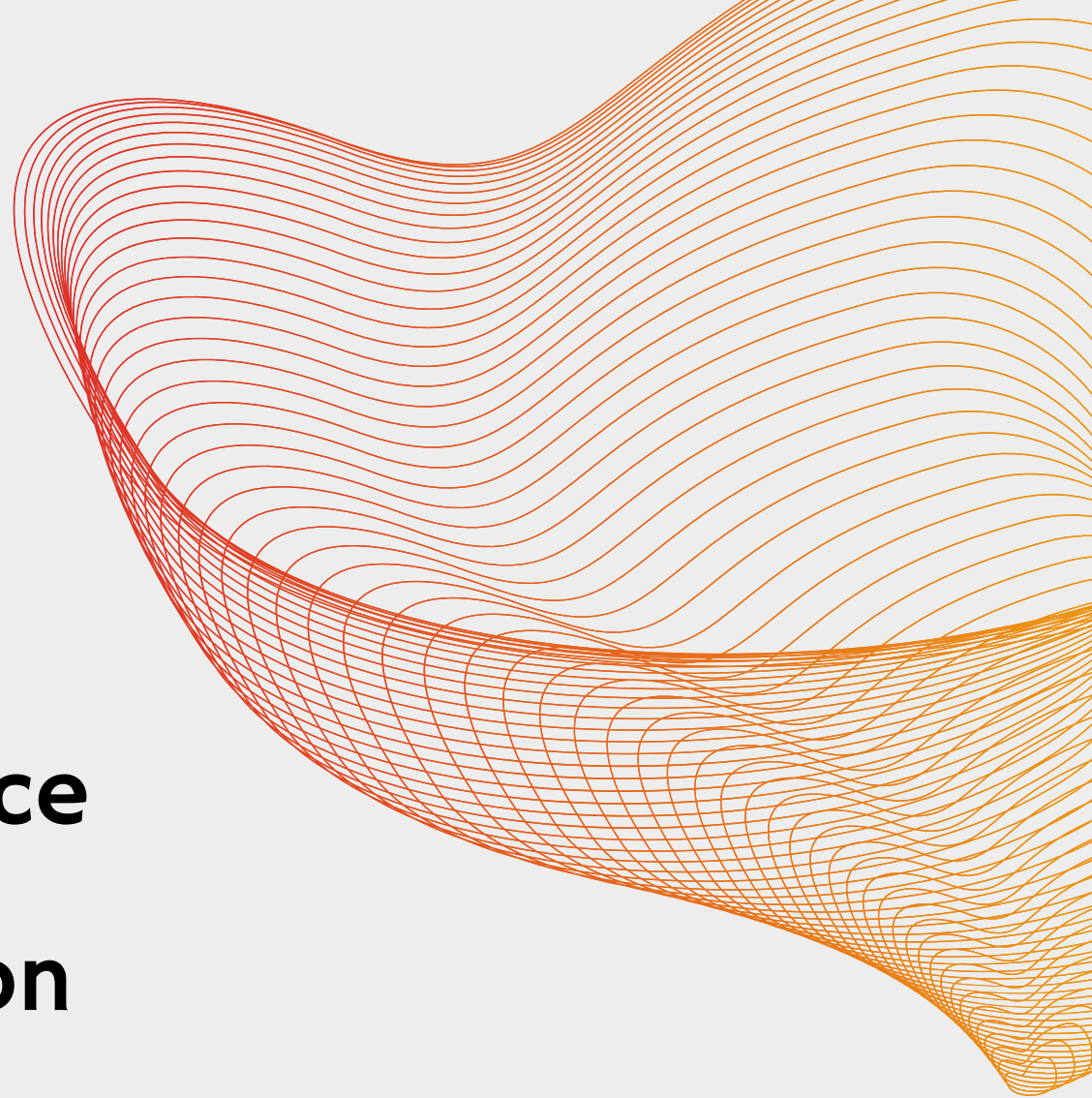




Université
de Poitiers



Retours d'expérience sur la rédaction d'un plan de gestion de données : le projet ANR DIVITAL

Marianne Vergez-Couret, FoReLLIS
5 décembre 2022, Université de Poitiers

Le projet ANR DIVITAL :

**Accroître la vitalité et la
visibilité numérique des
langues de France :
descriptions linguistiques et
corpus annotés**





