# Vers un jeu commun de métadonnées : quels besoins pour la recherche ?

CORpus Langues Interactions

Huma-Num

Carole Etienne, Nov 2017

# Métadonnées : différentes fonctions

- Distinguer différents niveaux
  - Orientées **recherche**
    - 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
  - Orientées archivage

# Construire son corpus d'études à partir de corpus existants

- Des besoins pourtant très simples : quelques exemples
  - des enregistrements en vidéo chez l'enfant de moins de six ans
  - des conversations privées d'adultes
  - des réunions de travail
  - des échanges entre non natifs
  - des appels téléphoniques entre jeunes

  - en prosodie : des enregistrements en audio chez l'enfant pour des études prosodiques : qualité, bruit de fond, …
  - en diachronie : des repas de famille enregistrés à une génération d'écart

# Construire son corpus d'études à partir de corpus existants

- Comment procéder ?
  - Recherche dans les plateformes d'archivage: quels critères ?
    - date d'enregistrement
    - auteur
    - version
    - nom
    - responsable, collecteur, transcripteur, …
    - …

# Construire son corpus d'études à partir de corpus existants

- Comment procéder ?
  - Recherche dans des sources spécialisées ou identifiées
    - Childes
    - autres corpus : attention, un jeu de critères par plateforme ou un fichier excel pour sélectionner directement certains des enregistrements du corpus
    - dans quel format sont les données : enregistrement, transcriptions ?
    - quelles annotations sont disponibles ?

# Construire son corpus d'études à partir de corpus existants

- Pour les corpus les plus anciens ➔ contenu connu à un instant donné
  - volume de données : alimentation ?
  - audio *vs* vidéo dans les enregistrements plus récents
  - qualité hétérogène
  - nature des enregistrements
    - enquêtes *vs* conversations
    - privé / professionnel
    - adulte / enfant
    - langue de l'enregistrement
    - …
- Pour les corpus plus récents ➔ peu connus donc peu réutilisés

6

# Un corpus d'étude dans un projet de recherche

- Constat : les projets de recherche concernent des corpus existants et peuvent impliquer  plusieurs sources de données
- De plus en plus de projets impliquant corpus oraux et corpus écrits (écrits non planifiés)
- En début de projet, au moins un "Work Package" dédié à la mise en commun de données … pourtant déjà décrites et annotées
- En fin de projet, de nouvelles annotations délivrées dans différents formats suivant les outils comment l'indiquer dans les sources

# IRCOM : Table ronde de juin 2014

- Les besoins et les difficultés rencontrées
  - un jeu de métadonnées commun
  - la citation obligatoire de la ressource
  - anonymisation : où trouver l'information
  - signal : problèmes de qualité, temps de téléchargement, format**s**
  - transcriptions disponibles dans un format lié à un logiciel de transcription
  - pas d'indications claires sur la personne à contacter si les données ne sont pas accessibles

# Exemple du projet Orféo : C.Benzitoun, C.Etienne, L.Bérard

- 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 millions de mots**
**350 heures de données**

# Projet Orféo : les métadonnées

- Très hétérogènes tant au niveau du format
  - Fichier texte (pdf, 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
    - contenu : nature, catégorie, domaine

# 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**

Pour les transcriptions

- format pivot pour les transcriptions 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

11

# Métadonnées : Structure

Métadonnées de niveau 0 : critères de recherche ou informations

Des modules additionnels suivant la thématique de recherche

- 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

12

# Métadonnées : le format TEI

- **Choix**
  - un seul fichier pour les métadonnées et la transcription
  - utilisé par les projets alipe, clapicolaje, , ciel-f, …
  - 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**

