Présentation des réflexions du groupe de travail TEI-Nakala : Faciliter l'insertion de l'entrepôt Nakala dans nos chaînes de traitement de corpus encodés en XML-TEI

Victoria Le Fourner, Gwenaëlle Patat, Guillaume Porte

Présentation des réflexions du groupe de travail TEI-Nakala

Faciliter l'insertion de l'entrepôt Nakala dans nos chaînes de traitement de corpus encodés en XML-TEI

Victoria Le Fourner (MESHS), Gwenaëlle Patat (MSHB), Guillaume Porte (ARCHE UR3400)
Journées EVEILLE, 06/05/2022

Le Groupe de Travail TEI-Nakala

Genèse du Groupe de Travail

- **Science Ouverte** et **principes FAIR** : la nécessité de déposer nos données dans des entrepôts certifiés

- Une nouvelle version de **Nakala** fin 2020

- Nakala comme **point d'entrée et de sortie des données** : faciliter l'import des données et des métadonnées selon les types de corpus/outils de traitement et de publication utilisés

- **Centraliser les initiatives** (s'y retrouver dans l'existant) et **harmoniser les pratiques** : gagner en efficacité et en qualité des données produites

Genèse du Groupe de Travail

Tour d'horizon des initiatives pour faciliter le dépôt de corpus en TEI dans Nakala :

- Les programmes de Michaël Nauge TeiHeaderCahierToSpreadSheet pour passer des métadonnées en TEI à un tableur et NakalaPyConnect pour générer des scripts json pour l'API Nakala.
- L'application web MyNkl, développée par Andrès Felipe Echavarria Pelaez et Ala Eddine Laouir sous la direction de Fatiha Idhmand dans le cadre du consortium Cahier.
- Solutions en cours de développement à la MRSH pour faciliter le dépôt de données sur Nakala.
- Retour d'expérience de Floriane Chiffoleau, ingénieure sur le projet DAHN, édition de la correspondance de Paul d'Estournelles de Constant (1914-1919), sur utilisation de Nakala pour déposer son corpus en XML-TEI.

⇒ **Former et informer** sur ce qui existe, ce qu'il faudrait améliorer pour prendre en compte nos besoins partagés.

Genèse du Groupe de Travail

Constitution du GT :

- 27 abonné·es sur la liste de diffusion suite au recueil de besoin (issus des correspondants Huma-Num et de la plateforme Scripto du RnMSH)
- 3 membres référents : Victoria Le Fourner, Gwenaëlle Patat, Guillaume Porte
- Mise en place d'une preuve de concept avec le soutien d'Huma-Num

Le GT s'est réuni 6 fois depuis juin 2021 : présentations des problématiques et réflexions existantes, définition des besoins partagés, des objectifs du GT, du workflow de la plateforme finale et avancées sur les livrables.

Besoins identifiés

- (Mieux) définir l'usage de Nakala pour les corpus TEI
- (Mieux) insérer Nakala dans les différents environnements de production
- Automatiser le mapping des données de fichiers TEI vers le Dublin Core et la récupération des DOI
- Permettre le dépôt en masse via une interface graphique intuitive où le code source est accessible

Besoins identifiés

- Modéliser les différents environnements éditoriaux (encodage → publication)
- Favoriser une communauté d'expertise autour des chaînes de production
- Améliorer les connexions / le passage entre différents outils
- Permettre la maintenance d'instances communes de publication

Nakala ?

Huma-Num
la TGIR des humanités numériques

Infrastructure de recherche financée par le Ministère de l'ESR dédiée à la **gestion des données en SHS**.

- Des services pour les données
- Des Consortiums
- Le HN Lab
- Coordination des communautés européennes et internationales

Maillage thématique et géographique du territoire


DISSÉMINATION
PUBLICATION
SIGNALEMENT
ANIMATION
COMMUNAUTÉS
NATIONALES &
INTERNATIONALES
HN
HN Lab
OUTILS COLLABORATIFS
STOCKAGE
TRAITEMENT
QUALITÉ DES DONNÉES
ACCOMPAGNEMENT

ORGANISATION

Des services pour organiser le travail collaboratif autour de vos données.

- ShareDocs
- GitLab
- Kanboard
- Mattermost

TRAITEMENT

Des services et outils spécifiques pour le traitement et l'analyse de vos données.

- Calcul statistique et environnements R
- Logiciels d'enquête et d'analyse de données
- Reconnaissance de caractères
- Puissance de calcul (+ CC-IN2P3)

...

PUBLICATION

Vos données peuvent être publiées depuis Nakala sur le web et signalées dans Isidore, moteur de recherche pour les SHS.

- Hébergement Web
- Machines Virtuelles
- Nakala
- Isidore

COLLECTE

Des services de stockage sécurisé pour la collecte et la création de vos données.

- ShareDocs
- Huma-Num Box

PRÉSERVATION

Huma-num vous accompagne pour le dépôt et la documentation de vos données dans Nakala, entrepôt pour les données en SHS.

- Nakala
- Huma-Num Box
- Préservation à long terme (+ CINES)

RÉUTILISATION

Vos données entreposées dans Nakala et signalées dans Isidore sont réutilisables.

- Portail web
- API
- Triplestore
- OAI-PMH

Les entrepôts de données

Stocker des données de recherche, y accéder et les réutiliser

- Disciplinaire, multi-disciplinaire
- Institutionnels, nationaux, européens, mondiaux…

Ortolang

nakala

DRYAD

Dropbox

figshare

Data INRAE

zenodo

Entrepôt de données de recherche destiné à accueillir, conserver, rendre visible et accessible les données de recherche.

Enregistrer des données, les décrire en vue de les exposer et les rendre réutilisables.

