

The European Language Resource Coordination

La Coordination Européenne
des Ressources Linguistiques

ELRC

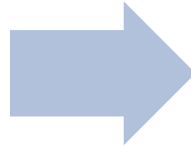


Préparer et partager les données
avec le service de dépôt ELRC
Khalid Choukri, Valérie Mapelli, Hélène Mazo
ELRA/ELDA



Données

- Tout contenu numérique



Données Linguistiques (Textuelles)

- Tout contenu textuel au format numérique

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-standard/ Other Licence/ Terms

Conditions: Non Commercial Use

Distribution Details

Attribution Details: NTU - Nederlandse Taalunie

Contact Person

Carole Tiberius 🇳🇱

text

Multilingual text corpus

Languages

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

Linguality

Linguality type: Multilingual

Text Format

TEX

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-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: 12/04/2017

Last Updated: 12/04/2017

Metadata Language: English (en)

Metadata Creator

Fraser Bowen 🇳🇱

Relations

Related Resource: Multilingual subtitle data 2BDutch (Processed)

Relation Type: Has Converted Version

People who looked at this resource also viewed the following:

- [Parallel Global Voices \(Greek - Spanish\)](#)
- [Parallel texts from Swedish Labour market agency](#)
- [Spanish-English website parallel corpus](#)
- [Spanish-Portuguese website parallel corpus](#)

Resources from the same project

La notion de données dans le context de eTranslation



ANR translation memory containing major publications, as well as several administrative documents and news 🗨

▶ View resource name in all available languages

Attribution details: ANR (Agence nationale de la recherche)

Documents / language resources from ANR –

Translation memory (.xliff) fr>en(uk) containing 9611 translation units (17 Mb)

Major publications

- Rapport d'activité 2014 (110 pages) (approximate synchronization)
- Rapport d'activité 2006 (only one report translated by a professional)
- Plan d'action 2... [Read More](#)

▶ View resource description in all available languages

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

This resource can be partially processed by ELRC Services. [Click here](#) to select service

text

<p>Distribution</p> <p>Availability: Available</p> <p>Licences</p> <p>Open Under-PSI</p> <p>Used for resources that fall under the scope of PSI (Public Sector Information) regulations, and for which no further information is required or available. For more information on the EU legislation on the reuse of Public Sector Information, see here: https://ec.europa.eu/digital-single-market/en/european-legislation-reuse-public</p>	<p>Bilingual text corpus</p> <p>Languages</p> <p>French (fr)</p> <p>English (en)</p> <p>Linguality</p> <p>Linguality type: Bilingual</p> <p>Text Format</p> <p>TM format of the SDL alignment tool</p>	<p>Resource Creation</p> <p>Funding Project</p> <p>Connecting Europe Facility - European Language Resource Coordination (CEF-ELRC - LANGUAGE RESOURCE COORDINATION - SMART 2014/1074 - 30-CE-0696785/00-64)</p> <p>URL: http://www.lr-coordi...</p> <p>Funding Type: Service Contract</p> <p>Funder: European Commission</p>
--	--	--

La notion de données dans le context de eTranslation



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

Trans.
Data

```
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>
```

The notion of data in the context of eTranslation



```
File01_nl.txt  
File01_en.txt  
File02_nl.txt  
File02_en.txt  
File03_nl.txt  
...
```

Trans.
Da

Amsterdam is de hoofdstad van ons land.

Nu is het een grote en drukke stad.

