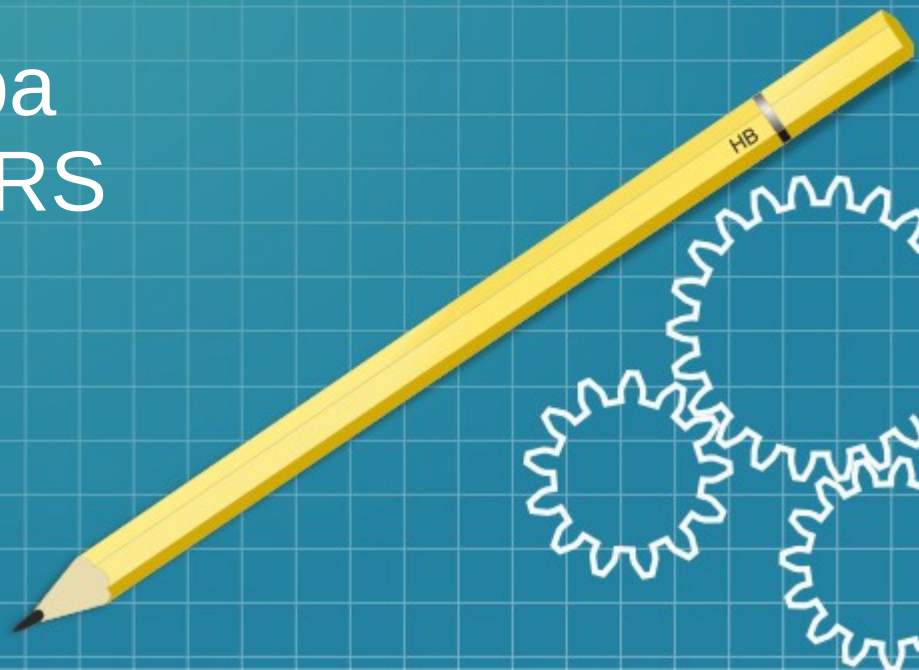
Projet AlimCorp

Atelier d'information 26/10/2023

Y. Stroppa
ASTN/CNRS



Sommaire

- Introduction
- Présentation de la plateforme
- Perspectives
- Démarche
- Conclusion

Introduction

- Naissance du projet
  - (Eslo - GFI perte de maîtrise
    et de savoir-faire pour le Labo)

Discussion entre ASTN et GB pour scinder le projet GFI en deux : alimentation et exploitation.

- Objectif initial
  - reprendre la main et la maîtrise de ce type d'outillage
- Evolution du projet
  - impulsion d'ASTN ==> orientation multi-corpus, messagerie, locuteurs, liens de parenté .....modèles plus libre, accompagner les utilisateurs, apporter du savoir-faire.
- Maîtrise du projet
  - développement interne (1 personne)
  - garantie de l'évolution
  - garantie de l'adaptation

Introduction

- Un des objectifs de ce projet est de se rapprocher des personnes qui collectent les corpus
- De leur faire profiter d'un retour d'expérience afin d'avoir une efficacité, une optimisation de son temps et d'éviter de faire des erreurs.
- D'éviter les oublis (consentement à récupérer au bon moment)
- De pouvoir profiter d'une copie supplémentaire qui doit être sécurisée (pour éviter les pertes)
- D'alimenter à plusieurs avec une validation et un suivi de différents niveaux

Introduction

- d'orienter les développements pour les adapter à vos besoins.
- de disposer de différentes versions de la plateforme : hébergée ou localisée
- de pouvoir exporter vers d'autres plateformes selon différents formats
- de pouvoir alimenter également la plateforme DeepCorp (en cours) pour l'exploitation et le traitement de vos corpus.