Une **donnée** : Un objet constitué de n fichier et d'une description via les métadonnées

Statut d’une donnée

- **Déposée** : espace de travail non publique (limité à 2Go ou 1000 données)

  2 métadonnées obligatoires : type + titre

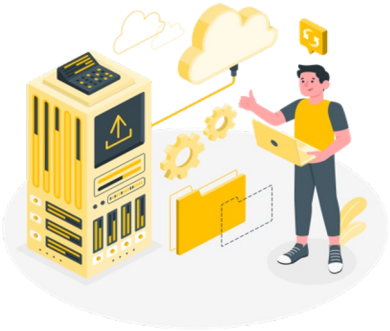
  Possibilité de la modifier et de la supprimer

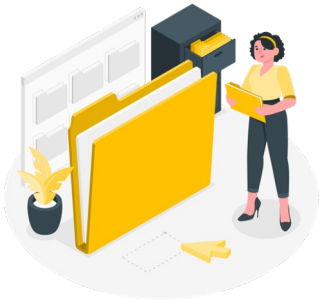
- **Publiée** : DOI publié

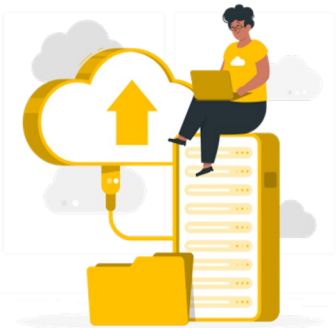
  5 métadonnées obligatoire : type + titre + auteur + date + licence

  Possibilité de modifier (version antérieure conservée) mais impossibilité de supprimer

Les 7 points à retenir sur Nakala







Simple à décrire

Déposer ses données via *drag and drop*

Description des métadonnées adossées sur le vocabulaire Dublin Core

5 métadonnées obligatoires

Gestion de tous les types de fichiers

Possibilité de mettre plusieurs fichiers dans une donnée

Page de présentation d'une donnée propose une visionneuse pour certains type de fichier

Gestion fine des droits d'accès aux données

Différents types de droits (administrateur, éditeur, lecteur)

Possibilité de mise sous embargo des fichiers

Organisation en collections

Organiser ses données en collections (ensemble de données)

Privée ou Publique

Exposer ses données dans un site dédié

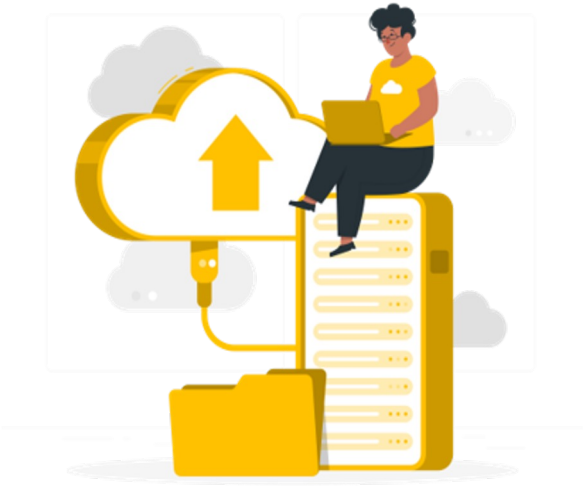
Plug-in Nakala_Press

Éditorialiser ses données dans un site dédié

Lié à une collection

Administrable via une interface graphique

Gestion du stockage par Huma-Num

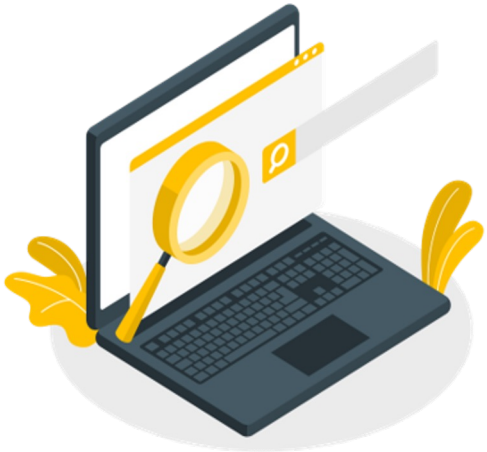


Lien avec le CINES sur demande pour un archivage pérenne

Stockage en France des données de la recherche

Pas de limite de dépôt

Recherche des données facile



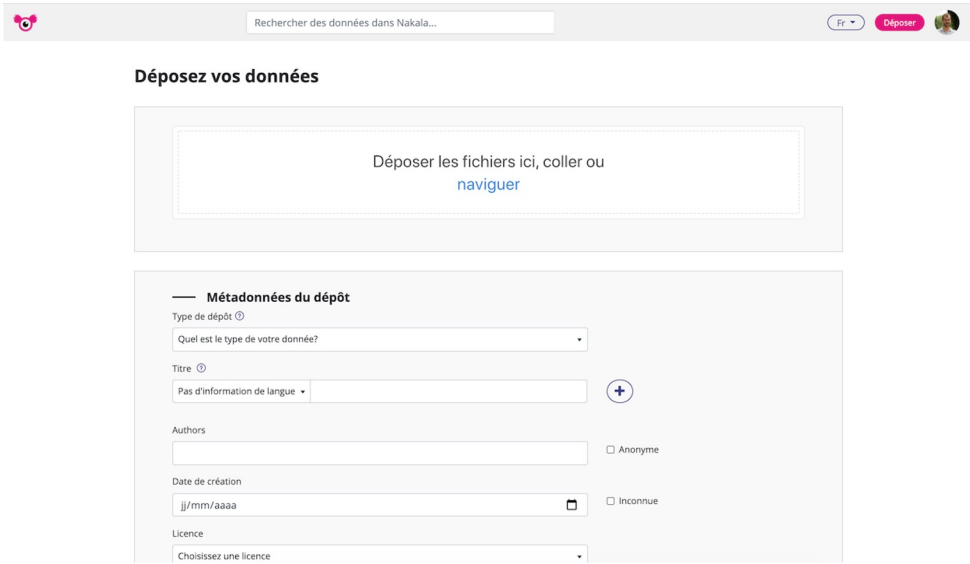
Moteur de recherche

Serveur OAI PMH

Triple Store RDF

API REST

Une interface facile d'utilisation



Rechercher parmi les 555 003 données de NAKALA...

