



GT Voix et Langage

Groupe DATA

Objectifs

- 1 Identifier les données et besoins existants
- 2 Définir les critères de qualité et d'exploitabilité d'un dataset
- 3 Sécuriser le cadre juridique de collecte et d'exploitation des données
- 4 Mutualiser les efforts de création de dataset de haute qualité

Principales actions & livrables

- 1 Recenser la liste des datas accessibles (particularités + limites):
Analyse de l'existant: Text to Speech, Speech to Text, NLP & NLU
Ateliers exploratoires Corpus *French Media Data Set*
- 2 Définir les prérequis du dataset : nbr speaker, durées, labels (...)
- 3 Valider le cadre juridique associé
- 4 Evaluer les opportunités de création d'un dataset et mutualiser les efforts associés