European Language Resource Coordination
Connecting Europe Facility

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

Khalid Choukri (ELDA)

aperçu

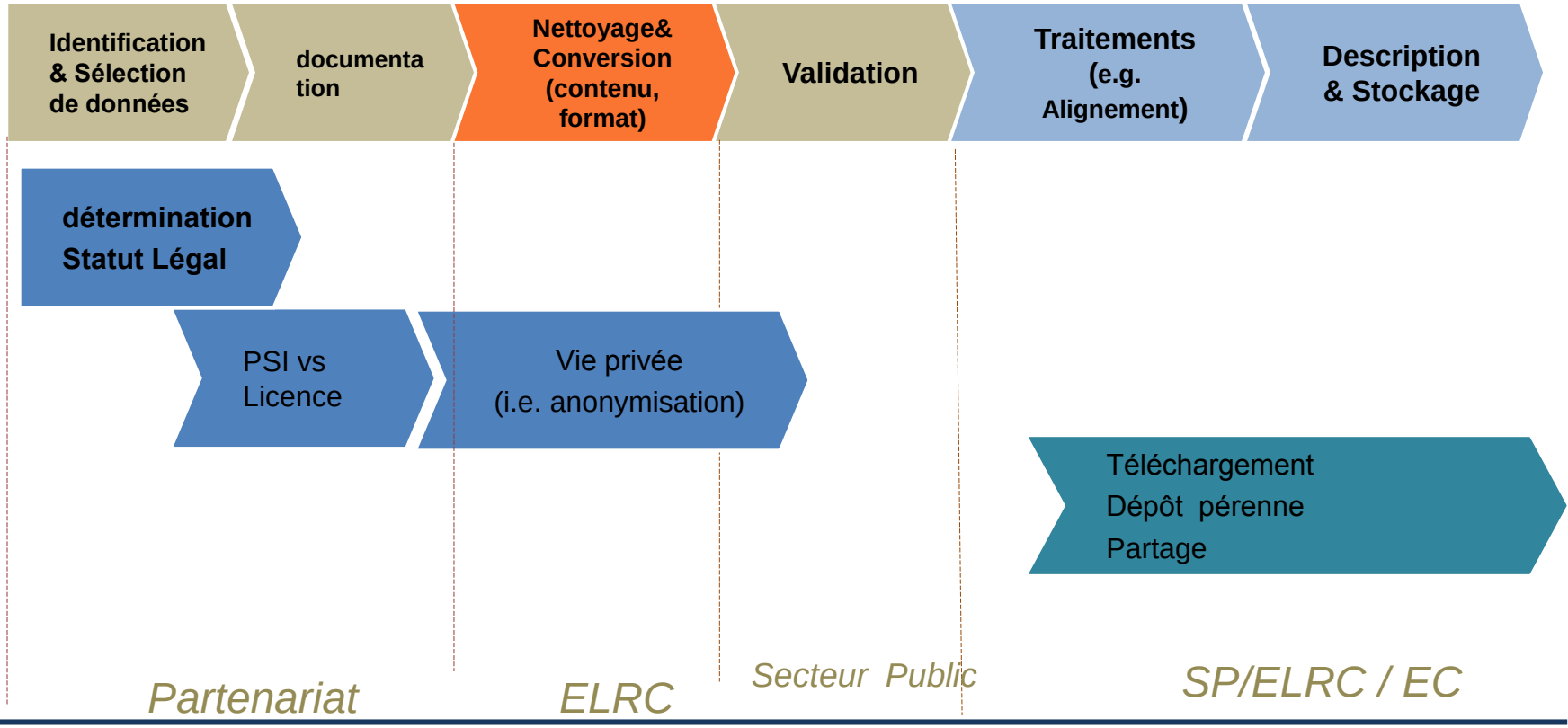
- Nous avons vu l'importance des données pour la Traduction Automatique
  - Apprentissage par les données (Data Driven Paradigm)
- Les données sont nécessaires dans toutes les langues (s)
- Où peut-on découvrir des données: Les acteurs du secteur public
  - Données Visibles par exemple des données Web (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

Illustration de la procédure flux de données

ex. Données existantes ➔RLs (Ressources Linguistiques)

