Laboratoire ligérien
de linguistique

L'utilisation du modèle du Linked Open Data dans la plateforme Cocoon

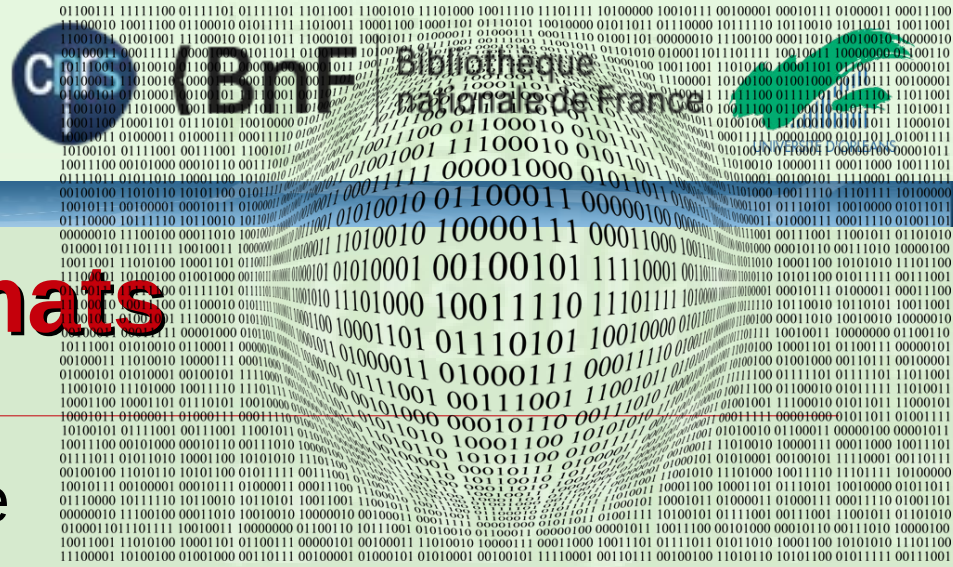
Michel Jacobson

Laboratoire ligérien de linguistique

Le Linked Open Data

- Faciliter la réutilisation
  - licences
  - formats des fichiers
  - identification
  - description (métadonnées)
  - liens entre les ressources

Les formats de fichiers dans Cocoon



Les formats

- L'information numérique est codée
- Pour être représentée et manipulée elle doit être décodée
- La manière de coder et de décoder est spécifiée dans un document (spécifications du format). La mise en œuvre de ces spécifications est faite dans des logiciels. Eux-même sont codés et dépendant pour leur exécution d'un environnement complexe et soumis à obsolescence rapide.
- Deux approches possibles pour la conservation
  - Émulation
  - Migration

Les formats

- Émulation
  - Privilégie la conservation des usages
    - Conservation des logiciels et de leurs conditions d'exécution
    - Peu de compétences
- Migration
  - Privilégie la conservation les informations contenus dans les fichiers
    - Identification des formats sur lesquels on peut engager sa responsabilité
    - Suivi des évolutions du marché et des usages
    - Vérification de la validité des données qu'on accepte (conformité des données avec les spécifications de son format)

Les formats

- Critères pour choisir un format :
  - Le format est-il spécifié indépendamment d'un outil ?
  - Les spécifications sont-elles accessibles ou secrètes ?
  - Existe-t-il des entraves juridiques à son utilisation ?
  - Le format peut-il être utilisé par différents outils ou est-on lié à un seul éditeur ?
  - Existe-t-il des outils de conversion de ce format vers d'autres formats équivalents ?
  - Existe-t-il des outils de validation pour vérifier la conformité des fichiers de données ?
  - Le format a-il fait l'objet d'une standardisation ou d'une normalisation ?

