European Language Resource Coordination
Connecting Europe Facility

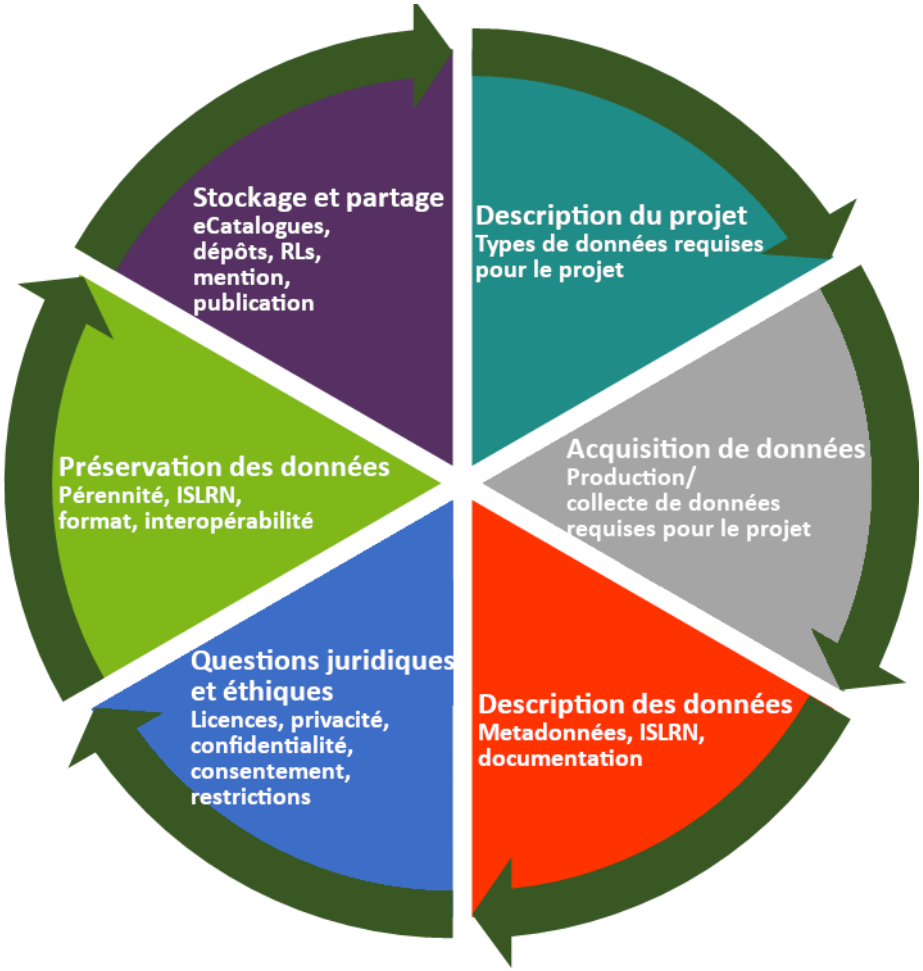
Atelier ELRC en France

Identifier et gérer vos données : questions-réponses

Khalid Choukri et Victoria Arranz
ELDA

Plan de gestion des données (PGD)

Le **PGD** décrit comment les données doivent être traitées pendant et après le flux de production. Il couvre l'ensemble du cycle de vie des données et définit une **politique de gestion** efficace et pérenne des données



Préoccupations liées à la création d'un PGD

- **Anticiper tous les problèmes juridiques potentiels**
  - Assurez-vous que la chaine des droits de propriété intellectuelle de vos données est validée
  - Assurez-vous que les parties productrices respectent votre droit de "propriété" (ex. relations avec les prestataires de services de traduction : assurez-vous que vous conservez tous les droits)
  - Assurez-vous que tous les documents intermédiaires produits sont à vous (p. ex. mémoires de traduction)
  - Vérifiez à l'avance les questions relatives à la protection de la vie privée et planifiez l'anonymisation si nécessaire
- **Définissez votre plan de gestion par rapport à la tâche**
  - Cela doit tenir compte de l'objectif principal (p. ex. la rédaction de documents, la traduction de documents, etc.)
- **Plan de réaffectation (des documents aux RLs)**
  - Demander des données dans un format utilisable (non seulement PDF mais aussi TMX/XML)
  - Assurez-vous que vos données utilisent des supports à jour (pas de CD ?)
- **Prévoyez la publication et le partage futurs** en tant qu'information du secteur public

European Language Resource Coordination
Connecting Europe Facility

Questions?

