

Critères de qualité pour les données de la recherche en général. Transposition aux corpus en linguistique.

23 SEPTEMBRE 2016,

CONSORTIUM CORPUS, LANGUES ET INTERACTIONS (CORLI)

TABLE RONDE : « CRITÈRES D'ÉVALUATION DES CORPUS »

THIERRY CHANIER

CC BY

De quoi parle-t-on ? Critères de qualité pour quoi ?



Publications

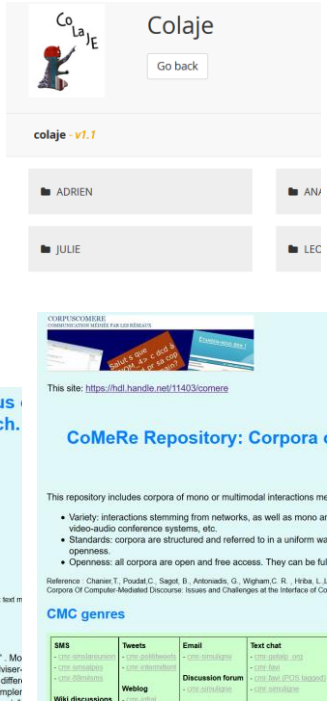


Données

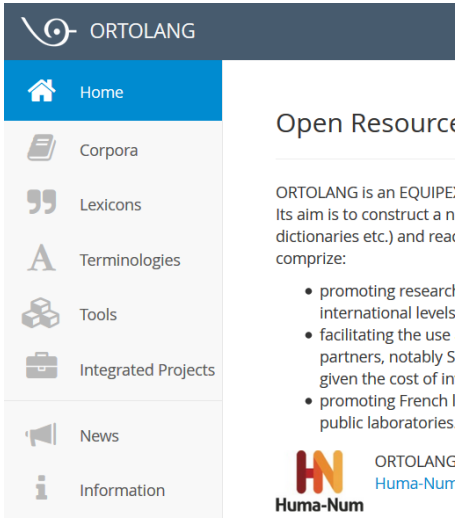
~~corpus~~



Corpus



Banque de Corpus



Infrastructure :
accès, services, préservation

Données de recherche en sciences

Au niveau européen : compétition et confusion

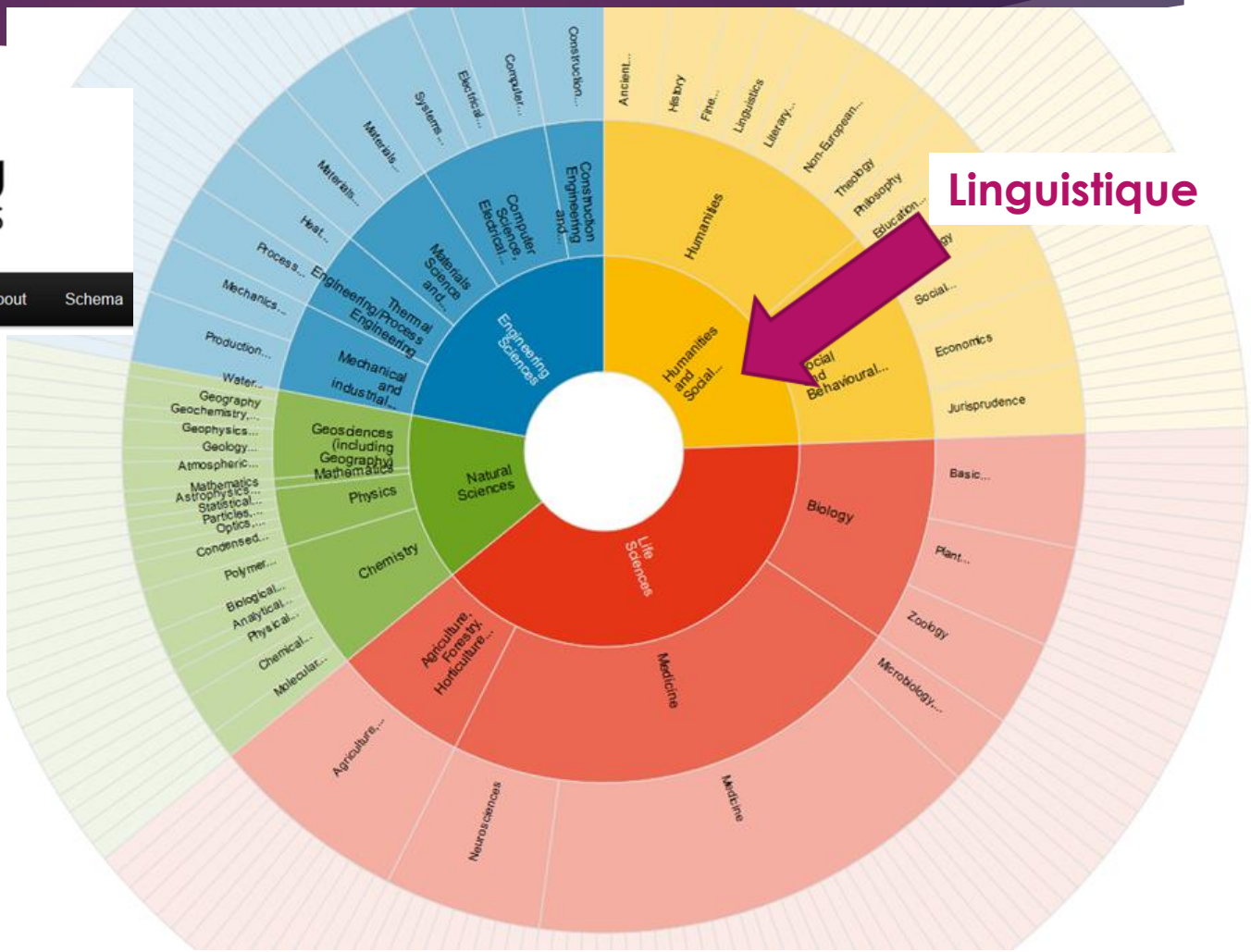
Exemple EUDAT



Recension des banques de données

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

Home Search Browse Suggest FAQ About Schema



Recension des banques de données

Filter

- Subjects ⊞
- Content Types ⊞
- Countries ⊞
- API ⊞
- Data access ⊞
- Database access ⊞
- Database licenses ⊞
- Data licenses ⊞
- Data upload ⊞
- Data upload restrictions ⊞
- Enhanced publication ⊞
- Institution responsibility type ⊞
- Institution type ⊞
- Keywords ⊞
- Metadata standards ⊞
- PID systems ⊞
- Provider types ⊞
- Quality management ⊞
- Repository languages ⊞
- Software ⊞
- Syndications ⊞
- Repository types ⊞
- Versioning ⊞

mulce

Search

Toogle short help

← Previous 1 Next →

Sort by

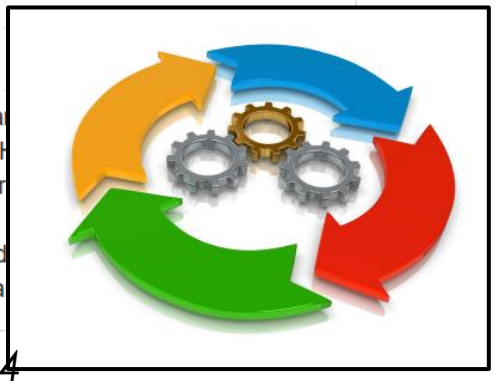
Found 1 result(s)

