NER4Archives (named entity recognition for archives) : méthodes et outils semi-automatiques pour reconnaître les entités nommées dans les instruments de recherche archivistiques encodés en XML/EAD

Pauline Charbonnier, Lucas Terriel, Florence Clavaud, Laurent Romary, Gaetano Piraino, Vincent Verdese

NER4Archives (named entity recognition for archives) : *méthodes et outils semi-automatiques pour reconnaître les entités nommées dans les instruments de recherche archivistiques encodés en XML/EAD*

Les Futurs Fantastiques - 3e Conférence Internationale sur l'Intelligence Artificielle appliquée aux Bibliothèques, Archives et Musées
#FF21

Bibliothèque nationale de France - 9 décembre 2021

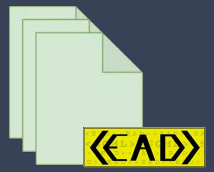
Pauline Charbonnier
ingénieure d'études au Lab des Archives nationales
pauline.charbonnier@culture.gouv.fr
@PaulineCharbo

Lucas Terriel
ingénieur R&D à ALMAnaCH (Inria)
lucas.terriel@inria.fr
@Lucaterre

Membres du projet : **Florence Clavaud** (Conservatrice générale du patrimoine et responsable du Lab des Archives nationales), **Laurent Romary** (Directeur de recherche à ALMAnaCH-Inria), **Gaetano Piraino** et **Vincent Verdese** (AN DSI)

Contexte NER4Archives

Types de métadonnées dans la salle des inventaires virtuelle (SIV) :



≈ 29 000 inventaires

≈ 15 200 notices producteurs (collectivités, personnes, familles) / en cours d'enrichissement

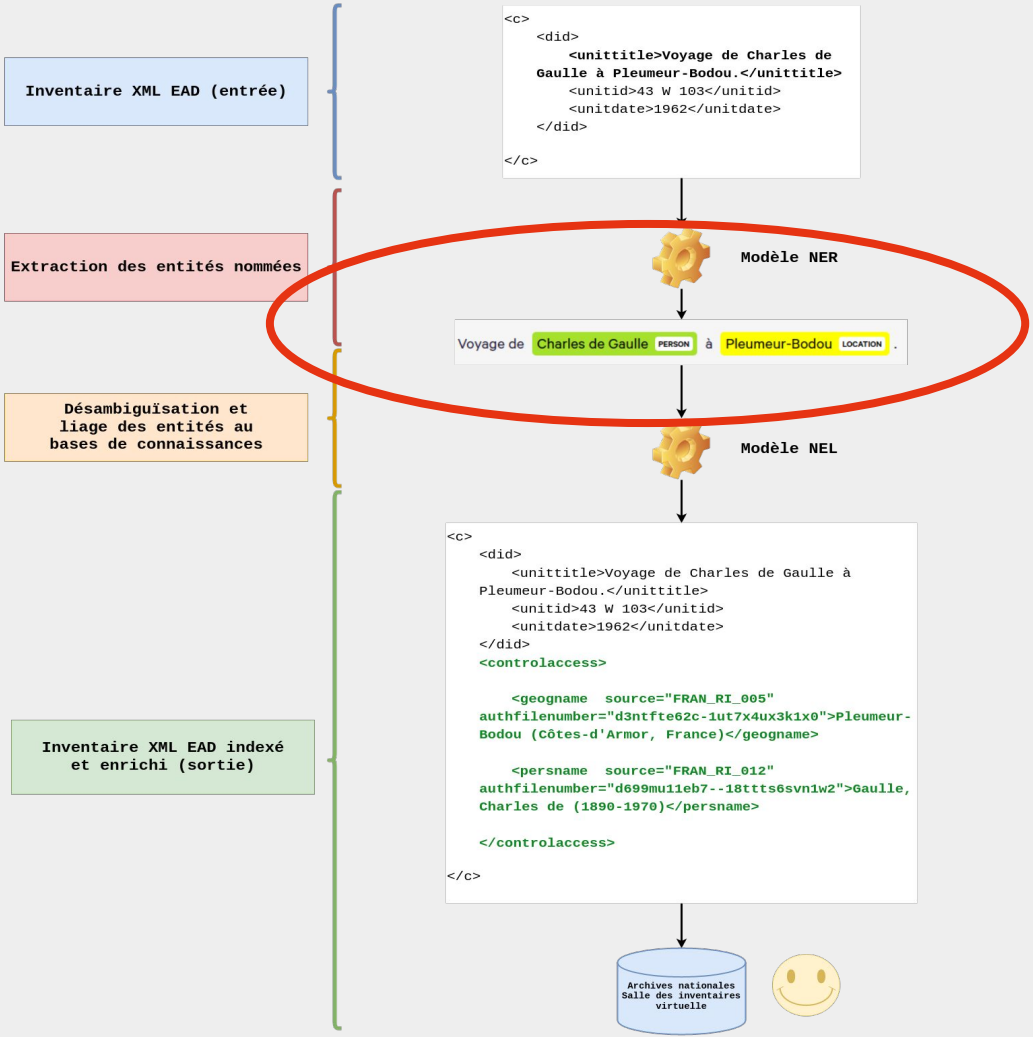
≈ 20 référentiels d'indexation / en cours d'enrichissement