Fr | + Uploader

‹ Retour

Le campement de Gergovia, lettre ouverte

FR | EN

ID : 11280/446befde

80 Consultations | 24 Téléchargements

Auteurs : Catherine Rioux-Milkovitch, Aurelia Vasile

Lettre ouverte des personnels administratifs et techniques, enseignant-e-s, chercheurs-chercheuses de l'université Clermont Auvergne, citoyen-ne-s à propos du campement de Gergovia à destination de Monsieur le Préfet, Monsieur le Président du Conseil Départemental, Monsieur le Maire de Clermont-Ferrand, Mesdames et messieurs les élu-e-s.

FR | EN

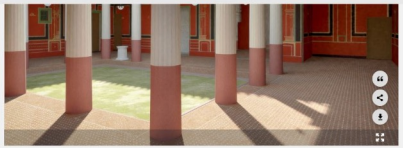
+ Ajouter à une collection

Données

- CAM_G_004.pdf
- IMG_001.jpg
- IMG_002.jpg
- IMG_003.jpg

Tout télécharger

Visualisation



IMG_003.jpg ID : 11280/446befde/003

Lettre ouverte des personnels administratifs et techniques, enseignant-e-s, chercheurs-chercheuses de l'université Clermont Auvergne, citoyen-ne-s à propos du campement de Gergovia à destination de Monsieur le Préfet, Monsieur le Président du Conseil Départemental, Monsieur le Maire de Clermont-Ferrand, Mesdames et messieurs les élu-e-s.

Mots-clés

Histoire | Lettre | Clermont Auvergne | Gergovia

Licence

CC BY 4.0

> Afficher la liste complète des métadonnées

Versions

26 - 09 - 2019

Collections

Fonds de la Galerie des Archives

Citer

Forme par défaut

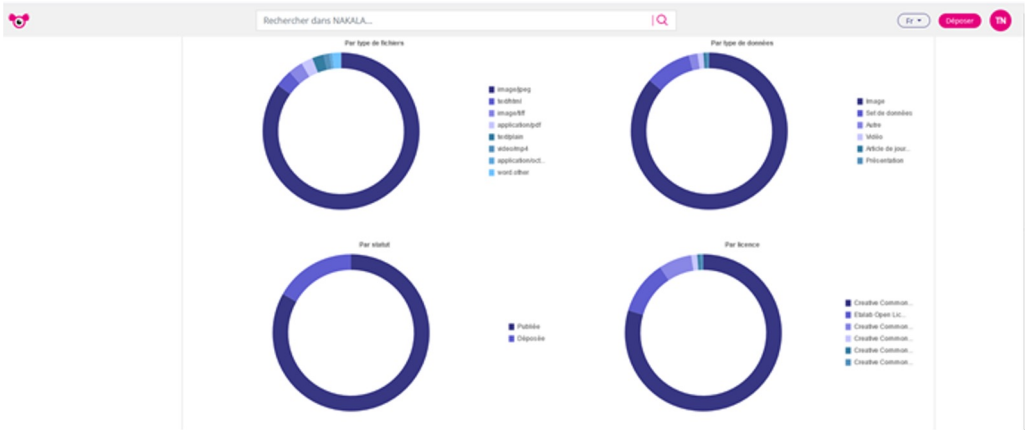
Catherine Rioux-Milkovitch, Le campement de Gergovia, lettre ouverte , 2019

Partager

Exporter

XML | PDF | BibTeX | DC | TEI | DCterms | EndNote | CSV

Publié par **Catherine Rioux-Milkovitch** le **26 - 09 - 2019**


Rechercher dans NAKALA...
Par type de fichiers
Par type de données
Par statut
Par licence
Déposer
TN

