

Formation Edition numérique

# OCR

Simon Gabay



# Une image numérique

Deux types d'images:

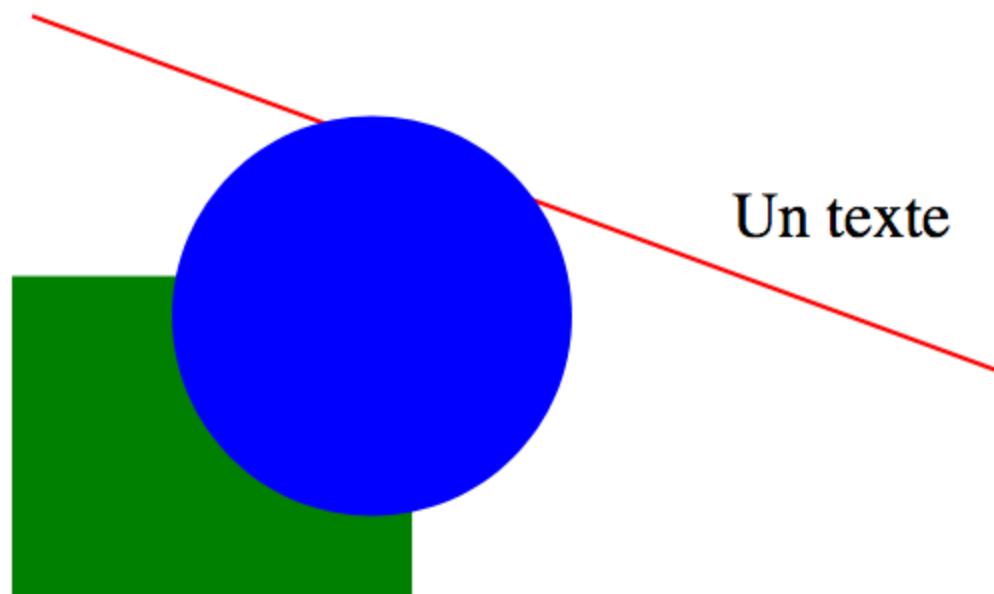
- Image vectorielle
- Image matricielle (ou bitmap)

## Image vectorielle (I)

- Représenter les données de l'image par des formules géométriques qui vont pouvoir être décrites d'un point de vue mathématique (abscisse et ordonnées)
- C'est notamment le format *svg* (pour *Scalable Vector Graphics*)
- En pratique : pas de problème si on zoom.

## Image vectorielle (II)

```
<svg>
  <rect width="100" height="80" x="0" y="70" fill="green"/>
  <line x1="5" y1="5" x2="250" y2="95" stroke="red" />
  <circle cx="90" cy="80" r="50" fill="blue" />
  <text x="180" y="60">Un texte</text>
</svg>
```



## Image vectorielle (III)

Ouvrez le fichier `image.svg` dans un navigateur et dans un éditeur de code: comparez!

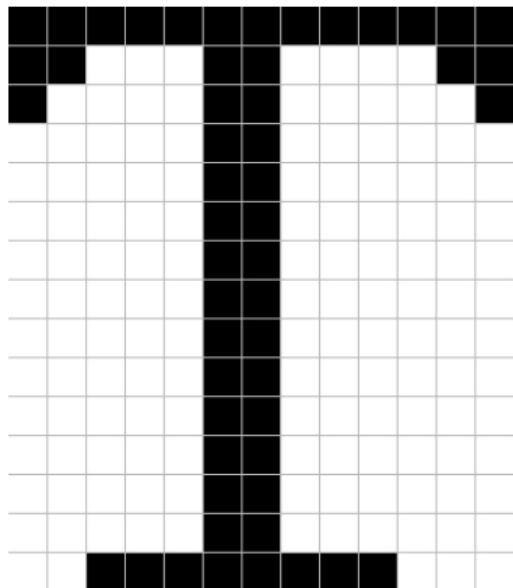
Pour plus d'exemples, allez regarder du côté de [w3schools](#).

## Une image bitmap (I)

- Composée d'une matrice (tableau) de points à plusieurs dimensions. Dans le cas des images à deux dimensions (le plus courant), les points sont appelés pixels.
- C'est notamment le format jpeg, gif, png outif.
- Ces différents formats se différencient par le nombre de couleurs, leur compression (avec ou sans perte), la possibilité d'un affichage progressif.
- En pratique : problème si on zoom.

# Une image bitmap (I)

Deux fois la même image matricielle

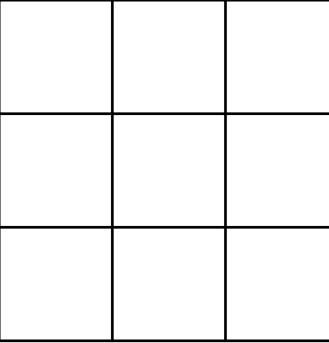
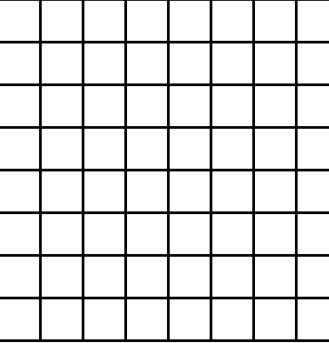
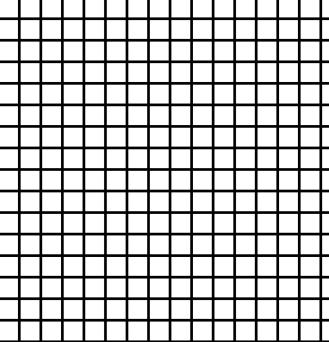


1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	0	0	0	1	1	0	0	0	0	1	1	0	0	0	0
1	0	0	0	0	1	1	0	0	0	0	0	1	1	0	0	0
0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0
0	0	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0

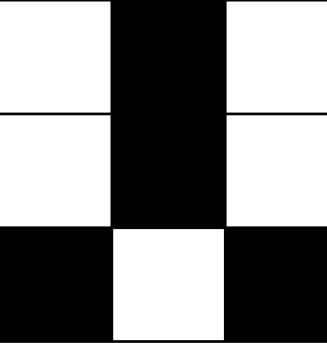
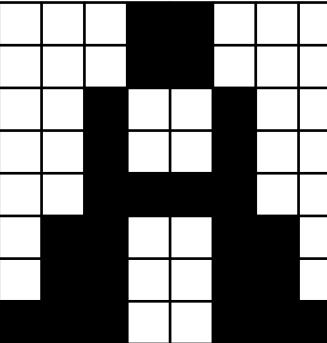
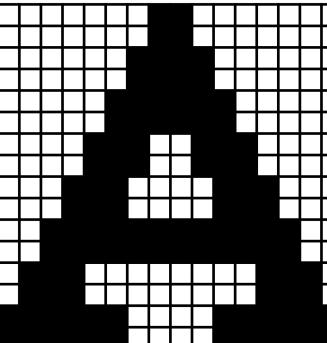
# Les caractéristiques techniques d'une image

- Sa taille en points (ou pixels)
- Ses dimensions réelles (en centimètres ou plus souvent en pouces)
  - un pouce faisant c. 2.4 cm
- On parle donc de *dpi* (*dot per inch*) ou *ppp* (*point par pouce*) pour la résolution, soit un nombre de pixels par unité de longueur.
- Meilleure est la résolution, meilleure est l'OCRisation

# PPI

Image	ppp
	3
	8
	16

# PPI en pratique: la lettre A

Image	ppp
	3
	8
	16

## Poids de l'image

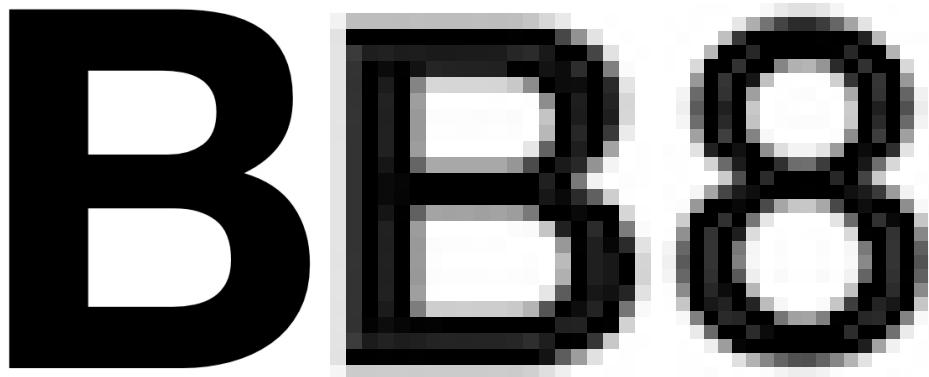
Résolution d'une page A4:  $(\text{dpi} * (21/2.54)) * (\text{dpi} * (29.7/2.54))$

dpi	pixels	total
100	826 x 1169	965 594
200	1650 x 2340	3 861 000
300	3500 x 2480	8 680 000

Il est louable de vouloir avoir de bonnes images pour l'OCR, mais attention au poids de l'image finale!

