

Les données orales du CRDO

Michel Jacobson

Direction des Archives de France

Plan

1. Les types d'objet manipulés par le CRDO
2. Les fonctions de pré-ingestion
3. Les fonctions d'accès

1. Les types d'objets manipulés

- Pour que les ressources soient éligibles il faut qu'il s'agisse:
 - D'enregistrements de la parole
 - Accompagnés éventuellement d'annotations
 - Dont le but est d'étudier la langue ou le langage
 - Constitués dans le cadre d'une démarche scientifique
 - Par ou pour la communauté SHS

Le périmètre

- Un enregistrement de parole:
 - Types de production orale: Récits, conversation, élicitations, textes lus, logatones...
 - Type de locuteurs: Vous et moi, des enfants avant 2 ans, des personnes atteintes de trouble de la parole (pas actuellement dans l'entrepôt)
 - Différentes langues. Il existe environ 6000 langues (ref. catalogue ethnologue) dont une infime partie possède une tradition d'écriture.
 - Type de situation: interactions au travail, interviews, expériences en chambre sourde...

Le périmètre

- Un enregistrement de parole:
 - Ca ne veut pas dire enregistrement audio
 - Langues des signes
 - Langage enfantin
 - Interactions dans un groupe
 - En général les enregistrements sont audio ou vidéo mais on rencontre aussi des enregistrements de mesures physiologiques (electroglotogramme, endoscopie, IRMf...)
 - Ces enregistrements peuvent être accompagnés d'annotations (transcriptions, traductions, indications scénographiques, analyses syntaxiques, prosodiques, etc.)

Description des ressources

- Les enregistrements
 - Audio
 - format = wav, codage = pcm, fréquence: ≥ 44.1 Khz, taille de l'échantillon ≥ 16 bits
 - Vidéo
 - En attente des prescriptions du CINES
- Les annotations
 - format = xml, encodage = utf-8, différents schémas et dtd
 - format text seul, encodage ascii ou utf-8
 - format jpeg (pour les scans de manuscripts)
- Des métadonnées
 - Format xml, encodage utf-8, schémas d'OLAC

~

Exemple d'une ressource type

Ethnologue > Web version > Country index > Asia > China > Jiarong

Jiarong

A language of China

ISO 639-3: [jya](#)

Population 83,000 (1999 Sun Hongkai). 25,000 monolinguals. Ethnic population: 151,197 including 139,000 in Situ Jiarong, 12,197 in Chabao and Sidaba (1993 Lin).

Region North central Sichuan. Situ is in the traditional territory of four chieftaincies: Zhuokeji, Suomo, Songgang, Dangba. Chabao is in the northeastern corner of Maerkang county, at Longerjia, Dazang, and Shaerzong townships in Chabao District. Sidaba is in Caodeng, Kangshan, and Ribu townships in Sidaba District of Maerkang County. Some outlying Sidaba communities are to the southwest in the Rongang townships, at the southwestern corner of the Aba County. The Jiarong River between Wuyi and Shili townships in Rangtang County, at the confluence of the Seda and Duke rivers in Seda County.

Alternate names Jyarung, Gyarong, Gyarung, Rgyarong, Chiarong, Jarong

Dialects Chabao (Dazang, Northeastern Jiarong), Sidaba (Caodeng, Northern Jiarong). Subdialects of Situ are: Maerkang, Lixian, Jinchuan, Xiaojin; of Chabao: Western and Northern are fairly similar and differ greatly from Eastern and Northern Jiarong, 60% between Western and Northern Jiarong.

Classification [Sino-Tibetan](#), [Tibeto-Burman](#), [Tangut-Qiang](#), [rGyarong](#)

