FAIR en linguistique de la langue orale : objectifs, méthode et outils

CORpus Langues Interactions

Huma-Num

Loïc Liégeois, Carole Etienne et Christophe Parisse Novembre 2018

CORLI : CORpus, Langues et Interactions

- Consortium dédié à la **linguistique de corpus** de l'infrastructure **Huma-Num**
- Comité de pilotage d'une vingtaine de personnes, 180 chercheurs
- **Large couvertur**e des différents domaines de la linguistique
- Groupes projet
  - **Inter-Explo** : Interopérabilité /Pratique et outils d'exploration de corpus
  - **Multicom** : Multimodalité et Nouvelles formes de communication
  - **Corpus multilingues et plurilingues**

CORLI : CORpus, Langues et Interactions

- Missions
  - **Fédérer** les laboratoires travaillant sur corpus pour **recenser** et **mutualiser** les méthodologies, les ressources, les pratiques et les besoins
  - **Établir**, **partager et diffuser** des bonnes pratiques
  - Diffuser dans des **standards** européens et internationaux : contributions CLARIN , DARIAH et au Consortium TEI
  - Établir des **critères d'évaluation des corpus** en tant que production scientifique

- Actions
  - Organisation de journées d'études
  - Organisation de formations (15 sessions en 2017)
  - Aide à la mise à disposition de corpus (13 projets en 2017)
  - Diffusion de nos réalisations dans les colloques
  - Concertation avec l'équipex ORTOLANG : conception de nouveaux outils

FAIR en linguistique de la langue orale

- **Findability : Différents publics donc différentes métadonnées**
  - **La situation :** professionnel/privé, face à face/distance, contexte, …
  - **Les locuteurs :** adulte/enfant, natif/non natif, caractéristiques socio-économiques, formation, …
  - **L'enregistrement :** format**S** diffusion / analyse
  - **Le mode de citation de la ressource**
  - **Les conditions d'accès :** diffusion, anonymisation
  - **Les annotations >>> la transcription :** nature et convention
  - **Le format des fichiers de transcription**
  - **Les projets de recherche incluant cette ressource ou une partie de cette ressource**

FAIR en linguistique de la langue orale

- **Accessibility**
  - Tradition de **plateformes d'archivage en linguistique** depuis les années 2000
  - Un corpus oral est **long** à constituer : recherche de terrain, problèmes de droit, enregistrement physique, recueil des métadonnées

    … et à annoter particulièrement quand plusieurs locuteurs interviennent
  - Des annotations phonologiques, prosodiques, interactionnelles, multimodales ➔ **une à plusieurs heures d'annotation pour 1 minute d'enregistrement** ou des corpus longitudinaux recueillis dans les mêmes conditions sur une période assez longue ➔ **phase d'enregistrement longue et beaucoup d'heures à annoter, partiellement automatisable**
  - **Vérification** du processus scientifique : besoin de **conserver** une version d'un corpus pour reproduire une analyse déjà effectuée en vue de l'améliorer
  - **Constituer un corpus d'étude** en regroupant plusieurs corpus existants

FAIR en linguistique de la langue orale

- **Interoperability :** Différentes disciplines de la linguistique pour
  - **explorer les mêmes données :** analyses syntaxiques, prosodiques ou interactionnelles d'une même donnée orale
  - **contraster le même objet d'étude dans une autre perspective** : écrit /oral, enfant/adulte, registre formel/informel, langues différentes, régions différentes, diachronie…
    - **identifier** précisément les données pour les **sélectionner** dans l'étude
    - rendre **homogène** le corpus d'étude constitué de données de plusieurs sources **sans le redécrire** (déjà fait dans chaque source)
    - **disposer d'informations** au moment des analyses
    - **ajouter** de nouvelles annotations

