Projet CITATION : CORPUCIT
Fournir des outils pour faciliter la citation de corpus ou d'extraits de corpus

Atelier "Citation des données : enjeux, pratiques, outils"
Université Paris Nanterre le vendredi 9 décembre 2022

Christophe Parisse (MoDyCo), **Driss Sadoun** (PostLab)

driss.sadoun@postlab.fr

Contexte et objectifs

Consortium HN Corpus, Langues et Interactions (CORLI)

Un réseau de laboratoires et de chercheurs travaillant sur les corpus de langue.

Objectif : Proposer des outils, de la documentation et des formations autour de l'utilisation scientifique des corpus de langue, en suivant les principes FAIR (Faciles à trouver, Accessibles, Interopérables et Réutilisables).

CORLI travaille activement sur trois projets :

- L'annotation collaborative
- Le Corpus Ouvert du Français: données de corpus et outils pour la langue française
- **La citation de corpus ou d'extraits de corpus**

Principes de citation des données scientifiques

- Depuis 2014, une déclaration commune des principes de citation des données (JDDCP) [Gro14] proposée plusieurs organisations
- Un ensemble de lignes directrices et des recommandations pour la citation des données scientifiques
  - Déposer ses données dans des archives pérennes
  - Utiliser des formats standards et répandus
  - Utiliser un identifiant pérenne qui dirige vers une page d'accueil
  - Utiliser des structures standards pour les citations

Archivages pérennes et Formats de citation

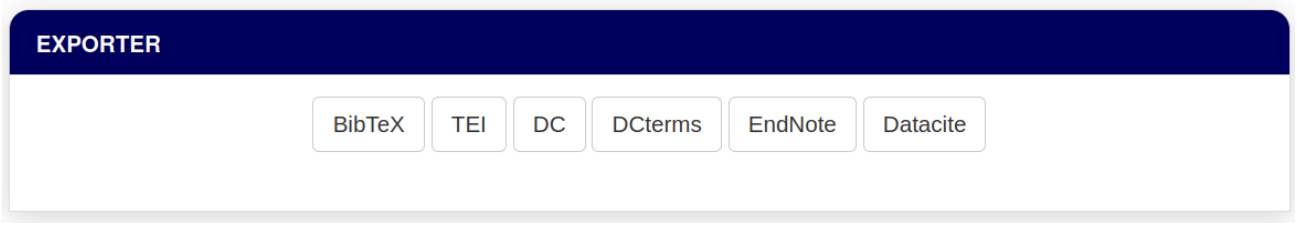

Figure: Formats d'export de citation sous HAL

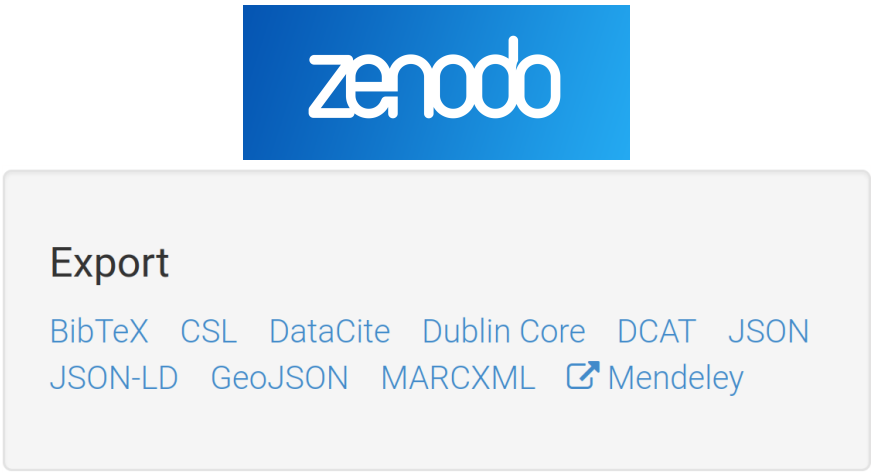

Figure: Formats d'export de citation sous Zenodo

Identifiant pérenne et page d'accueil

Un identifiant pérénne doit renvoyer vers une page d'accueil plutôt que directement vers les données [CKG+18].

- Les données ne sont pas toujours légalement accessibles
- Les données peuvent se présenter sous de multiples formats ou versions
- Les métadonnées de la page d'accueil peuvent être complémentaires à celles de la référence

  $\Rightarrow$ Les métadonnées doivent être lisible par les humains et les machines

Structure d'une citation de données

Préconisation pour la citation des données (corpus) numérisés.

