



HAL
open science

VERA - Vidéo, Ethnotextes et Ressources Associées

Pierre-Aurélien Georges

► **To cite this version:**

Pierre-Aurélien Georges. VERA - Vidéo, Ethnotextes et Ressources Associées. AG CORLI 2019, Jan 2019, Paris, France. . hal-03136050

HAL Id: hal-03136050

<https://hal.science/hal-03136050>

Submitted on 9 Feb 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - ShareAlike 4.0 International License

Des enregistrements inédits...

2006 – 2017 :

- + de 100 enquêtes de terrain
- Dialectes occitans et liguriens de 75 localités (France & Italie)
- 5 questionnaires de phrases totalisant 1 200 questions



- Bilan de ces 9 campagnes d'enquêtes :
- + de 120 locuteurs interviewés
- 58 ethnotextes récoltés
- 60 h de vidéos inédites à traiter (et autant d'enregistrements audio)
- env. 400 000 occurr. à lemmatiser

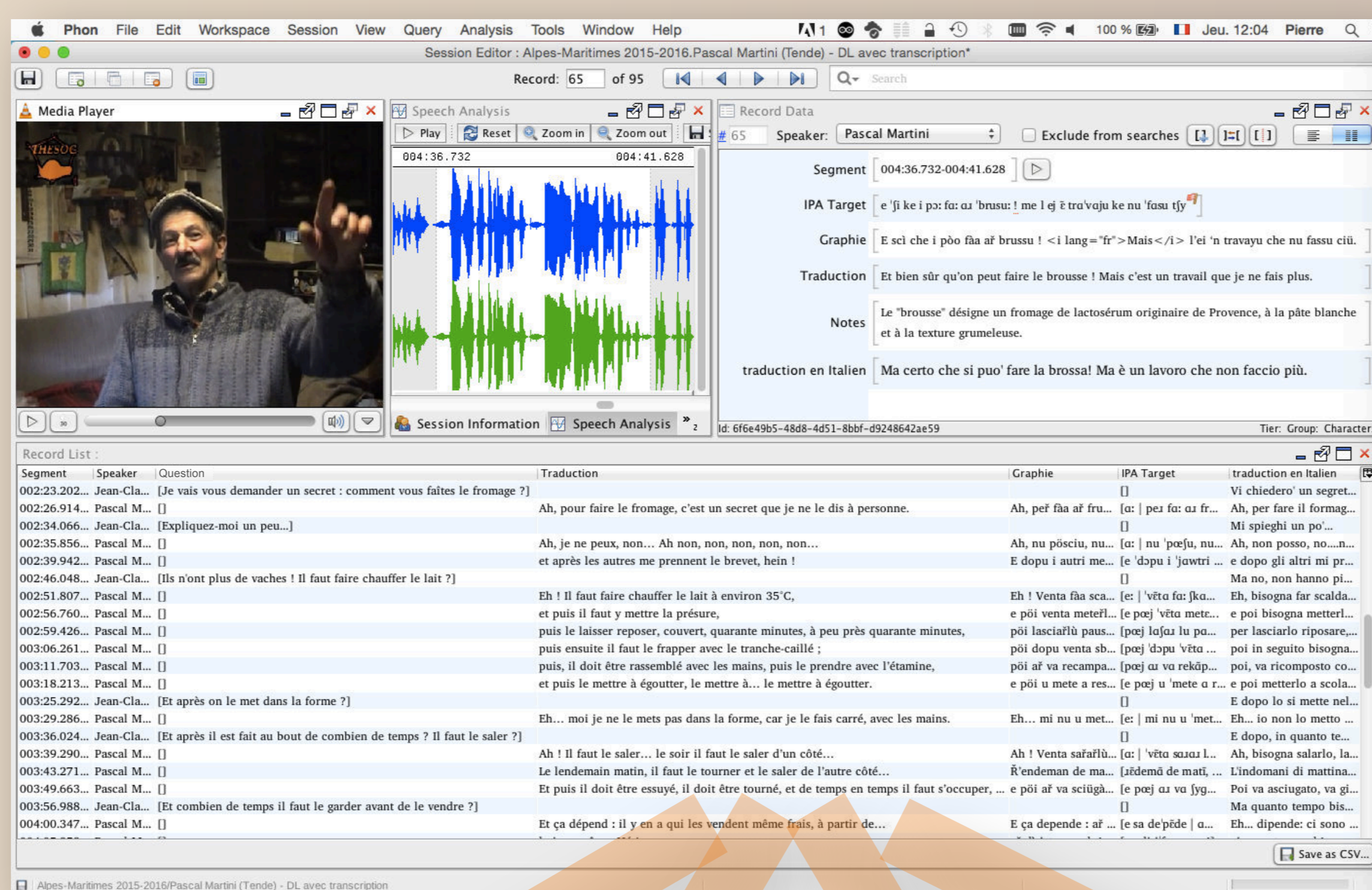
...accompagnés de nouvelles ressources...

