MINISTÈRE DE LA CULTURE
Liberté
Égalité
Fraternité

ARCHIVES NATIONALES

Inria

NER4Archives (named entity recognition for archives) : *Conception et réalisation d'un outil de détection, de classification et de résolution des entités nommées dans les instruments de recherche archivistiques encodés en XML/EAD.*

Florence Clavaud (Conservatrice générale du patrimoine et responsable du Lab des Archives nationales)

Laurent Romary (Directeur de recherche à ALMAnaCH-Inria)

Pauline Charbonnier (Ingénieure d'études au Lab des Archives nationales)

Lucas Terriel (Ingénieur R&D à ALMAnaCH-Inria)

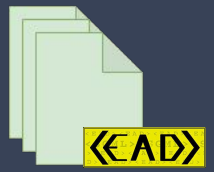
Gaetano Piraino et **Vincent Verdese** (AN DSI)

Archives nationales - 22 mars 2022

Atelier Culture - INRIA
#AtelierCultureInria2022

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

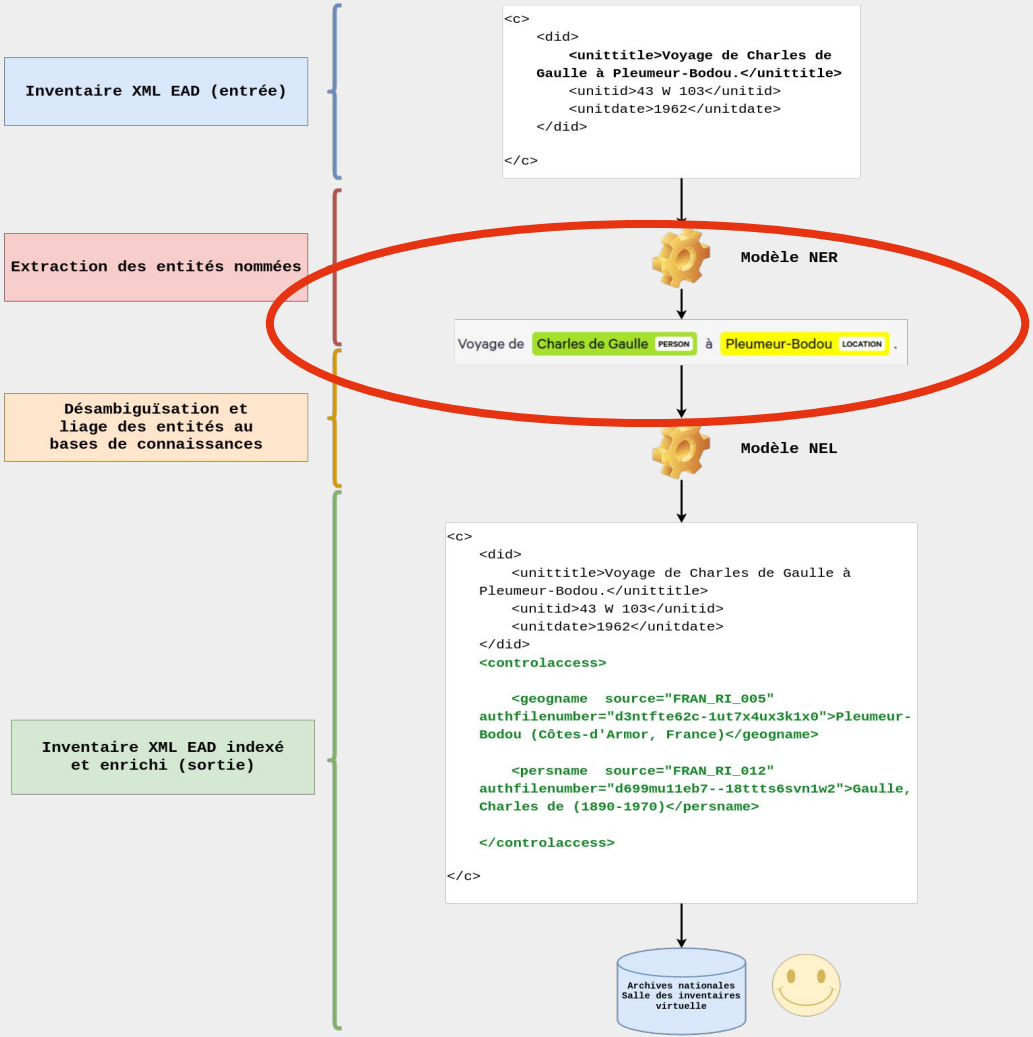
≈ 20 référentiels d'indexation / en cours d'enrichissement