## Le *B-test*

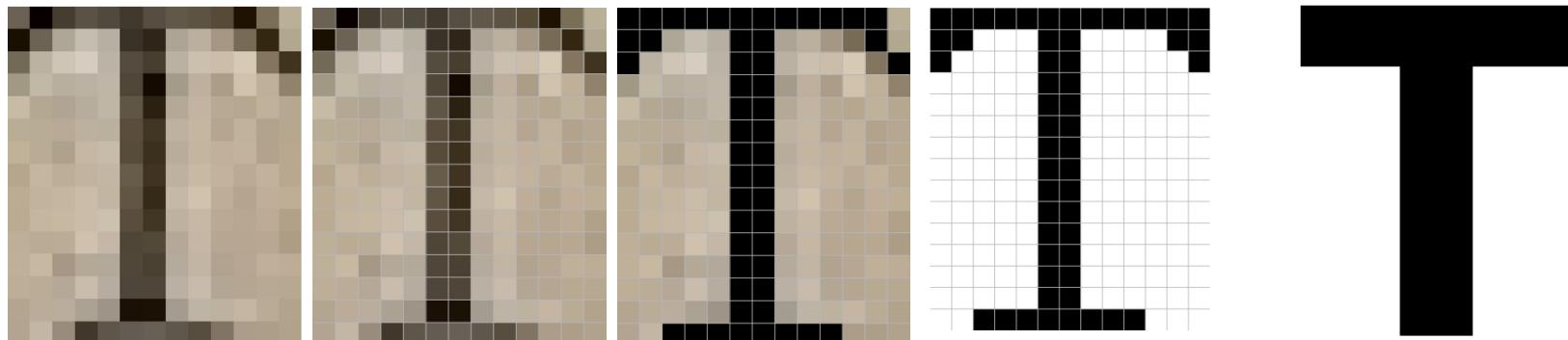
L'enjeu est de réussir ce que nous appellerons le *B-test*



## Résolution vs efficacité

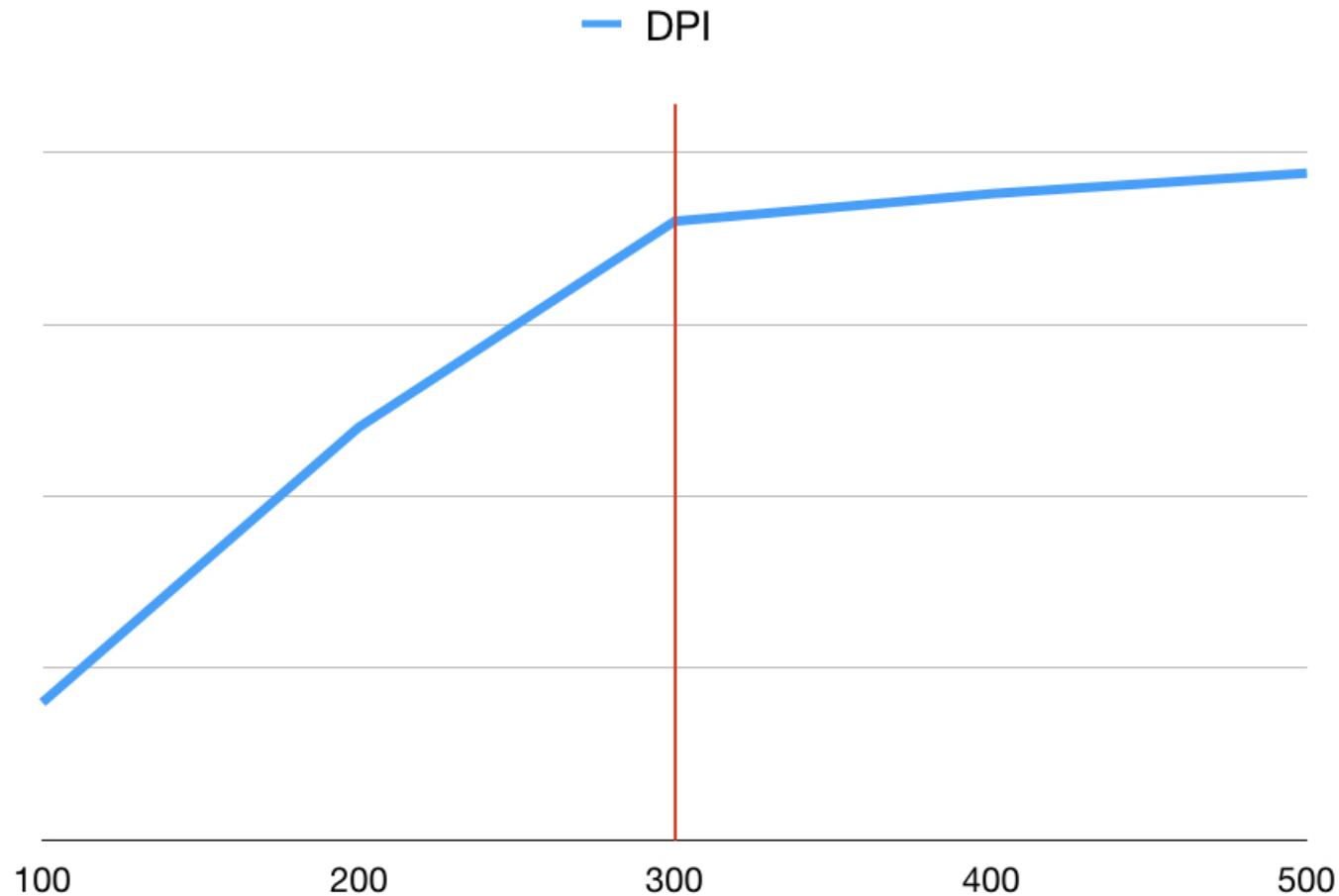
Il n'est pas nécessaire d'avoir un grand nombre de pixels (au contraire) pour bien faire fonctionner un OCR.

La schématisation de l'image obtenue par sa pixelisation est une force: trop d'information tue l'information.



# La bonne résolution (III)

300 dpi serait le meilleur rapport poids/qualité



# *Pre-processing*

# Rotation

## Original

— 6 —

Il lui fait part du décès de Gaston d'Orléans; détails sur la fin du prince.  
Chaleureuse recommandation en faveur d'une femme de chambre.

- 87545 BIGOTTINI (Émilie), célèbre danseuse, qui eut de grands succès à l'Opéra. — L. a. s. à M<sup>me</sup> Guyet; 29 août 1833, 10 » 2 p. in-folio.

Chaleureuse recommandation en faveur d'une femme de chambre.

- 87546 BLANQUI (Jérôme-Adolphe), économiste, membre de l'Académie des Sciences morales et politiques, né à Nice. — 30 l. a. s. à Jullien de Paris; 1818-1841, 50 p. env. in-4° ou in-8°. 30 »

Intéressante correspondance toute relative à ses travaux et à sa collaboration à la *Revue encyclopédique*... Il parle de ses frères dont il dirigeait l'éducation, des moyens d'échapper à la conscription. « Mes parents sont dans l'impossibilité de m'acheter un remplaçant, et il me serait bien désagréable d'interrompre ma carrière et de perdre le fruit de mes études en entrant comme simple soldat dans les rangs de l'armée. » Il demande la réintégration de son père, le conventionnel, un des 73 du 31 mai, dont il retrace une brève biographie. Il accepte de collaborer à la *Revue encyclopédique*, mais il refuse de s'occuper de la rédaction des tables; il demande des renseignements pour écrire une vie de Kosciusko, l'émule de Washington. Les lettres postérieures traitent plus particulièrement d'économie sociale.

- 87547 BONAPARTE (Jérôme), le fils du roi Jérôme. — L. a. s. N. à M. Gasc; Paris, 31 octobre 1848, 2 p. 1/2 in-8°. 35 »

Curieuse lettre politique entièrement relative à la préparation de l'élection de son cousin à la présidence de la République. « Tous les chefs sont mal pour nous, nos vrais appuis sont les paysans, les ouvriers et les soldats, la tête de tous les partis nous est hostile, cependant les hommes d'ordre reviennent à nous. Cavaignac a pour lui le pouvoir, la corruption et l'intimidation; son cousin a pour lui le nom de l'Empereur et la sympathie des masses » qui voit (*sic*) que lui, mieux que tout autre peut sauver la République, l'établir à tout jamais en France ! »

- 87548 BOSQUET (Pierre-Jean-François), maréchal de France, qui s'illustra à l'Alma, à Inkermann, et à l'assaut de la tour Malakoff, né à Mont-de-Marsan (Landes). — L. a. s. à M. Renoux; 12 juin 1851, 3 p. in-8°. 25 »

Intéressante lettre entièrement relative aux opérations militaires; il se plaint de la froideur de ses vieux camarades, qui jaloussent ses succès.

- 87549 BRUANT (Aristide), chansonnier montmartrois, né à Courtenay (Loiret). — L. a. s., à un confrère; Le Pouldu, 15 août 1911, 1 p. in-8°. 5 »

Il dit que sa pièce les *Bas-fonds de Paris* est déjà faite; il l'a écrite en collaboration avec Arthur Bernède.

- 87550 BRUEYS (François-Paul), vice-amiral, commandant en

## Résultat

— 6 —

Il lui fait part du décès de Gaston d'Orléans; détails sur la fin du prince.  
Chaleureuse recommandation en faveur d'une femme de chambre.

- 87545 BIGOTTINI (Émilie), célèbre danseuse, qui eut de grands succès à l'Opéra. — L. a. s. à M<sup>me</sup> Guyet; 29 août 1833, 10 » 2 p. in-folio.

Chaleureuse recommandation en faveur d'une femme de chambre.