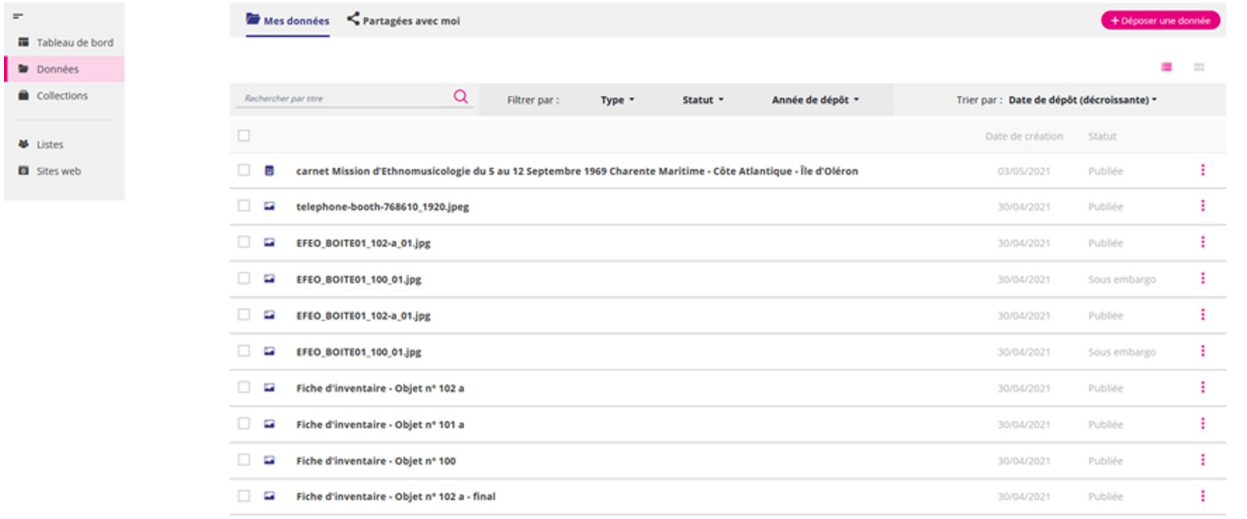
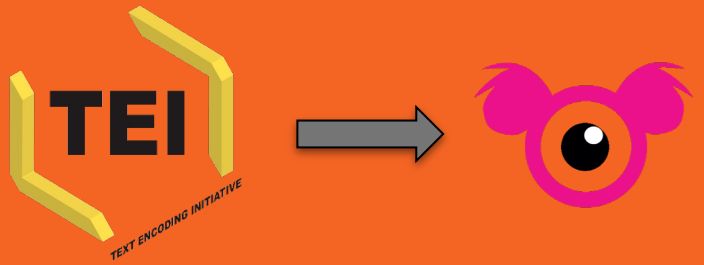


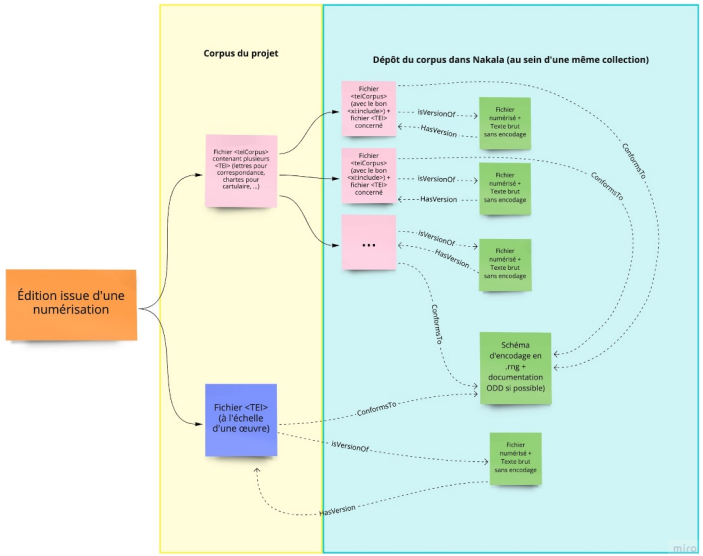
Tableau de bord
Données
Collections
Listes
Sites web
Mes données
Partagées avec moi
+ Déposer une donnée
Rechercher par titre
Filtrer par :
Type
Statut
Année de dépôt
Trier par : Date de dépôt (décroissante)
Date de création
Statut
carnet Mission d'Ethnomusicologie du 5 au 12 Septembre 1969 Charente Maritime - Côte Atlantique - Île d'Oléron 03/05/2021 Publiée
telephone-booth-768610_1920.jpeg 30/04/2021 Publiée
EFEO_BOITE01_102-a_01.jpg 30/04/2021 Publiée
EFEO_BOITE01_100_01.jpg 30/04/2021 Sous embargo
EFEO_BOITE01_102-a_01.jpg 30/04/2021 Publiée
EFEO_BOITE01_100_01.jpg 30/04/2021 Sous embargo
Fiche d'inventaire - Objet n° 102 a 30/04/2021 Publiée
Fiche d'inventaire - Objet n° 101 a 30/04/2021 Publiée
Fiche d'inventaire - Objet n° 100 30/04/2021 Publiée
Fiche d'inventaire - Objet n° 102 a - final 30/04/2021 Publiée

GT TEI-Nakala

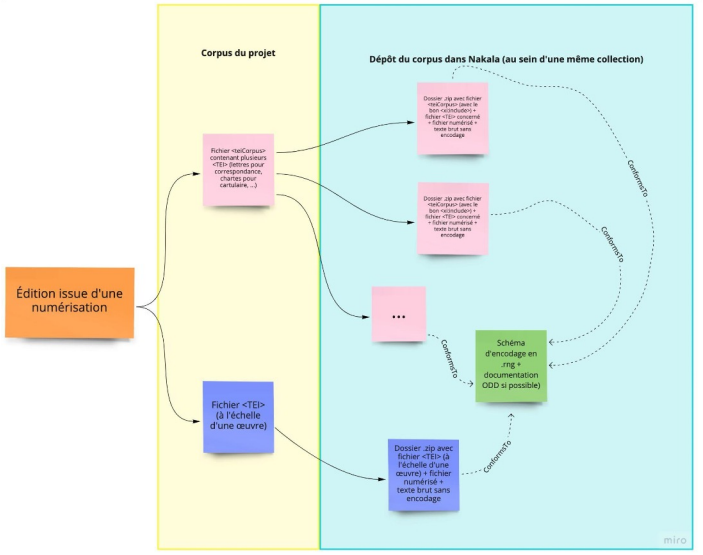
Pistes de réflexions et de développements



Déposer ses corpus dans Nakala : pistes de modélisation du dépôt



Séparer et relier l'édition numérique et les fichiers issus d'une numérisation → désambiguïser ce qui est décrit dans les métadonnées.



Regrouper tout ce qui se rapporte à une entité textuelle dans un dossier .zip : modélisation plus adaptée aux pratiques de citation ?