1. Auteur(s)
2. Titre
3. Date de publication
4. Editeur (ou archive où elles sont hébergées)
5. Identifiant pérenne (IDP) : *DOI* ou *HANDLE*

6. Type de document (optionnel)
7. Édition, volume ou version (optionnel)
8. Date du dernier accès (optionnel)

Structure d'une citation sur COCOON

Auteur(s) | Date | Titre | Version | Editeur | IDP

Danto Anatole | 2017 | Collection ApoliMer (Anthropologie Politique de la Mer) | Version 1 | Anthropologie politique de la Mer | https://doi.org/10.34847/COCOON.C7C3F56D-8CF4-3176-BD16-1DA83A8AFCF4

Danto, Anatole. (2017). Collection ApoliMer (Anthropologie Politique de la Mer) (Version 1). Anthropologie politique de la Mer. https://doi.org/10.34847/COCOON.C7C3F56D-8CF4-3176-BD16-1DA83A8AFCF4

Anne Zribi-Hertz, Elena Soare, Sarra El Ayari | 2016 | Langues et Grammaires en (Ile-de-)France (LGIDF) | Version 1 | Structures formelles du langage | https://doi.org/10.34847/COCOON.7A6A91B9-66ED-3099-BBED-FBB70361C226

Zribi-Hertz, Anne, Soare, Elena, & El Ayari, Sarra. (2016). Langues et Grammaires en (Ile-de-)France (LGIDF) (Version 1). Structures formelles du langage. hhttps://doi.org/10.34847/COCOON.7A6A91B9-66ED-3099-BBED-FBB70361C226

Structure d'une citation sur NAKALA

Auteur(s) | Date | Titre | Type | Editeur | IDP

Pierre Fasula | 2021 | Séminaire Philosophie et psychanalyse 21-05-29 | Sound | NAKALA | https://doi.org/10.34847/nkl.be9clr95

Fasula, Pierre (2021) "Séminaire Philosophie et psychanalyse 21-05-29 - Pierre-Henri Castel" [Sound] NAKALA. https://doi.org/10.34847/nkl.be9clr95

Heba Al Sakhel Amlou | 2019 | Hache 1 - 1 | Image | NAKALA | https://doi.org/10.34847/nkl.ca935q8w

Al Sakhel Amlou, Heba (2019) "Hache 1 - 1" [Image] NAKALA. https://doi.org/10.34847/nkl.ca935q8w

Structure d'une citation sur ORTOLANG

Auteur(s) | Titre | Editeur | Volume | Date | Page | IDP

Aliyah Morgenstern et Christophe Parisse | The Paris Corpus

Journal of French Language Studies | Volume 22 / Special Issue 01 | March 2012 | 7-12

https://hdl.handle.net/11403/colaje/v2.4

Aliyah Morgenstern and Christophe Parisse : The Paris Corpus, Journal of French Language Studies / Volume 22 / Special Issue 01 / March 2012, pp 7 - 12, https://hdl.handle.net/11403/colaje/v2.4

Projet CORPUCIT

Ambition du projet CORPUCIT

- Proposer des recommandations et des outils pour la création et la citation d'extraits de corpus langagiers.

Enjeux

- Lever les freins à la réutilisation et à l'émergence de nouveaux usages autour des données scientifiques
- Offrir aux corpus un statut plus clair de livrable scientifique
- Permettre aux chercheurs de mieux valoriser la conception, la collecte et le partage de corpus comme une activité scientifique à part entière.
- Favoriser la dynamique de sciences ouvertes et de respect des principes FAIR (Faciles à trouver, Accessibles, Interopérables et Réutilisables)

Objectif du projet CORPUCIT

Plateforme permettant de lier finement écrits scientifiques et extraits de données de langage (écrits, sons, vidéos, images), présentées dans leur contexte, facilitant la réflexion scientifique et la réutilisation des données.

- Éditer des portions de corpus pour générer des extraits pouvant être insérer dans des écrits scientifiques.
- Générer des identifiants pérennes et des citations pouvant être insérer dans des écrits scientifiques.
- Visualiser, contextualiser et manipuler des extraits de corpus.

Parties prenantes

PRISMES - Langues, Textes, Arts et Cultures du Monde Anglophone - EA 4398

Modèles, Dynamiques, Corpus
UMR 7114

Extraits de corpus : définition et exemples

Qu'est qu'un extrait de corpus

Définition d'un extrait dans le contexte de CORPUCIT

Passage ou portion tiré d'un document (écrit, audio, image, vidéo etc) faisant partie d'un corpus langagier.