- 87546 BLANQUI (Jérôme-Adolphe), économiste, membre de l'Académie des Sciences morales et politiques, né à Nice. — 30 l. a. s. à Jullien de Paris; 1818-1841, 50 p. env. in-4° ou in-8°. 30 »

Intéressante correspondance toute relative à ses travaux et à sa collaboration à la *Revue encyclopédique*... Il parle de ses frères dont il dirigeait l'éducation, des moyens d'échapper à la conscription. « Mes parents sont dans l'impossibilité de m'acheter un remplaçant, et il me serait bien désagréable d'interrompre ma carrière et de perdre le fruit de mes études en entrant comme simple soldat dans les rangs de l'armée. » Il demande la réintégration de son père, le conventionnel, un des 73 du 31 mai, dont il retrace une brève biographie. Il accepte de collaborer à la *Revue encyclopédique*, mais il refuse de s'occuper de la rédaction des tables; il demande des renseignements pour écrire une vie de Kosciusko, l'émule de Washington. Les lettres postérieures traitent plus particulièrement d'économie sociale.

- 87547 BONAPARTE (Jérôme), le fils du roi Jérôme. — L. a. s. N. à M. Gasc; Paris, 31 octobre 1848, 2 p. 1/2 in-8°. 35 »

Curieuse lettre politique entièrement relative à la préparation de l'élection de son cousin à la présidence de la République. « Tous les chefs sont mal pour nous, nos vrais appuis sont les paysans, les ouvriers et les soldats, la tête de tous les partis nous est hostile, cependant les hommes d'ordre reviennent à nous. Cavaignac a pour lui le pouvoir, la corruption et l'intimidation; son cousin a pour lui le nom de l'Empereur et la sympathie des masses » qui voit (*sic*) que lui, mieux que tout autre peut sauver la République, l'établir à tout jamais en France ! »

- 87548 BOSQUET (Pierre-Jean-François), maréchal de France, qui s'illustra à l'Alma, à Inkermann, et à l'assaut de la tour Malakoff, né à Mont-de-Marsan (Landes). — L. a. s. à M. Renoux; 12 juin 1851, 3 p. in-8°. 25 »

Intéressante lettre entièrement relative aux opérations militaires; il se plaint de la froideur de ses vieux camarades, qui jaloussent ses succès.

- 87549 BRUANT (Aristide), chansonnier montmartrois, né à Courtenay (Loiret). — L. a. s., à un confrère; Le Pouldu, 15 août 1911, 1 p. in-8°. 5 »

Il dit que sa pièce les *Bas-fonds de Paris* est déjà faite; il l'a écrite en collaboration avec Arthur Bernède.

- 87550 BRUEYS (François-Paul), vice-amiral, commandant en

# Niveau de gris

## Original

— 6 —

Il lui fait part du décès de Gaston d'Orléans; détails sur la fin du prince.  
Chaleureuse recommandation en faveur d'une femme de chambre.

- 87545** BIGOTTINI (Émilie), célèbre danseuse, qui eut de grands succès à l'Opéra. — L. a. s. à M<sup>me</sup> Guyet; 29 août 1833, 2 p. in-folio. 10 »

Chaleureuse recommandation en faveur d'une femme de chambre.

- 87546** BLANQUI (Jérôme-Adolphe), économiste, membre de l'Académie des Sciences morales et politiques, né à Nice. — 30 l. a. s. à Jullien de Paris; 1818-1841, 50 p. env. in-4° ou in-8°. 30 »

Intéressante correspondance toute relative à ses travaux et à sa collaboration à la *Revue encyclopédique*... Il parle de ses frères dont il dirigeait l'éducation, des moyens d'échapper à la conscription. « Mes parents sont dans l'impossibilité de m'acheter un remplaçant, et il me serait bien désagréable d'interrompre ma carrière et de perdre le fruit de mes études en entrant comme simple soldat dans les rangs de l'armée. » Il demande la réintégRATION de son père, le conventionnel, un des 73 du 31 mai, dont il retrace une brève biographie. Il accepte de collaborer à la *Revue encyclopédique*, mais il refuse de s'occuper de la rédaction des tables; il demande des renseignements pour écrire une vie de Kosciusko, l'émule de Washington. Les lettres postérieures traitent plus particulièrement d'économie sociale.

- 87547** BONAPARTE (Jérôme), le fils du roi Jérôme. — L. a. s. N. à M. Gasc; Paris, 31 octobre 1848, 2 p. 1/2 in-8°. 35 »

Curieuse lettre politique entièrement relative à la préparation de l'élection de son cousin à la présidence de la République. « Tous les chefs sont mal pour nous; nos vrais appuis sont les paysans, les ouvriers et les soldats, la tête de tous les partis nous est hostile, cependant les hommes d'ordre reviennent à nous. Cavaignac a pour lui le pouvoir, la corruption et l'intimidation; son cousin a pour lui le nom de l'Empereur et la sympathie des masses « qui voit (sic) que lui, mieux que tout autre peut sauver la République, l'établir à tout jamais en France ! »

- 87548** BOSQUET (Pierre-Jean-François), maréchal de France, qui s'illustra à l'Alma, à Inkermann, et à l'assaut de la tour Malakoff, né à Mont-de-Marsan (Landes). — L. a. s. à M. Renoux; 12 juin 1851, 3 p. in-8°. 25 »

Intéressante lettre entièrement relative aux opérations militaires; il se plaint de la froideur de ses vieux camarades, qui jaloussent ses succès.

- 87549** BRUANT (Aristide), chansonnier montmartrois, né à Courtenay (Loiret). — L. a. s., à un confrère; Le Pouldu, 15 août 1911, 1 p. in-8°. 5 »

Il dit que sa pièce les *Bas-fonds de Paris* est déjà faite; il l'a écrite en collaboration avec Arthur Bernède.

- 87550** BRUEYS (François-Paul), vice-amiral, commandant en

## Résultat

— 6 —

Il lui fait part du décès de Gaston d'Orléans; détails sur la fin du prince.  
Chaleureuse recommandation en faveur d'une femme de chambre.

- 87545** BIGOTTINI (Émilie), célèbre danseuse, qui eut de grands succès à l'Opéra. — L. a. s. à M<sup>me</sup> Guyet; 29 août 1833, 2 p. in-folio. 10 »

Chaleureuse recommandation en faveur d'une femme de chambre.

- 87546** BLANQUI (Jérôme-Adolphe), économiste, membre de l'Académie des Sciences morales et politiques, né à Nice. — 30 l. a. s. à Jullien de Paris; 1818-1841, 50 p. env. in-4° ou in-8°. 30 »

Intéressante correspondance toute relative à ses travaux et à sa collaboration à la *Revue encyclopédique*... Il parle de ses frères dont il dirigeait l'éducation, des moyens d'échapper à la conscription. « Mes parents sont dans l'impossibilité de m'acheter un remplaçant, et il me serait bien désagréable d'interrompre ma carrière et de perdre le fruit de mes études en entrant comme simple soldat dans les rangs de l'armée. » Il demande la réintégRATION de son père, le conventionnel, un des 73 du 31 mai, dont il retrace une brève biographie. Il accepte de collaborer à la *Revue encyclopédique*, mais il refuse de s'occuper de la rédaction des tables; il demande des renseignements pour écrire une vie de Kosciusko, l'émule de Washington. Les lettres postérieures traitent plus particulièrement d'économie sociale.

- 87547** BONAPARTE (Jérôme), le fils du roi Jérôme. — L. a. s. N. à M. Gasc; Paris, 31 octobre 1848, 2 p. 1/2 in-8°. 35 »

Curieuse lettre politique entièrement relative à la préparation de l'élection de son cousin à la présidence de la République. « Tous les chefs sont mal pour nous; nos vrais appuis sont les paysans, les ouvriers et les soldats, la tête de tous les partis nous est hostile, cependant les hommes d'ordre reviennent à nous. Cavaignac a pour lui le pouvoir, la corruption et l'intimidation; son cousin a pour lui le nom de l'Empereur et la sympathie des masses « qui voit (sic) que lui, mieux que tout autre peut sauver la République, l'établir à tout jamais en France ! »

- 87548** BOSQUET (Pierre-Jean-François), maréchal de France, qui s'illustra à l'Alma, à Inkermann, et à l'assaut de la tour Malakoff, né à Mont-de-Marsan (Landes). — L. a. s. à M. Renoux; 12 juin 1851, 3 p. in-8°. 25 »

Intéressante lettre entièrement relative aux opérations militaires; il se plaint de la froideur de ses vieux camarades, qui jaloussent ses succès.

- 87549** BRUANT (Aristide), chansonnier montmartrois, né à Courtenay (Loiret). — L. a. s., à un confrère; Le Pouldu, 15 août 1911, 1 p. in-8°. 5 »

Il dit que sa pièce les *Bas-fonds de Paris* est déjà faite; il l'a écrite en collaboration avec Arthur Bernède.

- 87550** BRUEYS (François-Paul), vice-amiral, commandant en

# Binarisation

## Original

— 6 —

Il lui fait part du décès de Gaston d'Orléans; détails sur la fin du prince.  
Chaleureuse recommandation en faveur d'une femme de chambre.

- 87545** BIGOTTINI (Émilie), célèbre danseuse, qui eut de grands succès à l'Opéra. — L. a. s. à M<sup>me</sup> Guyet; 29 août 1833, 2 p. in-folio. 10 »

Chaleureuse recommandation en faveur d'une femme de chambre.

