Interopérabilité et métadonnées quels besoins dans un projet de recherche ?

Carole Etienne, Sept 2016

Les corpus dans les projets de recherche

- Constat : les projets de recherche concernent des corpus existants et impliquent plusieurs sources de données
- A venir, plus de projets impliquant corpus oraux et corpus écrits (écrits non planifiés)
- En entrée, au moins un "Work Package" dédié à la mise en commun de données … pourtant déjà décrites et annotées
- En sortie, de nouvelles annotations délivrées dans différents formats suivant les outils

Exemple du projet Orféo

- Corpus de Français Parlé Parisien (S. Branca, F. Lefeuvre, M. Pires, S. Fleury)
- FLEURON (Vie étudiante – V. André)
- Corpus TUFS (Y. Kawaguchi)
- VALIBEL (A.-C. Simon)
- French oral narrative (J. Carruthers)
- Entretiens (S. Caddeo, J.-M. Debaisieux)
- Réunions de travail (M. Husianycia)
- Corpus de référence du Français parlé (DELIC)
- CORALROM (DELIC/Cresti/Moneglia)
- Corpus de Français Parlé à Bruxelles (A. Dister)
- CLAPI (V. Traverso)
- OFROM (M.-J. Béguelin, M. Avanzi, F. Démioz)
- TCOF (V. André, C. Benzitoun, E. Canut, J.-M. Debaisieux)

ORAL : 3.5 millions de mots
350 heures de données

Projet Orféo : les métadonnées

- Très hétérogènes tant au niveau du format
  - Fichier texte (Word)
  - Fichiers tabulaires (Excel, CSV)
  - XML (OLAC, TEI Header, CMDI)

- … que du contenu
  - champs basiques : durée, âge, lieu , nom, ..
  - critères subjectifs :
    - qualité, niveau de langue : bon/moyen/mauvais
    - niveau de spontanéité
    - nature, catégorie, domaine

Projet Orféo : l'enregistrement et la transcription

- Mutualisation de l'existant
  - Signal : vidéo, audio mp3/wav, pause, alignement à la transcription
  - Transcription : tours de parole imbriqués, orthographe, production langagière *vs* descriptions/rires/pauses, convention

    ➔ homogénéiser pour pouvoir annoter
  - Anonymisation vérification manuelle: noms, lieux, chiffres d'affaires en réunion de travail ➔ script D.Hirst pour conserver le signal

IRCOM : Table ronde de juin 2014

- Les besoins et les difficultés rencontrées
  - une transcription de base " niveau 0 " orthographique (pauses) réutilisable
  - un jeu de métadonnées commun
  - la citation obligatoire de la ressource

  - problèmes d'anonymisation (long et peu satisfaisant)
  - signal : problèmes de qualité, temps de téléchargement, format**s**
  - prise en main des logiciels de transcription
  - pas d'indications claires sur la personne à contacter si les données ne sont pas accessibles

Métadonnées : différentes fonctions

- Distinguer différents niveaux
  - Documentation : décrire les corpus **et** aider à l'analyse des résultats
  - Requêtes : constitution de sous-corpus -> quels critères
  - Annotations : enrichir les corpus de nouvelles annotations
  - Archivage

Un **format commun** pour les métadonnées et la transcription

- **format autonome d'échange de données utilisable pour des recherches en linguistique**, pas seulement dédié à l'archivage d'une ressource
- unique et normalisé
- paramétrable et évolutif
- déjà utilisé dans des projets concernant des corpus oraux et multimodaux
- portée européenne
- format pivot entre les logiciels d'annotations (Elan, Praat, Transcriber etc.)
- conçu pour être exploité par des outils automatiques d'annotations, des outils de requêtes , des outils de visualisation

Métadonnées : Structure

Métadonnées de niveau 0 : critères de recherche ou informations

Métadonnées expliciter le contexte

Métadonnées techniques signal audio/video

Métadonnées connaissances en langues des locuteurs/apprentissage

Métadonnées sociolinguistiques

Métadonnées annotations : conventions, outils (semi-)automatiques, versions

Métadonnées objet corpus : théorie, analyses, bibliographie

Métadonnées juridiques

Métadonnées : le format TEI

- **Choix**
  - un seul fichier pour les métadonnées et la transcription
  - utilisé par les projets alipe, clapi, ciel-f, colaje, …
  - personnalisation de son jeu de balises et documentation (ODD)

- **Granularité** suivant le niveau de définition minimaliste vs complète pour les métadonnées (speaker, setting) ou transcription (ex: utterance vs phonème)