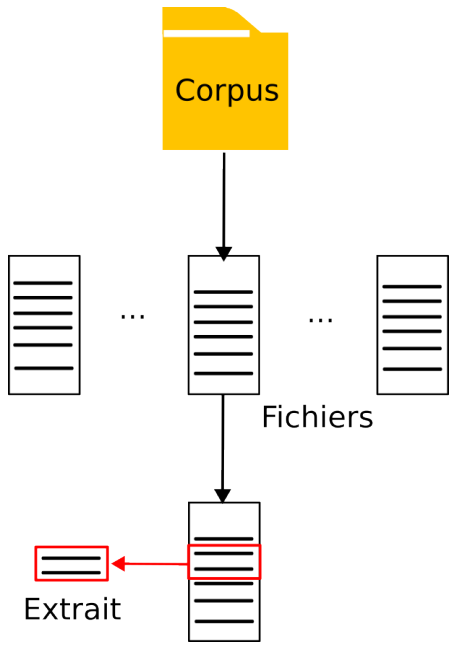

Figure: Illustration d'un extrait de corpus

Extraits issus d'un livre papier

N°1 : ***Boule de Suif, Guy de Maupassant***

Maupassant, dans ses contes et nouvelles, exerce son humour caricatural sur les femmes, particulièrement les vieilles-filles et les anglaises, et surtout sur les bourgeois. Il peint de préférence les petits bourgeois de province auxquels il reproche de trop bien vivre et de trop bien manger. Pour ce faire, il utilise la métaphore de l'homme-boule, métaphore utilisée dans *Boule de suif* à la fois pour la peinture des bourgeois que celle de la prostituée, ainsi que dans la peinture des prostituées de *La Maison Tellier*. D'autre part, son humour n'épargne pas les représentants de l'Eglise.

Fiche de lecture : L'humour dans les Contes et nouvelles

Extrait : Boule de Suif, p. 6, p.7 et p.7-8

Loiseau *: « De taille exiguë, il présentait un ventre en ballon surmonté d'une face rougeaude entre deux favoris grisonnants. »*

Les deux religieuses *: « L'une était vieille avec une face défoncée par la petite vérole comme si elle eût reçu à bout portant une bordée de mitraille en pleine figure. L'autre, très chétive, avait une tête jolie et maladive sur une poitrine de phtisique rongée par cette foi dévorante qui fait les martyrs et les illuminés. »*

Boule de Suif *: « Petite, ronde de partout, grasse à lard, avec des doigts bouffis, étranglés aux phalanges, pareils à des chapelets de courtes saucisses ; avec une peau luisante et tendue, une gorge énorme qui saillait sous sa robe, elle restait cependant appétissante et courue, tant sa fraîcheur faisait plaisir à voir. Sa figure était une pomme rouge, un bouton de pivoine prêt à fleurir ; et là-dedans s'ouvraient, en haut, deux yeux noirs magnifiques, ombragés de grands cils épais qui mettaient une ombre dedans ; en bas, une bouche charmante, étroite, humide pour le baiser, meublée de quenottes luisantes et microscopiques. »*

Figure: L'humour dans les Contes et nouvelles (Casden)