Les formats

- Choix des formats pour Cocoon

Audio	WAV ou FLAC avec un encodage PCM
Vidéo	MPEG4 avec encodages H264 et AAC
Annotations	Du plus structuré au moins structuré : XML, TEXT, PDF, JPEG

  - Cf. les guides du SIAF/HN/CINES

- Pour les autres formats des conversions seront envisagées
- Pour des usages web, des versions sont générées dans des formats plus légers
  - Audio : mp3 128Kbits
  - Vidéo : ogg/vorbis/theora 1500Kb audio 128Kb 320x240

Les formats

- Exemple de format utilisé par les producteurs « textGrid »
  - Le format est spécifié par l'éditeur du logiciel Praat.
  - Les spécifications ne suivent pas de formalisme ni de processus d'élaboration particulier.
  - Le format n'est pas reconnu par les outils d'identification qui vont y voir au mieux un simple document textuel
    - File →
      UTF-8 Unicode C++ program text, with CRLF line terminators
      ou ASCII C++ program text, with CRLF line terminators
    - DROID →
      inconnu
      ou si on change l'extension en .txt : Plain Text File (mimeType text/plain)

Les formats

- Le format « textGrid » (suite)
  - Il n'existe pas d'outil de validation
  - L'encodage des caractères est implicite
  - …
  - Le choix de ce format oblige à compléter les manques sois-même.
  - Par exemple, si on dépose des fichiers textGrid au CINES, ce dernier va les reconnaître comme de simples fichiers texte et les conserver et les décrire comme tels. Ce qui sera conservé sera donc uniquement les caractères qu'ils contiennent. Seule cette information fera par exemple l'objet de futures conversions.

Les formats

- Le format « textGrid » (suite et fin)
  - Traitement de ce format dans Cocoon
    - Stockage du fichier textGrid (pour réutilisation par des experts)
    - Conversion dans un format XML (pour faciliter la consultation)
    - Conversion en TEI par l'outil « tei_corpo » (pour faciliter l'interopérabilité)
    - Conservation
      - *Du format d'origine textGrid ?*
      - *Du format TEI si validation du résultat de la conversion ?*
      - *Du format XML de consultation ?*

Les identifiants pérennes dans Cocoon

Les identifiants

- Les ressources documentaires de Cocoon (enregistrements, annotations et collections) sont identifiées par :
  - Des identifiants OAI
  - Sont ajoutés dans les métadonnées d'autres identifiants :
    - Les identifiants ARK affectés par le CINES lors de leur archivage
    - Les identifiants HANDLE affectés par ISIDORE lors de de son moissonnage de l'entrepôt
    - Les autres identifiants apportés par le producteur: par exemple les anciennes cotes

Les identifiants

- Les identifiants OAI peuvent être utilisés pour citer les ressources en suivant le schéma POI (PURL-based Object Identifier)
  - Il s'agit d'une URL distincte de l'URL cible, mais dont l'activation redirige vers celle-ci. Ce mécanisme permet de palier les éventuels déménagements et changements de nom de domaine.
  - Les autres schémas associés aux identifiants ARK et HANDLE reposent sur les mêmes principes
    - Dissociation de l'adresse publiée et de l'adresse cible ;
    - Existence d'un mécanisme de redirection
      - *http://purl.org/poi/crdo.vjf.cnrs.fr/crdo-LEG_0014_SOUND*
      - *http://hdl.handle.net/hdl:10670/1.fqvtpe*
      - *http://cocoon.huma-num.fr/exist/crdo/ark:/87895/1.5-125831*

Les identifiants

- La pérennité d'un identifiant est généralement déterminée par le mode de gouvernance du système dans lequel il a été attribué, c'est-à-dire par l'organisation choisie par le producteur de données pour attribuer et maintenir ses identifiants

Type d'identifiant	Organisation qui attribue l'identifiant local	Organisation qui gère le système
POI	COCOON	Internet Archive
ARK	CINES	California Digital Library
HANDLE	ISIDORE	Corporation for National Research Initiatives

Les identifiants

- Les autres ressources sont identifiées dans Cocoon par des URI maintenus dans des référentiels externes
  - Les personnes (auteurs) : VIAF
  - Les lieux d'enregistrement : Geonames
  - Les thèmes abordés dans les enregistrements : RAMEAU
  - Les langues : Lexvo
  - …

Laboratoire ligérien
de linguistique

cnrs

BnF Bibliothèque nationale de France

UNIVERSITE D'ORLEANS

Les métadonnées dans Cocoon

Les métadonnées

- Les classes manipulées
  - Les ressources documentaires (enregistrements, annotations et collections)
    - Les collections sont des regroupements de ressources et ou de collections. Les collections sont ouvertes, peuvent s'accroître. Une même ressources peut appartenir à plusieurs collections à la fois.
  - Les autres ressources (personnes, organisations, lieux, concepts, langues...)

