PROJET CORPUCIT
CORPUS, CITATIONS ET VISUALISATIONS

Cahier des charges d'une plateforme de création et de citation d'extraits de corpus

Driss Sadoun

driss.sadoun@postlab.fr

Plan

Contents

1 Introduction

1.1 Résumé court

CORPUCIT permet de gérer les citations d'extraits de textes ou de corpus en générant un identifiant pérenne pour chaque citation, et de lier finement les écrits scientifiques et leurs données de langage (écrits, sons, vidéo, images) présentées dans leur contexte, facilitant la réflexion scientifique et la réutilisation des données.

Aujourd'hui les corpus sont peu ou pas cités et les extraits de corpus encore moins. Le crédit aux auteurs se fait le plus souvent par l'intermédiaire de la citation d'articles scientifiques en lien avec le corpus. Les citations de corpus dépendent des outils de diffusion des corpus, mais des normes existent allant dans ce sens. Par contre les citations de d'extraits de corpus n'existent pas, et si elles sont réalisées, elles le sont de manière non systématique et en utilisant souvent des liens web non pérennes.

Comment simplifier la création (extraction) d'un extrait à partir d'un document issu d'un corpus ?

À qui et comment donner du crédit aux créateurs, collecteurs, annotateurs ... ?

Se pose les questions de comment donner du crédit à la personne ayant découvert l'extrait et les phénomènes scientifiques qu'il illustre ?

1.2 Résumé long

Les corpus de langage sont des données incontournables des travaux de recherche dans de nombreuses disciplines comme la linguistique, la littérature, l'histoire, la psychologie, l'anthropologie. Dans ces travaux, on est amené à étayer ses démonstrations sur la base de corpus, à présenter des extraits de corpus comme exemples ou éléments de discussion scientifique, ou encore à fonder ses descriptions et ses modèles sur des corpus. Le lien entre les publications scientifiques utilisant les corpus et les corpus eux-mêmes est extrêmement fort, l'ensemble publication/corpus formant souvent une unité indivisible dans la recherche scientifique. Or, il est encore rare que les données de langage formant ces corpus et utilisées pour asseoir les analyses soient partagées et quand elles le sont, qu'elles soient liées aux publications scientifiques et traitées elles-mêmes comme partie intégrante de ces publications. Le projet CORPUCIT a pour but de permettre l'édition de textes scientifiques contenant des citations ou des extraits pointant directement (par un hyperlien) sur des corpus ou des extraits de corpus de langage. Le projet permettra aussi l'édition de corpus pour les structurer en exemples ou en citations afin de leur donner un statut scientifique clair et de les intégrer pleinement à la fois dans le processus scientifique et dans le champ de la science ouverte.

Il s'agit donc :

1. à partir de corpus, de permettre leur édition pour générer des identifiants pérennes (IDP) sur des parties de ces corpus et de construire des exemples

ou des citations extraits du corpus. Les IDP seront basés sur les standards existants de diffusion de données ouvertes. Les outils seront des services Web et seront libres, ce qui leur permettra d'être intégrés dans d'autres sites et disponibles pour d'autres services. Pour les corpus de format connus (TEI), il sera possible de créer des IDP pour des sous-parties de documents. Pour les autres formats, les IDP pointeront sur des documents complets. Au-delà de la création d'un IDP, l'outil permettra d'éditer les extraits ou les citations, et d'y associer toutes les métadonnées et informations complémentaires disponibles en fonction des besoins du chercheur. Pour les formats de corpus connus, il sera possible de visualiser la partie de corpus correspondante.

2. d'utiliser les IDP comme une citation dans les écrits scientifiques ou dans toute présentation de document sur Internet pour pointer sur les corpus et les extraits. Le mécanisme de citation respectera le format standard des citations scientifiques et les citations pourront donc être utilisées par des outils de gestion de citations comme par exemple Zotero. Cela donnera aux corpus un statut beaucoup plus clair de livrable scientifique et permettra aux chercheurs de valoriser la conception, la collecte et le partage de corpus comme une activité scientifique à part entière.

1.3 Contexte

La linguistique est une science qui est de plus en plus souvent basée sur l'utilisation de corpus ou d'exemples issus de données attestées, c'est-à-dire tirées de corpus, de situations expérimentales, ou de tout élément disponible sur Internet ou dans la littérature. Ce mode de fonctionnement existe aussi dans de nombreuses autres disciplines, comme l'histoire ou la littérature par exemple. De la même façon que l'on peut, dans un travail scientifique, citer les propos d'un auteur, en pointant sur la référence publiée, la page et éventuellement la ligne, on est amené à le faire pour un extrait issu d'un corpus. Également, il est souvent nécessaire de faire référence aux données ou aux outils sur lesquels on s'appuie. Dans certaines disciplines, ces références doivent désormais se faire sur un corpus ou une partie de corpus librement accessible. La différence entre la citation d'auteurs et la citation de corpus ou d'extraits de corpus est que les corpus ne sont pas systématiquement organisés ou publiés comme le sont les articles scientifiques. Pointer sur des données est intéressant si les données sont dans un format clair pouvant être traité ou visualisé automatiquement – ce que les travaux sur la science ouverte et le respect des principes FAIR cherchent à réaliser et qui est l'un des enjeux de CORPUCIT. L'objectif du projet CORPUCIT est de donner aux chercheurs et à toute personne utilisant ou citant des textes ou des corpus de langage, écrits, sonores ou visuels, la possibilité de citer facilement et de manière pérenne des extraits ou des parties issus de ces textes ou corpus. Les citations incluses dans des textes scientifiques ou des sites internet mèneront par un simple clic de souris à des visualisations permettant facilement de voir, comprendre, réutiliser, citer les éléments de textes ou de cor-