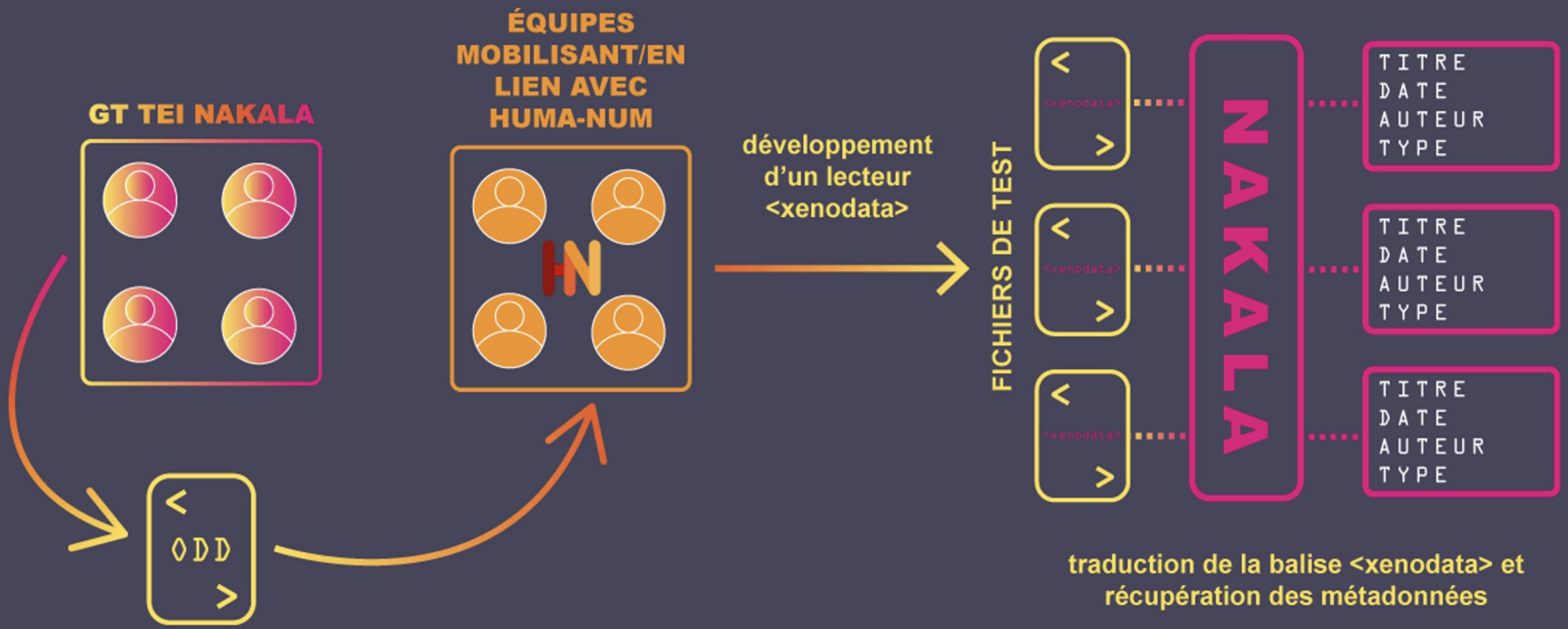
Production de recommandations pour les métadonnées : comment passer des métadonnées présentes dans les fichiers TEI à la notice Dublin core de Nakala ?

- Quelles métadonnées conserver, ajouter ou laisser de côté pour la consultation sur un entrepôt de données ? Un travail de désambiguïsation des champs Dublin Core pour les corpus déposés.

- Des métadonnées spécifiques selon les types de corpus traités (tableur partagé de correspondances entre métadonnées TEI et Dublin Core).

- Génération d'une balise <xenodata> dans le fichier TEI importé en fonction des métadonnées déjà présentes (repository sur GitLab pour partager nos scripts et réflexions, versionner et collaborer).

DÉVELOPPEMENT D'UN LECTEUR <xenodata>



traduction de la balise <xenodata> et récupération des métadonnées

Développement d'un lecteur <xenodata>

- Application d'un schéma d'encodage par le GT pour contrôler la conformité des informations entrées dans le <xenodata>
- Dialogue avec Huma-Num pour formaliser les besoins identifiés et les réponses possibles
- Développement du lecteur <xenodata> lisant le Dublin Core et remplissant automatiquement les champs dans Nakala (preuve de concept)

