



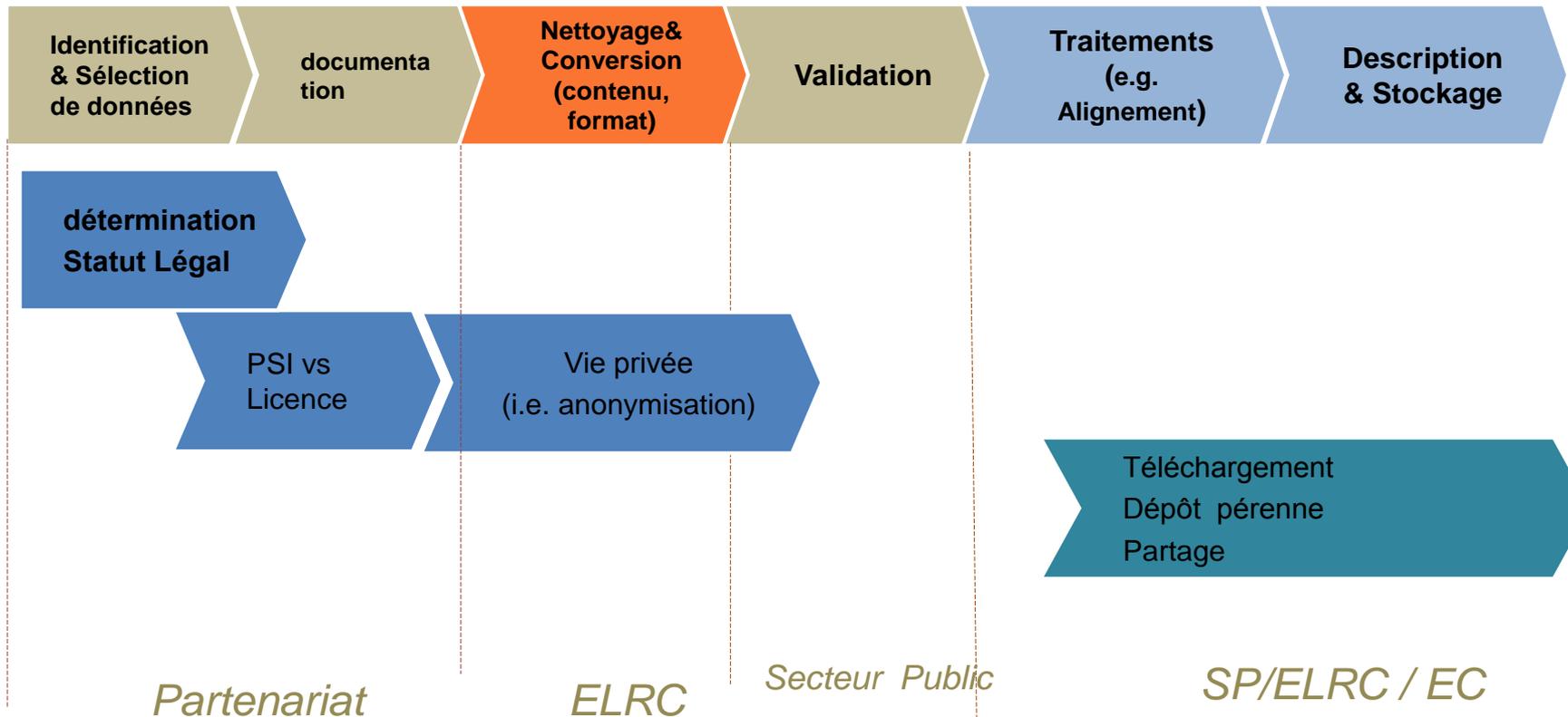
Ressources Linguistiques Bonnes pratiques pour l'avenir et Outils pour la fourniture de données

Khalid Choukri (ELDA)

- Nous avons vu l'importance des données pour la Traduction Automatique
 - Apprentissage par les données (Data Driven Paradigm)
- Les données sont nécessaires dans toutes les langues (s)
- Où peut-on découvrir des données: Les acteurs du secteur public
 - Données Visibles par exemple des données Web (pages HTML, rapports, etc.)
 - Données Invisibles: archives, web caché (profond), dépôts/référentiels internes
 - Par le biais des prestataires de services linguistiques
- Que peut-on faire pour capitaliser sur ces actifs (les données)
 - Notre expérience avec les "plans de gestion des données" (Data Management Plan)
 - Pour un partage durable et pérenne

ex. Données existantes → RLs (Ressources Linguistiques)