FAIR en linguistique de la langue orale

- **Interoperability : les solutions pour les métadonnées**
  - Un jeu **commun** de métadonnées à toutes les ressources pour faciliter la prise en main des métadonnées et la mise en commun des données
  - Une application **personnalisée teimeta** pour saisir ces métadonnées à partir d'un fichier TEI/ODD défini pour les corpus oraux et largement diffusé
  - Jeu commun mais **différents niveaux de granularité** :

    adulte >> tranche d'âge >> âge précis
  - **Vocabulaire contrôlé** et **application multilingue teimeta**

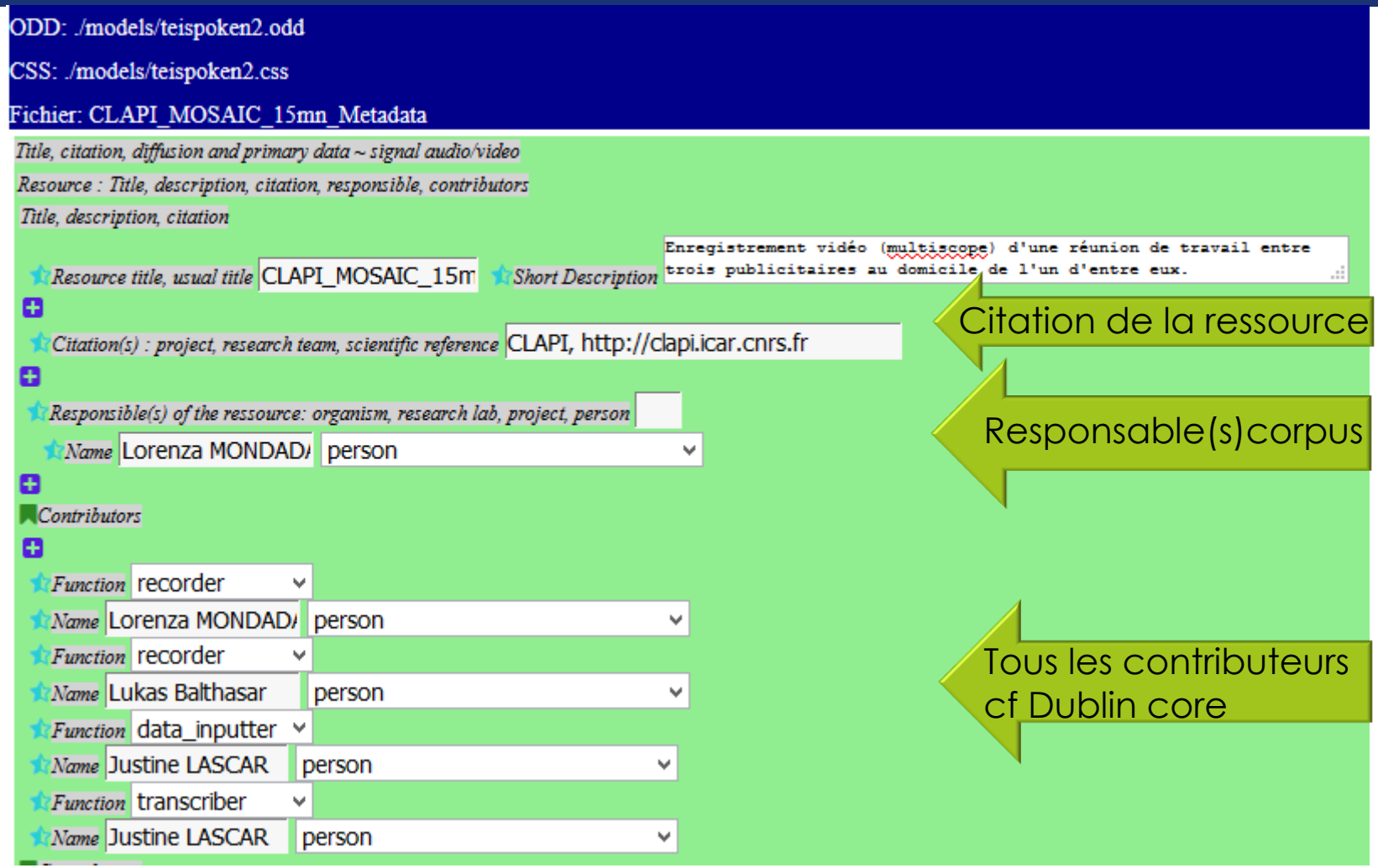
- **Interoperability : les solutions pour les annotations**
  - Un format de transcription **pivot**, indépendant des conventions, logiciels ou outils
  - Un standard **international TEI** pour ce format pivot