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)*

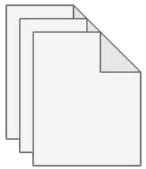


Sommaire

01

Cycle de vie du projet en première phase : méthodologie pour la tâche de reconnaissances en entités nommées

Données initiales pour l'annotation



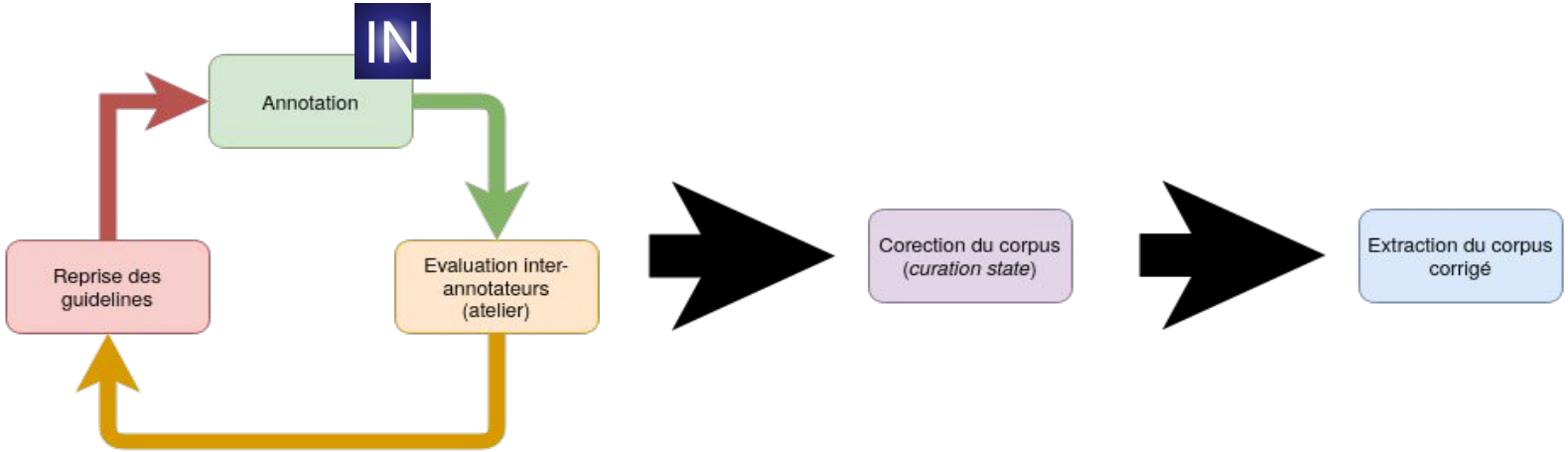
- Sélection de 8 instruments de recherche (V1) étendu à 12 (V2) sur 17 sélectionnés suivant des critères précis