- 87546** BLANQUI (Jérôme-Adolphe), économiste, membre de l'Académie des Sciences morales et politiques, né à Nice. — 30 l. a. s. à Jullien de Paris; 1818-1841, 50 p. env. in-4° ou in-8°. 30 »

Intéressante correspondance toute relative à ses travaux et à sa collaboration à la *Revue encyclopédique*... Il parle de ses frères dont il dirigeait l'éducation, des moyens d'échapper à la conscription. « Mes parents sont dans l'impossibilité de m'acheter un remplaçant, et il me serait bien désagréable d'interrompre ma carrière et de perdre le fruit de mes études en entrant comme simple soldat dans les rangs de l'armée. » Il demande la réintégRATION de son père, le conventionnel, un des 73 du 31 mai, dont il retrace une brève biographie. Il accepte de collaborer à la *Revue encyclopédique*, mais il refuse de s'occuper de la rédaction des tables; il demande des renseignements pour écrire une vie de Kosciuzko, l'émule de Washington. Les lettres postérieures traitent plus particulièrement d'économie sociale.

- 87547** BONAPARTE (Jérôme), le fils du roi Jérôme. — L. a. s. N. à M. Gasc; Paris, 31 octobre 1848, 2 p. 1/2 in-8°. 35 »

Curieuse lettre politique entièrement relative à la préparation de l'élection de son cousin à la présidence de la République. « Tous les chefs sont mal pour nous, nos vrais appuis sont les paysans, les ouvriers et les soldats, la tête de tous les partis nous est hostile, cependant les hommes d'ordre reviennent à nous. Cavaignac a pour lui le pouvoir, la corruption et l'intimidation; son cousin a pour lui le nom de l'Empereur et la sympathie des masses « qui voit (sic) que lui, mieux que tout autre peut sauver la République, l'établir à tout jamais en France! »

- 87548** BOSQUET (Pierre-Jean-François), maréchal de France, qui s'illustra à l'Alma, à Inkermann, et à l'assaut de la tour Malakoff, né à Mont-de-Marsan (Landes). — L. a. s. à M. Renoux; 12 juin 1851, 3 p. in-8°. 25 »

Intéressante lettre entièrement relative aux opérations militaires; il se plaint de la froideur de ses vieux camarades, qui jaloussent ses succès.

- 87549** BRUANT (Aristide), chansonnier montmartrois, né à Courtenay (Loiret). — L. a. s., à un confrère; Le Pouldu, 15 août 1911, 1 p. in-8°. 5 »

Il dit que sa pièce les *Bas-fonds de Paris* est déjà faite; il l'a écrite en collaboration avec Arthur Bernède.

- 87550** BRUEYS (François-Paul), vice-amiral, commandant en

## Résultat

— 6 —

Il lui fait part du décès de Gaston d'Orléans; détails sur la fin du prince.  
Chaleureuse recommandation en faveur d'une femme de chambre.

- 87545** BIGOTTINI (Émilie), célèbre danseuse, qui eut de grands succès à l'Opéra. — L. a. s. à M<sup>me</sup> Guyet; 29 août 1833, 2 p. in-folio. 10 »

Chaleureuse recommandation en faveur d'une femme de chambre.

- 87546** BLANQUI (Jérôme-Adolphe), économiste, membre de l'Académie des Sciences morales et politiques, né à Nice. — 30 l. a. s. à Jullien de Paris; 1818-1841, 50 p. env. in-4° ou in-8°. 30 »

Intéressante correspondance toute relative à ses travaux et à sa collaboration à la *Revue encyclopédique*... Il parle de ses frères dont il dirigeait l'éducation, des moyens d'échapper à la conscription. « Mes parents sont dans l'impossibilité de m'acheter un remplaçant, et il me serait bien désagréable d'interrompre ma carrière et de perdre le fruit de mes études en entrant comme simple soldat dans les rangs de l'armée. » Il demande la réintégRATION de son père, le conventionnel, un des 73 du 31 mai, dont il retrace une brève biographie. Il accepte de collaborer à la *Revue encyclopédique*, mais il refuse de s'occuper de la rédaction des tables; il demande des renseignements pour écrire une vie de Kosciuzko, l'émule de Washington. Les lettres postérieures traitent plus particulièrement d'économie sociale.

- 87547** BONAPARTE (Jérôme), le fils du roi Jérôme. — L. a. s. N. à M. Gasc; Paris, 31 octobre 1848, 2 p. 1/2 in-8°. 35 »

Curieuse lettre politique entièrement relative à la préparation de l'élection de son cousin à la présidence de la République. « Tous les chefs sont mal pour nous, nos vrais appuis sont les paysans, les ouvriers et les soldats, la tête de tous les partis nous est hostile, cependant les hommes d'ordre reviennent à nous. Cavaignac a pour lui le pouvoir, la corruption et l'intimidation; son cousin a pour lui le nom de l'Empereur et la sympathie des masses « qui voit (sic) que lui, mieux que tout autre peut sauver la République, l'établir à tout jamais en France! »

- 87548** BOSQUET (Pierre-Jean-François), maréchal de France, qui s'illustra à l'Alma, à Inkermann, et à l'assaut de la tour Malakoff, né à Mont-de-Marsan (Landes). — L. a. s. à M. Renoux; 12 juin 1851, 3 p. in-8°. 25 »

Intéressante lettre entièrement relative aux opérations militaires; il se plaint de la froideur de ses vieux camarades, qui jaloussent ses succès.

- 87549** BRUANT (Aristide), chansonnier montmartrois, né à Courtenay (Loiret). — L. a. s., à un confrère; Le Pouldu, 15 août 1911, 1 p. in-8°. 5 »

Il dit que sa pièce les *Bas-fonds de Paris* est déjà faite; il l'a écrite en collaboration avec Arthur Bernède.

- 87550** BRUEYS (François-Paul), vice-amiral, commandant en

# Segmentation I

## Original

— 6 —

Il lui fait part du décès de Gaston d'Orléans; détails sur la fin du prince.  
Chaleureuse recommandation en faveur d'une femme de chambre.

**87545** BIGOTTINI (Émilie), célèbre danseuse, qui eut de grands succès à l'Opéra. — L. a. s. à M<sup>me</sup> Guyet; 29 août 1833, 2 p. in-folio. 10 "

Chaleureuse recommandation en faveur d'une femme de chambre.

**87546** BLANQUI (Jérôme-Adolphe), économiste, membre de l'Académie des Sciences morales et politiques, né à Nice. — 30 l. a. s. à Jullien de Paris; 1818-1841, 50 p. env. in-4° ou in-8°. 30 "

Intéressante correspondance toute relative à ses travaux et à sa collaboration à la *Revue encyclopédique*... Il parle de ses frères dont il dirigeait l'éducation, des moyens d'échapper à la conscription. « Mes parents sont dans l'impossibilité de m'acheter un remplaçant, et il me serait bien désagréable d'interrompre ma carrière et de perdre le fruit de mes études en entrant comme simple soldat dans les rangs de l'armée. » Il demande la réintroduction de son père, le conventionnel, un des 73 du 31 mai, dont il retrace une brève biographie. Il accepte de collaborer à la *Revue encyclopédique*, mais il refuse de s'occuper de la rédaction des tables; il demande des renseignements pour écrire une vie de Kosciusko, l'émule de Washington. Les lettres postérieures traitent plus particulièrement d'économie sociale.

**87547** BONAPARTE (Jérôme), le fils du roi Jérôme. — L. a. s. N. à M. Gasc; Paris, 31 octobre 1848, 2 p. 1/2 in-8°. 35 "

Curieuse lettre politique entièrement relative à la préparation de l'élection de son cousin à la présidence de la République. « Tous les chefs sont mal pour nous, nos vrais appuis sont les paysans, les ouvriers et les soldats, la tête de tous les partis nous est hostile, cependant les hommes d'ordre reviennent à nous. Cavaignac a pour lui le pouvoir, la corruption et l'intimidation; son cousin a pour lui le nom de l'Empereur et la sympathie des masses « qui voit (sic) que lui, mieux que tout autre peut sauver la République, l'établir à tout jamais en France ! »

**87548** BOSQUET (Pierre-Jean-François), maréchal de France, qui s'illustra à l'Alma, à Inkermann, et à l'assaut de la tour Malakoff, né à Mont-de-Marsan (Landes). — L. a. s. à M. Renoux; 12 juin 1851, 3 p. in-8°. 25 "

Intéressante lettre entièrement relative aux opérations militaires; il se plaint de la froideur de ses vieux camarades, qui jaloussent ses succès.

**87549** BRUANT (Aristide), chansonnier montmartrois, né à Courtenay (Loiret). — L. a. s., à un confrère; Le Pouldu, 15 août 1911, 1 p. in-8°. 5 "

Il dit que sa pièce les *Bas-fonds de Paris* est déjà faite; il l'a écrite en collaboration avec Arthur Bernède.

**87550** BRUEYS (François-Paul), vice-amiral, commandant en

## Résultat

— 6 —

Il lui fait part du décès de Gaston d'Orléans; détails sur la fin du prince.  
Chaleureuse recommandation en faveur d'une femme de chambre.

**87545** BIGOTTINI (Émilie), célèbre danseuse, qui eut de grands succès à l'Opéra. — L. a. s. à M<sup>me</sup> Guyet; 29 août 1833, 2 p. in-folio. 10 "

Chaleureuse recommandation en faveur d'une femme de chambre.

**87546** BLANQUI (Jérôme-Adolphe), économiste, membre de l'Académie des Sciences morales et politiques, né à Nice. — 30 l. a. s. à Jullien de Paris; 1818-1841, 50 p. env. in-4° ou in-8°. 30 "

Intéressante correspondance toute relative à ses travaux et à sa collaboration à la *Revue encyclopédique*... Il parle de ses frères dont il dirigeait l'éducation, des moyens d'échapper à la conscription. « Mes parents sont dans l'impossibilité de m'acheter un remplaçant, et il me serait bien désagréable d'interrompre ma carrière et de perdre le fruit de mes études en entrant comme simple soldat dans les rangs de l'armée. » Il demande la réintroduction de son père, le conventionnel, un des 73 du 31 mai, dont il retrace une brève biographie. Il accepte de collaborer à la *Revue encyclopédique*, mais il refuse de s'occuper de la rédaction des tables; il demande des renseignements pour écrire une vie de Kosciusko, l'émule de Washington. Les lettres postérieures traitent plus particulièrement d'économie sociale.