1) Montage Vidéo

Normalisation du volume sonore,
Synchronisation audio – vidéo,
Choix des séquences vidéos,
Titres et logos.
(*Adobe Premiere*)

2) Segmentation

Repérage des tours de parole,
Minutage phrase par phrase,
Numérotation des réponses
aux questionnaires de phrases
(*PHON.ca – XML PhonBank*)



3) Transcriptions

en graphie et en Alphabet Phonétique International
- ou - dans 2 graphies différentes, suivant les enquêtes.

4) Traduction

alignée phrase par phrase, en français et parfois italien,
avec commentaires et explication des termes opaques.

5) Annotations

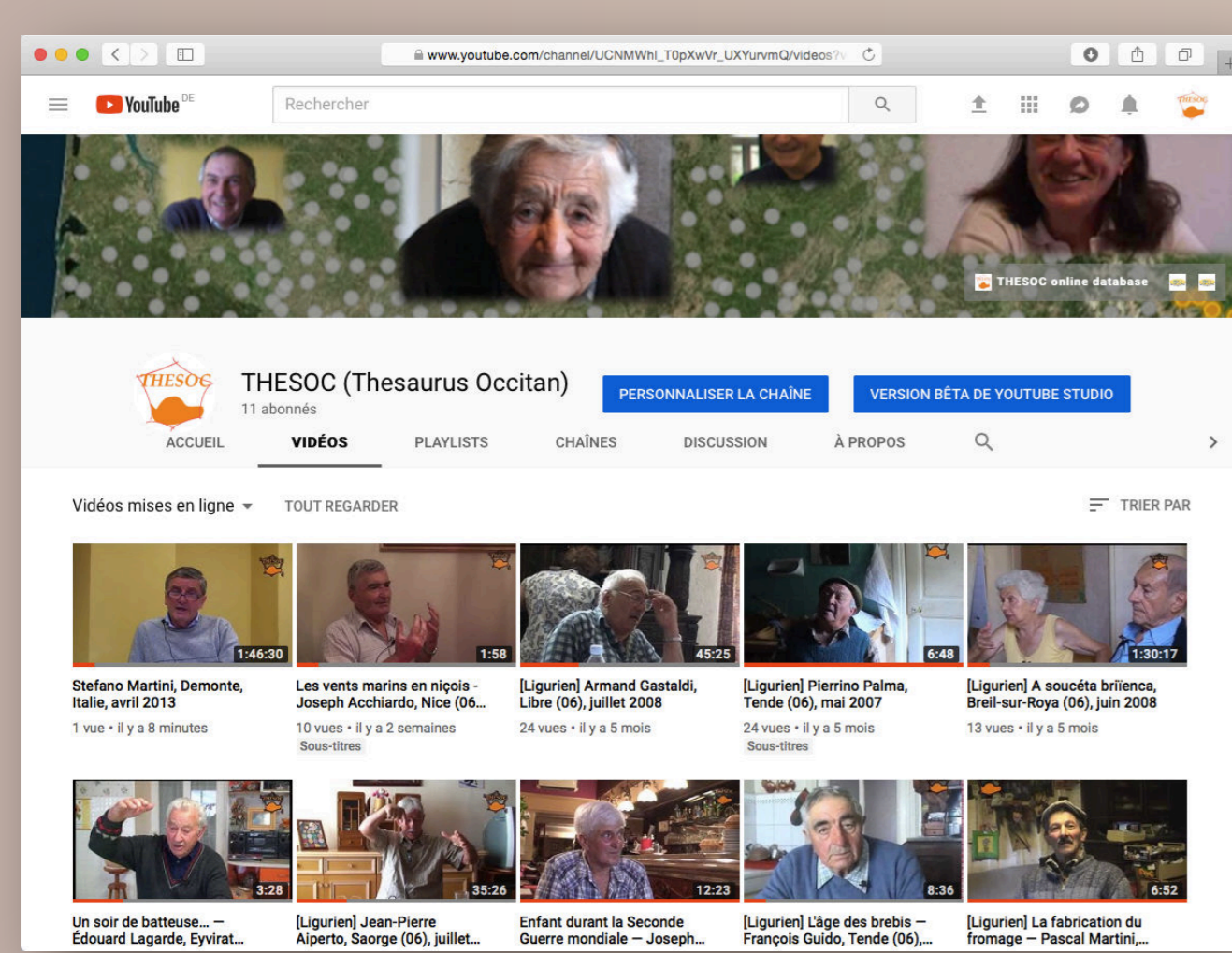
Lemmatisation et étiquettes P.O.S.
(base *MMS du Thesoc*)

