



HAL
open science

Les catalogues et GROBID

Simon Gabay, Mohamed Khemakhem, Laurent Romary

► **To cite this version:**

Simon Gabay, Mohamed Khemakhem, Laurent Romary. Les catalogues et GROBID. Doctorat. Du catalogue aux humanités numériques : quelles méthodes pour quels résultats?, Paris, France. 2018. cel-01951107

HAL Id: cel-01951107

<https://hal.science/cel-01951107>

Submitted on 11 Dec 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Les catalogues et GROBID

Simon Gabay
Université de
Neuchâtel

Mohamed
Khemakhem
ALMAAnCH
Paris 7
Centre Marc Bloch

Laurent Romary
ALMAAnCH
Centre Marc Bloch
BBAW

15 novembre 2018



L'encodage automatique de catalogues de vente en TEI avec *GROBID dictionaries*

GROBID (dictionaries)

La famille GROBID

- ▶ Bibliographie:
<http://cloud.science-miner.com/grobid>
- ▶ NERD: <http://cloud.science-miner.com/nerd>
- ▶ Quantities:
<http://cloud.science-miner.com/quantity>
- ▶ le petit nouveau: Dictionaries

GROBID dictionaries

- ▶ Développé par Mohamed Khemakhem (Paris VI-CMB/Humboldt)
- ▶ Initialement prévu pour les ressources lexicales numérisées (dictionnaires, glossaires, lexiques. . .)
- ▶ Détournement du système pour les sources *Encyclopedic-like*
- ▶ Annuaire, catalogues. . .

L'encodage TEI et *GROBID dictionaries*

La TEI de *GROBID dictionaries*

- ▶ Les options d'encodage sont limitées. . .
- ▶ . . . et basée sur la TEI pour les ressources lexicographiques
- ▶ la structure des dictionnaires et des catalogues est cependant similaire: on peut exploiter cette ressemblance. . .
- ▶ . . . ce qui entraîne une utilisation quelque peu «métaphorique» de la TEI

ABERDEEN [*aberdin'*], v. d'Ecosse, ch.-l. de comté; port sur la mer du Nord; 170.000 h. Université.

ABERDEEN (G. H. Gordon, *comte d'*), homme d'Etat anglais, né à Edimbourg. Premier ministre en 1852, il conclut avec la France une alliance contre la Russie (1784-1860).

ABER-VRACH, fl. côtier du Finistère (Atlantique); 34 kil. Station marémotrice d'essai.

ABGAR, nom de huit rois d'Edesse, en Mésopotamie (132 av. J.-C.-216 apr.).

ABIA, roi de Juda, fils de Roboam, vainqueur de Jéroboam, roi d'Israël (957-955 av. J.-C.).

ABIDJAN, ch.-l. de la Côte-d'Ivoire (A.-O. F.), sur une vaste lagune navigable; 15.000 h.

ABIMÉLECH [*lèk*], fils de Gédéon. Il devint Juge d'Israël, après avoir fait égorger ses frères; il établit son pouvoir sur Sichem et fut tué au siège de Thèbes, en Palestine (vers 1100 av. J.-C.).

ABIRON, lévite qui fut englouti dans la terre avec Coré et Dathan, tous trois révoltés contre Moïse et Aaron (*Bible*).

Le Petit Larousse illustré

ABERDEEN (G. H. Gordon, *comte d'*), homme d'Etat anglais, né à Edimbourg. Premier ministre en 1852, il conclut avec la France une alliance contre la Russie (1784-1860).

```
1 ▼ <entry>
2 ▼   <form type="lemma">
3     <persName>ABERDEEN</persName>
4     <addName>(G. H. Gordon, comte d')</addName><pc>,</pc>
5     <desc>homme d'Etat anglais, né à Edimbourg</desc>
6   </form>
7   <pc>.</pc>
8 ▼   <sense>
9     <def>Premier ministre en 1852, il conclut avec la France
10      une alliance contre la Russie (1784-1860)</def>
11   </sense>
12   <pc>.</pc>
13 </entry>
```