la chaîne de valeur

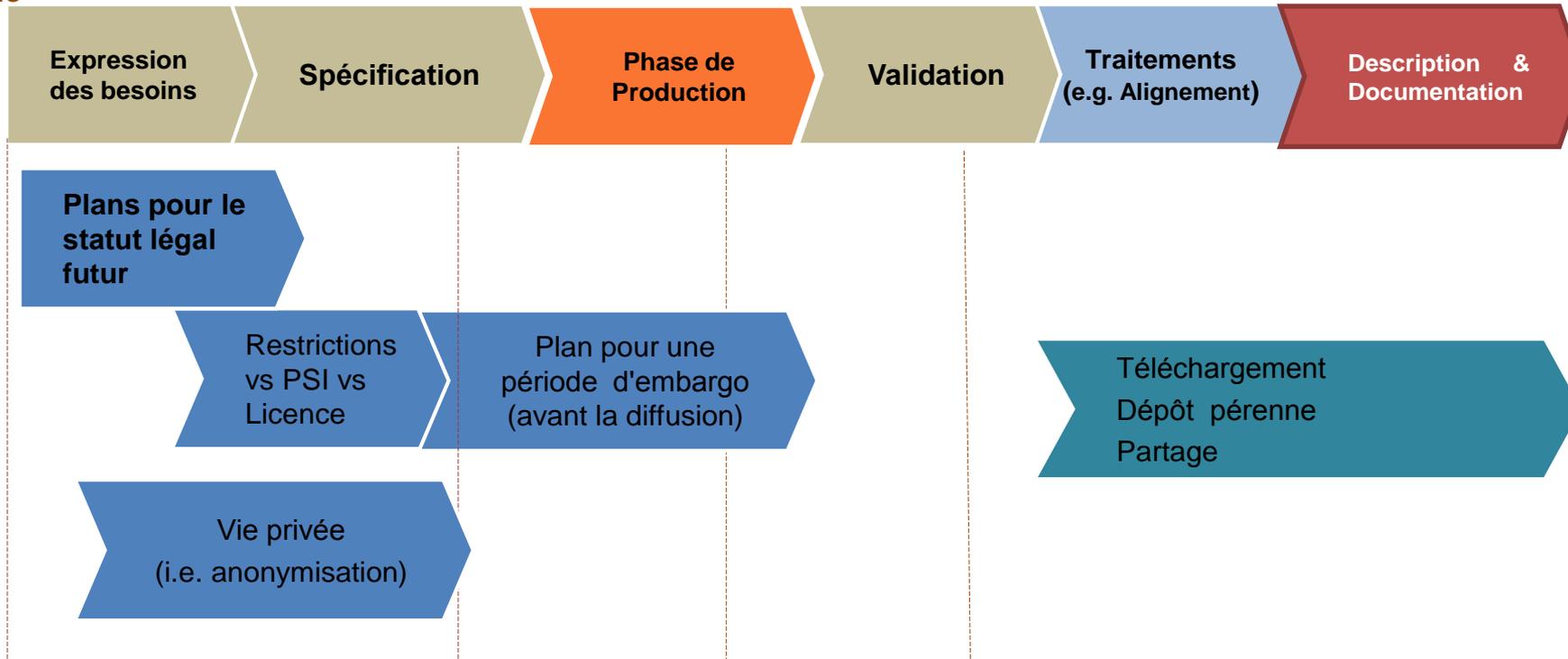




- Analyser toutes les phases de développement de données
- Sur la base de 1), créer un plan de gestion des données considérant les aspects
 - Juridique, flux de données, formats, publication comme PSI,
 - Relations avec les sous-traitants et autres partenaires
- Envisager la pérennité des données
 - Spécification des données, la production, la validation, le partage et la distribution, l'entretien et la préservation
- Utiliser le Web comme un canal supplémentaire de publication (voir comment ELRC peut aider)

Données nouvelles → RL (Ressources Linguistiques)

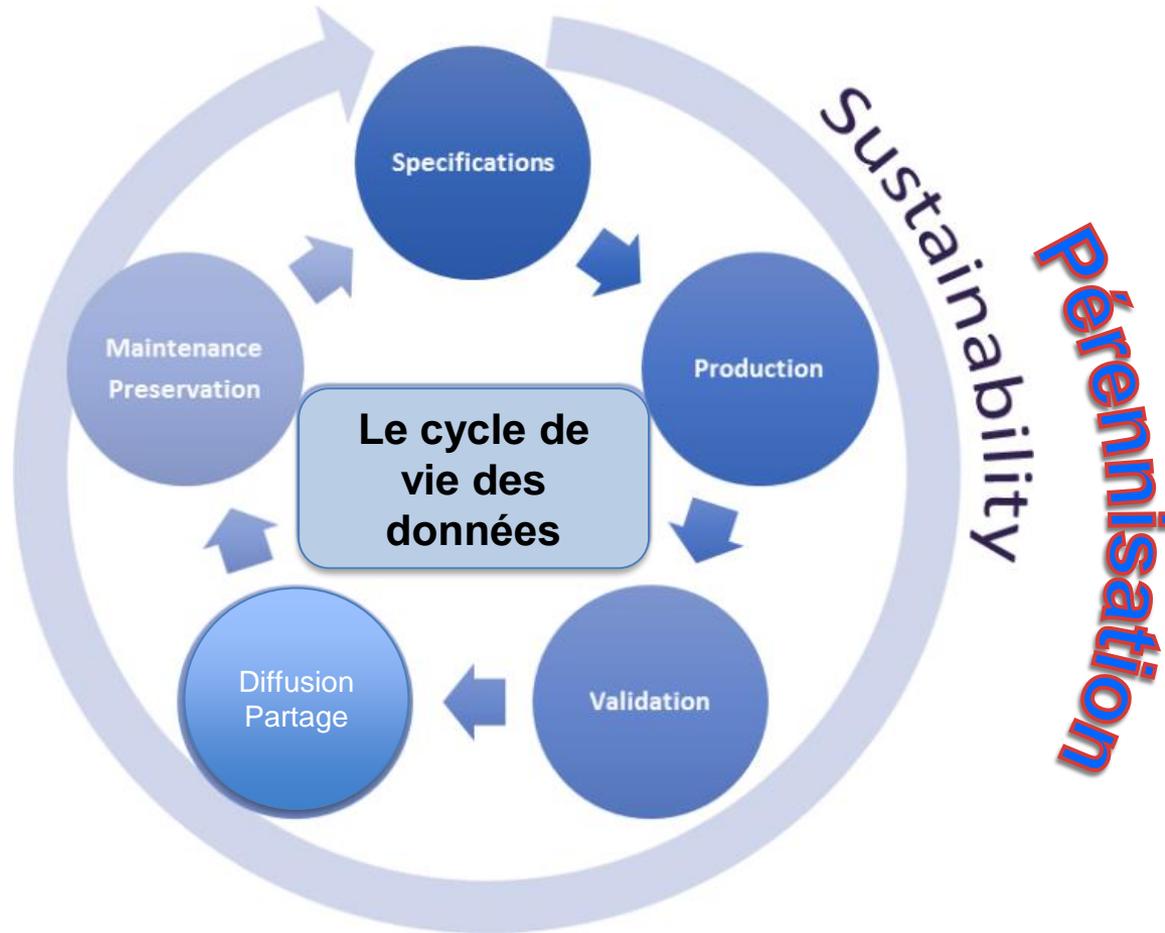
Chaine de
Valeur



Plan de gestion de données // data management plan (DMP)