Mais une sous-indexation des inventaires
→ Peu de points d'accès simples et intuitifs aux descriptions pour les utilisateurs

Le problème est le même dans de nombreux services d'archives français (où EAD est très utilisé et les inventaires sous-indexés)

Contexte NER4Archives

- Projet débuté en novembre 2020
- Financé par le Ministère de la Culture, les AN et INRIA (accord-cadre MC/INRIA) pour une durée de 1 an
- Archives nationales et équipe-projet ALMAnaCH (Inria)
- Mise en oeuvre : outils du TAL avec la tâche de *named entity recognition* (NER) / *entity linking (EL)*



Sommaire

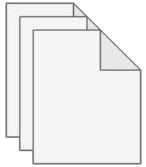
01. Constitution d'un premier corpus de données annotées pour l'entraînement

02. Première évaluation pour la reconnaissance des entités nommées (NER - *named entity recognition*) dans les inventaires

01

Constitution d'un premier corpus de données annotées pour l'entraînement

Données initiales pour l'annotation



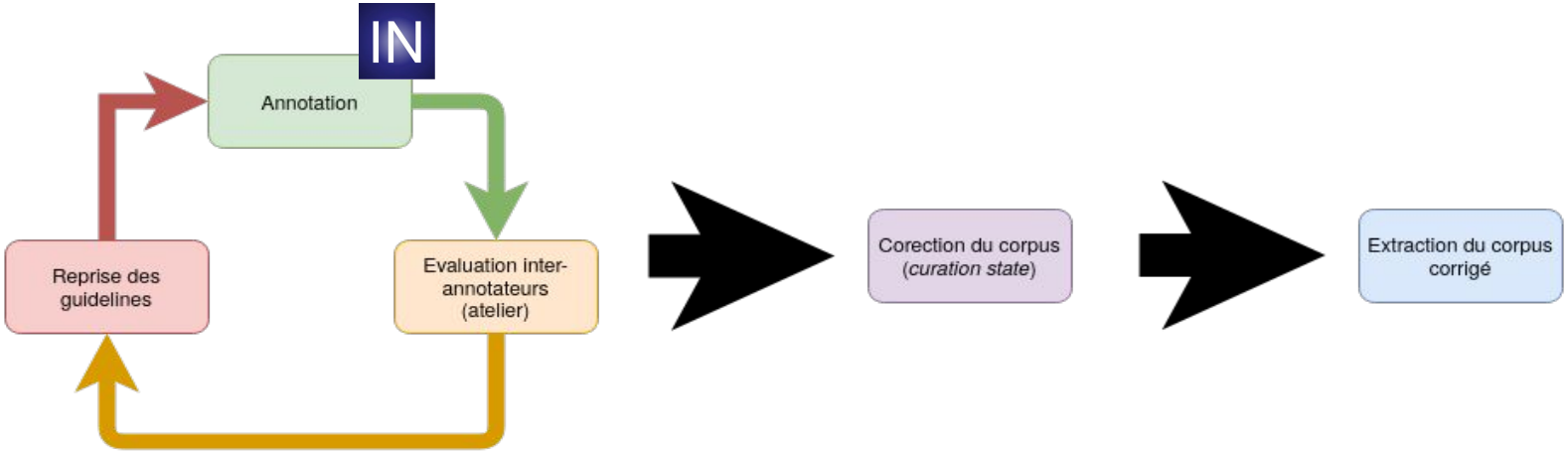
- Sélection de 8 instruments de recherche sur 17 sélectionnés suivant des critères précis