Le *Petit Larousse illustré*

- 49 **Kourakin** (le prince Alexis B.), frère du précédent, homme d'Etat russe. — Billet aut. sig., en français, à M. Monférand, 1 p. in-8. 2 »
- 50 **Labanoff** (le prince Alex.), célèbre général et écrivain russe, historien de Marie Stuart. — L. a. s., en français, 1835, 1 p. in-4. 3 »
- 51 **Ladislas IV**, roi de Pologne, célèbre par ses succès contre les Russes, époux de Marie de Gonzague. — L. sig., en latin, au cardinal de Montalte; Varsovie, 1645, 1 p. in-f. 8 »
- 52 **Lafayette**, illustre général. — L. a. sig. de ses initiales à M. Masclet; Washington, 13 août 1825, 1 p. 1/4 in-4. Un peu fatiguée. 15 »
- Très-curieuse lettre sur le voyage qu'il fit en Amérique, de 1824 à 1825. « C'est avec de bien tendres regrets que je quitterai cette terre américaine, le bon, grand et heureux peuple des Etats-Unis auquel je suis amalgamé depuis près d'un demi-siècle, et qui vient encore de me combler de ses bontés. J'y ai vu les miracles de l'indépendance, de la liberté, égalité et *self government*; le problème des institutions républicaines a été résolu ici sur une grande échelle et jamais expérience n'a si bien réussi. » Il comptait retourner comme il était venu, sur un paquebot-poste, mais le peuple et le gouvernement en ont disposé autrement. On a donné le nom de *Brandywine* à une superbe frégate qui est chargée de le ramener en France.

Revue des autographes

49 **Kourakin** (le prince Alexis B.), frère du précédent, homme d'Etat russe. — Billet aut. sig., en français, à M. Monférand, 1 p. in-8. 2 »

```

1 ▼ <entry>
2     <num>49</num>
3 ▼     <form type="lemma">
4         <surName>Kourakin</surName>
5         <addName>(le prince Alexis B.),</addName>
6         <desc> frère du précédent, homme d'Etat russe.</desc>
7     </form>
8 ▼     <sense>
9         <pc>-</pc>
10 ▼         <def>
11             <bibl>Billet auto sig., en francais, à M. Monférand, 1 p, in-
12                 8.</bibl> <num type="price">2 »</num>
13         </def>
14     </sense>
15 </entry>

```

Revue des autographes

Installer *GROBID dictionaries*

Quelques remarques

- ▶ *GROBID dictionaries* utilise un *container docker* (qu'il faut télécharger)
- ▶ Cela permet de fournir simplement un espace de travail avec toutes les librairies, dépendances et outils nécessaires
- ▶ L'installation se fait *via* le terminal

Ligne de commande

```
$ docker pull medkhem/grobid-dictionaries
```

Téléchargement du *container*

```
docker pull medkhem/grobid-dictionaries
Using default tag: latest
latest: Pulling from medkhem/grobid-dictionaries
bc9ab73e5b14: Pull complete
193a6306c92a: Pull complete
e5c3f8c317dc: Pull complete
a587a86c9dcb: Pull complete
a4c7ee7ef122: Pull complete
a7c0dad691e9: Pull complete
367a6a68b113: Pull complete
60c0e52d1ec2: Pull complete
c9d22bc43935: Pull complete
e3de5a273367: Pull complete
c8a45fcf02c1: Pull complete
a9cba52ae4f5: Pull complete
9a69063e63f8: Pull complete
7189b29e6a6d: Pull complete
e59d50656f46: Pull complete
de8d2ebe6e42: Pull complete
Digest: sha256:e7b30f49637da3c8ce86462418e11e2667b35e728d8a8036ee640d34bbd23e16
Status: Downloaded newer image for medkhem/grobid-dictionaries:latest
```

container docker

Les catalogues et GROBID

Short Name (U ABC)

GROBID

Encodage TEI

Intaller GROBID

Machine teaching

Visualiser le résultat

Conclusion

On lance docker

Ligne de commande

```
$ docker run -it medkhem/grobid-dictionaries bash
```


Données d'entraînement

Les catalogues et GROBID

Short Name (U ABC)

GROBID

Encodage TEI

Intaller GROBID

Machine teaching

Visualiser le résultat

Conclusion

Principe d'entraînement

- ▶ le *container* doit être synchronisé avec un dossier *toydata* (disponible sur <https://github.com/MedKhem/grobid-dictionaries>)
- ▶ C'est dans ce dossier *toydata* que vont être placées les données d'entraînement
- ▶ On exécute la commande ¹

Ligne de commande

```
$ docker run -v  
PATH_TO_YOUR_TOYDATA/toyData:/grobid/grobid-dictionaries/r  
-p 8080:8080 -it medkhem/grobid-dictionaries bash
```

¹Nous sommes sur mac, pour les commandes sous windows, cf. <https://github.com/MedKhem/grobid-dictionaries>

Les catalogues et GROBID

Short Name (U ABC)