Partenaires et langues régionales

Alsacien : Université de Strasbourg,

UR Linguistique, Langues et Parole

Corse : Université de Corse Pascal Paoli,

UMR Lieux, Identités, Espaces, Activités

Occitan : Université Toulouse Jean Jaurès,

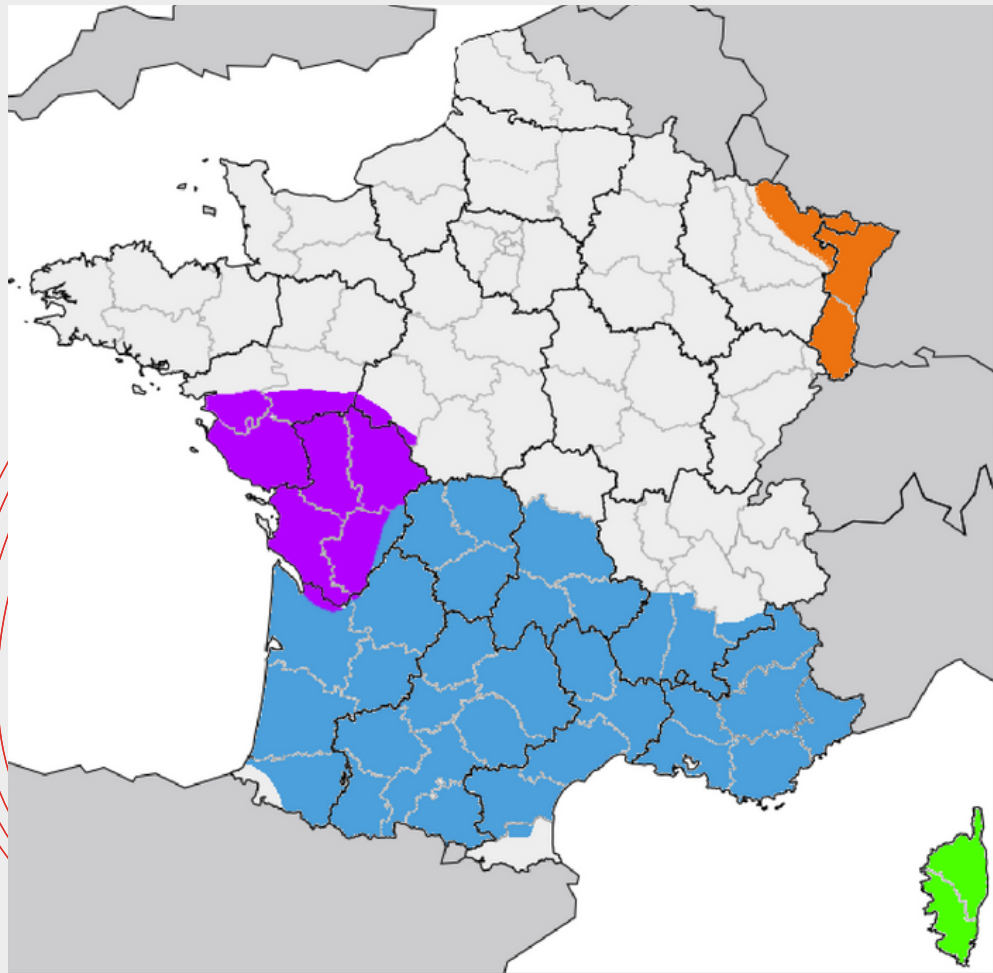
UMR Cognition, Langues, Langage, Ergonomie