Présentation de la plateforme

- Frontal web de type Wordpress
- Possibilité d'installation sous différents environnements à prévoir en local ou pas
  - Linux
  - Mac OS
  - Windows à prévoir
- La plateforme s'inscrit dans un ensemble autour d'un socle commun

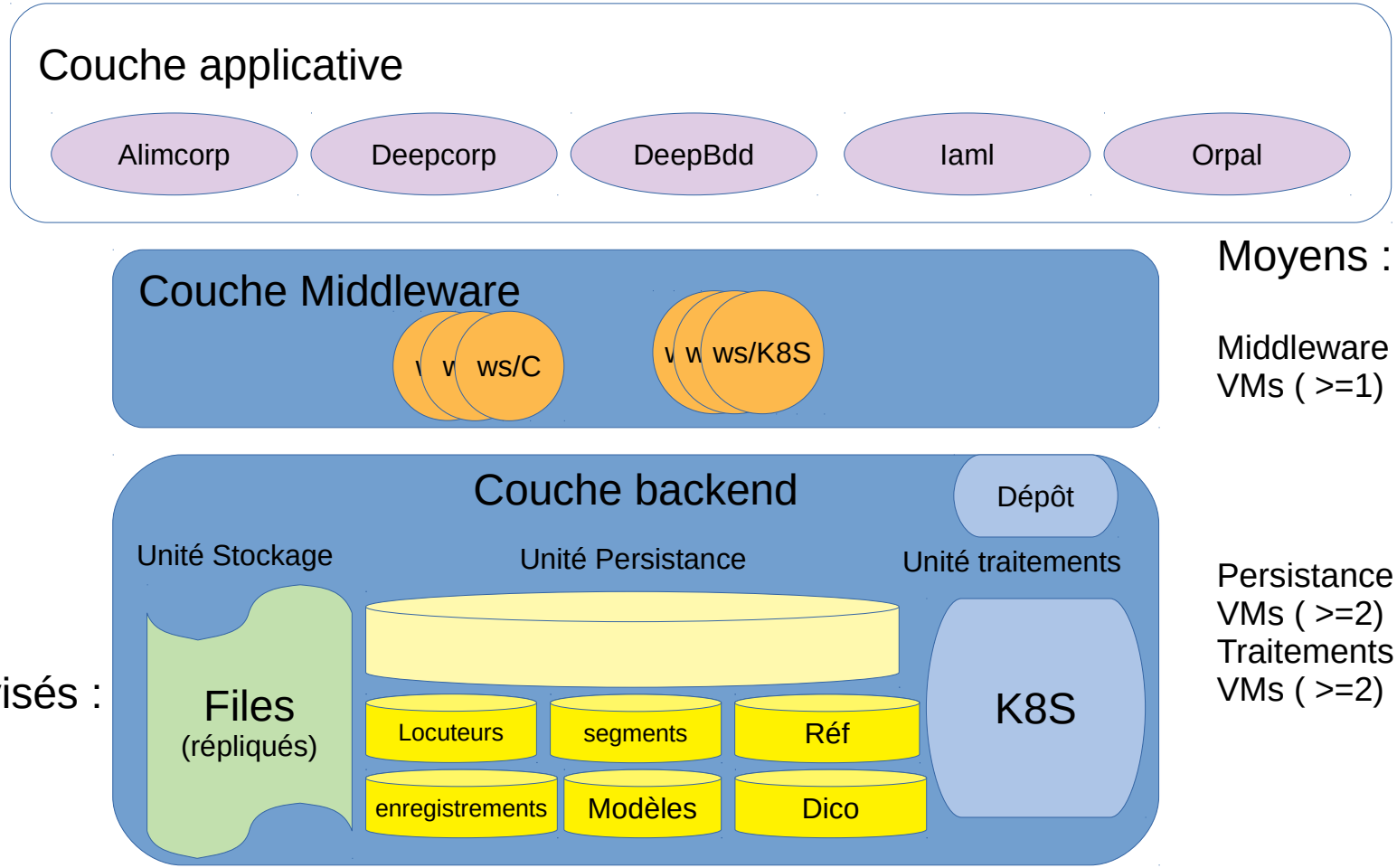
Présentation de la plateforme

Objectifs/Mots clés :

Capitaliser
Mutualiser
Partager
Réutiliser
Sécuriser

Exécuter

Pérenniser
Stocker



Moyens :