pus. L'objectif du projet est de créer les outils d'édition permettant de proposer ces fonctionnalités. Les outils ciblent deux éléments principaux :

- les citations, éléments de bibliographie inclus dans des textes scientifiques ou des sites internet;
- les visualisations des extraits ou des parties de corpus. Il s'agit de présenter les extraits de manière visuelle ou sonore sur des sites de dépôts de données scientifiques. Ces présentations pourront être accompagnées de métadonnées et d'informations scientifiques dans le respect des principes FAIR.

1.4 Parties prenantes

CORLI (Corpus, Langue et Interaction), Consortium Huma-Num

Responsables : Christophe Parisse (CR INSERM) et Céline Poudat (MCF Université Côte d'Azur)

Participants :

- Stéphanie Caët (MCF Université de Lille)
- Élisabeth Delais-Roussarie (DR CNRS)
- Sarra El Ayari (Ingénieur CNRS)
- Carole Etienne (Ingénieur CNRS)
- Thomas Gaillat (MCF Université de Rennes)
- Julie Glickman (MCF Université de Strasbourg)

Le consortium CORLI est le porteur du projet. En conséquence, sa tâche principale sera de déterminer lors du lancement du projet les besoins exacts des acteurs de la communauté et des linguistes. Pour cela, des interviews d'utilisateurs potentiels, à la fois membres de CORLI ou non, seront réalisées. Avec l'équipe responsable d'ORTOLANG (UMR CNRS ATILF), le cahier des charges technique sera spécifié. Le consortium sera également chargé d'organiser les séances du comité de pilotage qui gérera la gouvernance du projet. Lors du déroulement, CORLI pourra présenter l'avancement des travaux à la communauté pour obtenir des retours d'expérience et travailler en collaboration avec les laboratoires Prismes et Modyco. Le consortium CORLI, par la variété des matériaux sur lesquels il travaille, pourra orienter les partenaires de manière à couvrir tous les types de corpus de langage. Le consortium sera chargé de créer du matériel pédagogique pour la formation des futurs utilisateurs, en collaboration encore une fois avec Prismes et Modyco. Cette tâche comportera une partie d'étude d'ergonomie et de design.

ATILF (Analyse et Traitement Informatique de la Langue Française), UMR CNRS 7118 et Université de Lorraine

Responsable : Etienne Petitjean (Ingénieur CNRS)

Participants :

- Christophe Benzitoun (MCF Université de Lorraine)
- Cyril Pestel (Ingénieur CNRS)

L'ATILF est le laboratoire porteur principal de l'Equipex ORTOLANG et le laboratoire responsable du suivi technique de cet Equipex. Comme les outils développés dans le cadre du projet CORPUCIT seront d'abord testés et mis en œuvre sur le site de dépôt de données ORTOLANG, l'ATILF sera le laboratoire dans lequel travaillera un ingénieur informaticien chargé du développement des outils, édition et création des extraits, repérage et insertion des IDP, visualisation de corpus, conversion de format de fichiers, mise à niveau des fichiers TEI et autres formats de l'Equipex, création des éléments de bibliographie. Les aspects édition de documents seront réalisés en collaboration avec les laboratoires Prismes et Modyco et le consortium CORLI. Les aspects liens pérennes, formats de données, métadonnées et liens seront conçus et réalisés en collaboration avec la TGIR Huma-Num.

Prismes (Langues, Textes, Arts et Cultures du Monde Anglophone), EA 4398, PARIS III Sorbonne Nouvelle

Responsable : Aliyah Morgenstern

Participants :

- Charlotte Danino (MCF Université Sorbonne Nouvelle)
- Celine Horgues (MCF Université Sorbonne Nouvelle)
- Hélène Josse - De La Gorce (MCF Université Sorbonne Nouvelle)
- Sylwia Scheuer (MCF Université Sorbonne Nouvelle)

Le laboratoire Prismes a élaboré lors des années 2019 et 2020 un prototype de présentation et de diffusion d'extraits issus de corpus de langage oral, VALANGE (voir https://ct3.ortolang.fr/valange/). De nombreux chercheurs de ce laboratoire, et en particulier ceux qui participent au projet, manipulent dans leur travail de recherche des corpus et des extraits de corpus. Prismes apportera son expérience dans ce domaine (manipulation de corpus et retour des utilisateurs) et travaillera à enrichir les exemples d'extraits et d'utilisation des outils développés. PRISMES étant un laboratoire pluridisciplinaire, des chercheurs non-linguistes (en Lettres et Arts) pourront également devenir utilisateurs des outils conçus par le projet CORPUCIT.