En dit is 'De Dam',

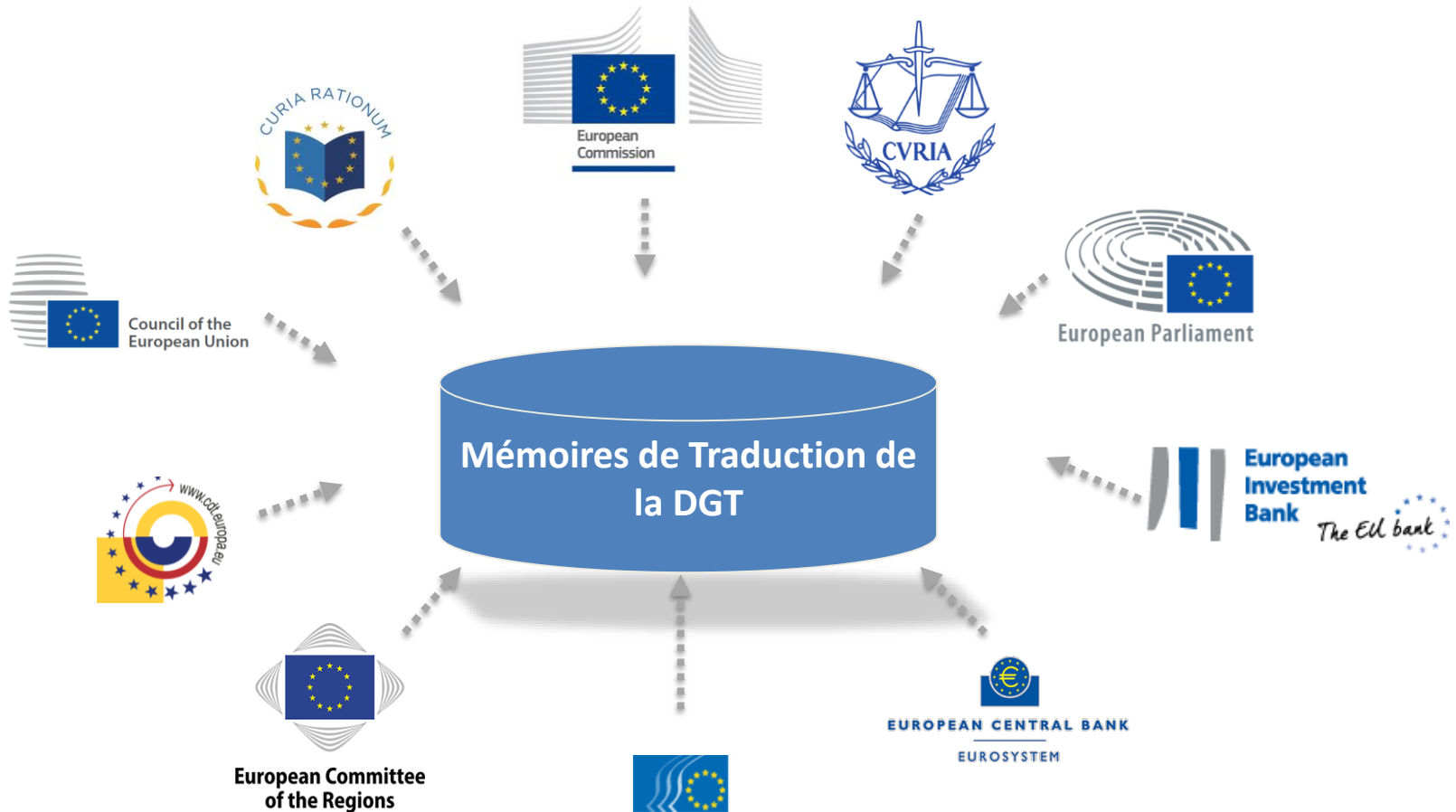
het beroemde plein in Amsterdam.

Amsterdam is the capital of
our country

It is a big and busy city

This is Dam square

It is the famous square of
Amsterdam



Ces données sont déjà utilisées par la DGT
mais
il en faut beaucoup plus

- Toutes données détenues par des organismes publics,
- Produites en interne ou en sous-traitance
 - Rapports
 - Documents de Communication
 - Documents d'information (Nouvelles)
 - Contenu Web produit en plusieurs langues
 - Politiques
 - Terminologies
 - Archives
 - Formulaires
 - Foires aux Questions (FAQs)
 -

Quelles données sont utiles pour eTranslation selon le type | 1



- Tout texte numérique dans une des langues de l'UE + Norvège/Icelande
- **Des textes et leurs traductions (par exemple, bilingues ou multilingues parallèles)**

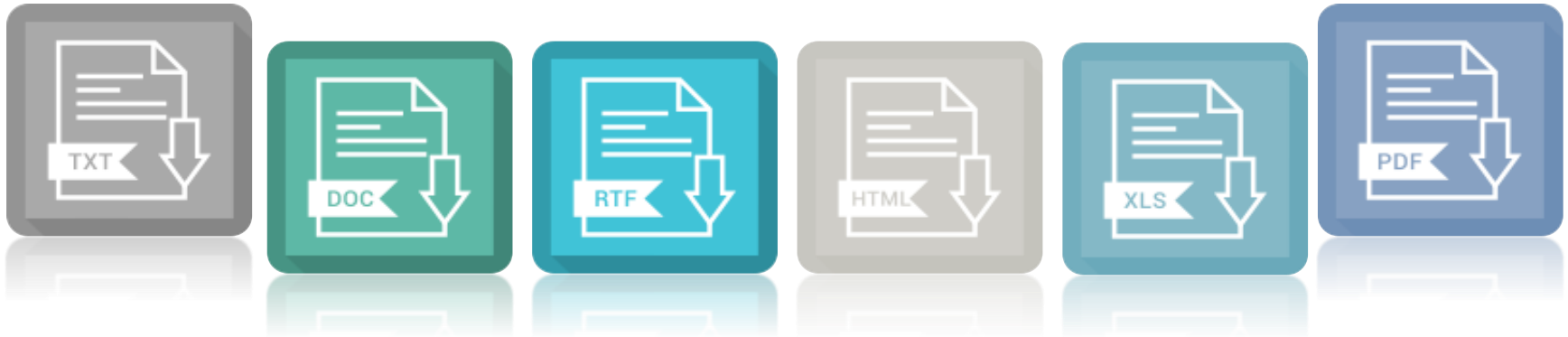
Texte en NL

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

Traduction en Français

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.

Quelles données sont utiles pour eTranslation selon le format | 1



- En principe, tout texte au format numérique (format que peut lire des ordinateurs),
- Mais certains formats sont plus "prêts" pour le traitement que d'autres, c'est-à-dire qu'ils nécessitent moins de traitement (manuel ou automatique).
- Plus de traitement introduit plus d'erreurs dans le résultat final, le rendant moins utile pour eTranslation



Les formats suivants sont particulièrement utiles:

- Pour les textes parallèles bilingues / multilingues
 - Mémoires de traduction (.tmx)
 - Fichiers de traduction XML (.xliff)
 - Texte brut (.txt, .csv)
 - Feuilles de calcul (par exemple xlsx)
- Pour les terminologies
 - TermBase eXchange (.tbx)
 - Texte brut (.txt, .csv)
 - Feuilles de calcul (par exemple xlsx)
- Pour les textes monolingues
 - Texte brut (.txt, .csv)



Formats de fichiers de textes parallèles et leur manipulation



A ne pas faire!!