Extraits issus d'un corpus journalistique numérisé

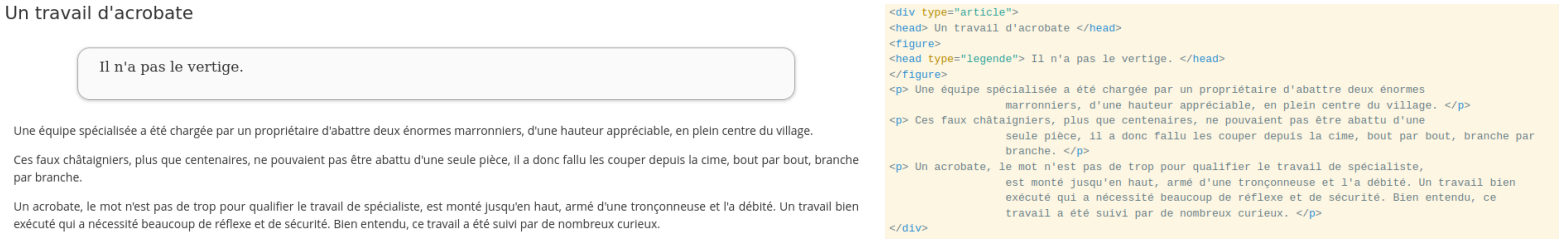

Figure: Extraits tirés du corpus de l'Est Républicain de ORTOLANG

Extraits issus d'un corpus de transcriptions audios

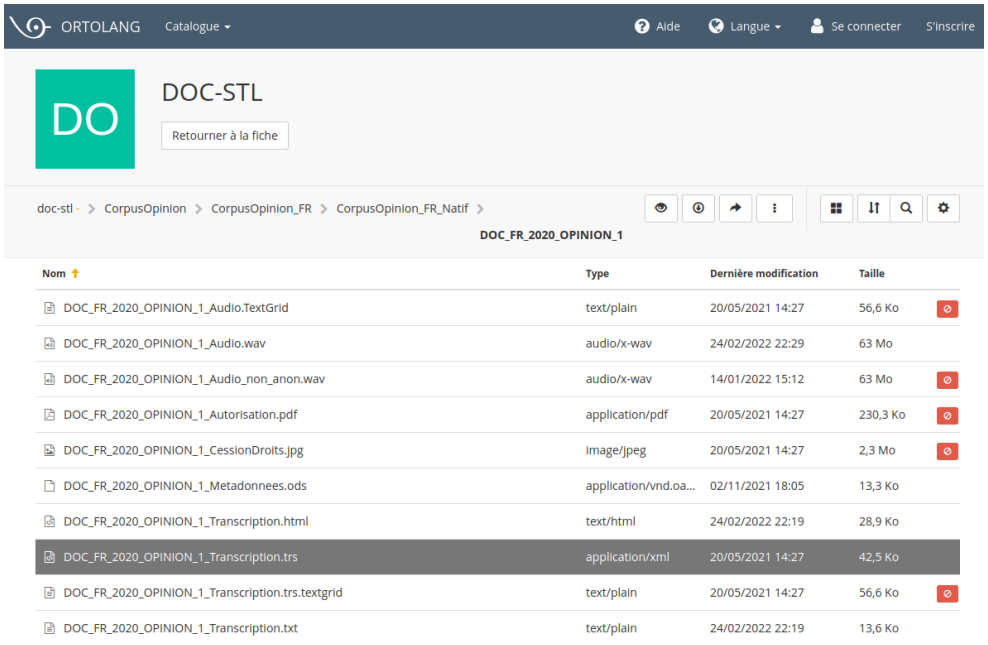

Figure: Extrait au format Transcriber tiré du corpus DOC-STL de ORTOLANG

Extraits issus d'un corpus de transcription vidéos

Figure: Extrait vidéo et sa transcription héberger sur VaLangE

Extraits issus d'un corpus d'images

Figure: UCD Image Cropper (University College Dublin)

Citation d'extraits de corpus

Pourquoi citer un extrait de corpus

- Question d'intégrité scientifique
- Crédit aux personnes ayant contribué à la constitution du corpus
- Donner au lecteur un accès direct à l'extrait
- Situer l'extrait dans le document ou le corpus original
- Aider à la validation scientifique
- Simplifier la visualisation, la manipulation et la réutilisation de l'extrait

Éléments d'une citation d'extrait de corpus

- Mention de l'extrait dans l'écrit scientifique : exemple ou illustration
- Citation de l'extrait : référence bibliographique

- Identifiant pérenne du document ou corpus d'origine (IDPC)
- Identifiant pérenne de l'extrait (IDPE)
- URL sur laquelle pointe l'IDPE

Différences entre un IDP et une URL

IDP

- Identifie de manière unique et pérenne un document numérique
- Est invariant
- Pointe vers une URL

URL

- Correspond à l'adresse d'un document numérique
- Peut évoluer dans le temps (ex évolution de la structure du corpus)
- Peut être éditée de manière dynamique
- Peut avoir des paramètres

Exemple d'un IDP qui pointe vers une URL :
https://doi.org/10.34847/cocoon.ce8a4a03-5083-3144-88ce-7cc4606e8352

Outils de création d'extraits de corpus

1. Visualiser le document d'origine
2. Sélectionner l'extrait à partir de paramètre(s) délimiteur(s)
3. Visualiser la sélection
4. Enregistrer et exporter la sélection
5. Créer un identifiant pérenne pour l'extrait
6. Créer une citation (référence bibliographique)

Exemples de sélection d'extraits d'images

Sélection via un outil en ligne

Sélection via les paramètres d'un URL

Conclusion

Proposer des bonnes pratiques et développer des outils pour la citation d'extraits de corpus.

Afin de favoriser la longévité et la réutilisation de ces outils :

- Choisir des formats et structures de citations standards
- Travailler avec des données ouvertes
- Utiliser et produire du code ouvert

Feuille de route :

- Lancement du projet
- Développement incrémental d'outils et bonnes pratiques
- Construire ces outils pour et avec la communauté scientifique

Discussion

References

- Helena Cousijn, Amye Kenall, Emma Ganley, Melissa Harrison, David Kernohan, Fiona Murphy, Patrick Polischuk, Maryann Martone, and Tim Clark, A data citation roadmap for scientific publishers, bioRXiv **5** (2018).
- Martin Fenner, Mercè Crosas, Jeffrey S. Grethe, David Kennedy, Henning Hermjakob, Phillippe Rocca-Serra, Gustavo Durand, Robin Berjon, Sebastian Karcher, Maryann Martone, and Tim Clark, A data citation roadmap for scholarly data repositories, Scientific Data **6** (2019), no. 28.
- Data Citation Synthesis Group, Joint declaration of data citation principles, 2014.