Multimodal Learning Corpus Exchange

MULCE

Subject(s)	General and Applied Linguistics, Education Sciences, Linguistics, Humanities, Humanities and Social Sciences, Social and Behavioural Sciences
Content type(s)	Standard office documents, Audiovisual data, Scientific and statistical data formats, Raw data, Structured graphics
Country	France, European Union, International

Mulce (MUltimodal contextualized Learner Corpus Exchange) is a resea
National Research Agency (ANR programme: "Corpus and Tools in the H
CORP-006). A teaching corpus (LETEC - Learning and Teaching Corpor
structured data set, particularly of interactional data, and traces left by a
experimentation, conducted partially or completely online and completed
human, pedagogical and scientific information to enable the data to be a

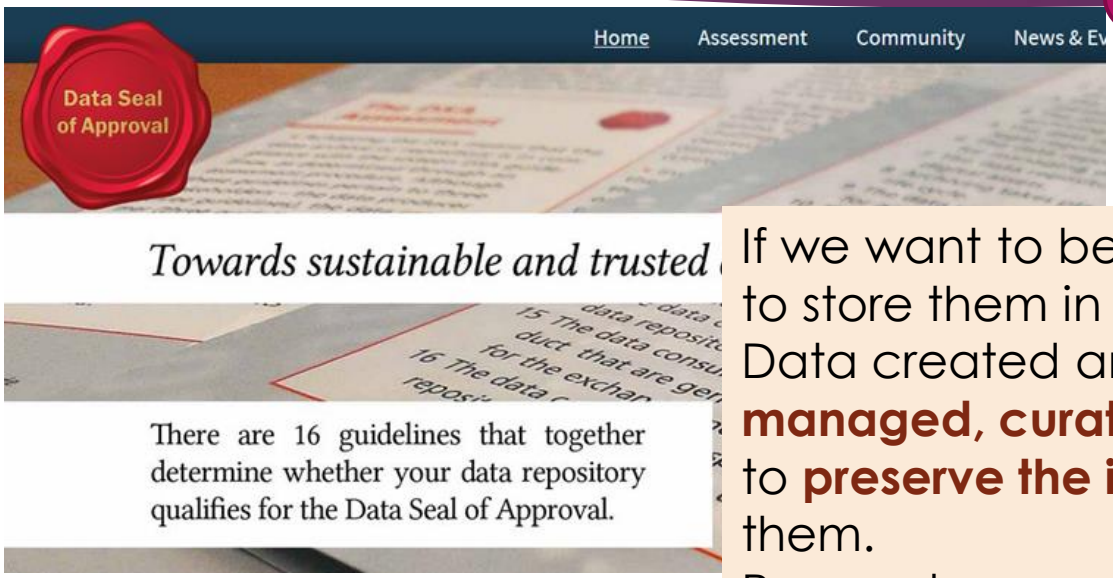


http://repository.mulce.org *depuis 2014*

Cycle de vie des données

Label de qualité pour les serveurs de données, les infrastructures

Linguistique justifier intérêts pour société (cf. projets COST, éval SHS)



If we want to be able to share data, we need to store them in a **trustworthy digital repository**. Data created and used by scientists should be **managed, curated, and archived** in such a way to **preserve the initial investment** in collecting them.

Researchers must be certain that data held in archives **remain useful and meaningful into the future**.

Funding authorities increasingly require **continued access to data** produced by the projects they fund, and have made this an important element in Data Management Plans.

Label de qualité pour les serveurs de données, les infrastructures

- The 16 Data Seal of Approval guidelines are based on the following five criteria:
- The data can be found on the Internet
- The data are accessible (clear rights and licences)
- The data are in a usable format
- The data are reliable
- The data are identified in a unique and persistent way so that they can be referred to

Acquired Seals

Repository	DSA
3TU.Datacentrum	2014-2017
Archaeology Data Service	2014-2017
BABS - Long Term Preservation Library Archiving and Access	2010
Banco de Información para l Sociales (BIIACS)	2014-2017
BAS CLARIN	2014-2017
CINES : Long-term Preservation Platform	2014-2017
CISER Data Archive	2014-2017
CLARIN Center BBAW	2014-2017
CLARIN Center INL	2014-2017
CLARIN Centre Vienna	2014-2017
CLARIN-D Resource Center Leipzig	2014-2017
Clarin-PL Repository	2014-2017
CLARIN.SI Repository	2014-2017

Et ORTOLANG ?

Corpus : distinguons les choses

RARETÉ ?

DONNÉES = CORPUS ?