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.

Ne pas fusionner le texte source et le texte traduit (cible) en un seul document



A ne pas faire!



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'agrération 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.



A ne pas faire!



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.



A faire

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

Utilisation d'un fichier par langue et la même façon de nommer les paires de fichiers (source – traduction)



A faire

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

Inclure le nom de la langue
(ou son identifiant) dans le
nom de fichier

Quelques critères de regroupement de vos données



- Un ensemble de données est un ensemble de données "regroupées selon certains critères"
- Afin de renforcer et d'adapter le système eTranslation, deux critères sont essentiels:
 - **Langue (s)**: chaque collection est définie par la ou les paires de langues de ses données, par ex:
 - Recueil de textes en [anglais - allemand](#)
 - Documents en [anglais - norvégien - finnois](#)
 - **Domaine**: chaque collection appartient idéalement à un seul domaine, par exemple:
 - Recueil de textes en anglais - allemand dans [le domaine de la culture](#)
 - Documents de [sécurité sociale](#) en anglais - norvégien - finnois

- Domaines des écrits Administratifs / réglementations et tout ce qui concerne les CEF DSI (Infrastructures de Services Numériques)

CEF DSI	Domaines
Résolution des litiges en ligne (ODR)	Droits des consommateurs, litiges
Échange électronique d'informations sur la sécurité sociale,	Sécurité Sociale, Assurance
eProcurement (passation électronique des marchés)	Commande Publique,
Portail e-Justice	Justice, Loi
Portail eHealth	Santé, Médecine,
Business Registers Interconnection System (Registres de Commerce Interconnectés)	Affaires, Marché
Internet Sans Crainte (Safer Internet)	
Cybersécurité	
Données Publics Ouvertes	
Europeana	Culture

Comment contribuer vos données à eTranslation

Un guide en quelques étapes

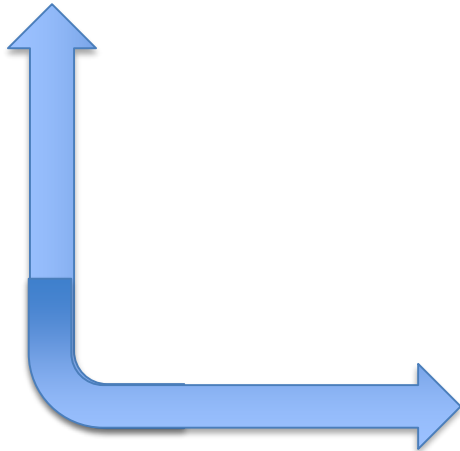
European Language Resource Coordination — supporting Multilingual Europe

ELRC-SHARE

Subscribe to Newsletter

- Sur le portail ELRC, cliquez sur le bouton «Soumission de ressources linguistiques». Ou Tapez l'adresse URL:

elrc-share.eu



Type in your keywords, please...

Recherche ▶

Qu'est-ce que les ressources linguistiques?

Les termes ressources linguistiques désignent des séries de données et de descriptions de langues sous une forme lisible par machine, y compris les bases de données terminologiques et écrites et orales, les grammaires et les bases de données terminologiques. Les ressources linguistiques peuvent être utilisées pour construire, améliorer ou évaluer des systèmes linguistiques naturels tels que les moteurs de traduction automatique.

Afin de développer les systèmes de traduction automatique pour la plateforme de traduction automatique MIE, l'initiative ELRC vise à rassembler des ressources linguistiques dans toutes les langues officielles de l'UE. L'initiative vise un vaste corpus de grands domaines, qu'il s'agisse d'une langue monolingue (par exemple un corpus de langues nationales) ou d'une langue multilingue, ainsi que de ressources linguistiques spécifiques à un domaine dans les domaines des droits des consommateurs, de la culture, du domaine juridique, de la sécurité sociale, de la santé, des marchés publics, etc.

[En savoir plus sur les ressources linguistiques nécessaires](#)

Comment contribuer?

Tout contributeur peut nous présenter des ressources linguistiques à n'importe quel stade de l'exploitation: Des liens internet simples vers des sites web (sources), des données brutes ou des données entièrement emballées (ressources linguistiques).