6) Encodage XML-TEI et géolocalisation

(métadonnées Dublin-Core et OLAC)

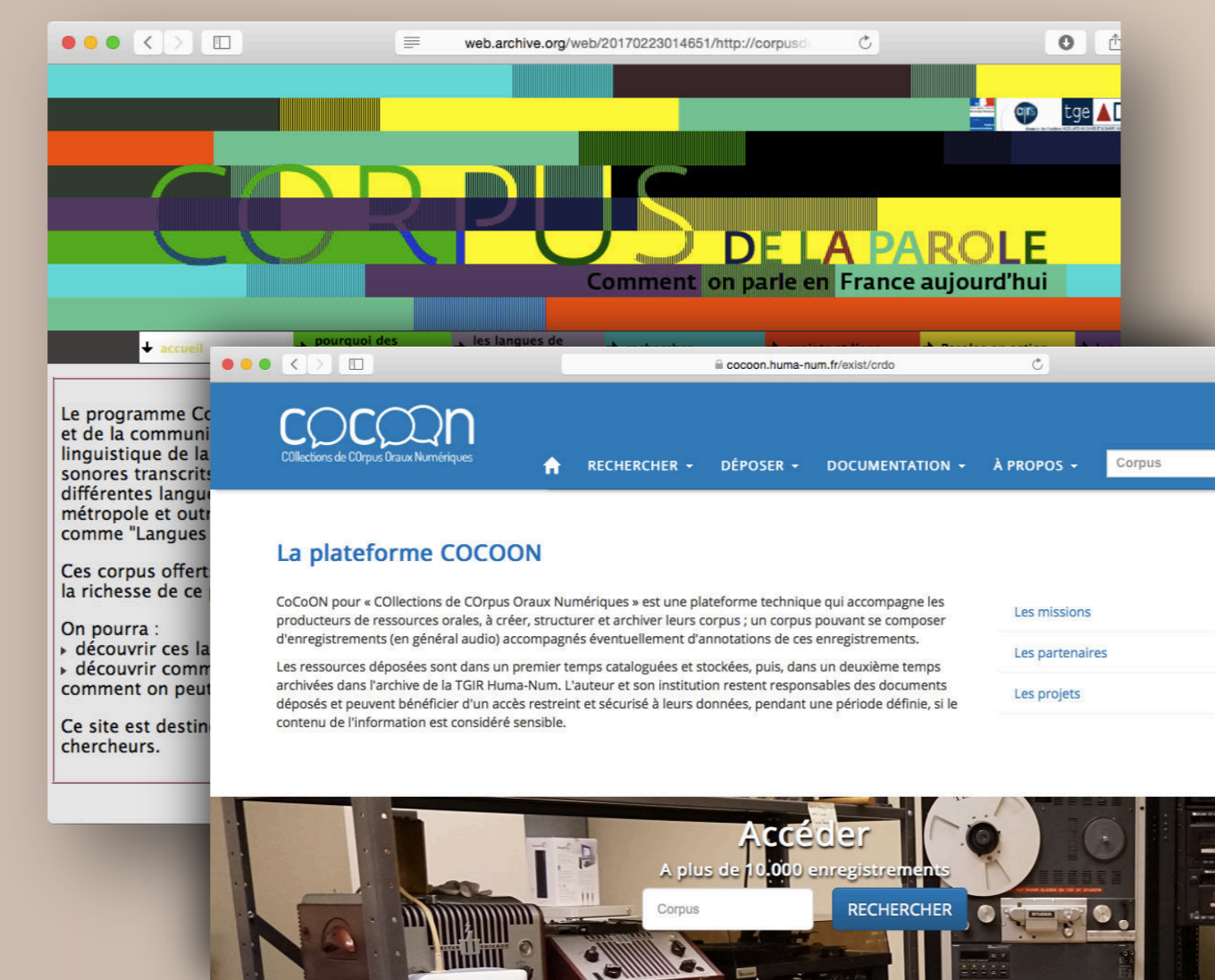
...consultables selon 4 modalités différentes...