FAIR en linguistique de la langue orale

- **Reusability**
  - les licences d'utilisation des données
  - teimeta : un jeu commun de **métadonnées orientées recherche**
  - teiconvert : des outils de conversion pour passer **sans perte d'information** d'un logiciel d'annotation à un autre

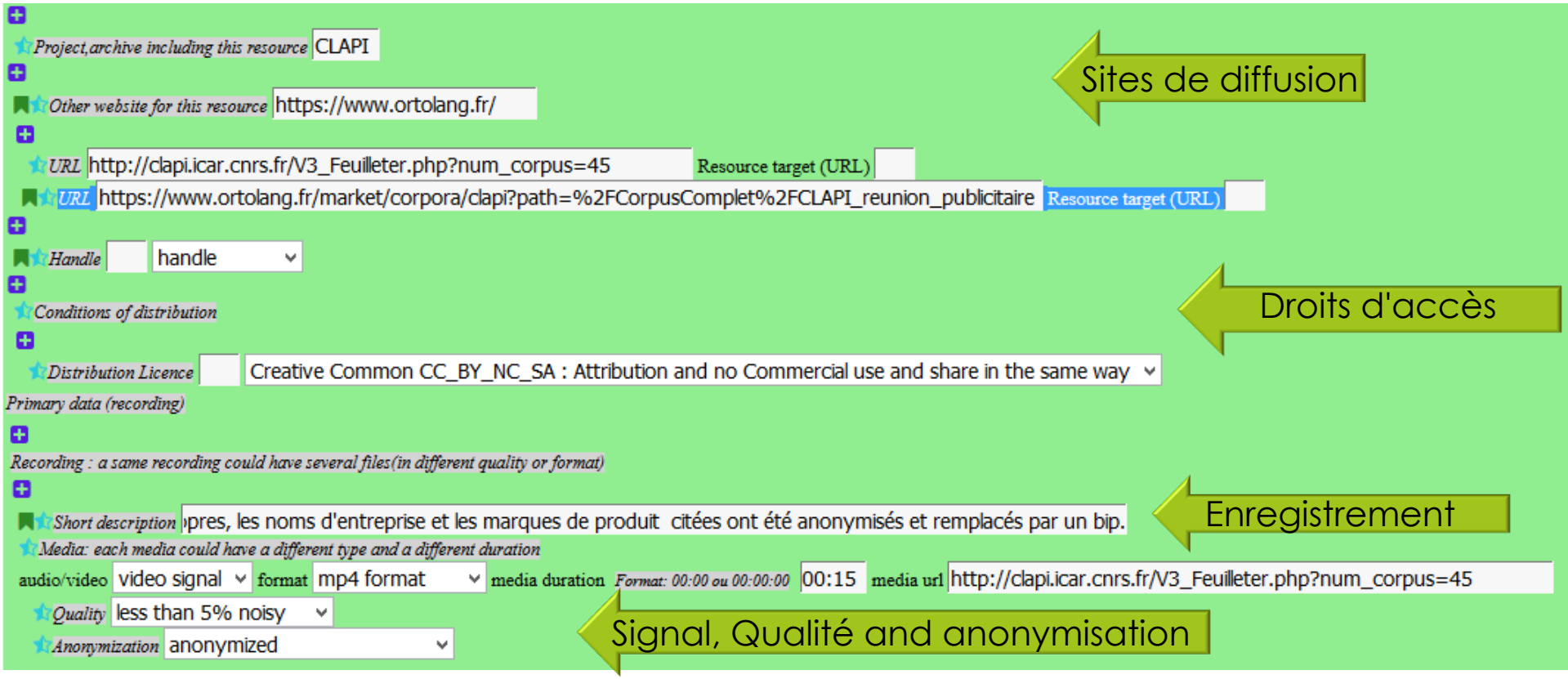
teimeta : exemples

TEIMETA application : exemple d'une réunion



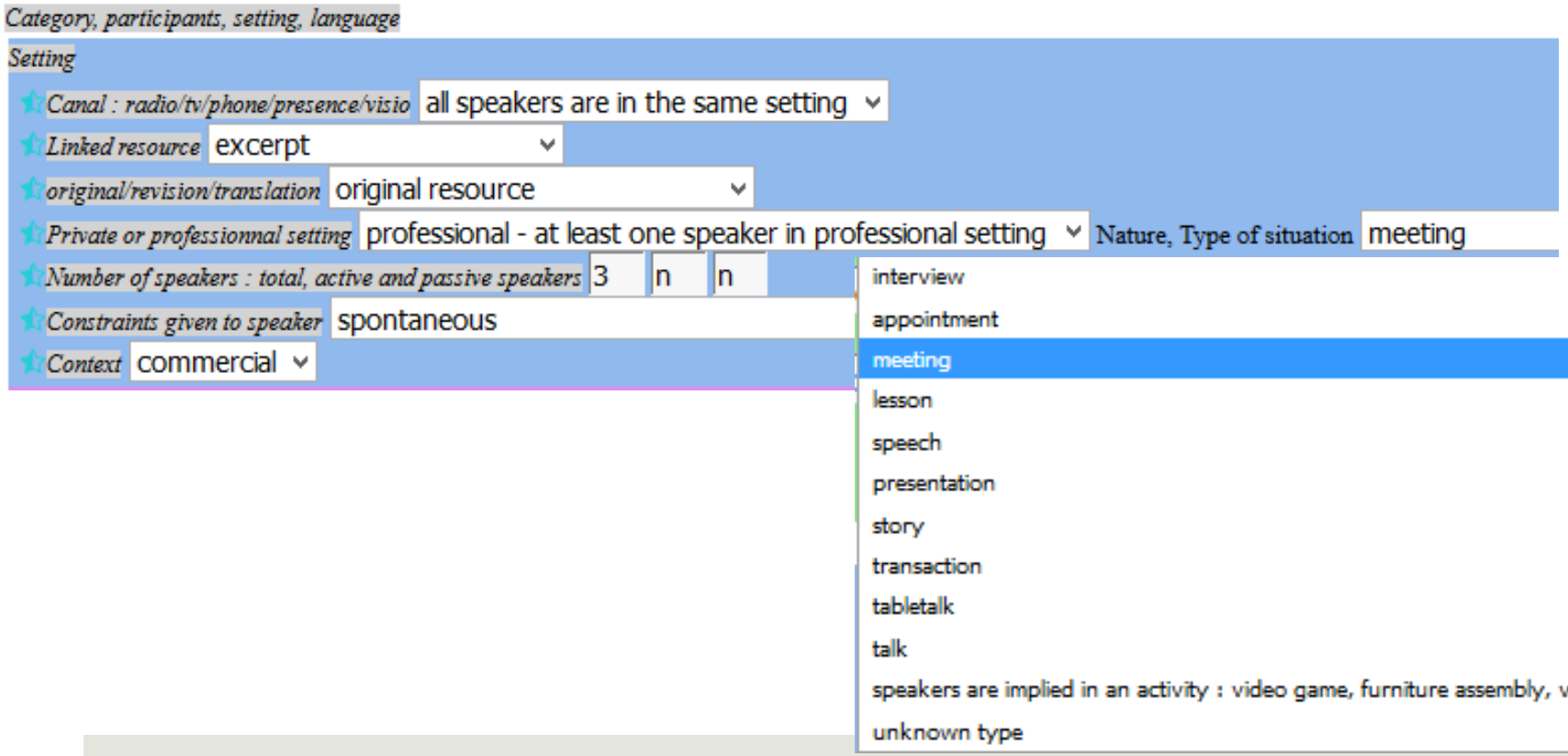
TEIMETA application : exemple d'une réunion