GROBID

Encodage TEI

Intaller GROBID

Machine teaching

Visualiser le résultat

Conclusion

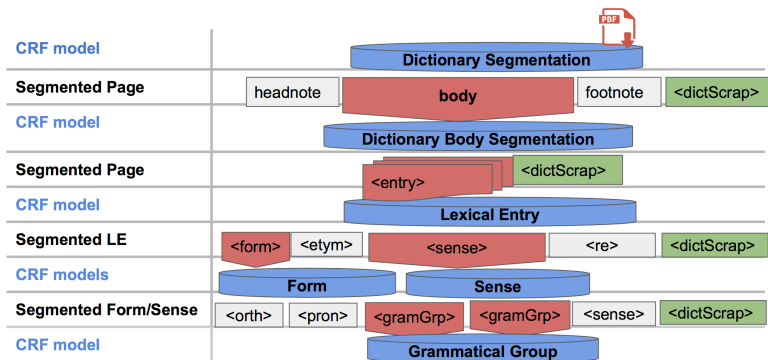
Teaching GROBID dictionaries

Sous le capot

- ▶ Développé en Java
- ▶ Utilise des champs aléatoires conditionnels (*conditional random fields*), des modèles statistiques utilisés en reconnaissance des formes et plus généralement en apprentissage statistique
- ▶ Ces champs fonctionnent en cascade
- ▶ C'est donc du *machine learning*...
- ▶ ...et ça commence donc par du *machine teaching*

Premier niveau d'entraînement

Dictionary Segmentation



Dictionary Segmentation

L'objectif est de séparer le corps du texte du reste (titre courant, numéro de page. . .)

On crée les données d'entraînement

ligne de commande

```
$ java -jar  
/grobid/grobid-dictionaries/target/grobid-dictionaries-0.4.3-SNAPSHOT.one-jar.jar -dIn  
resources/dataset/dictionary-segmentation/corpus/pdf/ -dOut resources -exe  
createTrainingDictionarySegmentation
```

On annote

headnote	-14 -	headnote
lb	lb	body
lb	lb	43 Foroade-Laroquette (Ad. de), ministre de Napoléon III. -
lb	lb	L. a. s. à Stéphen de la Madeleine, 1840, 1 p. in-8. 2 50
lb	lb	44 Garcia (Eugénie), célèbre cantatrice italienne. -L. a. s. à
lb	lb	Stéphen de la Madeleine. 1850, 3 p. in-8. 2 »
lb	lb	45 Gautier (Théop.), le célèbre critique. -L. a. s. à Stéphen de
lb	lb	la Madeleine, 1850, 1 p. in-8. 2 50
lb	lb	46 Godefroy (Théodore), savant historien et généalogiste. -
lb	lb	Huit. a. s. sur vélin, 1604, pet. in-4 oblong. 2 »
lb	lb	47 Hadick (le comte André), général hongrois au service de l'Au-
lb	lb	triche, célèbre pas ses exploits dans la guerre de Sept ans. -
lb	lb	L. s. à M. de Balsch, en français; Vienne, 9 janv. 1782, 2 p.

Entraînement

Les catalogues et GROBID

Short Name (U ABC)

GROBID

Encodage TEI

Intaller GROBID

Machine teaching

Visualiser le résultat

Conclusion

On entraîne à partir des données annotées

ligne de commande

```
$ mvn generate-resources -P train_dictionary_segmentation -e
```

L'entraînement est lancé

[301]	obj=133.17	act=332	err= 0.02%/42.86%	time=0.21s/51.04s
[302]	obj=133.17	act=341	err= 0.01%/14.29%	time=0.21s/51.25s
[303]	obj=133.17	act=332	err= 0.02%/42.86%	time=0.21s/51.46s
[304]	obj=133.17	act=341	err= 0.01%/14.29%	time=0.21s/51.67s
[305]	obj=133.16	act=332	err= 0.02%/42.86%	time=0.21s/51.88s
[306]	obj=133.16	act=340	err= 0.01%/14.29%	time=0.21s/52.09s
[307]	obj=133.16	act=333	err= 0.02%/42.86%	time=0.21s/52.29s
[308]	obj=133.16	act=336	err= 0.01%/14.29%	time=0.21s/52.50s
[309]	obj=133.15	act=333	err= 0.01%/28.57%	time=0.21s/52.71s
[310]	obj=133.15	act=333	err= 0.00%/ 0.00%	time=0.21s/52.92s
[311]	obj=133.15	act=330	err= 0.01%/28.57%	time=0.21s/53.13s
[312]	obj=133.15	act=318	err= 0.00%/ 0.00%	time=0.21s/53.34s
[313]	obj=133.14	act=318	err= 0.01%/14.29%	time=0.21s/53.55s
[314]	obj=133.14	act=318	err= 0.00%/ 0.00%	time=0.21s/53.76s
[315]	obj=133.14	act=319	err= 0.01%/14.29%	time=0.21s/53.97s
[316]	obj=133.14	act=329	err= 0.00%/ 0.00%	time=0.21s/54.17s
[317]	obj=133.13	act=318	err= 0.01%/14.29%	time=0.21s/54.38s
[318]	obj=133.13	act=317	err= 0.00%/ 0.00%	time=0.21s/54.59s
[319]	obj=133.13	act=318	err= 0.01%/28.57%	time=0.14s/54.73s
[320]	obj=133.12	act=326	err= 0.00%/ 0.00%	time=0.14s/54.87s
[321]	obj=133.12	act=317	err= 0.01%/14.29%	time=0.14s/55.00s
[322]	obj=133.12	act=325	err= 0.00%/ 0.00%	time=0.14s/55.14s
[323]	obj=133.12	act=326	err= 0.01%/14.29%	time=0.15s/55.29s
[324]	obj=133.12	act=324	err= 0.00%/ 0.00%	time=0.14s/55.43s
[325]	obj=133.11	act=325	err= 0.01%/14.29%	time=0.14s/55.57s
[326]	obj=133.11	act=325	err= 0.00%/ 0.00%	time=0.15s/55.72s
[327]	obj=133.11	act=326	err= 0.01%/28.57%	time=0.14s/55.86s
[328]	obj=133.10	act=326	err= 0.01%/28.57%	time=0.22s/56.07s
[329]	obj=133.10	act=324	err= 0.01%/28.57%	time=0.14s/56.21s
[330]	obj=133.10	act=326	err= 0.01%/14.29%	time=0.15s/56.36s
[331]	obj=133.10	act=326	err= 0.02%/42.86%	time=0.15s/56.51s
[332]	obj=133.10	act=335	err= 0.01%/14.29%	time=0.14s/56.64s