la chaîne de valeur



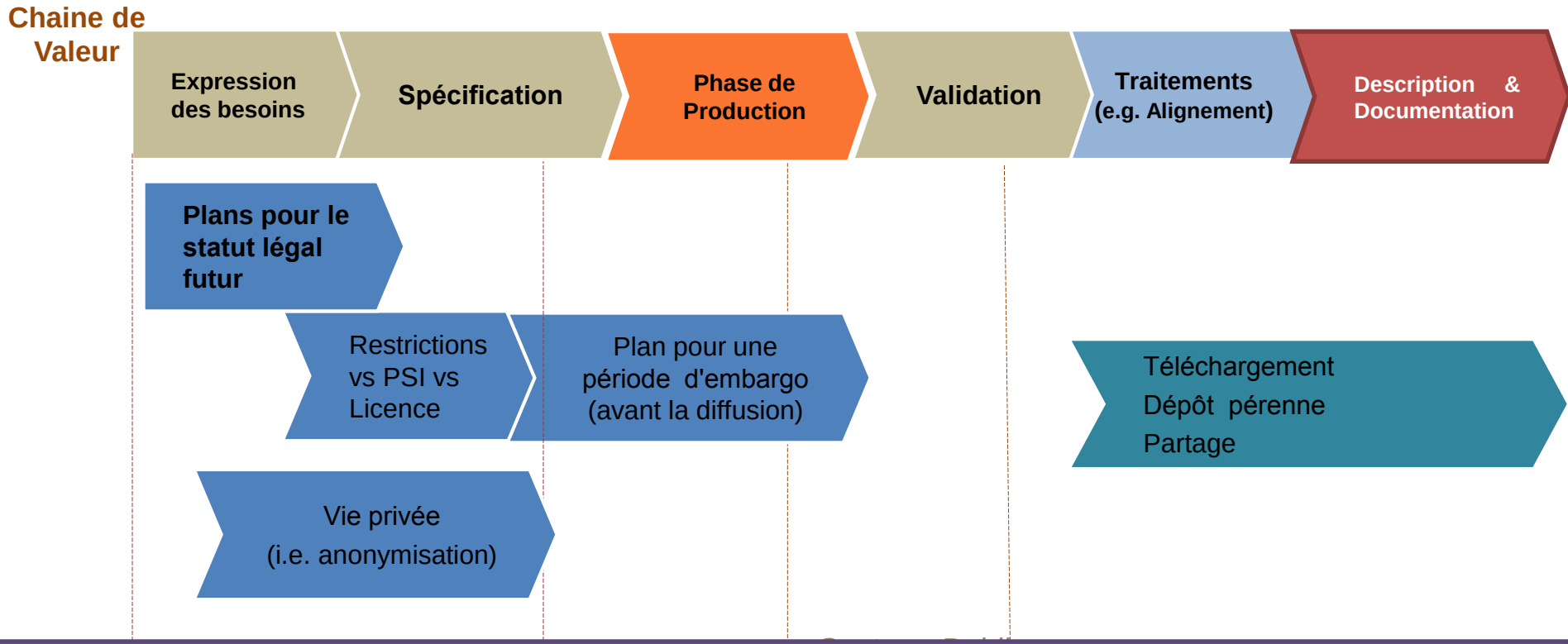
Des données nouvelles
Bonnes pratiques pour la gestion de données

- 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)

European Language Resource Coordination
Connecting Europe Facility

Illustration de la procédure flux de données

Données nouvelles➔RL (Ressources Linguistiques)

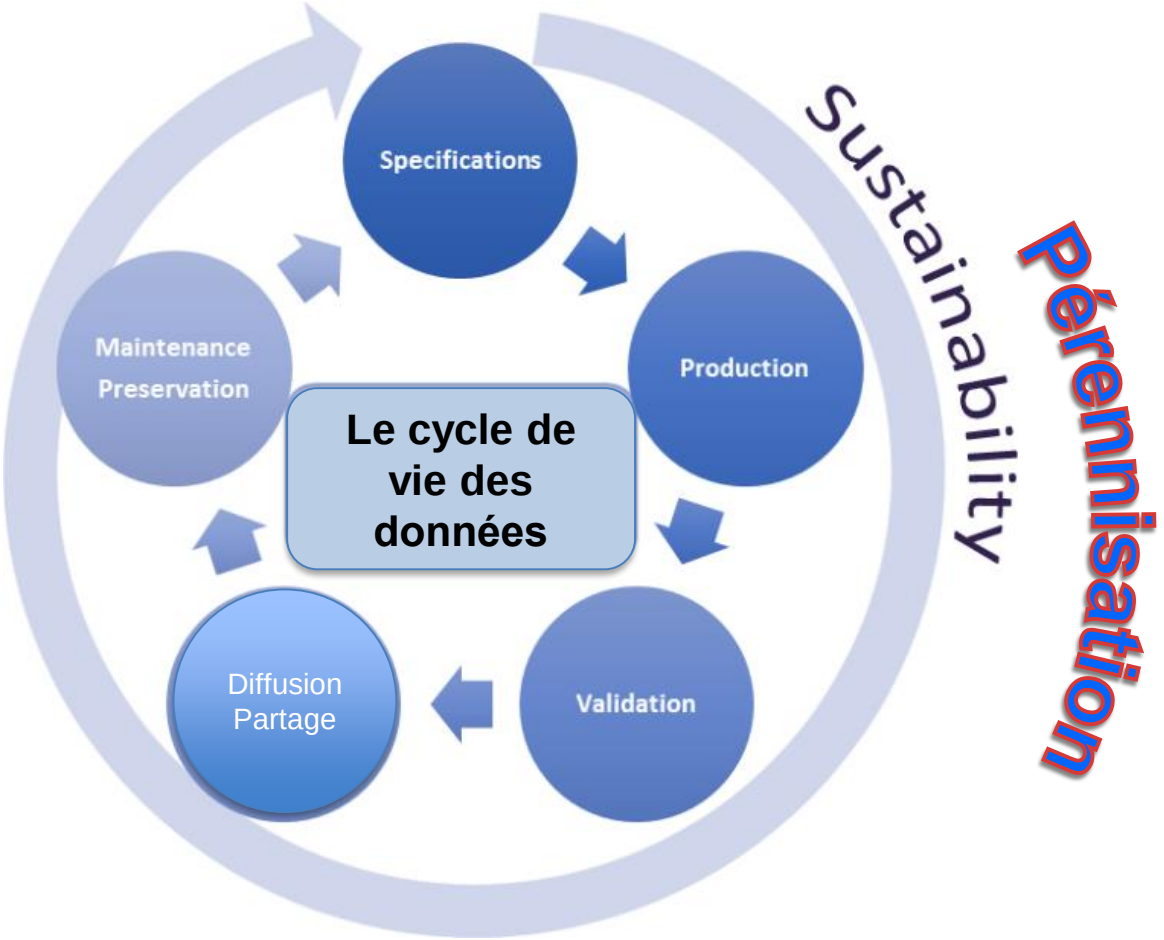


Plan de gestion de données // data management plan (DMP)

Etapes de Création d'un 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



Les principaux éléments d'un plan de gestion des 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)

Les principaux éléments d'un plan de gestion des 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)

Les actions de coopérations

- 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

Votre Implication est essentielle

- ✓ 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).