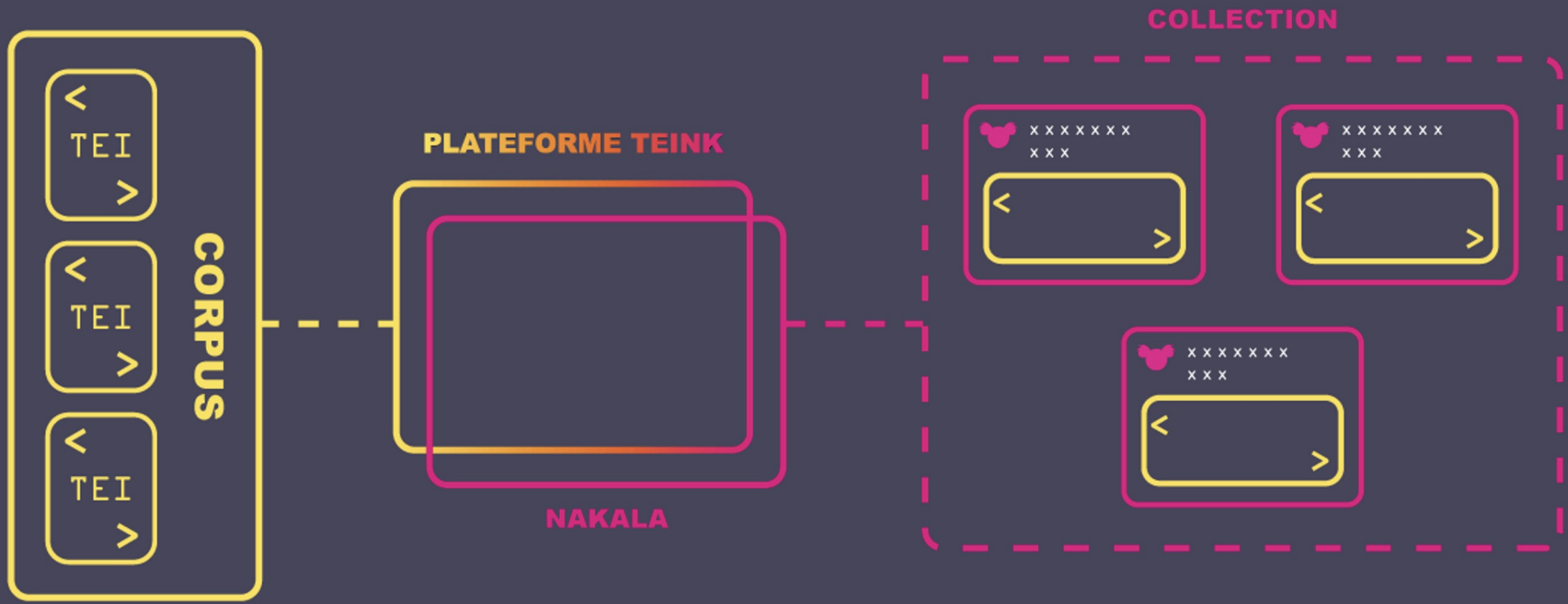
DÉVELOPPEMENT DE LA PLATEFORME DE RAFFINAGE TEINK - possibilité d'interface

Développement d'une plateforme de raffinage

- Application d'un schéma d'encodage pour contrôler les métadonnées obligatoires à mettre dans le <xenodata>
- Alerte si omission et correction possible depuis l'interface graphique
- Possibilité d'ajouter des métadonnées supplémentaires (par exemple, la chaîne de traitement appliquée au corpus)

CRÉATION D'UNE RECHERCHE AVANCÉE DANS NAKALA

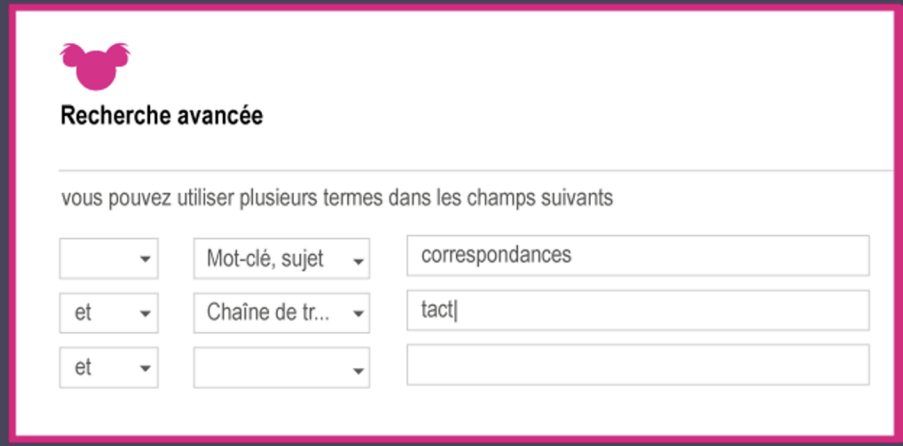
LECTURE D'UN CORPUS PAR LA PLATEFORME TEINK



Gestion des corpus par la plateforme de raffinage

- Besoin de créer les collections souhaitées en amont sur Nakala
- Drag and drop des fichiers constituant un corpus
- Gestion des métadonnées
- Gestion des DOI