On contrôle la qualité de l'entraînement

```

createTrainingDictionaryBodySegmentation
=====
label      accuracy  precision  recall    f1
<entry>   100         100        100       100
<pc>      100         100        100       100
-----
all fields ENTRY  100         100        100       100   (micro average)
          100         100        100       100   (macro average)

===== Field-level results =====
label      accuracy  precision  recall    f1
<entry>   100         100        100       100
<pc>      100         100        100       100
-----
all fields  100         100        100       100   (micro average)
          100         100        100       100   (macro average)

===== Instance-level results =====

Total expected instances:  2
Correct instances:         2
Instance-level recall:     100
  
```

Les catalogues et GROBID

Short Name (U ABC)

GROBID

Encodage TEI

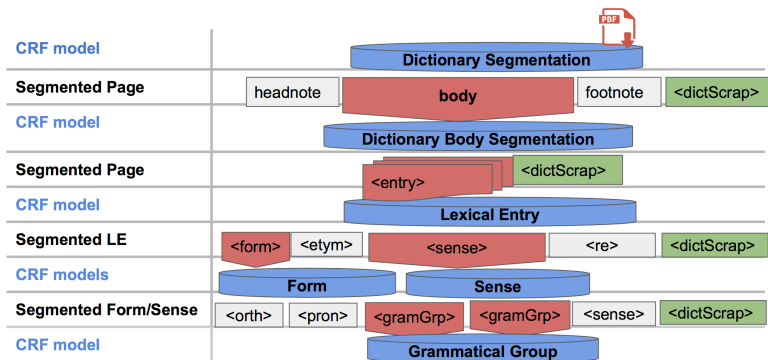
Intaller GROBID

Machine teaching

Visualiser le résultat

Conclusion

Deuxième niveau d'entraînement *Body Segmentation*



Body Segmentation

L'objectif est de séparer les différentes entrées

```
1 ▼ <entry>
2     54
3     |
4         Lassalle
5         (A.-Ch.-L. de),
6         le plus brillant général de cavalerie des guerres de la République et de
7         l'Empire, né à Metz, tué à la bataille de Wagram
8
9         .-
10
11         L. a. s. au général Dugua; 1 p. in-f.
12         10 »
13
14         Superbe lettre sur la campagne d'Egypte. Il profite du départ du général Desaix
15         pour lui donner des nouvelles. Desaix lui laisse le commandement de la colonne
16         qui doit poursuivre Mourad-Bey, et qui se compose de 400 hommes de cavalerie, 4
17         pièces de canon et 160 dromadaires. Le général Boyer a, dans une petite
18         affaire, tué 10 mameloucks et 40 arabes, etc.
```

Body Segmentation

On crée les données d'entraînement

ligne de commande

```
$ java -jar  
/grobid/grobid-dictionaries/target/grobid-dictionaries-0.4.3-SNAPSHOT.one-jar.jar -dIn  
resources/dataset/dictionary-segmentation/corpus/pdf/ -dOut resources -exe  
createTrainingDictionaryBodySegmentation
```

