CLARIN
Common Language Resources and Technology Infrastructure

Presentation on CLARIN in France (CLARIN-FR)

Patrice Bellot (Université Aix-Marseille)
Nicolas Larrousse (Huma-Num)

patrice.bellot@univ-amu.fr / nicolas.larrousse@huma-num.fr

CLARIN-FR Roadmap

- ***Observer*** in CLARIN
  with the endorsement of the FR Minist. of Research
- French situation in 2016 :
  - A *very large facility* dedicated to Humanities and Social Sciences that leads the participation of France in CLARIN and in DARIAH
    - Huma-Num
  - and a network infrastructure for Language data and tools :
    - Ortolang
- The first aim for France in CLARIN :
  - to gather the French structures already involved (two C centres in 2016) + Ortolang + some labs
  - to propose a road map before becoming a full CLARIN member (economic model, existing scientific communities)

http://www.huma-num.fr

https://www.ortolang.fr

MINISTÈRE DE L'ÉDUCATION NATIONALE, DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE

Speech & Language Data Repository

Website	http://sldr.org/
Consortium	*none*
Type(s)	C
Description	SLDR is a repository for oral and linguistic resources aimed at their long-term preservation and sharing
DSA	*none*
PID status	Handle
Repository system	Self-developed
Strict versioning?	✓

Organisational information

Organisation name	Laboratoire Parole et Langage (LPL)
Institution	Centre National de la Recherche Scientifique (CNRS) and Aix-Marseille University
Working unit	UMR 7309
Shorthand	SLDR
Postal address	5 avenue Pasteur 13100 Aix-en-Provence France
Expertise	oral and linguistic resources

In 2016 : 2 CLARIN C-Centres

ORTOLANG Deposit and sharing

Speech and Language Data Repository (SLDR/ORTOLANG)

Huma-Num CLARIN

Open archives (OAI-PMH)

- ▸ Home
- Browse
- Detailed query
- Submit/edit
- Users' community
- FAQ
- Wiki
- Maintenance
- Contact us
- RSS : RSS / Atom :

Lost in migration ? This document may help you : infoMigration-en_0.2.pdf

New deposits and registration of new users must now be done on Ortolang platform.

Deposit and sharing of oral/multimodal linguistic data

Visible items : 343
Documents : 790785
Downloadings : 19078
Members : 622 (52 countries)
Spoken languages : 175

As of 01/12/2015, registration of new users and deposit of data on SLDR website will be suspended to allow the public opening of Ortolang platform.

Now engaged with the CNRTL and Nanterre Orléans Centre in building ORTOLANG, SLDR continues its work of gathering and sharing language data. All services currently offered will remain part of the new platform.

Thus, SLDR/ORTOLANG allows you to browse data already collected in the area of **speech/multimodal linguistics**. Resources are grouped into four main types of items: primary data, secondary data, tools and collections. **Downloading** is possible on the basis of various provisions of archival law and intellectual property rights.

After authenticating on the site, you can also **deposit** your data and create their descriptive notice via a dedicated interface. This will notably allow you to specify access to data based on its status, nature, or professional categories of users.

Descriptive records created are consistent with the core **metadata standards**: they are eligible for harvesting by international directories and search engines. Meanwhile, items and their contents are assigned **persistent identifiers** to facilitate their retrieval regardless of physical location.

These steps are intended to finally prepare items for their **long-term preservation** by CINES, an institutional archive site.

Do not hesitate to contact us for additional information, or read our guidelines for the sharing and archiving of linguistic resources.

Detailed query

The latest deposits (169) >> more page 1 >>

Primary data (corpus) ortolang-000939
SITAF (tandems anglais/français) (Céline HORGUES)
Individual contribution

Video-recorded corpus of tandem interactions between French-speaking students and English-speaking students at the University Sorbonne Nouvelle- Paris 3. Each of the 21 tandem pairs performs the collaborative speaking tasks (story-telling, debating, reading) on 2 occasions (2 recording sessions) 3 months apart. The corpus also comprises L1-L1 control interactions for the participants.
Metada is [...]