Conclusions

- 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" fourni 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)

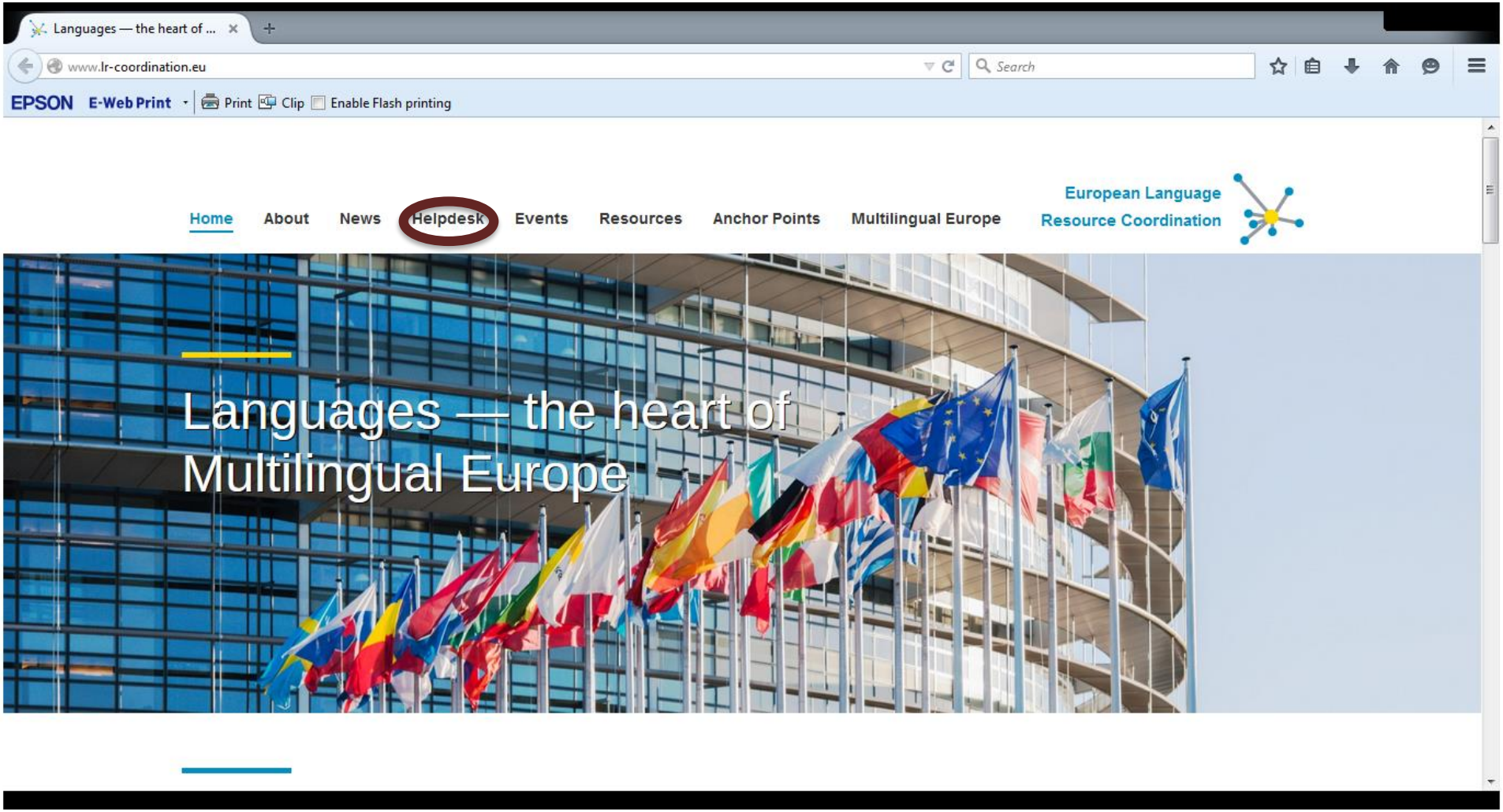
- **Ne pas sous-estimer la valeur de vos ressources linguistiques, prévoir un plan de gestion des données**

