Ortolang
Open Resources and Tools for LANGuage
Outils et Ressources pour un Traitement Optimisé de la LANGue

Quels critères pour une ressource linguistique pour mieux être prise en compte dans les processus d'évaluation ?

Jean-Marie.Pierrel@atilf.fr

ORTOLANG bénéficie d'une aide de l'État au titre du programme « Investissements d'avenir » (ANR-11-EQPX-0032)

Pourquoi une telle question (1) ?

- La notion de ressources linguistiques (corpus, lexiques, dictionnaires, outils de traitement) est aujourd'hui incontournable en humanités numériques et spécifiquement en linguistique et en TAL ;
- Or la constitution et la normalisation de corpus ou de ressources de qualité sont très couteuses
- Sans une véritable mutualisation chaque équipe de recherche se verrait dans l'obligation de tout réinventer
- Une telle mutualisation passe par le dépôt sur des plateformes institutionnelles facilement accessibles pour
  - Ne pas perdre de nombreuses ressources
  - Permettre leurs réutilisations faciles par d'autres

Pourquoi une telle question (2) ?

- Pour aider cette mutualisation, il faut mieux valoriser l'effort de constitutions de corpus qui nécessité
  - un investissement humain en temps de chercheurs ou d'ingénieurs important
  - Un investissement financier (en budget consolidé) important pour les laboratoires supports

⇒ Comment permettre de mieux prendre en compte cette tâche de constitution de corpus dans l'évaluation

  - des chercheurs
  - des laboratoires

⇒ La réponse à cette question passe nécessairement par la définition de critères d'évaluation de ces ressources

Les questions à se poser

1. Quelles ressources doivent être déposées ?
2. Comment décrire une ressource ?
3. Quelles formes de dépôt de ces ressources ?
4. Quels critères de qualité d'un corpus ou d'une ressource ?

Dans le suite :

j'aborderai rapidement les points 1 et 2

Puis je me focaliserai sur les points 3 et 4

Quelles ressources doivent être déposées ?

- Toutes ressources fruits d'un travail de recherche méritent d'être pérenniser au travers d'un dépôt sur une plateforme institutionnelle.
- Pourquoi ?
  - Certes une ressource en recherche est réalisée dans un objectif précis
    - C'est particulièrement vrai pour les corpus
  - Mais, sous certaines conditions, elle doit pouvoir être réutiliser dans d'autres cadres
- La question est plus celle de ses critères de qualité

Comment décrire une ressource ?

- Bien sûr au travers de métadonnées normalisées
  - Cf. intervention de Carole Etienne
- Mais de plus sur une plateforme de dépôt ou de mutualisation, il convient de respecter les recommandations de cette plateforme :
  - Description,
  - Référence,
  - Auteurs et contributeurs,
  - …

Où déposer des ressources pour assurer cette mutualisation ?

- Un dépôt sur sa machine (ou dans un coin d'une machine de son labo) n'assure aucune pérennité de la ressource développée
  - Trop de ressources financées sur budget public ont ainsi été perdues !
- Pourquoi privilégier des plateformes institutionnelles ?
  - Pour assurer la pérennité de la ressource
  - Parce qu'une équipe de gestion de cette plateforme assure cette pérennisation
  - Pour faciliter les recherches par d'autres car elles capitalisent un ensemble important de ressources

Que peut-on attendre d'une plateforme institutionnelle de dépôt (1) ?

- Une procédure simple de dépôt.
  - Le dépôt d'une ressources ne doit pas être un parcours du combattant !
  - Avec si possible l'appui d'équipes compétentes pour aider et orienter le déposant
- Des procédures de recherche et de découverte permettant des accès diversifiés aux ressources
- Des procédures de sauvegarde sécurisée solides
- Une gestion des droits d'accès aux ressources
  - Le respect des droits (d'auteur, d'éditeurs et des personnes) fait que beaucoup de ressources issues de la recherche ne peuvent être réutilisées que dans ce cadre.

Que peut-on attendre d'une plateforme institutionnelles de dépôt (2) ?

- Des procédures transparentes de catalogage et d'indexation à l'international pour mieux valoriser et signaler les ressources
  - A minima production et gestion de métadonnées OAI-PMH
- Des identifiants pérennes pour les ressources
  - Rien de pire qu'un message indiquant « la page n'existe plus »
- Une procédure d'archivage pérenne indispensable pour les ressources qu'on ne pourrait pas recréer

Un exemple d'une telle plateforme

- Dans le domaine de la langue, l'Equipex ORTOLANG répond à ces exigences de plateforme de dépôt institutionnelle

Ortolang
Open Resources and Tools for LANGuage
Outils et Ressources pour un Traitement Optimisé de la LANGue

  - Service spécialisé pour la langue complémentaire de l'offre généraliste proposée par Huma-Num
  - Partie prenante du consortium CORLI

Quels critères de qualité d'une ressource ?

- Tous les ressources (corpus) ne sont pas équivalentes
- elles diffèrent par
  - Leur taille
  - Leur rareté
  - Leur réutilisabilité
  - Leur diffusion
  - Leur contenu (ressources brutes versus ressources annotées)
- Analysons par ordre d'importance (selon moi) ces critères qu'il est souvent difficile d'objectiver.

Critère 1 : la disponibilité

- En lecture (simple ou au travers d'une interface spécialisée) : permet d'accéder à la ressource
- En téléchargement : permet d'exploiter la ressource

⇒Importance d'identifiants pérennes et de leur résolution

⇒Cette disponibilité peut se faire moyennant éventuellement identification de l'utilisateur

  - Ortolang gère 4 types de disponibilités :
    - Pour tous
    - Pour les utilisateurs connectés
    - Pour les membres de l'ESR
    - Pour les membres de l'espace de travail lié à la ressource

Critère 2 : la réutilisabilité

- Cela passe par
  - Des métadonnées standardisées.
  - Un codage des données respectant les standards utilisés par la communauté
  - Une documentation minimale

⇒Eviter les formats propriétaires pour les données

⇒Expliciter le codage des annotations, si possible en rendant disponible le guide d'annotation utilisé

Critère 3 : la rareté et le coût de définition de la ressource

- Ce qui est rare et cher
- Ce qui est « Rare » : des ressources que l'on ne peut pas reproduire à l'identique
  - Exemple des corpus « véritables », oraux ou journalistiques par exemple
  - ESLO 1, Est républicain….
- Ce qui est « cher » est difficilement accessible par les chercheurs
  - Par « cher » il faut entendre un coût important de définition ou d'annotation de la ressource
  - pas un coût d'achat car on se situe ici dans le domaine des ressources partagées

Critère 4 : la réutilisation effective par d'autres

- Nécessité de disposer de statistiques d'usage
  - Dans Ortolang les déposants disposeront de telles statistiques : nb d'accès, nb de téléchargement
- Nécessité de citer les ressources que l'on utilise
  - Offrir un mode de citation clair, stable et unifié aux utilisateurs (compatible avec les structures de citations des articles
  - Respecter une charte d'utilisation demandant de citer les ressources utilisées
- *Je pense que ce sera sous ces conditions que petit à petit des ressources pourront être prise en considération dans l'évaluation des chercheurs ou des laboratoires comme des articles*