Poitevin-Saintongeais : Université de Poitiers,

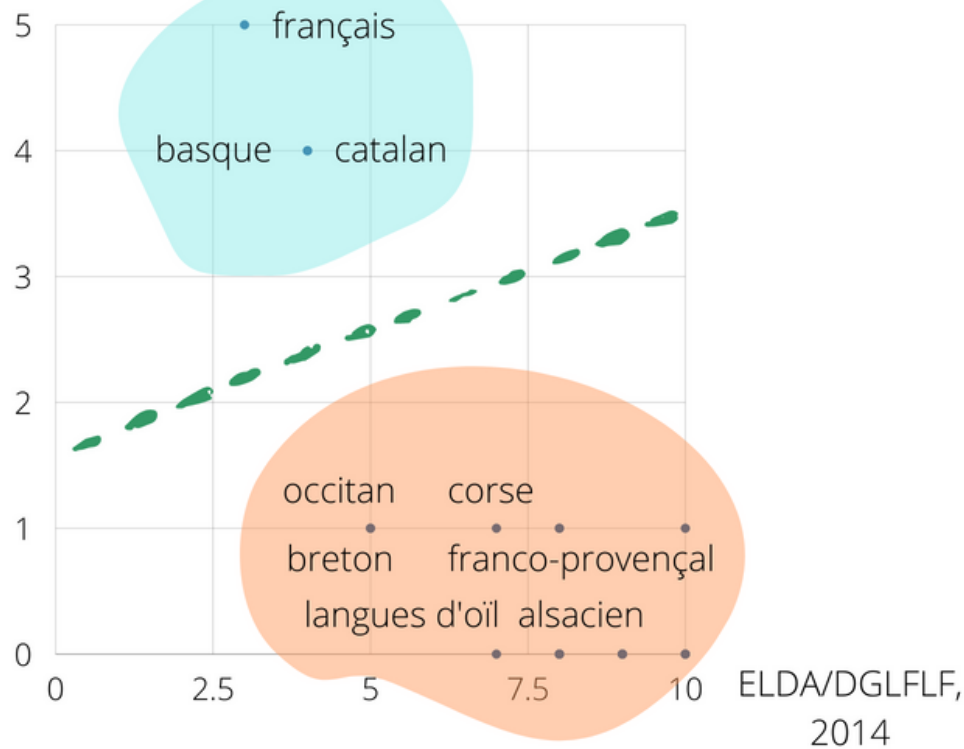
UR Formes et Représentations en Linguistique,

Littérature et dans les arts de l'Image et de la

Scène



Joshi et al.,
2020



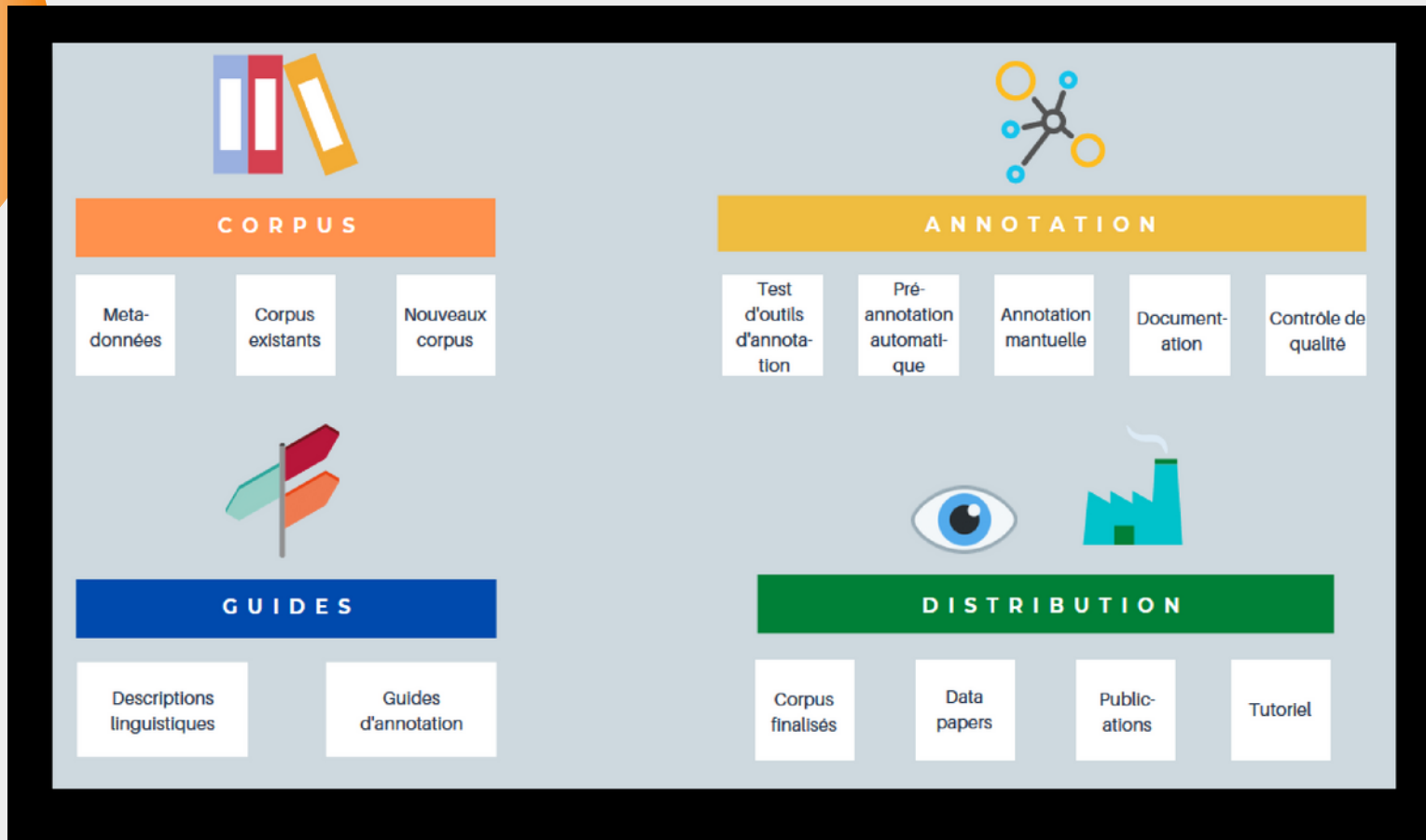
Ressources linguistiques pour les langues de France

