# Le projet AGODA. Annoter et publier les débats parlementaires français de la fin du XIX e siècle : défis et solutions

Nicolas Bourgeois, Fanny Lebreton, Aurélien Pellet, Marie Puren, Pierre Vernus

# LE PROJET AGODA

## Annoter et publier les débats parlementaires français de la fin du XIX$^{e}$ siècle : défis et solutions

Nicolas Bourgeois[1], Fanny Lebreton[2], Aurélien Pellet[1], Marie Puren[1 3], Pierre Vernus[4 5]

17 juin 2022 - Présentation des projets AGODA et Gallicorpora, Bibliothèque nationale de France (Paris)

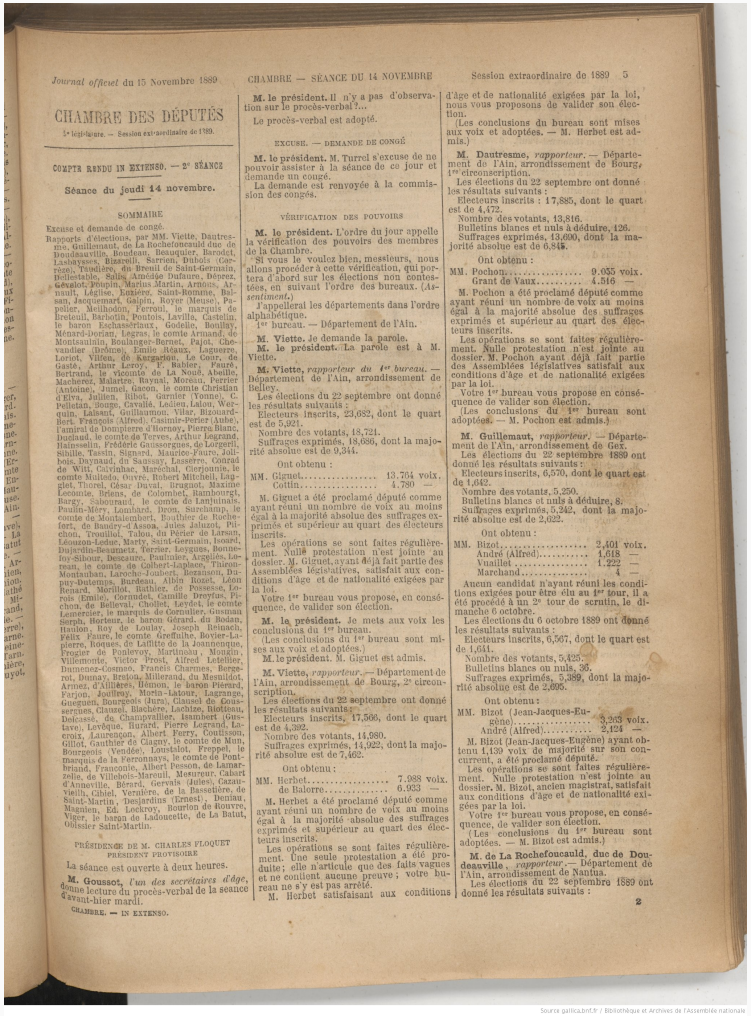
[1]MNSHS, Epitech [2]Ecole nationale des chartes [3]CJM, Ecole nationale des chartes [4] LARHRA [5]Université Lyon 2

# PRÉSENTATION DU PROJET

# LE PROJET AGODA

- AGODA : **A**nalyse sémantique et **G**raphes relationnels pour l'**O**uverture et l'étude des **D**ébats à l'**A**ssemblée nationale
- Collaboration entre Epitech (MNSHS), l'Université Lumière Lyon 2 (LARHRA) et Inria (ALMAnaCH)

# LES DÉBATS PARLEMENTAIRES DURANT LA TROISIÈME RÉPUBLIQUE



*Journal officiel* du 15 Novembre 1889 CHAMBRE — SÉANCE DU 14 NOVEMBRE Session extraordinaire de 1889 5

CHAMBRE DES DÉPUTÉS

1re législature. — Session extraordinaire de 1889.

COMPTE RENDU IN EXTENSO. — 2e SÉANCE

Séance du jeudi 14 novembre.

SOMMAIRE

Excuse et demande de congé.

Rapports d'élections, par MM. Viette, Dautresme, Guillemaut, de La Rochefoucauld duc de Doudeauville, Boudeau, Beauquier, Barodet, Lasbaysses, Rizerolt, Sarrien, Dubuis (Cantal), Tondière, du Breuil de Saint-Germain, Bellestable, Sans, Amédée Dufaure, Dèpres, Gévelot, Poupin, Marius Martin, Arnous, Armault, Lagisse, Ramirez, Saint-Romme, Bassan, Jacquemart, Galpin, Royer (Meuse), Papolier, Meilhodon, Ferroul, le marquis de Breteuil, Barbetin, Poutois, Lavilla, Castelin, le baron Eschasseriaux, Godelle, Boulay, Ménard-Dorian, Legras, le comte Armand, de Montsaulnin, Boulanger-Bernet, Pajot, Chevandier (Drôme), Emile Réaux, Laguerre, Lorel, Viton, de Kergariou, Le Cour, de Gasté, Arthur Leroy, P. Babier, Fauré, Bertrand, le vicomte de La Nouë, Abeille, Macherez, Malartre, Raynal, Moreau, Perrier (Antoine), Junel, Gicon, le comte Christian d'Elva, Julien, Ribot, Garnier (Yonne), C. Pelletan, Goupy, Cavaillé, Leclerc-Lalon, Vercquin, Laisant, Guillaumou, Vilar, Bizouard-Bert, François (Alfred), Casimir-Perier (Aube), l'amiral de Dompierre d'Hornoy, Pierre Blanc, Duclaud, le comte de Torres, Arthur Legrand, Hainseliin, Frédéric Gaussorgues, de Lorgeril, Sibille, Tassin, Signard, Maurice-Faure, Jolibois, Duvaud, du Saussay, Lasserre, Conrad de Wit, Calvinhac, Maréchal, Clerjounie, le comte Mathieu, Ouvré, Robert Mitchell, Laggiet, Thorel, César Duval, Brugnot, Maxime Lecomte, Brians, de Colombet, Rambourgt, Rargy, Jahorrand, le comte de Lanjuinais, Paulin-Méry, Lombard, Dron, Surchamp, le comte de Montalembert, Boutiller de Rochefort, de Baudry-d'Asson, Jules Jaluzot, Pichon, Trouillot, Talou, du Périer de Larsan, Léouzon-Leduc, Marty, Saint-Germain, Isnard, Dujardin-Beaumetz, Terrier, Leygues, Bonnefoy-Sibour, Descaure, Faulmier, Argelès, Lerreau, le comte de Gilbert-Lapisse, Thieon-Montauban, Larroche-Joubert, Rezanson, Dupuy-Dutemps, Burdeau, Albin Rozet, Léon Renard, Morillot, Rathier, de Possesse, Lerois (Emile), Cormdet, Camille Dreyfus, Fichon, de Bétleval, Chollet, Leydet, le comte Lemercier, le marquis de Cornulier, Cousmaa Sirph, Horteur, le baron Gérard, du Bodan, Hanlon, Roy de Loulay, Joseph Reinach, Félix Faure, le comte Greffulhe, Bovier-Lapierre, Rogues, de Lafitte de la Joannequie, Frezier de Fontevoy, Martinaud, Mangin, Villemonte, Victor Prost, Alfred Letellier, Dumesnil-Cosmao, Francis Charmes, Bergerot, Dunay, Breton, Millerand, du Mesnildot, Arnez, d'Ulmès, Démon, le baron Pierard, Farjon, Joulleoy, Morin-Latour, Lagrange, Gueguen, Bourgeois (Jura), Clausel de Coussergues, Clauzel, Blachère, Lachize, Rotteau, Delcassé, de Champvallier, Isambert (Gustave), Levêque, Rurard, Pierre Legrand, Lecroix, Laurençon, Albert Ferry, Coutisson, Gillot, Gauthier de Clagny, le comte de Mun, Bourgeois (Vendée), Lonstalot, Freppel, le marquis de la Ferronnays, le comte de Pontbriand, Francoulie, Albert Pesson, de Lamarzelle, de Villebois-Mareuil, Mesureur, Cabart d'Anneville, Bérard, Gervais (Julien), Cazauvieilh, Chiet, Verninre, de la Bassetière, de Saint-Martin, Desjardins (Ernest), Denian, Magnien, Ed. Lockroy, Pupine de Rouvre, Vigor, le baron de Ladoucette, de La Batut, Odiisier Saint-Martin.

PRÉSIDENCE DE M. CHARLES FLOQUET
PRÉSIDENT PROVISOIRE

La séance est ouverte à deux heures.

**M. Gousset,** *l'un des secrétaires d'âge,* donne lecture du procès-verbal de la séance d'avant-hier mardi.

CHAMBRE. — IN EXTENSO.

**M. le président.** Il n'y a pas d'observation sur le procès-verbal ?...

Le procès-verbal est adopté.

EXCUSE. — DEMANDE DE CONGÉ

**M. le président.** M. Turrel s'excuse de ne pouvoir assister à la séance de ce jour et demande un congé.

La demande est renvoyée à la commission des congés.

VÉRIFICATION DES POUVOIRS

**M. le président.** L'ordre du jour appelle la vérification des pouvoirs des membres de la Chambre.

Si vous le voulez bien, messieurs, nous allons procéder à cette vérification, qui portera d'abord sur les élections non contestées, en suivant l'ordre des bureaux. (*Assentiment.*)

J'appellerai les départements dans l'ordre alphabétique.

1er bureau. — Département de l'Ain.

**M. Viette.** Je demande la parole.

**M. le président.** La parole est à M. Viette.

**M. Viette,** *rapporteur du 1er bureau.* — Département de l'Ain, arrondissement de Belley.

Les élections du 22 septembre ont donné les résultats suivants :

Electeurs inscrits, 23,682, dont le quart est de 5,921.

Nombre des votants, 18,721.

Suffrages exprimés, 18,636, dont la majorité absolue est de 9,344.

Ont obtenu :

MM. Gignet..... 13,764 voix.
Cottin..... 4,780 —

M. Gignet a été proclamé député comme ayant réuni un nombre de voix au moins égal à la majorité absolue des suffrages exprimés et supérieur au quart des électeurs inscrits.

Les opérations se sont faites régulièrement. Nulle protestation n'est jointe au dossier. M. Gignet, ayant déjà fait partie des Assemblées législatives, satisfait aux conditions d'âge et de nationalité exigées par la loi.

Votre 1er bureau vous propose, en conséquence, de valider son élection.

**M. le président.** Je mets aux voix les conclusions du 1er bureau.

(Les conclusions du 1er bureau sont mises aux voix et adoptées.)

**M. le président.** M. Gignet est admis.

**M. Viette,** *rapporteur.* — Département de l'Ain, arrondissement de Bourg, 2e circonscription.

Les élections du 22 septembre ont donné les résultats suivants :

Electeurs inscrits, 17,506, dont le quart est de 4,282.

Nombre des votants, 14,980.

Suffrages exprimés, 14,922, dont la majorité absolue est de 7,462.

Ont obtenu :

MM. Herbet..... 7,988 voix.
de Balorre..... 6,933 —

M. Herbet a été proclamé député comme ayant réuni un nombre de voix au moins égal à la majorité absolue des suffrages exprimés et supérieur au quart des électeurs inscrits.

Les opérations se sont faites régulièrement. Une seule protestation a été produite; elle n'articule que des faits vagues et ne contient aucune preuve ; votre bureau ne s'y est pas arrêté.

M. Herbet satisfaisant aux conditions d'âge et de nationalité exigées par la loi, nous vous proposons de valider son élection.