European Language Resource Coordination
Connecting Europe Facility

Helpdesk et Support

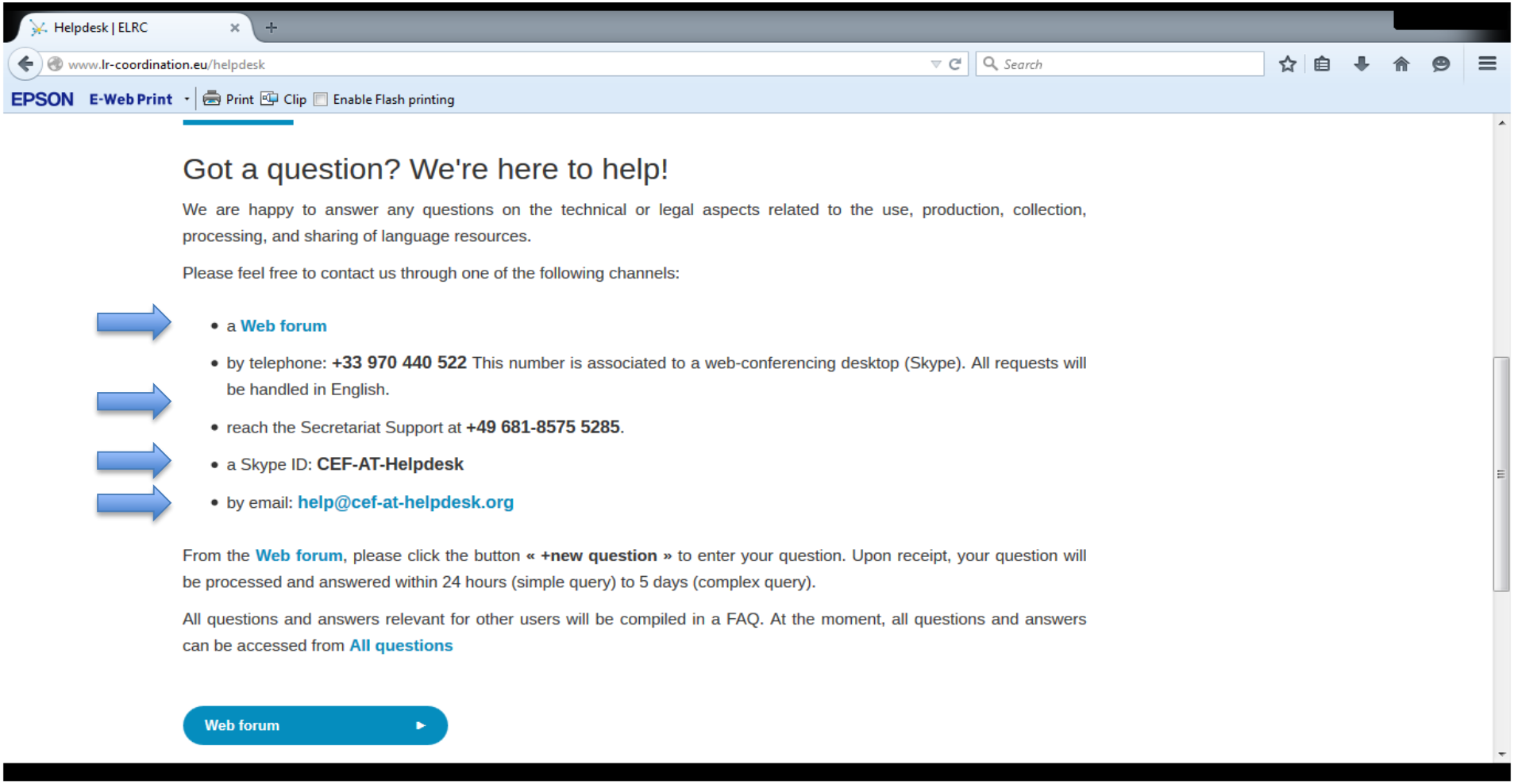
European Language Resource Coordination
Connecting Europe Facility