Sites de diffusion, droits d'accès et enregistrement



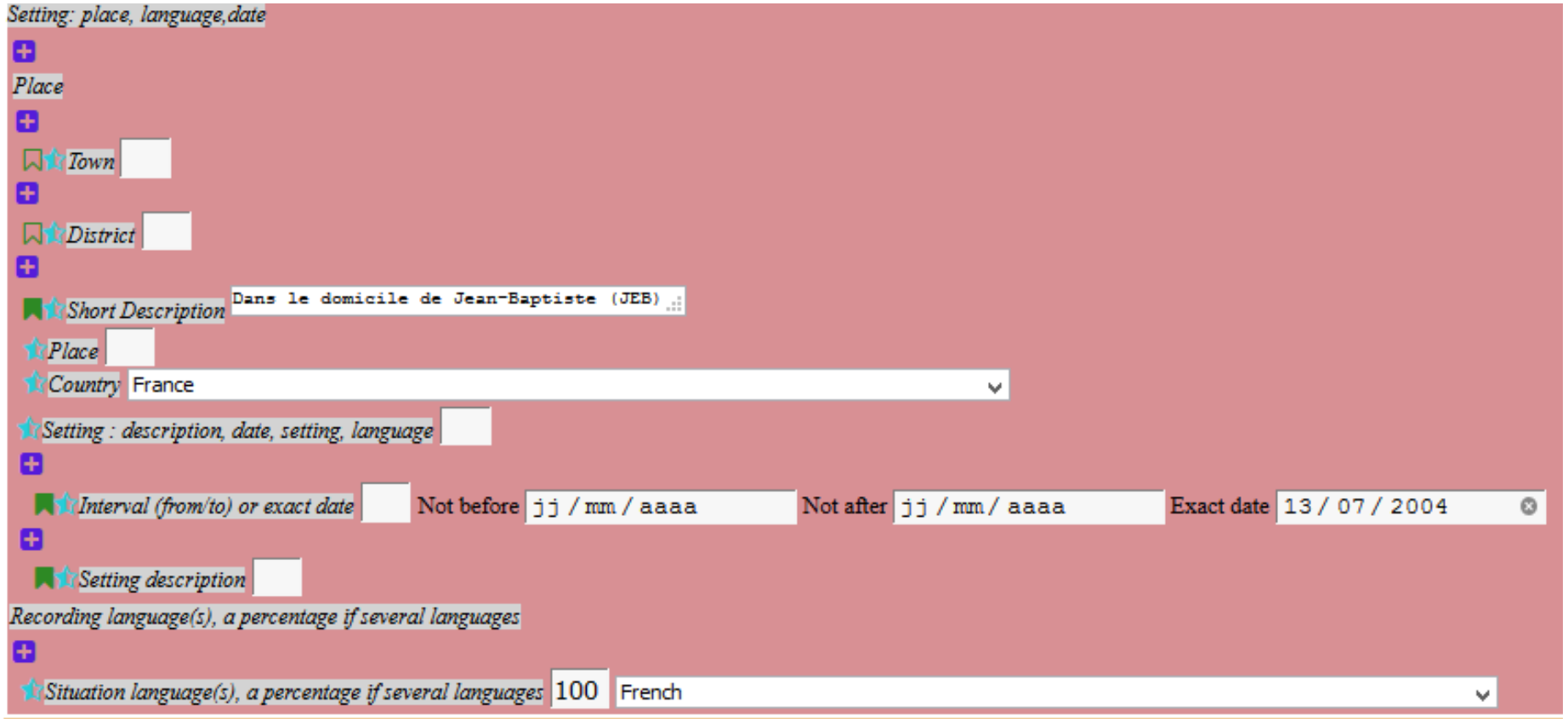
TEIMETA application : exemple d'une réunion