- Anticiper toutes les questions juridiques potentielles
 - Veiller à ce que vos droits de propriété intellectuelle de données sont effacée
 - Veiller à ce que les parties productrices adhèrent à votre droite "propriété" (par exemple, les relations avec les LSP: assurez-vous de garder tous les droits
 - Veiller à ce que tous les documents intermédiaires produits vous appartiennent
 - (par exemple des mémoires de traduction)
 - Vérifiez les questions de confidentialité à l'avance et prévoir plan d'anonymisation si nécessaire
- Définir votre plan de gestion par rapport à la tâche:
 - Cela doit tenir compte de l'objectif principal (par exemple production de documents, traduction de doc, etc.)
- Plan de réorientation/réaffectation
 - Convertir la documentation en ressource linguistiques pour les Technologies de la langue
 - Demander les données dans un format éditable
 - Pas seulement des fichiers PDF, mais aussi TMX / Word / XML / TXT/...
 - Assurez-vous que vos archivages utilisent des "medium" à jour (CD?)
 - Prévoir une publication future et le partage de l'information (PSI)

Eléments Clés d'un plan de gestion de données





- Spécifications
 - Veiller à ce que les documents originaux soient décrits
 - Veiller à ce que vos besoins soient décrits
 - Anticiper ce que vous pouvez obtenir des ressources précieuses (un effet secondaire)
- Production
 - Que ce soit en interne ou externalisé, vérifier que les outils utilisés sont compatibles avec vos besoins et au-delà (par exemple CAT, MT, etc.)
 - Demandez la liste des outils et des logiciels de production
 - Vérifiez si vous pouvez obtenir des textes dans les différentes langues , alignés les uns aux autres
 - Gardez une claire bonne documentation des données en cours de production (méta-données)



- Validation
 - En plus de votre Contrôle Qualité, vous pouvez utiliser certains outils de validation e.g. analyse cohérence lexicale, analyse syntaxique, etc.
 - Souvent des outils libre de droit , open-source
- Partage / diffusion
 - Assurez vous que vos données relève bien de la directive PSI telle que transposée dans la législation de votre pays
 - Envisagez une licence ouverte et permissive s'il en existe pas une dans votre administration
 - Le respect/la protection de la vie privée est crucial, planifiez les procédures nécessaires pour gérer cela
- Maintenance / préservation
 - La meilleure option est souvent partenariat avec un centre de données (ELRA)
 - Voir dans le présent cadre comment ELRC peut vous aider
 - Il y a aussi l' «option» du portail de données ouvertes national
 - mettre les données sur Internet est rarement suffisant (référencement, pérennisation)

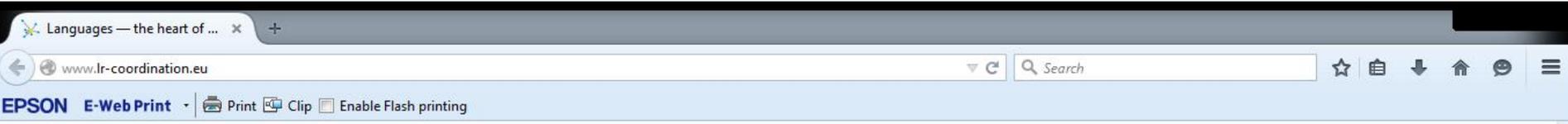
- Identification des sources, l'identification et la sélection des ensembles de données (données brutes)
 - Les données peuvent être obtenues à partir des sources visibles (par exemple récoltées sur le Web),
 - Les données peuvent être remises par les acteurs du secteur public,
 - Les acteurs du secteur public peuvent stimuler l'identification des sources visibles,
- Le traitement indiqué ci-dessus peut être réalisé en coopération par le ELRC et le fournisseur de données



- ✓ Vous connaissez bien vos données: Visibles et invisibles
- ✓ Accès aux archives, web caché/profond, etc. est souvent impossible de l'extérieur
- ✓ Toutes les données ne sont pas déjà en PSI ou sous licence permissive
- ✓ L'accès à des formes dérivées (par exemple, PDF) est moins efficace que l'accès aux contenus "source" (interne).

- Re-"cibler" /Réorienter les données existantes (traductions humaines) est la meilleure façon d'améliorer la qualité de la traduction automatique
- Le paradigme "apprentissage par les données" fournit un moyen efficace de tirer parti de la valeur des ressources existantes
- ELRC peut aider à l'examen de la pertinence des données (à toutes les phases)
- **Ne pas sous-estimer la valeur de vos ressources linguistiques, prévoir un plan de gestion des données**