(Les conclusions du bureau sont mises aux voix et adoptées. — M. Herbet est admis.)

**M. Dautresme,** *rapporteur.* — Département de l'Ain, arrondissement de Bourg, 1re circonscription.

Les élections du 22 septembre ont donné les résultats suivants :

Electeurs inscrits : 17,885, dont le quart est de 4,472.

Nombre des votants, 13,816.

Bulletins blancs et nuls à déduire, 126.

Suffrages exprimés, 13,690, dont la majorité absolue est de 6,845.

Ont obtenu :

MM. Pochon..... 9,055 voix.
Grant de Vaux..... 4,516 —

M. Pochon a été proclamé député comme ayant réuni un nombre de voix au moins égal à la majorité absolue des suffrages exprimés et supérieur au quart des électeurs inscrits.

Les opérations se sont faites régulièrement. Nulle protestation n'est jointe au dossier. M. Pochon ayant déjà fait partie des Assemblées législatives satisfait aux conditions d'âge et de nationalité exigées par la loi.

Votre 1er bureau vous propose en conséquence de valider son élection.

(Les conclusions du 1er bureau sont adoptées. — M. Pochon est admis.)

**M. Guillemaut,** *rapporteur.* — Département de l'Ain, arrondissement de Gex.

Les élections du 22 septembre 1889 ont donné les résultats suivants :

Electeurs inscrits, 6,570, dont le quart est de 1,642.

Nombre des votants, 5,250.

Bulletins blancs et nuls à déduire, 8.

Suffrages exprimés, 5,242, dont la majorité absolue est de 2,622.

Ont obtenu :

MM. Bizot..... 3,401 voix.
André (Alfred)..... 1,618 —
Vuaillet..... 1,222 —
Marchand..... 1 —

Aucun candidat n'ayant réuni les conditions exigées pour être élu au 1er tour, il a été procédé à un 2e tour de scrutin, le dimanche 6 octobre.

Les élections du 6 octobre 1889 ont donné les résultats suivants :

Electeurs inscrits, 6,567, dont le quart est de 1,641.

Nombre des votants, 5,425.

Bulletins blancs ou nuls, 36.

Suffrages exprimés, 5,389, dont la majorité absolue est de 2,695.

Ont obtenu :

MM. Bizot (Jean-Jacques-Eugène)..... 3,263 voix.
André (Alfred)..... 2,124 —

M. Bizot (Jean-Jacques-Eugène) ayant obtenu 1,139 voix de majorité sur son concurrent, a été proclamé député.

Les opérations se sont faites régulièrement. Nulle protestation n'est jointe au dossier. M. Bizot, ancien magistrat, satisfait aux conditions d'âge et de nationalité exigées par la loi.

Votre 1er bureau vous propose, en conséquence, de valider son élection.

(Les conclusions du 1er bureau sont adoptées. — M. Bizot est admis.)

**M. de La Rochefoucauld, duc de Doudeauville,** *rapporteur.* — Département de l'Ain, arrondissement de Nantua.

Les élections du 22 septembre 1889 ont donné les résultats suivants :

2

Source gallica.bnf.fr / Bibliothèque et Archives de l'Assemblée nationale

**Figure –** Séance parlementaire du 14 novembre 1889

- Débats à la Chambre des députés (chambre basse du parlement) transcrits en détail dans le Journal officiel de la République française. Débats parlementaires (1881-1940)
- Disponible en ligne via Gallica (bibliothèque numérique de la Bibliothèque nationale de France)
- Difficile de travailler sur ce corpus, pourtant intéressant pour diverses disciplines (histoire, sociologie, science politique, linguistique)

# OBJECTIFS

- Donner plus facilement accès aux retranscriptions anciennes des débats parlementaires
- Faciliter la recherche dans ce corpus
- Permettre la constitution de sous-corpus
- Offrir de nouveaux modes de visualisation des documents

# RÉSULTATS ATTENDUS

- Créer une plateforme de consultation
- Produire des données textuelles structurées et sémantiquement enrichies à partir de ces débats numérisés
- Contribuer à la conception d'un workflow adapté à la préparation, à la publication et à l'analyse de gros corpus de documents historiques

# UNE PREUVE DE CONCEPT

Traitement d'une sous-partie du corpus : législature 1889-1893 soit **10418 images à traiter**

- Renouvellement partiel du personnel politique (boulangisme et scandale de Panama)
- Premières manifestations du Ralliement des catholiques à la République
- Tournant de la politique douanière (lois Méline)
- Essor du socialisme et du syndicalisme (Fourmies)
- Premiers attentats anarchistes

# WORKFLOW


océriser
Annoter et enrichir sémantiquement le corpus
Publier (TEI publisher)
Extraire et Analyser

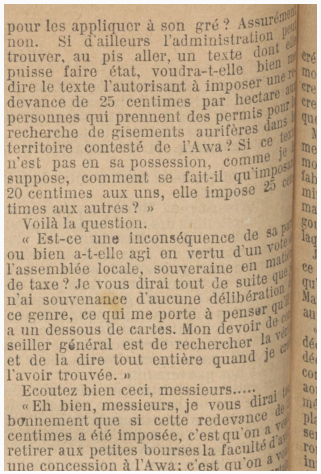
# OCÉRISER LES DÉBATS

# LE CAS DES DÉBATS PARLEMENTAIRES

- Récupération des textes océrisés via API Document de Gallica => qualité inégale de l'OCR
- Erreurs dues à :
  - qualité du document : tâches et surimpression
  - la **courbure de la page** au niveau de la reliure


**Figure –** Evaluation de la qualité de l'OCR fourni par Gallica

# EFFET DE LA COURBURE SUR LES RÉSULTATS DE L'OCR



pour les appliquer à son gré ? Assure.

non. Si d'ailleurs l'administration fj, trouver, au pis aller, un texte (1011t eu puisse faire état, voudra-t-elle dire le texte l'autorisant à imposer u devance de 25 centimes par hectar , personnes qui prennent des permis Pour1 recherche de gisements aurifères a teSI territoire contesté de l'Awa ? Si ce tes

n'est pas en sa possession, cornas le suppose, comment se fait-il qulIWjs3j 20 centimes aux uns, elle impose 20 cet times aux autres ? » , Voilà la question. é

« Est-ce une inconséquence de sa te ou bien a-t-elle agi en vertu d'un v l'assemblée locale, souveraine en OlqallC de taxe? Je vous dirai tout de suite Il n'ai souvenance d'aucune délibératl°ll'V ce genre, ce qui me porte à penser a un dessous de cartes. Mon devoir seiller général est de rechercher la cJ'!" et de la dire tout entière quand Je l'avoir trouvée. »

Ecoutez bien ceci, messieurs. l «Eh bien, messieurs, je vous dIra de bonnement que si cette redevance J centimes a été imposée, c'est qu'on a, retirer aux petites bourses la faculté" une concession à l'Awa; c'est qu'on a r¡;:i\

**Figure –** 20 octobre 1890 (p.1718)

**Figure –** Résultat de l'OCR

## Décision de ré-océriser le corpus

# Conséquences sur la chaîne de traitement

1. Segmentation incorrecte
2. OCR moins performant
3. Texte très courbé et même coupé : peu lisible par un être humain, encore moins lisible par une machine
4. Annotation automatique en XML-TEI plus difficile car liée à la segmentation
5. Texte trop fautif : impossible à publier tel quel en ligne
6. Influence négative des erreurs de l'OCR sur la performance des analyses TAL (NER, topic modeling, word embedding) (cf. Assessing the Impact of OCR Quality on Downstream NLP Tasks)

# QUELLES SOLUTIONS ?

# TRAITER LES IMAGES

Améliorer la qualité de l'image avec une méthode de "dewarping" => résultats peu probants

- Gérer la courbure des pages avec le dewarping ?
- Utiliser des outils plus avancés ?

(a) Image d'origine

(b) Image "dewarpée"

**Figure –** Dewarping : pas adapté à nos documents

# OUTIL DE NETTOYAGE DÉVELOPPÉ PAR L'ANR SODUCO

Année 1881 DÉBATS ET DOCUMENTS PARLEMENTAIRES 25 Février [CHAMBRE] 345

CHAMBRE DES DÉPUTÉS
Session ordinaire de 1881

COMPTE RENDU IN EXTENSO. — 23e SÉANCE

Séance du jeudi 24 février 1881

SOMMAIRE

Procès-verbal : M. Devès.
Excuse.
Dépôt, par M. Marquiset, du rapport de la commission chargée d'examiner le projet de loi ayant pour objet de faire déclarer d'utilité publique la construction d'un canal de navigation dans la vallée de la Chiers, entre Longwy et la Meuse, à Mouzon.
Dépôt, par M. Morel, d'un rapport de la 2e commission d'intérêt local, sur le projet de loi tendant à autoriser le département de la Haute-Marne à contracter un emprunt pour l'amélioration de l'asile des aliénés de Saint-Dizier.
Dépôt, par M. Sarrien, du rapport de la commission chargée d'examiner le projet de loi ayant pour objet de faire déclarer d'utilité publique l'amélioration du canal du centre.
Suite de la 1re délibération sur le projet de loi, adopté par le Sénat, concernant l'administration de l'armée : M. le comte de Roys, rapporteur. — Ajournement à samedi.
Adoption en 2e délibération du projet de loi ayant pour objet l'établissement d'un canal de Dombasle à Saint-Dié.
Discussion des conclusions du rapport de la 2e commission des pétitions sur la pétition du sieur Jean Dupuis, citoyen français, demeurant à Han-Kow (Chine) : M. Emile Bouchet, rapporteur. Adoption.
1re délibération sur les propositions de loi : 1° de M. Germain Casse et de plusieurs de ses collègues, tendant à régler les rapports entre les compagnies de chemins de fer et leurs agents commissionnés ; 2° de M. le baron de Janzé, relative à la réglementation des compagnies de chemins de fer : MM. René Brice, le baron de Janzé.
Règlement de l'ordre du jour : MM. Latrade, Serrette, de Gasté, Bardoux.
Reprise de la discussion des propositions tendant à régler les rapports entre les compagnies de chemins de fer et leurs agents commissionnés : M. Trarieux.
Présentation, par M. le ministre de l'intérieur et des cultes, d'un projet de loi tendant à autoriser le département des Côtes-du-Nord à contracter deux emprunts pour les travaux des chemins vicinaux ordinaires.
Demande, par M. le ministre de la marine, de discussion immédiate du projet de loi, adopté par la Chambre des députés, adopté avec modifications par le Sénat, ayant pour objet d'ouvrir au ministre de la marine et des colonies un crédit de 8,352,751 fr., sur l'exercice 1881, pour les chemins de fer du Haut-Sénégal. Adoption. = Article 1er : M. Des Rotours. Adoption. = Adoption de l'article 2. = Adoption, au scrutin, de l'ensemble du projet de loi.
Règlement de l'ordre du jour : MM. le président, de Gasté.
Présentation par M. le sous-secrétaire d'Etat des travaux publics :
1° D'un projet de loi ayant pour objet : 1° L'incorporation dans le réseau d'intérêt général du chemin de fer d'intérêt local de Mézidon à Dives ; 2° l'approbation d'une convention passée avec la compagnie des chemins de fer de la vallée d'Auge pour le rachat de ladite ligne ;
2° D'un projet de loi ayant pour objet : 1° De déclarer d'utilité publique l'établissement, dans le département des Landes, de plusieurs chemins de fer d'intérêt local et d'en autoriser l'exécution ; 2° d'approuver un traité passé entre la compagnie du Midi et les concessionnaires de ces chemins.
Dépôt par M. Cavallié d'un rapport de la 2e commission d'intérêt local sur le projet de loi tendant à autoriser le département de la Haute-Saône à contracter un emprunt pour les travaux des chemins vicinaux.