"Inventaire des ressources linguistiques des langues de France".

Leixa, Mapelli et Choukri, ELDA/DGLFLF – 2014

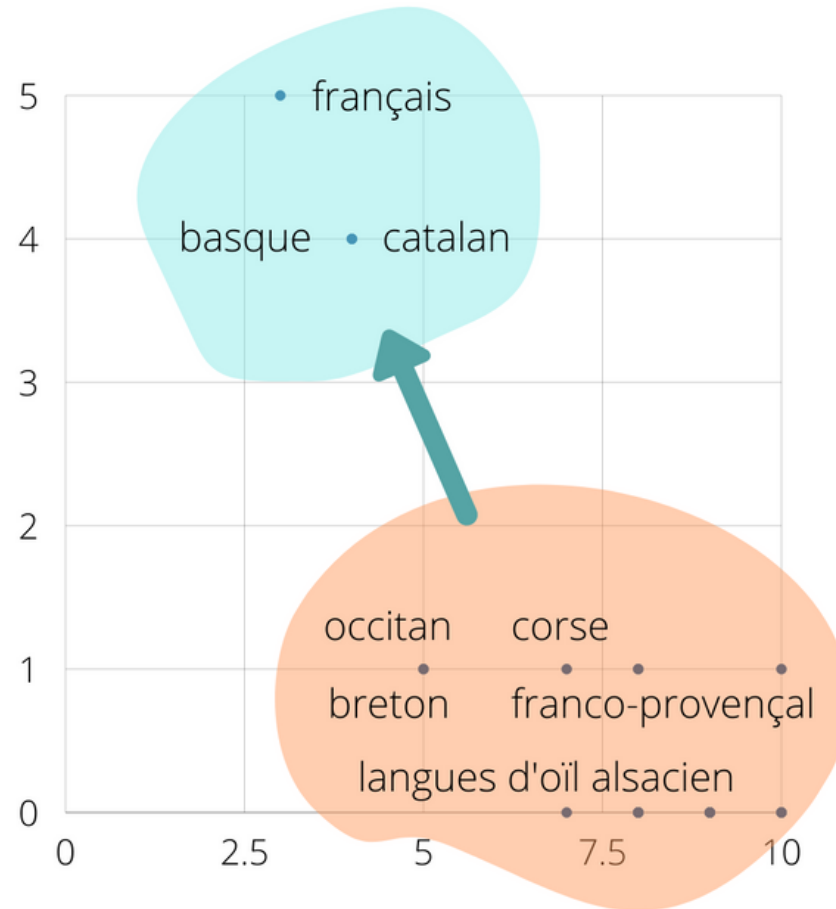
"The State and Fate of Linguistic Diversity and Inclusion in the NLP World". Joshi, Santy, Budhiraja, Bali & Choudhury, ACL 2020





4 tâches autour de la constitution, l'annotation, la documentation et la diffusion de ressources