SITE DE DOCUMENTATION

CORPUS GLOBAL

CORPUS DISTINGUABLE

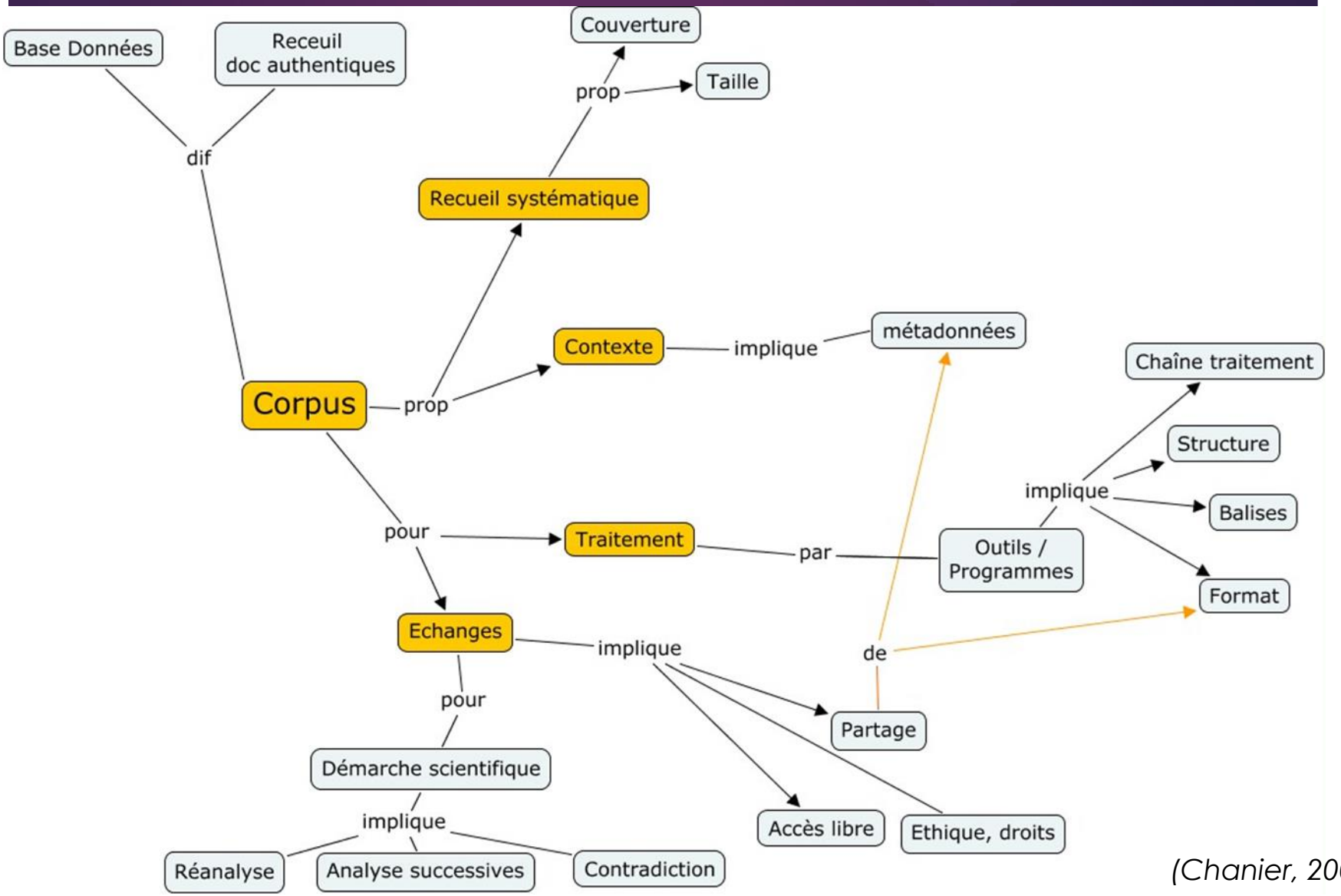
Rareté : pas un critère significatif pour corpus

- Même si archivage et pérennité sont importants (cycle de vie courts sur nos projets / formats électroniques), notre consortium n'a pas une vocation première de bibliothèque
- Le premier patrimoine à conserver (et pas en priorité pour des besoins patrimoniaux, cf. la suite) est celui que nous créons au cours de nos projets récents
  - Les travaux de constitution de corpus nécessitent des investissements importants non renouvelables (rend nos corpus "rares" ?)
- Conserver des données rares (langues en voies d'extinction, situation d'élocution en voie d'extinction) : oui mais **ces données ne constituent pas pour autant des "corpus**".
- Donc rareté, pas un bon critère évaluation corpus
- (Critère de variété à prendre en compte)

Tout ensemble de données ne constitue pas un corpus

- En linguistique, tout le monde parle sans arrêt de corpus dès qu'il collecte / manipule de simples données
  - Exemple : Confusion fréquente entre exemplier et corpus (cf. *datum* vs *exemplum*, Laks)
  - Ensemble de vidéos, de fichiers audio (même avec métadonnées simples) ne constitue pas un corpus
- Paradigme de corpus

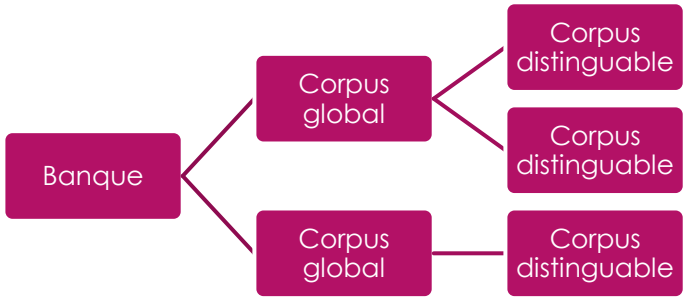
Paradigme de corpus



(Chanier, 2006)

Distinguer les choses

- Pas des corpus :
  - Corpus ≠ ensemble de données
  - Corpus ≠ base de données (jargon MSH, oubli de ce qu'est une BD, modèle conceptuel)
- Ce qui a rapport aux corpus :
  - Site de documentation
  - Banque de corpus (TCF, Colaje, Mulce, CoMeRe, etc.)
  - Corpus global
  - Corpus distinguable



Rapport site de documentation et corpus

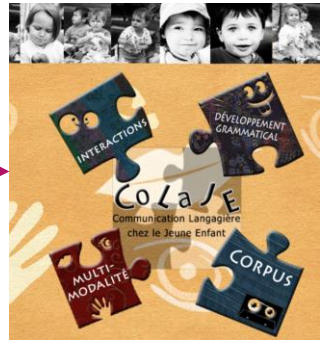
Banque de corpus **Site de documentation**

- Site important :
  - méthodo constitution corpus, accès aux rapports, animer communauté autour du projet, etc.
- Mais cela ne peut remplacer le travail d'explicitation dans le corpus lui-même (cf. plus ci-après)
- Existence du site (avec maintenance) critère d'évaluation d'un projet de recherche, pas d'un corpus













Granularités autour d'un LETEC (corpus d'apprentissage, cf. Mulce)

LETEC : Learning & Teaching corpus, 2007-2013

- Corpus global
  - Données + contexte liées à une formation considérée dans son ensemble
- Corpus distinguable
  - Sous-corpus rassemblant un ensemble de données d'une taille suffisante pour être l'objet d'une recherche spécifique, données dont on veut caractériser la cohésion d'ensemble par rapport au corpus de références (global de base) :



  - associer publication scientifique et données (type 1) ;
  - rassembler des données prêtes à l'analyse avec mise en forme pour outils/logiciels libres (type 2) ;
  - partager des analyses avec des outils associés (type 3).