Cliquez ci-dessous si vous pouvez indiquer une source potentielle de données pertinentes

Soumission des sources de données▶

Cliquez ci-dessous si vous êtes un propriétaire de ressources linguistiques et souhaitez les partager pour les besoins du CEFR

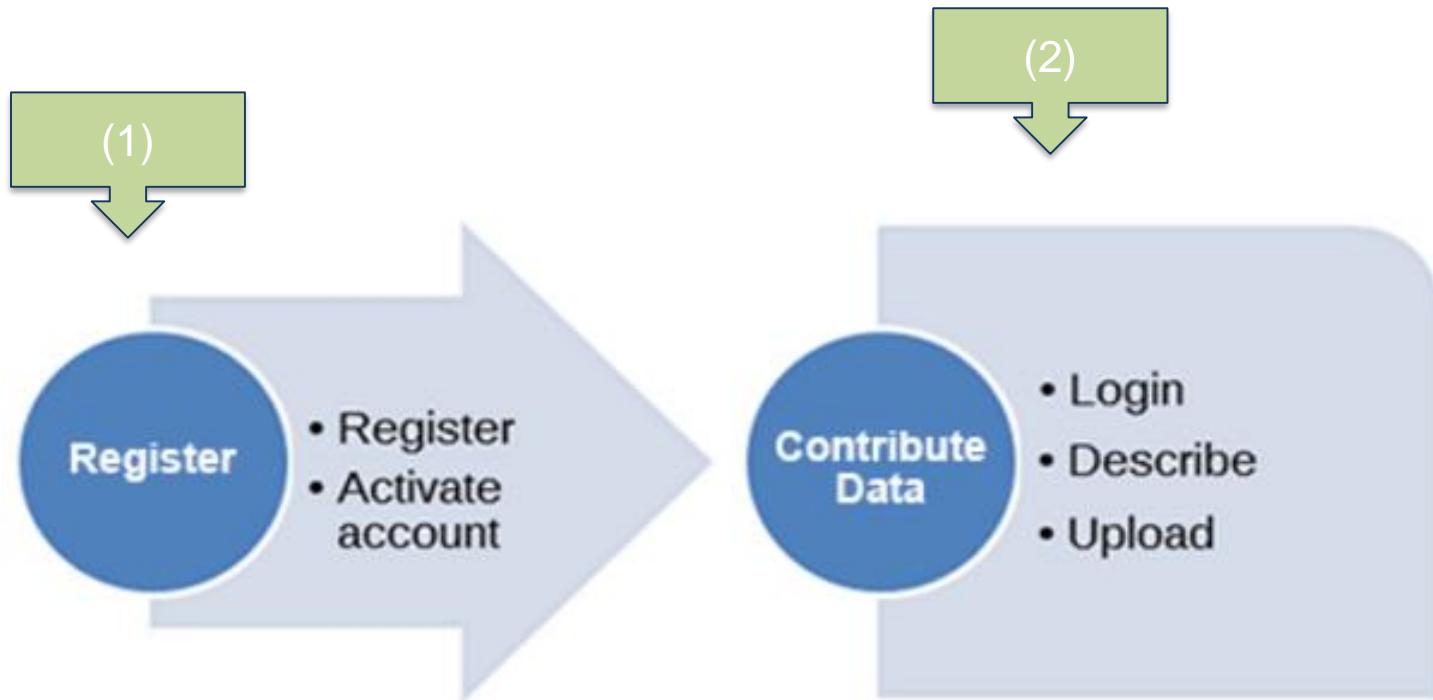
Présentation des ressources linguistiques▶



ELRC-SHARE Repository



Welcome to the ELRC-SHARE repository!





 Register

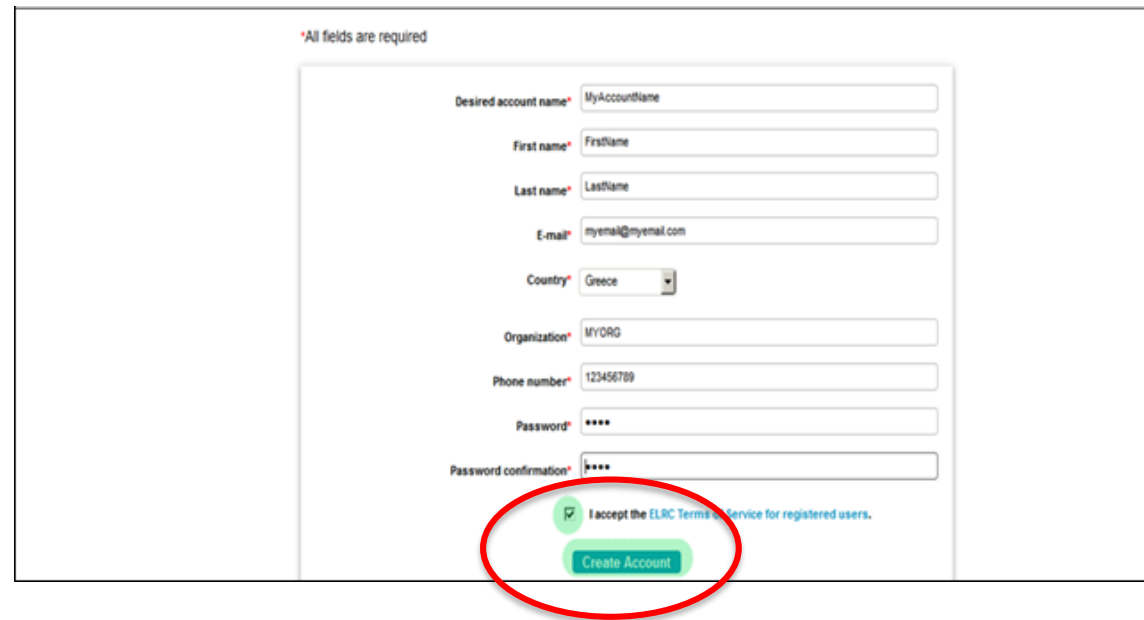
ELRC-SHARE Repository



Welcome to the ELRC-SHARE repository!