Situation : plusieurs éléments pour l'identifier ➔ Vocabulaire contrôlé
Utilisation de css pour repérer le type de métadonnées



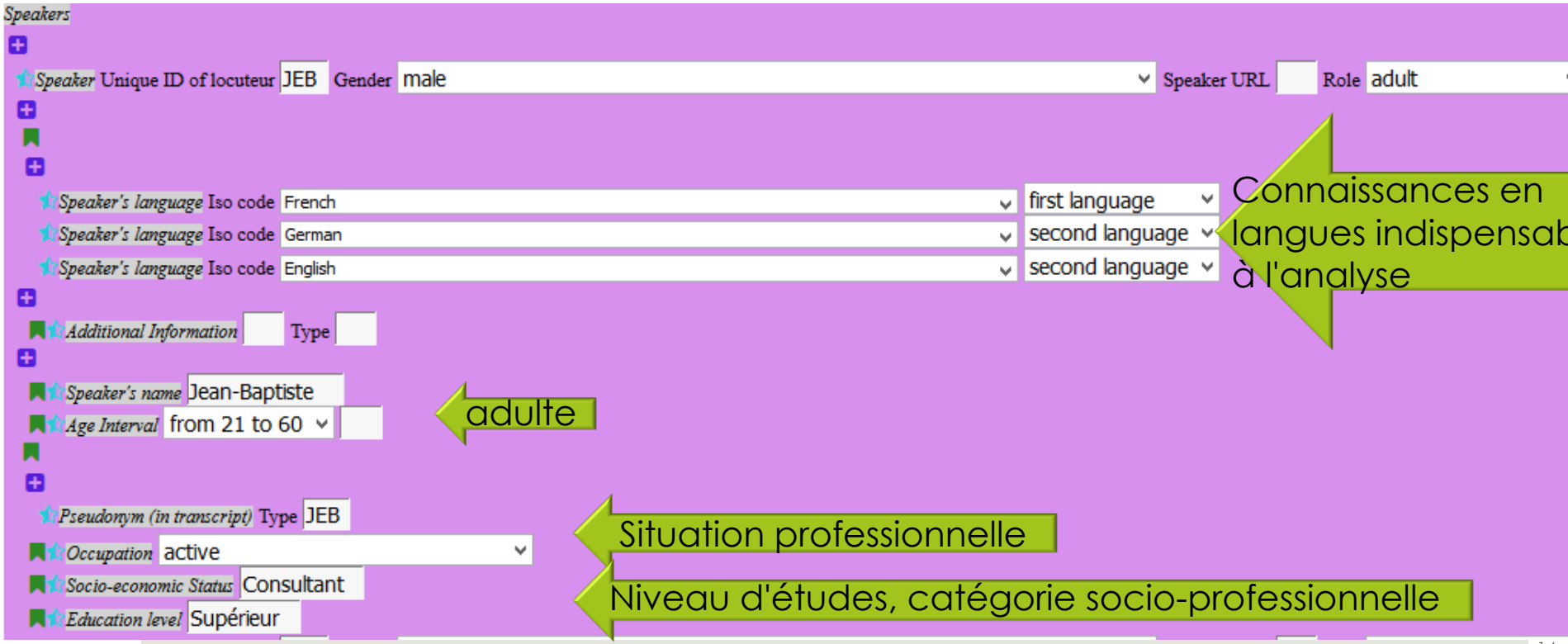
TEIMETA application : exemple d'une réunion