**87547** BONAPARTE (Jérôme), le fils du roi Jérôme. — L. a. s. N. à M. Gasc; Paris, 31 octobre 1848, 2 p. 1/2 in-8°. 35 "

Curieuse lettre politique entièrement relative à la préparation de l'élection de son cousin à la présidence de la République. « Tous les chefs sont mal pour nous, nos vrais appuis sont les paysans, les ouvriers et les soldats, la tête de tous les partis nous est hostile, cependant les hommes d'ordre reviennent à nous. Cavaignac a pour lui le pouvoir, la corruption et l'intimidation; son cousin a pour lui le nom de l'Empereur et la sympathie des masses « qui voit (sic) que lui, mieux que tout autre peut sauver la République, l'établir à tout jamais en France ! »

**87548** BOSQUET (Pierre-Jean-François), maréchal de France, qui s'illustra à l'Alma, à Inkermann, et à l'assaut de la tour Malakoff, né à Mont-de-Marsan (Landes). — L. a. s. à M. Renoux; 12 juin 1851, 3 p. in-8°. 25 "

Intéressante lettre entièrement relative aux opérations militaires; il se plaint de la froideur de ses vieux camarades, qui jaloussent ses succès.

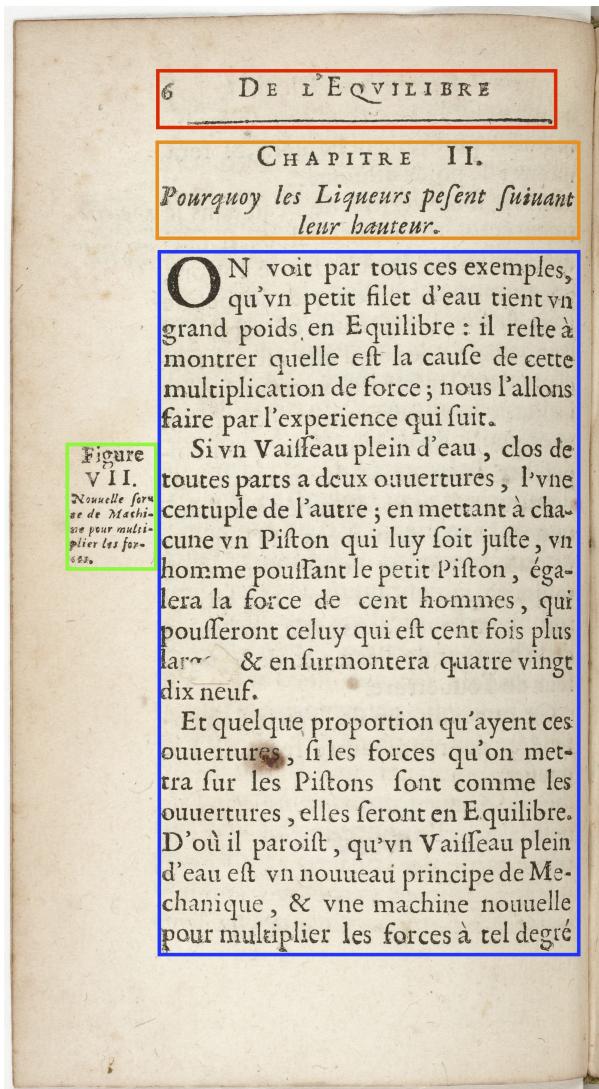
**87549** BRUANT (Aristide), chansonnier montmartrois, né à Courtenay (Loiret). — L. a. s., à un confrère; Le Pouldu, 15 août 1911, 1 p. in-8°. 5 "

Il dit que sa pièce les *Bas-fonds de Paris* est déjà faite; il l'a écrite en collaboration avec Arthur Bernède.

**87550** BRUEYS (François-Paul), vice-amiral, commandant en

# **Segmeteur**

# Segmentation: mise en page

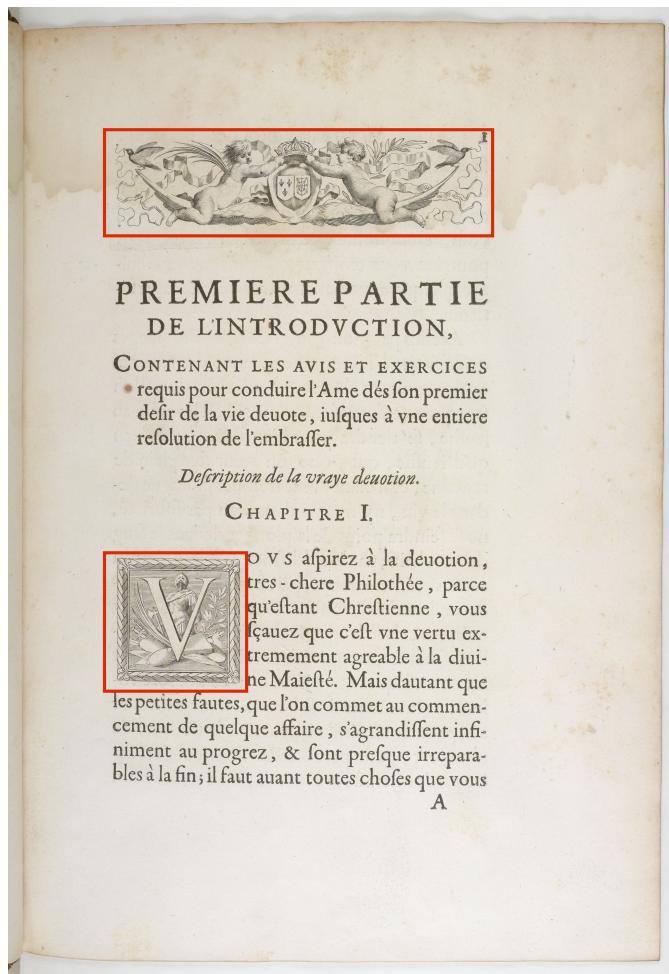


## Segmentation: lignes

87545 BIGOTTINI (Émilie), célèbre danseuse, qui eut de grands succès à l'Opéra. — L. a. s. à M<sup>me</sup> Guyet; 29 août 1833, 2 p. in-folio. 10 » Chaleureuse recommandation en faveur d'une femme de chambre.

L'OCR fonctionnant au niveau des lignes, il est fondamental de les extraire au mieux. L'utilisation d'outil dédié à la segmentation, et non d'un segmenteur intégré à l'OCR, peut être intéressant

# Segmentation: ornements



Le segmenteur peut extraire plus que des lignes: il peut extraire, par exemple, des ornements (bandeaux, initiales, culs-de-lampe...)

# Données

Il faut réussir décrire les documents OCRisés, afin de reconstruire l'apparence originelle sur la base des informations conservées. On privilégie pour cela des documents XML, *page driven*.

```
<document>
  <page>
    <zone>
      <ligne coordonnées="points">
        <mot coordonnées="points">exemple</mot>
      <ligne>
    </zone>
  </page>
</document>
```

# Données

Afin de faire le lien entre l'image et le texte, on doit donner une information géométrique. Celle-ci peut être de trois ordres: ligne, bloc, ou polygone.

Il existe des documents de niveau page, paragraphe, ligne ou mot.

Il existe aussi plusieurs formats: hOCR, Alto, PageXML... Ces formats sont normalement utilisés avec METS (*Metadata Encoding and Transmission Standard*) pour la description de l'objet numérisé.

# Exemple 1: Alto

ALTO: *Analyzed Layout and Text Object*

Développé lors du projet européen METAe (*Meta Data engine*, 2000-2003) et publié en 2004

Trois éléments centraux:

- <Description> contient les métadonnées
- <Styles> contient le texte
  - <TextStyle> contient les informations sur les fontes (famille, type, taille...)
  - <ParagraphStyle> contient la description des paragraphes (alignement gauche/droite, intelrigne)
- <Layout> contient le contenu, divisé en <Page>

```
<?xml version="1.0"?>
<alto>
    <Description>
        <MeasurementUnit/>
        <sourceImageInformation/>
        <Processing/>
    </Description>
    <Styles>
        <TextStyle FONTSIZE="10.0"/>
        <ParagraphStyle ALIGN="Left"/>
    </Styles>
    <Layout>
        <Page ID="P1" WIDTH="123" HPOS="123" VPOS="123">
            <PrintSpace WIDTH="123" HPOS="123" VPOS="123">
                <TextBlock ID="P1_TB1" WIDTH="123" ...>
                    <TextLine WIDTH="123" HPOS="123" ...>
                        <String WIDTH="123" ... CONTENT="Un">
                            <sp WIDTH="123" HPOS="123" VPOS="123">
                                <String WIDTH="123" ... CONTENT="Exemple">
                            </TextLine>
                        </TextBlock>
                    </PrintSpace>
                </Page>
            </Layout>
        </alto>
```

## Exemple 2: PageXML

PAGE: *Page Analysis and Ground-truth Elements*

Format créé lors du projet IMPACT EU (2010)