(Reffay, Chanier, Noras & Betbeder, 2008 ; Chanier & Ciekanski, 2010)

Corpus global : un niveau nécessaire

- Exemple de situation actuelle : ensemble de répertoires et sous-répertoires avec dans chacun un entretien, une séquence d'interaction avec transcription, vidéo, fiche métadonnées
  - Plein de corpus ? Cohérence ? Réutilisabilité ?
- Problème augmente encore avec multimodalité (multiplicité des données et types de données)
- Besoin d'un niveau d'empaquetage / conditionnement (*packaging*) où le chercheur décrit l'ensemble
  - LETEC / éducation : structure XML IMS-CP utilisé dans Mulce
  - TEICorpus ?
  - Cela fait partie travail de constitution du corpus
- **Publication n'apporte pas** la cohérence pour utiliser données (cf. présentation Ortolang) : article court, autre vocation, pas exhaustif
  - **Comprendre relation article et données de recherche**

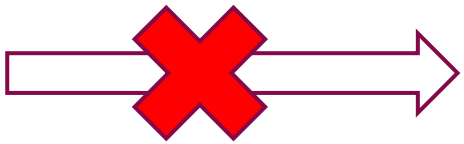
Article vs données , corpus



Publications



Données





(Banque de)Corpus



Article d'analyses





Données associées



Corpus distinguable

Publication et données de recherche

- CRITÈRE ESSENTIEL DE SCIENTIFICITÉ
- NOTRE COMMUNAUTÉ DOIT AVANCER VITE
- CRÉER CORPUS DISTINGUABLES

Publication et réplication

"Replication data sets include the original data and any other information needed to reproduce the numerical results in a published work.

[…] making publicly available a replication data set for each of their empirical articles or books.

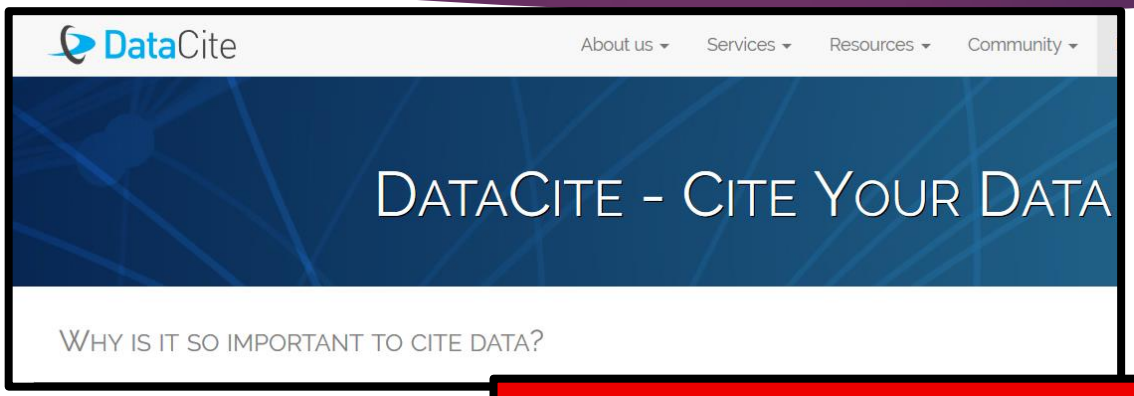
Citation credit should be apportioned both for the original article and separately for the data. "

Gary King (2007). "An Introduction to the Dataverse Network as an Infrastructure for Data Sharing," *Sociological Methods and Research*, Vol. 32, No. 2

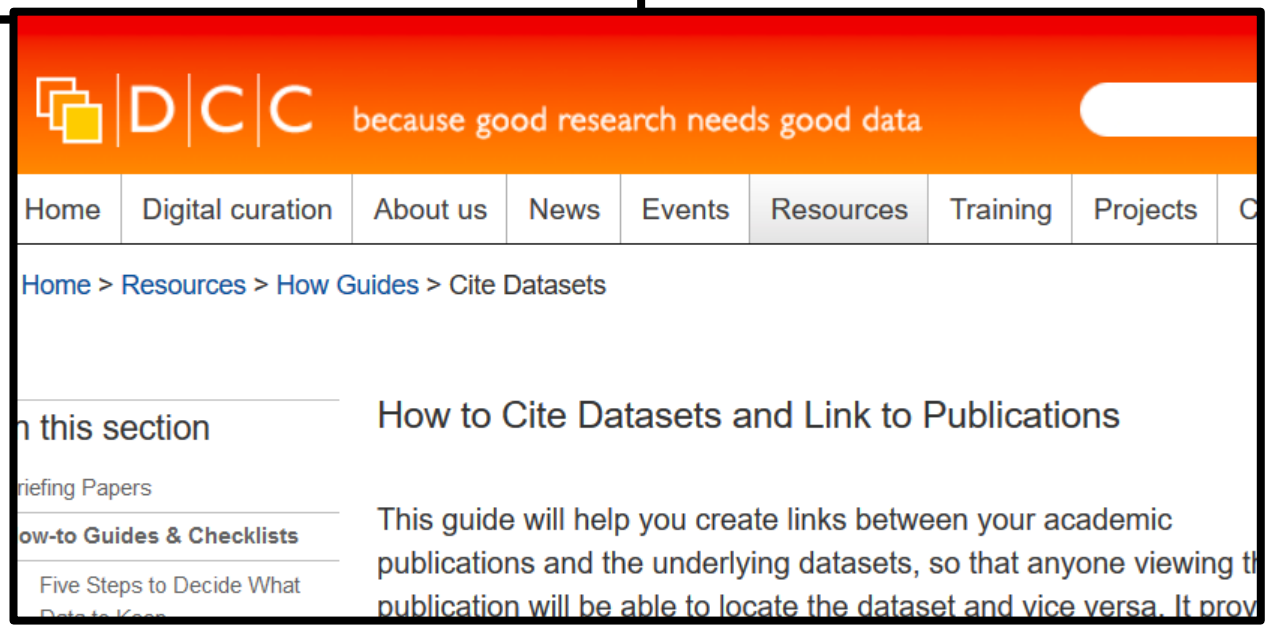
Publication et données de recherche : évaluation / valorisation du chercheur

- Enjeux :
  - scientificité pour notre discipline
  - Évaluation de notre travail (reconnaissance du temps passé dessus)
- Chemin à ouvrir dans nos disciplines : soumission d'article accompagnée d'une soumission des données de recherche associées
  - Comité de relecture prend les deux en compte
  - Publication des deux ensembles : donc double publications avec référencements croisés.
  - Référencement et citations dans les listes de travaux des chercheurs (HCERES) et dans les articles
  - Développer notion corpus distinguable