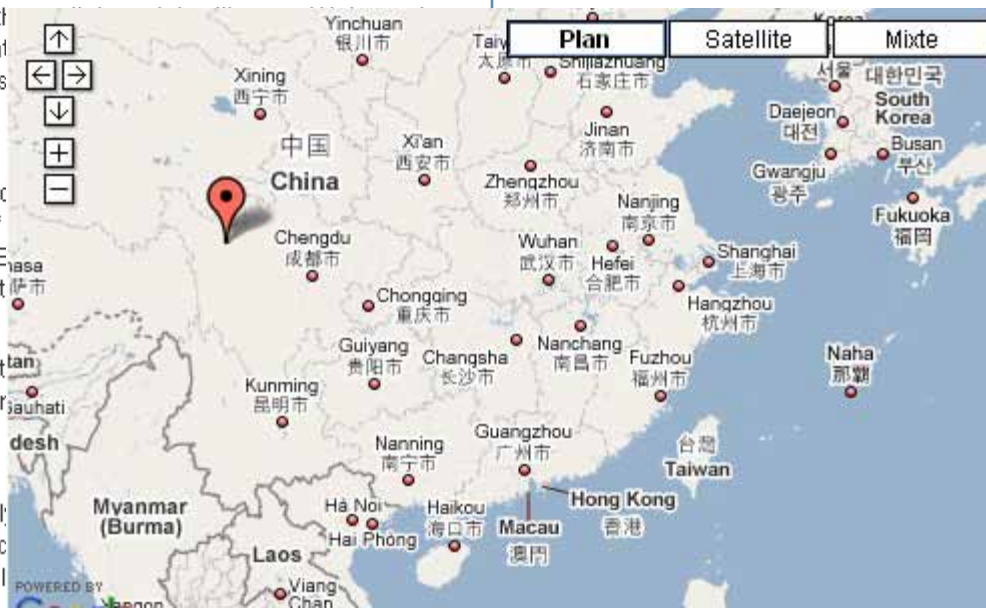
Language use Vigorous. All domains. All ages. Positive language attitude, but declining use of the language. 56,000 use Chinese, 950 Tibetan, 50 Qiang.

Language development Radio programs. Dictionary.

Comments Part of the Tibetan nationality. SOV; phonologically and lexically similar to Pumi and Qiang; complex consonant clusters; limited riverine. Agriculturalists: apples, pears; lumbermen. Traditional

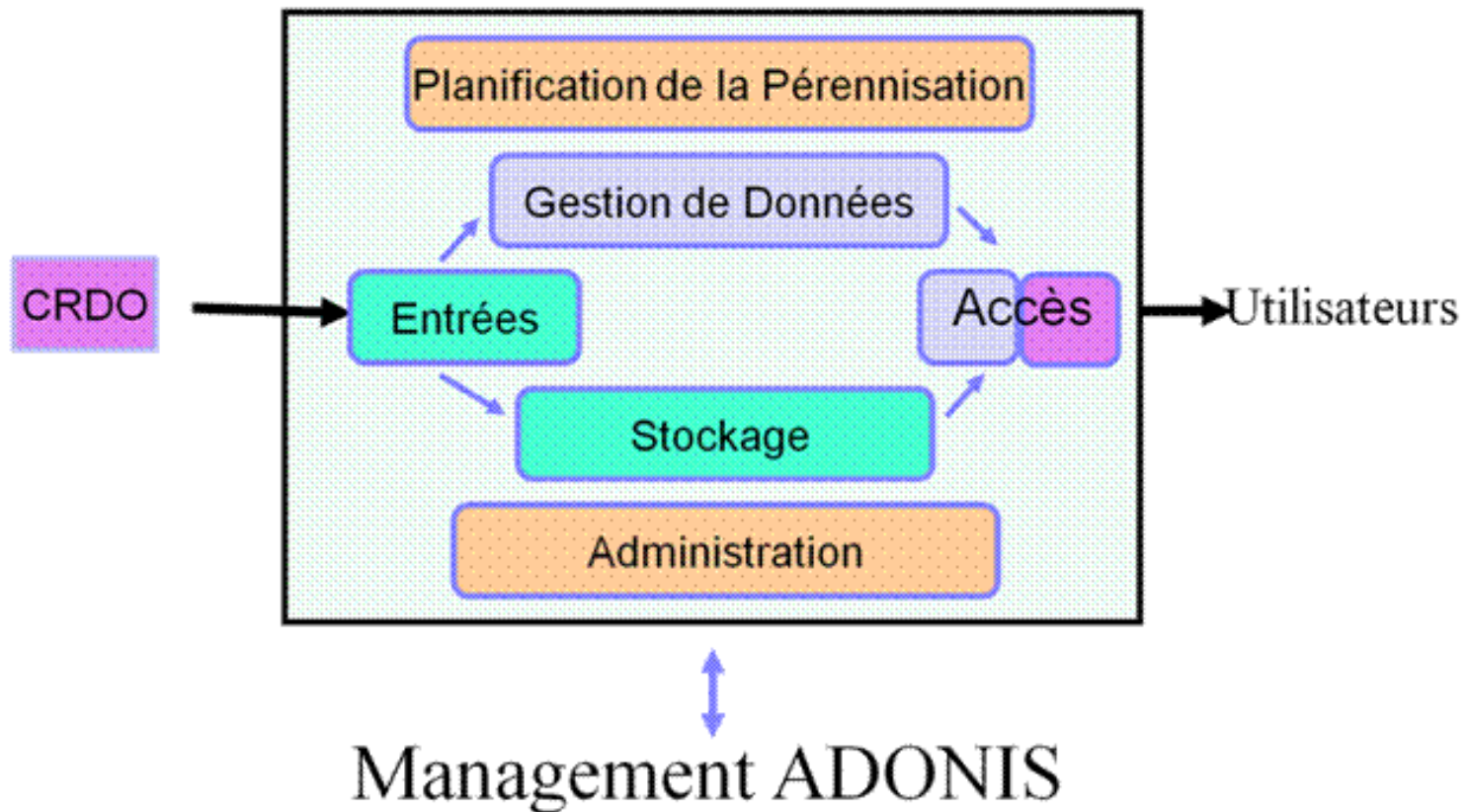
Récit de 7 minutes, enregistré en 2005, transcrits en 2007 par Guillaume Jacques (CNRS/CRLAO). Langue Japhug. Chine (Sichuan, Maerkang, Bar-khams)