Les métadonnées

- Un modèle de description unique
  - Europeana data model (EDM)
  - Embarque des vocabulaires pré-établis largement connus
    - Dublin-Core
    - SKOS
    - OAI-ORE
    - FOAF, …

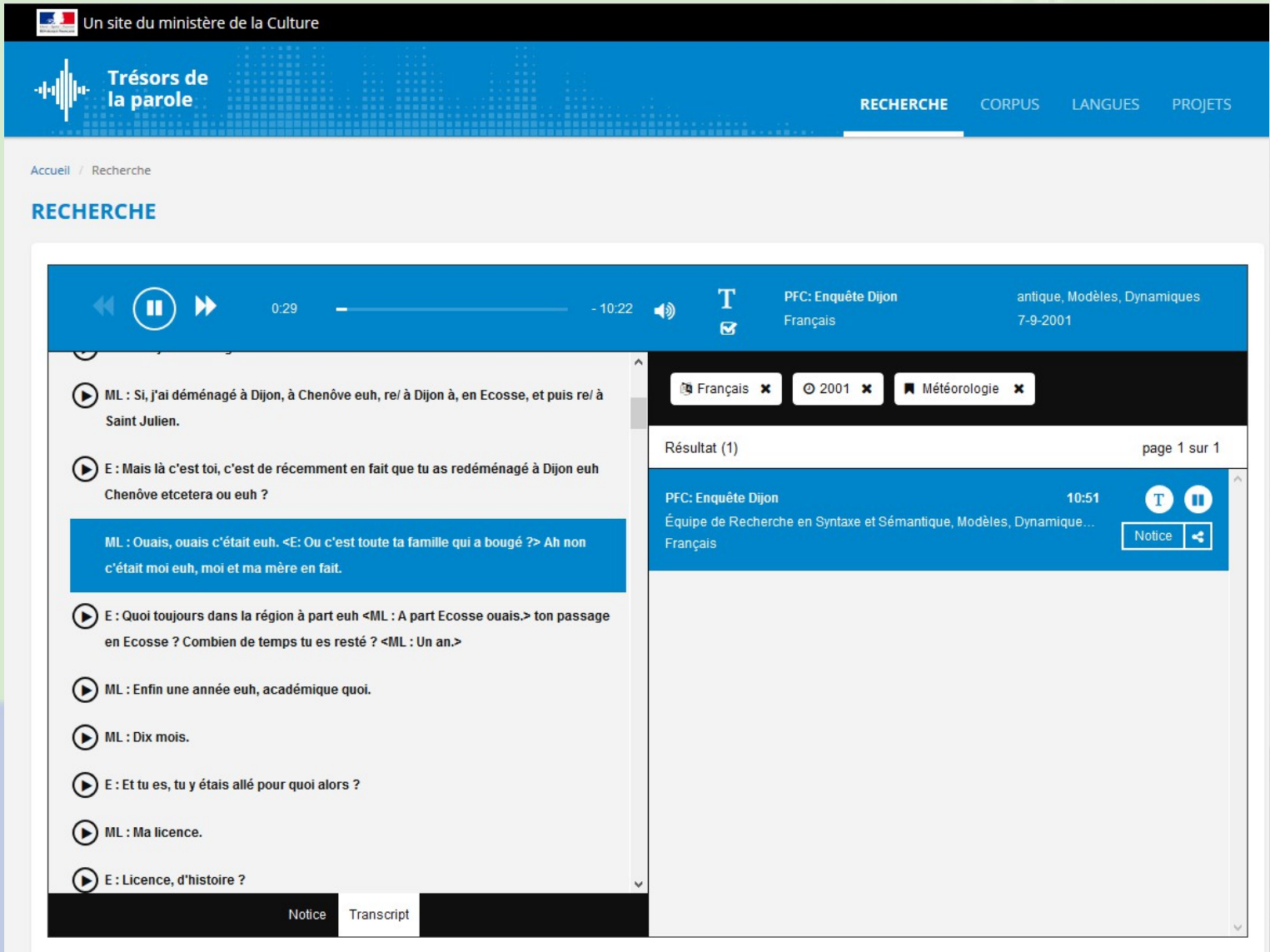
Les métadonnées

- Pour faciliter la réutilisation des métadonnées
  - Elles sont exprimées en RDF