Citation & références des données : plein développement en Europe



Cf. Horizon 2020

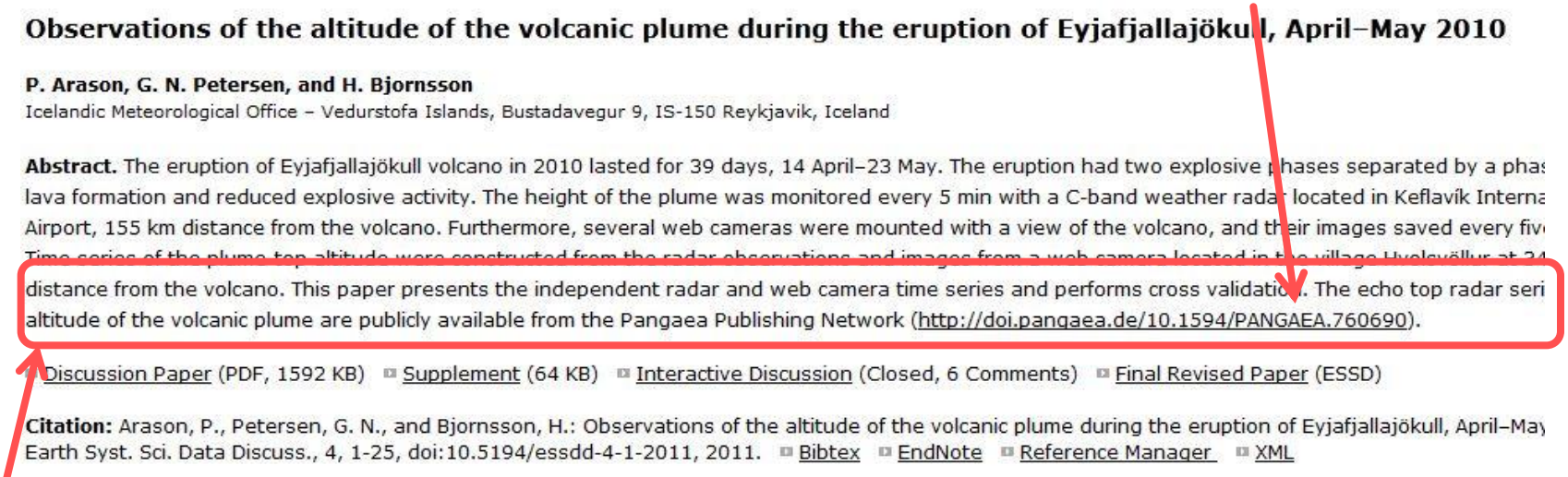


The paper and its reference on the Journal web site

Sciences de la terre

Arason, P., Petersen, G. N., and Bjornsson, H.: Observations of the altitude of the volcanic plume during the eruption of Eyjafjallajl, April?May 2010, Earth Syst. Sci. Data Discuss., 4, 1-25, doi:10.5194/essdd-4-1-2011, 2011.

2 separate permalinks



Observations of the altitude of the volcanic plume during the eruption of Eyjafjallajökull, April–May 2010

P. Arason, G. N. Petersen, and H. Bjornsson
Icelandic Meteorological Office – Vedurstofa Islands, Bustadavegur 9, IS-150 Reykjavik, Iceland

Abstract. The eruption of Eyjafjallajökull volcano in 2010 lasted for 39 days, 14 April–23 May. The eruption had two explosive phases separated by a phas
lava formation and reduced explosive activity. The height of the plume was monitored every 5 min with a C-band weather radar located in Keflavík Interna
Airport, 155 km distance from the volcano. Furthermore, several web cameras were mounted with a view of the volcano, and their images saved every fiv
Time series of the plume top altitude were constructed from the radar observations and images from a web camera located in the village Hvolsvöllur at 24
distance from the volcano. This paper presents the independent radar and web camera time series and performs cross validation. The echo top radar seri
altitude of the volcanic plume are publicly available from the Pangaea Publishing Network (http://doi.pangaea.de/10.1594/PANGAEA.760690).

Discussion Paper (PDF, 1592 KB) Supplement (64 KB) Interactive Discussion (Closed, 6 Comments) Final Revised Paper (ESSD)

Citation: Arason, P., Petersen, G. N., and Bjornsson, H.: Observations of the altitude of the volcanic plume during the eruption of Eyjafjallajökull, April–May
Earth Syst. Sci. Data Discuss., 4, 1-25, doi:10.5194/essdd-4-1-2011, 2011. Bibtex EndNote Reference Manager XML

Link to the data

Data on the data site (Pangaea)

PANGAEA®
Data Publisher for Earth & Environmental Science

Data Description

Citation: Arason, P et al. (2011): Plume-top altitude time-series during 2010 volcanic eruption of Eyjafjallajökull. *Icelandic Meteorological Office, Reykjavik*, doi:10.1594/PANGAEA.760690, *Supplement to:* **Arason, Pordur; Petersen, G N; Bjornsson, H (2011):** Observations of the altitude of the volcanic plume during the eruption of Eyjafjallajökull, April-May 2010. *Earth System Science Data*, **3**, 9-17, doi:10.5194/essd-3-9-2011

Data reference

Paper reference



Abstract: The eruption of Eyjafjallajökull volcano in 2010 lasted for 39 days, 14 April-23 May. The eruption had two explosive phases separated by a phase with lava formation and reduced explosive activity. The height of the plume was monitored every 5 min with a C-band weather radar located in Keflavík International Airport, 155 km distance from the volcano. Furthermore, several web cameras were mounted with a view of the volcano, and their images saved every five seconds. Time series of the plume-top altitude were constructed from the radar observations and images from a web camera located in the village Hvolsvöllur at 34 km distance from the volcano. This paper presents the independent radar and web camera time series and performs cross validation.

Coverage: *Latitude:* 63.630000 * *Longitude:* -19.620000

Date/Time Start: 2010-04-14T00:00:00 * *Date/Time End:* 2010-05-24T00:00:00

Event(s): **Eyjafjallajökull** * *Latitude:* 63.630000 * *Longitude:* -19.620000 * *Date/Time Start:* 2010-04-14T00:00:00 * *Date/Time End:* 2010-05-24T00:00:00 * *Location:* Iceland * *Device:* Observation

Comment: The plume-top altitude during the volcanic eruption of Eyjafjallajökull in Iceland, April-May 2010, was measured by a weather radar and web camera. This h many researchers as input into models as constraints on volume erupted and distribution of ash.