Organisation du projet

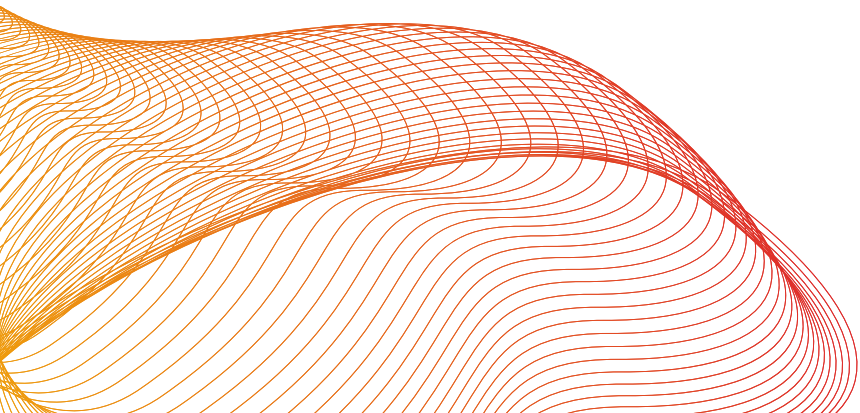


Objectifs





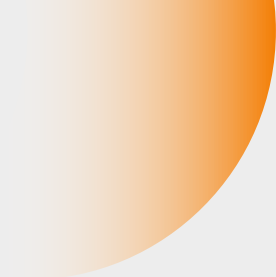
Plan de Gestion des données





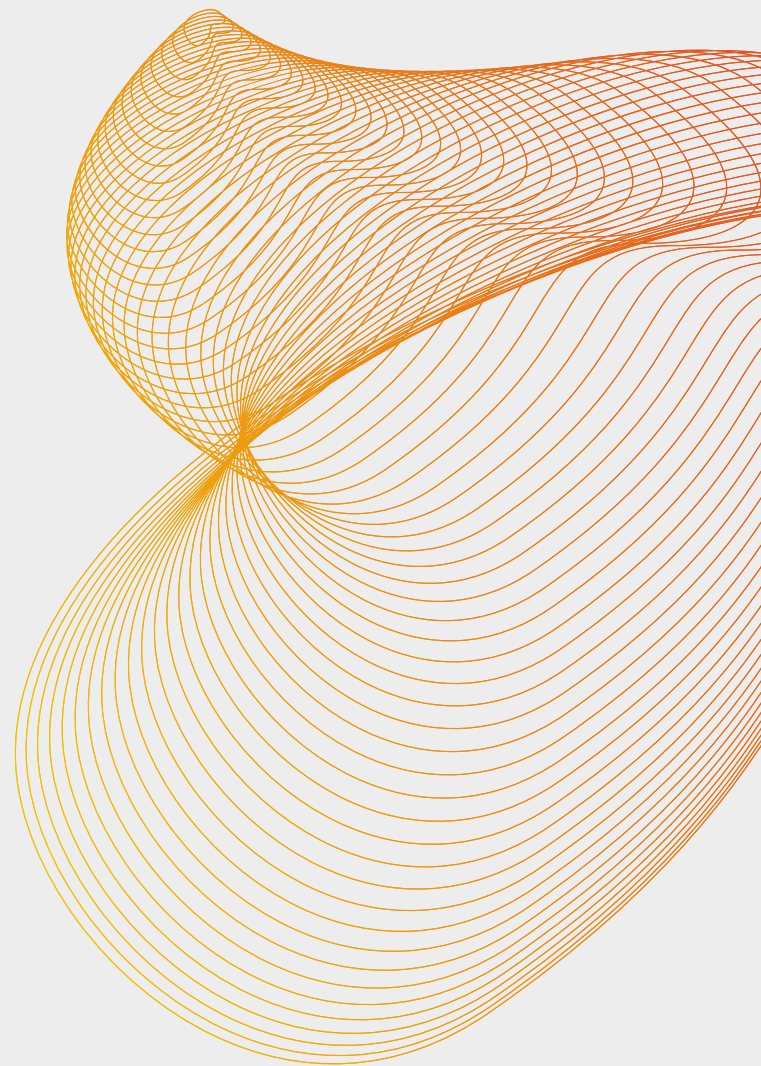
Préparation

- Modèle recommandé par l'ANR sur DMP OPIDOR
- Des exemples de PGD dans la discipline visée
- Mise en place d'outil pour le partage et la sauvegarde des données : ShareDocs (Huma-Num)



Qui ?

- Delphine Bernhard, coordinatrice et "data contact" pour les corpus en alsacien
- Marianne Vergez-Couret, "data contact" pour les corpus en poitevin-saintongeais
- Myriam Bras, "data contact" pour les corpus en occitan
- Stella Medori et Laurent Kevers, "data contact" pour les corpus en corse





Collecte des données : principes généraux

- Critères de sélection des données comparables
 - genres qui représentent ou transcrivent le langage oral : pièce de théâtre, ethnotextes narratifs, sous-titres
 - domaines : agriculture, vie rurale, savoir-faire...
 - autres métadonnées : lieu, dialecte, genre du locuteur, période de temps...
- Constitution de données parallèles
- Disponibilité des ressources : demande d'autorisation d'utilisation





