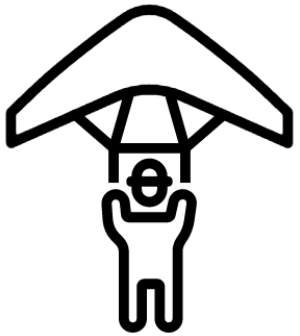
Datenverarbeitung vor der Auslieferung an die EK

- 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

Unterstützung vor Ort

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

Unterstützung vor Ort



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

European Language Resource Coordination
Connecting Europe Facility

Unterstützung vor Ort

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

European Language Resource Coordination
Connecting Europe Facility

Unterstützung vor Ort

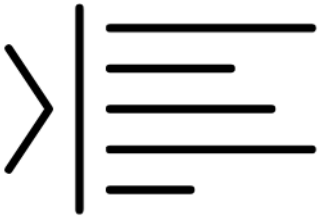
Wir verarbeiten Ihre Sprachdaten und geben die bereingten Daten wieder an Sie zurück. Wir können auch helfen, Ihre Datenmanagementprozesse zu optimieren. Sprechen Sie uns einfach an!

Textverarbeitungsdienstleistungen vor Ort

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.

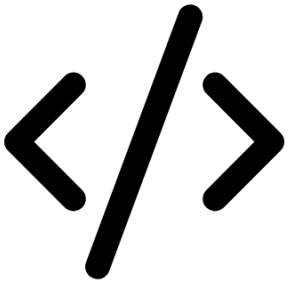
Wenn Ihre Daten “unsauber” sind, werden wir sie bereinigen.



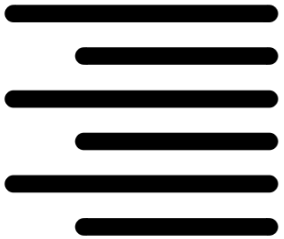
Müssen Sie DOCX in XML oder PDF in WORD umformatieren? Wir übernehmen das gerne für Sie!

Textverarbeitungsdienste

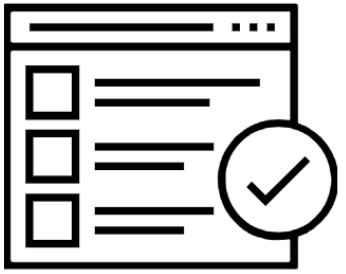
Wenn Ihre Daten nicht in den richtigen Formate konvertiert sind, können wir Ihnen helfen, sie zu konvertieren.



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



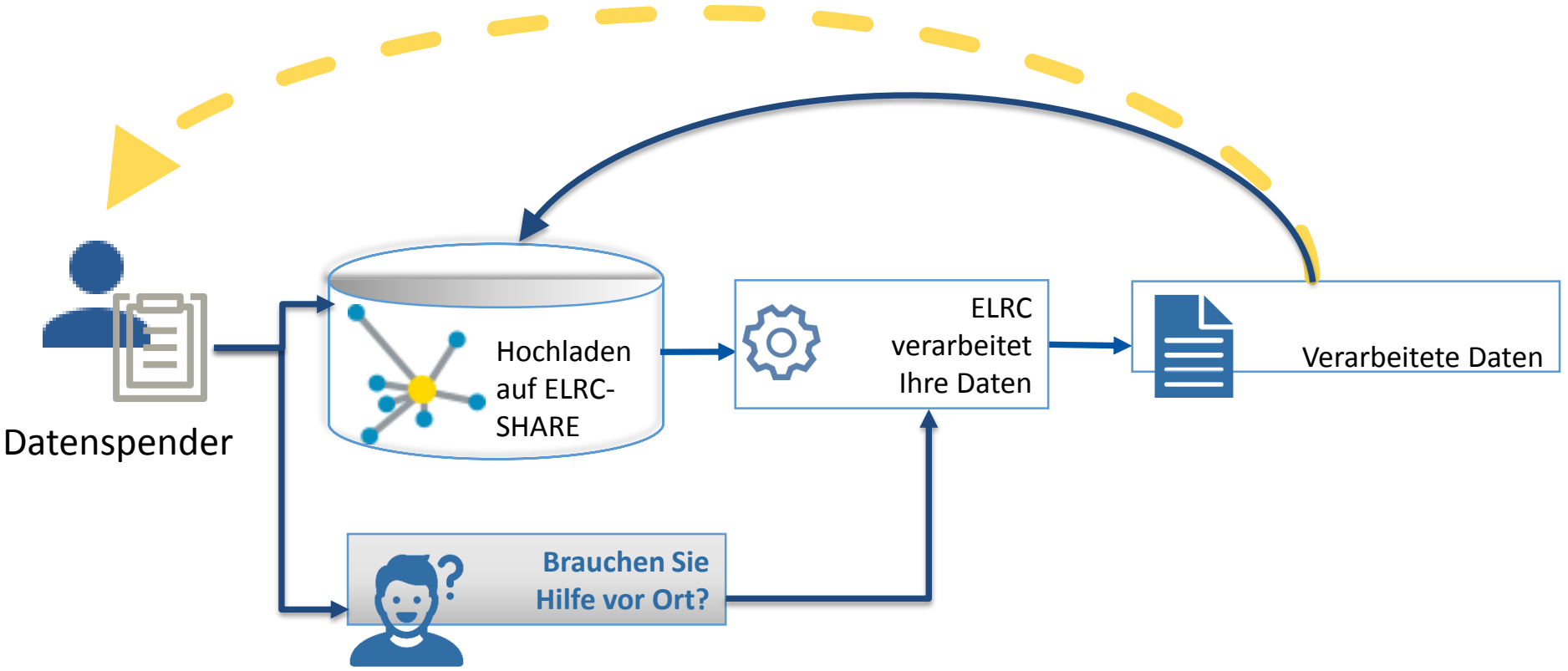
Sind die Übersetzungen nicht aligniert? Wir erledigen das für Sie mit unseren Werkzeugen!



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

European Language Resource Coordination
Connecting Europe Facility

Was passiert mit Ihren Daten?



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

Wie können Sie Hilfe vor Ort anfordern?

European Language Resource Coordination
Connecting Europe Facility

Unterstützung 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

European Language Resource Coordination
Connecting Europe Facility

ELRC Helpdesk

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

European Language Resource Coordination
Connecting Europe Facility

ELRC Konsortium – sprechen Sie uns an!

European Language Resource Coordination
Connecting Europe Facility

Vielen Dank für Ihre Aufmerksamkeit!

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