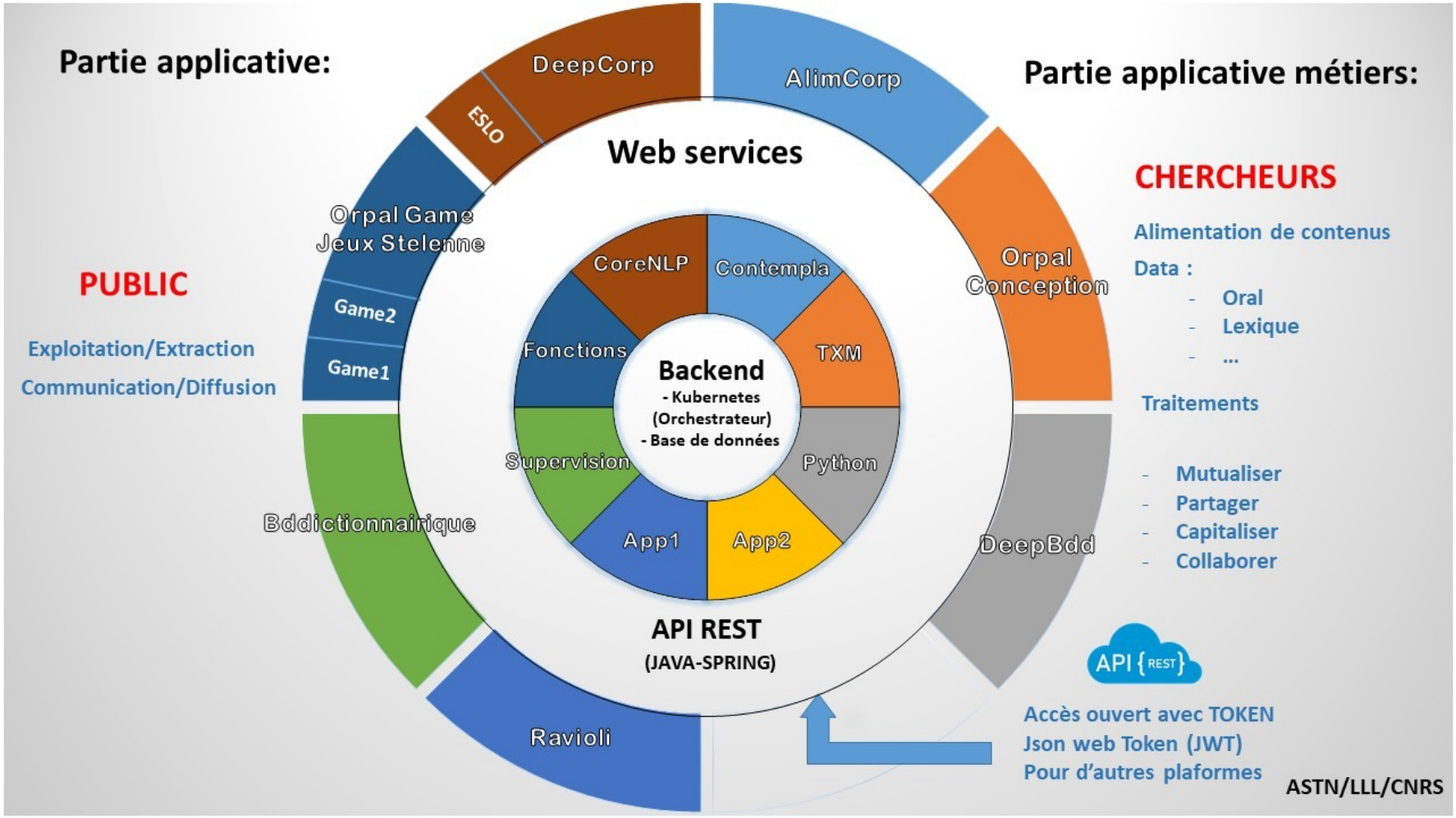
Middleware
VMs (>=1)

Persistance
VMs (>=2)
Traitements
VMs (>=2)