Démo



La nouvelle organisation

- Fonctions sous responsabilité CINES
- Fonctions sous responsabilité IN2P3
- Fonctions sous responsabilité CINES ou IN2P3, chacun dans son domaine
De fonctionnalités principales
- Fonctions sous responsabilité CRDO

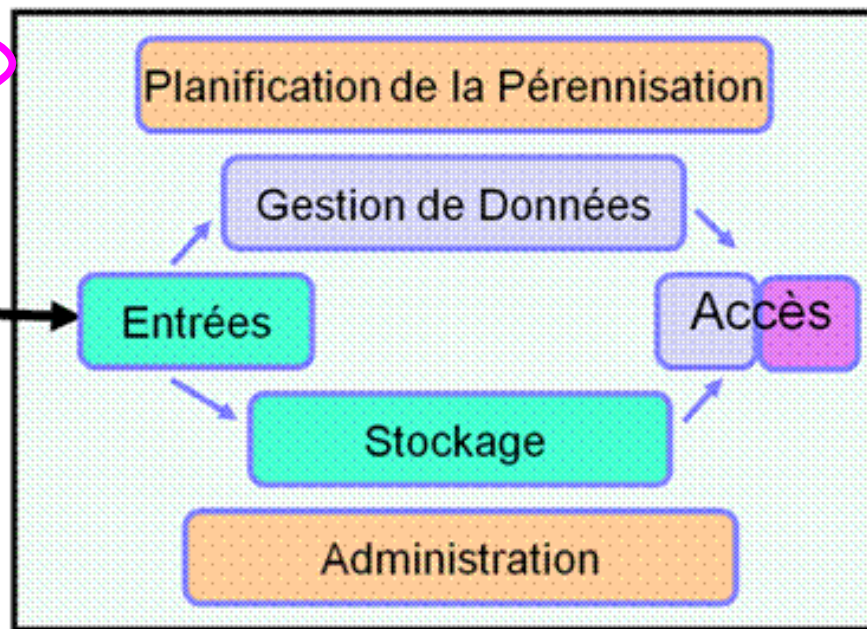


2. Les fonctions d'entrée

- Fonctions sous responsabilité CINES
- Fonctions sous responsabilité IN2P3
- Fonctions sous responsabilité CINES ou IN2P3, chacun dans son domaine
De fonctionnalités principales
- Fonctions sous responsabilité CRDO