Helpdesk et Support

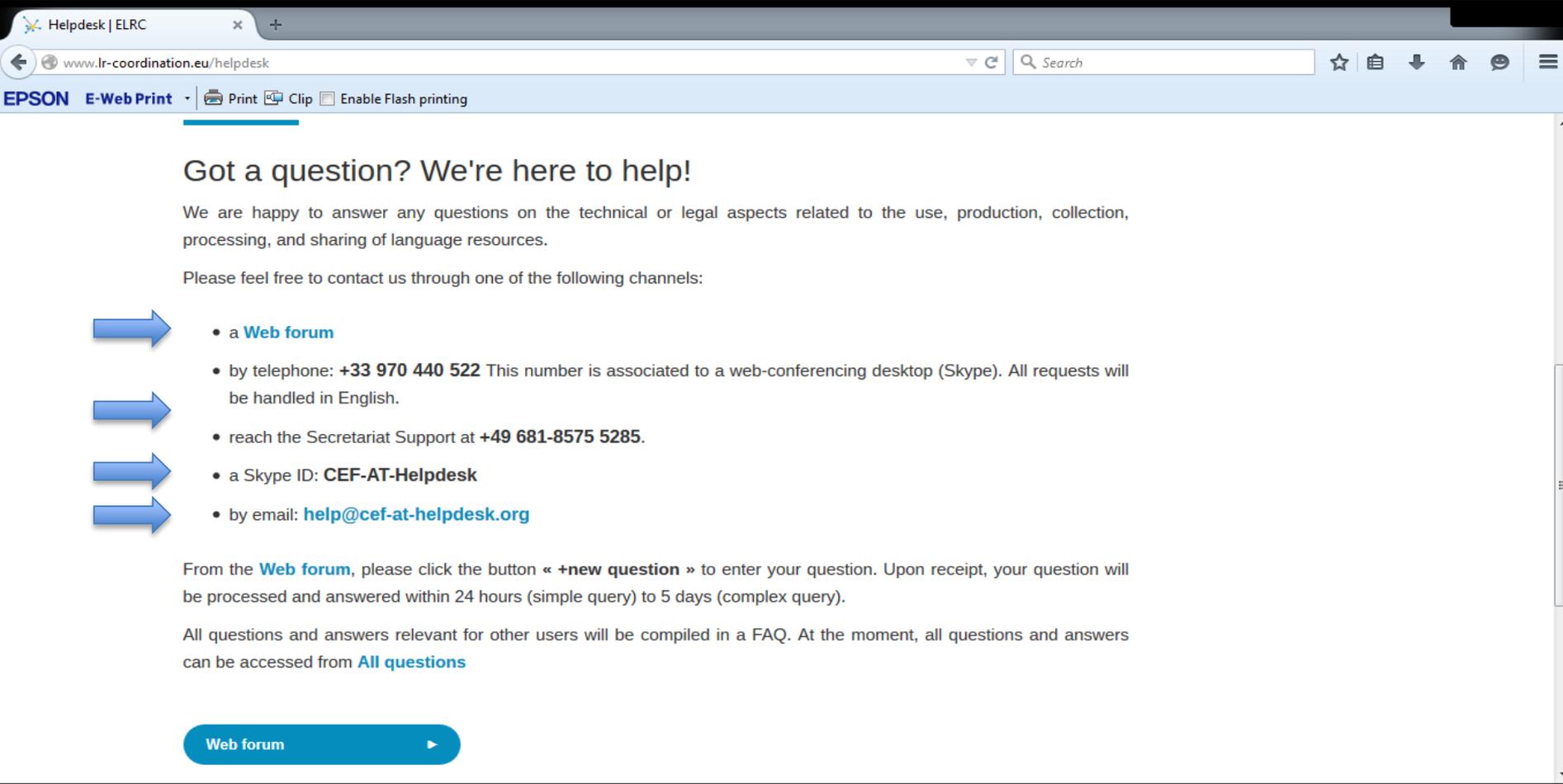


- [Home](#)
- [About](#)
- [News](#)
- [Helpdesk](#)
- [Events](#)
- [Resources](#)
- [Anchor Points](#)
- [Multilingual Europe](#)



Languages — the heart of
Multilingual Europe





The screenshot shows a web browser window with the address bar displaying 'www.lr-coordination.eu/helpdesk'. The page content includes a heading 'Got a question? We're here to help!', a paragraph stating the helpdesk's purpose, and a list of contact channels: a Web forum, telephone (+33 970 440 522), Secretariat Support (+49 681-8575 5285), a Skype ID (CEF-AT-Helpdesk), and email (help@cef-at-helpdesk.org). A 'Web forum' button is located at the bottom of the content area.

Got a question? We're here to help!

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.