  - Les données sont liées à des référentiels largement connus et utilisés (VIAF, Geonames, RAMEAU, Lexvo)
  - Les URI sont « déréférençables »
  - Un SparqlEndpoint permet de formuler des requête sur tout l'entrepôt.

Quelques exemples de ce que ça permet de faire...

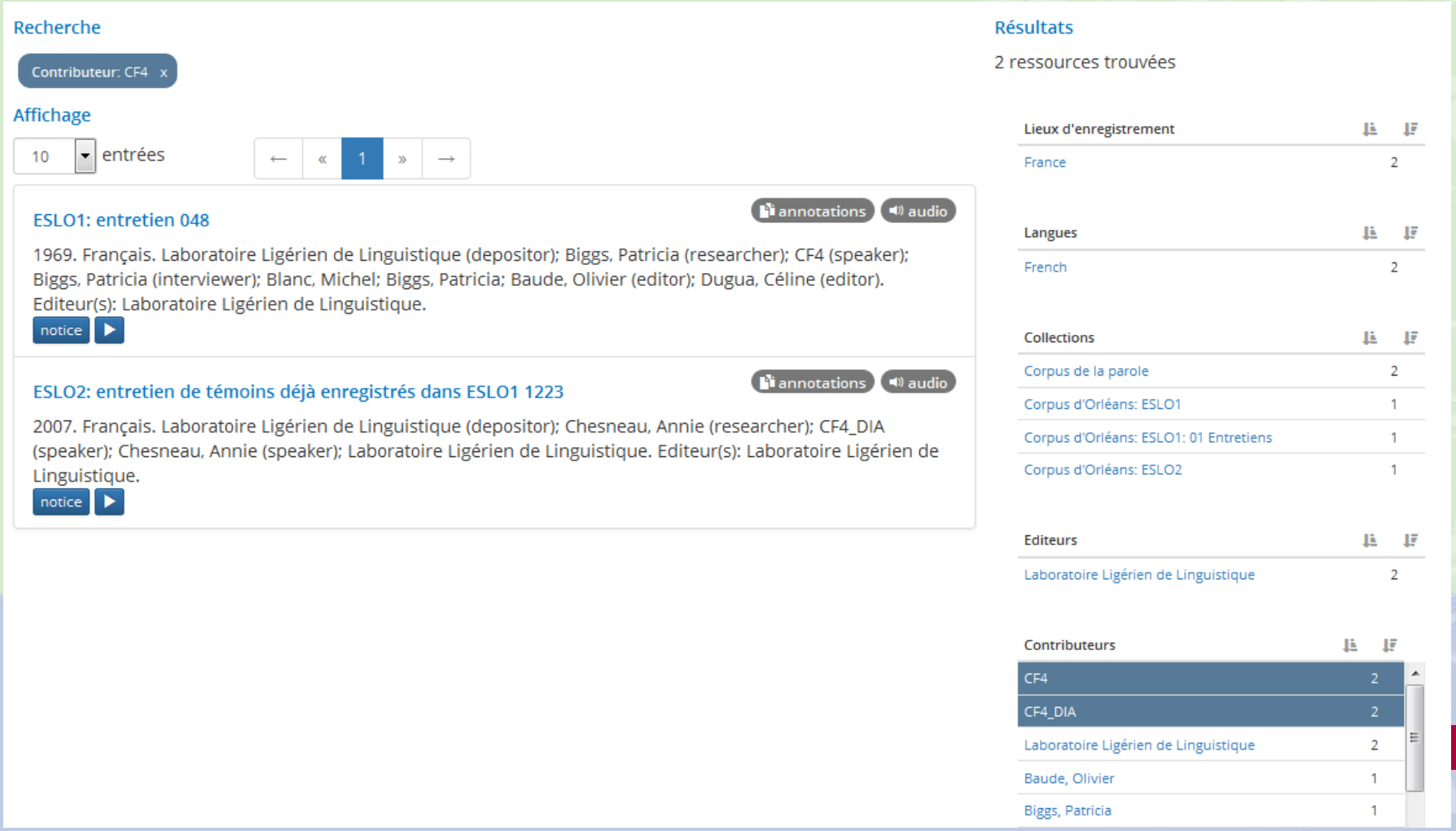
1. Faciliter la réutilisation par des tiers