Modyco (Modèle, Dynamique, Corpus), UMR CNRS 7114 et Université Paris Nanterre

Responsable : Sophie de Pontonx (Ingénieur CNRS)

Participants :

- Aude Da Cruz Lima (Ingénieur CNRS)

Le laboratoire Modyco a participé, conjointement au laboratoire Prismes, à l'élaboration du prototype VALANGE. Il apportera ses compétences sur ces aspects techniques. Il collaborera avec le laboratoire Prismes à enrichir les exemples d'utilisation et obtenir des retours d'utilisation. Enfin le laboratoire Modyco pourra utiliser certains de ses corpus spécifiques (par exemple corpus mixte de langage et d'expérimentation psycholinguistique) pour travailler sur des extensions du modèle de base du projet.

Huma-Num, TGIR - UMS 3598

Responsable : Michel Jacobson (Ingénieur CNRS)

Participants :

- Nicolas Larrousse (Ingénieur CNRS)

La TGIR apportera son expérience dans la manipulation des identifiants et les liens avec les projets visant à outiller la science ouverte. La TGIR possède également une très bonne compétence dans la gestion et les choix techniques pour les projets informatiques destinés à la communauté des utilisateurs de sciences sociales. La compétence de la TGIR dépasse le cadre de la linguistique, ce qui est important pour que le projet actuel puisse rapidement proposer des applications au-delà des Sciences du Langage.

2 Projet CORPUCIT

2.1 Description

Ce projet a pour but de fournir aux linguistes et à toute personne réalisant un travail scientifique sur la base de corpus langagiers, des outils pour créer et insérer dans un document scientifique des citations d'extraits de corpus.

Définition 1 (*Corpus*)
Dans le cadre de ce projet, un corpus est un ensemble de documents, regroupés dans un but scientifique précis. Il peut servir, en tout ou partie, de base d'un travail de recherche autour du langage. Les documents (ou fichiers) d'un corpus peuvent être multimodaux (écrit, oral, transcription, audio, vidéo etc) et sous différents formats : texte (txt, tei, trs, eaf, ...), audio (mp3, wav ...), vidéo (avi,), image (png, jpg ...) etc.

Définition 2 (*Extrait de corpus*)
Passage ou portion tiré d'un document (écrit, audio, image, vidéo etc) faisant partie d'un corpus langagier (cf. Figure 2.1).

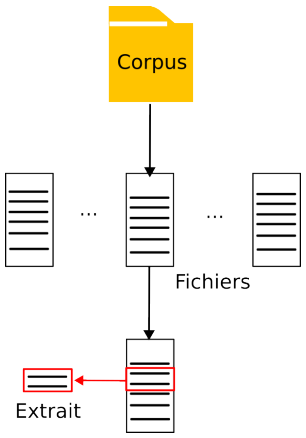

Figure 1: Illustration d'un extrait de corpus

2.2 Objets manipulés

Le projet propose de manipuler trois types d'objets :

- **Partie de corpus** : (ensemble de mots, de phrases, de paragraphes, d'énoncés, chapitre, etc.) tirés de corpus de langage (écrit, audio, vidéo) ;

- **Extrait/Illustration** : comprenant une présentation visuelle des partie de corpus écrits, audio ou vidéo, et accompagnés de toutes informations et métadonnées utiles.
- **Citation** : figurant dans des textes scientifiques et faisant référence à des parties de corpus ou des extraits.

Ces objets seront appréhendés différemment selon qu'on est :

- **Créateur et/ou utilisateur de corpus** : dans ce cas les utilisateurs de corpus feront le chemin qui va des corpus aux extraits puis aux citations dans les textes scientifiques;
- **Lecteur de texte scientifique** : dans ce cas les lecteurs feront le chemin des citations dans les textes scientifiques vers les extraits de corpus et les corpus.

Ci-dessous, les trois objets sont présentés dans l'ordre A-Corpus, B-Extrait, C-citation, mais sont à lire dans l'ordre A-B-C pour les *utilisateurs de corpus* ou l'ordre C-B-A pour les *lecteurs de texte scientifique*.