Contrairement à l'ALTO, PageXML conserve des informations sur le *pre-processing* (binarisation, deskew, dewarping...) et l'évaluation du layout.

```
<PcGts>
  <Metadata>...</Metadata>
  <page>
    <TextRegion type="paragraph" id="r_1">
      <Coords points="1474,486 3684,486 3684,900...">
      <TextLine id="l_1">
        <Coords points="1475,487 3683,487 3683,635...">
        <Baseline points="1475,635 1587,635 2061...">
        <Word id="w1">
          <Coords points="1475,497 1587,497 1587...">
          <TextEquiv>
            <Unicode>Un</Unicode>
          </TextEquiv>
        </Word>
        <Word id="w2">
          <Coords points="1935,497 2061,497 2061,619...">
          <TextEquiv>
            <Unicode>exemple</Unicode>
          </TextEquiv>
        </Word>
        <TextEquiv>
          <Unicode>Un exemple</Unicode>
        </TextEquiv>
      </TextLine>
    </TextRegion>
  </page>
</PcGts>
```

## Exemple 3: hOCR

Format XML *embedded* dans du XHTML/HTML

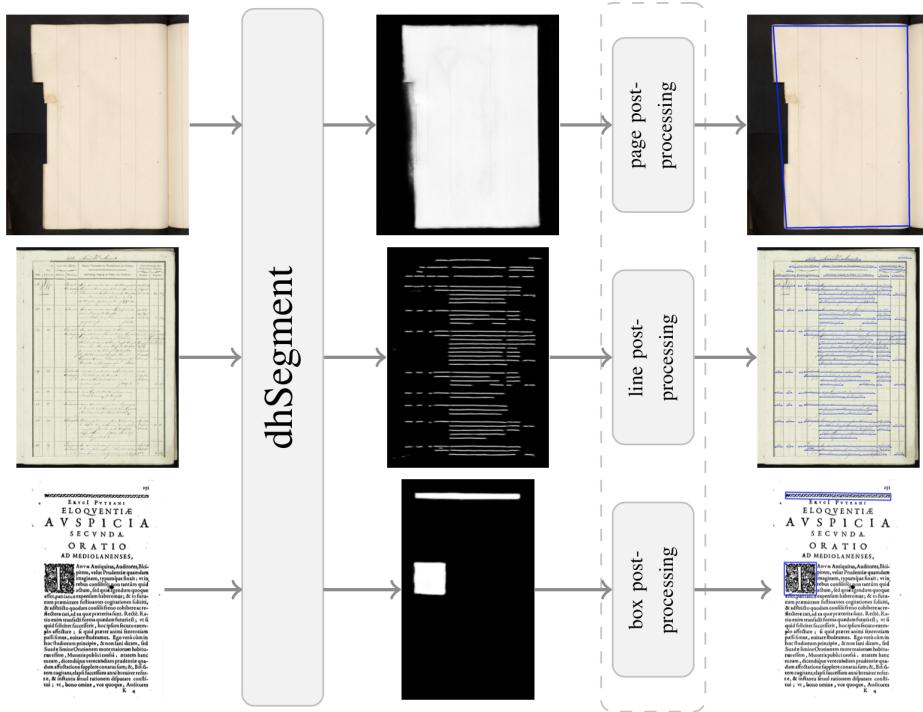
Trois grandes classes associées aux éléments html `<div>` , `<p>` ,  
`<span>`

- `ocr_page` pour les pages
- `ocr_par` pour les paragraphes
- `ocrx_line` pour les lignes
- `ocrx_word` pour les mots

L'information géométrique est stockée dans une bbox

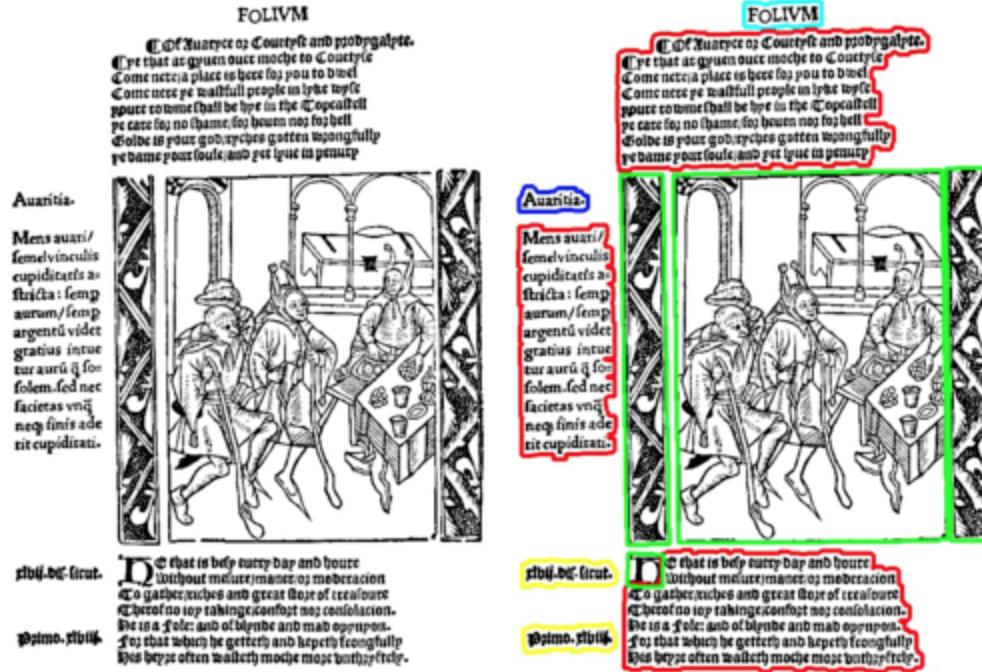
```
<?xml version="1.0" encoding="UTF-8"?>
<html xmlns="http://www.w3.org/1999/xhtml">
  <head>
    <title></title>
    <meta name='ocr-system' content='tesseract'/>
  </head>
  <body>
    <div class='ocr_page' id='page_1'
        title='bbox 0 0 1926 3102'>
      <div class='ocr_carea' id='block_1_1'
          title="bbox 638 108 756 147">
        <p class='ocr_par' id='par_1_1' lang='eng'
            title="bbox 638 108 756 147">
          <span class='ocr_line' id='line_1_1'
              title="bbox 638 108 756 147;
                      baseline 0 0">
            <span class='ocrx_word' id='word_1_1'
                title='bbox 638 108 756 147'>
              exemple
            </span>
          </span>
        </p>
      </div>
    </div>
  </body>
```

# DHsegment



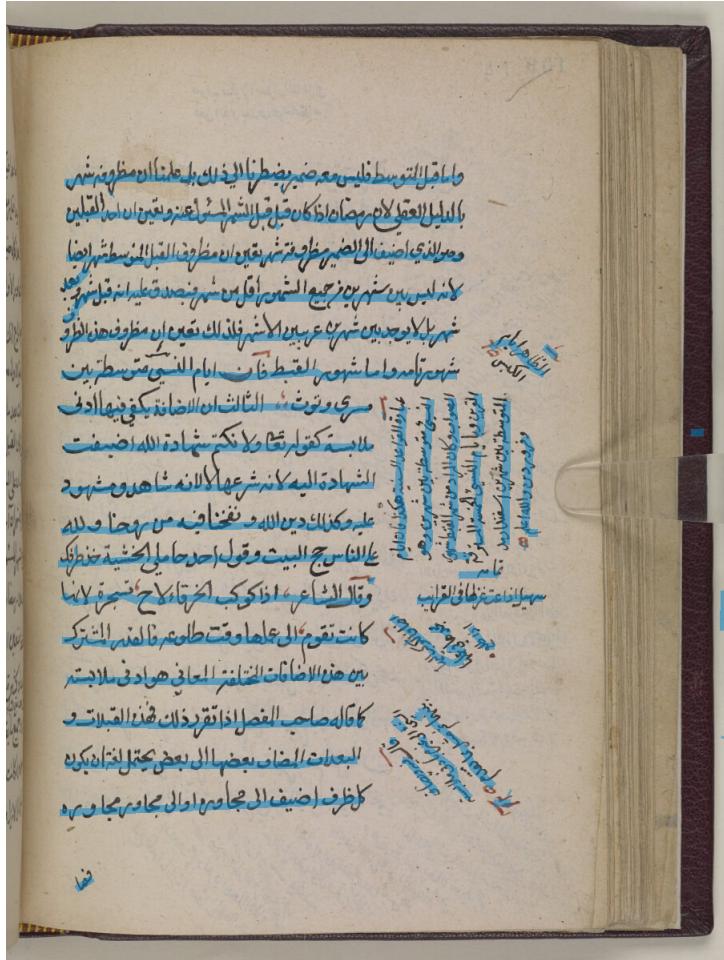
Sofia Ares Oliveira, Benoit Seguin, Frederic Kaplan, "dhSegment: A generic deep-learning approach for document segmentation", v.2, [arXiv:1804.10371](https://arxiv.org/abs/1804.10371)

# Larex



Reul, C., Springmann, U., and Puppe, F., "LAREX - A semi-automatic open-source Tool for Layout Analysis and Region Extraction on Early Printed Books", [arXiv:1701.07396](https://arxiv.org/abs/1701.07396)

# Kraken



Kiessling, B., Stökl Ben Ezra, D., Miller M., "BADAM: A Public Dataset for Baseline Detection in Arabic-script Manuscripts", HIP@ICDAR 2019.  
[arXiv:1907.04041](https://arxiv.org/abs/1907.04041)

**OCR**

# OCR

- *Optical character recognition*
- En français ROC (Reconnaissance optique de caractères).
- Extraire le texte d'une image.