13

# Le format TEI : Les métadonnées

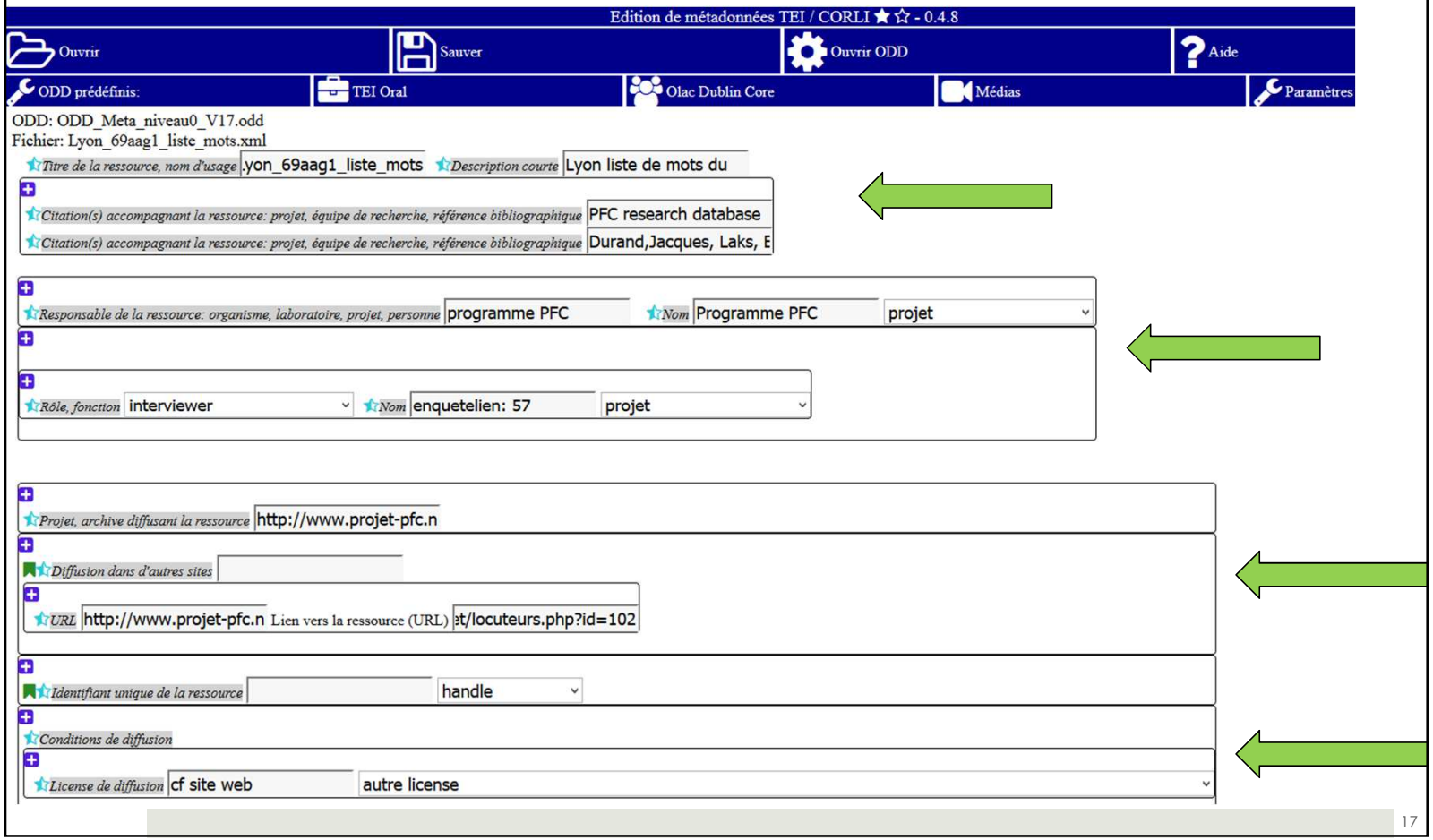
- Le niveau commun "niveau 0"
  - Informations générales sur le corpus: **citation**, **diffusion**, version,…
  - Informations sur les données primaires: enregistrement, collecte des textes
  - Informations sur les données secondaires: transcription, **annotations**
  - Informations sur les locuteurs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