- Remplissez les informations requises
- Lisez les conditions d'utilisation et cliquez sur "Accepter", si vous êtes d'accord.
- Cliquez sur le bouton "Créer un compte".
- Activez votre compte conformément aux directives qui vous sont envoyées par courriel.

*All fields are required



Desired account name* MyAccountName

First name* FirstName

Last name* LastName

E-mail* myemail@myemail.com

Country* Greece

Organization* MYORG

Phone number* 123456789

Password* ****

Password confirmation* ****

I accept the ELRC Terms of Service for registered users.

Create Account



 **Contribute Resources**

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

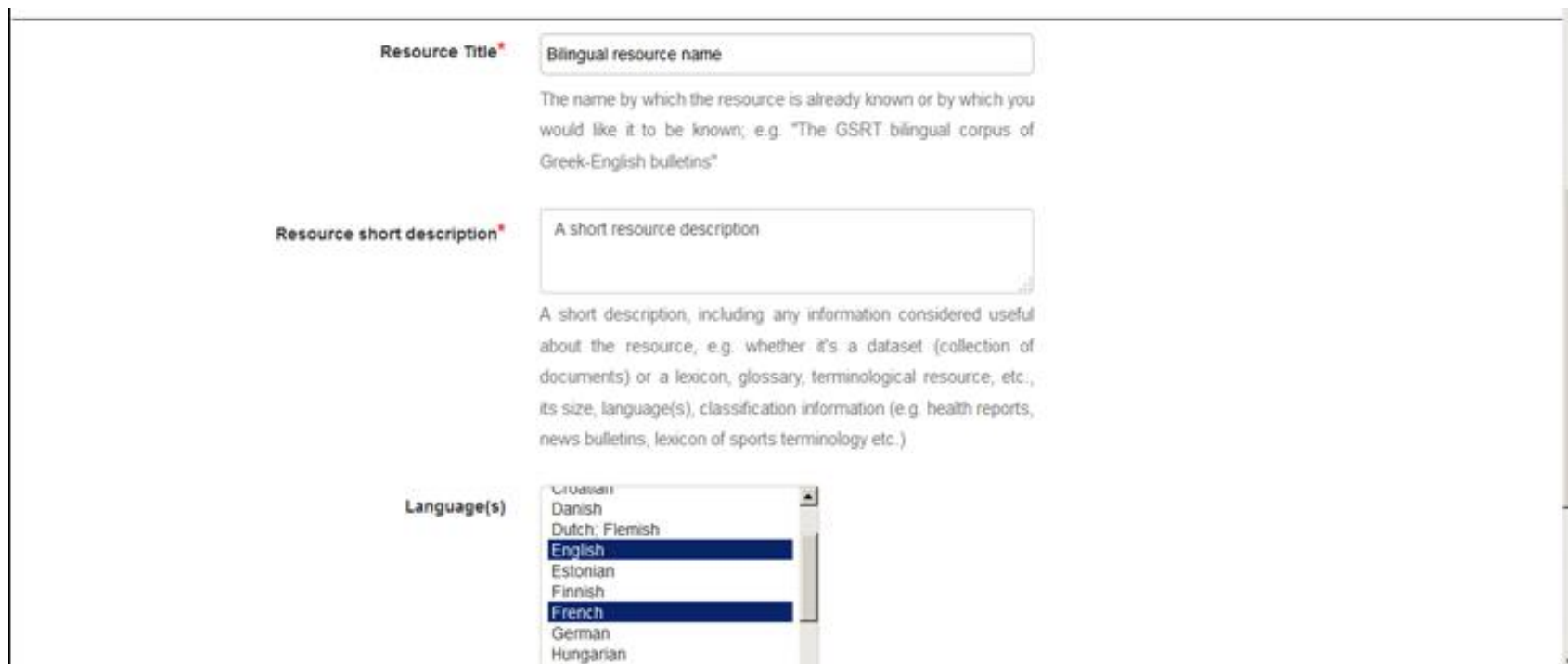
Data Contribution

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"

- Remplissez les détails décrivant l'ensemble de données



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)

- Crusian
- Danish
- Dutch/Flemish
- English
- Estonian
- Finnish
- French
- German
- Hungarian



