

Transcrire (automatiquement)

Alexandre Bartz, 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>
```

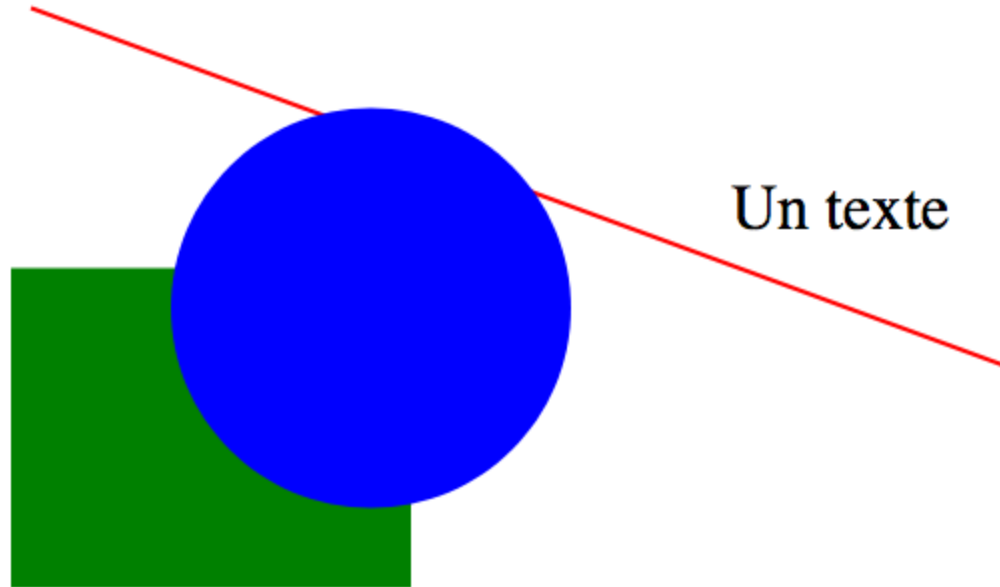


Image vectorielle (III)

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

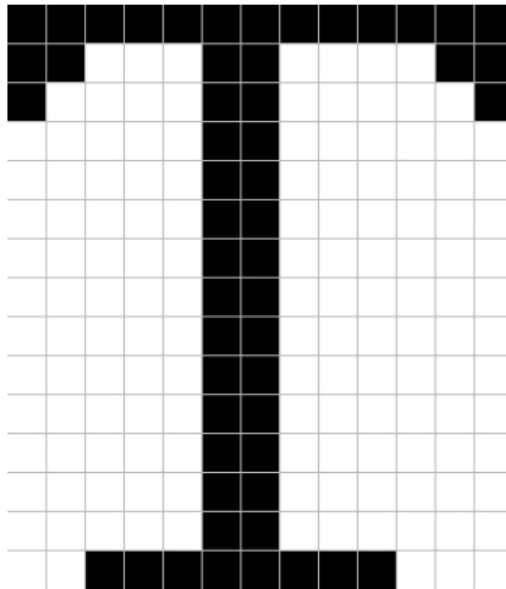
Pour plus d'exemples, allez regarder du côté de [w3schools](https://www.w3schools.com).