Description des ressources existantes

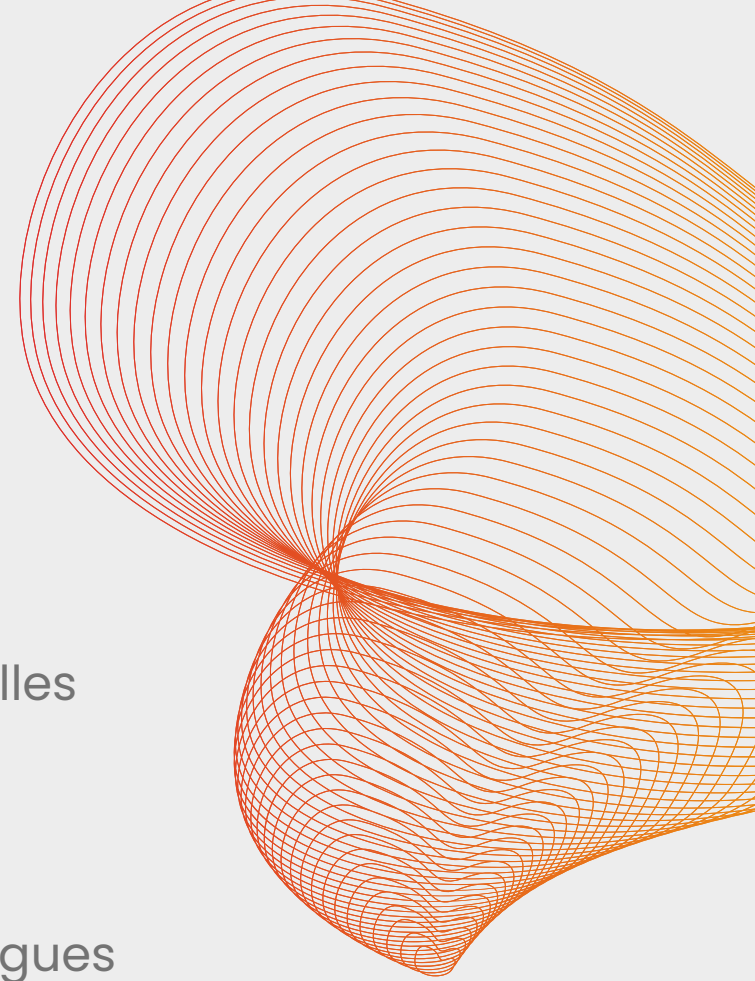
- Listées pour chaque langue
 - Description du contenu
 - Taille/Format
 - Disponibilité/Accessibilité (lien vers les ressources)
 - Références





Description des nouvelles données prévues

- Listées pour chaque langue à propos de nouvelles données brutes :
 - Description des sources disponibles
- Méthodes et outils partagés pour toutes les langues
 - Taille et format des données brutes : XML TEI
 - Taille et format des données annotées : CONLL-U
 - Outils : Arborator Grew, INECpTION, Analog, Varialog
 - Description
 - Lien





Documentation et qualité des données

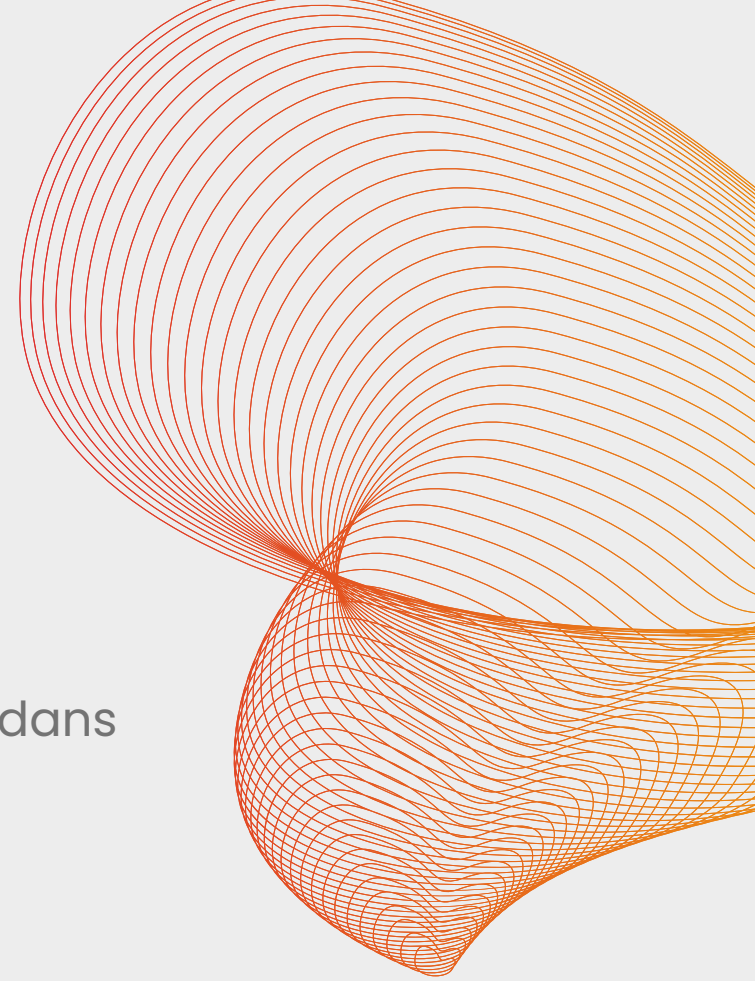
- Standards pour les corpus : XML TEI
- Métadonnées riches : auteur/locuteur, genre, date, lieu, dialecte, graphie, acquisition, disponibilité...
- Méthodologie d'annotation
 - Guides d'annotation : Universal dependencies
 - Description du processus d'annotation