PRÉSIDENCE DE M. GAMBETTA

La séance est ouverte à deux heures.
M. Frémiet, *l'un des secrétaires*, donne lecture du procès-verbal de la séance de mardi dernier.
M. le président. La parole est à M. Devès sur le procès-verbal.
M. Paul Devès. Messieurs, MM. Tassin, de Sonnier et Deniau auraient voté contre la proposition d'enquête faite par M. Lorgé s'ils eussent assisté à la séance de mardi. Je le déclare en leur nom.
M. le président. Il n'y a pas d'autre observation ?...
Le procès-verbal est adopté.

EXCUSE

M. le président. M. Dremel s'excuse de ne pouvoir assister à la séance de ce jour.

DÉPÔT DE RAPPORTS

M. le président. Je donne la parole à ceux de MM. les députés qui ont des rapports à déposer.
M. Sarrien. Au nom de la commission des voies navigables, j'ai l'honneur de déposer sur le bureau de la Chambre un rapport sur le projet de loi ayant pour objet de faire déclarer d'utilité publique l'amélioration du canal du Centre.
M. Marquiset. J'ai l'honneur de déposer sur le bureau de la Chambre, au nom de la commission des voies navigables, un rapport sur le projet de loi ayant pour objet de faire déclarer d'utilité publique la construction d'un canal de navigation dans la vallée de la Chiers, entre Longwy et la Meuse, à Mouzon.
M. Morel. Au nom de la 2e commission d'intérêt local, j'ai l'honneur de déposer sur le bureau de la Chambre un rapport sur le projet de loi tendant à autoriser le département de la Haute-Marne à contracter un emprunt pour l'amélioration de l'asile des aliénés de Saint-Dizier.
M. le président. Ces rapports seront imprimés et distribués.

AJOURNEMENT DE LA SUITE DE LA DISCUSSION DU PROJET DE LOI SUR L'ADMINISTRATION DE L'ARMÉE.

M. le président. L'ordre du jour appelle la suite de la première délibération sur le projet de loi, adopté par le Sénat, concernant l'administration de l'armée.
M. le rapporteur a la parole.
M. le comte de Roys, *rapporteur*. Messieurs, M. le ministre de la guerre étant appelé aujourd'hui au Sénat par une interpellation, la commission vous prie de vouloir bien ajourner à la séance de samedi la suite de la discussion sur l'administration de l'armée, en lui maintenant son rang en tête de l'ordre du jour de cette séance.
M. le président. Il n'y a pas d'opposition ?...
L'ajournement à samedi est prononcé.

DISCUSSION DU PROJET DE LOI CONCERNANT LE CANAL DE DOMBASLE A SAINT DIÉ

M. le président. L'ordre du jour appelle la deuxième délibération sur le projet de loi ayant pour objet l'établissement d'un canal de Dombasle à Saint-Dié.
Art. 1er. — Sont déclarés d'utilité publique les travaux à faire pour la construction du canal latéral à la Meurthe, de Dombasle à Saint-Dié, et son raccordement par voie ferrée avec la gare du chemin de fer de l'Est à Saint-Dié, conformément aux dispositions de l'avant-projet adopté par le conseil général des ponts et chaussées dans ses délibérations des 7 novembre 1878 et 29 janvier 1880. »
(L'article 1er est mis aux voix et adopté.)
« Art. 2. — Un péage dont le produit est garanti, comme il est dit à l'article 3, par les départements des Vosges et de Meurthe-et-Moselle, sera établi sur la voie navigable, au fur et à mesure de la mise en exploitation de ses diverses sections.
« Ce péage se composera, par tonne, d'une constante de 0.30 et d'un élément variable de 0.0005 par kilomètre.
« Il prendra fin vingt ans après l'ouverture du canal sur toute sa longueur. » — (Adopté.)
« Art. 3. — Le ministre des travaux publics est autorisé à accepter, au nom de l'Etat :
« 1° Les subventions souscrites par les industriels intéressés à titre de libération du péage ;
« 2° L'offre faite par les deux départements de Meurthe-et-Moselle et des Vosges, ainsi qu'il résulte des délibérations des conseils généraux en date des 22 et 23 août 1879, de garantir solidairement et à parts égales les subventions précitées ainsi qu'un produit annuel de 140,000 francs pour le péage à partir de la mise en exploitation du canal sur toute sa longueur, sauf déduction, au taux de 4 1/2 p. 100, de l'annuité d'amortissement en vingt ans de ces subventions ;
« 3° Les offres de concours des villes de Lunéville, Baccarat, Raon l'Etape, Saint-Dié et Fraize montant ensemble à la somme de 510,000 francs. » — (Adopté.)
« Art. 4. — La somme de 20 millions de francs à laquelle sont évalués les travaux sera imputée sur les ressources extraordinaires inscrites au budget de chaque année.
« Viendra en déduction de la dépense à la charge du Trésor le montant des subventions mentionnées à l'article précédent. » — (Adopté.)
(L'ensemble du projet de loi est mis aux voix et adopté.)

DISCUSSION D'UN RAPPORT DE LA 2e COMMISSION DES PÉTITIONS

M. le président. L'ordre du jour appelle la discussion des conclusions du rapport fait au nom de la 2e commission des pétitions sur la pétition du sieur Jean Dupuis, citoyen français, demeurant à Han-Kow (Chine).
La parole est à M. le rapporteur.
M. Emile Bouchet, *rapporteur*. Messieurs, sur cette question qui aurait comporté de longs développements, je n'aurai que très-peu de mots à prononcer : je demande simplement à la Chambre de vouloir bien accueillir les conclusions de la commission, tendant au renvoi de la pétition au ministre de la marine, conclusions acceptées par M. le ministre de la marine et par le Gouvernement.
M. de Mahy. Très-bien ! très-bien !
M. La Vieille. Quelles sont ces conclusions ?
M. le président. La commission conclut au renvoi de la pétition au ministre de la marine ; le renvoi est accepté par le Gouvernement.
M. le rapporteur. Les conclusions sont celles de toutes les commissions de pétitions qui, après avoir examiné les faits, concluent, quand elles les trouvent dignes d'une attention particulière, au renvoi de la pétition au ministre compétent. C'est ce qu'a fait la deuxième commission des pétitions.
M. de Larochefoucauld, duc de Bisaccia. Quelle est cette pétition ?
M. le président. Le rapport a été imprimé et distribué, et la discussion est inscrite au feuilleton depuis deux mois.
M. de Larochefoucauld, duc de Bisaccia. Avant de voter, je demande des explications.
M. le président. La commission conclut au renvoi à M. le ministre de la marine.



(a) Image d'origine

Année 1881 DÉBATS ET DOCUMENTS PARLEMENTAIRES 25 Février [CHAMBRE] 345

CHAMBRE DES DÉPUTÉS
Session ordinaire de 1881

COMPTE RENDU IN EXTENSO. — 23e SÉANCE

Séance du jeudi 24 février 1881

SOMMAIRE

Procès-verbal : M. Devès.
Excuse.
Dépôt, par M. Marquiset, du rapport de la commission chargée d'examiner le projet de loi ayant pour objet de faire déclarer d'utilité publique la construction d'un canal de navigation dans la vallée de la Chiers, entre Longwy et la Meuse, à Mouzon.
Dépôt, par M. Morel, d'un rapport de la 2e commission d'intérêt local, sur le projet de loi tendant à autoriser le département de la Haute-Marne à contracter un emprunt pour l'amélioration de l'asile des aliénés de Saint-Dizier.
Dépôt, par M. Sarrien, du rapport de la commission chargée d'examiner le projet de loi ayant pour objet de faire déclarer d'utilité publique l'amélioration du canal du centre.
Suite de la 1re délibération sur le projet de loi, adopté par le Sénat, concernant l'administration de l'armée : M. le comte de Roys, rapporteur. — Ajournement à samedi.
Adoption en 2e délibération du projet de loi ayant pour objet l'établissement d'un canal de Dombasle à Saint-Dié.
Discussion des conclusions du rapport de la 2e commission des pétitions sur la pétition du sieur Jean Dupuis, citoyen français, demeurant à Han-Kow (Chine) : M. Emile Bouchet, rapporteur. Adoption.
1re délibération sur les propositions de loi : 1° de M. Germain Casse et de plusieurs de ses collègues, tendant à régler les rapports entre les compagnies de chemins de fer et leurs agents commissionnés ; 2° de M. le baron de Janzé, relative à la réglementation des compagnies de chemins de fer : MM. René Brice, le baron de Janzé.
Règlement de l'ordre du jour : MM. Latrade, Serrette, de Gasté, Bardoux.
Reprise de la discussion des propositions tendant à régler les rapports entre les compagnies de chemins de fer et leurs agents commissionnés : M. Trarieux.
Présentation, par M. le ministre de l'intérieur et des cultes, d'un projet de loi tendant à autoriser le département des Côtes-du-Nord à contracter deux emprunts pour les travaux des chemins vicinaux ordinaires.
Demande, par M. le ministre de la marine, de discussion immédiate du projet de loi, adopté par la Chambre des députés, adopté avec modifications par le Sénat, ayant pour objet d'ouvrir au ministre de la marine et des colonies un crédit de 8,352,751 fr., sur l'exercice 1881, pour les chemins de fer du Haut-Sénégal. Adoption. = Article 1er : M. Des Rotours. Adoption. = Adoption de l'article 2. = Adoption, au scrutin, de l'ensemble du projet de loi.
Règlement de l'ordre du jour : MM. le président, de Gasté.
Présentation par M. le sous-secrétaire d'Etat des travaux publics :
1° D'un projet de loi ayant pour objet : 1° L'incorporation dans le réseau d'intérêt général du chemin de fer d'intérêt local de Mézidon à Dives ; 2° l'approbation d'une convention passée avec la compagnie des chemins de fer de la vallée d'Auge pour le rachat de ladite ligne ;
2° D'un projet de loi ayant pour objet : 1° De déclarer d'utilité publique l'établissement, dans le département des Landes, de plusieurs chemins de fer d'intérêt local et d'en autoriser l'exécution ; 2° d'approuver un traité passé entre la compagnie du Midi et les concessionnaires de ces chemins.
Dépôt par M. Cavallié d'un rapport de la 2e commission d'intérêt local sur le projet de loi tendant à autoriser le département de la Haute-Saône à contracter un emprunt pour les travaux des chemins vicinaux.

PRÉSIDENCE DE M. GAMBETTA

La séance est ouverte à deux heures.
M. Frémiet, *l'un des secrétaires*, donne lecture du procès-verbal de la séance de mardi dernier.
M. le président. La parole est à M. Devès sur le procès-verbal.
M. Paul Devès. Messieurs, MM. Tassin, de Sonnier et Deniau auraient voté contre la proposition d'enquête faite par M. Lorgé s'ils eussent assisté à la séance de mardi. Je le déclare en leur nom.
M. le président. Il n'y a pas d'autre observation ?...
Le procès-verbal est adopté.

EXCUSE

M. le président. M. Dremel s'excuse de ne pouvoir assister à la séance de ce jour.

DÉPÔT DE RAPPORTS

M. le président. Je donne la parole à ceux de MM. les députés qui ont des rapports à déposer.
M. Sarrien. Au nom de la commission des voies navigables, j'ai l'honneur de déposer sur le bureau de la Chambre un rapport sur le projet de loi ayant pour objet de faire déclarer d'utilité publique l'amélioration du canal du Centre.
M. Marquiset. J'ai l'honneur de déposer sur le bureau de la Chambre, au nom de la commission des voies navigables, un rapport sur le projet de loi ayant pour objet de faire déclarer d'utilité publique la construction d'un canal de navigation dans la vallée de la Chiers, entre Longwy et la Meuse, à Mouzon.
M. Morel. Au nom de la 2e commission d'intérêt local, j'ai l'honneur de déposer sur le bureau de la Chambre un rapport sur le projet de loi tendant à autoriser le département de la Haute-Marne à contracter un emprunt pour l'amélioration de l'asile des aliénés de Saint-Dizier.
M. le président. Ces rapports seront imprimés et distribués.