Les fonctions de
pré-ingestion

CRDO

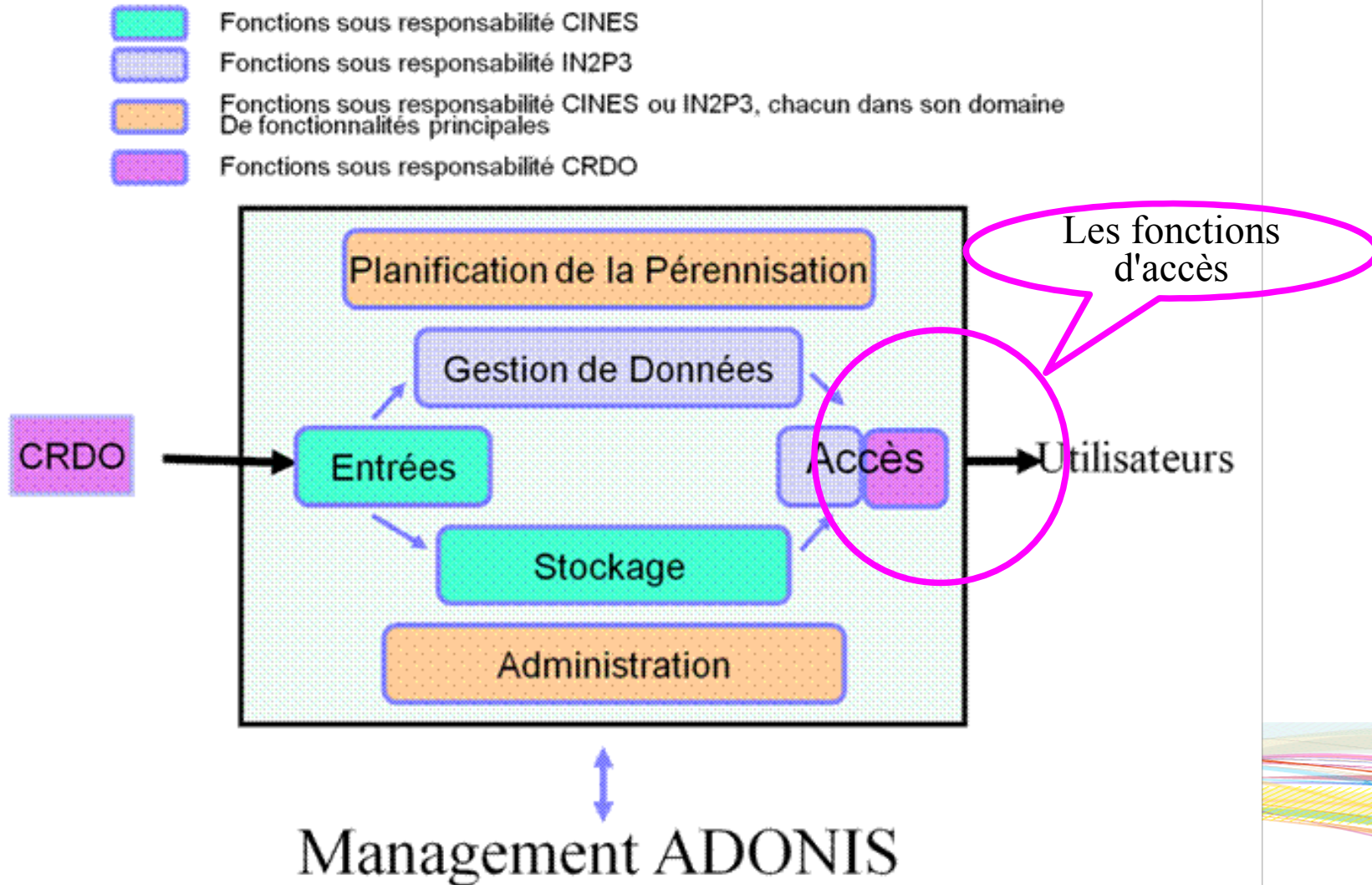


Management ADONIS

Pré-ingestion

- Interface avec les producteurs pour
 - Aide à l'édition des données (outils d'édition, formations)
 - Aide à la normalisation des données
 - Aide à la construction des métadonnées
- Prévalidation des données
 - un outil a été crée pour les fichiers audio: vérification du format, du codage, de la qualité utilisée...
 - Xerces pour la validation des fichier xml
 - Validation de l'exactitude et de la complétude des métadonnées par des documentalistes
- Constitution et versement du SIP au CINES lorsque tout est ok

3. Les fonctions d'accès



3. Les fonctions d'accès

- ✓ Des moteurs de recherche
- ✓ Des outils de consultation
- ✓ La Diffusion

3.1. Des moteurs de recherche

- Moteurs de recherche accessibles sur le site web du CRDO
 - Par simple mot-clef
 - Par catégories Dublin-core
 - Par une carte géographique (googlemaps)
 - Par un axe temporel (simile timeline)
- La recherche peut être effectuée par les moteurs de recherche des fournisseurs de services qui ont moissonné l'entrepôt du CRDO
 - En général exploite les catégories Dublin-Code (OAIster) ou OLAC (LinguistList)

Recherche par mot-clef

Groupe Parisien du Centre de Ressources pour la Description de l'Oral - Mozilla Firefox

Fichier Édition Affichage Historique Marque-pages Outils ?

http://crdo.risc.cnrs.fr/exist/crdo/

Centre de Ressources pour la Description de l'Oral

Le CNRS Accueil SHS Autres sites CNRS

CRDO
centre de ressources pour la description de l'oral

CRDO groupe parisien

Les projets

L'archive ouverte

Documentation

Communauté

Annuaire

Rechercher

dans l'archive

nemi

Centre de Ressources pour la Description de l'ORAL (CRDO)
Le groupe parisien "Gestion documentaire et réservoir de données"

Les centres de ressources numériques (CRN) sont une initiative conjointe de la Direction de l'Information Scientifique et du département Sciences Humaines et Sociales - SHS.