License: Creative Commons Attribution 3.0 Unported

Size: 7 datasets

Download Data

Download **ZIP** file containing all datasets as tab-delimited text *(use the following character encoding:* ISO-8859-1: ISO Western (PANGAEA default) *)*

Datasets listed in this Collection

Access to data

1. **Arason, P; Petersen, GN; Bjornsson, H (2011):** Altitude estimates of the volcanic plume-top from web camera photographs. doi:10.1594/PANGAEA.760684
2. **Arason, P; Petersen, GN; Bjornsson, H (2011):** Altitude estimates of the volcanic plume-top measured by C-band weather radar. doi:10.1594/PANGAEA.760683

Premiers pas dans ex TGE-Adonis en 2012

http://datapublication.org

Data publication - publish articles and research data -

Données de l'enquête sur les usages des TIC par les lycéens.

Information

Year: 2012

Author(s): Nicolas Guichon

Provide by: Nicolas Guichon

Keywords: usages numériques, fracture numérique, compétences, Technologies Technologies (ICT), ICT usage, digital divide, competences

Description

Un questionnaire concernant l'origine (profession des parents, lycée, équipen administré au printemps 2011 à 1300 élèves de première en France métropoli lycéens de l'enseignement secondaire général en France. Elle a concerné 10 l périphérie urbaine ou centre d'une grande ville.

Restrictions

Les données issues de ce questionnaire sont destinées à la communauté scientifi faire l'objet d'une demande particulière à l'auteur.
L'utilisation commerciale n'est pas autorisée.

Data file

To download the files click on their name. The downloadable files are the forma request must be do from the author.

Questionnaire_FINAL-Guichon.pdf application/pdf - 25.4 ko	Oth
Enquête Guichon2011.xls application/vnd.ms-excel - 1.23 Mo	Oth
Codage1.csv application/vnd.ms-excel - 3.74 ko	Da
Codage2.csv application/vnd.ms-excel - 6.07 ko	Oth
Etablissement.csv application/vnd.ms-excel - 510 o	Data

sticef .org Sciences et Technologies de l'Information et de la Communication pour l'Éducation et la Formation

version pleine page
version à télécharger (pdf)

Volume 19, 2012
Article de recherche

Sommaire

Rechercher
- auteur
- année
- titre
- résumé
- abstract
- rubrique

Contact : infos@sticef.org

Les usages des TIC par les lycéens - déconnexion entre usages personnels et usages scolaires

Nicolas GUICHON (ICAR, Lyon 2)

RÉSUMÉ : Cette recherche, qui adopte la perspective de la sociologie des usages, s'appuie sur une enquête à la fois par questionnaire et par entretiens pour sonder les usages numériques des lycéens de l'enseignement secondaire général en France. Deux objectifs sont visés : d'une part, grâce aux données empiriques obtenues, un état des lieux des usages numériques des jeunes est conduit. D'autre part, cette étude investigue de quelles façons les outils numériques sont utilisés pour le travail à la maison et pour l'apprentissage des langues étrangères. Les résultats mettent au jour une déconnexion entre usages des Technologies de

Fiche du corpus sur Corpus distinguable : repository.mulce.org



- Briefly
- Context where interactions happened
- Analysis on interactions
- Resources linked to analysis
- Software used for analysis
- References

Social Network Analysis, online group c

[This corpus is based on data extracted from the global Learning & Teaching Corpus Si
here referred to (mainly Reffay, C. & Chanier, T. (2003a)) have been built. It has been
Management System or CSCL environment, such cues are ab
on data extracted from a 10-week distance-learning experimer
that such processing, embodied in monitoring tools, can display global properties both a
and detailed textual analysis of messages and the statistical distribution of participants' co

online collaborative learning ; Social Network Analysis ; group communication structure ;

Education -- Data processing, Computer-assisted instruction, Language and languages --

Context where interactions happened

Description of Simuligne course

Simuligne is a distance French as a foreign language learning session and was born in a tra
natives –French teacher trainees from the Université de Franche-Comté–, 4 tutors –teachers
experiment: learners, experts –natives– and teachers –tutors and the coordinator. All agents we
to adapt the method called "Simulation globale" for the first time to a distance situation. The gl
no one had ever met before Simuligne, except the natives from Besançon. Two other importan
acquisition of technical skills, 2. (3 weeks): designing the place of the simulation, –city, campu
weeks): discussing and voting for their favoured project. Three groups out of four achieved th
produced by the three remaining groups presented rich and high quality language productions (

données

données

Resources linked to analysis

Resource "resource-mce-simu-sna-01-00"

Documentation on the corpus which can be read in a web browser

- content/documentation/bando_mulce_seul.jpg :
- content/documentation/letec_disting_affiche.css :
- content/documentation/letec_disting_affiche_en.xsl :
- content/documentation/index.xml : Double click on this file in order to display information on this

Resource "resource-mce-simu-sna-01-01"

Slide presentation in PDF format at conference CSCL 2003, see also bibliographicCitation (Reffay, C. &

- content/simu-reffay_chanier-CSCL2003.pdf :

Resource "resource-mce-simu-sna-01-02"

Data and graphical representation corresponding to section 3.1 of (Reffay & Chanier, 2003a) and (Reff

- content/mce-simu-sna-01-01/simu-sna-matrices-1.xls : Matrices corrsponding to forums and ema
- content/mce-simu-sna-01-01/simu-Gallia-email-1.jpg : Image drawn with Graphviz of emails in Ga
- content/mce-simu-sna-01-01/simu-Gallia-email-2.jpg : Image drawn with Graphviz of emails in Ga
- content/mce-simu-sna-01-01/C_Galia.txt : input format of GraphViz for emails in Gallia group
- content/mce-simu-sna-01-01/F_Galia.txt : input format of GraphViz for forums in Gallia group
- content/mce-simu-sna-01-01/simu-Gallia-email-3.jpg : Output of GraphViz of file C_Galia.txt
- content/mce-simu-sna-01-01/simu-Gallia-forum3.jpg : Output of GraphViz of file F_Galia.txt
- See also resource "resource-mce-simu-sna-01-01" :

Resource "resource-mce-simu-sna-01-03"

Data and graphical representation corresponding to section 3.2 of (Reffay & Chanier, 2003a)

- content/mce-simu-sna-01-02/simu-sna-matrices-2.xls : Matrices (simple, symmetric and threshol
- content/mce-simu-sna-01-02/simu-all-forum-cliques.jpg : Image of cliques on the forums of the 4
- content/mce-simu-sna-01-02/simu-forum-clusters.jpg : Image of clusters computed with USENET
- See also resource "resource-mce-simu-sna-01-01" :

Software used for analysis

Computer Programs for Social Network Analysis : Graphviz - Graph Visualization Software
http://www.graphviz.org/

Computer Programs for Social Network Analysis : UCINET
http://www.insna.org/software/ http://www.analytictech.com/ucinet6/ucinet.htm

Références dans la banque de corpus CoMeRe

- Riou, S. & Sagot, B. (2016). *Etiquetage morpho-syntaxique du corpus FAVI [corpus]. D'après Yun, H. & Chanier, T. (2014). Corpus d'apprentissage FAVI (Français académique virtuel international) [cmr-favi-tei-v1]*. Banque de corpus CoMeRe. Ortolang.fr : Nancy. http://hdl.handle.net/11403/comere/cmr-favi/cmr-favi-tei-v2

  7 780 textchat turns ; 77 605 tokens ; 31 participants ; 29 POS tags.