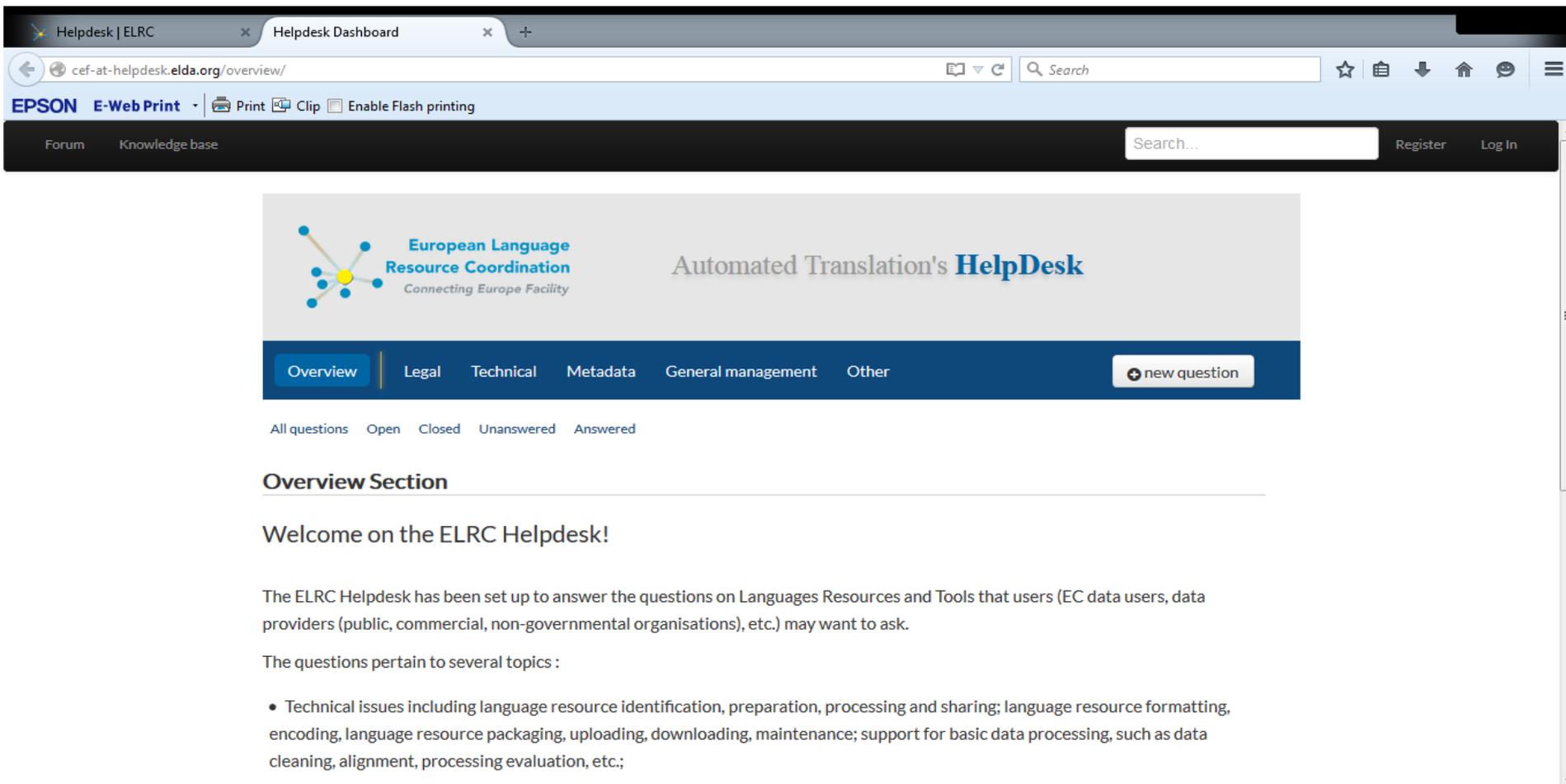
Please feel free to contact us through one of the following channels:

- a **Web forum**
- by telephone: **+33 970 440 522** This number is associated to a web-conferencing desktop (Skype). All requests will be handled in English.
- reach the Secretariat Support at **+49 681-8575 5285**.
- a Skype ID: **CEF-AT-Helpdesk**
- by email: **help@cef-at-helpdesk.org**

From the **Web forum**, please click the button « **+new question** » to enter your question. Upon receipt, your question will be processed and answered within 24 hours (simple query) to 5 days (complex query).

All questions and answers relevant for other users will be compiled in a FAQ. At the moment, all questions and answers can be accessed from **All questions**

[Web forum](#)



The screenshot shows a web browser window with two tabs: 'Helpdesk | ELRC' and 'Helpdesk Dashboard'. The address bar shows 'cef-at-helpdesk.elda.org/overview/'. The browser interface includes a search bar, navigation icons, and a toolbar with 'EPSON E-Web Print', 'Print', 'Clip', and 'Enable Flash printing' options. Below the browser, a dark navigation bar contains 'Forum', 'Knowledge base', a search input field, and 'Register' and 'Log In' links. The main content area features the ELRC logo and the title 'Automated Translation's HelpDesk'. A horizontal menu includes 'Overview' (selected), 'Legal', 'Technical', 'Metadata', 'General management', and 'Other', along with a '+ new question' button. Below the menu, there are filters for 'All questions', 'Open', 'Closed', 'Unanswered', and 'Answered'. The 'Overview Section' is titled 'Welcome on the ELRC Helpdesk!' and contains a paragraph explaining the helpdesk's purpose and a list of topics it covers.

Helpdesk | ELRC Helpdesk Dashboard

cef-at-helpdesk.elda.org/overview/

EPSON E-Web Print Print Clip Enable Flash printing

Forum Knowledge base Search... Register Log In

European Language Resource Coordination
Connecting Europe Facility

Automated Translation's HelpDesk

Overview Legal Technical Metadata General management Other + new question

All questions Open Closed Unanswered Answered

Overview Section

Welcome on the ELRC Helpdesk!

The ELRC Helpdesk has been set up to answer the questions on Languages Resources and Tools that users (EC data users, data providers (public, commercial, non-governmental organisations), etc.) may want to ask.

The questions pertain to several topics :

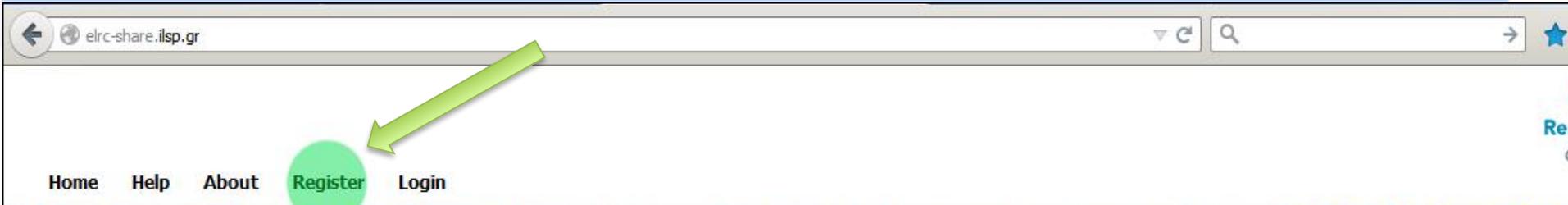
- Technical issues including language resource identification, preparation, processing and sharing; language resource formatting, encoding, language resource packaging, uploading, downloading, maintenance; support for basic data processing, such as data cleaning, alignment, processing evaluation, etc.;



Languages — the heart of
Multilingual Europe



- Dépôt ELRC (ELRC Repository)