A – Parties de corpus: Position exacte dans un corpus

On est amené pour des raisons scientifiques à trouver dans un corpus de langage écrit ou oral, des éléments, parties, passages, énoncés, phrases, mots, etc. qui sont particulièrement intéressants. Dans un corpus de langage multimodal, il peut s'agir d'un contour prosodique, d'un geste, une expression faciale, ou d'une séquence de gestes ou de regards de plusieurs participants. Ces éléments ont une position précise (numéro de ligne, numéro de paragraphe, minutage, identifiant xml).
L'information *position exacte* dans un corpus permet à un outil de localiser un extrait automatiquement dans ce corpus. Dans les cas où l'on maîtrise le format des données, alors il est possible de présenter ces données à l'écran et de permettre à l'utilisateur de localiser l'extrait qui l'intéresse. Techniquement pour le projet CORPUCIT, on peut réaliser cette localisation et l'insertion de balises permettant de trouver automatiquement l'extrait pour les fichiers qui sont dans des formats connus, ou dans des formats convertibles automatiquement dans des formats connus. Les formats connus sont la TEI (Text Encoding Initiative, https://tei-c.org/) et les fichiers en texte brut (qui seront traités par une conversion automatique vers un format TEI minimal). On citer en exemple en particulier la représentation de l'écrit (balises div et p) ou celle de l'oral (balises annotationBlock et u).

B – Présentation d'extraits

Dans certains cas, l'accès direct à une partie du document, même si elle permet de balayer le document, de voir ce qui figure autour de l'extrait, n'est pas suffisant car on a besoin d'ajouter des informations. Par ailleurs, il n'est pas toujours possible de visualiser une partie, soit parce que le document lui-même n'est pas pérennisé, soit parce qu'il n'est pas disponible de manière libre sans être anonymisé. Tout un ensemble de raisons peut amener à créer un extrait de manière autonome sans qu'il soit réellement un extrait de corpus (à partir d'élément fournis par le créateur d'extraits donc), et/ou à l'accompagner d'une transcription, d'une image, d'un son, d'un film, ou d'un lien vers un site pérenne.

C – Citations figurant dans un texte scientifique : article, chapitre ou page internet

Les citations dans un article, un ouvrage, un site Internet peuvent apparaître sous deux formes :

1. citation bibliographique, qui contient un lien hypertexte vers l'extrait ou l'élément de corpus
2. un exemple visuel (texte et/ou image) associé à une citation bibliographique et qui contient un lien hypertexte vers une présentation d'extrait.

Quelque soit la forme d'une citation, on comprend bien qu'elle doit permettre en un clic de souris d'accèder aux informations complètes sur l'extrait et grâce à un deuxième clic de souris de passer de cet extrait au corpus original.

2.3 Périmètre

Les développements prévus s'inscrivent dans le cadre d'un des trois projets du réseau CORLI (https://corli.huma-num.fr/wp-content/uploads/2022/05/corpucit-lancement.pdf). Ils s'adressent en particulier aux personnes menant des activités de recherche autour de corpus linguistiques. Néanmoins, L'objectif est de développer des outils suffisamment ouverts pour être utilisées par d'autres communautés scientifiques.

2.4 Contraintes

- Les développements doivent être réalisés en respectant les standards et les bonnes pratiques définis pour la citation de données [**?**].
- L'ensemble des développements et des livrables devront s'inscrire dans une démarche de sciences ouvertes (codes et données ouvertes).
- Les corpus et extraits de corpus devront être stockés et identifiés de manière pérenne grâce à des identifiants pérennes (ex DOI).
- Se servir autant que possible des outils existants.

- Viser à être complémentaire et interopérable avec d'autres plateformes telles que Cocoon, Nakala ou Ortolang.

2.5 Besoins

Nous pouvons distinguer trois niveaux de besoins :

2.5.1 Création d'extraits

Un extrait de corpus peut avoir différents usages : objet d'étude (phénomène particulier), illustration de cours, présentation, publication scientifique, validation d'hypothèse scientifique, reproduction scientifique. La création d'un extrait peut se faire de plusieurs manières :

- copier-coller de l'extrait dans un champs de formulaire.
- téléverser l'extrait à partir d'un fichier ou d'un lien web.
- créer l'extrait à partir de la sélection d'une portion (ou fragment) d'un fichier (texte, image, audio ...).
  - sélection manuelle d'une partie de fichier.
  - sélection automatique via des filtres/requêtes dans une collection de documents. Filtres sur des balises XML ou sur des attributs de balises, pour alléger l'extrait ou mettre en relief un phénomène particulier. Par exemple, éliminer les pauses ou les longs silences dans un extrait audio.

2.5.2 Accès, visualisation, manipulation et utilisation d'extraits

Une fois l'extrait créé, il doit pouvoir être utilisé par la communauté scientifiques. Voici ci-dessous, les besoins exprimés par les membres de la communauté CORLI concernant l'usage des extraits de corpus.

- avoir un accès simple vers l'extrait et ses métadonnées (humains et machines).
- accéder à une page web décrivant l'extrait.
- avoir un accès aux différentes versions (visualisation et téléchargement).
- avoir accès aux différents formats (visualisation et téléchargement).
- avoir accès à plus ou moins de contexte. Par exemple, 5 lignes avant et après pour du texte ou 20s avant et après pour de l'audio.
- pouvoir maintenir le lien entre l'extrait et son origine (corpus et fichier).
- avoir un référencement des extraits dans une base de données.

