

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

*C*ORpus  
*L*angues  
*I*nteractions



The logo for Huma-Num features the letters 'H' and 'N' in a stylized, overlapping font. The 'H' is dark red and the 'N' is orange. Below the letters, the text 'Huma-Num' is written in a bold, black, sans-serif font.

**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, transcripateur, ...
    - ...

# 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

## 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, formats
  - 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

# 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

## 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**

## 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

Edition de métadonnées TEI / CORLI ★ ☆ - 0.4.8

Ouvrir Sauver Ouvrir ODD Aide

ODD prédéfinis: TEI Oral Olac Dublin Core Médias Paramètres

ODD: ODD\_Meta\_niveau0\_V17.odd  
Fichier: Lyon\_69aag1\_liste\_mots.xml

★ Titre de la ressource, nom d'usage: Lyon\_69aag1\_liste\_mots ★ Description courte: Lyon liste de mots du

★ Citation(s) accompagnant la ressource: projet, équipe de recherche, référence bibliographique PFC research database

★ Citation(s) accompagnant la ressource: projet, équipe de recherche, référence bibliographique Durand, Jacques, Laks, E

★ Responsable de la ressource: organisme, laboratoire, projet, personne programme PFC ★ Nom: Programme PFC projet

★ Rôle, fonction: interviewer ★ Nom: enquetelien: 57 projet

★ Projet, archive diffusant la ressource: http://www.projet-pfc.n

★ Diffusion dans d'autres sites

★ URL: http://www.projet-pfc.n Lien vers la ressource (URL): t/locuteurs.php?id=102

★ Identifiant unique de la ressource: handle

★ Conditions de diffusion

★ License de diffusion: cf site web autre license

17

# Interface teimeta : situation et enregistrement

## <http://ct3.ortolang.fr/teimeta>

+ Description courte Liste de mots

★ Média : chaque média peut avoir un type et une durée différente audio/vidéo signal audio format autre format durée du média Format: 00:00 ou 00:00:00 02:46 url du média

★ Qualité moins de 5% de bruit ★ Anonymisation non anonymisée

★ Canal de l'interaction : radio/tv/téléphone/présence/visio tous les locuteurs sont présents ★ Ressource liée ressource autonome

★ original ou adaptation d'une autre ressource(révision, traduction...) ressource d'origine

★ Domaine privé ou professionnel

privé: tous les locuteurs sont en situation privée Genre interactionnel nature inconnue

★ Nombre de locuteurs : au total, actifs et passifs 1 n n

★ Consignes, instructions lecture, lecture de mots ← Contexte inconnu

+ Ville Lyon 4ième

+ Région Rhône-Alpes

+ Description courte

★ Lieu appartement ★ Pays fr

+ Session enregistrée : description, date, environnement, langue

+ Intervalle (depuis/jusqu'à) ou date exacte Depuis jj / mm / aaaa Jusqu'à jj / mm / aaaa Date exacte 12 / 05 / 1999

+ Description de la session enregistrée

+ Langue(s) de la situation, si plusieurs pourcentage d'utilisation français ←

18

# Interface teimeta : le locuteur

<http://ct3.ortolang.fr/teimeta>

+ **Locuteur** Identifiant unique du locuteur  Sexe  URL Locuteur  Rôle

+ **Langue des locuteurs**

Code Iso	fra	maternelle
Code Iso	allemand	seconde
Code Iso	anglais	seconde
Code Iso	espagnol	seconde

+ **Information supplémentaire** Type

+ **Nom du locuteur**  **Tranche d'âge**  **Pseudonyme (dans la transcription)** Type

**Situation**  **Indice socio-économique**  **Niveau de scolarité**

19

# Interface teimeta : les annotations

<http://ct3.ortolang.fr/teimeta>

The screenshot displays the teimeta interface with two main sections. The top section is a filter panel with four rows of dropdown menus. A green arrow points to the first dropdown, which is set to 'phonologique'. The second dropdown is set to 'transcription non anonymisée', the third to 'praat', and the fourth to 'cf site web'. The bottom section is a conversion tool with a dropdown set to 'autre convertisseur' and a button labeled 'export des tiers' with a green arrow pointing to it.