- Choix de 5 catégories : LOCATION, PERSON, ORGANISATION, EVENT, TITLE

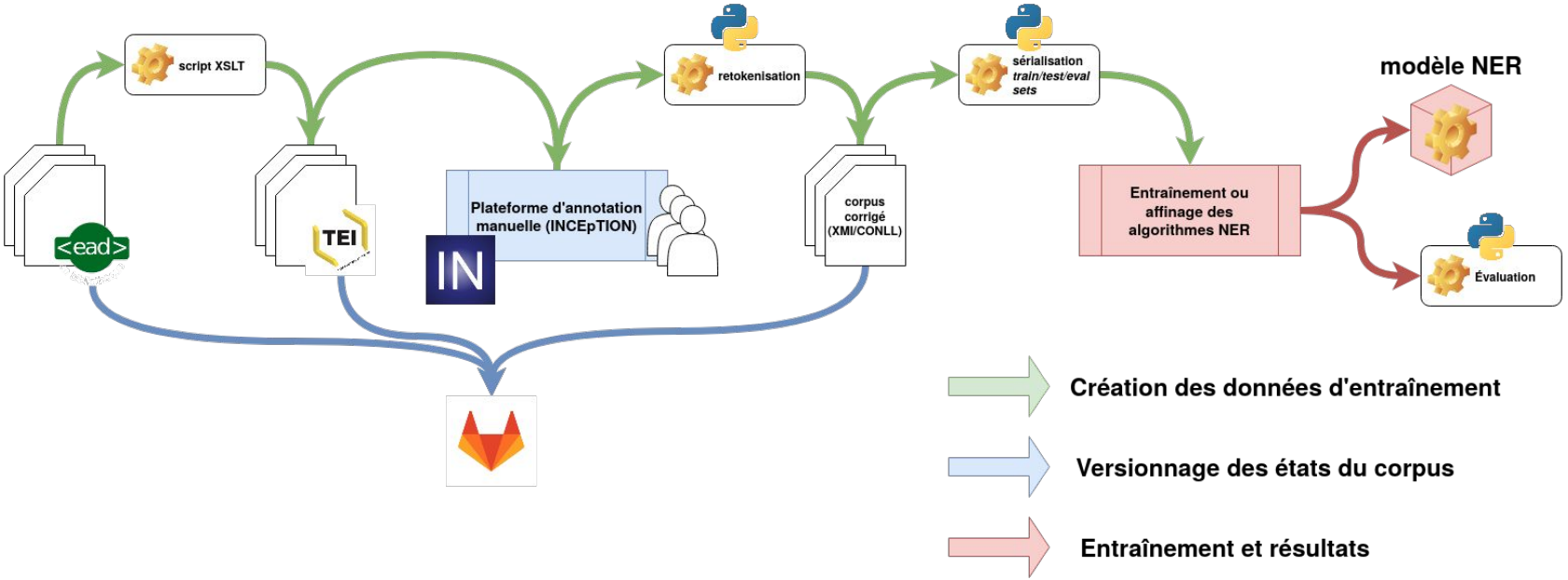
- Création de conventions d'annotations (*guidelines*) avec quelques exemples d'entités annotées dans leur contexte

- 4 annotateurs

La campagne d'annotation (vue micro)



La chaîne de traitement complète (vue macro)



La chaîne de traitement complète (étiquetage IOB, *Ramshaw & Marcus, 1995*)