Chaîne vidéo YouTube™
permettant au grand public de découvrir tous nos discours libres

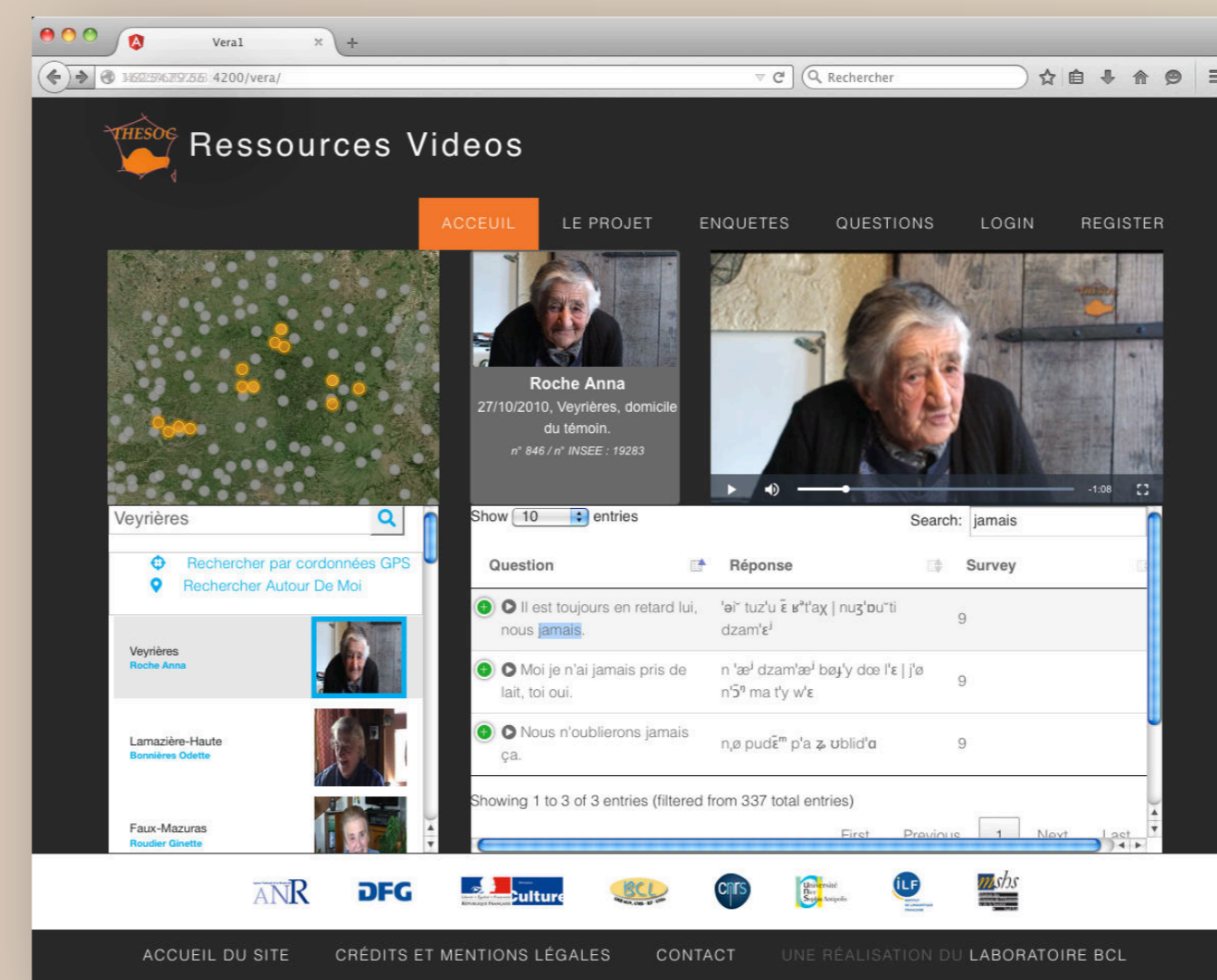


<https://bit.ly/thesoc>

Archivage pérenne
et versement au **Corpus de la Parole**



Plateforme web dédiée
avec cartographie interactive
et fonctions d'exploration du corpus



Interface de requête API REST
(interopérabilité, notamment avec les moteurs de recherche spécialisés)



...avec de multiples applications possibles :

- Industries de la Langue** : corpus d'entraînement en traduction automatique, Text2Speech, reconn. vocale, ...
- Enseignement des langues régionales** : support pédagogique pour les enseignants et étudiants
- Linguistique** : lexicque, phonologie, microvariation syntaxique, toponymie, ...
- Communication non verbale** : analyse de la gestuelle et expressions du visage
- Histoire, sociologie, ethnologie** (ethnotextes aux contenus très variés)

Ressources linguistiques
sous licence **LGPL-LR**
-
Vidéos sous licence
CC-BY-NC-ND