AJOURNEMENT DE LA SUITE DE LA DISCUSSION DU PROJET DE LOI SUR L'ADMINISTRATION DE L'ARMÉE.

M. le président. L'ordre du jour appelle la suite de la première délibération sur le projet de loi, adopté par le Sénat, concernant l'administration de l'armée.
M. le rapporteur a la parole.
M. le comte de Roys, *rapporteur*. Messieurs, M. le ministre de la guerre étant appelé aujourd'hui au Sénat par une interpellation, la commission vous prie de vouloir bien ajourner à la séance de samedi la suite de la discussion sur l'administration de l'armée, en lui maintenant son rang en tête de l'ordre du jour de cette séance.
M. le président. Il n'y a pas d'opposition ?...
L'ajournement à samedi est prononcé.

DISCUSSION DU PROJET DE LOI CONCERNANT LE CANAL DE DOMBASLE A SAINT DIÉ

M. le président. L'ordre du jour appelle la deuxième délibération sur le projet de loi ayant pour objet l'établissement d'un canal de Dombasle à Saint-Dié.
Art. 1er. — Sont déclarés d'utilité publique les travaux à faire pour la construction du canal latéral à la Meurthe, de Dombasle à Saint-Dié, et son raccordement par voie ferrée avec la gare du chemin de fer de l'Est à Saint-Dié, conformément aux dispositions de l'avant-projet adopté par le conseil général des ponts et chaussées dans ses délibérations des 7 novembre 1878 et 29 janvier 1880. »
(L'article 1er est mis aux voix et adopté.)
« Art. 2. — Un péage dont le produit est garanti, comme il est dit à l'article 3, par les départements des Vosges et de Meurthe-et-Moselle, sera établi sur la voie navigable, au fur et à mesure de la mise en exploitation de ses diverses sections.
« Ce péage se composera, par tonne, d'une constante de 0.30 et d'un élément variable de 0.0005 par kilomètre.
« Il prendra fin vingt ans après l'ouverture du canal sur toute sa longueur. » — (Adopté.)
« Art. 3. — Le ministre des travaux publics est autorisé à accepter, au nom de l'Etat :
« 1° Les subventions souscrites par les industriels intéressés à titre de libération du péage ;
« 2° L'offre faite par les deux départements de Meurthe-et-Moselle et des Vosges, ainsi qu'il résulte des délibérations des conseils généraux en date des 22 et 23 août 1879, de garantir solidairement et à parts égales les subventions précitées ainsi qu'un produit annuel de 140,000 francs pour le péage à partir de la mise en exploitation du canal sur toute sa longueur, sauf déduction, au taux de 4 1/2 p. 100, de l'annuité d'amortissement en vingt ans de ces subventions ;
« 3° Les offres de concours des villes de Lunéville, Baccarat, Raon l'Etape, Saint-Dié et Fraize montant ensemble à la somme de 510,000 francs. » — (Adopté.)
« Art. 4. — La somme de 20 millions de francs à laquelle sont évalués les travaux sera imputée sur les ressources extraordinaires inscrites au budget de chaque année.
« Viendra en déduction de la dépense à la charge du Trésor le montant des subventions mentionnées à l'article précédent. » — (Adopté.)
(L'ensemble du projet de loi est mis aux voix et adopté.)

DISCUSSION D'UN RAPPORT DE LA 2e COMMISSION DES PÉTITIONS

M. le président. L'ordre du jour appelle la discussion des conclusions du rapport fait au nom de la 2e commission des pétitions sur la pétition du sieur Jean Dupuis, citoyen français, demeurant à Han-Kow (Chine).
La parole est à M. le rapporteur.
M. Emile Bouchet, *rapporteur*. Messieurs, sur cette question qui aurait comporté de longs développements, je n'aurai que très-peu de mots à prononcer : je demande simplement à la Chambre de vouloir bien accueillir les conclusions de la commission, tendant au renvoi de la pétition au ministre de la marine, conclusions acceptées par M. le ministre de la marine et par le Gouvernement.
M. de Mahy. Très-bien ! très-bien !
M. La Vieille. Quelles sont ces conclusions ?
M. le président. La commission conclut au renvoi de la pétition au ministre de la marine ; le renvoi est accepté par le Gouvernement.
M. le rapporteur. Les conclusions sont celles de toutes les commissions de pétitions qui, après avoir examiné les faits, concluent, quand elles les trouvent dignes d'une attention particulière, au renvoi de la pétition au ministre compétent. C'est ce qu'a fait la deuxième commission des pétitions.
M. de Larochefoucauld, duc de Bisaccia. Quelle est cette pétition ?
M. le président. Le rapport a été imprimé et distribué, et la discussion est inscrite au feuilleton depuis deux mois.
M. de Larochefoucauld, duc de Bisaccia. Avant de voter, je demande des explications.
M. le président. La commission conclut au renvoi à M. le ministre de la marine.



(b) Image nettoyée

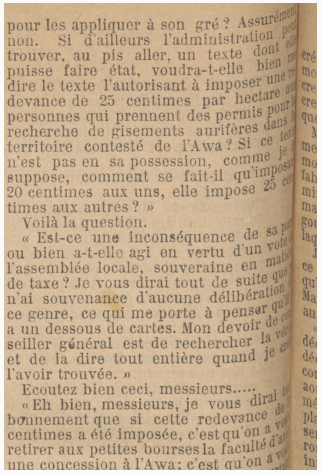
**Figure –** Démonstration de l'outil SODUCO sur une page de débat

# UNE SOLUTION PLUS ADAPTÉE À DÉVELOPPER

- Redresser la courbure => redresser le texte tout entier et peut recourber les parties du texte non courbées.
- Solution envisagée : redresser uniquement la partie de l'image concernée (colonne en bord de reliure)

# UTILISER UN MODÈLE DE LANGUE

Quand le texte n'est pas lisible par l'humain et donc par la machine :


pour les appliquer à son gré ? Assurément
non. Si d'ailleurs l'administration peut
trouver, au pis aller, un texte dont elle
puisse faire état, voudra-t-elle bien me
dire le texte l'autorisant à imposer une re
devance de 25 centimes par hectare aux
personnes qui prennent des permis pour la
recherche de gisements aurifères dans le
territoire contesté de l'Awa ? Si ce terri
n'est pas en sa possession, comme je le
suppose, comment se fait-il qu'imposant
20 centimes aux uns, elle impose 25 cen
times aux autres ? »

Voilà la question.

« Est-ce une inconséquence de sa part
ou bien a-t-elle agi en vertu d'un vote de
l'assemblée locale, souveraine en matière
de taxe ? Je vous dirai tout de suite que je
n'ai souvenance d'aucune délibération de
ce genre, ce qui me porte à penser qu'il
a un dessous de cartes. Mon devoir de con
seiller général est de rechercher la vérité
et de la dire tout entière quand je crois
l'avoir trouvée. »

Ecoutez bien ceci, messieurs.....

« Eh bien, messieurs, je vous dirai fran
chement que si cette redevance de 25
centimes a été imposée, c'est qu'on a voulu
retirer aux petites bourses la faculté d'ar
une concession à l'Awa ; c'est qu'on a voulu

**Figure –** Mots coupés par la forte courbure

- Utiliser un modèle de langue pour « prédire » le ou les mot(s) manquants
- Modèle de langue pour le français CamemBERT
  - … Mais entraîné sur du français (trop ?) contemporain
  - Peut-être nécessaire de réentraîner sur du français de l'époque (utiliser la vérité de terrain, des corpus de presse corrigé…)

# « RESTAURER » LE TEXTE

Deux problèmes récurrents :

- Mots coupés à cause de la courbure
- Mots océrisés mais apparaissant dans le désordre : conséquence de la mauvaise segmentation due à la courbure de la page

Utiliser une IA pour « restaurer » le texte, comme le fait Ithaca de Deep Mind (cf. Restoring and attributing ancient texts using deep neural networks)

# PAS UNE SOLUTION UNIQUE

DES solutions

Meilleure approche = combiner plusieurs approches

# OCÉRISER ET POST-CORRIGER

# « DIVISER » LE CORPUS

Dans le cadre du financement, le traitement de l'image est impossible.

Décision de travailler d'abord sur les numéros les moins affectées :

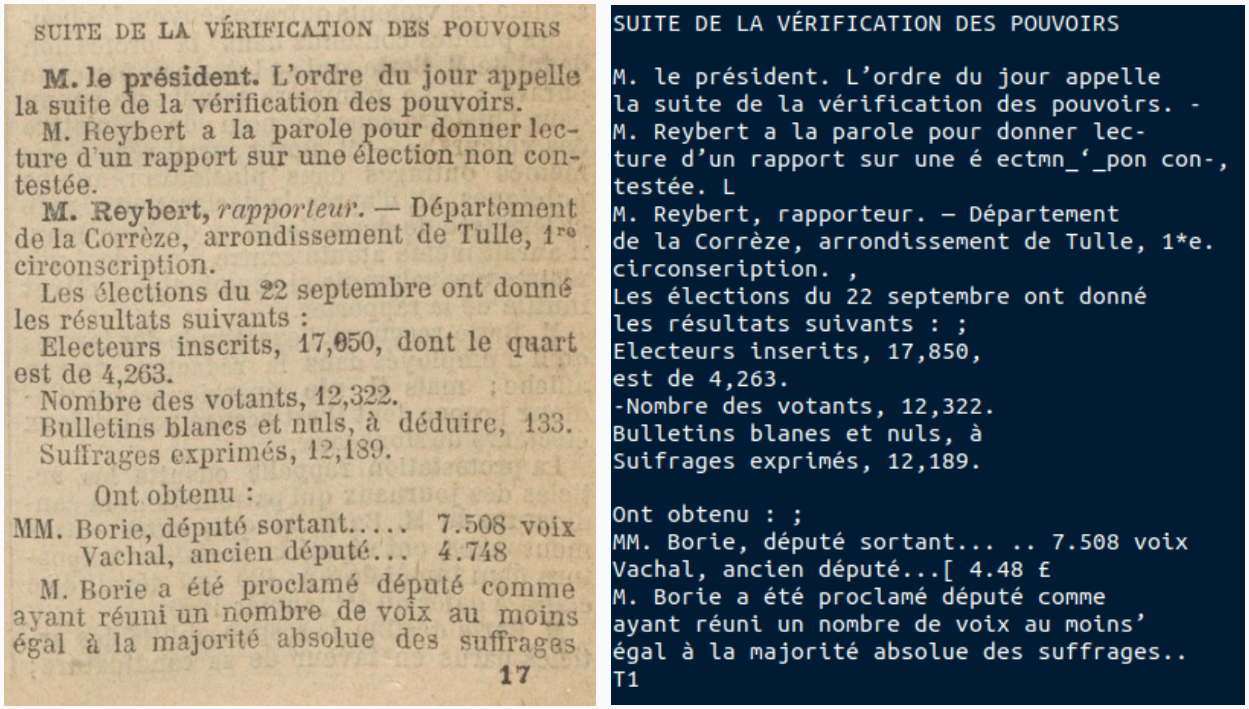
- Océriser ces documents en priorité
- Corriger les textes extraits (cf. post-correction)

Utiliser le taux d'erreur dans l'OCR pour repérer les documents les plus problématiques

# LES SOLUTIONS À NOTRE DISPOSITION

Comparer les performances des moteurs OCR offerts par Huma-Num (Sharedocs) :

- Tesseract (ocr-tesseract ou pytesseract)
- ABBYY FineReader



SUITE DE LA VÉRIFICATION DES POUVOIRS

**M. le président.** L'ordre du jour appelle la suite de la vérification des pouvoirs.

M. Reybert a la parole pour donner lecture d'un rapport sur une élection non contestée.

**M. Reybert,** *rapporteur.* — Département de la Corrèze, arrondissement de Tulle, 1re circonscription.

Les élections du 22 septembre ont donné les résultats suivants :

Electeurs inscrits, 17,050, dont le quart est de 4,263.