Mainlevée O
par O
Louis B-PERSON
Paul I-PERSON
Blondel I-PERSON
, O
huissier B-TITLE
, O
demeurant O
51 B-LOCATION
, I-LOCATION
rue I-LOCATION
de I-LOCATION
Richelieu I-LOCATION
, O
au O
profit O
de O
Jean-Baptiste B-PERSON
Charles I-PERSON
Plisson I-PERSON
, O
agent B-TITLE
d' I-TITLE
affaires I-TITLE
, O
9 B-LOCATION
, I-LOCATION
rue I-LOCATION
Thérèse I-LOCATION
. O

extrait de l'inventaire FRAN_IR_041253 (service du département du Minutier Central des notaires parisiens)

Lettres O
de O
rémission O
accordées O
à O
Jean B-PERSON
Barrault I-PERSON
, O
écuyer B-TITLE
, O
homme B-TITLE
d' I-TITLE
armes I-TITLE
des I-TITLE
ordonnances I-TITLE
sous O
la O
charge O
de O
Jacques B-PERSON
Galiot I-PERSON
de I-PERSON
Genoilhac I-PERSON
, O
seigneur B-TITLE
d' I-TITLE
Assier I-TITLE
, O
sénéchal B-TITLE
d' I-TITLE
Armagnac I-TITLE
. O

extrait de l'inventaire FRAN_IR_000061 (service du Département du Moyen-Âge et de l'Ancien Régime)

État du premier corpus pour l'entraînement et les tests

- **Total de phrases annotées** : 872 / 106 513 séquences
- **Total d'entités** : 2 241 entités
- **Coefficient de Kappa de Fleiss** (évaluation inter-annotateurs):
  > 68.6 % (mesure du 07/10/2021)
  > + 17.3 % entre le 06/2021 et le 10/2021
  (sur l'interprétation du Kappa : Landis & Koch, 1977)
- **Temps d'annotation** (mise en place et cycle annotation-reprise-correction) ≈ 4 mois

Répartition des entités dans le corpus V1 de NER4Archives	
LOCATION	797
PERSON	495
ORGANISATION	490
TITLE	431
EVENT	28

02

Première évaluation pour la reconnaissance des entités nommées (NER - *named entity recognition*) dans les inventaires

Approches : Les outils et les architectures testées

- **Utilisation de modèles état de l'art pour le NER :**

Approches différentes : affinage d'un modèle existant (*finetuned*) ou entraînement de zéro (*from scratch*)

spaCy

- **CNN** (*Convolutional Neural Networks (CNN)* - Réseau neuronal convolutif)
- **Transformer basé sur BERT** (*camembert-base*)
- **BiLSTM-CRF** (*Recurrent neural network (RNN)* - Réseau de neurones récurrents)

DeLFT (Deep Learning Framework for Text)

Validation des méthodes : premiers résultats (a)

Résultats préliminaires obtenus avec des modèles entraînés ou affinés sur le corpus NER4Archives V1

Scores NER généraux

Architecture	F1-score
CNN (SpaCy)	0.53
Transformer BERT (SpaCy - camembert-base)	0.82
BiLSTM-CRF (DelfT)	0.58

Le score pour EVENT n'a pu être calculé en raison du nombre insuffisant d'entités dans l'ensemble d'entraînement

Scores NER par entités (F1-score)

Type de modèle / entités	PERSON	LOCATION	ORGANISATION	TITLE
CNN (SpaCy)	0.67	0.74	0.41	0.49
Transformer BERT (SpaCy - camembert-base)	0.92	0.83	0.78	0.92
BiLSTM-CRF (DelfT)	0.81	0.79	0.69	0.64

Validation des méthodes : modèle générique vs. modèle affiné (b)

Comparaison des prédictions sur des extraits d'inventaires :

- **modèle Transformer (affinage avec SpaCy sur corpus NER4Archives | F-score : 0.82)**
- **modèle générique CNN (entraîné avec SpaCy sur WikiNerFR | F-score : 0.39)**
- **Erreurs**

Exemple 1

Montereau-Fault-Yonne LOCATION (suite) à Neufmoutiers-en-Brie LOCATION Montereau-Fault-Yonne LOCATION , Centre d'hémodialyse ORGANISATION , 2009-2010 Nemours LOCATION , Centre hospitalier ORGANISATION , 2007-2009 Neufmoutiers-en-Brie LOCATION , Centre médical et pédagogique pour adolescents ORGANISATION , 2006-2008.