# Les métadonnées : les outils

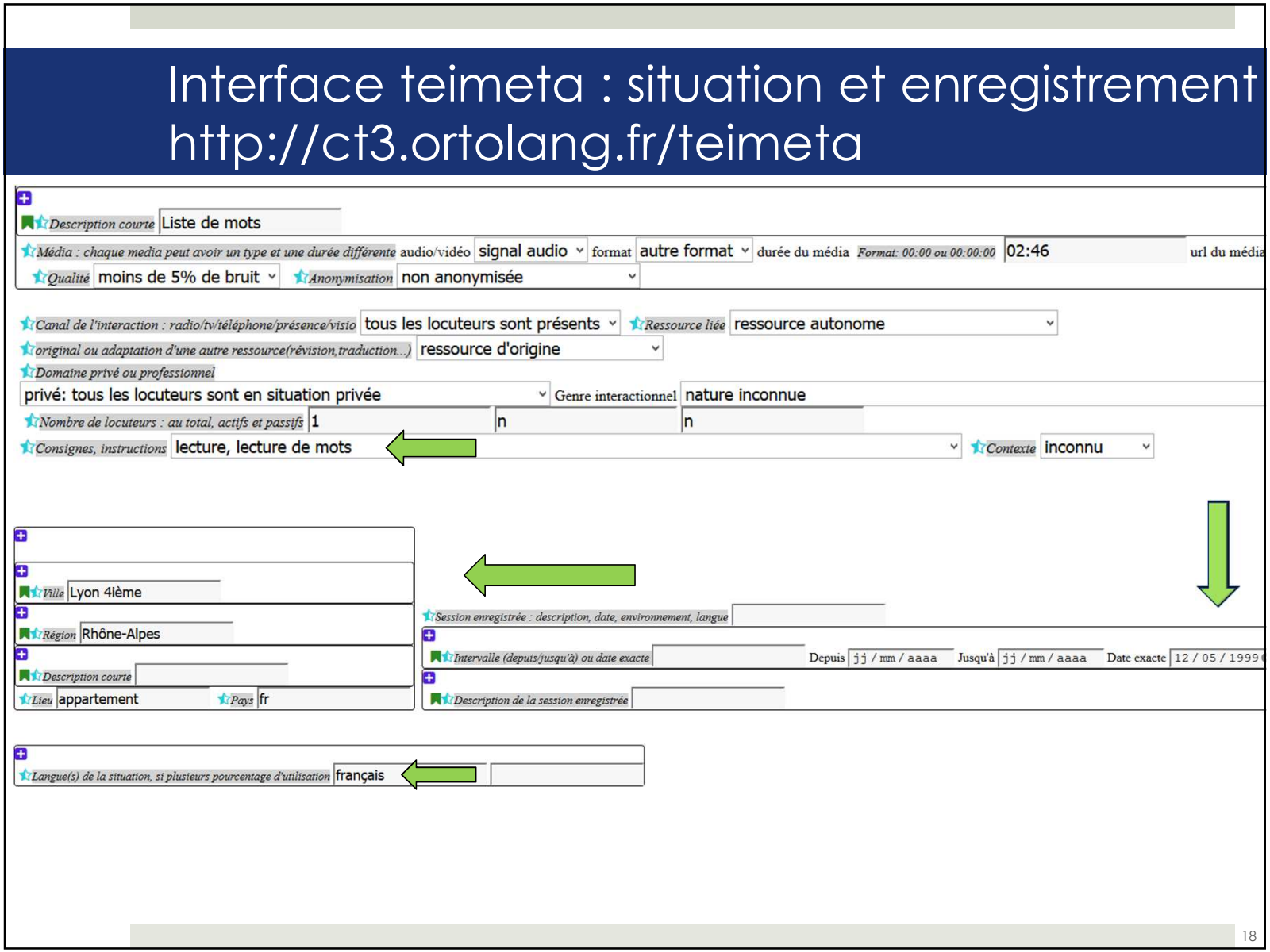
- Outil(s) de saisie des métadonnées
  - en ligne, dans une interface web guidée et documentée, sans installer aucun logiciel avec téléchargement du résultat, duplication, …
  - éditeur TEI : schéma et le mode auteur d'oXygen
  - en exportant directement en TEI les métadonnées d'une banque de données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