- Panckhurst R., Détrie C., Lopez C., Moïse C., Roche M., Verine B. (2016). *88milSMS. A corpus of authentic text messages in French*. In Chanier T. (ed) Banque de corpus CoMeRe. Ortolang : Nancy. https://hdl.handle.net/11403/comere/cmr-88milsms

  nbparticipants=422 ; nbmessages=88 522; nbemoticons-emojis=29 563

- Yun, H. & Chanier, T. (2014). *Corpus d'apprentissage FAVI (Français académique virtuel international)*. Banque de corpus CoMeRe. Ortolang.fr : Nancy. https://hdl.handle.net/11403/comere/cmr-favi

  7 780 textchat turns ; 77 605 tokens ; 31 participants.

https://hdl.handle.net/11403/comere

- Objectifs dans CORLI et ORTOLANG
  - Adopté et encourager système de références corpus (type APA)
  - Développer notion de corpus distinguable comme unité correspondante
  - Relier publication et corpus distinguable

Critères de qualité pour les corpus

POUR QUOI FAIRE ?

POUR QUI ?

LESQUELS ?

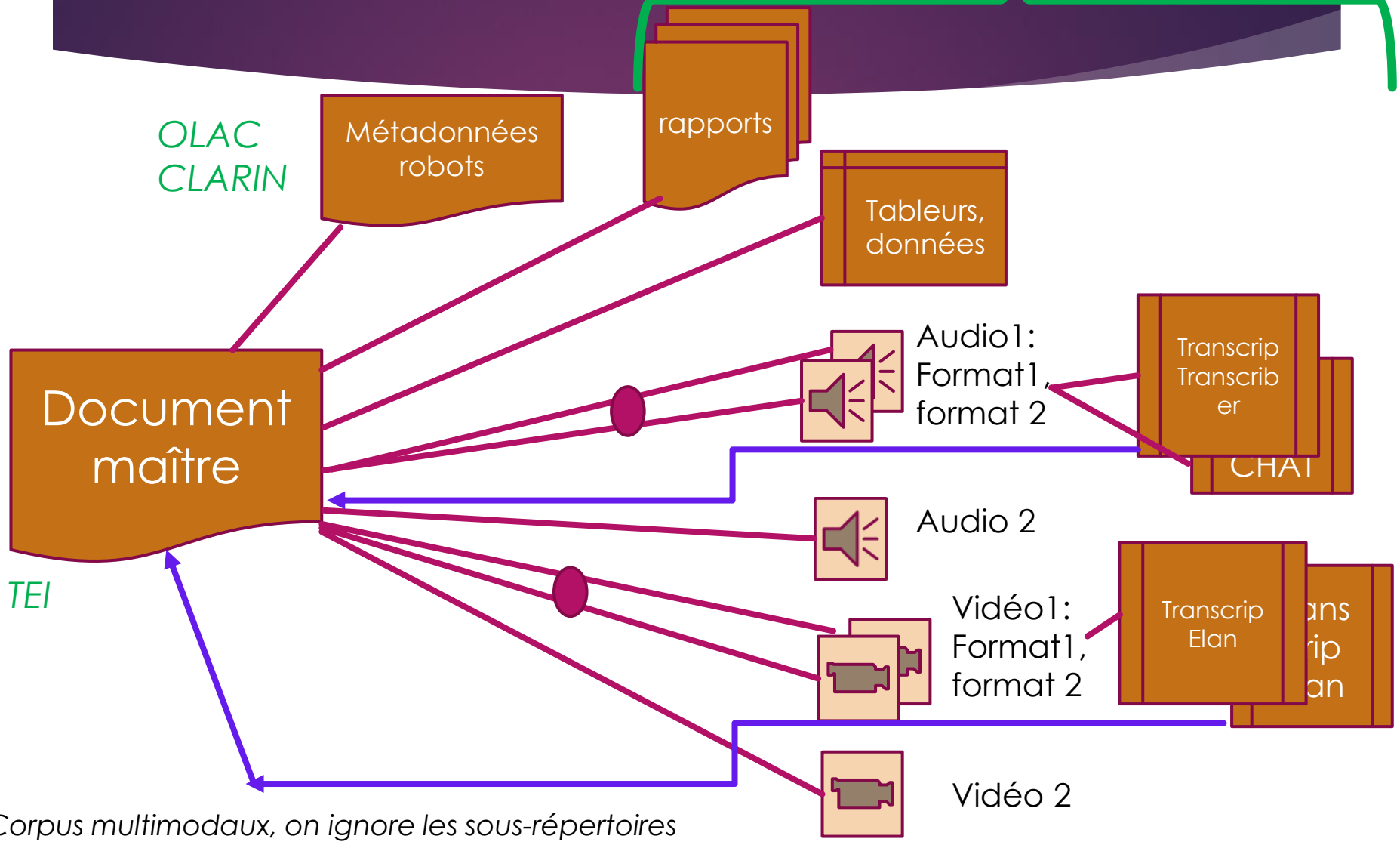
Qualité, pour quoi ? Pour qui ?

- En priorité : Réutilisabilité (réplicabilité)
  - Donc critères à déterminer en fonction de l'usage à venir ("les autres vont-ils pouvoir se servir de ce que j'ai fait ?")
  - Comment ?
    - Consignes / conseils de CORLI / ORTOLANG
    - Attitude pro-active parmi les chercheurs (cf. débat dans table ronde)
- Puis : Evaluation / valorisation travail du chercheur, de son unité.

Réutilisabilité pour qui ?