Montereau-Fault-Yonne (suite) à Neufmoutiers-en-Brie Montereau-Fault-Yonne LOCATION , Centre d'hémodialyse LOCATION , 2009-2010 Nemours, Centre LOCATION hospitalier 2007-2009 Neufmoutiers-en-Brie LOCATION , Centre LOCATION médical et pédagogique pour adolescents, 2006-2008.

Validation des méthodes : modèle générique vs. modèle affiné (b)

Comparaison des prédictions sur des extraits d'inventaires :

- **modèle Transformer (affinage avec SpaCy sur corpus NER4Archives | F-score : 0.82)**
- **modèle générique CNN (entraîné avec SpaCy sur WikiNerFR | F-score : 0.39)**
- **Erreurs**

Exemple 2

planche 131 Portraits des Coquelin PERSON , acteurs TITLE au Théâtre-Français ORGANISATION sans date 7 photographies.

planche 131 Portraits des Coquelin OTHER , acteurs au Théâtre-Français sans date 7 photographies.

Validation des méthodes : modèle générique vs. modèle affiné (b)

Comparaison des prédictions sur des extraits d'inventaires :

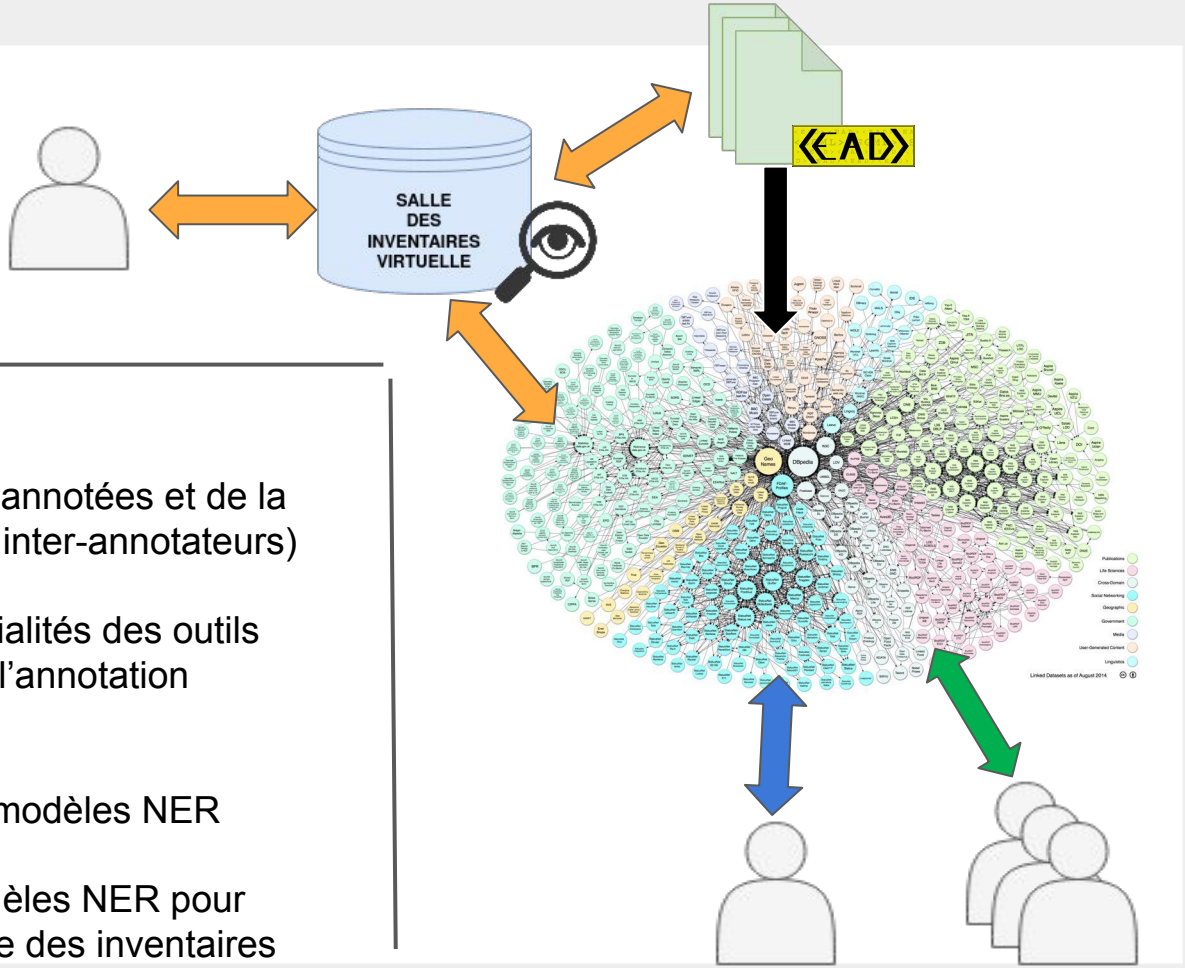
- modèle Transformer (affinage avec SpaCy sur corpus NER4Archives | F-score : 0.82)
- modèle générique CNN (entraîné avec SpaCy sur WikiNerFR | F-score : 0.39)
- Erreurs