• Trois modes pour contribuer vos données

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

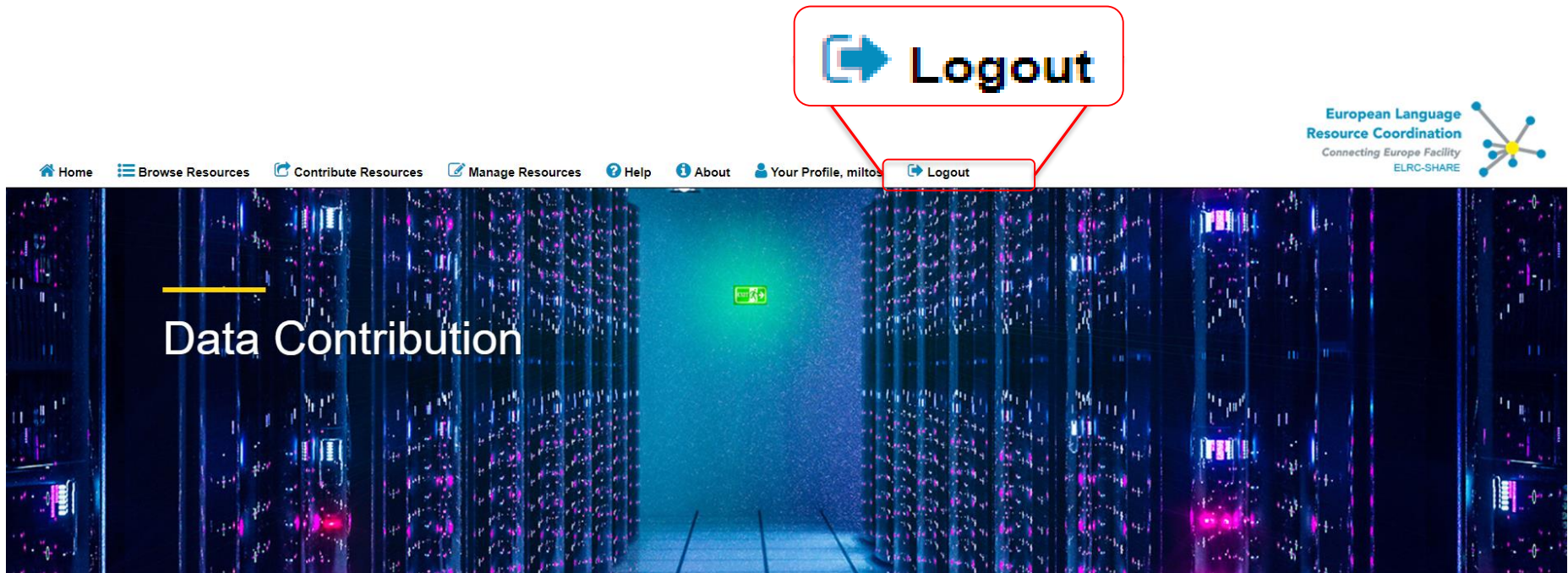
Please upload a **.zip file** up to 100MB.

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

Submit

Reset

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



The screenshot displays the top navigation bar of the ELRC-SHARE website. The menu items are: Home, Browse Resources, Contribute Resources, Manage Resources, Help, About, Your Profile, milto, and Logout. The 'Logout' link is highlighted with a red box, and a red callout bubble with a blue arrow icon points to it with the text 'Logout'. The background of the page is a server room with blue and purple lighting. The text 'Data Contribution' is overlaid on the left side of the server room image.



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:

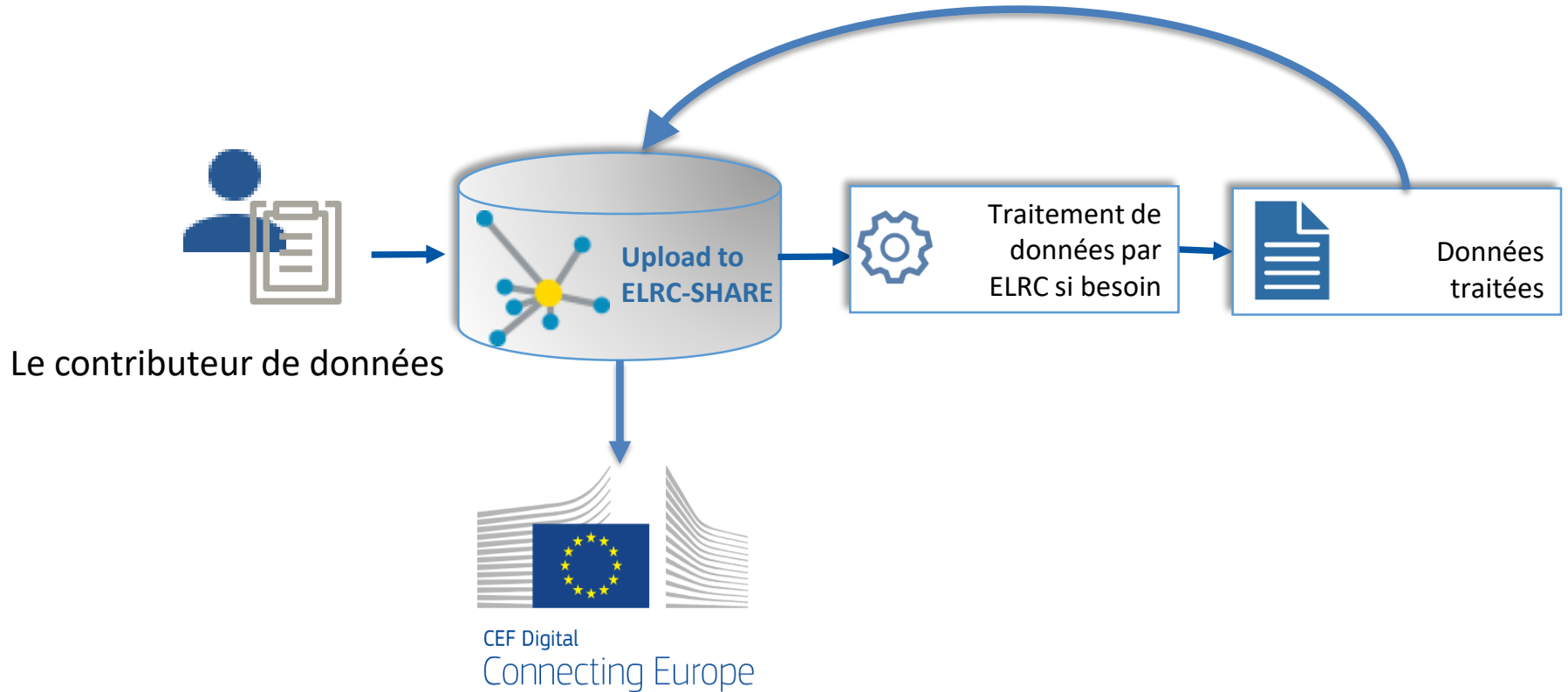
- [Walkthrough for contributors](#)
- [Walkthrough for editors](#)