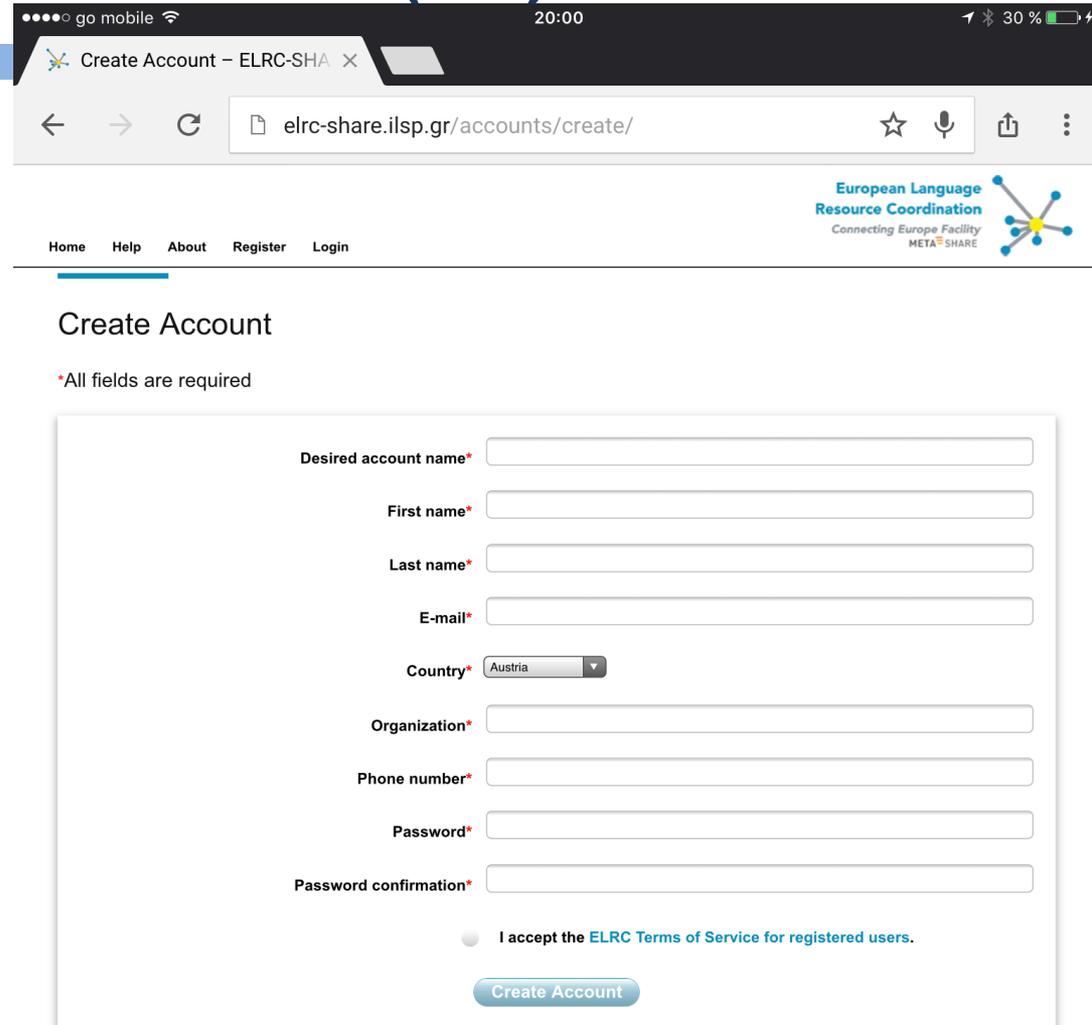


- Choisir "*Register*" pour créer un compte

Comment contribuer des données (2/7)



- Remplissez le formulaire
- Lisez les conditions d'utilisation et cliquez sur "Accepter"
- Cliquez sur le bouton Créer compte