Nombre des votants, 12,322.

Bulletins blancs et nuls, à déduire, 133.

Suffrages exprimés, 12,189.

Ont obtenu :

MM. Borie, député sortant..... 7.508 voix
Vachal, ancien député... 4.748

M. Borie a été proclamé député comme ayant réuni un nombre de voix au moins égal à la majorité absolue des suffrages

17

SUITE DE LA VÉRIFICATION DES POUVOIRS

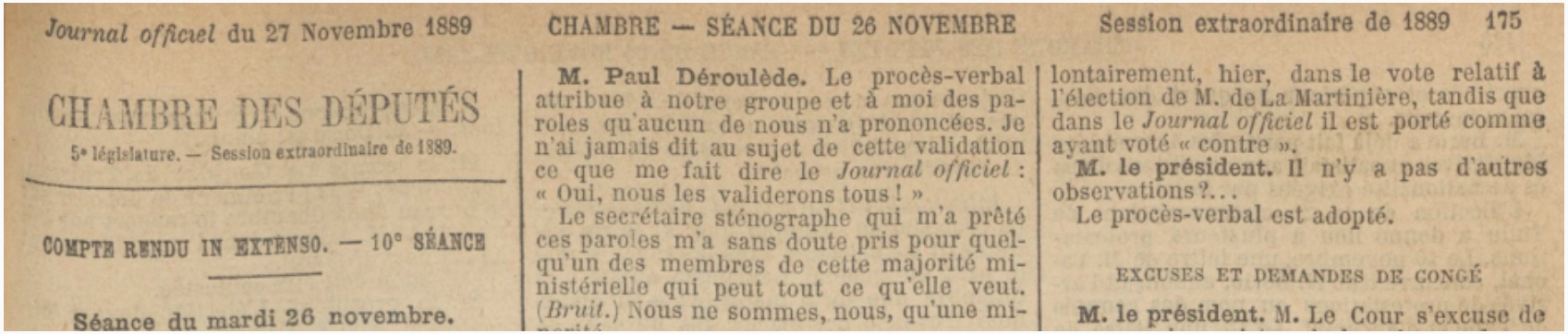
M. le président. L'ordre du jour appelle
la suite de la vérification des pouvoirs. -
M. Reybert a la parole pour donner lec-
ture d'un rapport sur une é ectmn_'_pon con-,
testée. L
M. Reybert, rapporteur. — Département
de la Corrèze, arrondissement de Tulle, 1*e.
circonseription. ,
Les élections du 22 septembre ont donné
les résultats suivants : ;
Electeurs inserits, 17,850,
est de 4,263.
-Nombre des votants, 12,322.
Bulletins blanes et nuls, à
Suifrages exprimés, 12,189.

Ont obtenu : ;
MM. Borie, député sortant... .. 7.508 voix
Vachal, ancien député...[ 4.48 £
M. Borie a été proclamé député comme
ayant réuni un nombre de voix au moins'
égal à la majorité absolue des suffrages..
T1

**(a)** Image d'origine **(b)** OCR

**Figure –** Zoom sur un bloc de texte + OCR (tesseract)

# LES SOLUTIONS À NOTRE DISPOSITION



*Journal officiel* du 27 Novembre 1889

CHAMBRE DES DÉPUTÉS

5e législature. — Session extraordinaire de 1889.

COMPTE RENDU IN EXTENSO. — 10e SÉANCE

Séance du mardi 26 novembre.

CHAMBRE — SÉANCE DU 26 NOVEMBRE

**M. Paul Déroulède.** Le procès-verbal attribue à notre groupe et à moi des paroles qu'aucun de nous n'a prononcées. Je n'ai jamais dit au sujet de cette validation ce que me fait dire le *Journal officiel* : « Oui, nous les validerons tous ! »

Le secrétaire sténographe qui m'a prêté ces paroles m'a sans doute pris pour quelqu'un des membres de cette majorité ministérielle qui peut tout ce qu'elle veut. (*Bruit.*) Nous ne sommes, nous, qu'une mi-

Session extraordinaire de 1889 175

lontairement, hier, dans le vote relatif à l'élection de M. de La Martinière, tandis que dans le *Journal officiel* il est porté comme ayant voté « contre ».

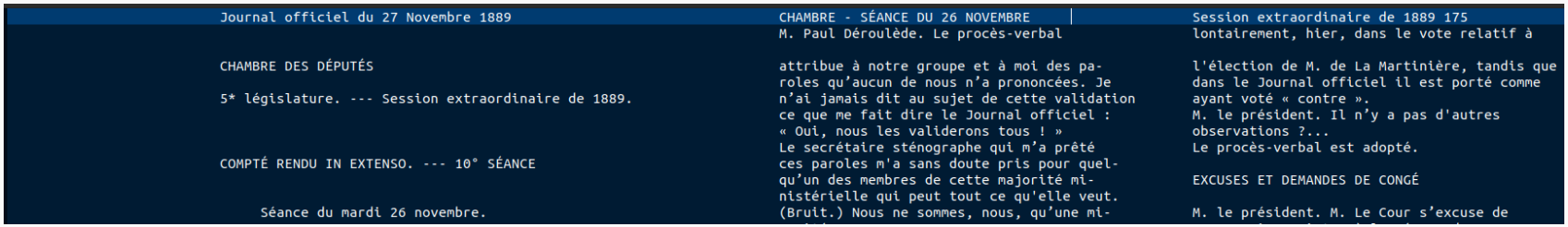
**M. le président.** Il n'y a pas d'autres observations ?...

Le procès-verbal est adopté.

EXCUSES ET DEMANDES DE CONGÉ

**M. le président.** M. Le Cour s'excuse de

(a) Image d'origine



```
Journal officiel du 27 Novembre 1889

CHAMBRE DES DÉPUTÉS

5* législature. --- Session extraordinaire de 1889.

COMPTÉ RENDU IN EXTENSO. --- 10° SÉANCE

    Séance du mardi 26 novembre.

CHAMBRE - SÉANCE DU 26 NOVEMBRE
M. Paul Déroulède. Le procès-verbal

attribue à notre groupe et à moi des pa-
roles qu'aucun de nous n'a prononcées. Je
n'ai jamais dit au sujet de cette validation
ce que me fait dire le Journal officiel :
« Oui, nous les validerons tous ! »
Le secrétaire sténographe qui m'a prêté
ces paroles m'a sans doute pris pour quel-
qu'un des membres de cette majorité mi-
nistérielle qui peut tout ce qu'elle veut.
(Bruit.) Nous ne sommes, nous, qu'une mi-

Session extraordinaire de 1889 175
lontairement, hier, dans le vote relatif à

l'élection de M. de La Martinière, tandis que
dans le Journal officiel il est porté comme
ayant voté « contre ».
M. le président. Il n'y a pas d'autres
observations ?...
Le procès-verbal est adopté.

EXCUSES ET DEMANDES DE CONGÉ

M. le président. M. Le Cour s'excuse de
```

(b) OCR

**Figure –** Zoom sur un bloc de texte + OCR (ABBYY)

# OUTIL DÉVELOPPÉ PAR LRDE (EPITA) - 1

Outil basé sur le moteur d'OCR PERO OCR : très efficace sur les textes historiques

Développé dans le cadre de l'ANR SODUCO

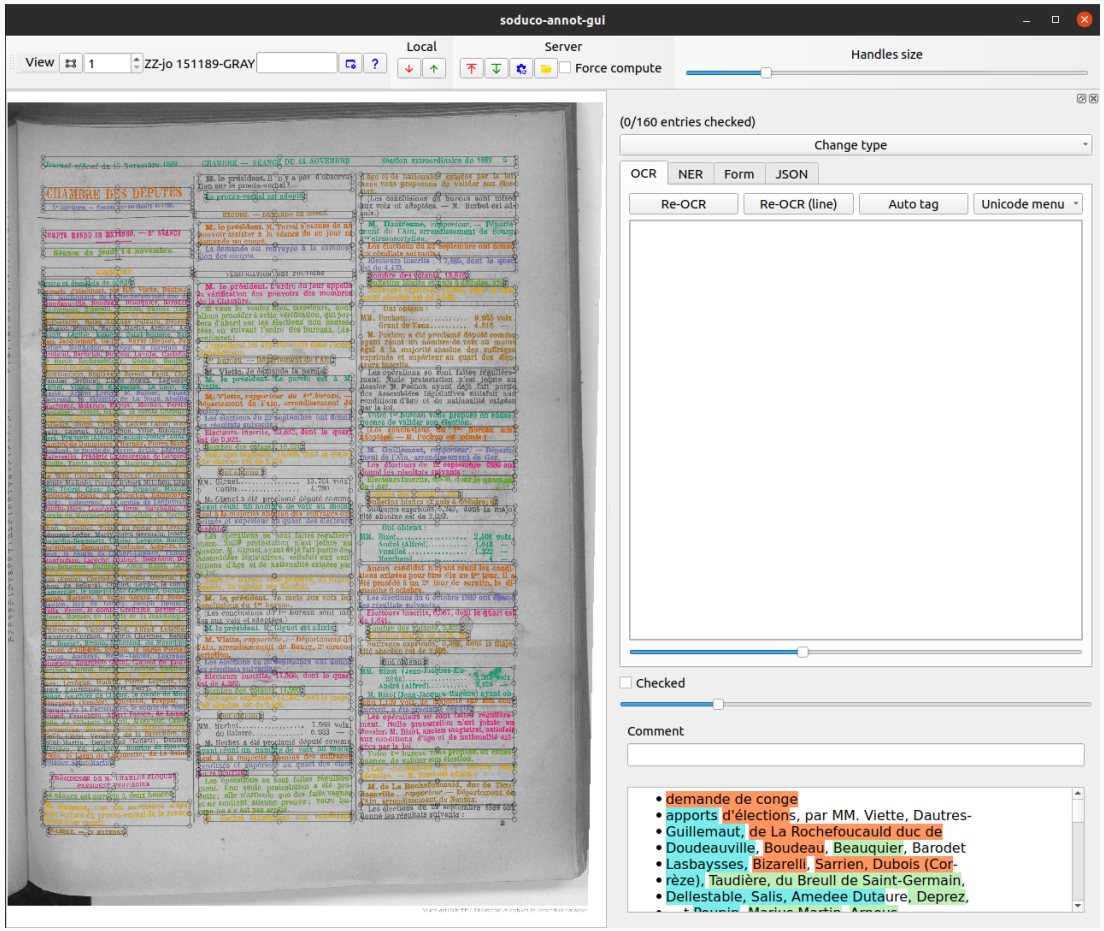
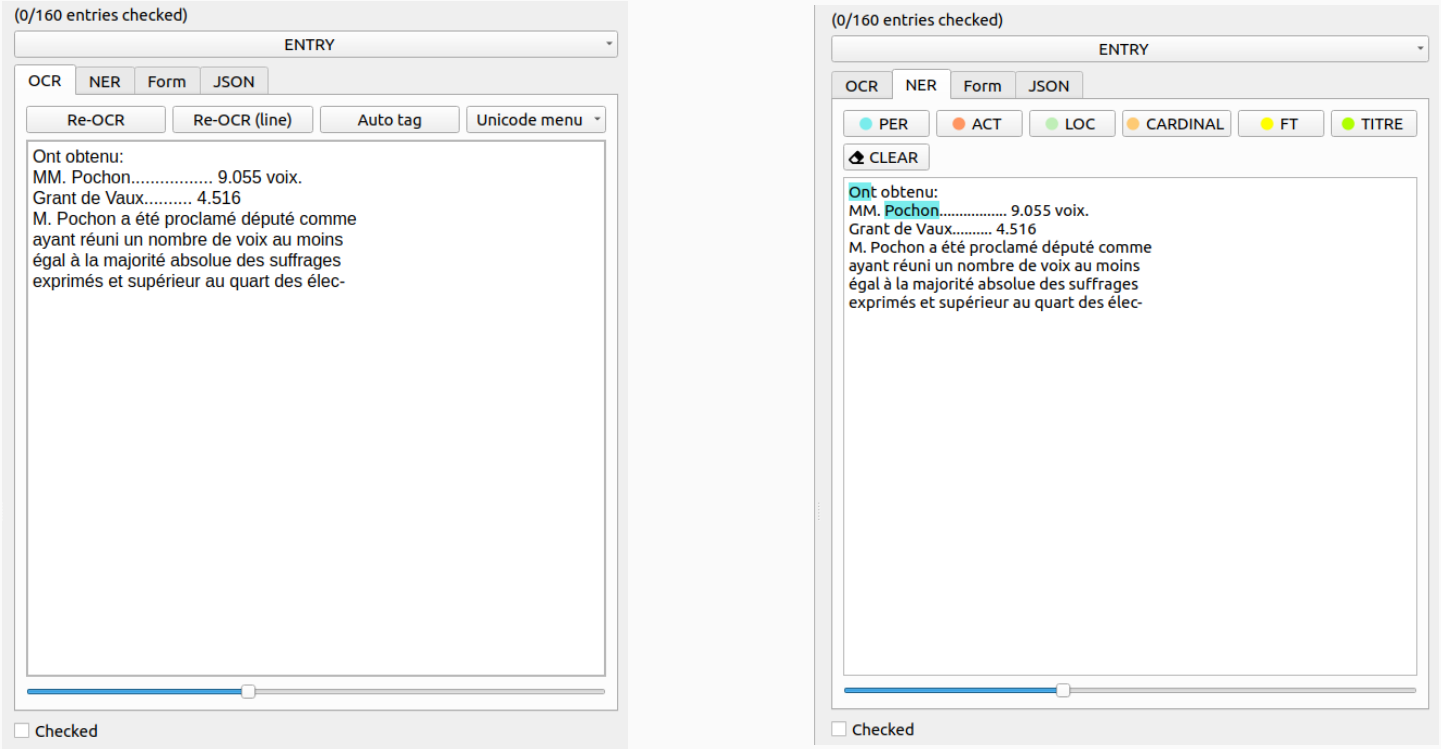