- pouvoir lister et parcourir les extraits.
- rechercher des extraits en fonction de filtres : base d'extraits thématisés pour que les gens les trouvent : dans laquelle on peut rechercher des extraits via des filtres. (ex : interactionnel + 'papa et enfant qui lisent ensemble').
- avoir des métriques sur les extraits. Par exemple, pour les auteurs de corpus ou d'extrait pouvoir faire une recherche inverse, i.e. savoir où et par qui est cité son corpus ou son extrait.
- pouvoir naviguer au sein d'un extrait.
- pouvoir manipuler un extrait (modifier, recadrer, redimensionner, ...).
- avoir des explications simples concernant les droits et restrictions associés à la licence.
- avoir une transparence sur le statut des données, à savoir le type (ou la source) des données, où elles se trouvent, et comment on peut y avoir accès.
- pouvoir générer un lien de partage de l'extrait.

2.5.3 Citations des extraits

Les citations doivent permettre d'identifier les personnes ayant contribué à la constitution des corpus, et de quelle manière celles-ci ont contribué. Elles doivent également permettre de faire le lien entre documents scientifiques et extraits de corpus. Enfin, il doit être possible d'identifier le créateur de l'extrait, qui peut ête différent de l'auteur du corpus ou de l'auteur de la publication utilisant les extraits. Ci-dessous, les besoins exprimés par les membres de la communauté CORLI interrogés pour la citation d'extraits.

- avoir un lien vers l'extrait à insérer dans les documents scientifiques (articles, présentations, cours ...).
- avoir une citation bibliographique à insérer dans mes documents scientifiques (articles, présentations, cours ...).
- avoir un exemple/illustration à insérer dans les documents scientifiques (articles, présentations, cours ...).
- pouvoir insérer une image contenant l'extrait i.e. l'extrait sous forme d'image. Par exemple, dans le cas de limite de caractères dans une publication scientifique.
- crédit aux auteurs : lister les contributeurs et leur(s) rôle(s) dans la citation.

  - au sein d'une citation dans un document scientifique (collaborateurs avec un apport scientifique).
  - au sein d'une page web décrivant de l'extrait (l'ensemble des collaborateurs ou intervenant dans la création, la récolte).

- crédit à la personne ayant identifié l'extrait (phénomène particulier)
- traçabilité des personnes à créditer (cycle de vie de l'extrait).

3 Développements

Les développements devront se faire de manière incrémentale, suivant une démarche centrée utilisateur, permettant à chaque étape de développement d'impliquer et de mobiliser un cercle d'utilisateurs plus élargi.

Phases de développements

Les développements se feront en 3 phases décrites ci-dessous :

- phase 1 : représentant une preuve de concept (PoC).
- phase 2 : représentant un prototype alpha (de test), avec une visualisation des fonctionnalités pas encore développées.
- phase 3 : représentant un outil fonctionnel et complet.

Profils (rôles) utilisateurs

Ces profils permettent de définir les droits et les restrictions pour chaque catégorie d'utilisateurs dans une base d'extraits de corpus. Les rôles décrits ci-dessous, seront créés en fonction des besoins de chaque phase de développement.

- *Super Administrateur* : possède tous les droits d'administration, dont la création et la suppression d'*administrateur*.
- *Administrateur* : possède les droits d'administration dont la création, l'édition ou la suppression de compte. Il peut créer, modifier, publier et supprimer les extraits de corpus et leurs métadonnées.
- *Contributeur* : peut créer, modifier, publier et supprimer ses propres extraits de corpus. Il peut également enregistrer ses extraits favoris pour les retrouver facilement.
- *Modérateurs* : contrôle la qualité et la validité des extraits déposés. Il vérifie qu'il ne manque pas d'information et peut éditer les extraits si nécessaire.
- *Utilisateur* : navigue sur le site et visualise les extraits de corpus sans être connecté.
- *Lecteur* : lit les documents scientifiques pouvant contenir les exemples, liens et citations liés aux extraits.

3.1 Phase 1 : Preuve de concept

La preuve de concept (PoC) a pour objectif de donner un aperçu des possibilités de visualisation des extraits de corpus. Il s'agit de construire un site web illustrant des exemples d'extraits construit à la main.

3.1.1 Scénarios

Pour les besoins du PoC, seuls les profils *Administrateur* et *Utilisateur* seront nécessaires.