RECHERCHE DE COMMUNAUTÉ VIA LES MÉTADONNÉES - exemple avec TACT

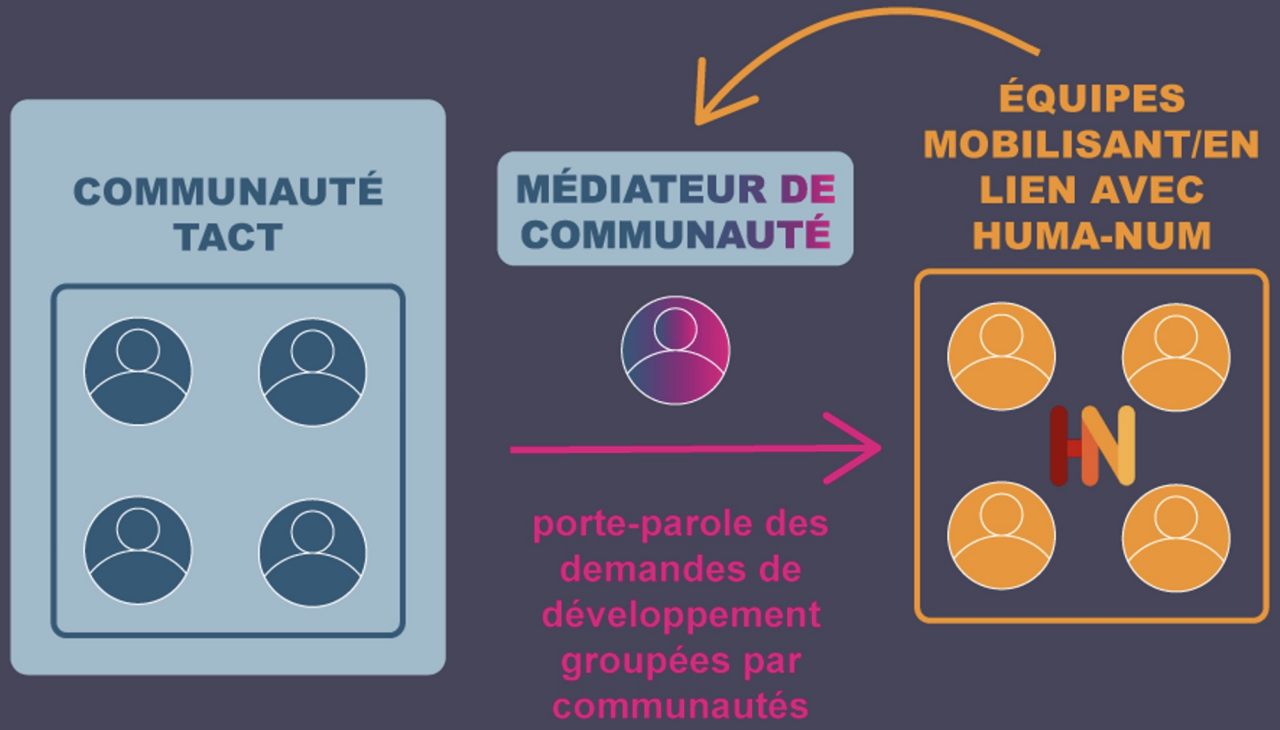


Nakala et ses communautés

- Par la recherche avancée, possibilité de trouver des exemples de dépôts similaires → harmoniser les pratiques

- Le nom du déposant, associé à l'annuaire interne à Huma-Num, pourrait être cliquable afin de le contacter

RECHERCHE DE COMMUNAUTÉ VIA LES MÉTADONNÉES - exemple avec TACT



Gestion Nakala et ses communautés par Huma-Num

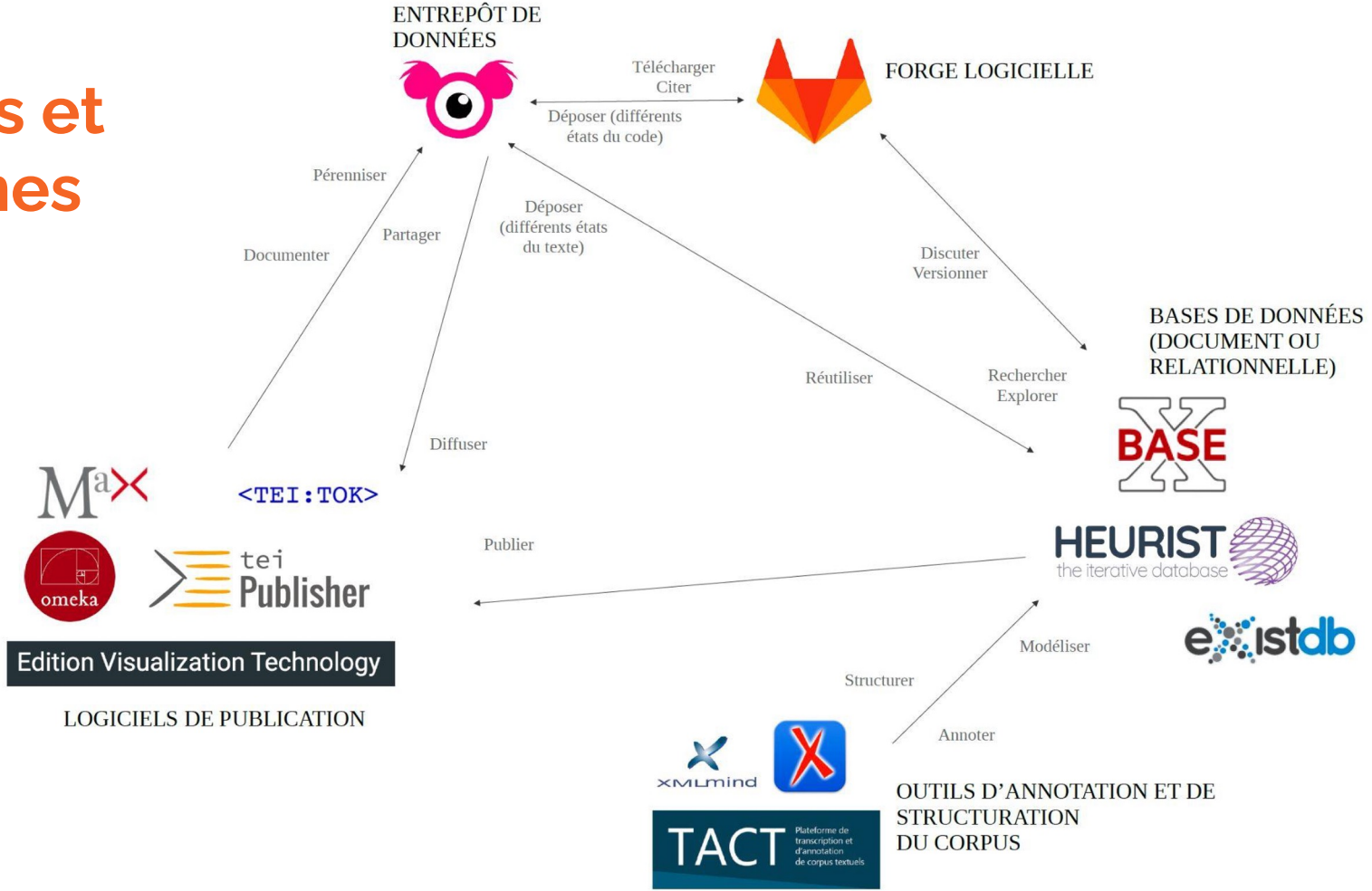
- Communautés autour de la TEI mieux identifiées en fonction des chaînes de traitements mises en place
- Demandes de développements par ces communautés modérées par un médiateur de communautés
- Lien des médiateurs de communautés avec la TGIR pour faciliter les échanges