- Un portail sur les langues de France



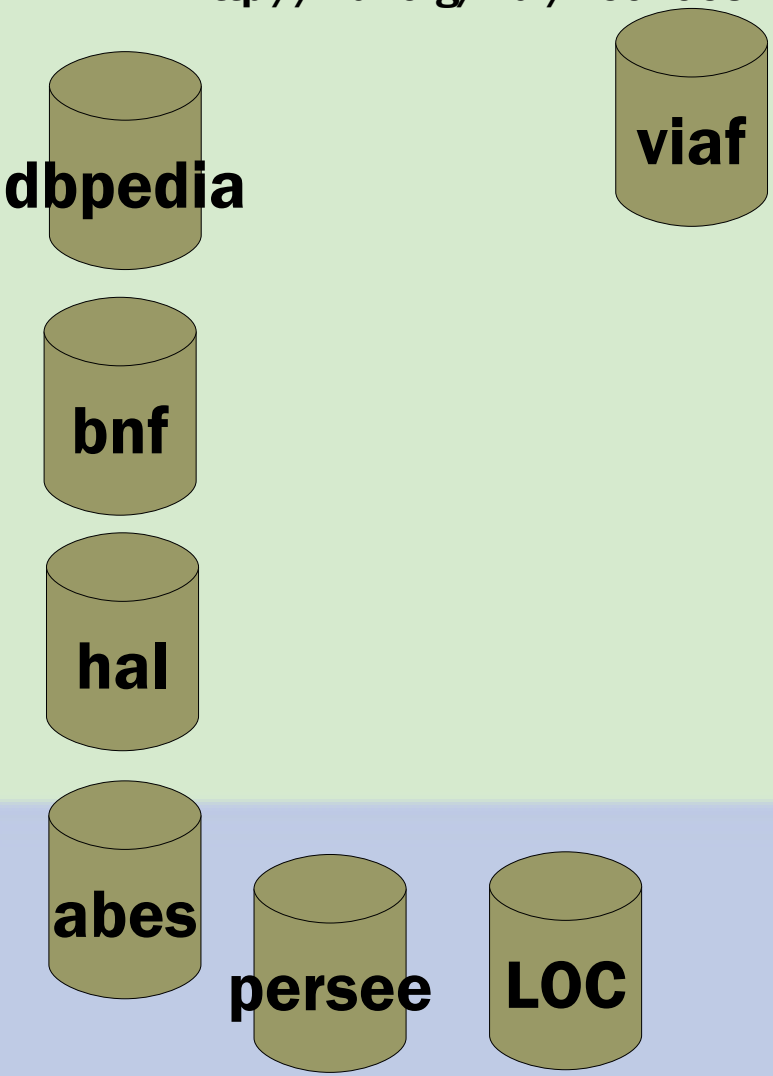
2. Améliorer la recherche

- <CF4> owl:sameAs <CF4_DIA>



3. Enrichir les descriptions

<http://viaf.org/viaf/2602065>



Tournadre, Nicolas

Résumé [en]: Nicolas Tournadre is a professor at the University of Provence specializing in morphosyntax and typology. He is a member of the LACITO lab of the CNRS. His research mainly deals with ergative morphosyntax and grammatical semantics of tense, aspect, mood and evidentiality. N. Tournadre is a specialist of Tibetic languages. Since 1986, he has carried out fieldwork on the Tibetan High Plateau, in the Himalayas and the Karakoram in China, India, Bhutan, Nepal and Pakistan. N. Tournadre taught at the Institute of Oriental Languages (Inalco), at the Paris 8 University, at the University of Virginia and conducted research in the Tibet Academy of Social Sciences. He obtained his Ph.D. in 1992 at the University of Paris III: Sorbonne Nouvelle under the supervision of Claude Hagège. In 2000, he was awarded the CNRS Bronze medal.He