Scénario 1 (Administrateur)**.** *L'administrateur doit pouvoir créer, visualiser, éditer ou supprimer un extrait de corpus.*

Scénario 2 (Utilisateur)**.** *L'utilisateur doit pouvoir parcourir l'ensemble des extrait de corpus (cf. Figure 2). Il peut sélectionner un extrait pour aller sur sa page de description (cf. Figure 3).*

Scénario 3 (Utilisateur)**.** *L'utilisateur doit pouvoir copier/télécharger le lien vers l'extrait et sa citation (bibtex).*

Scénario 4 (Lecteur)**.** *Un lecteur doit pouvoir utiliser un lien vers l'extrait pouvant être inséré dans un document scientifique.*

3.1.2 Fonctionnalités

Ce PoC doit permettre de :

- renseigner un extrait (plusieurs types texte image ...) et ses métadonnées via un formulaire.
- créer une citation pour l'extrait via les informations du formulaire.
- éditer un extrait et ses métadonnées via un formulaire.
- créer un identifiant pérenne associé à chaque extrait (ex DOI créé via Nakala).
- avoir des liens vers le corpus et le fichier d'origine.
- sauvegarder l'extrait dans la base de données.
- afficher la liste des extraits.
- visualiser un extrait et ses métadonnées associées (ex comment citer).
- supprimer un extrait.

3.1.3 Développements

Afin d'atteindre les objectifs du PoC, un site web devra être développé et mis en ligne. Ci-dessous, les grandes lignes des développements attendus.

Base de données Les exemples d'extraits de corpus devront être contenus dans une base de données.

Gestion des extraits La mise à jour des extraits, se fera via des formulaires web permettant de renseigner (créer), modifier et supprimer un extrait. Seul un

administrateur aura accès à la gestion des extraits

Page d'accueil L'ensemble des informations décrivant le projet CORPUCIT et ses objectifs devra être accessible au sein de la page d'accueil du site.

Index des extraits Une page listant l'ensemble des extraits de corpus contenus dans la base de données devra être accessible simplement. À partir de cette liste, les utilisateurs pourront sélectionner un extrait.

Page d'un extrait La sélection d'un extrait devra diriger l'utilisateur vers une page contenant l'extrait, sa description et ses métadonnées.

Lien pérenne La création d'un extrait devra s'accompagner de la création d'un lien pérenne (DOI). Ce lien pérenne sera créé sur l'entrepôt de données Nakala. Il pourra être copié ou téléchargé par les utilisateurs.

Citation d'un extrait Les utilisateurs devront pouvoir copier ou télécharger une représentation (soit au format bibtex pour être utilisée par les outils de gestion de bibliographie, soit au format texte pour ête collé dans un document) d'un extrait à partir de sa page de description. Cette représentation aura été renseigné à la création ou à l'édition d'un extrait.