Exemple 3

Déclaration de privilège de second ordre de Eugène Hippolyte Rascal PERSON demeurant 4 rue Vide-Gousset LOCATION , au profit de Charles Adolphe Schneider PERSON demeurant au 12 rue Gaillon LOCATION 8 mars 1838.

Déclaration de privilège OTHER de second ordre de Eugène Hippolyte Rascal PERSON demeurant 4 rue Vide-Gousset LOCATION au profit de Charles Adolphe Schneider PERSON demeurant au 12 rue Gaillon LOCATION 8 mars 1838.

Pistes

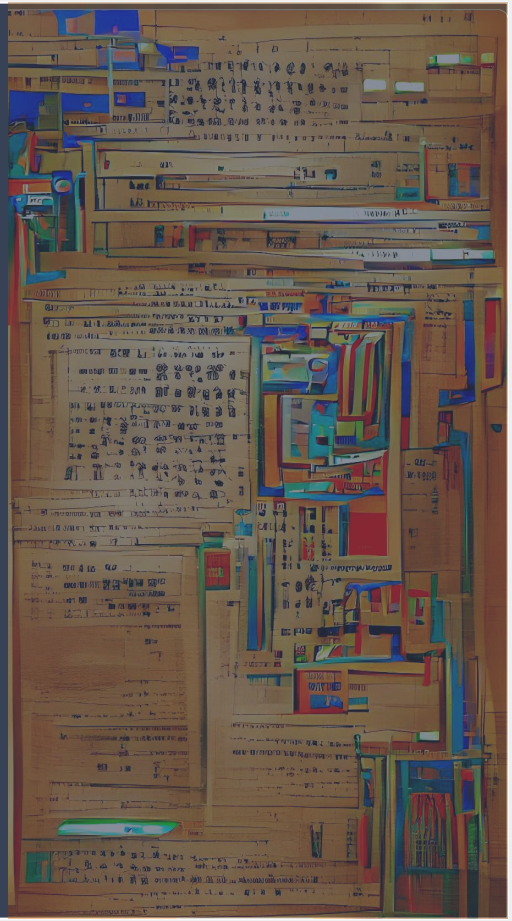


- Accroissement des données annotées et de la qualité d'annotation (mesure inter-annotateurs)
- Utilisation des pleines potentialités des outils existants (INCEpTION) pour l'annotation semi-automatique
- Nouvelles configurations de modèles NER
- Prototypage à partir des modèles NER pour l'indexation semi-automatique des inventaires

Merci de votre attention !