is a polyglot and has some knowledge (ranging from fluency to basic conversation) of languages belonging to 7 families (Romance, Slavic, Germanic, Tibetic, Sinitic, Indo-Iranian, Sign Language) : Russian, Polish, Ukrainian, Slovakian, English, German, Swedish, French, Spanish, Portuguese, Italian, Catalan, Hebrew, Mandarin Chinese, Standard Tibetan, Classical Tibetan, Kham, Amdo, Ladakhi, Balti, Dzongkha, Sherpa, Drejongke (or Lhoke) Hindi-Urdu, Persian, French Sign Language (LSF).

Résumé [fr]: Nicolas Tournadre est professeur de linguistique à l'université de Provence, spécialiste de morphosyntaxe et de typologie. Il est membre du laboratoire Lacito (CNRS).

En savoir plus sur la page wikipedia: http://en.wikipedia.org/wiki/Nicolas_Tournadre

« Tournadre, Nicolas » (Personne)
Voir la notice d'autorité sur Virtual International Authority File (VIAF): http://viaf.org/viaf/2602065

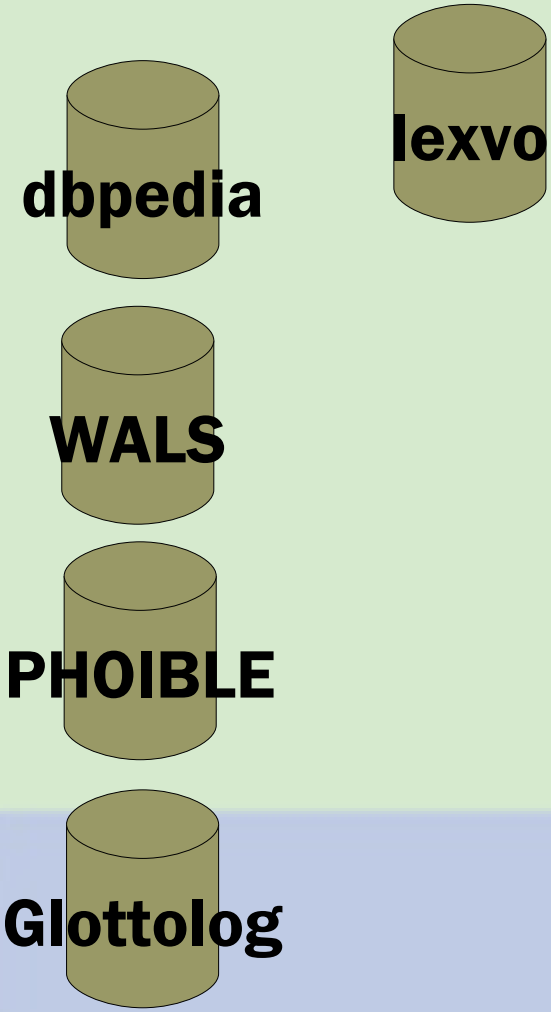
Rechercher s'il existe dans l'entrepôt, des ressources cet acteur en tant que contributeur ou éditeur

Références BnF (data.bnf.fr)