**Figure –** Outil LRDE (EPITA)

# OUTIL DÉVELOPPÉ PAR LRDE (EPITA) - 2




**(a)** OCR

**(b)** NER

**Figure –** Zones d'OCR et de NER

# OUTIL DÉVELOPPÉ PAR LRDE (EPITA) - 3


```
      "box": [
           827.8018264183805,
           2374.55023568486,
           589.3638270234404,
           207.5897599646105
       ],
       "checked": true,
       "comment": "u-par",
       "id": 352,
       "ner_xml": "<PER>M. Francis Laur</PER>. Je n'ai pas l'intention\u2029d'apporter \u00e0 la Chambre d'autr\u2029S de",
       "origin": "computer",
       "parent": 274,
       "persons": ["M. Francis Laur"],
       "text_ocr": "M. Francis Laur. Je n'ai pas l'intention\nd'apporter \u00e0 la Chambre d'autres documents\nque ceux qui me sont fournis par la lettre\nde
d\u00e9mission de M. Lev\u00eaque, adress\u00e9e \u00e0\nM. le ministre des finances. Je demanderai\n\u00e0 la Chambre si elle conna\u00eet cette lettre ou\nsi
elle veut que j'en donne lecture.",
       "type": "ENTRY"
   },
   {
       "activities": [],
       "addresses": [],
```


**Figure –** Sortie Json (extrait)

# EVALUATION DE LA QUALITÉ DE L'OCR

2 Groupes de métriques :

- Méthodes non-supervisées : Dictionnaire...
- Méthodes supervisées
  - Bag Of Words : hérite de métrique classique, des limites
  - Distance de Levenshtein

**Distance de Levenshtein** : Nombre minimal d'insertion, délétions et substitutions sur charactères seuls pour transformer un texte en un autre. Selon les implémentations, les espaces peuvent être pris en compte ou ignorés

| OCR (string\|token) | Ground Truth \| Dict | $LEV_{1,1,1}$ |
|---|---|---|
| «  des ates sont accomplis » | « Des actes sont accomplis » | 3 |
| « tensor » | [tenseur, trois, cube] | 2 |

**Table –** Exemples de distances de Levenshtein

# MÉTRIQUE : CHARACTER ACCURACY

$$\text{CharacterErrorRate(CER)} = \frac{\text{Lev}(\text{text}_{gt}, \text{text}_{o}cr)}{\text{len}(\text{text}_{gt})}$$

$$\text{CharacterAccuracy} = \max(0, 1 - \text{CER})$$

| OCR | Ground Truth | LEV | CharacterAccuracy |
|---|---|---|---|
| «  des ates sont accomplis » | « Des actes sont accomplis » | 3 | 0.88 |
| « as aboue s0 belo » | « as above so below » | 3 | 0.82 |

**Table –** Calcul de l'Accuracy

# EXEMPLE DE RÉSULTATS

| OCR | Document | CharAccuracy |
|---|---|---|
| ocr-tesseract (eng,psm3) | 26 Novembre 1889 | 0.93 |
| ocr-tesseract (fr,psm3) | 26 Novembre 1889 | 0.95 |
| ABBYY | 26 Novembre 1889 | 0.22 |

**Table –** Résultats sur un de nos document

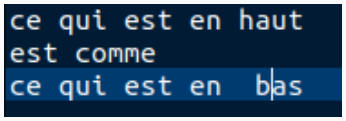
On remarque la limite de la distance de Lenvenshtein, elle est trop sensible à l'odre de détection.

# Flexible Character Accuracy (flexCharAccuracy)

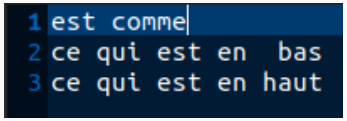
Une métrique moins sensible à l'ordre de détection du texte :

1. Sépare les deux textes en sacs de lignes (GT + OCR)
2. Ordonne les lignes de la GT de la plus longue à la plus petite
3. Pour la première ligne, chercher le meilleur match dans les lignes de l'OCR
4. 
   - Si match complet on sort la ligne
   - Sinon on sous-divise et on rajoute à la pile de lignes
5. On recommence l'étape précédente
6. On compte le nombre de modifications nécessaires pour les lignes qu'on ne peut pas matcher
7. Calcul de l'accuracy

# MÉTRIQUE : FLEXIBLE CHARACTER ACCURACY

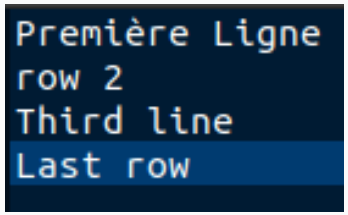



(a) Ground Truth

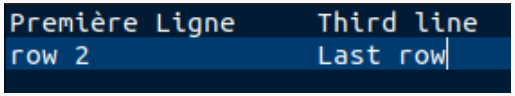



(b) OCR

**Figure –** Exemple A




(a) Ground Truth




(b) OCR

**Figure –** Exemple B

| Exemple | CharAccuracy | FlexCharAccuracy |
|---|---|---|
| A | 0.45 | 1 |
| B | 0.39 | 0.95 |

**Table –** Character Accuracy vs Flexible Character Accuracy

# EXEMPLE DE RÉSULTATS

| OCR | Document | CharAccuracy | FlexCharAccuracy |
|---|---|---|---|
| ocr-tesseract (fr,psm3) | 26 Novembre 1889 | 0.95 | 0.95 |
| ABBY | 26 Novembre 1889 | 0.22 | 0.97 |

**Table –** Résultats sur un de nos documents

- La flexible character accuracy est moins sensible à l'ordre de détection
- Résultats sensiblement équivalents quand l'ordre est le bon
- Comparaison de la qualité de détection des caractères toutes choses égales par ailleurs
- Utile si on n'utilise des algorithmes qui ne s'occupent pas de l'ordre des mots (Clustering via frequency matrix, topic modelling : LDA...)

# POST-CORRECTION

- Dictionnaire de post-correction
- Utilisation d'expressions régulières : gérer les espaces multiples, passage à la ligne, les « - »
- Corrections endogènes
- Modèle de langue CamemBERT pour corriger les mots fautifs (cf. ci-dessus)

# Annoter les débats en XML-TEI

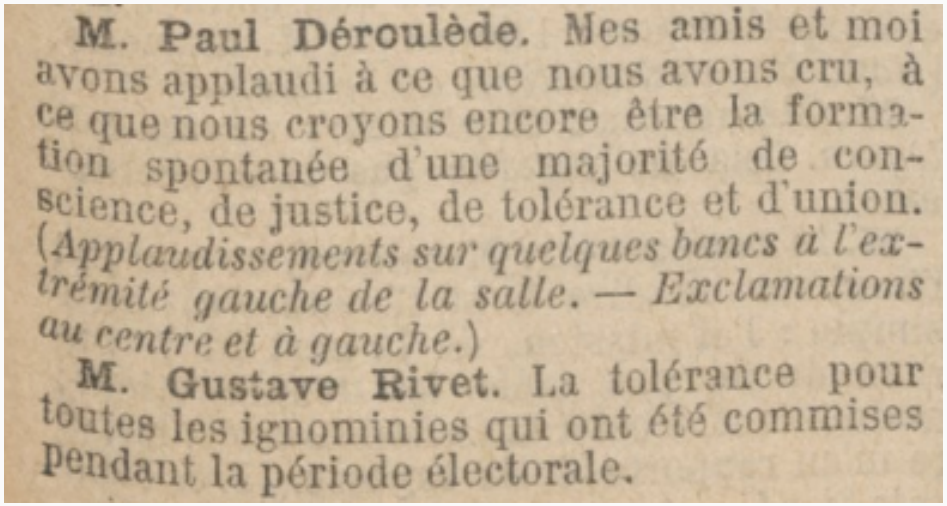
# CADRE DE LA MODÉLISATION

Encodage pensé selon 4 principes :

- Les différentes exploitations des textes
- Les particularités de la source
- Les projets similaires : ParlaClarin et ParlaMint
- Le processus de balisage automatique

Principes permettant d'orienter, de déterminer, d'influencer, et de contraindre nos choix.