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 (II)

Deux fois la même image matricielle

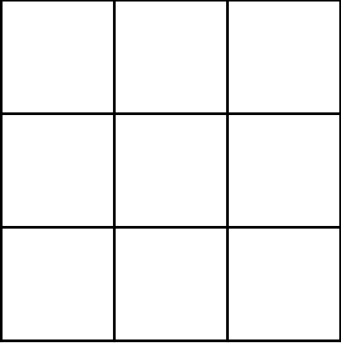
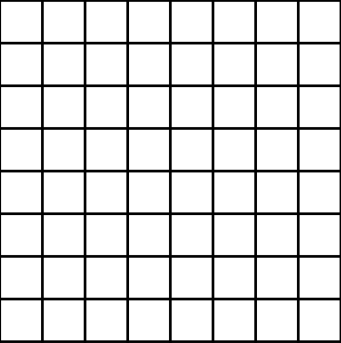
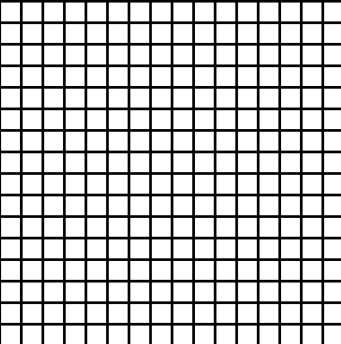


1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	0	0	0	1	1	0	0	0	0	1	1		
1	0	0	0	0	1	1	0	0	0	0	0	1		
0	0	0	0	0	1	1	0	0	0	0	0	0		
0	0	0	0	0	1	1	0	0	0	0	0	0		
0	0	0	0	0	1	1	0	0	0	0	0	0		
0	0	0	0	0	1	1	0	0	0	0	0	0		
0	0	0	0	0	1	1	0	0	0	0	0	0		
0	0	0	0	0	1	1	0	0	0	0	0	0		
0	0	0	0	0	1	1	0	0	0	0	0	0		
0	0	0	0	0	1	1	0	0	0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

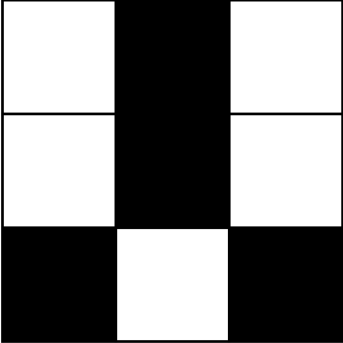
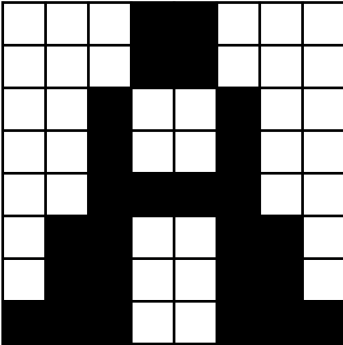