Portail ELRC
www.lr-coordination.eu

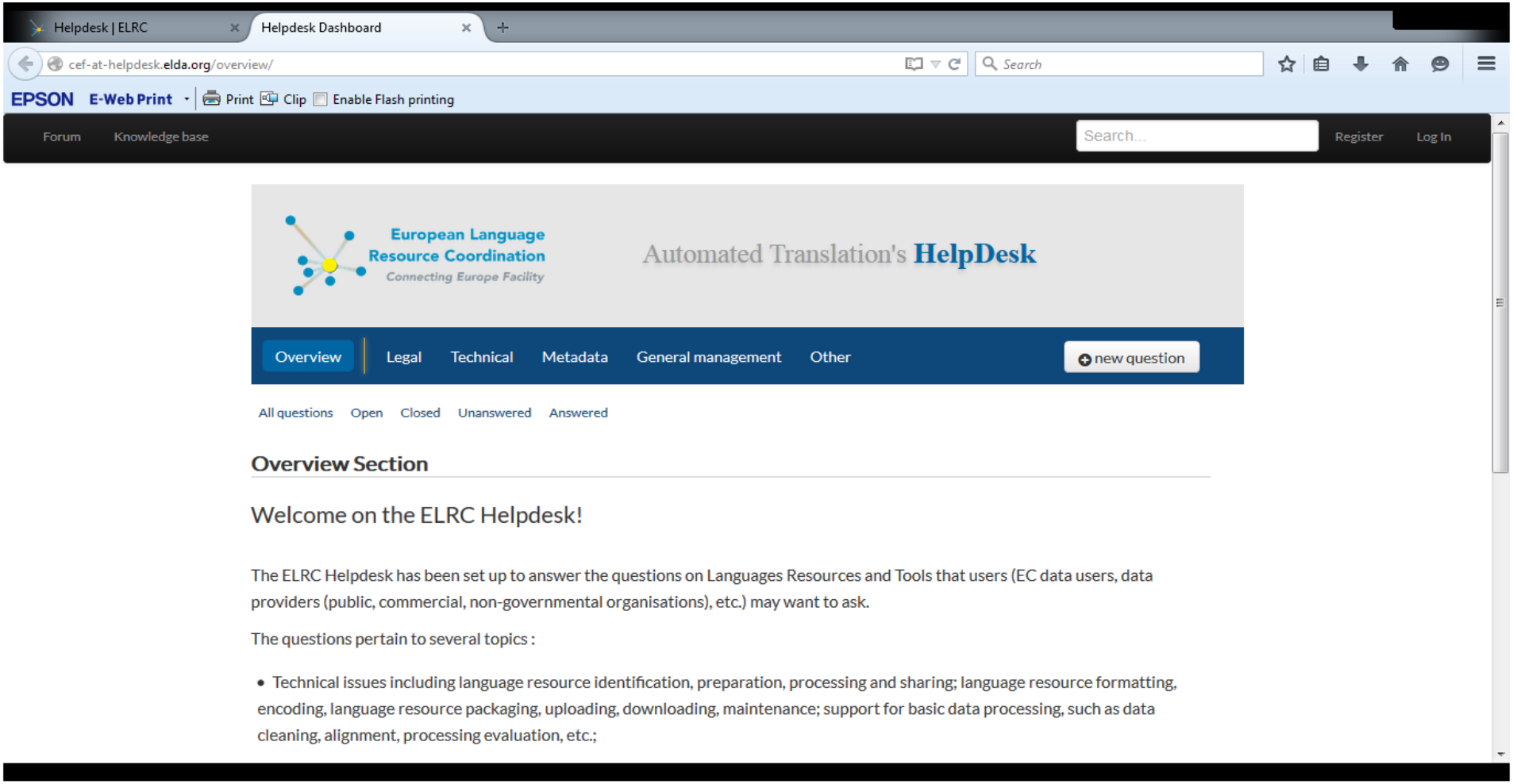


European Language Resource Coordination
Connecting Europe Facility

Portail ELRC

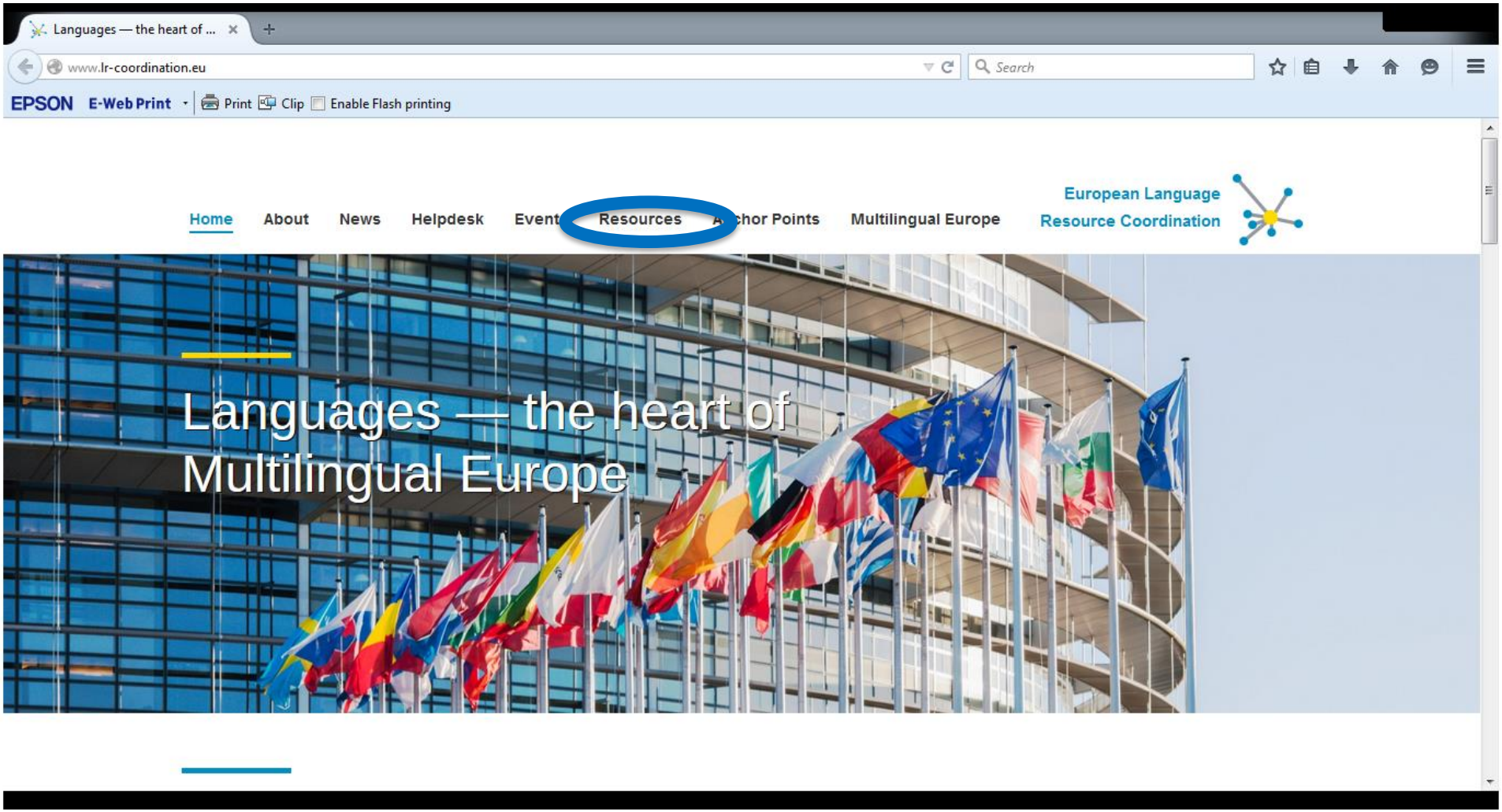


Portail ELRC , le Forum