(*applied_linguistics*, *language_acquisition*, *phonetics*)
French; English

hdl:11041/ortolang-000939
2015-09-30
Version 1
source data
(Misc publications)

Primary data (corpus) blri-000940
Projet Noms Verbes (MEG) (ALARIO François-Xavier, LPC, BADIER Jean-Michel, INS, STRIJKERS Kristof, LPC, CHANOINE

hdl:11041/blri-000940
2015-09-30

http://sldr.org/

https://www.ortolang.fr

Collections de corpus oraux numeriques

Website	http://cocoon.huma-num.fr
Consortium	*none*
Type(s)	C
Description	Speech recordings repository
DSA	*none*
PID status	ARK and PURL
Repository system	self-developed
Strict versioning?	✗

Organisational information

Organisation name	CNRS/COCOON
Institution	Centre National de la Recherche Scientifique
Working unit	Laboration ligérien de linguistique (LLL) & Langues et civilisations à tradition orale (LACITO)
Shorthand	COCOON
Postal address	Bibliothèque nationale de France, Quai François Mauriac, 75013 Paris 75013 Paris France


cocoon.huma-num.fr
cnrs
dépasser les frontières
coCOon
COllections de COrpus Oraux Numériques
Présentation
Corpus
Documentation
Corpus
guest
http://cocoon.huma-num.fr/
COllections de COrpus Oraux Numériques
coCOon
CoCoON pour « COllections de COrpus Oraux Numériques » est une plateforme technique qui accompagne les producteurs de ressources orales, à créer, structurer et archiver leurs corpus ; un corpus pouvant se composer d'enregistrements (en général audio) accompagnés éventuellement d'annotations de ces enregistrements.
Les ressources déposées sont dans un premier temps cataloguées et stockées, puis, dans un deuxième temps archivées dans l'archive de la TGIR Huma-Num. L'auteur et son institution restent responsables des documents déposés et peuvent bénéficier d'un accès restreint et sécurisé à leurs données, pendant une période définie, si le contenu de l'information est considéré sensible.
La plateforme COCOON est gérée conjointement par deux unités mixtes de recherche: le Laboratoire de Langues et civilisations à tradition orale (LACITO - UMR7107 - Université Paris3 / INALCO / CNRS) et le Laboratoire Ligérien de Linguistique (LLL - UMR7270 - Universités d'Orléans et de Tours, BnF, CNRS).
Statistiques
Actuellement sont consultables dans l'archive: 9019 enregistrements pour un total de 153 langues différentes.
Déposer
Rechercher
Derniers dépôts...
Khakcalop
Proto-history : Orphans Lahaussois, Aimée (depositor) Lahaussois, Aimée (researcher) Dhana (speaker) Projet HimalCo (ANR-12-corp-0006) (sponsor). Editeur(s): Histoire des Théories Linguistiques. 2011....
Fri, 20 May 2016 00:00:00 +0100
The shelter and the lovers
Tale-telling. Session organised in the language assistant's house. Some members of his family were present. A man is caught by the rain chasing his camel and takes refuge in a shelter which happens to be the meeting place of two lovers. In the dark, two caseq of mistaken identity ensue. Recording c...
Thu, 12 May 2016 00:00:00 +0100
The farmer and the djinn
Tale-telling. Session organised in the language assistant's house. Some members of his family were present. A farmer's son finds cold in a field. After the father has exhausted it, he decides to stop farming and go fishing instead. One day he fishes a copper can containing a djinn who fulfils all h...
Thu, 12 May 2016 00:00:00 +0100
The camel race
Story-telling. Session organised in the language assistant's house. Some members of his family were present. An old man is boasting about his ability to ride a camel during a race, competing with members of another tribe. Recording conditions: Edirol 09 Digital recorder ; Sony MS907 microphone Lang...
Thu, 12 May 2016 00:00:00 +0100
The cat-djinn
Story-telling. Session organised in the language assistant's house. Some members of his family were present. A man asks a shopkeeper to provide goods to his children and to grant him facilities for payment. The shopkeeper never gets his money back as the man turns out to be a cat-djinn.