Le Centre de Ressources sur la Description de l'Oral (CRDO) est le Centre de Ressources Numériques centré sur les ressources orales. Il a été composé à partir de deux propositions portées respectivement par les deux laboratoires Lacito et LPL.

Le groupe parisien du CRDO correspond à la proposition portée par le Lacito et à laquelle sont associés différents laboratoires membres. Ce groupe est plus particulièrement centré sur la gestion documentaire des ressources orales et la constitution d'un réservoir de données.

L'archive ouverte
Rôle d'une archive ouverte

Une *archive ouverte* est un réservoir dans lequel le dépôt des données est effectué par l'auteur lui-même ou bien par une personne ou une institution, à partir du moment où celle-ci peut garantir qu'elle détient les droits ou l'autorisation des ayants-droits pour disposer des données de la sorte. Le dépositaire doit documenter son dépôt par l'intermédiaire d'un formulaire en-ligne. Une fois ce formulaire rempli et les données transmises, et après avoir passé un contrôle qualité, ces ressources sont entreposées dans un

Actualité...

Statistiques:
Actuellement sont consultables dans l'archive: **1371** enregistrements public et **2836** en accès protégé pour un total de **90** langues différentes.

Annnonce du colloque CatCod 2008 "Premières rencontres internationales pour le Catalogage et le Codage de corpus oraux" les 4-5 Décembre 2008 à l'Université d'Orléans, France
<http://www.catcod.org/>

Le TGE-Adonis lance une expérimentation sur le stockage, la conservation et l'accès à des données orales. Cette expérimentation devrait relier le CRDO, le centre de calcul de l'IN2P3 et le CINES dans une organisation sur le modèle de la norme OAIS... ([en savoir plus](#))

Terminé

démarrer

TextPad - [C:...]

Courrier entra...

Communications

2 OpenOffic...

Groupe Parisi...

2 gimp-2.2

FR

19:21

Recherche par categorie D-C

Groupe Parisien du Centre de Ressources pour la Description de l'Oral - Mozilla Firefox

Fichier Édition Affichage Historique Marque-pages Outils ?

http://crdo.risc.cnrs.fr/exist/crdo/rechercher.xq

copie fenêtre

Centre de Ressources pour la Description de l'Oral

Le CNRS Accueil SHS Autres sites CNRS

CRDO

centre de ressources pour la description de l'oral

CRDO groupe parisien

Les projets

L'archive ouverte

Documentation

Communauté

Annuaire

Rechercher dans l'archive

Rechercher: vanuatu

Catégories: Coverage

Opérateurs: near

Langues: Toutes

Résultats affichés: 15

Valider

dans vos recherches les jokers *, ? ainsi que les rangées de caractères du type [a-zA-Z]

titre	déposant	langue	fichiers liés
-------	----------	--------	---------------

W3C XHTML 1.0

Site hébergé par le RISC

Terminé

démarrer

TextPad - [C:\apac...

Courrier entrant po...

Communications

OpenOffice.org...

Groupe Parisien du ...

FR

20:03

Résultats d'une recherche

Groupe Parisien du Centre de Ressources pour la Description de l'Oral - Mozilla Firefox

Fichier Édition Affichage Historique Marque-pages Outils ?

http://crdo.risc.cnrs.fr/exist/crdo/rechercher.xq?term1=vanuatu&field1=Coverage&mode1=near&langue=*&showmany=15

copie fenêtre

Centre de Ressources pour la Description de l'Oral

Le CNRS Accueil SHS Autres sites CNRS

CRDO

centre de ressources pour la description de l'oral

CRDO groupe parisien

Les projets

L'archive ouverte

Documentation

Communauté

Annuaire

Rechercher dans l'archive

Rechercher: vanuatu

Catégories: Coverage Opérateurs: near Langues: Toutes Résultats affichés: 15

Vous pouvez utiliser dans vos recherches les jokers *, ? ainsi que les rangées de caractères du type [a-zA-Z]

Ressources 1 à 10 (total: 10 trouvées)

fichier	titre	déposant	langue	fichiers liés
	A myth of origin: t...	François, Alexandre	Araki	consulter
	How the village of ...	François, Alexandre	Araki	consulter
	Hunting for coconut...	François, Alexandre	Araki	consulter
	The Rat, the Hawk a...	François, Alexandre	Araki	consulter
	The elves	François, Alexandre	Dorig	consulter
	The danse of the Sp...	François, Alexandre	Koro	consulter
	Wuwut, the orphan	François, Alexandre	Lakon	consulter
	Get the trickster	François, Alexandre	Mwesen	consulter
	Romanmangan, the fa...	François, Alexandre	Mwotlap	consulter
	The gecko and the t...	François, Alexandre	Mwotlap	consulter

W3C XHTML 1.0

Site hébergé par le RISC

Terminé

démarrer

TextPad - [C:\apac...