# PROBLÉMATIQUE : CONSERVER LA MISE EN PAGE ? - 1



M. Paul Déroulède. Mes amis et moi avons applaudi à ce que nous avons cru, à ce que nous croyons encore être la formation spontanée d'une majorité de conscience, de justice, de tolérance et d'union. (*Applaudissements sur quelques bancs à l'extrémité gauche de la salle. — Exclamations au centre et à gauche.*)

M. Gustave Rivet. La tolérance pour toutes les ignominies qui ont été commises pendant la période électorale.

**Figure –** Source numérisée - Séance parlementaire du 26 novembre 1889 (extrait)

# PROBLÉMATIQUE : CONSERVER LA MISE EN PAGE ? - 2

```
<lb/><u who="#pers_ID" xml:id="CR_1889-11-26_u5" ana="#speaker">
   <seg xml:id="CR_1889-11-26_u5.1">
      <persName ref="#pers_ID">M. Paul Déroulède</persName>. Mes amis et moi
      <lb/>avons applaudi à ce que nous avons cru, à
      <lb/>ce que nous croyons encore être la forma-
      <lb/>tion spontanée d'une majorité de con-
      <lb/>science, de justice, de tolérance et d'union.
      <lb/><incident><desc>(Applaudissements sur quelques bancs à l'ex-
      <!-- Pas de lb possible dans incident --> trémité gauche de La salle. – Exclamations
      <!-- Pas de lb possible dans incident --> au centre et à gauche.)</desc></incident>
   </seg>
</u>

<lb/><u who="pers_ID" xml:id="CR_1889-11-26_u6" ana="#speaker">
   <seg xml:id="CR_1889-11-26_u6.1">
      <persName ref="#pers_ID">Gustave Rivet</persName>. La tolérance pour
      <lb/>toutes Les ignominies qui ont été commises
      <lb/>Pendant la période électorale.
   </seg>
</u>
```

**(a)** Modèle d'encodage 1 : logique, sémantique et formel

```
        <u who="#pers_ID" xml:id="CR_1889-11-26_u5" ana="#speaker">
           <seg xml:id="CR_1889-11-26_u5.1"><persName ref="#pers_ID">M. Paul Déroulède</persName>. Mes amis et moi
avons applaudi à ce que nous avons cru, à ce que nous croyons encore être la formation spontanée d'une majorité de
conscience, de justice, de tolérance et d'union. <incident><desc>(Applaudissements sur quelques bancs à l'extrémité
gauche de la salle. –Exclamations au centre et à gauche.)</desc></incident></seg>
        </u>

        <u who="#pers_ID" xml:id="CR_1889-11-26_u6" ana="#speaker">
           <seg xml:id="CR_1889-11-26_u6.1"><persName ref="#pers_ID">Gustave Rivet</persName>. La tolérance pour
toutes les ignominies qui ont été commises pendant la période électorale.</seg>
        </u>
```

**(b)** Modèle d'encodage 2 : logique et sémantique

**Figure –** Encodages - Séance parlementaire du 26 novembre 1889 (extrait)

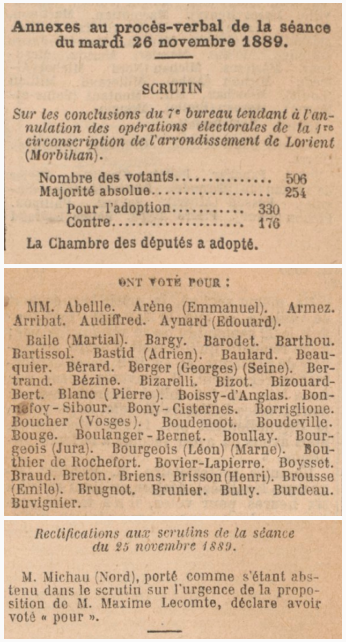
# PROBLÉMATIQUE : CONSERVER LA MISE EN PAGE ? - 3

```
        <div type="part" corresp="#rapportelections">
           <head>SUITE DE LA VÉRIFICATION DES POUVOIRS</head>
           <u who="#pers_ID" xml:id="CR_1889-11-26_u23" ana="#chair">
              <seg xml:id="CR_1889-11-26_u23.1"><persName ref="#pers_ID">M. le <roleName ref="#pers_ID">président</roleName></persName>. L'ordre du jour
appelle la suite de la vérification des pouvoirs.</seg>
              <seg xml:id="CR_1889-11-26_u23.2"><persName ref="#pers_ID">M. Reybert</persName> a la parole pour donner lecture d'un rapport sur une
élection non contestée.</seg>
           </u>

           <u who="#pers_ID" xml:id="CR_1889-11-26_u24" ana="#rapporteur">
              <!-- Lecture d'un rapport -->
              <quote>
                 <seg xml:id="CR_1889-11-26_u24.1"><persName ref="#pers_ID">M. Reybert, <roleName ref="#pers_ID">rapporteur</roleName></persName>.
—<placeName ref="#lieu_ID">Département de la Corrèze, arrondissement de Tulle, <num>1re</num> circonscription</placeName>.</seg>
                 <seg xml:id="CR_1889-11-26_u24.2">Les élections du <date when="1889-09-22">22 septembre</date> ont donné les résultats suivants :</seg>
                 <seg xml:id="CR_1889-11-26_u24.3">Electeurs inscrits, <num>17,950</num>, dont le quart est de <num>4,263</num>.</seg>
                 <seg xml:id="CR_1889-11-26_u24.4">Nombre des votants, <num>12,322</num>.</seg>
                 <seg xml:id="CR_1889-11-26_u24.5">Bulletins blancs et nuls, à déduire, <num>133</num>.</seg>
                 <seg xml:id="CR_1889-11-26_u24.6">Suffrages exprimés, <num>12,189</num>.</seg>
                 <seg xml:id="CR_1889-11-26_u24.7">Ont obtenu : MM. <persName ref="#pers_ID">Borie, <roleName ref="#pers_ID">député
sortant</roleName></persName>... <num>7.508</num> voix</seg>
                 <seg xml:id="CR_1889-11-26_u24.8"><persName ref="#pers_ID">Vachal, <roleName ref="#pers_ID">ancien député</roleName></persName>...
<num>4.748</num></seg>
                 <seg xml:id="CR_1889-11-26_u24.9"><persName ref="#pers_ID">M. Borie</persName> a été proclamé député comme ayant réuni un nombre de voix
au moins égal à la majorité absolue des suffrages</seg>
              </quote>
           </u>

           <floatingText><body><div><pb n="176"/></div></body></floatingText>

           <!-- [...] -->
        </div>
```

(a) Modèle d'encodage 3 : solution médiane

**Figure –** Encodage - Séance parlementaire du 26 novembre 1889 (extrait)

# PROBLÉMATIQUE : ENCODER LES ANNEXES



**Annexes au procès-verbal de la séance du mardi 26 novembre 1889.**

SCRUTIN

*Sur les conclusions du 7e bureau tendant à l'annulation des opérations électorales de la 1re circonscription de l'arrondissement de Lorient (Morbihan).*

Nombre des votants..... 506
Majorité absolue..... 254
Pour l'adoption..... 330
Contre..... 176

La Chambre des députés a adopté.

ONT VOTÉ POUR :

MM. Abeille. Arène (Emmanuel). Armez. Arribat. Audifred. Aynard (Edouard).

Baile (Martial). Bargy. Barodet. Barthou. Bartissol. Bastid (Adrien). Baulard. Beauquier. Bérard. Berger (Georges) (Seine). Bertrand. Bézine. Bizarelli. Bizot. Bizouard-Bert. Blanc (Pierre). Boissy-d'Anglas. Bonnefoy-Sibour. Bony-Cisternes. Borriglione. Boucher (Vosges). Boudenoot. Boudeville. Bouge. Boulanger-Bernet. Boullay. Bourgeois (Jura). Bourgeois (Léon) (Marne). Bouthier de Rochefort. Bovier-Lapierre. Boysset. Braud. Breton. Briens. Brisson (Henri). Brousse (Emile). Brugnot. Brunier. Bully. Burdeau. Buvignier.

*Rectifications aux scrutins de la séance du 25 novembre 1889.*

M. Michau (Nord), porté comme s'étant abstenu dans le scrutin sur l'urgence de la proposition de M. Maxime Lecomte, déclare avoir voté « pour ».

(a) Source numérisée

```
<!-- ANNEXES -->
<back>
   <head>Annexes au procès-verbal de la séance du <date when="1889-11-26">mardi 26 novembre 1889</date>.</head>

   <div xml:id="vot18891126">
      <!-- VOTE 1 -->
      <div xml:id="vot18891126_vot1" type="voting" corresp="#discussion7ebureau">
         <head>
            <label>SCRUTIN</label>
            <note><seg>Sur les conclusions du <num>7e</num> bureau tendant à l'annulation des opérations électorales de la
<placeName ref="#lieu_ID"><num>1re</num> circonscription de l'arrondissement de Lorient (Morbihan)</placeName>.</seg></note>
         </head>
         <!-- Détail du vote -->
         <desc>
            <measure type="nbvotants" quantity="506">Nombre des votants <num>506</num></measure>
            <measure type="maj" quantity="254">Majorité absolue <num>254</num></measure>
            <measure type="ayes" quantity="330">Pour l'adoption <num>330</num></measure>
            <measure type="noes" quantity="176">Contre <num>176</num></measure>
         </desc>

         <note type="result"><seg>La <orgName ref="#org_ID">Chambre des députés</orgName> a adopté.</seg></note>

         <floatingText><body><div><pb n="192"/></div></body></floatingText>

         <!-- Liste des votants -->
         <note type="voterslist">
            <desc>Ont voté pour :</desc>
            <seg> MM.<persName ref="#pers_ID">Abeille</persName>. <persName ref="#pers_ID">Arène (Emmanuel)</persName>. <persName
ref="#pers_ID">Armez</persName>. <persName ref="#pers_ID">Arribat</persName>. <persName ref="#pers_ID">Audifred</persName>. <persName ref
="#pers_ID">Aynard (Edouard)</persName>.</seg>
            <!-- [...] -->
         </note>
      </div>

      <!-- RECTIFICATIONS -->
      <div corresp="#vot18891125" type="rectification">
         <head>Rectifications aux scrutins de la séance du <date>25 novembre 1889</date>.</head>
         <note corresp="#vot18891125_vot1"><seg><persName ref="#pers_ID">M. Michau</persName> <placeName
ref="#lieu_ID">(Nord)</placeName>, porté comme s'étant abstenu dans le scrutin sur l'urgence de la proposition de <persName
ref="#pers_ID">M. Maxime Lecomte</persName>, déclare avoir voté « pour ».</seg></note>
         <!-- [...] -->
      </div>
   </div>
</back>
```

(b) Modèle d'encodage

**Figure –** Séance parlementaire du 26 novembre 1889 - votes, liste des votants, rectifications (extrait annexes)

# STRUCTURE GÉNÉRALE DES FICHIERS XML TEI

```xml
<teiCorpus xmlns="http://www.tei-c.org/ns/1.0" xml:id="FR_3R_5L" xml:lang="fr">

   <!-- Métadonnées du corpus (non développées) -->
   <teiHeader>
      <fileDesc>
         <titleStmt>
            <title></title>
         </titleStmt>
         <publicationStmt>
            <p></p>
         </publicationStmt>
         <sourceDesc>
            <p></p>
         </sourceDesc>
      </fileDesc>
   </teiHeader>

   <!-- Données liées et informations contextuelles -->
   <standOff>
      <listPerson>
         <person></person>
      </listPerson>

      <listOrg>
         <org></org>
      </listOrg>

      <listPlace>
         <place></place>
      </listPlace>
   </standOff>

   <!-- Stockage du composant correspondant à la séance du 26 novembre 1889 -->
   <xi:include xmlns:xi="http://www.w3.org/2001/XInclude" href="FR_3R_5L_1889-11-26.xml"/>


   <!-- Stockage des autres composants du corpus de façon identique -->
   <!-- ... -->

</teiCorpus>
```