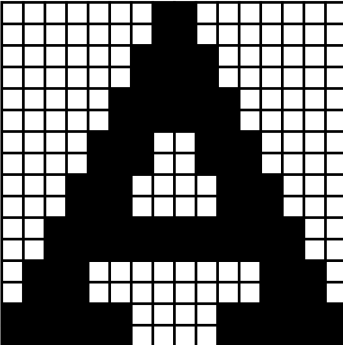
Les caractéristiques techniques d'une image

- 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

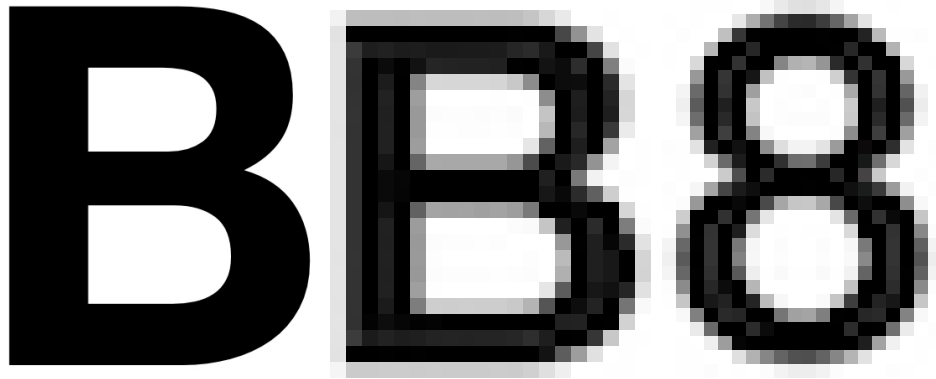
Résolution d'une page A4: $(\text{dpi} \times (21/2.54)) \times (\text{dpi} \times (29.7/2.54))$

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

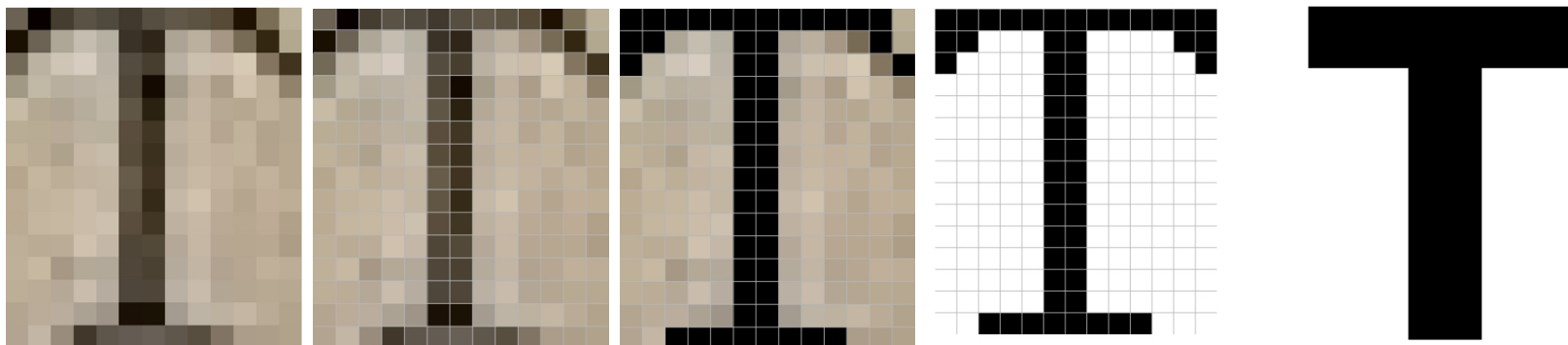
L'enjeu est de réussir ce que nous appellerons le *B*-test



Résolution vs efficacité

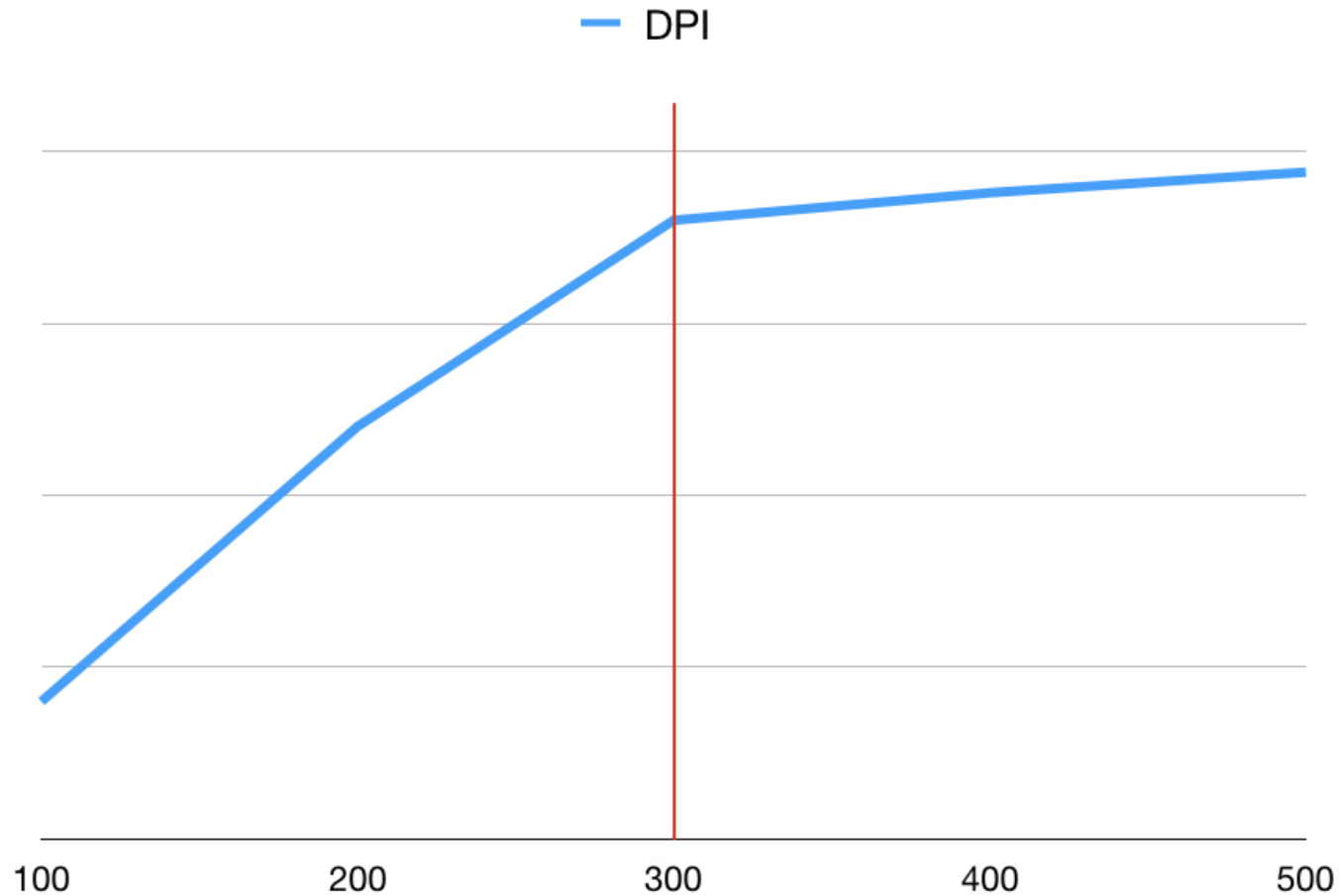
Il n'est pas nécessaire d'avoir un grand nombre de pixels (au contraire) pour bien faire fonctionner un OCR.

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