Courrier entrant po...

Communications

OpenOffice.org...

Groupe Parisien du ...

FR

20:04

Recherche par point d'enquête

CRDO: Rechercher une ressource avec une carte géographique - Mozilla Firefox

Fichier Édition Affichage Historique Marque-pages Outils ?

http://crdo.risc.cnrs.fr/exist/crdo/geo.htm

copie fenêtre

Centre de Ressources pour la Description de l'Oral

Le CNRS Accueil SHS Autres sites CNRS

CRDO

centre de ressources pour la description de l'oral

CRDO groupe parisien

Les projets

L'archive ouverte

Documentation

Communauté

Annuaire

Rechercher dans l'archive

Terminé

Accueil > Chercher une ressource

Chercher une ressource par la géographie

Nous disposons pour un certain nombre de ressources, d'informations sur le lieu précis où a été effectuée l'enquête. Pour ces ressources là, et uniquement celles-là, il vous est possible d'y accéder directement par la carte géographique en cliquant sur les points rouges. Lorsque vous cliquez sur l'un d'eux, une bulle d'informations s'ouvre et vous donne les différentes ressources disponibles pour ce point précis. Pour consulter une de ces ressources, cliquez sur son titre.

[patiencez jusqu'à ce que les points rouges apparaissent sur la carte]

Breton

Submit

Plan Satellite Mixte

- Breton : [NALBB: Guerlesquin \(3\)](#)
- Breton : [NALBB: Guerlesquin \(4\)](#)

Plymouth

Brest

Ren

démarrer

TextPad - [C:\apac...

Courrier entrant po...

Communications

2 OpenOffice.org...

CRDO: Rechercher...

FR

20:56

Recherche par date de création

CRDO: Rechercher une ressource avec sa date d'enregistrement - Mozilla Firefox

Fichier Édition Affichage Historique Marque-pages Outils ?

http://crdo.risc.cnrs.fr/exist/crdo/time.htm

copie fenêtre

CRDO
centre de ressources pour
la description de l'oral

CRDO groupe parisien

Les projets

L'archive ouverte

Documentation

Communauté

Annuaire

Rechercher dans l'archive

* []

Accueil > Chercher une ressource

Chercher une ressource par sa date d'enregistrement

Pour la plupart des ressources d'enregistrement nous disposons d'informations la date de création. Pour ces ressources là, et uniquement celles-là, il vous est possible d'y accéder directement par la présentation temporelle en cliquant sur les points d'ancrage. Lorsque vous cliquez sur l'un d'eux, une bulle d'informations s'ouvre et vous donne les différentes ressources disponibles pour ce moment précis. Pour consulter une de ces ressources, cliquez sur son titre.

Timeline © SIMILE

2004 2005 2006

Bedja

Bedja: [Le Bedja et le lait](#)
Bedja: [Le Bedja et le lait](#)
Bedja: [Adam, le roi des animaux et le diable](#)
Bedja: [Adam, le roi des animaux et le diable](#)
Bedja: [La chouette et Suleyman](#)
Bedja: [La chouette et Suleyman](#)
Bedja: [Le joueur](#)

Terminé

démarrer

TextPad - [C:\apac...

Courrier entrant po...

Communications

2 OpenOffice.org...

CRDO: Rechercher...

FR

20:58

3.2. Des outils de consultation

- Feuilles de styles xslt pour présenter les transcriptions et les métadonnées
- Outil de synchronisation (plugin + javascript)
- Outil de découpage d'extrait sonore (servlet java)

Consultation des métadonnées

Groupe Parisien du Centre de Ressources pour la Description de l'Oral - Mozilla Firefox

Fichier Édition Affichage Historique Marque-pages Outils ?

http://crdo.risc.cnrs.fr/exist/crdo/details.xq?id=crdo-AKR_HISTOIRE_ARAKI_SOUND

copie fenêtre

Centre de Ressources pour la Description de l'Oral

Le CNRS Accueil SHS Autres sites CNRS

CRDO
centre de ressources pour la description de l'oral

CRDO groupe parisien
Les projets
L'archive ouverte
Documentation
Communauté

Annuaire

Rechercher dans l'archive

oai:crdo.vjf.cnrs.fr:crdo-AKR_HISTOIRE_ARAKI_SOUND

General Description:

Title: A myth of origin: the island of Araki

Publisher(s): CNRS/LACITO

Contributor(s): CNRS/Fédération de recherche 2559: Typologie et universaux linguistiques (sponsor)
François, Alexandre (researcher)
François, Alexandre (depositor)
Lele Moli (speaker)

Coverage: [spatial] Vanuatu, Espiritu Santo, Luganville; [spatial] TGN:7023855; [spatial] Point: east=167.1667; north=-15.5333

Date(s): created: 1997-11

Type(s): Sound
(primary text)
(narrative)

Subject(s): Araki (Ethnologue: akr)

Language(s): Araki (Ethnologue: akr) (for translations and/or glosses)

Format(s): (IANA MIME Media Type: audio/x-wav)
duration: 0:1:24

Access Description:

Rights: Copyright (c) François, Alexandre

Identifier: http://crdo.risc.cnrs.fr/data/francois/AKR_HISTOIRE_ARAKI.wav

Relation(s): [isRequiredBy] [oai:crdo.vjf.cnrs.fr:crdo-AKR_HISTOIRE_ARAKI](http://crdo.vjf.cnrs.fr:crdo-AKR_HISTOIRE_ARAKI)

W3C XHTML 1.0

Site hébergé par le RISC

Terminé

démarrer

TextPad - [C:\apac...

Courrier entrant po...

Communications

2 OpenOffice.org...

Groupe Parisien du ...

FR

20:53

Consultation audio + transcription

Groupe Parisien du Centre de Ressources pour la Description de l'Oral - Mozilla Firefox

Fichier Édition Affichage Historique Marque-pages Outils ?

http://crdo.risc.cnrs.fr/exist/crdo/leconte-tilouine/nep/crdo-NEP_DAURE.xhtml?plugin=real&mediaUrl=http://crdo.risc.cnrs.fr/data/divers/DAURE.wav

Centre de Ressources pour la Description de l'Oral

Le CNRS Accueil SHS Autres sites CNRS

CRDO
centre de ressources pour la description de l'oral

CRDO groupe parisien
Les projets
L'archive ouverte
Documentation
Communauté

Annuaire

Rechercher dans l'archive

Site hébergé par le RISC

Accueil > Chercher une ressource > Métadonnées > Voir le text

Texte interlinéaire: ☒ Traduction: ☒ [basculer sur l'interface standard]

[karaoke: ☒

S1 **Laphie:** एक ठाउमा त्पो दाउरा बड्डो दाउराई चीरिर -
Dans un pays, un vieux bûcheron coupait du bois...
एक ठाउँमा त्पो दाउरे बुद्धो दाउरा चीरिर -

S2 **Prem:** अछामी भाषाबाट भने पनि हुन्छ ।
Vous pouvez la raconter en langue achhami.
प्रेम - अछामी भाषाबाट भने पनि हुन्छ ।

S3 **Laphie:** त्पो दउरा चीरिर आफ्तु गुजेरा गर्छ ।
D'accord. Il vivait en coupant du bois.
हजुर है, त्पो दाउरा चीरिर आफ्तु गुजेरा गर्छ ।

S4 **Laphie:** बड्डी बाही [थे-] कुच्चा बाटुपा गाउमा लीने गाउकै एउटा -
Sa vieille, elle fabriquait des balais, elle les emmenait au village
बुद्धीचाही, थे- कुचाहरू बाटुने, गाउँमा ल्याउने, गाउँकै एउटा -

S5 **Prem:** अछामी भाषा, अछामी भाषाबाट भने पनि हुन्छ ।
Vous pouvez la raconter en langue achhami.
अछामी भाषा, अछामी भाषाबाट भने पनि हुन्छ ।

S6 **Laphie:** गाउमा लीएर त्पो भु कुच्चा बेच्चाको ल्पाएर खर्च बनाउने दुई जना बसन्पा खाने पीने तेस्तै गरेर
बसिहेका छीया ।
Elle les emmenait au village et là elle les vendait, et avec l'argent les deux trouvaient à se
loger, à manger et à boire

Consultation vidéo + transcription

[Fichier](#)
[Édition](#)
[Affichage](#)
[Historique](#)
[Marque-pages](#)
[Outils](#)

[<](#)
[>](#)
[C](#)
[X](#)
[H](#)
[P](#)
[G](#)
[Google](#)

[Centre de Ressources pour la Description de l'Oral](#)
[Le CNRS](#)
[Accueil SHS](#)
[Autres sites CNRS](#)



[CRDO](#)
 centre de ressources pour la description de l'oral

[CRDO groupe parisien](#)
[Les projets](#)
[L'archive ouverte](#)
[Documentation](#)
[Communauté](#)

[Annuaire](#)
[Rechercher dans l'archive](#)

[W3C XHTML 1.0](#)

Site hébergé par le [RISC](#)

Terminé

2008-12-15 IN2P3...