# Transcrire (I)

Ligne de commandes + interface dans un navigateur

truc/0001/010001.bin.png

P R E F A C E.

PREFACE

truc/0001/010002.bin.png

où il estoit tombé, apres le refus qu'on luy avoit

ou il étoit tombé, apres le refus qu'on luy avoit

truc/0001/010003.bin.png

fait des armes d'Achille. Ils ont admiré le Phi-

fait des armes d'Achille. Ins ont admiré le Phi-

truc/0001/010004.bin.png

loctete , dont tout le sujet est Ulysse, qui vient

loctete, dont tout le sujet est Ulysse, qui vient

truc/0001/010005.bin.png

pour surprendre les fleches d'Hercule. L'Oe-

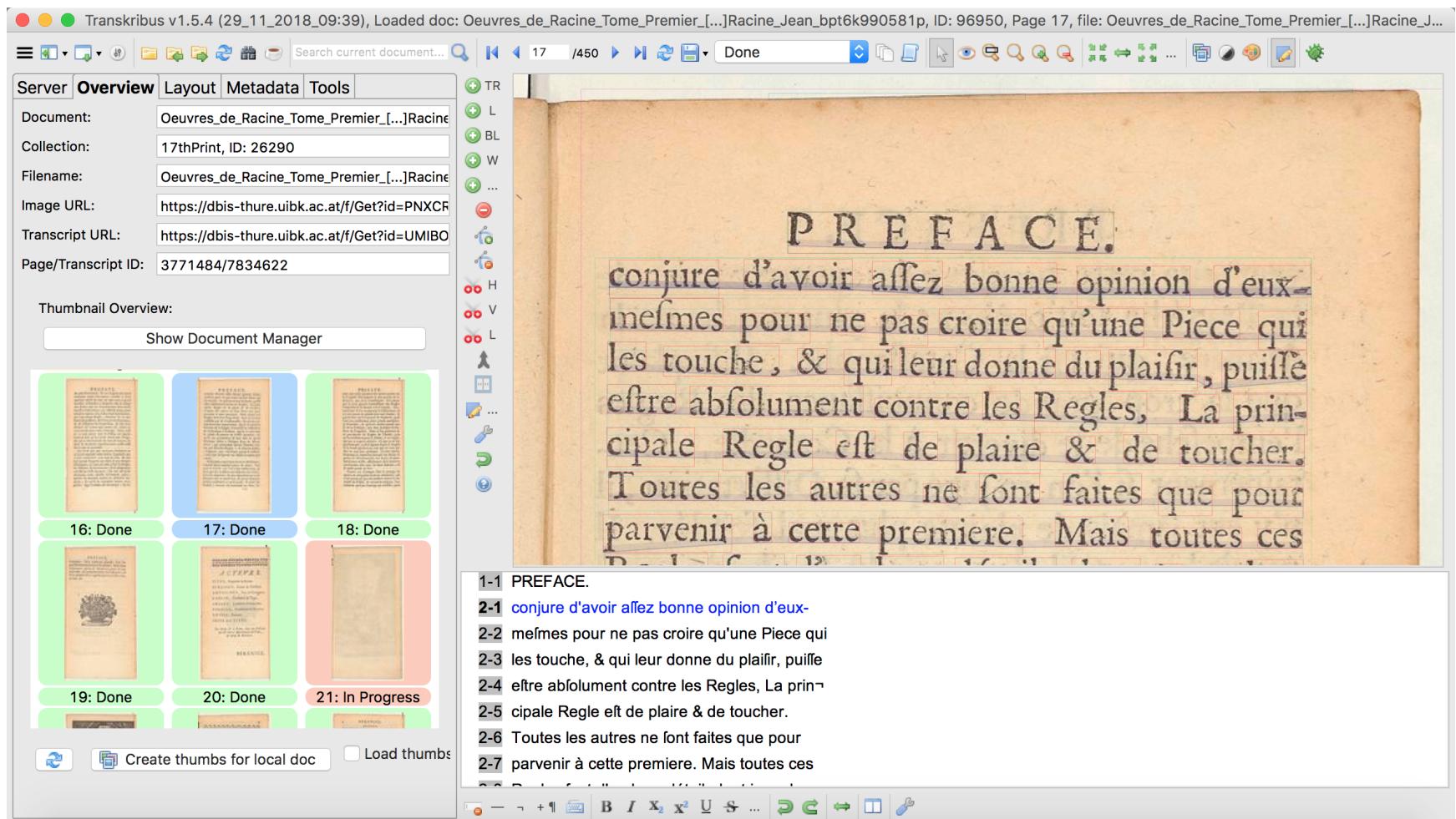
pour surprendre les fleches d'Hercule. L'Oe-

truc/0001/010006.bin.png

dipe mesme , quoy que tout plein de recon-

# Transcrire (II)

## Transkribus (Innsbruck)



## Transcire (III)

## eScriptorium (EPHE/PSL)

test Description Images Edit Element 1

Pause 00:00:00 Select Area Audio Record Pointer

kraken:Josephus\_abbrev\_with\_majuscules\_best.pronn

111 127

habent penitulam. Et faciem angusta  
fistula cantharis calamo patior praeberens  
laxitudinem quecumque maxima adstipendit  
spissitudine scilicet non iuxta hebreia lingua  
vacuar morta. factum est tales diuina curia  
quidem uerolens adiuuionem excollecentis  
populi mediocriterbrandi. Meritorum  
opere res adiectiandam dederunt necessariis  
principiis conuenire. Ambabus autem multi  
eudo colligebatur. Cum exordiendum  
miseretur haec fuit sumpniorum  
qui ab aliis adserentes admodum excedebat  
parum surgebant. Cum vero secunda fons  
iste quadamstrum refeldebatur apparabat  
ut censimur inuicta tabernacula  
dies procedendi scriribul percolabat  
ex aliis sequitur. Lemaueru  
omnes circuiterbancu erant. Tenuaueru  
sonans parique conseruantes adorante  
ferabant. Ex quatuor secessione  
Bucbanum autem laetulam inservientibus  
locais offerentes et tabulari exortavimus  
debet. Sacrificare ut rite prius perficit  
estum exagypio quod puthobat  
indeserto. & post pululat pectoris amato  
suis oculis quedam committit de quibus  
dicimus adhuc quidam oculis uenit  
ubimodo rursum fedime fecit xal  
fando moyser necessestis pectoris  
dicendo. & quoddam et bonis offiiciis  
exire. Iam illa pudorem pectoris  
quod, pectoris praebusum duerit  
inserit conseruans ex quo pro  
uera letitiamque uerum manefi  
cert enim lintonis solitudo perseruit.  
Cum malex feminis illius edocere  
anquidem moneret est uenit moyser  
laborum eius exhortat. pectoris canceru  
luminorebat debet. non doli adiuu  
re desperare. plebatus adhuc ponit  
et commone exponit moyser maioritur  
confusus. Moyserus confortans  
tali desperacione diceret. Licer turpiter

Aber fuisse consumelam passus exponimus  
est caro multitudinem plauri nonno  
de die plurimi illis autem hoc ni creditis  
et diuinorum unde posset haec antea  
multibz quae pectoris ministrare. Dicimus  
ergo malas uidentes auctor non tam pro rubeo  
edicimus operari est diuinum tardie pueris  
Cum dixit felix auctor etiam repletus est  
omnes exorti collegibz sicut exordium san  
cti. Dicimus nonne haec uerba de huius muri  
etem adhuc esse commissum non parus  
siquidem in tristitudine mortuus est. tene  
tent locuplere cognominatus ab aliis. Aliis  
quod interpræta potest defideris sepulturam  
ed dicuntur enim vel in terra que uocatur  
camillus uice dananorum cernuntur. sed  
tempore exadiuverandam diffidem populi  
meda congregatis citi incunione loquunt  
et dicens. Unde nobis dubio, pomerit  
liberorum est felicitas etas pectoris  
aliud edante unipossestis aliud et  
peccatum celeste. Insinuabitur namq[ue] fide  
mut cananeori acer brev[us] non decet p[ro]p[ter]e  
debet neq[ue] cuius felicitas omni  
corum gentes. Exercitum ergo p[ro]p[ter]e  
ad bellum. Non enim sine labore nobis  
terram conciderit fedimmo etiam p[ro]p[ter]e  
ut obremdimus. Mox etiam autem explo  
ratores qui huius terrae bona considerant  
et quae uirtus nescit habitantum. Ante  
omnia uenit animus tumur et deinde quiet  
membrum nobis adiutor expugandi et  
fecit hominem. Hac itaq[ue] candens  
fet moyser multitudine etiam horum exultat  
elegerat exploratores ducendum noscitur  
utroq[ue] unum deinceps tribu circumuenient  
omne terram dananorum usq[ue] apud eam  
aegyptum iacentibus usq[ue] dictu[m] et  
ex istis. Admetum libanum puerem  
considerant ualde conseruantes qui  
digna debet omne hoc p[ro]p[ter]e splendet  
Jup[iter] et fructus qui proficit illa terra