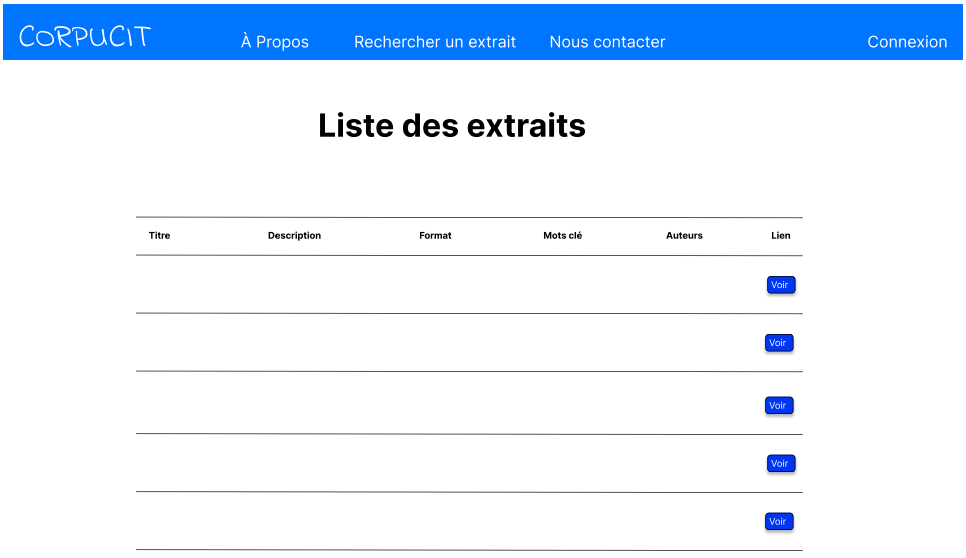

Figure 2: Page listant les extraits de corpus

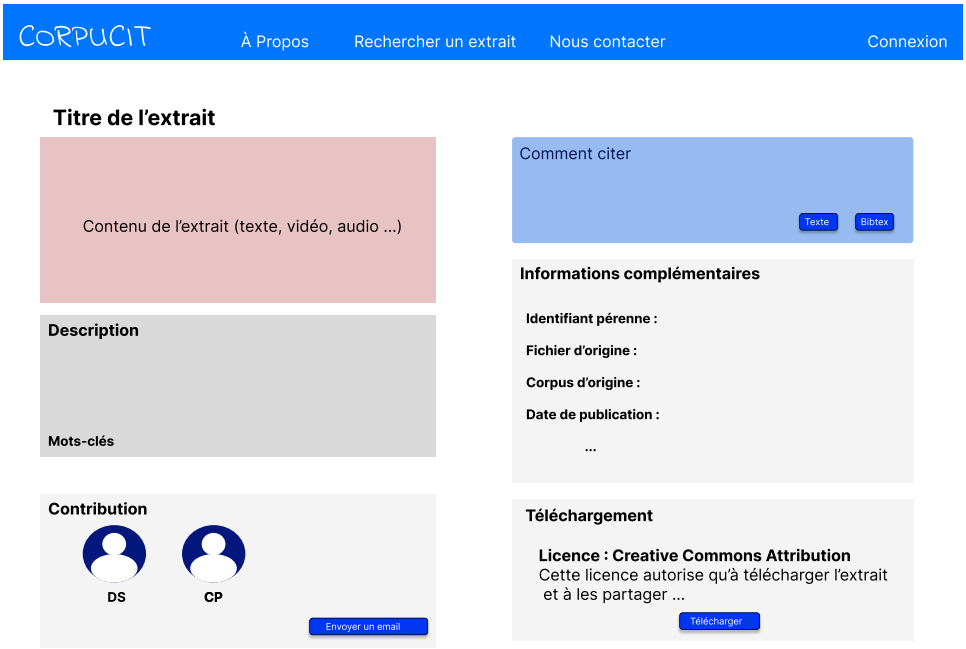

Figure 3: Page de description d'un extrait de corpus

3.2 Prototype alpha

Ce Prototype alpha doit apporter deux nouvelles fonctionnalités par rapport au PoC :

1. ouvrir l'application à des contributeurs (autres que les administrateurs) pouvant manipuler leurs propres extraits en y associant les métadonnées qui leurs semblent les plus pertinentes.
2. permettre une recherche des extraits en fonction de différents critères.
3. permettre d'exporter une citation sous un format aisément compatible avec un outil de bibliographie comme Zotero.

3.2.1 Scénarios

Afin de permettre la gestion des contributeurs, l'application doit disposer d'un système de gestion des comptes : s'inscrire, se connecter et gérer son profil. Pour les besoins de ce prototype, le profil *Contributeur* viendra s'ajouter aux profils *Administrateur* et *Utilisateur*.

Scénario 5 (Contributeur)**.** *Un contributeur doit pouvoir créer, voir, éditer, et supprimer ses propres extraits.*

Scénario 6 (Contributeur)**.** *Un contributeur peut choisir le niveau de visibilité (public [tous, recherche, groupe] ou privé) de ses extraits pour les autres contributeurs et utilisateurs. Un contributeur ne peut pas voir l'extrait d'un autre contributeur s'il est privé.*

Scénario 7 (Utilisateur)**.** *Un utilisateur ne peut voir que les extraits publics.*

Scénario 8 (Utilisateur)**.** *Un utilisateur doit pouvoir créer un compte sur le site. Une fois connecté il a le rôle de Contributeur.*

Scénario 9 (Administrateur)**.** *Un administrateur peut créer, éditer et supprimer un Contributeur. Il peut visualiser, éditer ou supprimer tous les extraits publics ou privés.*

3.2.2 Fonctionnalités

En plus des fonctionnalités offertes par le PoC, le prototype doit permettre de :

1. se créer un compte.
2. se connecter à son compte.
3. éditer son profil.
4. visualiser les profils d'autres contributeurs (s'ils sont publics).
5. créer ses propres extraits.
6. associer des tags ou des catégories à ses extraits.
7. lister ses propres extraits.
8. éditer ses extraits.
9. supprimer ses extraits.
10. rechercher des extraits via une barre de recherche.
11. rechercher des extraits via des filtres.

3.2.3 Développements

Panel d'administration des utilisateurs Afin de permettre la gestion de comptes et des rôles d'utilisateurs, un panel d'administration des utilisateurs devra être mis en place. Ce panel permettra de créer, modifier, visualiser ou supprimer un utilisateur. Il permettra également de gérer le rôle d'un utilisateur.

Formulaires d'enregistrement et de connexion Un formulaire d'enregistrement et un formulaire de connexion devront respectivement permettre à un utilisateur de créer un compte et de s'y connecter. Les informations des utilisateurs telles que l'identifiant, le mail et le mot de passes seront stockées en base de données. Si besoin, l'utilisateur devra pouvoir réinitialiser son mot de passe. L'utilisateur

pourra également supprimer son compte et ses données.

Profil Une page profil permettra à l'utilisateur d'éditer ses informations. Il pourra notamment spécifier s'il souhaite que son profil soit public ou privé.

Droits d'édition des extraits Tous les extraits pourront être visualisés par n'importe quel utilisateur mais seul le créateur d'un extrait (ou un administrateur) pourra le modifier ou le supprimer. La gestion des droits d'édition nécessitera donc de distinguer les extraits en fonction de leur créateur.