- Les autres chercheurs : accès direct, compréhension aisée, possibilité de reproduire les analyses, mélanger les données du corpus avec les siennes, compléter / étendre (analyses, données)
- Les acteurs sociaux : individus, institutions, entreprises (ex: compréhension interactions réseaux sociaux, développer nouveaux programmes / services ; éducation : comprendre évolution littératie)
- Les directions de recherche (nationales, européennes) : financement pas perdus, investissement qui sert à d'autres recherches ; utilité sociale de notre secteur SHS

Vue simplifiée d'un corpus



Corpus multimodaux, on ignore les sous-répertoires

Quelques critères de qualité

- Données recueillies :
  - Couverture significative / objet d'étude
  - Contexte décrit et explicité pour personnes hors projet
  - Formats non propriétaires
  - Formats pour pérennisation (consignes CINES)
  - Analyses / transcriptions présentes réalisées avec programmes libres
  - Toutes données reliées au document maître (identifiées aussi dans doc maître)
- Licence posée sur tout le corpus garantissant accès libre et réutilisation directe sans entrave (*Opendata*) : CC-BY et pas CC-BY-NC, par exemple

Quelques critères de qualité

- Fiche métadonnées (simple, standard, type OLAC) :
  - Complète
  - Notamment rôles acteurs (cf. OLAC : dépositeur, collecteur, éditeur, etc.)
  - Référence claire et standard du corpus pour citation (BibliographicCitation)
- Et le principal, le document maître

Document maître : critères pour réutilisation

Sharing oral data involves the diffusion of available and reusable corpora for different research purposes. It requires researchers :

- To share recordings and transcripts in open formats,
- to define a common subset of relevant metadata to explain the context in which the data has been collected or provide information about the speakers,
- to deal with juridical restrictions concerning ethics and data reuse,
- to describe and give examples of annotations
- and more generally to keep in mind that all the components of an oral corpus need to be stored, described and converted into ready-to-use open formats.

(Liégeois, Etienne, Benzitoun & Parisse, à paraître)

Document maître

- Il informe sur tout, contient le corps du corpus, rassemble les informations. L'utilisateur ne doit pas avoir à chercher des infos hors corpus pour pouvoir l'exploiter.
- Écrit dans un format standard, structuré avec validation (TEI)
- Composé de 3 parties ppales :
  - p1) Rapport de recherche
  - p2) Espace verbal (le "texte" proprement dit du corpus)
  - p3) Ressources : recension de tous les fichiers/ documents composant le corpus avec lien et métadonnées (format, longueur, URN, etc.)
- Respect des standards et normes : dates, codage langues, éducation, lieux, etc.
  - on ne réinvente pas la roue et cela doit être compréhensible / traitable au niveau international)
- éthique : en partie1, explication contrat (détail contrat consentement signé dans les docs annexes), codage anonymisation ; en partie2 : anonymisation qui conserve info sémantique

Document maître : partie 1, rapport de recherche

- Ne plus appeler cette partie (en TEI, il s'agit du *teiHeader*) ~~"métadonnées"~~ par risque de confusion avec les métadonnées au sens de la documentation bibliophile (cf. fichier méta *OLAC*)
- Le chercheur rédige sur plusieurs pages, suivant une terminologie standard (ex: modèle du *teiHeader*) un rapport complet de recherche sur la constitution, l'édition, les systèmes de codage, l'organisation du corpus.
  - Voir exemples dans *(Liégeois, Etienne, Benzitoun & Parisse, à paraître)*
- Les informations sur les auteurs d'actes de langage (*participants*) avec leur biographie sociolinguistique. Chaque participant est identifié de façon unique dans tous les corpus de la série, voir dans toute la banque de corpus (comparaisons, calculs)

Document maître : partie 2, Espace verbal

- Le "texte" proprement dit du corpus
- Toute la partie textuelle du verbal et non verbal contenus dans les ressources (texte des tapuscrits, transcriptions, analyses, annotation),
  - Assemblée de façon structurée, cohérente et uniforme (plusieurs TEI ? Ou <text> , divisions, ? , etc.)
- Cet assemblage oblige à créer un espace scripturale ou dialogique cohérent
  - Exemple : Alignement temporel absolu de toutes les transcriptions de tous les sujets, organisation des lieux d'énoncés (sessions, entretiens, etc.), identification du locuteur / scripteur unique

Document maître : partie 3, ressources

- Ressources : recension de tous les fichiers/ documents composant le corpus avec lien et métadonnées (format, longueur, URN, etc.)
- Placée en TEI dans le *teiHeader*, mais mauvaise conception dans le modèle TEI
- Modèle plus intéressant en IMS-CP
  - voir exemples dans la banque de corpus Mulce (http://repository.mulce.org, voir *manifest.xml*, documents maitre des corpus globaux).

Nouvelles perspectives de réutilisation

- Lorsque ces critères seront respectés (et c'est déjà le cas pour plusieurs banques de corpus Ortolang)
- Alors , du fait des collaborations entres les acteurs (chercheurs, consortium Corli, Ortolang), la réutilisation deviendra possible
- Mais il faudra aussi être pro-actif (cf. discussion)
- Exemple de nouveaux services aux chercheurs

Exemple Colaje

C'est déjà demain ☺

Corpus disponibles

♦ COL

Adrien



Interrogatio

Accès à d'in

Visualiser la transcription

Visualisation texte : ADRIEN-01-1_03_09.tei_corpo.xml

Téléchargement des données

Transcriptions

Transcription au format TEI/Ortolang : ADRIEN-01-1_03_09.tei_corpo.

Transcription au format CLAN : ADRIEN-01-1_03_09.cha (24 ko)

Enregistrements

ADRIEN-01-1_03_09-480p.webm (234 Mo)
ADRIEN-01-1_03_09-240p.mp4 (serveur www.ortolang.fr)
ADRIEN-01-1_03_09-720p.mp4 (serveur www.ortolang.fr)
ADRIEN-01-1_03_09-original.mov (serveur www.ortolang.fr)
ADRIEN-01-1_03_09-240p-anonym.mp4 (serveur www.ortolang.fr)
ADRIEN-01-1_03_09-480p-anonym.mp4 (serveur www.ortolang.fr)
ADRIEN-01-1_03_09-720p-anonym.mp4 (serveur www.ortolang.fr)

ssion régulière) :

s :

☑ Colaje
☑ TCOF_adulte
☑ TCOF_enfant
☑ ESLO-1
☑ ESLO-2

Expression régulière ☐

Les mots à rechercher peuvent-ils être séparées par un ou
☑ Oui
Si oui, par combien de mots peuvent-ils être séparés au m
3

Recherche dans :
☑ Lignes principales
Nom du locuteur recherché (Si non-précisé, recherche da
☐ Annotations (tiers)
Nom de l'annotation recherchée (Si non-précisé, recherch