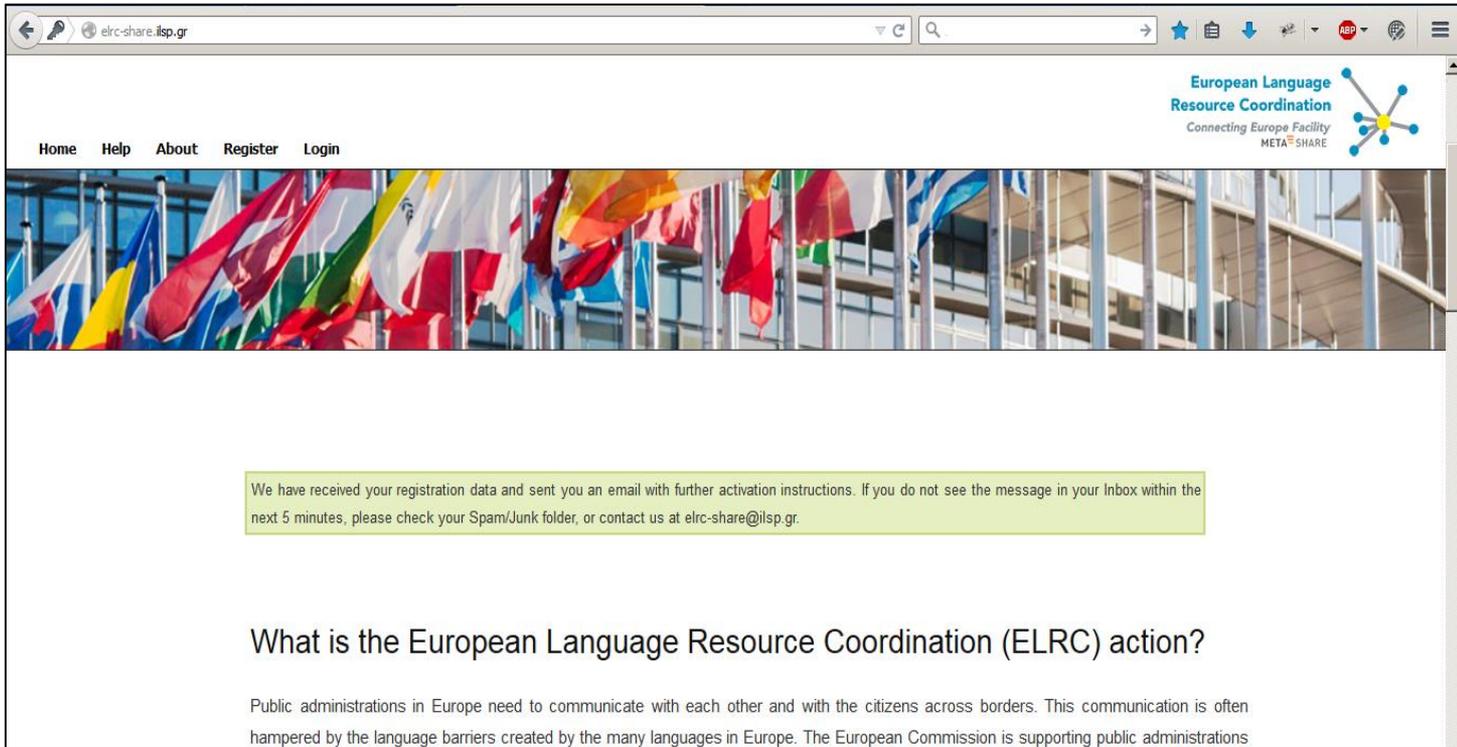


The screenshot shows a mobile browser interface for the 'Create Account' page of the ELRC. The browser address bar shows 'elrc-share.ilsp.gr/accounts/create/'. The page header includes the ELRC logo and navigation links: Home, Help, About, Register, Login. The main heading is 'Create Account' with a note '*All fields are required'. The form contains the following fields: 'Desired account name*', 'First name*', 'Last name*', 'E-mail*', 'Country*' (with a dropdown menu showing 'Austria'), 'Organization*', 'Phone number*', 'Password*', and 'Password confirmation*'. Below the form is a radio button for 'I accept the ELRC Terms of Service for registered users.' and a 'Create Account' button.

Comment contribuer des données (3/7)

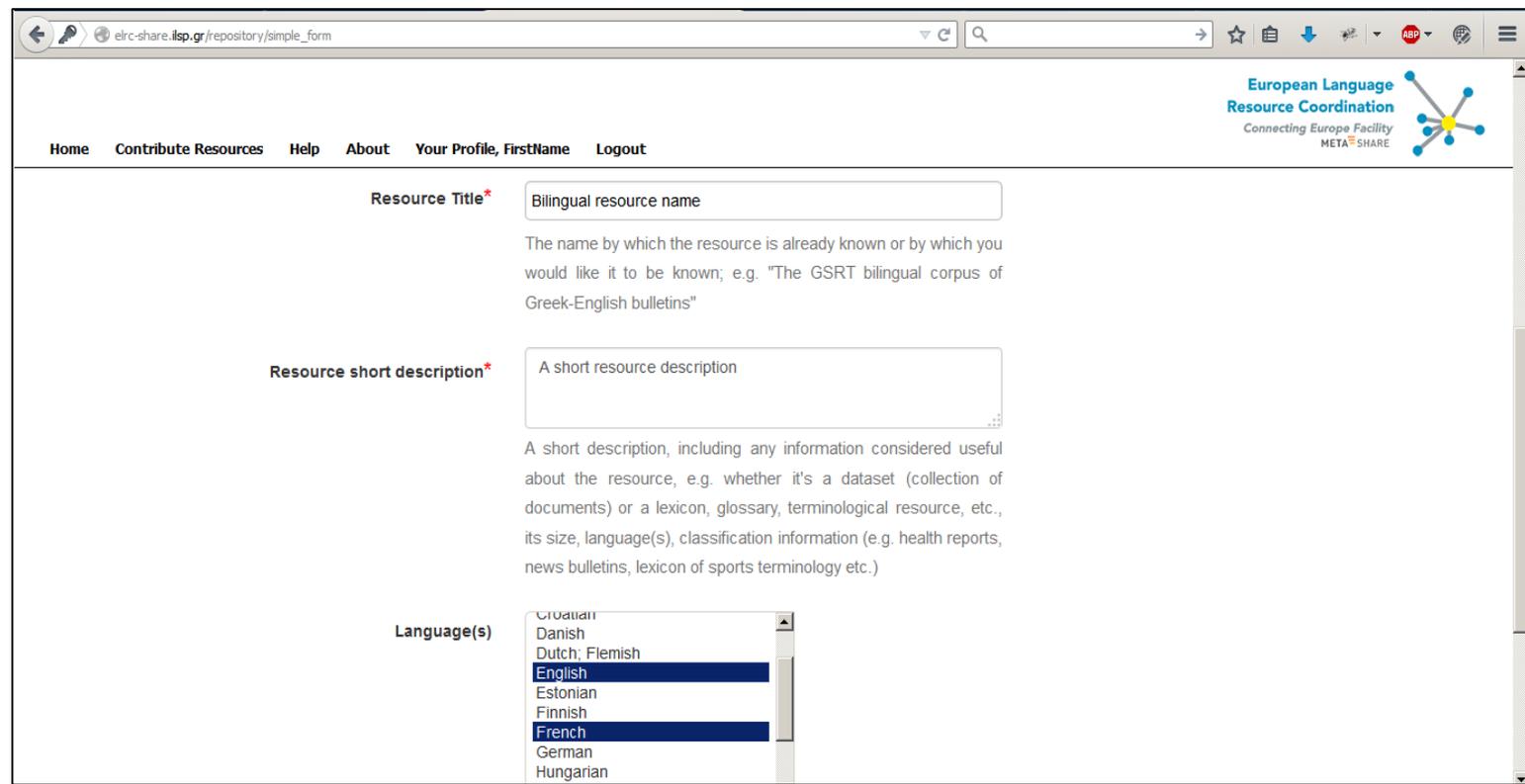


- Votre demande est reconnue et un e-mail d'activation est envoyé à l'adresse que vous avez indiquée
- Vérifiez votre email et cliquez sur le lien d'activation