Stockage des données

- Comment sont stockées les métadonnées ?
 - Base de données Access, Heurist...
- Comment sont stockées les données partagées dans le projet ?
 - ShareDocs





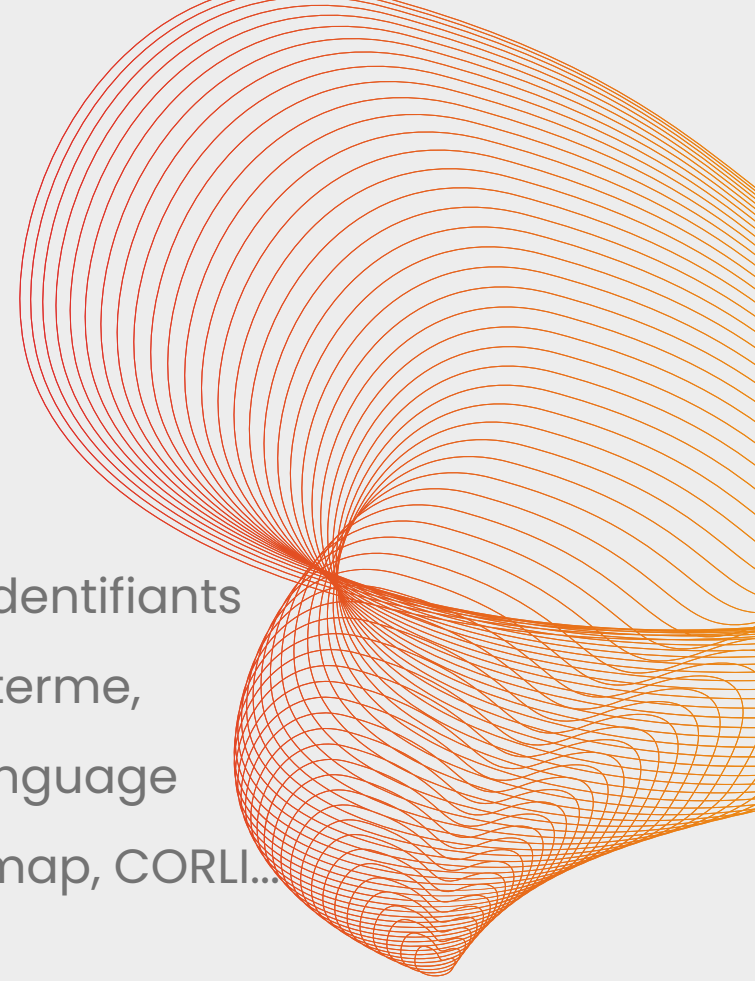
Sécurité et protection des données

- Comment est assurée la sécurité et la protection des données sensibles et la protection des données personnelles ?
 - Dire qu'on était pas directement concernés a priori
 - Montrer qu'on est conscient des cas de figure où cela pourraient devenir nécessaire
 - Montrer qu'on connaît les organismes à contacter en cas de besoin
 - Droits d'auteur
 - On était directement concernés
 - Identification des éditeurs et ayant-droits et demande d'autorisation
 - Sans autorisation : "Fouille de textes et analyse de données" du groupe
- Développement des bonnes pratiques, Comité pour la science ouverte, 2019