On annote

entry 43 Foroadé-Laroquette (Ad. de), ministre de Napoléon III. -
lb lb L. a. s. à Stéphen de la Madeleine, 1840, 1 p. in-8. 2 50 entry
lb lb entry 44 Garcia (Eugénie), célèbre cantatrice italienne. -L. a. s. à
lb lb Stéphen de la Madeleine. 1850, 3 p. in-8. 2 entry »
lb lb entry 45 Gautier (Théop.), le célèbre critique. -L. a. s. à Stéphen de
lb lb la Madeleine, 1850, 1 p. in-8. 2 50 entry
lb lb entry 46 Godefroy (Théodore), savant historien et généalogiste. -
lb lb Huit. a. s. sur vélin, 1604, pet. in-4 oblong. 2 entry »

Entraînement

Les catalogues et GROBID

Short Name (U ABC)

GROBID

Encodage TEI

Intaller GROBID

Machine teaching

Visualiser le résultat

Conclusion

On entraîne à partir des données annotées

ligne de commande

```
$ mvn generate-resources -P train_dictionary_body_segmentation -e
```

Et on passe au modèle suivant

Les catalogues et GROBID

Short Name (U ABC)

GROBID

Encodage TEI

Intaller GROBID

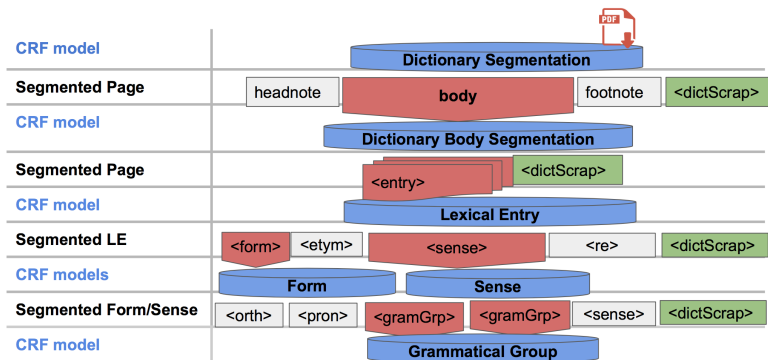
Machine teaching

Visualiser le résultat

Conclusion

Troisième niveau d'entraînement

Lexical Entry



Lexical Entry

L'objectif est de séparer les grandes articulations de chaque entrée du catalogue

```
1 ▾ <entry>
2   <num>54</num>
3 ▾   <form type="lemma">
4     Lassalle
5     (A.-Ch.-L. de),
6     le plus brillant général de cavalerie des guerres de la République et de
       l'Empire, né à Metz, tué à la bataille de Wagram
7   </form>
8 ▾   <sense> .-
9
10      L. a. s. au général Dugua; 1 p. in-f.
11      10 »
12
13      Superbe lettre sur la campagne d'Egypte. Il profite du départ du général Desaix
       pour lui donner des nouvelles. Desaix lui laisse le commandement de la colonne
       qui doit poursuivre Mourad-Bey, et qui se compose de 400 hommes de cavalerie, 4
       pièces de canon et 160 dromadaires. Le général Boyer a, dans une petite
       affaire, tué 10 mameloucks et 40 arabes, etc.
14   </sense>
15 </entry>
16
```

Lexical Entry

On crée les données d'entraînement

ligne de commande

```
$ java -jar  
/grobid/grobid-dictionaries/target/grobid-dictionaries-0.4.3-SNAPSHOT.one-jar.jar -dIn  
resources/dataset/dictionary-segmentation/corpus/pdf/ -dOut resources -exe  
createTrainingLexicalEntry
```

On annote

num 43 **num** **form** Foroade-Laroquette (Ad. de), ministre de Napoléon III **form** . -
lb **lb** **sense** L. a. s. à Stéphen de la Madeleine, 1840, 1 p. in-8. 2 50 **sense**
lb **lb**

num 44 **num** **form** Garcia (Eugénie), célèbre cantatrice italienne **form** . - **sense** L. a. s. à
lb **lb** Stéphen de la Madeleine. 1850, 3 p. in-8. 2 **sense**

num 45 **num** **form** Gautier (Théop.), le célèbre critique **form** . - **sense** L. a. s. à Stéphen de
lb **lb** la Madeleine, 1850, 1 p. in-8. 2 50 **sense**
lb **lb**

Entraînement

Les catalogues et GROBID

Short Name (U ABC)

GROBID

Encodage TEI

Intaller GROBID

Machine teaching

Visualiser le résultat

Conclusion

On entraîne à partir des données annotées

ligne de commande

```
$ mvn generate-resources -P train_lexicalEntries -e
```

Et on passe au modèle suivant

Les catalogues et GROBID

Short Name (U ABC)