The screenshot shows a web browser window with the URL `elrc-share.ilsp.gr`. The page header includes the ELRC logo and navigation links: Home, Help, About, Register, Login. Below the header is a banner image of various national flags. A green message box states: "We have received your registration data and sent you an email with further activation instructions. If you do not see the message in your Inbox within the next 5 minutes, please check your Spam/Junk folder, or contact us at elrc-share@ilsp.gr." Below this, the heading "What is the European Language Resource Coordination (ELRC) action?" is followed by a paragraph: "Public administrations in Europe need to communicate with each other and with the citizens across borders. This communication is often hampered by the language barriers created by the many languages in Europe. The European Commission is supporting public administrations

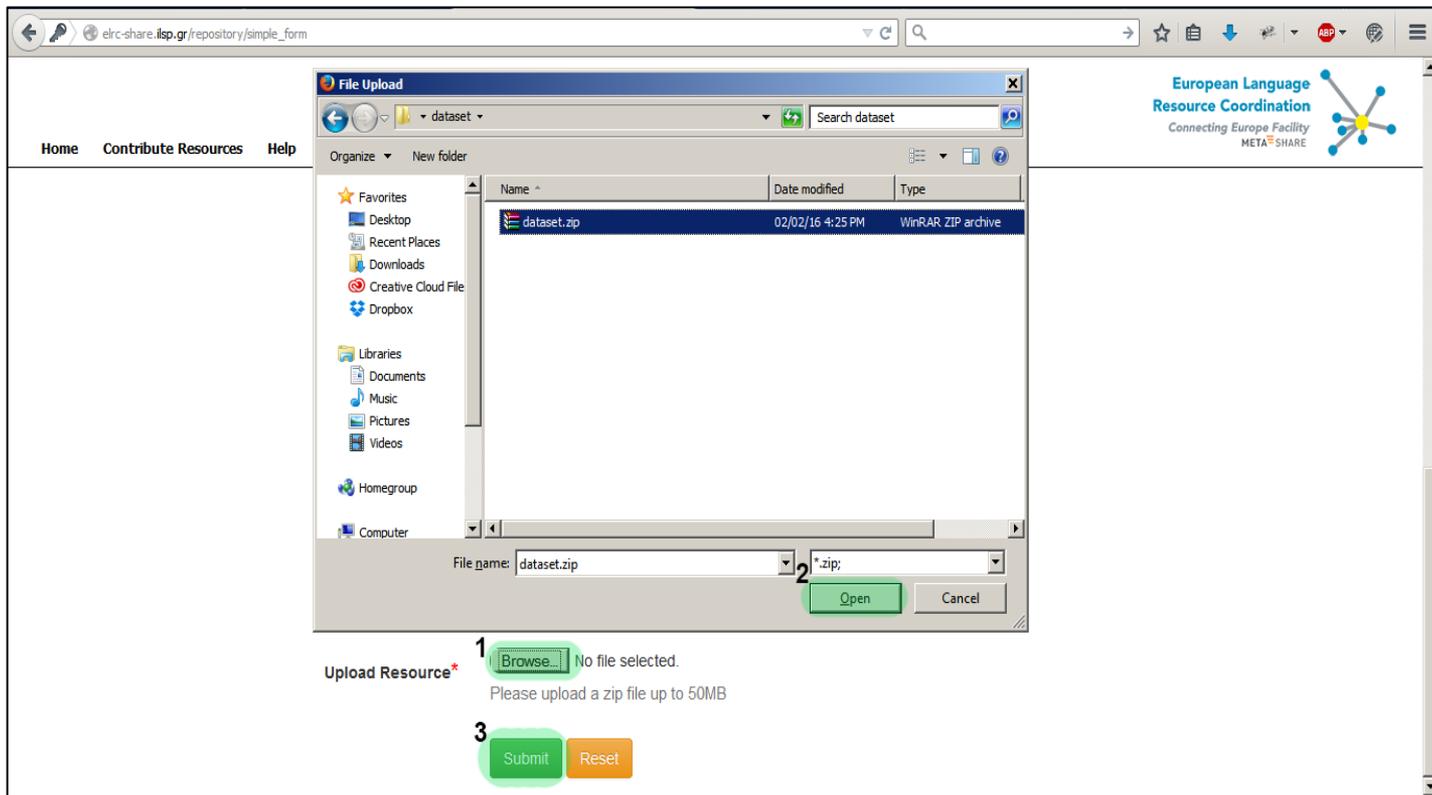
- Une fois sur le site, documenter l'ensemble des données



The screenshot shows a web browser window with the URL `elrc-share.ilsp.gr/repository/simple_form`. The page header includes the ELRC logo and navigation links: Home, Contribute Resources, Help, About, Your Profile, FirstName, and Logout. The main content area contains three form fields:

- Resource Title***: A text input field containing "Bilingual resource name". Below it is a description: "The name by which the resource is already known or by which you would like it to be known; e.g. 'The GSRT bilingual corpus of Greek-English bulletins'".
- Resource short description***: A text area containing "A short resource description". Below it is a 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)**: A dropdown menu with the following options: Croatian, Danish, Dutch; Flemish, English (highlighted), Estonian, Finnish, French (highlighted), German, and Hungarian.

- Choisir "Parcourir" sur votre ordinateur pour trouver le fichier .zip contenant vos données çà télécharger
- Cliquez sur "Sent"



1 Browse... No file selected.

Please upload a zip file up to 50MB

3 Submit Reset