ELRC-SHARE schema

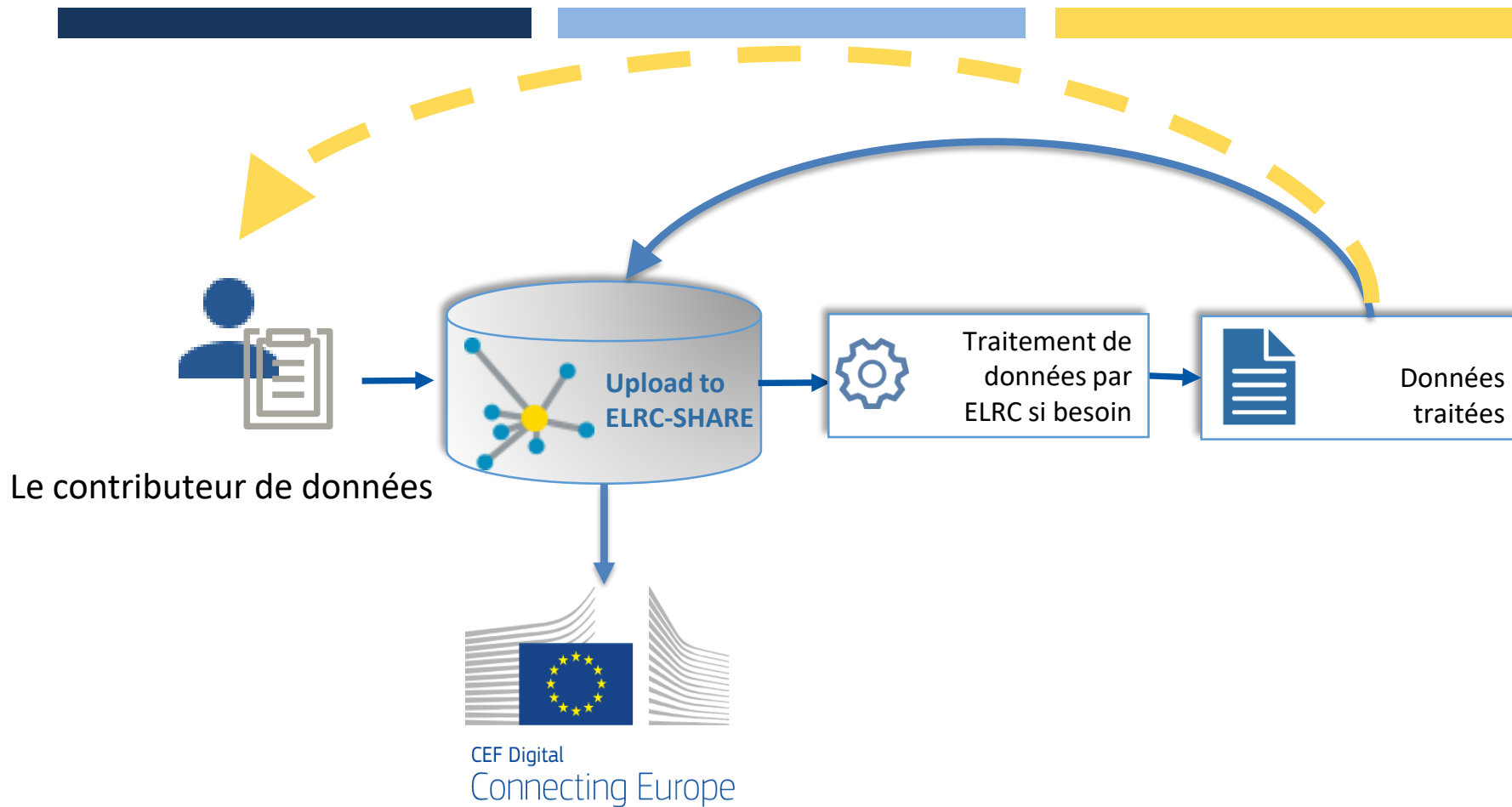
- [ELRC-SHARE schema XSD](#) (based on the META-SHARE Schema)
- [Documentation about the schema](#)

et la suite !!