# teiMeta : la liste de mots

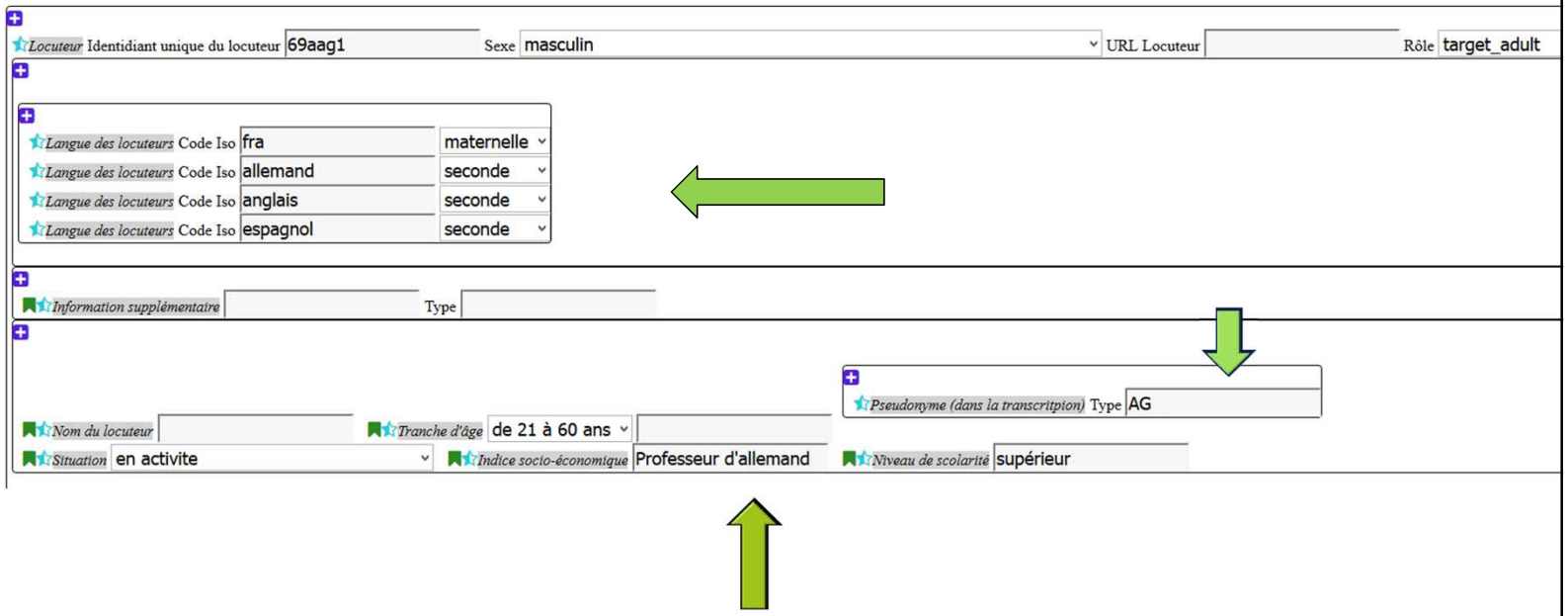
# Interface teimeta : la ressource http://ct3.ortolang.fr/teimeta



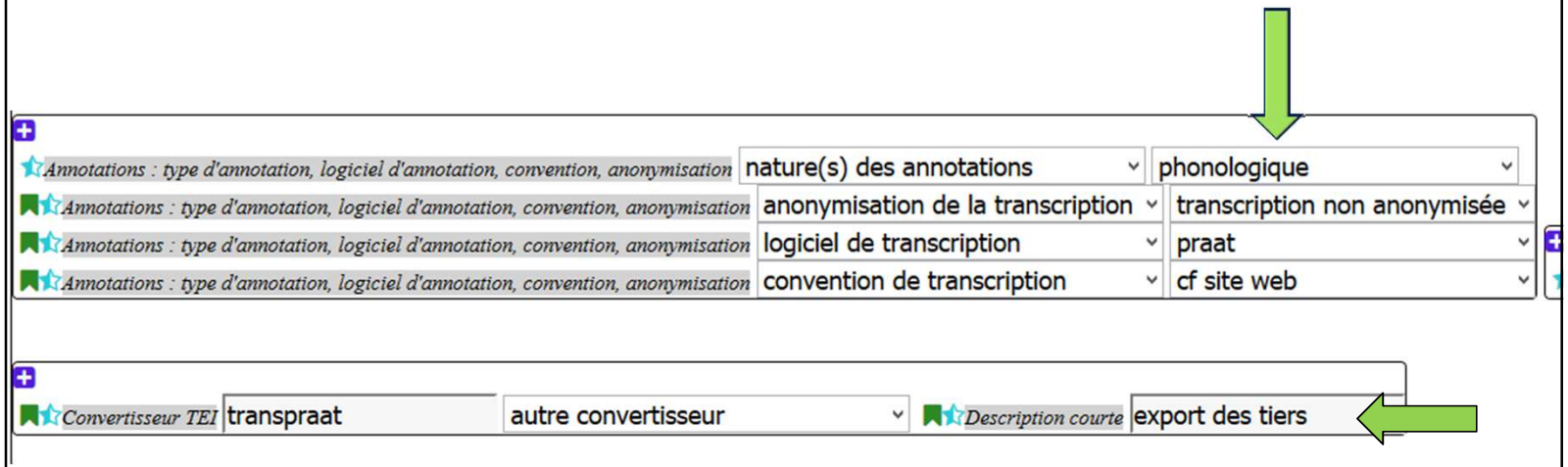
# Interface teimeta : situation et enregistrement
# http://ct3.ortolang.fr/teimeta