Questions-réponses : droits d'auteur

Si un organisme public sous-traite la traduction d'un texte dont il détient les droits, à qui appartient le droit d'auteur de la version traduite ? La traduction peut-elle être partagée ?

Cela dépend de ce que le contrat d'externalisation établit en matière de droits de propriété intellectuelle. Les agences publiques devraient veiller à ce que le contrat d'externalisation leur accorde le droit de réutiliser et de partager librement les mémoires de traduction.

Questions-réponses : droits d'auteur

J'ai créé un corpus de textes littéraires pour mes recherches. Puis-je en faire don à ELRC ?

La chaine de droits de tous les textes inclus dans le corpus doit être validée. Certaines d'entre eux, en particulier les œuvres anciennes, peuvent être dans le domaine public (par exemple, si le droit d'auteur a expiré). Pour le reste, une licence doit être obtenue auprès des ayants droit autorisant la redistribution à des tiers.

Questions-réponses : protection de données

Que faisons-nous de données contenant des données personnelles ?

Toutes les données personnelles ne doivent pas être anonymisées.
Si vous avez des doutes sur la manière de traiter des ensembles de données contenant des données personnelles, contactez l'équipe du ELRC. ELRC offre un service d'assistance juridique ainsi qu'un service d'anonymisation pour les données collectées.

Questions-réponses : protection de données

J'ai une collection de documents bilingues de mon organisation du secteur public qui sont accessibles au public, p. ex. des manifestations d'intérêt, des appels d'offres, etc. Ils comprennent les noms de personnes, p. ex. les noms des directeurs et des membres des comités. Sont-elles soumises aux restrictions en matière de données à caractère personnel ? Devrait-on les rendre anonymes ?

Il s'agit d'activités de la fonction publique qui n'appartiennent donc pas au domaine privé.

Questions-réponses : traitement de données

Nous disposons de données, mais nous n'avons pas les ressources nécessaires pour identifier les données pertinentes et les traiter.

ELRC peut vous aider à identifier les ensembles de données pertinents. Il offre également des services de traitement linguistique aux administrations publiques (conversion de données, suppression d'étiquettes, reformatage, nettoyage, alignement, validation de métadonnées, etc.). Un service d'assistance sur place est également proposé pour fournir une assistance technique. Ces services sont gratuits.

Questions-réponses : traitement de données

Nous avons une grande collection de fichiers PDF scannés. Pouvons-nous demander de l'aide sur place ? Allons-nous récupérer des formulaires lisibles par la machine (résultat de la reconnaissance optique de caractères ?

Les résultats de cette reconnaissance sur des fichiers PDF scannés diffèrent en qualité (selon les langues, la qualité du papier, etc.). Certains d'entre eux peuvent être utiles pour un traitement ultérieur afin d'obtenir des textes lisibles par machine, et peuvent donc être traités ultérieurement par ELRC pour produire des corpus parallèles.
ELRC offre un service d'assistance sur place.

Questions-réponses : pérennité des données

La plupart de nos données sont numériques (p. ex. National Bank, Statistics office) et accompagnées de quelques textes. Vous en avez encore besoin ?

ELRC se concentre principalement sur les données textuelles. Cependant, si votre jeu de données numériques contient du texte, cela peut quand même être utile, surtout dans le cas de textes bilingues ou multilingues.

Questions-réponses : pérennité des données

Un extrait de notre corpus national est disponible dans un dépôt différent, par exemple CLARIN. Dois-je également contribuer à ELRC ?

Seulement s'il s'agit de parties différentes du corpus (ELRC a accès aux ensembles de données des centres de données).

European Language Resource Coordination
Connecting Europe Facility

Merci de votre attention

Courriel : info@lr-coordination.eu
Site web: www.lr-coordination.eu