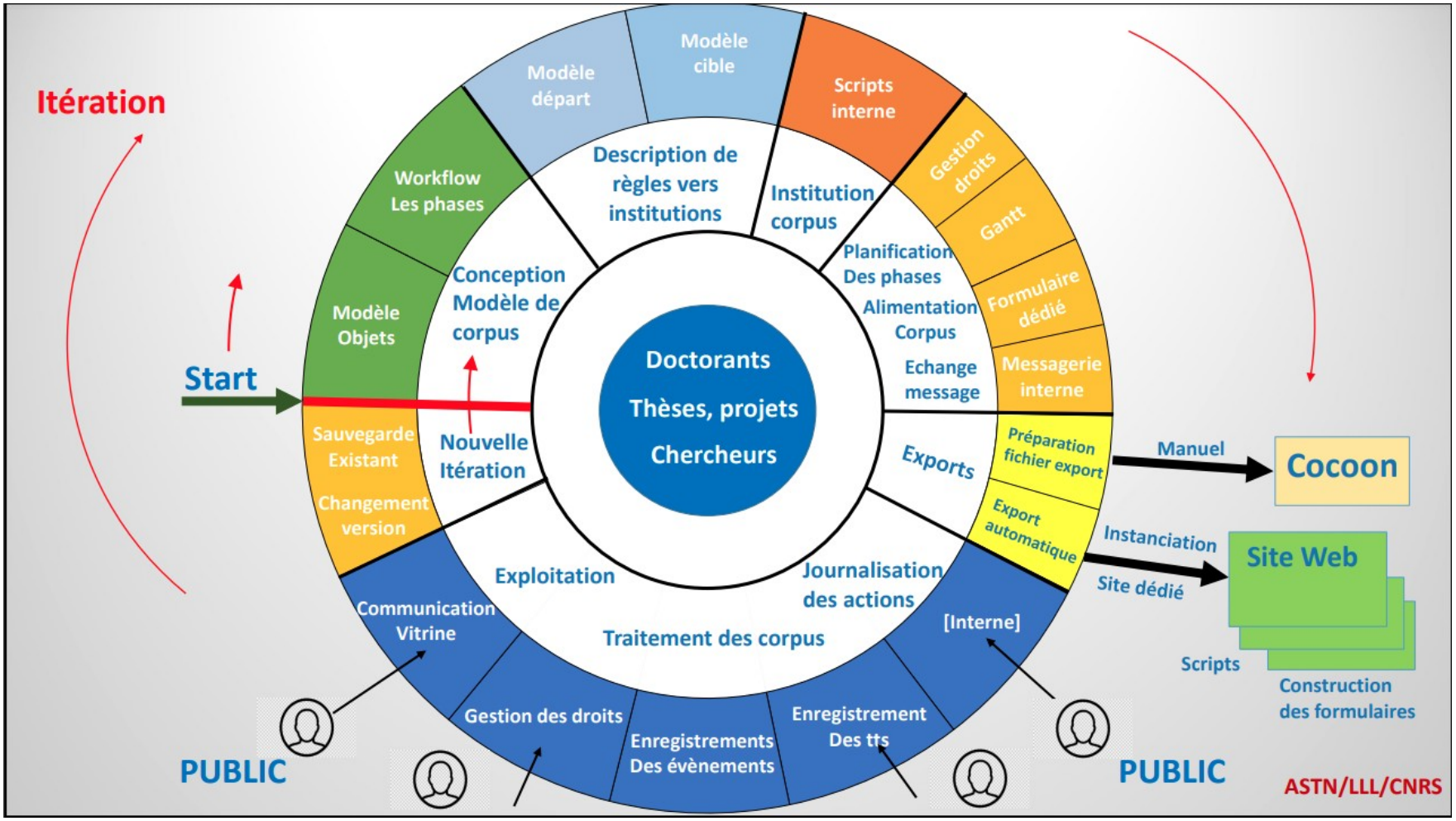
Partenaires :
LLL

Utilisateurs visés :
ASTN

Contexte global



Cycle possible



Les étapes dans AlimCorp

- Inscription sous la plateforme
  - Libre sans modérateur pour le moment
- Création d'un Corpus
  - Association par défaut des droits à l'utilisateur courant
  - seul l'utilisateur peut modifier et alimenter son corpus