GROBID

Encodage TEI

Intaller GROBID

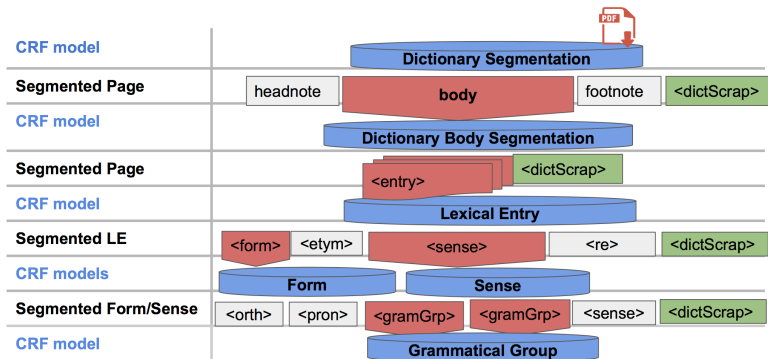
Machine teaching

Visualiser le résultat

Conclusion

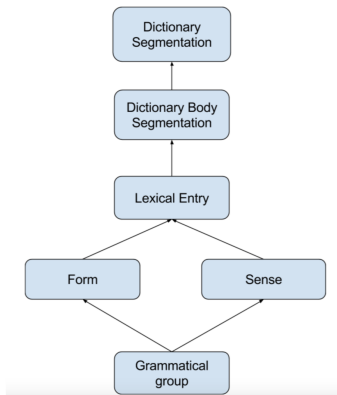
Quatrième niveau d'entraînement

Form



Form

On va traiter séparément le niveau *Form* et le niveau *Sense*, l'un après l'autre



Form

L'objectif est de séparer les différentes informations dans `<form>` (nom et informations biographiques)

```
1 ▾ <entry>
2     <num>54</num>
3 ▾   <form type="lemma">
4       <surname>Lassalle</surname>
5       <addName>(A.-Ch.-L. de)</addName>,
6       <desc>le plus brillant général de cavalerie des guerres de la République et de
7         l'Empire, né à Metz, tué à la bataille de Wagram</desc>
8     </form>
9     <sense> .-
10         L. a. s. au général Dugua; 1 p. in-f.
11         10 »
12
13         Superbe lettre sur la campagne d'Egypte. Il profite du départ du général Desaix
14         pour lui donner des nouvelles. Desaix lui laisse le commandement de la colonne
15         qui doit poursuivre Mourad-Bey, et qui se compose de 400 hommes de cavalerie, 4
16         pièces de canon et 160 dromadaires. Le général Boyer a, dans une petite
17         affaire, tué 10 mameloucks et 40 arabes, etc.
18     </sense>
19 </entry>
```

Form

On crée les données d'entraînement

ligne de commande

```
$ java -jar  
/grobid/grobid-dictionaries/target/grobid-dictionaries-0.4.3-SNAPSHOT.one-jar.jar -dIn  
resources/dataset/dictionary-segmentation/corpus/pdf/ -dOut resources -exe  
createTrainingForm
```

On annote

`<name>` Foroadé-Larouquette (Ad. de `<name>`), `<desc>` ministre de Napoléon III `<desc>`

`<name>` Garcia (Eugénie `<name>`), `<desc>` célèbre cantatrice italienne `<desc>`

`<name>` Gautier (Théop. `<name>`), `<desc>` le célèbre critique `<desc>`

Entraînement

Les catalogues et GROBID

Short Name (U ABC)

GROBID

Encodage TEI

Intaller GROBID

Machine teaching

Visualiser le résultat

Conclusion

On entraîne à partir des données annotées

ligne de commande

```
$ mvn generate-resources -P train_form -e
```

Et on passe au modèle suivant

Les catalogues et GROBID

Short Name (U ABC)