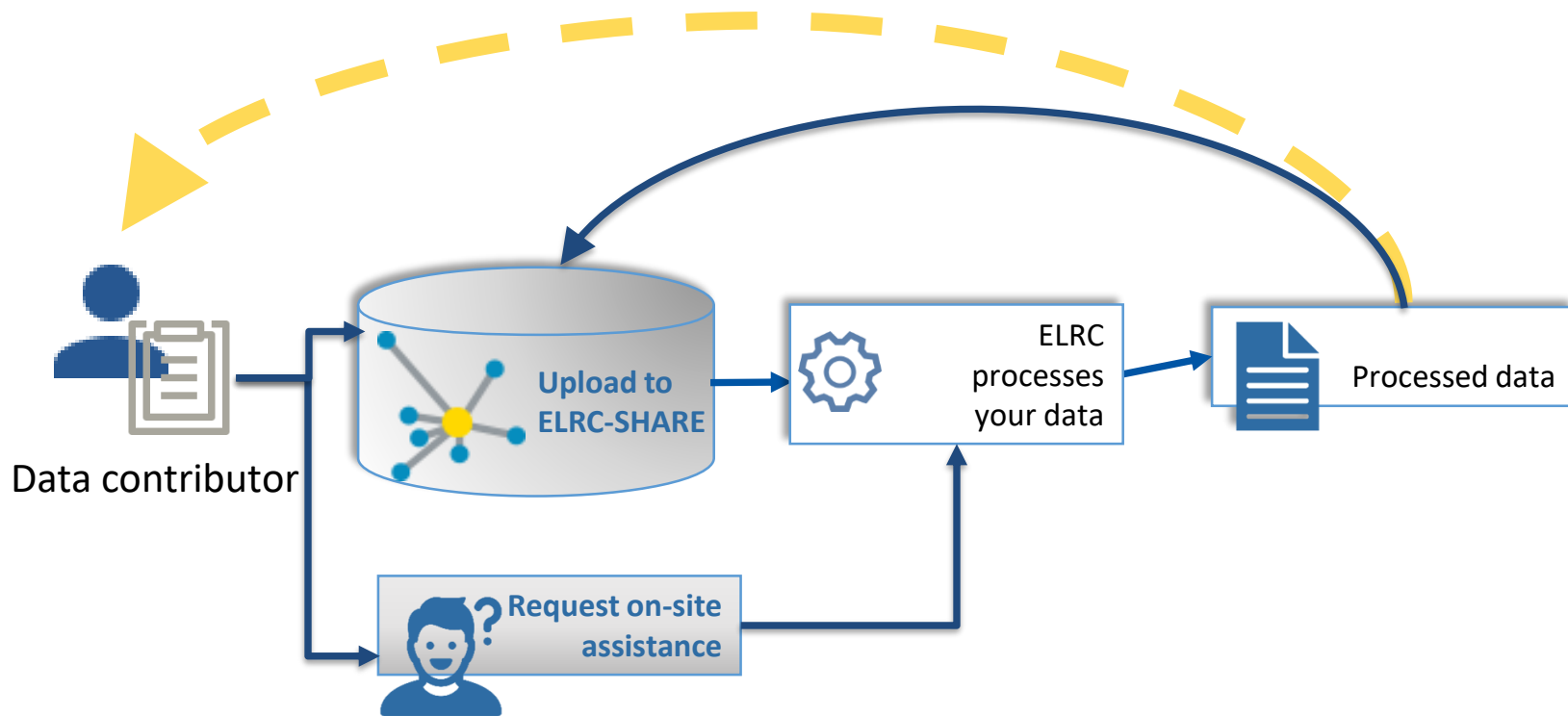
Qu'advient-il de vos données?



Partenariat avec les autres initiatives e.g. ELRI



Partenariat avec les autres initiatives e.g. ELRI





Tous les services de préparation de données
peuvent également être offerts gratuitement
sur votre site
Pour tous les contributeurs de données.



L'assistance sera fournie en coopération étroite
avec un vaste réseau d'experts



Nous pouvons vous assister pour résoudre vos problèmes de données

Une fois "préparées", vous obtenez une copie de cette version.

Nous pouvons également vous aider à améliorer vos processus de gestion des données. Il suffit de demander!



Extraction de données

Si vos données sont dans des archives et des bases de données, nous pouvons vous aider à les extraire.



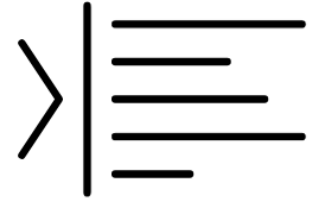
Anonymisation

Vos données contiennent-elles des informations personnelles? Nous pouvons aider à les anonymiser



Nettoyage

Si vos données contiennent beaucoup de "bruit", nous pouvons les nettoyer.



Re-formatage

Besoin de reformater DOCX en XML ou PDF en WORD? On peut vous aider

Comment demander une assistance sur site services et de l'aide



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

[Ir-coordination.eu/request-onsite-assistance](https://ir-coordination.eu/request-onsite-assistance)



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-coordantion.eu

lr-coordination.eu/helpdesk

Merci de votre attention

Courriel : info@lr-coordination.eu

Site web: www.lr-coordination.eu