aben penecubitum. st autem a nigista  
fistula cantatoria calamo patior præbebris  
latitudinem quoem uenatori ad suis cptionem  
sp[iritu]s et clasicis sono uia que hebreia lingua  
uocatur asorsa. factesunt tales duæat una  
quidem utebantis adiessionem et collectionem  
populi inediis caelebrandam itera uero  
opertent ad cogitandum dederunt d[omi]n[u]s  
principes conuenire. Ambabus autem mul  
tudo colligebatur. Cumque tabernaculum  
moueretur a et flabunt isson ante p[ro]p[ter]e.  
qui habitabant ad orientem len tabernaculum  
partem surgebant. Cumvero secunda son  
isset qui adiustum residebant apparabat  
et ita inuenit in uoluata a b[ea]tissima  
qui d[omi]n[u]s n[ost]ris habet. inuenit  
tenus locuplise cognominatur cabrath. alia  
quod interpræta potest desideris sepultrum  
euens autem eo no[n] intersam queuoati  
cumillus iuxtagana neorum terminos edis  
ent et ad habitandum diffidem populi  
ineda congruavit eius incontione loquutus  
est dicens. Unde nobis duo bona p[ro]misit  
liberatem et felicitatem possessionem.  
aliud et dante. iam posseditis. aliud estis  
pectupi celesti. P[ri]mibus namque sede  
mus cananeorum. acerbitate nos decete p[re]  
dictes. neq[ue] rex neq[ue] ciu[us] sedneque. omnes  
eorum gentes. exercitum ergo pre parem  
a bellum. Non enim sine labore. nobis ha  
tem conceditur. sed ma ximis eam p[ro]p[ter]e  
illis ostendimus. M[er]ita tamen autem expli  
catores qui huius terrae bona considerent  
et. quae uirtus in easit habitantium. nt  
omnia ueo uni animis sumus. et d[omi]n[u]s quis  
in omnibus nobis ad iutis et pugnandis  
socius honestus. Haec itaque. cum dixis  
set mos. multitudine ei honorem exhibuit  
elegitque. ex ploratores ducendum notis simos  
uiros. unum deum queque tribu circum eunt  
mnem terram chana et n[ost]ra partibus cir  
aegyptum iacentibus usque. a dicitu ate  
eithin. ad montem libanum per uenient  
Naturamque. terrae et in colariu hominum  
considerantibus ualide conseruuntur sunt.  
dra ginta diebus omne hoc opus expletus  
super et fructus quos proficit illa terra

## Création d'une vérité de terrain (*ground truth*)

Les images transcrites sont alors associées à leur transcription

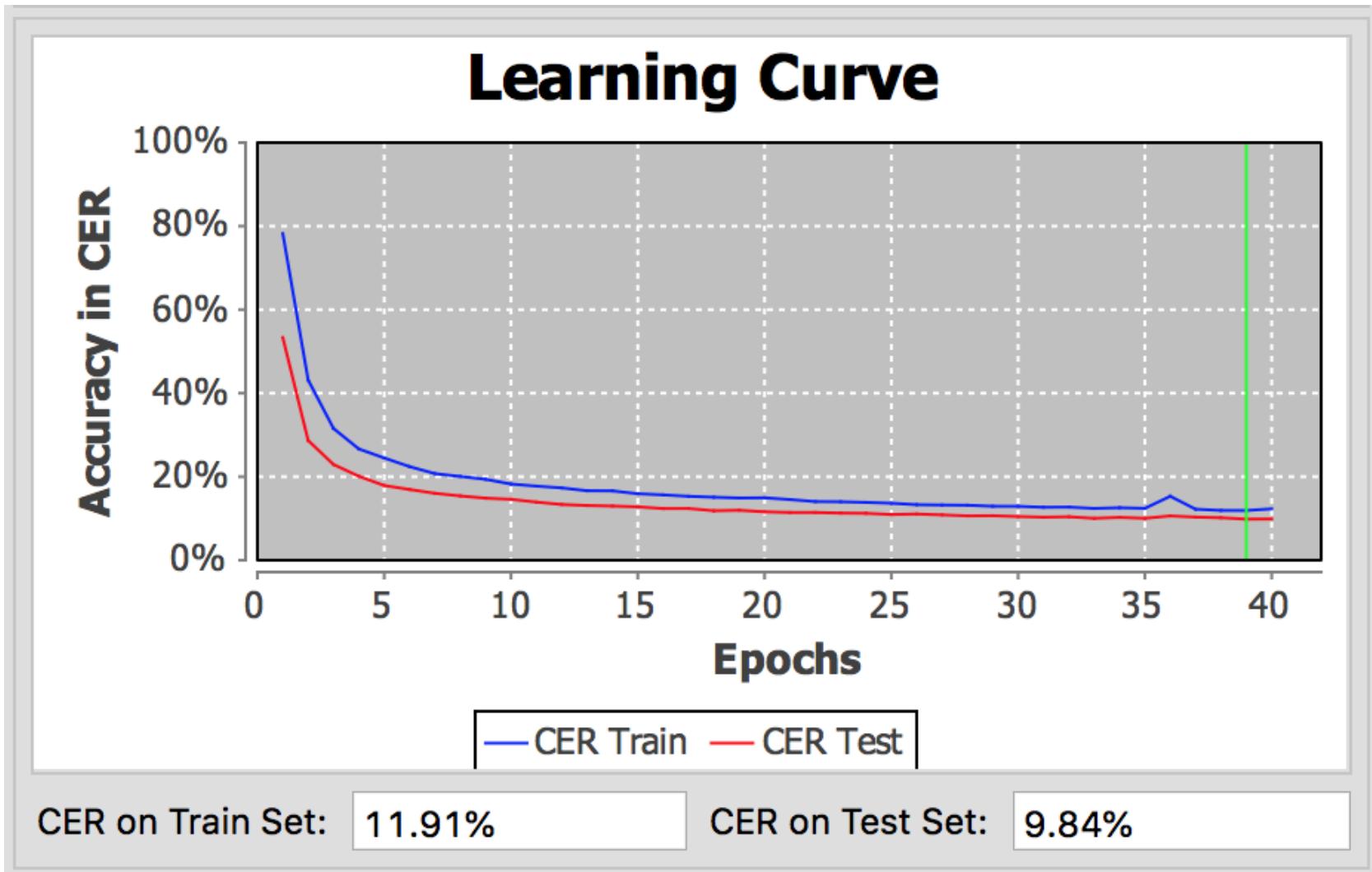


# Entraînement (I)

Comme c'est du *machine learning*, on va répéter l'entraînement une multitude de fois (on parle d'*epochs*, de *stages* ...). À chaque fois un modèle est créé: celui qui performe le mieux est conservé

```
Accuracy report (17429) 0.9610 4825 188
stage 15/∞ [#####
Accuracy report (18591) 0.9621 4825 183
stage 16/∞ [#####
Accuracy report (19753) 0.9606 4825 190
stage 17/∞ [#####
Accuracy report (20915) 0.9615 4825 186
stage 18/∞ [#####
Accuracy report (22077) 0.9602 4825 192
stage 19/∞ [#####
Accuracy report (23239) 0.9617 4825 185
stage 20/∞ [#####]
```

## Entraînement (II)



# Scores

- On parle de CER (*Character Error Recognition*) et parfois de WER (*Word Error Recognition*).
- Distance de Levenshtein : combien d'opérations pour retrouver le résultat attendu (par exemple entre tonte et toute) ?
- **Une seule lettre fausse crée un mot faux ! Le WER est donc toujours supérieur au CER !**
- Ces scores peuvent être calculés sur deux jeux de données :
  - Le train set (on OCRise les images qui servent pour l'entraînement)
  - Le test set (on OCRise des images qui n'ont pas servi pour l'entraînement)

## L'amélioration des scores: données artificielles

- Avec Baskerville

C'est ceux dont il est écrit au commen-

- Avec IM FELL English SC

VOUS ESTIMÉS QUELQUE CHOSE

- Avec JSL Ancient

reZ Lecteur (si je ne me trompe,) la

- Avec Chapbook

Tandis qu'autour de moy vostra Cour assemblée,

## L'amélioration des scores: bruit

- Original

C'est ceux dont il est écrit au commen-

- Bruit faible

C'est ceux dont il est écrit au commen-

- Bruit fort

C'est ceux dont il est écrit au commen-

## L'amélioration des scores: modification du cadre

- Cadre normal

C'est ceux dont il est écrit au commen-

- Cadre élargi

C'est ceux dont il est écrit au commen-

# Dans la jungle des outils

# Outils

- Tesseract
- Ocropy
- Kraken
- Calamari
- DHsegment
- ...

Il est souhaitable de préférer une solution qui intègre les différentes étapes nécessaires à l'OCrisation

## *Pipeline*

Il existe plusieurs solutions qui articulent tous les éléments nécessaires pour l'OCRisation

- Web: [eScriptorium](#)
- Docker: [ocr4all](#)
- Java: [Transkribus](#)

## *Pipeline : Comment choisir?*

Il existe plusieurs solutions qui articulent tous les éléments nécessaires pour l'OCRisation

- Web: [eScriptorium](#), *open source*
- Docker: [ocr4all](#), *open source*
- Java: [Transkribus](#), *non open source*

# Bibliographie

- Sami Nousiainen, *Report on File Formats for Hand-written Text Recognition (HTR) Material*, 2016, [en ligne](#)
- Kiessling, B., Stökl Ben Ezra, D., Miller M., "BADAM: A Public Dataset for Baseline Detection in Arabic-script Manuscripts", HIP@ICDAR 2019. [arXiv:1907.04041](#)
- Sofia Ares Oliveira, Benoit Seguin, Frederic Kaplan, "dhSegment: A generic deep-learning approach for document segmentation", v.2, [arXiv:1804.10371](#)
- Reul, C., Springmann, U., and Puppe, F., "LAREX - A semi-automatic open-source Tool for Layout Analysis and Region Extraction on Early Printed Books", [arXiv:1701.07396](#)