During the conference

- You can meet some of the French labs / structures that may (wish to) become CLARIN centres :
  - LIMSI (CNRS - Paris Orsay)
  - LINA (Université de Nantes)
  - LPL, LIF-LSIS (Université Aix-Marseille)
  - Ortolang
  - ELDA

lina.univ-nantes.fr

LABORATOIRE d'INFORMATIQUE de NANTES ATLANTIQUE UMR 6241

Rechercher

Présentation | Equipes | Plateformes | Formations | Valorisation | Partenariats | Manifestations | International | Liens

Accueil > Equipes > TALN > Ressources logicielles

TALN

- Actualités
- Les publications (via hal)
- Prix et récompenses
- Projets
- Ressources langagières

RESSOURCES LOGICIELLES

Liste non exhaustive des réalisations logicielles produites par l'équipe TALN

https://www.lina.univ-nantes.fr

Béatrice Daille

Termsuite

TermSuite is a Java UIMA-based toolbox for terminology extraction and multilingual term alignment.
It extracts monolingual terminologies and generates bilingual dictionaries from these terminologies by the means of distributional and compositional methods.
The languages covered are : English, French, German, Spanish, and Russian.
More information here

Apopsis

Un **détecteur d'opinions qui explore les tweets** sur le sujet qui vous intéresse ! Une personne, un produit, un sujet d'actualité. Entrez le sujet et visualisez en temps réel les opinions qui sont émises positives, négatives ou neutres.
Jouer avec un démonstrateur ici

Apache OpenNLP models for processing French

The Apache **OpenNLP** library is a machine learning based toolkit for the processing of natural language text. It supports the most common NLP tasks, such as tokenization, sentence segmentation, part-of-speech tagging, named entity extraction, chunking, parsing, and coreference resolution.
More information here

Code source and Apache UIMA components

Various contributions to the NLP and Apache **UIMA** (Unstructured Information Management Architecture) communities to facilitate the development of NLP components and pipelines, to connect with various data formats, to solve interoperability issues (within UIMA workflow or by integrating third-party tools) and also to perfom some NLP analysis tasks.

Here two source code repositories dev-star et jules-star ; both should be merged soon.
Some components are available under Java Apache Maven dependencies.

UNIVERSITÉ DE NANTES — MINES Nantes — cnrs — Inria

Limsi

https://www.limsi.fr/

Gilles Adda

Speech Recognition

Textual Documents

Sign Language

Speaker Characterization

Language Variation

Paraphrasing

Machine Translation

Modeling Language

Detection of Affects, Sentiments, Opinions

Corpora, Annotation, Evaluation

Machine Learning

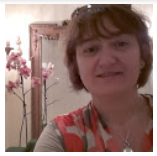
Information Extraction

Question-Answering

Human-Machine Dialogue

	ILES	TLP	total
Researchers / Professors	14	14	28
Engineers, Technicians	5	3	8
PhDs, Post-Docs	21	23	44
total	40	40	80
among whom HDRs	6	10	16

* ILES: Written and Signed Language Information Processing
* TLP: Spoken Language Processing



Valérie Mapelli

The Evaluations and Language resources Distribution Agency

ELDA was created parallel to ELRA, European Language Resources Association, in February 1995.

ELDA is the association's operational body, and is in charge of the development and the execution of ELRA's missions and tasks as defined by the Board of the association.

ELDA is incorporated as a company in order to handle all the commercial and business-oriented tasks of the association.

The CEO of the agency, Khalid Choukri, also acts as ELRA Secretary General and currently works with 12 full-time and permanent members. Some temporary staff may join from time to time to help on some specific projects.

Structure of the company

Links

Tags

association
collection
distribution
identification
language resources
promotion
validation

Latest News

New LRs in ELRA Catalogue Oct. 19, 2016

Tag Cloud

association catalogue computational linguistics conference

ELRA Tweets

ELRA @ELRAnews

Huma-Num
la TGIR des humanités numériques

Infrastructure

VLRI (Very Large Research Infrastructure)?

European level … and further

ERICs (European Research Infrastructure Consortium)

Huma-Num's Ecosystem

SUPPORTING RESEARCH COMMUNITIES

Huma-Num

Huma-Num's Consortia

CAHIER
(Corpus d'Auteurs pour les Humanités : Informatisation, Édition, Recherche)

CONSORTIUM ARCHIVES DES ETHNOLOGUES

CORLI
(Corpus, langues, interactions)

3D-SHS

ARCHIPOLIS
(Archives des sciences sociales du politique)

MASA
(Mémoires des archéologues et des sites archéologiques)