  Que se passe t-il lors de la création :
    - Préparation sur le serveur d'un espace dédié pour le stockage par la suite des différents fichiers de transcription et sonore.
  - Saisie de locuteurs
  - saisie des enregistrements selon les différentes catégories
    - Liste de références
  - Possibilité d'édition de la saisie pour un contrôle
    - En plus du taux de remplissage

Gestion des droits

- Idée est de pouvoir fixer les droits associés à son corpus pour permettre un travail collaboratif ou une vision externe (par le Directeur de thèse par exemple)
- Chaque propriétaire de Corpus définit et contrôle l'accès aux données.
- Un droit affecté peut être révoqué.
- On doit pouvoir s'inscrire, demander un accès à un corpus et d'être affecter des droits spécifiques en fonction du profil désiré pour un laps de temps donné
-

Etape de validation

- Objectif est de pouvoir dans le cadre d'une saisie répartie ou déléguée d'effectuer un contrôle et de bloquer la saisie pour permettre par la suite son exportation
  - Exportation vers des guichets institutionnels
  - et vers DeepCorp (pour permettre une exploitation/traitements)
  - et conserver un état des données.

Précision sur le stockage des locuteurs

- Suite au contexte ESLO
  - locuteur présent dans plusieurs enregistrements mais sur une période de temps longue (dans ce type d'action)
  - avec un changement de statut possible du locuteur sur cette période
- Solution relationnelle classique
  - on a prévu initialement un attribut multi-varié pour porter cette évolution
  - sur quel attribut s'applique cette évolution
  - il faut adapter l'application pour cette prise en compte
  - un peu complexe et pas très flexible

Précision sur le stockage des locuteurs

- Solution retenue dans ESLO
  - On duplique le locuteur et on lui affecte une codification spécifique
  - Attention comment fait-on pour savoir que c'est le même et que l'on souhaite regrouper les informations (utilisateurs????) et surtout que l'on applique systématiquement les mêmes règles ;
  - Autre problème, si on modifie les propriétés d'un locuteur, comme les enregistrements sont reliés, il y aura implicitement répercussion dans les extractions.

Précision sur le stockage des locuteurs

- Solution retenue pas ASTN
  - on découple et on isole en même temps les différentes informations associées à un enregistrement
  - les informations de référence sont insérées directement dans l'enregistrement pour constituer un tout cohérent et intègre.
  - ceci à plusieurs effets :
    - la vision d'un enregistrement se rapproche du concept de Document en NoSQL.
    - les modifications des locuteurs peuvent s'effectuer de façon désynchronisés des enregistrements et n'auront pas d’impact implicite
    - les exports et manipulations des enregistrements sont facilités et plus simples (pas de liaisons a reconstituer)
    - les enregistrements sont donc indépendants des éléments de référence
    - ce qui permet de faire évoluer plus facilement
    - cela impose que dans les mécanismes de recherche, on fasse des introspections pour remonter les différents éléments.