GROBID

Encodage TEI

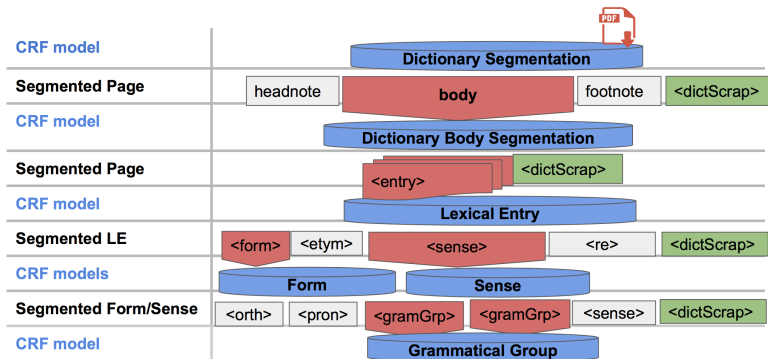
Intaller GROBID

Machine teaching

Visualiser le résultat

Conclusion

Quatrième niveau d'entraînement *bis* *Sense*



Sense

L'objectif est de séparer la description du manuscrit (<def>) des éventuelles notes additionnelles (<note>)

```
1 ▾ <entry>
2     <num>54</num>
3 ▾   <form type="lemma">
4       Lassalle
5       (A.-Ch.-L. de),
6       le plus brillant général de cavalerie des guerres de la République et de
7       l'Empire, né à Metz, tué à la bataille de Wagram
8     </form>
9     <sense> .-
10      <def>
11        L. a. s. au général Dugua; 1 p. in-f.
12        10 »
13      </def>
14      <note>Superbe lettre sur la campagne d'Egypte. Il profite du départ du général
15      Desaix pour lui donner des nouvelles. Desaix lui laisse le commandement de la
16      colonne qui doit poursuivre Mourad-Bey, et qui se compose de 400 hommes de
17      cavalerie, 4 pièces de canon et 160 dromadaires. Le général Boyer a, dans une
18      petite affaire, tué 10 mameloucks et 40 arabes, etc.</note>
19    </sense>
20  </entry>
```

Sense

On crée les données d'entraînement

ligne de commande

```
$ java -jar  
/grobid/grobid-dictionaries/target/grobid-dictionaries-0.4.3-SNAPSHOT.one-jar.jar -dIn  
resources/dataset/dictionary-segmentation/corpus/pdf/ -dOut resources -exe  
createTrainingSense
```

On annote

▷▷▷ L. a. s. à M. Delaunay,
▷▷▷ direct, de l'Artiste, 1/2 p. in-8. 4▷▷ »
▷▷▷ Il invite Jules Janin à venir fumer sa pipe au sortir de la comédie. « Il ne serait
▷▷▷ pas du tout mal reçu ; personne ne lui ferait de sottise, ha ! mon Dieu non !▷▷▷
▷▷▷ Relative à un dessin, vue prise sur le Righi, en Suisse, qu'il exécute pour son
▷▷▷ abonnement à l'Artiste▷▷▷ L. a. s., 1 p. in-8.
▷▷▷ Au sujet de sa pièce d'Henriette. 3▷▷▷ L. a. s., 1848, 1 p. 1/2 in-8. 6▷▷ »

Entraînement

Les catalogues et GROBID

Short Name (U ABC)

GROBID

Encodage TEI

Intaller GROBID

Machine teaching

Visualiser le résultat

Conclusion

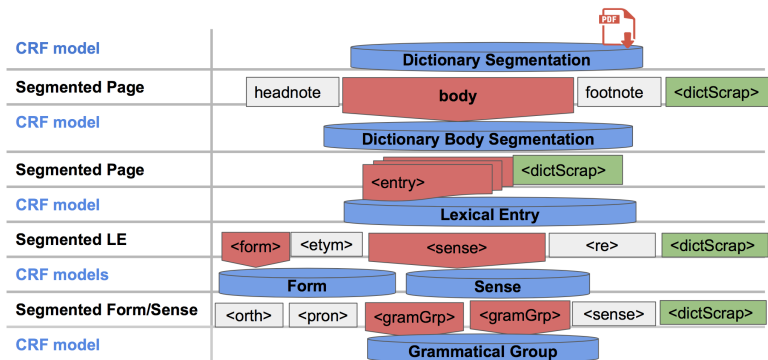
On entraîne à partir des données annotées

ligne de commande

```
$ mvn generate-resources -P train_sense -e
```

Et on passe au modèle suivant

Cinquième niveau d'entraînement (à construire)



Revue des autographes

```
1 ▾ <entry>
2   <num>54</num>
3 ▾   <form type="lemma">
4     <surname>Lassalle</surname>
5     <addName>(A.-Ch.-L. de)</addName>,
6     <desc>le plus brillant général de cavalerie des guerres de la République et de
7       l'Empire, né à Metz, tué à la bataille de Wagram</desc>
8   </form>
9 ▾   <sense> .-
10    <def>
11      <bibl>L. a. s. au général Dugua; 1 p. in-f.</bibl>
12      <num type="price">10 »</num>
13    </def>
14    <note>Superbe lettre sur la campagne d'Egypte. Il profite du départ du général
15      Desaix pour lui donner des nouvelles. Desaix lui laisse le commandement de la
16      colonne qui doit poursuivre Mourad-Bey, et qui se compose de 400 hommes de
17      cavalerie, 4 pièces de canon et 160 dromadaires. Le général Boyer a, dans une
18      petite affaire, tué 10 mameloucks et 40 arabes, etc.</note>
19  </sense>
20 </entry>
```

