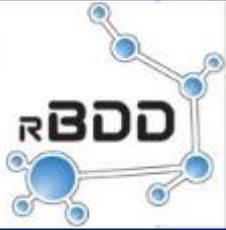




FAIR en linguistique de la langue orale : objectifs, méthode et outils

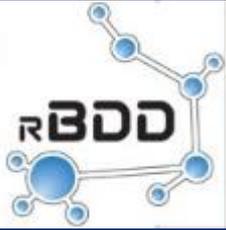
*C*ORpus
*L*angues
*I*nteractions

HN
Huma-Num



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 couverture** 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**



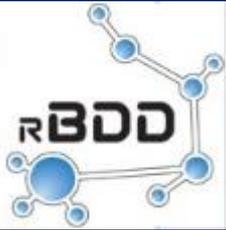
CORLI : CORpus, Langues et Interactions

□ 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

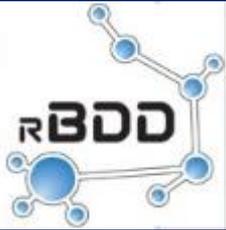
- **Findability : Différents publics donc différentes métadonnées**
 - **La situation** : professionnel/privé, face à face/distance, contexte, ...
 - **Les locuteurs** : adulte/enfant, natif/non natif, caractéristiques socio-économiques, formation, ...
 - **L'enregistrement** : formatS diffusion / analyse
 - **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

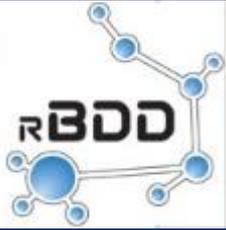
□ Accessibility

- 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



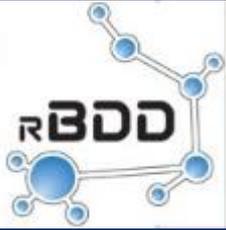
FAIR en linguistique de la langue orale

□ 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 diffusés
- 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



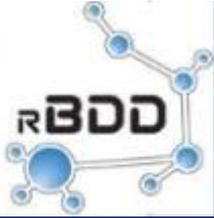
FAIR en linguistique de la langue orale

□ 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

ODD: ./models/teispoken2.odd

CSS: ./models/teispoken2.css

Fichier: CLAPI_MOSAIC_15mn_Metadata

Title, citation, diffusion and primary data ~ signal audio/video

Resource : Title, description, citation, responsible, contributors

Title, description, citation

★ *Resource title, usual title* CLAPI_MOSAIC_15m ★ *Short Description*

Enregistrement vidéo (multiscopes) d'une réunion de travail entre trois publicitaires au domicile de l'un d'entre eux.



★ *Citation(s) : project, research team, scientific reference* CLAPI, <http://clapi.icar.cnrs.fr>



★ *Responsible(s) of the resource: organism, research lab, project, person*

★ *Name* Lorenza MONDAD/ person



Contributors



★ *Function* recorder

★ *Name* Lorenza MONDAD/ person

★ *Function* recorder

★ *Name* Lukas Balthasar person

★ *Function* data_inputter

★ *Name* Justine LASCAR person

★ *Function* transcriber

★ *Name* Justine LASCAR person

Citation de la ressource

Responsable(s) corpus

Tous les contributeurs
cf Dublin core



TEIMETA application : exemple d'une réunion

Sites de diffusion, droits d'accès et enregistrement

Sites de diffusion

Droits d'accès

Enregistrement

Signal, Qualité and anonymisation

Project, archive including this resource CLAPI

Other website for this resource <https://www.ortolang.fr/>

URL http://clapi.icar.cnrs.fr/V3_Feuilleter.php?num_corpus=45 Resource target (URL)

URL https://www.ortolang.fr/market/corpora/clapi?path=%2FCorpusComplet%2FCLAPI_reunion_publicitaire Resource target (URL)

Handle handle

Conditions of distribution

Distribution Licence Creative Common CC_BY_NC_SA : Attribution and no Commercial use and share in the same way

Primary data (recording)

Recording : a same recording could have several files (in different quality or format)

Short description ...pres, les noms d'entreprise et les marques de produit citées ont été anonymisés et remplacés par un bip.