Intégrer Nakala à une chaîne éditoriale

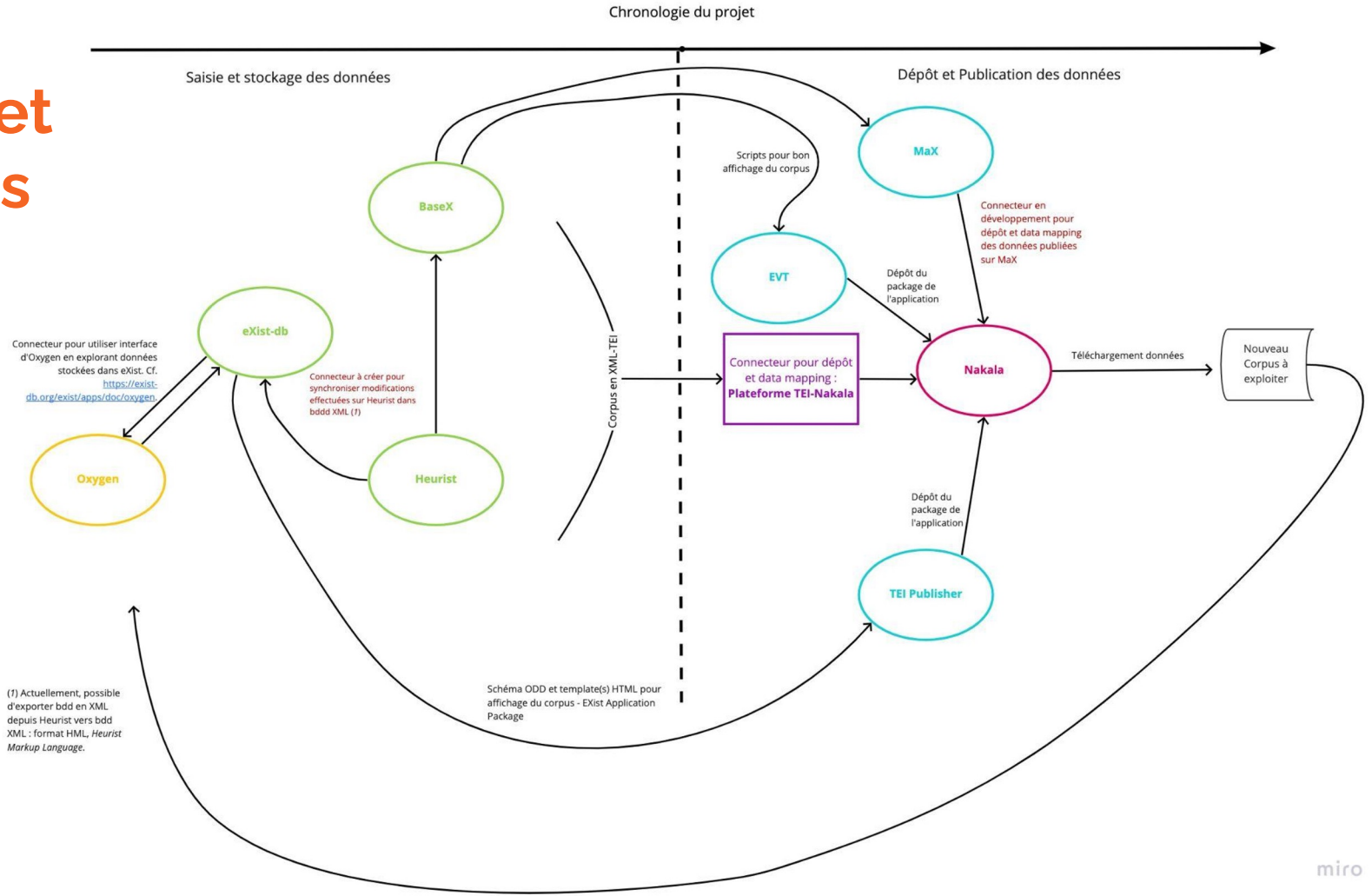
Nakala

- quel usage ?
  - dépôt strict / fin de chaîne
  - support à une publication => data paper, édition (Nakala Press)...
- quelles métadonnées pour les ressources / corpus déposés ? Que décrit-on ?
- quelle place dans un environnement éditorial ?

Outils et chaînes



Outils et chaînes



Les productions actuelles du GT

- Cartographier les outils en mettant en avant :
  - Les compétences requises
  - La licence
  - Le niveau de difficulté pour la prise en main
  - Le niveau de personnalisation
  - Les avantages
  - Les inconvénients

Les productions actuelles du GT

- Créer et éditer son corpus avec TACT, Oxygen, XML Mind XML Editor
- Explorer et analyser son corpus avec eXistdb, BaseX, Heurist
- Collaborer et versionner avec GitLab, GitHub
- Déposer et pérenniser son corpus avec Nakala
- Publier son corpus avec TEITOK, MaX, TEIPublisher, EVT, Omeka, NakalaPress

Cartographies en cours de constitution…

Mise à disposition et diffusion des réflexions et travaux

- gt-tei-nakala@listes.huma-num.fr
- GitLab documenté pour permettre la ré-appropriation les scripts
- Une documentation collaborative en édition continue (GitLab Pages, *à venir*)

Merci de votre attention !

Des questions ?