# Interface teimeta : le locuteur
# http://ct3.ortolang.fr/teimeta



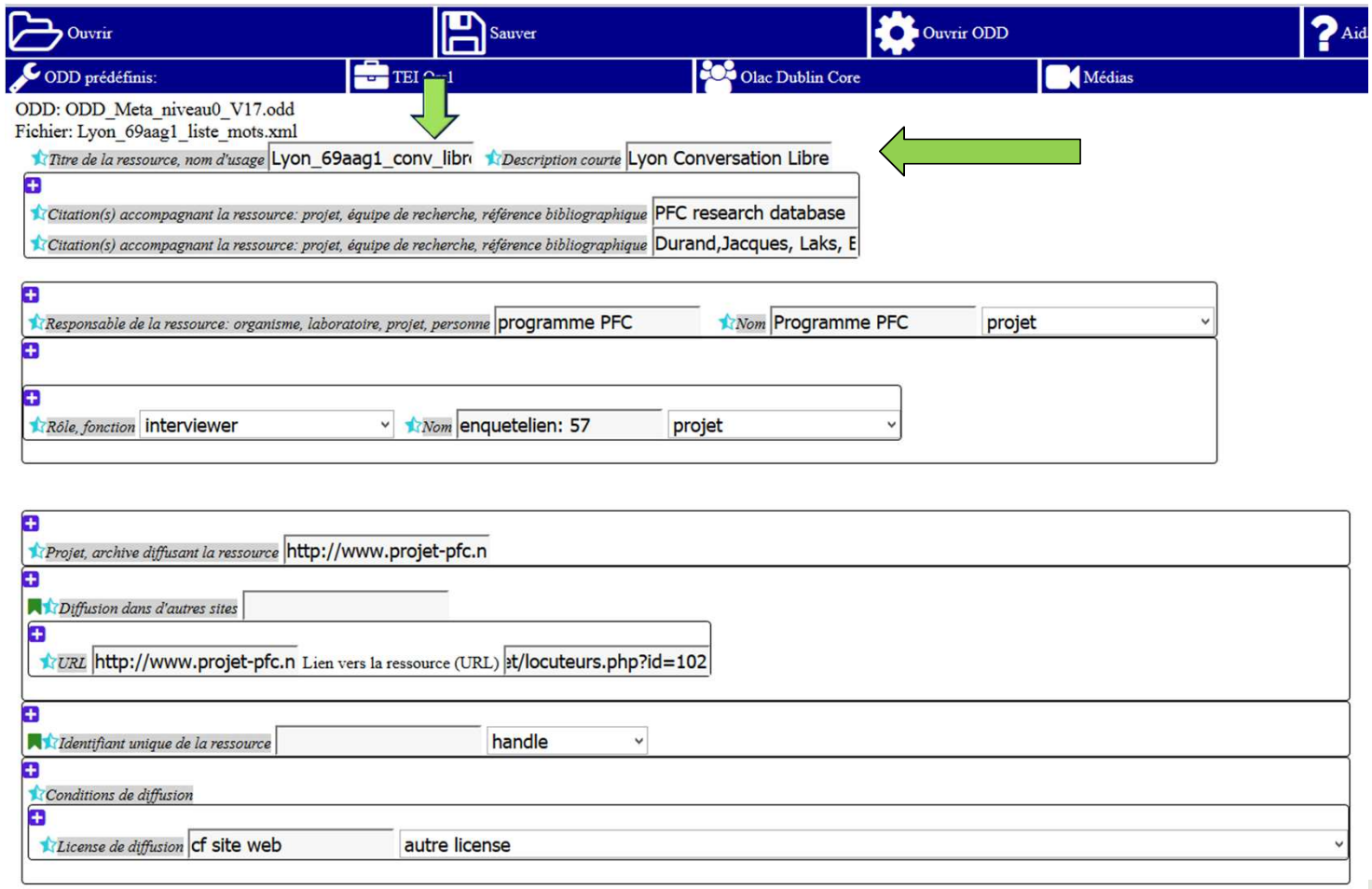
# Interface teimeta : les annotations
# http://ct3.ortolang.fr/teimeta