- Manuel de tibétain standard : langue et civilisation. Paris : Institut national des langues et civilisations orientales (Paris) , 1998 (cop.). 1 livre (567 p.) - 2 disques compacts : ill., couv. ill. en coul. ; 32 cm http://catalogue.bnf.fr/ark:/12148/cb38444231s
- Le clair miroir : enseignement de la grammaire tibétaine. Arvillard : Ed. Prajñā , 1992. VIII-271-XXXVII p. http://catalogue.bnf.fr/ark:/12148/cb35514438w
- Manuel de tibétain standard : langue et civilisation. Paris : l'Asiathèque-Maison des langues du monde , 2009. 1 vol. (606 p.-XVI p. de pl.) - 2 disques compacts (1 h 12 min 45 s, 44 min 45 s) : ill. en coul., couv. ill. en coul. ; 24 cm http://catalogue.bnf.fr/ark:/12148/cb421192732
- Le prisme des langues : essai sur la diversité linguistique et les difficultés des langues. Paris : l'Asiathèque-Maison des langues du monde , 2014. 1 vol. (349 p.) http://catalogue.bnf.fr/ark:/12148/cb43761949d
- L'ergativité en tibétain : approche morphosyntaxique de la langue parlée. Louvain ; Paris : Peeters , 1996. 393 p. http://catalogue.bnf.fr/ark:/12148/cb35850057f
- Le grand livre des proverbes tibétains. Paris : Presses du Châtelet , impr. 2006. 1 vol. (235 p.-[32] p. de pl.)

4. Déporter la maintenance de la description

<http://www.lexvo.org/resource/iso639-3/lzz>



Laz

Code ISO 639-3: lzz

Classification: Kartvelian > Georgian-Zan > Zan > Laz

Résumé [fr]: Le laze (en laze : Lazuri nena, ლაზური ნენა ; en géorgien : ლაზური ენა en turc : Lazca) est une langue caucasienne de la famille des langues kartvéliennes proche du géorgien, dont elle s'est séparée, avec le mingrélien, un millénaire avant l'ère chrétienne. Le laze n'a plus de forme écrite actuellement utilisée, les Lazes de Turquie (220 000) parlent laze mais utilisent le turc comme langue de communication écrite et interculturelle, tandis que ceux de Géorgie (30 000) utilisent le géorgien.

Abstract [en]: The Laz language (ლაზური ნენა, lazuri nena; Georgian: ლაზური ენა, lazuri ena, or ჭანური ენა, Çanuri ena, also chanuri ena; Turkish: Lazca) is a Kartvelian language spoken by the Laz people on the southeastern shore of the Black Sea. It is estimated that there are around 20,000 native speakers of Laz in Turkey, in a strip of land extending from Melyat to the Georgian border (officially called Lazistan until 1925), and about 2,000 in Georgia.

En savoir plus sur la page wikipedia: http://en.wikipedia.org/wiki/Laz_language

Ressources dans l'entrepôt

Pays ou régions dans lesquelles l'entrepôt dispose d'enregistrements

	Région	Code	Nombre
ℹ	Turkey	TR	1

Rechercher toutes les ressources de l'entrepôt sur cette langue: Laz

Propriétés (phonologiques, grammaticales, lexicales) de cette langue trouvées dans WALS

- *Nominal Categories*:Comitatives and Instrumentals = "Differentiation" (réf. Stolz 1996)
- *Nominal Categories*:Occurrence of Nominal Plurality = "All nouns, optional in inanimates" (réf. Kutscher 2001)
- *Nominal Syntax*:Nominal and Verbal Conjunction = "Identity" (réf. Kutscher 2001)
- *Simple Clauses*:Comparative Constructions = "Locational" (réf. Dirr 1928)
- *Simple Clauses*:Ditransitive Constructions: The Verb 'Give' = "Secondary-object construction"

Conclusion : Pourquoi on fait tout ça ?

Conclusion

- On fait tout ça...
  - Parce que ça évite de faire un travail déjà fait par d'autres (les référentiels)
  - Pour enrichir et contextualiser ses propres données par celles des autres
  - Parce que l'utilisation de ces standards permet d'utiliser des outils plutôt que de les développer
  - Pour faciliter les réutilisations des données par les autres
  - Pour améliorer la visibilité