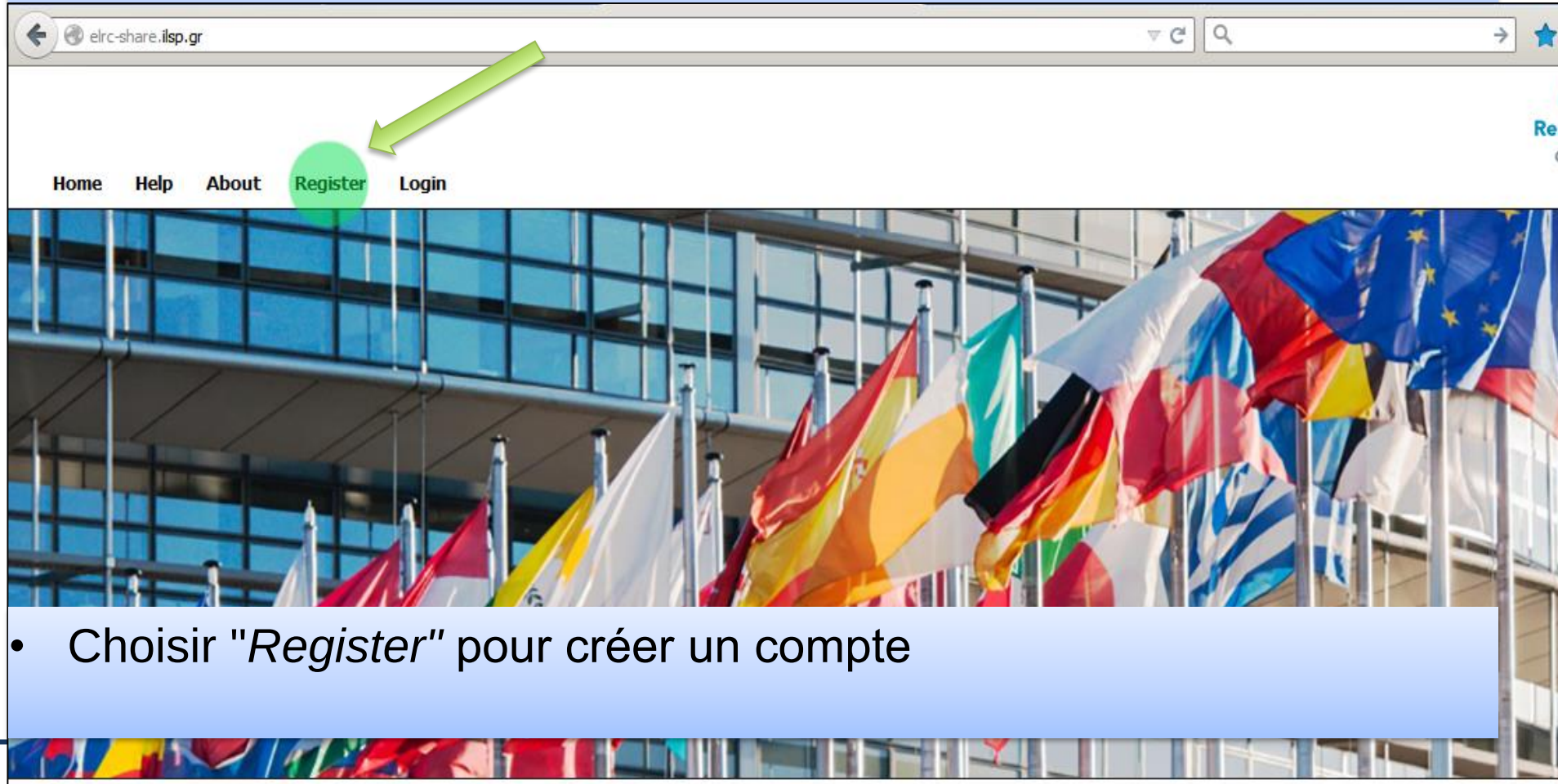
European Language Resource Coordination
Connecting Europe Facility

Portail ELRC … Ressources
www.lr-coordination.eu



Comment Contribuer Ressources Linguistiques (1/7)