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

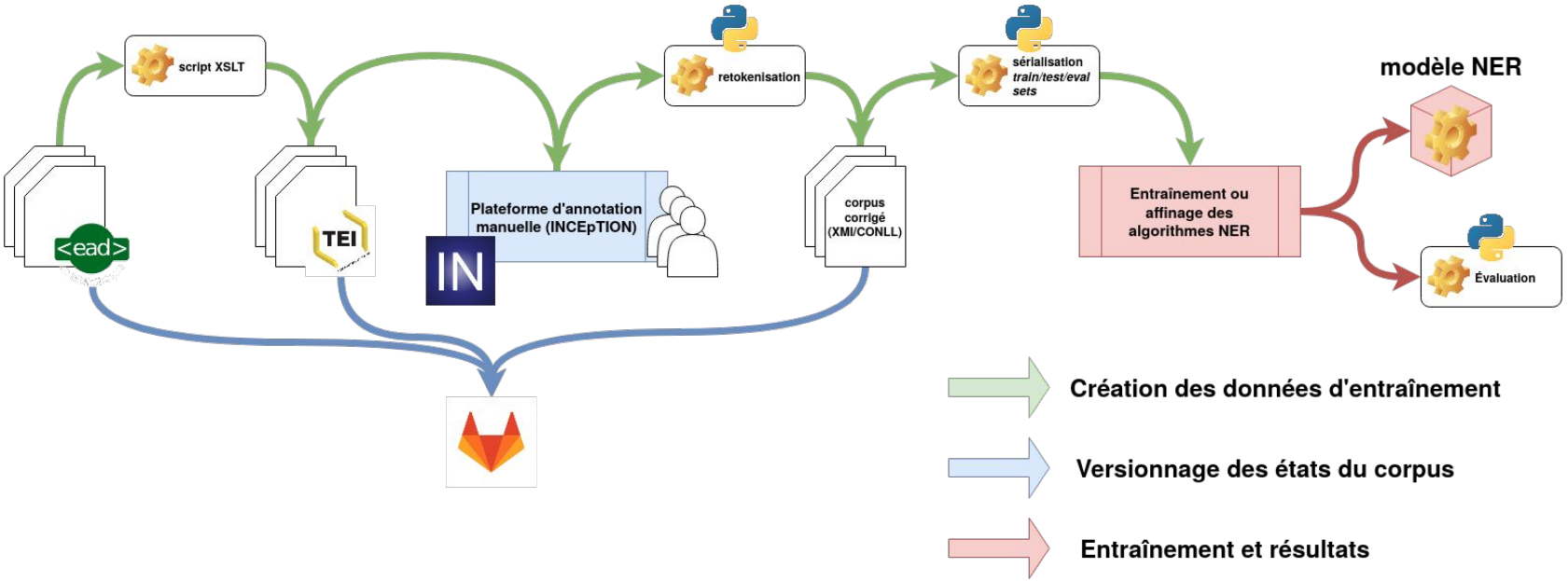
- Création de conventions d'annotations (*guidelines*) avec quelques exemples d'entités annotées dans leur contexte

- 4 annotateurs

Déroulement de la campagne d'annotation : itérations et évaluations



La chaîne de traitement complète



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 (responsable : 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 (responsable : Département du Moyen Âge et de l'Ancien Régime)

02

Évolutions des résultats NER sur les instruments de recherche

Évolution de l'état du corpus

- **Total de phrases annotées** :

  2179 / 120 115 phrases (+1307 par rapport à la version 1)

- **Total d'entités annotées** : 5 911 entités (+ 3670)

- **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**

  cycle annotation-reprise-correction ≈ 4 mois

Répartition des entités dans le corpus de NER4Archives (dernière version)	
LOCATION	2158 (+1361)
PERSON	1427 (+932)
ORGANISATION	1095 (+605)
TITLE	1177 (+746)
EVENT	54 (+26)

Évaluation de la tâche NER (1) : approches, architectures et outils

- **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)

Évaluation de la tâche NER (2) : validation des méthodes

Évolutions des résultats obtenus avec des modèles entraînés ou affinés sur le corpus NER4Archives (comparaison avec la version 1 du corpus)

Scores NER généraux

Architecture	F1-score
CNN (SpaCy)	0.72 (+0.19)
Transformer BERT (SpaCy - camembert-base)	0.91 (+0.09)
BiLSTM-CRF (DelfT)	0.76 (+0.18)

Scores NER par entités (F1-score)

Type de modèle / entités	PERSON	LOCATION	ORGANISATION	TITLE	EVENT
CNN (SpaCy)	0.68 (+0.01)	0.70 (-0.04)	0.77 (+0.36)	0.77 (+0.28)	0.27
Transformer BERT (SpaCy - camembert-base)	0.96 (+0.04)	0.93 (+0.10)	0.88 (+0.10)	0.88 (-0.04)	0.40
BiLSTM-CRF (DelfT)	0.81	0.80 (+0.01)	0.68 (-0.01)	0.70 (+0.06)	-

Évaluation de la tâche NER (2) : validation des méthodes (modèle générique vs. modèle affiné)

- *Prédictions réalisées sur des données n'appartenant pas au corpus d'entraînement afin de tester la généralisation des modèles sur d'autres données (extraits des inventaires en EAD)*