#FF21 #NER4Archives

Outils, méthodes, expérimentations, code-source du projet : https://gitlab.inria.fr/almanach/ner4archives



Bibliographie NER

- Cohen, C. (1960) A coefficient of agreement for nominal scales, Educational and psychological measurement.
- Dupont, Y. (2019) Un corpus libre, évolutif et versionné en entités nommées du français, TALN 2019 -Traitement Automatique des Langues Naturelles.
- Dupont, Y. (2017) Exploration de traits pour la reconnaissance d'entités nommées du Français par apprentissage automatique, 19e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL).
- Ehrmann, M., Romanello, M., Flückiger, A., Clematide, S. (2020), Extended Overview of CLEF HIPE 2020: Named Entity Processing on Historical Newspapers, CLEF 2020 Working Notes. Conference and Labs of the Evaluation Forum.
- Ehrmann, M. (2008) Les entités nommées, de la linguistique au TAL: statut théorique et méthodes de désambiguıisation, PhD thesis. Paris 7.
- Fort, K. (2012), Les ressources annotées, un enjeu pour l'analyse de contenu: vers une méthodologie de l'annotation manuelle de corpus, PhD thesis. Université Paris-Nord-Paris XIII.
- Fort, K., Sagot, B. (2010) Influence of pre-annotation on POS-tagged corpus development, The fourth ACL linguistic annotation workshop.
- Krippendorff, K. (2011) Computing Krippendorff's alpha-reliability.
- Landis, J. R., Koch, G. G. (1977) The Measurement of Observer Agreement for Categorical Data. Biometrics.
- Maurel, D., Friburger, N., Antoine, J.-Y., Eshkol, I., Nouvel, D. (2011) Cascades de transducteurs autour de la reconnaissance des entités nommées, Traitement automatique des langues.
- Neudecker C., Wilms L., Faber W. J., van Veen T. (2014) Large-scale refinement of digital historic newspapers with name entity recognition, Digital transformation and the changing role of news media in the 21st Century, IFLA Newspapers/GENLOC Pre-Conference Satellite Meeting Geneva.
- Ramshaw, L., Marcus, M. (1995) Text Chunking Using Transformation-Based Learning. Third ACL Workshop on Very Large Corpora. MIT.
- *Sagot, B., Richard, M., Stern, R. (2012). Annotation référentielle du Corpus Arboré de Paris 7 en entités nommées.*
- Javier Ortiz Suárez, P., Dupont, Y., Muller, B., Romary, L., Sagot, B. (2020) Establishing a New State-of-the-Art for French Named Entity Recognition, LREC 2020 - 12th Language Resources and Evaluation Conference.

Ressources

Outils

- **SpaCy** : Honnibal, M., Montani, I., Van Landeghem, S., Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python. https://spacy.io/
- **DeLFT** (Deep Learning Framework for Text) : Lopez, P. (2020), DeLFT. https://github.com/kermitt2/delft
- **INCEpTION** : Klie, J.-C., Bugert, M., Boullosa, B., Eckart de Castilho, R., Gurevych, I. (2018), The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. https://inception-project.github.io/

Projets cités :

- **Impresso. Media Monitoring of the Past :** https://impresso-project.ch/
- **Europeana Newspapers :** https://github.com/EuropeanaNewspapers/ner-corpora / https://api.bnf.fr/fr/texte-de-presse-annote-en-entites-nommees-du-projet-europeana-newspapers

Crédits images

[Slide 1]
- Logo Ministère de la Culture (https://commons.wikimedia.org/wiki/File:Minist%C3%A8re_de_la_Culture.svg)
- Logo Archives nationales (https://www.wikimedia.fr/appel-au-don-pour-wikipedien-en-residence/ob_b40c9f_logo-archives-nationales-gt-2/)
- Logo Inria (https://www.inria.fr/en/charter-use-visual-identity-inria)
- enseigne Almanach (© 2020 Alix Chagué)

[Slide 2]
- Logo EAD (https://francearchives.fr/file/0def64f5a10f3f1ae03fdea59399a3e0755ef157/static_1066.pdf)

[Slide 7/8]
- Logo INCEpTION (https://inception-project.github.io/)

[Slide 12]
- Mascotte Camembert (© 2020 Alix Chagué)
- Logo Delft (https://github.com/kermitt2/delft)
- Logo Keras-Tensorflow (https://blog.keras.io/keras-as-a-simplified-interface-to-tensorflow-tutorial.html)

[Slide 17]
- Carte des bases de données ouvertes du projet Linked Open Data CC

[Slide 8/18]
- Logo Gitlab (https://about.gitlab.com/press/press-kit/)
- “Named entity recognition in historical archives” image générée sur Dream Wombo art

L’ensemble des schémas ont été réalisés sur Diagrams.net