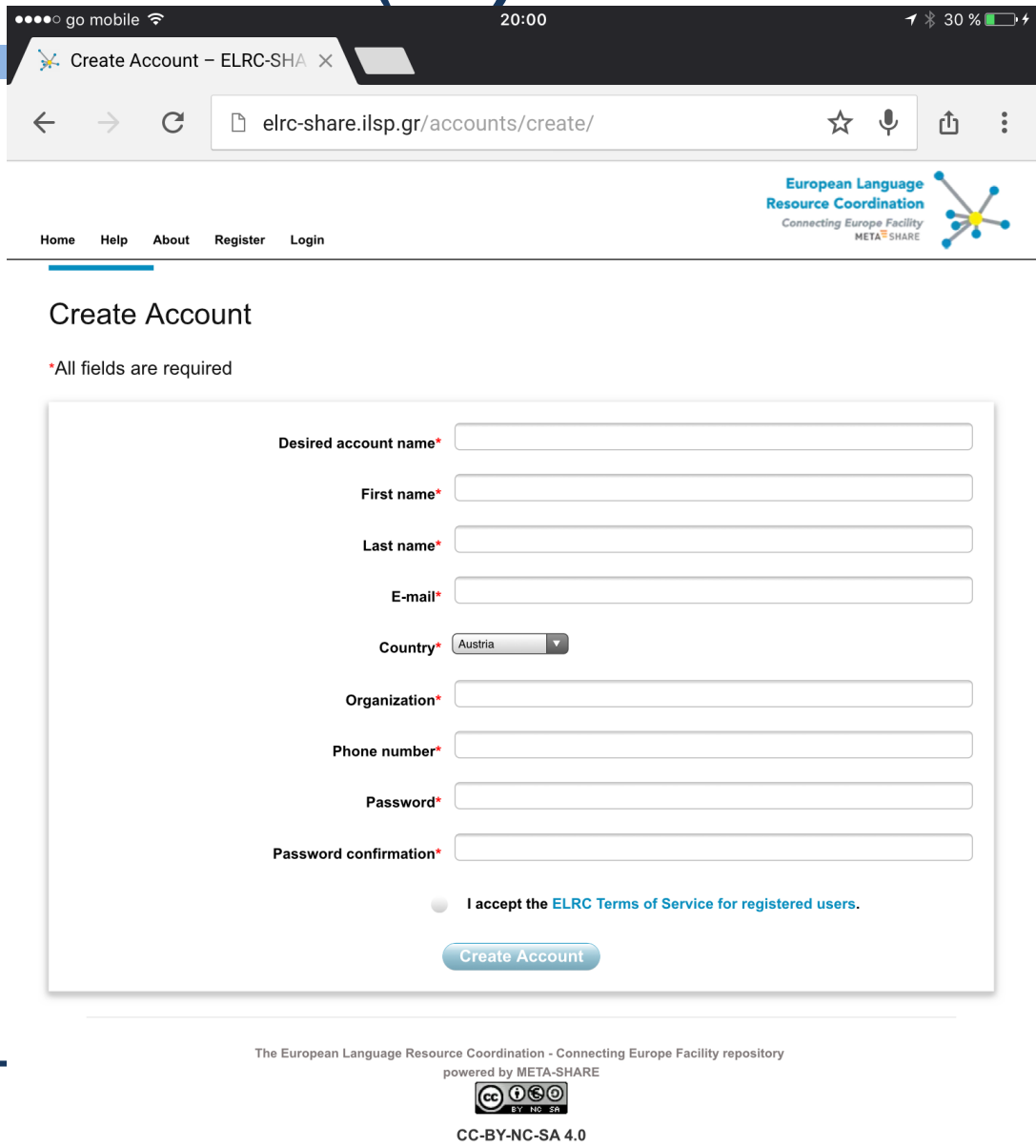
- Dépôt ELRC (ELRC Repository)



- Choisir "*Register*" pour créer un compte

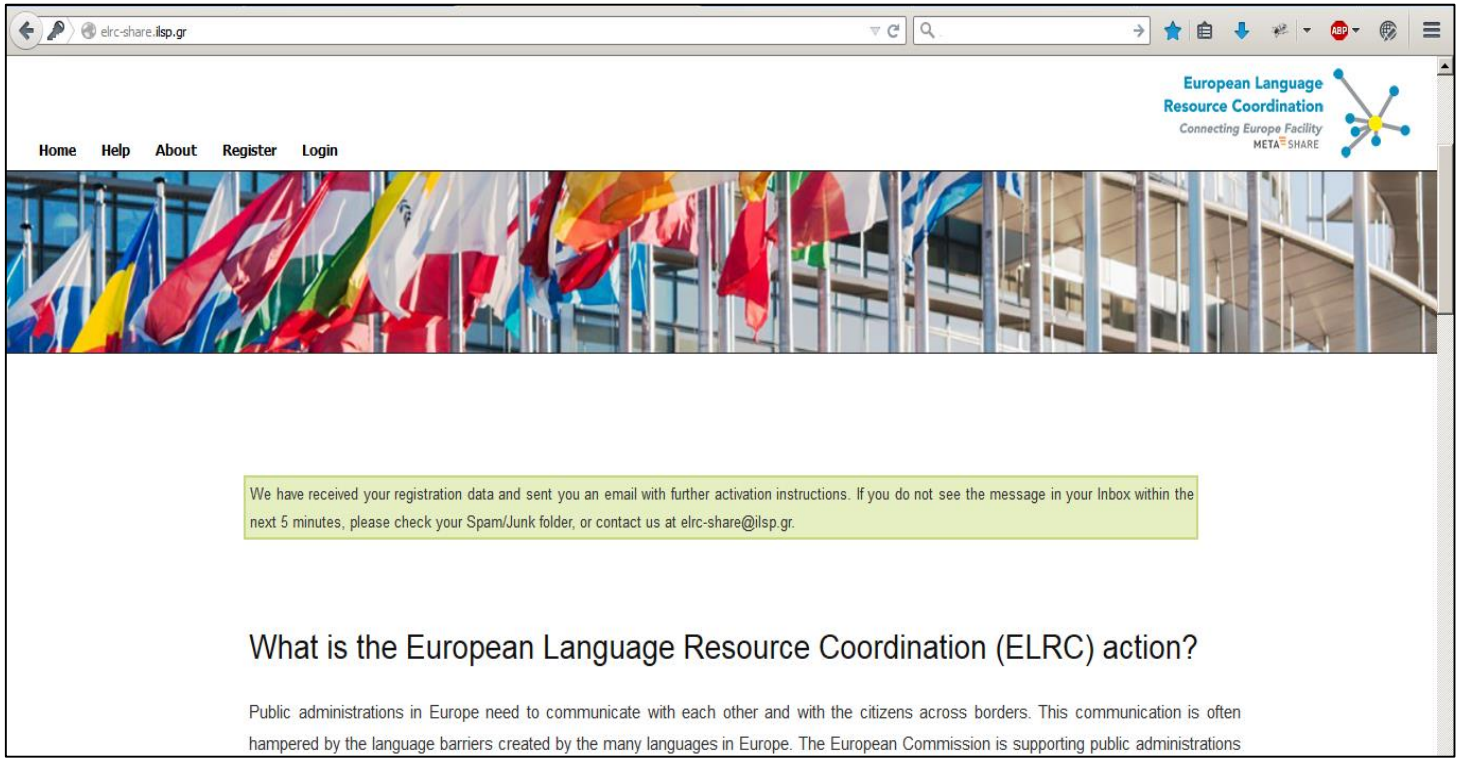
Comment contribuer des données (2/7)