Diffusion des données

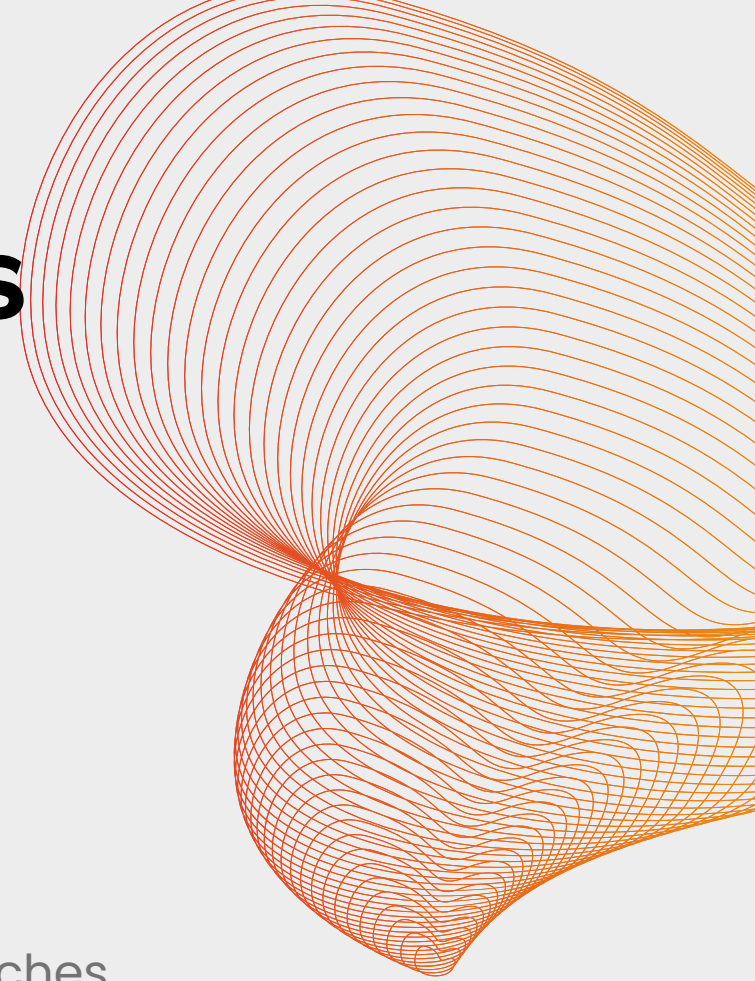
- Trouvable :
 - Plateforme d'entrepôt de données Nakala, identifiants uniques DOI, demande d'archivage à long-terme,
 - Dépôts github, UD release, CLARIN Virtual Language Observatory, European Language Grid, LRE map, CORLI...
- Accessible :
 - mise à disposition des métadonnées
 - data papers (procédure de constitution et d'annotation des corpus)
- Interopérable : formats standards de données et d'annotation (CSV, XML, UD, CONNL-U...)
- Réutilisable : licence CC-BY 4.0





Ressources humaines et financières

- Moyens humains : Ingénieurs d'étude et de recherche recrutés sur le projet
 - Constitution et annotation des corpus
 - Fairisation
- Moyens financiers : Budgets pour différentes tâches du projet
 - Numérisation, transcription, traductions, acquisition de droits d'auteurs auprès des éditeurs...



DiViTal

DE REGIONÀLSPROCHE IN FRANKRICH MEH **VITÀLITÄIT** UN SICHTBÀRKEIT IM **DIGITÀLE** RAUM GANN
LINGUISTISCHI BSCHRIEWUNG UN KOMMENTIERTI TEXTSÄMMLUNGE

FÀ CRESCE A **VITALITÀ** È A **VISIBLITÀ** NUMERICA DI E LINGUE DI FRANCIA
DESCRIZIONE LINGUISTICHE È CORPORA ANNUTATI

FAR CRÉISSER LA **VITALITAT** E LA **VISIBILITAT** NUMERICA DE LAS LENGAS DE FRANÇA
DÉSCRIPCIONS LINGÜISTICAS E CORPUSSES ANOTATS

ENBOUNESI LA **VITALITAI** LIMÉRIQUE É PI LA VEYABLLETAI DAUS PARLANJHES DE FRANCE
DESCRIPCIONS LENGHISTIKES É CORPUS ANNOUTAIS

ACCROÎTRE LA **VITALITÉ** ET LA **VISIBILITÉ** NUMÉRIQUE DES LANGUES DE FRANCE
DESCRIPTIONS LINGUISTIQUES ET CORPUS ANNOTÉS

INCREASE THE **DIGITAL VITALITY** AND **VISIBILITY** OF LANGUAGES OF FRANCE
LINGUISTIC DESCRIPTIONS AND ANNOTATED CORPORA