Description JSON

| 10 | 2 | public | Enregistrement | Frapeor_enr_17 | {"Duree": "420", "Sujet": "(text_and_corpus_linguistics) Français (Ethnologue: fra)", "Droits": "Copyright (c) 2023 Université d'Orléans/LLL", "Format": "mp3", "Langue": "Fr", "Statut": "valid", "Created": "2023-09-01", "Sommaire": "", "Categorie": "Entretien", "Locuteurs": [{"AM": "", "Nom": "Elahounana", "Date": "", "Sexe": 1, "Droits": "", "Niveau": "Secondaire complet", "Prenom": "Abdel", "Adresse": "", "Anonyme": 1, "Created": "2021-10-08", "Langues": [""], "Tranche": "25/35", "Updated": "16/09/2021", "Domicile": "", "Nom_fille": "", "Situation": "Ouvrier", "NB_enfants": 0, "Identifiant": "FRAPEOR_Loc_01", "categ_socio": "", "Anneearrivee": 0, "Commentaires": "", "Info_enfants": "", "Age_finetudes": 0, "Lieu_naissance": "", "Annee_naissance": 1982, "Prof_terme_propre": "", "Remarques_diverses": ""}], "Remarques": "", "Acoustique": 2, "Chercheurs": ["Farah El-Mouhrad"], "File_enreg": "", "Identifiant": "Frapeor_enr_17", "Localisation": {"TGN": "7008337", "Point": "east=1.886; north=47.906", "Lieu_spatial": "Orléans"}, "Participants": "1 participant", "VersionCorpus": 2, "Precisioncategorie": "Entretien", "Precision_acoustique": "Son faible", "Description_participants": "", "Description_enregistrement": ""}, "Transcriptions": [{"Date": "2023-09-01", "Mode":"Automatique","Guide": 5, "Droits": "Copyright (c) 2023 Université d'Orléans/LLL", "Statut": "valid", "Created": "2013-07-11", "Fichier": "Frapeor_enr_17.json", "Updated": "2018-06-13", "Sommaire": "", "Remarques": "", "Description": "", "Identifiant": "Frapeor_enr_17_1", "Transcripteur": "Whisper_UI"}]} | NULL |

Suivi de la saisie

- Comme on souhaite accompagné les utilisateurs, il faut donner des indications sur plusieurs métriques :
  - Métrique d'un taux de remplissage par enregistrement
    - approche quantitative : indication enregistrement par enregistrement du taux de remplissage des méta-données, fichier transcription et enregistrement (reste à faire) -- pas d'obligation mais juste attirer l'attention de l'utilisateur
    - approche qualitative : à définir ????
  - Métrique d'un taux de remplissage pour le corpus
    - Analyse globale des méta-données du corpus et mise en évidence des zones +/- bien saisies
  - lors de la validation :
    - Contrôle des métriques pour vérifier la pertinence des informations fournies
- Ces différents contrôles doivent être définis au niveau du Corpus comme des objectifs que l'on s'impose (pas de blocage mais juste des rappels)

Perspectives

- Comment s'organiser pour être efficace
  - Travailler avec des personnes qui ont réellement des corpus à constituer
  - Tester et alerter des défauts ou dysfonctionnements rencontrés (site mantis)
  - Documenter et élaborer les tutos associés à ce type de plateforme (à poser sur le site ASTN)
  - Elaborer les parties manquantes :
    - construction de modèle et de workflows
    - consentement

Démarches

- La proposition est de parier sur les doctorant(e)s pour avancer sur ce sujet.
  - Plus réactifs car plus concernés
- Mettre en place un processus d'amélioration continue avec une réelle implication et ne pas juste compter sans tester ce qui va et ce qui ne va pas ....
- L'étape de mise au point d'un logiciel est longue d'autant plus qu'il intègre un processus métier qui peut être complexe avec des étapes définies de validation et de contrôle.
- Trouver des ressources complémentaires ou passer en open-source et ouvrir la plateforme dans un github ASTN/CNRS.
- Documenter et faciliter l'utilisation de la plateforme
- Finaliser la partie Modèle de la plateforme : modèle de son corpus
- Insertion du module édition des transcriptions pour annotations (reprendre la partie en cours de développement du projet FRAPéOR)
  - axe enrichissement du corpus par des annotations de différentes natures (Sémantique, Syntaxique, Prosodique ...)

Conclusion

- Le projet nous appartient, il nous est possible de le façonner selon nos besoins ou pas.
- Cette opération peut apporter efficacité et laisser du temps pour l'analyse et traitements.
- Possibilité d'enrichir les données par la suite afin de les faire vivre grâce à l'éditeur à incorporer
- Possibilité d'utiliser de la transcription automatique (en délégation via l'appel à des Web services complémentaires locaux ou distants)
- Développer les projets connexes qu vont s'appuyer sur ce socle de données évolutives.
- Autres ????