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



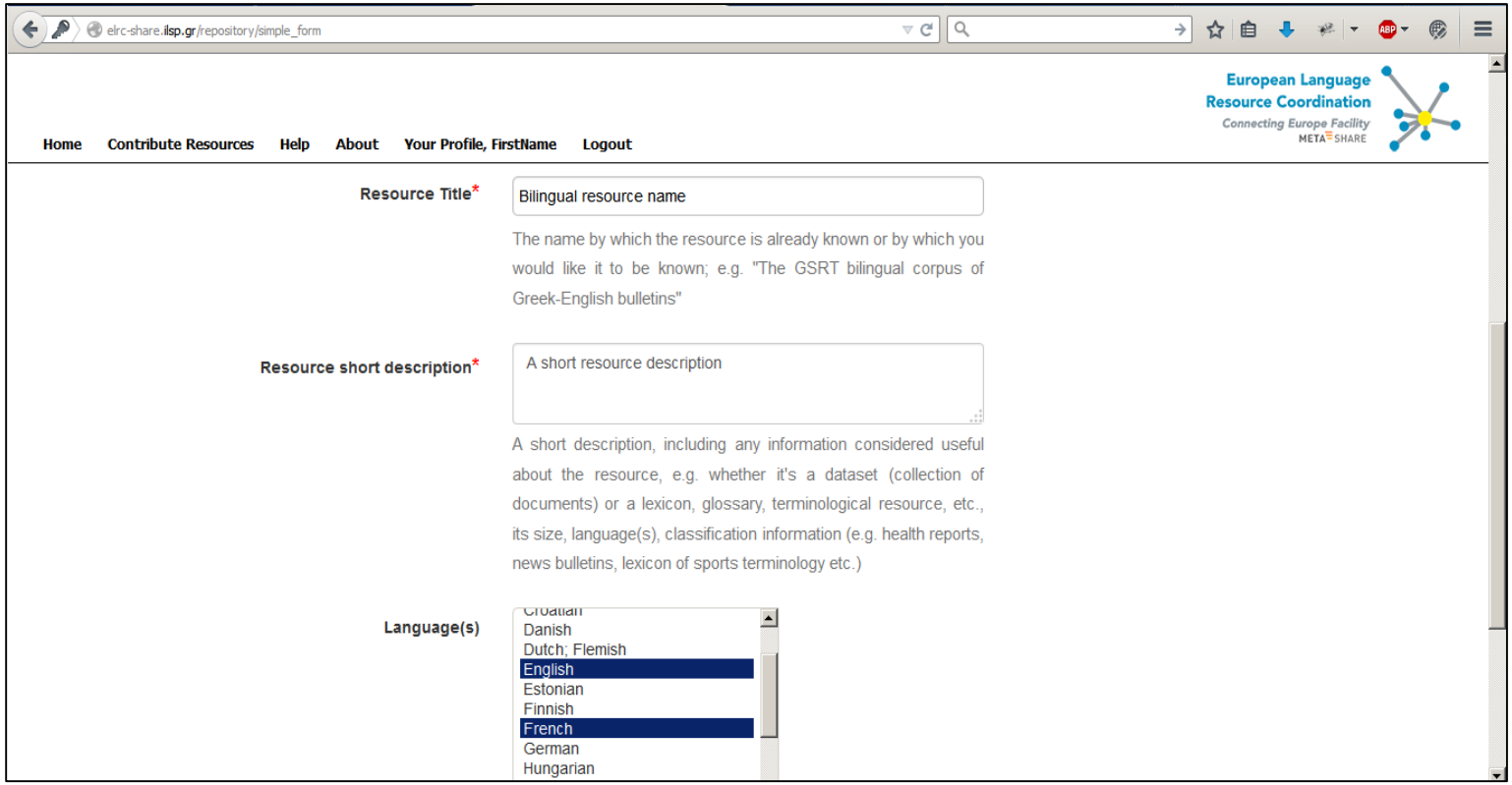
Comment contribuer des données (3/7)

- 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



Comment contribuer des données (5/7)

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



Comment contribuer des données (6/7)

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