Types d'erreurs	✖ modèle générique CNN (entraîné avec SpaCy sur corpus WikiNerFR)	✔ modèle Transformer (affinage avec SpaCy sur corpus NER4Archives)
• **Bruit**	'terres, vignes, prez LOCATION , isles et sausayes deppendant PERSON	"terres, vignes, prez, isles et sausayes deppendant
• **Classe erronée**	pour le collège de Navarre LOCATION 1639.	pour le collège de Navarre ORGANISATION 1639.
• **Segmentation incomplète ou erronée du token**	Sarry (Marne LOCATION) Voir : Châlons-sur-Marne LOCATION . appartenant au couvent des Mathurins LOCATION de Paris LOCATION veuve Paul Antoine Maurey PERSON , née Aubier LOCATION (23 avril 1925).	Sarry (Marne) LOCATION Voir : Châlons-sur-Marne LOCATION . appartenant au couvent des Mathurins de Paris ORGANISATION veuve Paul Antoine Maurey, née Aubier PERSON (23 avril 1925).

03

Le NER, première étape vers le liage des entités nommées

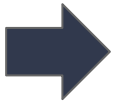
Approche pour la tâche de liage des entités nommées (*Entity linking*)

Extraction de l'entité (NER)

Génération de candidats (recherche approximative - fuzzy search dans la base de connaissance)

***Classement des candidats* (modèle de désambiguïsation lexicale et sémantique)**

Ford est un industriel automobile américain du XXe siècle.



- **ID1 : Harrison Ford** (acteur)
- **ID2** : **Henry Ford** (industriel)
- **ID3** : **Ford** (marque de voiture)

1. **ID2** : **Henry Ford** : 0.86
2. **ID3** : **Ford** : 0.09
3. **ID1** : **Harrison Ford** : 0.05

- Difficultés de la tâche : **variantes orthographiques** (ex. “Plato” pour “Platon”), **variantes de noms** (ex. “Ville Lumière” pour “Paris”) et **ambiguïtés** (ex. “empereur des Français” (quel empereur ?) ou “pape de Rome” (quel pape ?)).

Focus sur l'outil *Entity-Fishing* : une stratégie pour le liage des entités ?

Exemple d'Entity-linking avec Entity-fishing pour un extrait d'inventaire (FRAN_054605 - DAEAA) des Archives nationales.

Requête JSON


```
{
  "text": "19/07/2000 TOKYO : Point de
presse conjoint du Président de la
République de M Yoshiro MORI Premier
ministre du Japon et de M Romano PRODI
Président de la Commission européenne,
durée : 00:23:10",
  "language": {"lang": "fr"},
  "entities": [
        {
        "rawName": "TOKYO",
        "type": "LOCATION",
        "offsetStart": 11,
        "offsetEnd": 16
        },
        {
        "rawName": "Yoshiro MORI",
        "type": "PERSON",
        "offsetStart": 79,
        "offsetEnd": 91
        },
        {
        "rawName": "Romano PRODI",
        "type": "PERSON",
        "offsetStart": 126,
        "offsetEnd": 138
        }
  ],
  "mentions": [],
  "nbest": false,
  "sentence": false,
  "customisation": "generic"
}
```


Réponse JSON traitée par le client

Les mentions "Tokyo", "Yoshiro Mori", et "Romano Prodi" (après extraction du modèle NER) sont confrontées à la base de connaissances Wikidata qui renvoie le candidat avec le score le plus important.

Patrice Lopez (INRIA/Science Miner), *Entity-fishing*, 2016-2022, https://github.com/kermitt2/entity-fishing

Conclusion

Ce qui reste à faire

- 2022 :
  - ateliers internes aux Archives nationales
  - accroissement des données annotées et de la qualité de l'annotation (mesure inter-annotateurs)
  - utilisation des pleines potentialités des outils existants (INCEpTION) pour l'annotation semi-automatique

- 2023 (sous réserve d'un nouveau financement apporté par le ministère) :
  - nouvelles configurations de modèles NER
  - mise en place de dispositifs de liage des entités avec les entités des bases de connaissances sélectionnées (Wikidata, référentiels AN) et d'un workflow d'enrichissement des référentiels des Archives nationales (eux-mêmes déjà sémantisés, voir https://github.com/ArchivesNationalesFR/Referentiels)
  - prototypage à partir des modèles NER pour l'indexation semi-automatique des inventaires

Merci de votre attention !

#AtelierCultureInria2022 **#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 désambiguısation, 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
- **Entity Fishing** : Lopez, P. (2022). https://github.com/kermitt2/entity-fishing
- **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 entities in the Web” image générée sur Dream Wombo art

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