**(a)** Structure générale d'un fichier corpus

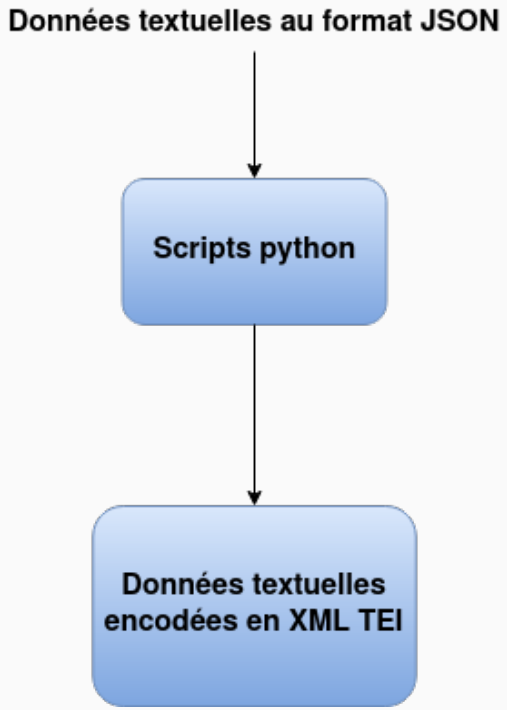
```xml
<TEI xmlns="http://www.tei-c.org/ns/1.0" xml:id="FR_3R_5L_1889-11-26" xml:lang="fr">

   <!-- Métadonnées du composant (non développées) -->
   <teiHeader>
      <fileDesc>
         <titleStmt>
            <title></title>
         </titleStmt>
         <publicationStmt>
            <p></p>
         </publicationStmt>
         <sourceDesc>
            <p></p>
         </sourceDesc>
      </fileDesc>
   </teiHeader>

   <text>

      <!-- Transcription du compte rendu -->
      <body>
         <div></div>
      </body>

      <!-- Annexes du compte rendu -->
      <back></back>

   </text>

</TEI>
```

**(b)** Structure générale d'un fichier composant

**Figure –** Encodages - Structure générale des fichiers corpus et composant

# APPLIQUER L'ENCODAGE : AUTOMATISATION


```
19:
  activities:        []
  addresses:         []
  box:
    0:               220.12998972250773
    1:               282.1299897225077
    2:               561.8527187504491
    3:               80.4418437486525
  checked:           true
  comment:           "seg"
  id:                276
  ner_xml:           "<PER>M. Borie</PER> a <ACT>déjà fait partie des Assemblées\u2029législatives et satisfait aux conditions d</ACT>'<ACT>âge\u2029et de
                     nationalité</ACT> exigées par la loi.\u2029L'élection de la 2e circonscription de"
  origin:            "computer"
  parent:            261
  persons:
    0:               "M. Borie"
  text_ocr:          "M. Borie a déjà fait partie des Assemblées\nlégislatives et satisfait aux conditions d'âge\net de nationalité exigées par la loi."
  type:              "ENTRY"
```

# PUBLIER UN CORPUS ÉDITORIALISÉ

- Publication avec TEI Publisher : outil de publication de corpus en TEI
- Basé sur la technologie Web Components : permet de développer de nouvelles fonctionnalités
- Développer de nouveaux modes de visualisation : par exemple, explorer les débats en fonction des sujets qu'ils évoquent , et pas seulement par des mots-clés

# VERS LES LINKED DATA

# VERS LES LINKED DATA

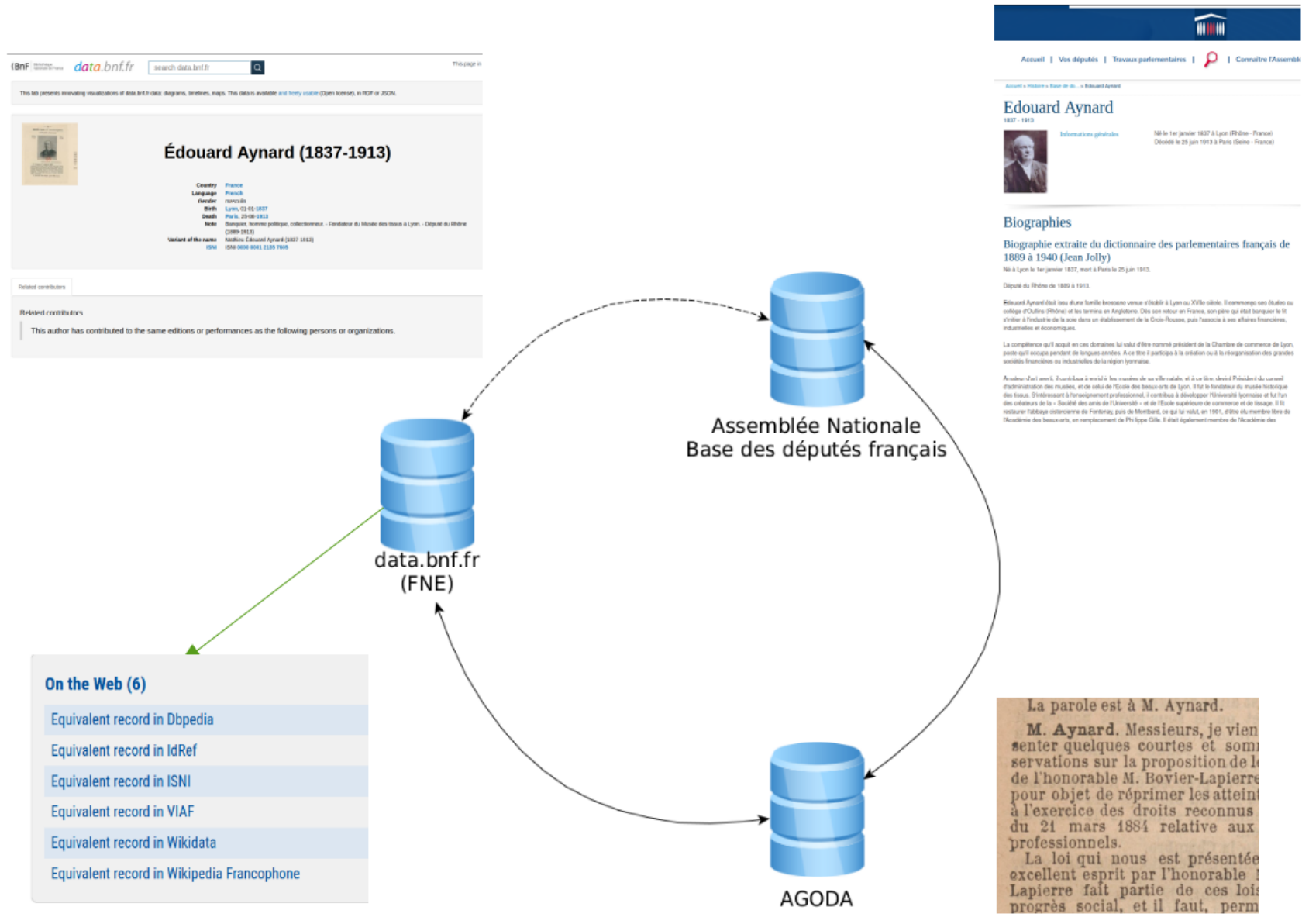


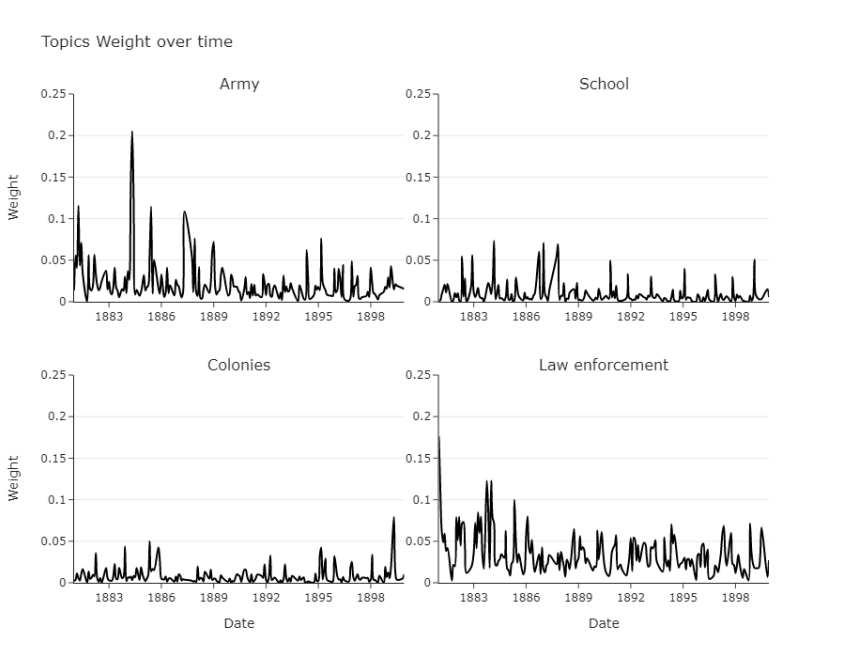

**Figure –** Linked data

# TOPIC MODELING ET WORD EMBEDDING

# EXPLORER LES DÉBATS AVEC LA MODÉLISATION DE SUJETS

| Topic 8 | Topic 11 | Topic 15 |
|---|---|---|
| salaire | général | pari |
| question | commission | télégraphe |
| gouvernement | régiment | faire |
| jour | troupe | ingénieur |
| patron | monsieur | train |
| chambre | année | ligne |
| droit | jeune | chambre |
| syndicat | temps | personnel |
| délégué | faire | etat |
| monsieur | corps | administration |
| travail | soldat | employé |
| travaux | ministre | poste |
| ministre | homme | public |
| grève | loi | travaux |
| faire | an | service |
| mineur | guerre | agent |
| mine | service | ministre |
| loi | militaire | fer |
| compagnie | officier | chemin |
| ouvrier | armée | compagnie |

**(a)** 3 sujets parmi 40 : classe ouvrière (8), armée (11) et infrastructures (15)




**(b)** Evolution de quatre sujets au cours du temps

**Figure –** Résultats de la modélisation de sujets

# LES PLONGEMENTS DE MOTS : WORD2VEC ET TOP2VEC


(a) Projection t-SNE des centroïdes des vecteurs (word2vec)

| Cluster 55 | Cluster 68 | Cluster 70 |
|---|---|---|
| victimes | divorce | enveloppes |
| inondations | epoux | timbres |
| secourir | mariage | poste |
| eprouvees | conjugal | postale |
| orages | divorces | timbre |
| sinistres | adultere | recepisses |
| grele | conjugale | postes |
| secours | remarier | postaux |
| venir | separation | telegraphes |
| infortunes | indissolubilite | colis |
| ravages | conjoints | fixe |
| miseres | mutuel | recouvrements |
| catastrophe | separations | graphes |
| evenements | mari | postales |
| repartition | mariages | taxe |
| incendies | femme | decide |
| soulager | conjoint | soit |

(b) 3 clusters parmi les 113 obtenus avec top2vec : tempêtes (55), divorce (68) et poste (70)

**Figure –** Résultats obtenus avec word2vec et top2vec

# POUR EN SAVOIR PLUS


- Nicolas Bourgeois, Aurélien Pellet, Marie Puren. ”Using Topic Generation Model to explore the French Parliamentary Debates during the early Third Republic (1881-1899)”. (hal-03526254v2)
- Marie Puren, Nicolas Bourgeois, Aurélien Pellet, Pierre Vernus, Fanny Lebreton. ”Between History and Natural Language Processing : Study, Enrichment and Online Publication of French Parliamentary Debates of the Early Third Republic (1881-1899)”. ParlaCLARIN III at LREC2022 - Workshop on Creating, Enriching and Using Parliamentary Corpora, Jun 2022, Marseille, France. (hal-03623351)

# MERCI POUR VOTRE ATTENTION !

Nicolas Bourgeois : nicolas.bourgeois@epitech.eu

Fanny Lebreton : fanny.lebreton@chartes.psl.eu

Aurélien Pellet : aurelien.pellet@epitech.eu

Marie Puren : marie.puren@epitech.eu

Pierre Vernus : pierre.vernus@msh-lse.fr

Répertoire Github du projet : https://github.com/mpuren/agoda