Situation : lieu, date et langues



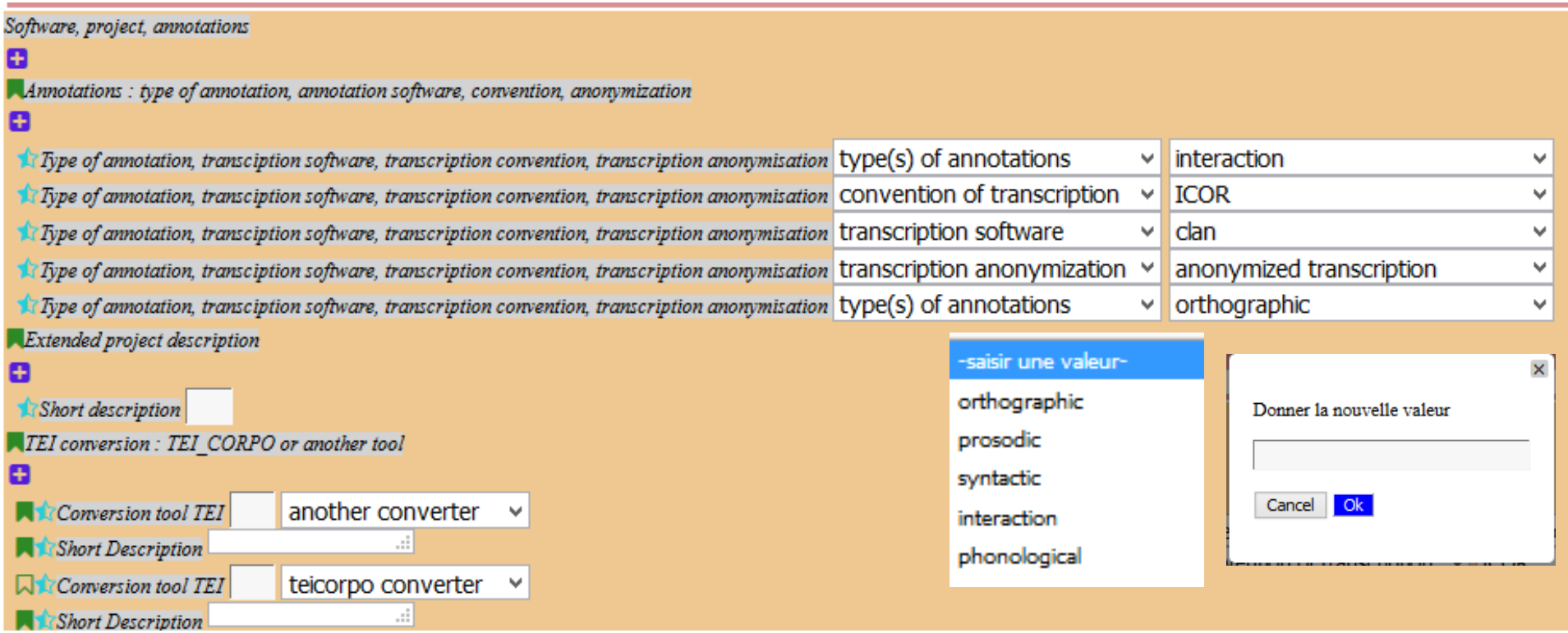
TEIMETA application : exemple d'une réunion

Locuteur : différents niveaux de granularité depuis un adulte parlant français jusqu'à une description plus complète avec l'âge précis, la situation professionnelle, le niveau d'études et les connaissances en langues



TEIMETA application : exemple d'une réunion

Transcription : logiciel, anonymisation, **annotations** qui peuvent être enrichies au fil des analyses, utilisation de **vocabulaire contrôlé avec une possibilité d'ajouter des valeurs**



teicorpo : Format pivot pour les transcriptions de l'oral

Format pivot pour les transcriptions de l'oral