Filtres et recherche d'extraits En plus, de pouvoir lister l'ensemble des extraits contenus dans la base de données, il faudra permettre aux utilisateurs de filtrer les extraits. Par exemple, par auteur (créateur), par type (texte, image, audio ...) ou par format (txt, wav, tei ...). Une barre de recherche sera également mise à disposition des utilisateurs. Elle permettra d'afficher les extraits correspondant au contenu de la requête rédigée par l'utilisateur.

3.3 Outil fonctionnel et complet

Cette dernière version doit favoriser l'utilisation et la citation des extraits en simplifiant leur intégration au sein de documents scientifiques. Elle facilitera la collaborations scientifiques entre contributeurs, en leur fournissant la possibilité de partager extraits ou collections d'extraits.

Ses contours devront s'affiner grâce aux retours utilisateurs issus du test du prototype alpha. Ces retours permettrons notamment de statuer sur la nécessité de mise en place pour les utilisateurs d'outils de manipulation d'extraits, par exemple redimensionner, changer l'orientation ou la luminosité pour les images, étendre avec plus ou moins de contexte pour le texte (ex 5 lignes avant et après) ou l'audio (ex 20s avant et après).

3.3.1 Scénarios

Le site permettra aux contributeurs de partager leurs extraits de corpus avec d'autres contributeurs. Les utilisateurs doivent pouvoir filtrer et trier les résultats en fonction de différents critères tels que les contributeurs ou groupes de contributeurs, les équipes ou laboratoires etc.

Le profil *Modérateur* viendra s'ajouter aux profils *Administrateur*, *Contributeur* et *Utilisateur*.

Scénario 10 (Contributeur)**.** *Un contributeur doit pouvoir partager ses extraits ou collections d'extraits avec d'autres contributeurs. Soit en les invitants, soit en acceptant leur demande de d'accès.*

Scénario 11 (Contributeur)**.** *Un contributeur doit pouvoir afficher la liste de ses extraits partagés avec d'autres contributeurs.*

Scénario 12 (Contributeur)**.** *Un contributeur peut visualiser tous les extraits de son groupe, même s'ils sont privés.*

Scénario 13 (Contributeur)**.** *Un contributeur peut télécharger ses extraits, ainsi que les extraits publics d'autres contributeurs.*

Scénario 14 (Utilisateur)**.** *Un utilisateur peut visualiser uniquement les extraits et collections d'extraits publics.*

Scénario 15 (Utilisateur)**.** *Un utilisateur peut avoir accès à des métriques concernant l'utilisation et la citation de l'extrait.*

Scénario 16 (Modérateur)**.** *Un modérateur est garant de la qualité et de la validité des extraits créés. Il doit pouvoir visualiser les extraits en mode révision. Ce mode lui permet d'éditer, de valider, de rejeter ou de supprimer un extrait.*

Scénario 17 (Modérateur)**.** *Un modérateur pourra avoir accès soit à la révision de tout extrait existant ou à la révision d'un sous-ensemble d'extraits (ex de son labo ou de son champs de discipline). Cela dépendra d'une hiérarchie de droit de révision à définir.*

3.3.2 fonctionnalités

- enregistrer ses extraits favoris.
- créer des collections d'extraits.
- partager un extrait ou une collections d'extraits avec d'autres contributeurs.
- regrouper des extraits au sein d'un exemplier. Lien vers l'exemplier avec un chemin vers chaque extrait.
- afficher les extraits similaires à chaque extrait visualisé.
- télécharger l'extrait, ses métadonnées, sa citation, ses différents formats.
- présenter des métriques (quantitatives et qualitative ex pour quelle usage) d'utilisation et de citation de chaque extrait.

3.3.3 Développements

Panel d'administration Le panel d'administration devra être étendu pour gérer les notions de modérateurs, de groupes de contributeurs et de collections d'extraits.

Modération des extraits Chaque modérateur devra avoir accès à une page listant l'ensemble des extraits qu'il peut réviser (modérer). En sélectionnant un extrait, il devra être redirigé vers un formulaire lui permettant d'éditer, de valider, de rejeter ou de supprimer l'extrait.

Collection d'extraits En plus de pouvoir gérer leurs extraits, les contributeurs devront également pouvoir gérer leurs collections d'extraits. Cela suppose de mettre à leur disposition la possibilité de créer, visualiser, modifier ou supprimer

une collection et de lister l'ensemble de leurs collections.

Groupes de contributeurs Un contributeur devra pouvoir lister l'ensemble des groupes auxquels il contribue. Il pourra accéder à des formulaires pour créer, visualiser, modifier et supprimer un groupes de contributeurs. Un mécanisme de gestion des d'invitation ou de demande d'accès à un groupe devra également être implémenté. chaque contributeur.

Métriques d'utilisation La page de chaque extrait devra également présenter des métriques d'utilisation et de citation. Il pourrait être envisagé de mettre en ligne une page de métriques globales à l'ensemble des extraits.