- **Lien** avec les autres formats : Dublin Core, Imdi (rhapsodie), …

- **Portée** groupe européen ISO-TEI, listes de diffusion de la Tei

- Utilisation par les corpus de l'**écrit**

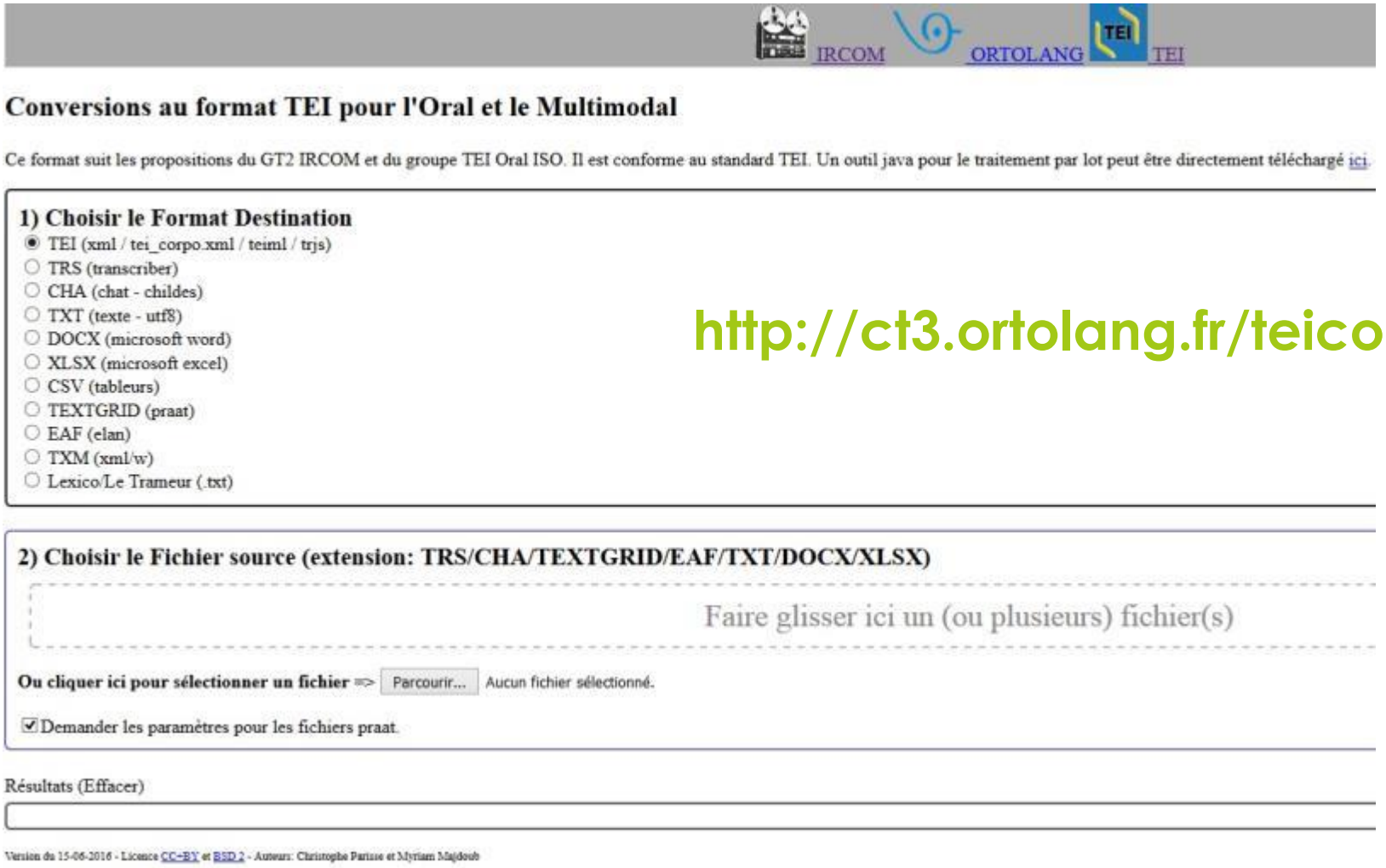
Le format TEI : Les métadonnées

- Le niveau commun "niveau 0"
  - Informations générales sur le corpus: citation, diffusion, version,…
  - Informations sur les données primaires: enregistrement, collecte des textes
  - Informations sur les données secondaires: transcription, annotations
  - Informations sur les locuteurs : nombre, âge, profil sociolinguistique…
- Vocabulaire contrôlé
- Personnalisation ODD, un schéma de données

Les métadonnées : les outils

- Outil(s) de saisie des métadonnées
  - en ligne, dans une interface web guidée et documentée, sans installer aucun logiciel avec téléchargement du résultat, duplication, …
  - éditeur TEI : schéma et le mode auteur d'oXygen
  - en exportant directement en TEI les métadonnées d'une banque de données existantes (à la charge du producteur de données)

L'interopérabilité des transcriptions



http://ct3.ortolang.fr/teiconvert/

Diffusion

- Journées TEI
  - communication aux journées TEI de novembre 2015
  - article dans le journal de la TEI (JTEI)
- Métadonnées et catalogage, Poitiers (Juin 2016)
- Soumission d'un atelier au colloque FLORAL (Mars 2017)
- Formation Archivage ET Interopérabilité (2017)

  CORLI : InterExplo-Corli et Dépôt conservation évaluation et diffusion
- Journées Linguistique de Corpus (2017)