<i>Annotations : type d'annotation, logiciel d'annotation, convention, anonymisation</i>	nature(s) des annotations	phonologique
<i>Annotations : type d'annotation, logiciel d'annotation, convention, anonymisation</i>	anonymisation de la transcription	transcription non anonymisée
<i>Annotations : type d'annotation, logiciel d'annotation, convention, anonymisation</i>	logiciel de transcription	praat
<i>Annotations : type d'annotation, logiciel d'annotation, convention, anonymisation</i>	convention de transcription	cf site web

<i>Convertisseur TEI</i>	transpraak	autre convertisseur	<i>Description courte</i>	export des tiers
--------------------------	------------	---------------------	---------------------------	------------------

# 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

Ouvre Ouvrir Sauver Ouvre ODD ? Aid  
 ODD prédéfinis: TEI 1 Olac Dublin Core Médias  
 ODD: ODD\_Meta\_niveau0\_V17.odd  
 Fichier: Lyon\_69aag1\_liste\_mots.xml  
 ★ Titre de la ressource, nom d'usage Lyon\_69aag1\_conv\_libri ★ Description courte Lyon Conversation Libre  
 ★ Citation(s) accompagnant la ressource: projet, équipe de recherche, référence bibliographique PFC research database  
 ★ Citation(s) accompagnant la ressource: projet, équipe de recherche, référence bibliographique Durand, Jacques, Laks, E  
 ★ Responsable de la ressource: organisme, laboratoire, projet, personne programme PFC ★ Nom Programme PFC projet  
 ★ Rôle, fonction interviewer ★ Nom enquetelien: 57 projet  
 ★ Projet, archive diffusant la ressource http://www.projet-pfc.n  
 ★ Diffusion dans d'autres sites  
 ★ URL http://www.projet-pfc.n Lien vers la ressource (URL) t/locuteurs.php?id=102  
 ★ Identifiant unique de la ressource handle  
 ★ Conditions de diffusion  
 ★ License de diffusion cf site web autre license

# Interface teimeta : situation et enregistrement

<http://ct3.ortolang.fr/teimeta>

+

★ Description courte

★ Média : chaque média peut avoir un type et une durée différente audio\vidéo  format  durée du média Format: 00:00 ou 00:00:00  url du média

★ Qualité  ★ Anonymisation

★ Canal de l'interaction : radio/tv/téléphone/présence/visio  ★ Ressource liée

★ original ou adaptation d'une autre ressource(révision, traduction...)

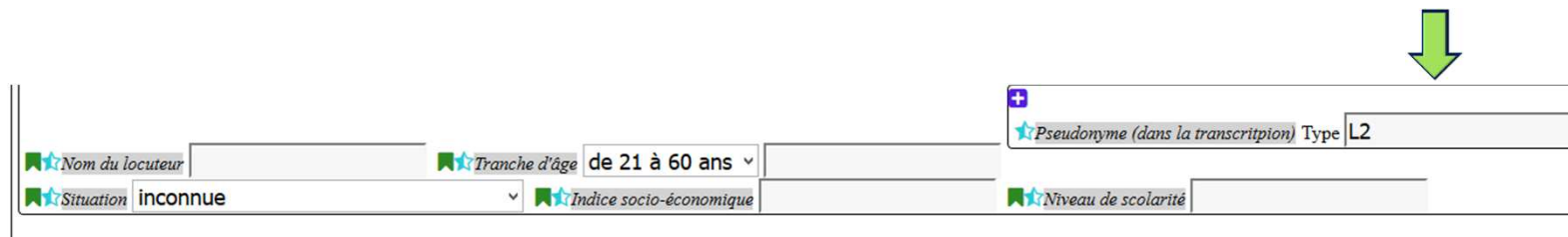
★ Domaine privé ou professionnel  Genre interactionnel

★ Nombre de locuteurs : au total, actifs et passifs

★ Consignes, instructions  ★ Contexte

# Interface teimeta : le locuteur

<http://ct3.ortolang.fr/teimeta>



The screenshot shows a search interface with the following fields and values:

- Nom du locuteur**: [Empty text box]
- Tranche d'âge**: de 21 à 60 ans (dropdown menu)
- Situation**: inconnue (dropdown menu)
- Indice socio-économique**: [Empty text box]
- Niveau de scolarité**: [Empty text box]
- Pseudonyme (dans la transcription) Type**: L2 (dropdown menu)

A green arrow points to the search button, which is a small purple square with a white plus sign.



## 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