CONSORTIUM MUSICA

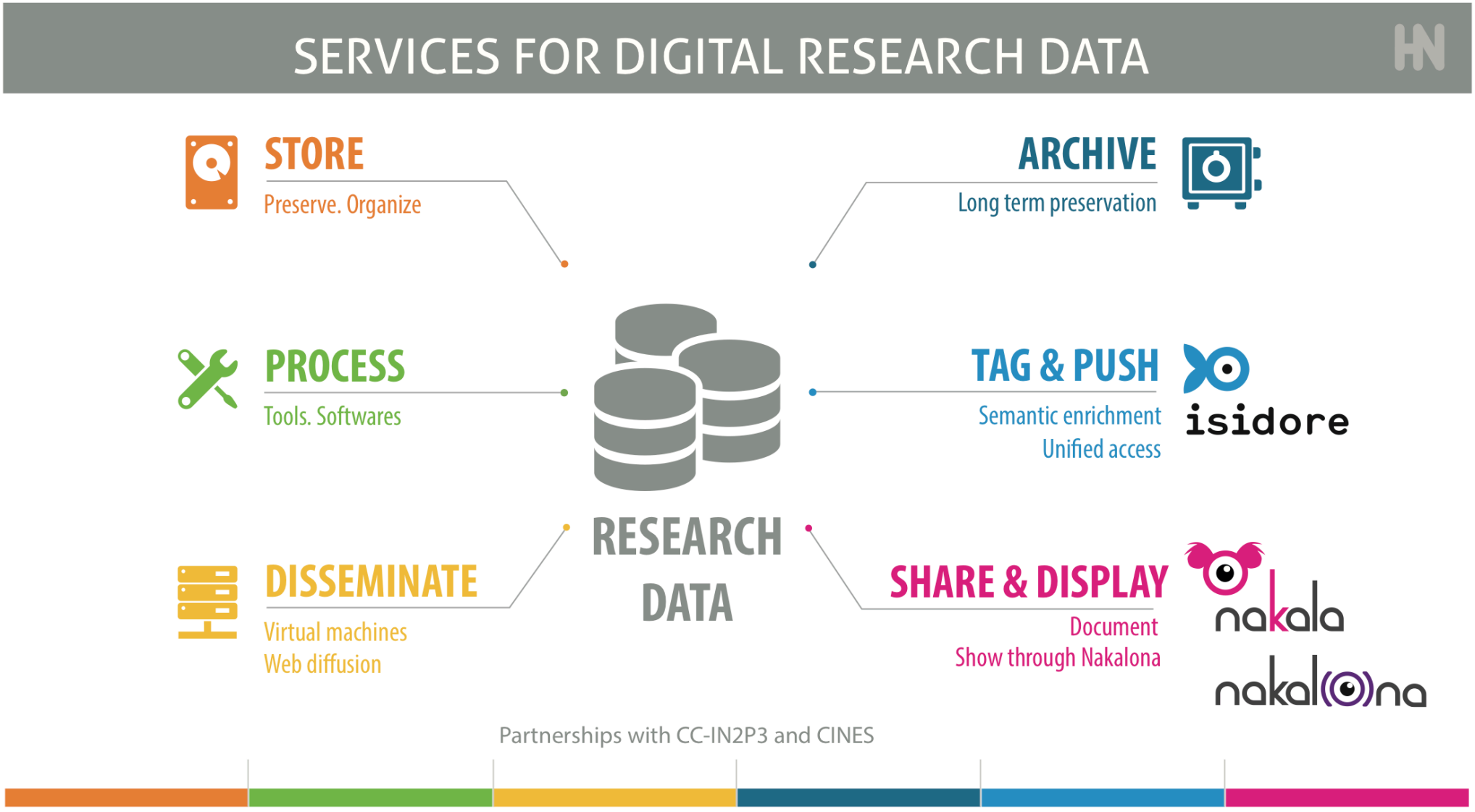
IMAGEO
(Cartes et photographies pour les géographes)

COSME
(Consortium sources médievales)

ARCMC
(Archives des mondes contemporains)

Huma-Num

Huma-Num's Core Services



Huma-Num

In a few words…

- Observer for the next 2 or 3 years….
- … even if we are still waiting for the signature

- Candidatures for new CLARIN Centres to come

patrice.bellot@univ-amu.fr / nicolas.larrousse@huma-num.fr

Huma-Num
la TGIR des humanités numériques

MINISTÈRE
DE L'ÉDUCATION NATIONALE,
DE L'ENSEIGNEMENT SUPÉRIEUR
ET DE LA RECHERCHE

CLARIN
Common Language Resources and Technology Infrastructure