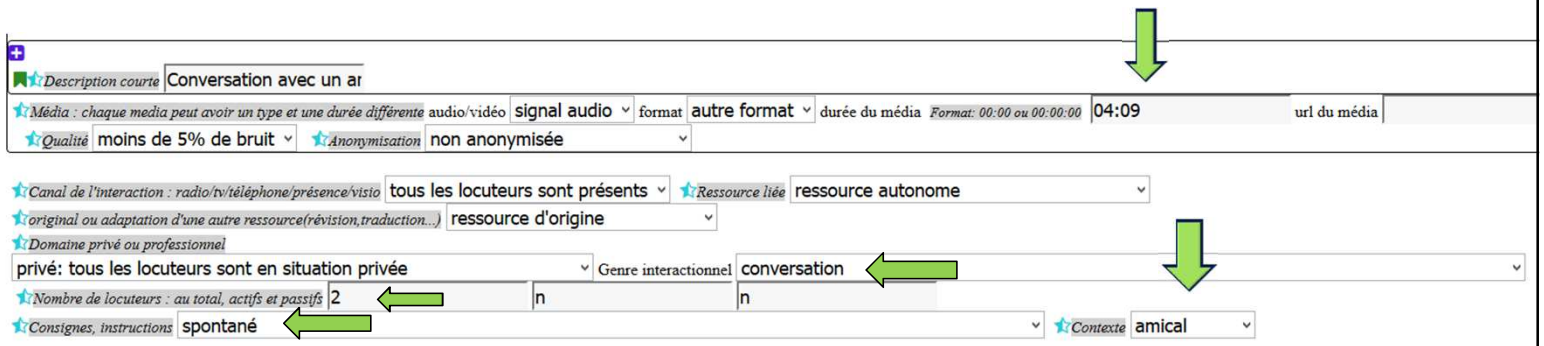
# teiMeta : la conversation libre

Seuls le titre et la situation  sont modifiés, un second locuteur est ajouté

# Interface teimeta : la ressource http://ct3.ortolang.fr/teimeta

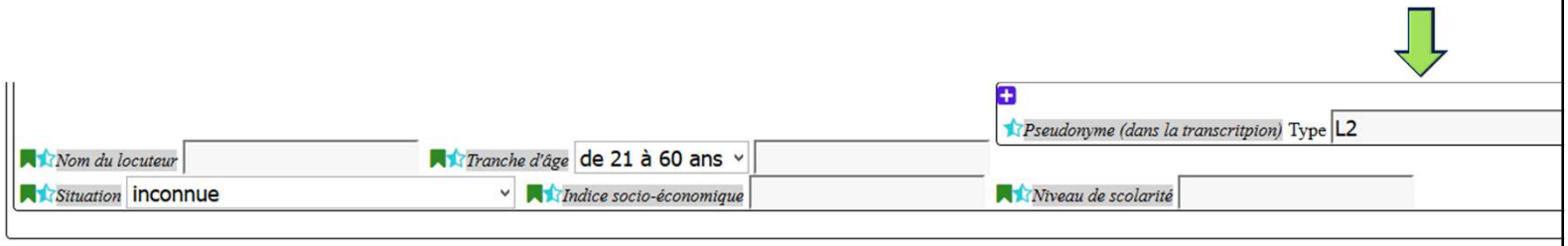


# Interface teimeta : situation et enregistrement
# http://ct3.ortolang.fr/teimeta




23

# Interface teimeta : le locuteur
# http://ct3.ortolang.fr/teimeta




24

# 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)
- Journées Linguistique de Corpus (2017): atelier
- Formation Archivage ET Interopérabilité (2018)

  CORLI : InterExplo-Corli et Dépôt conservation évaluation et diffusion

25