Media: each media could have a different type and a different duration

audio/video video signal format mp4 format media duration Format: 00:00 ou 00:00:00 00:15 media url http://clapi.icar.cnrs.fr/V3_Feuilleter.php?num_corpus=45

Quality less than 5% noisy

Anonymization anonymized



TEIMETA application : exemple d'une réunion

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

Category, participants, setting, language

Setting

★ *Canal : radio/tv/phone/presence/visio* all speakers are in the same setting ▾

★ *Linked resource* excerpt ▾

★ *original/revision/translation* original resource ▾

★ *Private or professional setting* professional - at least one speaker in professional setting ▾ Nature, Type of situation meeting

★ *Number of speakers : total, active and passive speakers* 3 n n

★ *Constraints given to speaker* spontaneous

★ *Context* commercial ▾

interview

appointment

meeting

lesson

speech

presentation

story

transaction

tabletalk

talk

speakers are implied in an activity : video game, furniture assembly, visit, task ..

unknown type



TEIMETA application : exemple d'une réunion

Situation : lieu, date et langues

Setting: *place, language, date*

+ Place

+
★ Town

+
★ District

+
★ Short Description Dans le domicile de Jean-Baptiste (JEB) ...
★ Place

★ Country France

★ Setting : *description, date, setting, language*

+
★ Interval (from/to) or exact date Not before Not after Exact date 13 / 07 / 2004

+
★ Setting description

Recording language(s), a percentage if several languages

+
★ Situation language(s), a percentage if several languages 100 French



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**

Software, project, annotations



Annotations : type of annotation, annotation software, convention, anonymization



- ★ Type of annotation, transcription software, transcription convention, transcription anonymisation
- ★ Type of annotation, transcription software, transcription convention, transcription anonymisation
- ★ Type of annotation, transcription software, transcription convention, transcription anonymisation
- ★ Type of annotation, transcription software, transcription convention, transcription anonymisation
- ★ Type of annotation, transcription software, transcription convention, transcription anonymisation

type(s) of annotations	interaction
convention of transcription	ICOR
transcription software	clan
transcription anonymization	anonymized transcription
type(s) of annotations	orthographic

Extended project description



★ Short description

TEI conversion : TEI_CORPO or another tool



★ Conversion tool TEI another converter

★ Short Description

★ Conversion tool TEI teicorpo converter

★ Short Description

- saisir une valeur-
- orthographic
- prosodic
- syntactic
- interaction
- phonological

Donner la nouvelle valeur

Cancel Ok



teicorpo : Format pivot pour les transcriptions de l'oral



Format pivot pour les transcriptions de l'oral



IRCOM



ORTOLANG



TEI

Conversions au format TEI pour l'Oral et le Multimodal

1) Choisir le Format Destination

- TEI (xml / tei_corpo.xml / teiml / trjs)
- TRS (transcriber)
- CHA (chat - childes)
- TXT (texte - utf8)
- DOCX (microsoft word)
- XLSX (microsoft excel)
- CSV (tableurs)
- TEXTGRID (praat)
- EAF (elan)
- TXM (xml/w)
- Lexico/Le Trameur (.txt)



- Conserver ces locuteurs/champs dans la sortie
 - Supprimer ces locuteurs/champs de la sortie
- Valeur du locuteur ou du champ (caractères génériques acceptés)
- Supprimer les marqueurs spécifiques de l'oral

2) Choisir le Fichier source (extension: TRS/CHA/TEXTGRID/EAF/TXT/DOCX/XLSX)

Faire glisser ici un (ou plusieurs) fichier(s)

Ou cliquer ici pour sélectionner un fichier => Aucun fichier sélectionné.

- Demander les paramètres pour les fichiers praat.

Résultats (Effacer)

Le format TEI_CORPO suit les propositions du GT2 IRCOM et du groupe TEI Oral ISO. Il est conforme au standard TEI.