Groupe Parisien du ...

FR

22:55

3.3. Publication référencement

- Le CRDO s'est organisé sous la forme d'un entrepôt OAI. Le protocole OAI-PMH est implémenté sous forme d'une servlet java.
- Des fournisseurs de ressources moissonnent régulièrement l'entrepôt et offrent des moteurs de recherche dans un ensemble plus large d'entrepôts
- Des portails web moissonnent régulièrement l'entrepôt pour offrir des consultations particulières des ressources ou des vues partielles de l'entrepôt
 - Corpus de la parole: portail sur les langues de France
 - Dallith: portail sur les langues de la zone himalayenne (archives du CRDO + Univ of Virginia)
 - Lacito: portail sur les ressources des chercheurs de ce labo

http://corpusdelaparole.culture.fr/

Corpus de la parole - Nemi - Mozilla Firefox

Fichier Édition Affichage Historique Marque-pages Outils ?

http://corpusdelaparole.culture.fr/spip.php?article27

CRDO: L'OAIS - Reference Model for a... Centre de Calcul IN2P3 - CNRS Corpus de la parole - Nemi

CORPUS DE LA PAROLE

Comment on parle en France aujourd'hui

→ accueil → pourquoi des corpus ? → sons et images → rechercher → projets et liens

Nemi

Némi, ou langue de Hienghène. Dans la langue, nemi est un pronom personnel inclusif de première personne du pluriel qui signifie « nous autres (vous exclus) ». Selon M. Leenhardt ce nom avait autrefois une valeur englobante pour désigner tout un ensemble de langues parlées sur la Grande Terre, des deux côtés de la chaîne centrale, à hauteur de Voh et de Hienghène : « la langue de Némi comprend les dialectes des régions de Hienghène, Pinjé, Temala, Faténawé, Pakèp, l'île Gatope, et tout le territoire de Voh à Hienghène s'appelle, chez les indigènes, le pays du parler Némi. » (LDAM 1946, p.xiv).

Écoutez !

Suivez la traduction :



- ▶ **Hingoo-ng bac ma dangem.**
Je vais raconter l'histoire de deux lochons d'eau douce, bac et dangem.
- ▶ **Yelu moo, moo... raan yelu pe-pala.**
Ils sont là et discutent.
- ▶ **Ye peei ru bac peei "Pia! dangem, pmwane hun davec o do li pmwaa?".**
Bac demande "Dis donc ! dangem, en cas d'inondation, que ferais-tu ?"

Rechercher : gare

Terminé

démarrer

TextPad - [C:\apa... Courrier entrant p... Communications 2008-12-15_IN2P... Corpus de la parol... FR 08:20

http://dallith.vjf.cnrs.fr/

DALLITH - Mozilla Firefox

Échier Édition Affichage Historique Marque-pages Outils ?

http://dallith.vjf.cnrs.fr/fr/show_text.php?path=%2Fthdl%2F00008_01-a-lucky-dream_07.xml&lg=bod& G jacobson

Portail Culture Semaphore MCC Pages jaunes SNCF

Espace S.O.L.O.N. Place de Marché Int... Capture d'écran / C... Gmail - Boîte de réce... these-japhug.pdf (... DALLITH

  **DALLITH**  

[ACCUEIL](#) | [RESSOURCES PAR LANGUE](#) | [A PROPOS DU DALLITH](#) | [AIDE & INSTRUCTIONS](#)

Tibétain : " 01 A Lucky Dream "
2:14 mn:sec - 1989

> **Metadonnées** < [Index des ressources](#)

Glose: ☒ Traduction: ☒ - Karaoke: ☒

S1   **phur sgron/_ltoś dang {da }**
Phurdroen, look!
普珍, 看!

S2   **de ring gnam yag po 'dra cig {ra cig}mi**
'dug kas/
Isn't it a nice day today?
今天的天气多好啊, 是吧!

S3   **'ong red zhag_nga yang {ya'i }gnam la**
bltas bśdad yod/
Yes, it is. I've also been looking at the sky.
是啊, 我也在往天上看呢.

S4   **'tshig yang {ya'i }ma 'tshig 'khyag yang**

Rechercher : Japhug Suivant Précédent Surligner tout Respecter la casse Bas de la page atteint, poursuite au début

Terminé