Grammatical group

Visualiser

Les catalogues et GROBID

Short Name (U ABC)

GROBID

Encodage TEI

Intaller GROBID

Machine teaching

Visualiser le résultat

Conclusion

Une API permet d'utiliser *GROBID dictionaries* une fois qu'il a été entraîné

- ▶ Il faut exécuter le service

ligne de commande

```
mvn -Dmaven.test.skip=true jetty:run-war
```

- ▶ Et aller à l'adresse <http://localhost:8080/> sur son navigateur

Les catalogues et GROBID

Short Name (U ABC)

GROBID

Encodage TEI

Intaller GROBID

Machine teaching

Visualiser le résultat

Conclusion

On utilise *GROBID dictionaries* via l'API

The screenshot shows a web browser window with the title "Grobid Dictionaries Web Applic: X". The address bar shows "localhost:8080". The page title is "GROBID-Dictionaries". The navigation menu includes "About", "Dictionary services", "Bibliography services", "Admin", and "Doc".

The main content area has a "Service to call" dropdown menu set to "Parse Lexical Entries". Below it is a text input field containing "D:\train_set_1(N025).pdf" with "Change" and "Remove" buttons. There are "Submit Query" and "Download TEI Result" buttons.

The output area displays the following XML code:

```
<?xml version="1.0" encoding="UTF-8"?>
<TEI
  xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
    <encodingDesc>
      <appInfo>
        <application version="0.4.2-SNAPSHOT" ident="GROBID" when="2018-12-08T11:52+0000">
          <ref target="https://github.com/medkhem/grobid-dictionaries">GROBID-Dictionaries - A machine learning software for structuring digitized dictionaries</ref>
        </application>
      </appInfo>
    </encodingDesc>
    <fileDesc>
      <titleStmnt>
        <title level="a" type="main"/>
      </titleStmnt>
    </fileDesc>
  </teiHeader>
```

Conclusion

Problèmes

Les catalogues et GROBID

Short Name (U ABC)

GROBID

Encodage TEI

Intaller GROBID

Machine teaching

Visualiser le résultat

Conclusion

Quelques problèmes à résoudre

- ▶ Grobid s'appuie sur la mise-en-page (gras, italique, retour à la ligne. . .)
- ▶ Il s'appuie aussi sur le texte: point, virgule, cadratin. . .
- ▶ Cela signifie qu'il lui faut un bon système d'OCR, car la qualité de l'OCRisation du PDF est cruciale
- ▶ Mais quel OCR?
- ▶ En plus de la question de l'entrée des données se pose celle de la sortie. Étant donnée la simplicité du schéma, ce problème est le moins complexe
- ▶ Simplifier l'utilisation au moyen d'un *pipeline*?

Aller plus loin

Les catalogues et GROBID

Short Name (U ABC)

GROBID

Encodage TEI

Intaller GROBID

Machine teaching

Visualiser le résultat

Conclusion

Nos données d'entraînement sont disponibles en ligne

- ▶ <https://github.com/gabays/grobid>

Télécharger GROBID

- ▶ <https://github.com/MedKhem/grobid-dictionaries>

Utiliser GROBID

- ▶ https://github.com/MedKhem/grobid-dictionaries/wiki/Docker_Instructions
- ▶ <https://github.com/MedKhem/grobid-dictionaries/wiki/How-to-Annotate>

Suivez-nous

Les catalogues et GROBID

Short Name (U ABC)

GROBID

Encodage TEI

Intaller GROBID

Machine teaching

Visualiser le résultat

Conclusion

Suivez-nous

- ▶ <http://editiones-hypotheses.org>
- ▶ https://twitter.com/e_ditiones

Bibliographie

- ▶ Mohamed Khemakhem, Luca Foppiano, Laurent Romary, "Automatic Extraction of TEI Structures in Digitized Lexical Resources using Conditional Random Fields", *electronic lexicography, eLex 2017*, Leiden (Netherlands), [<https://hal.archives-ouvertes.fr/hal-01508868>]
- ▶ Mohamed Khemakhem, Axel Herold, Laurent Romary, "Enhancing Usability for Automatically Structuring Digitised Dictionaries", *GLOBALEX workshop at LREC 2018*, Miyazaki (Japan) [<https://hal.archives-ouvertes.fr/hal-01708137>]
- ▶ Mohamed Khemakhem, Laurent Romary, Simon Gabay, Hervé Bohbot, Francesca Frontini, Giancarlo Luxardo, "Automatically Encoding Encyclopedic-like Resources in TEI", *TEI 2018*, Tokyo (Japan) [<https://hal.inria.fr/hal-01819505>]