La bonne résolution

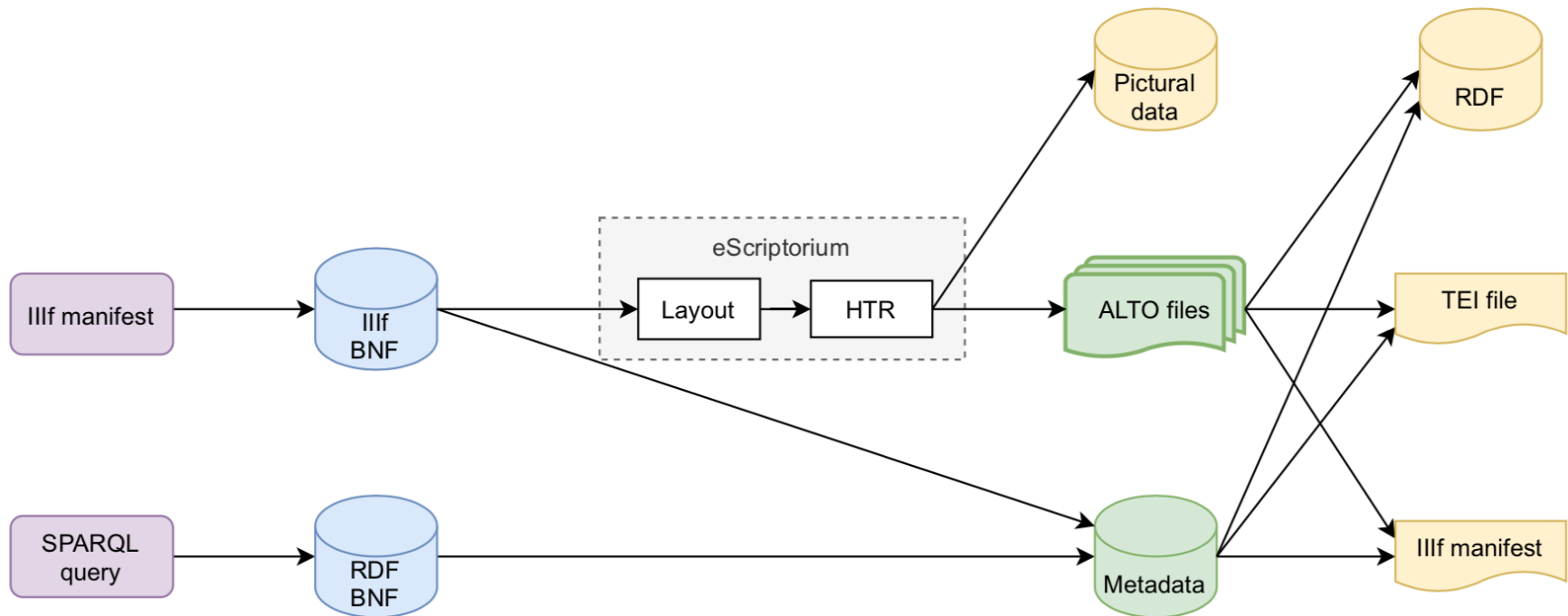
300 dpi serait le meilleur rapport poids/qualité



Récupération d'images automatisée

Chaîne de traitement

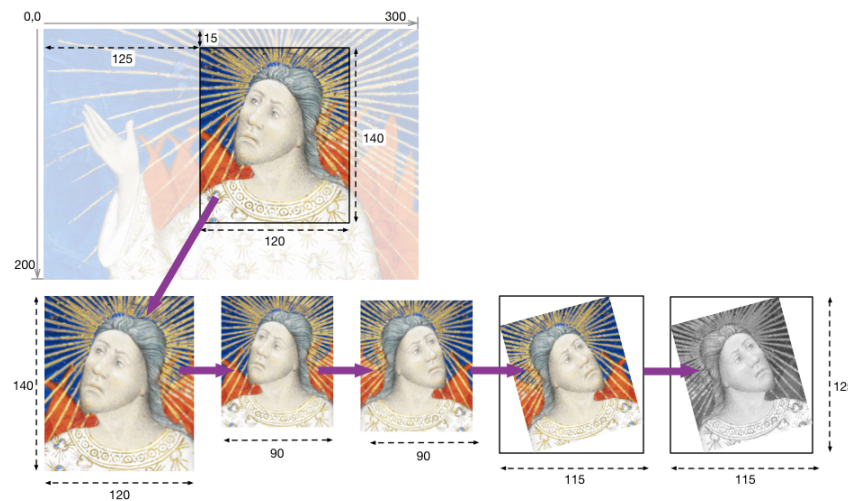
Projet Gallic(orpor)a: A. Pinche, J.-B. Camps, S. Gabay, B. Sagot, R. Bawden, P. Ortiz Suárez



IIIF

IIIF est un ensemble de spécifications techniques dont l'objectif est de définir un cadre d'interopérabilité pour la diffusion et l'échange d'images haute résolution sur le web.

The IIIF Image API specifies a web service that returns an image in response to a standard HTTP or HTTPS request. The URI can specify the region, size, rotation, quality characteristics and format of the requested image. (cf. [IIIF documentation](https://iiif.io))



IIIF: métadonnées

L'URI d'un manifeste est la suivante:

{scheme}://	{server}	/ {prefix}	/ {identifiant}
https://	gallica.bnf.fr	/iiif	/ark:/12148/bpt6k313644d

<https://gallica.bnf.fr/iiif/ark:/12148/bpt6k313644d/manifest.json>

Ici l'identifiant est un ark (cf. explications [ici](#)).

Pour récupérer les métadonnées d'une image:

{scheme}://	{server}	/ {prefix}	/ {identifiant}
https://	gallica.bnf.fr	/iiif	/ark:/12148/bpt6k313644d/f1

<https://gallica.bnf.fr/iiif/ark:/12148/bpt6k313644d/f1/info.json>

IIIF: données

<code>/ {region}</code>	<code>/ {size}</code>	<code>/ {rotation}</code>	<code>/ {quality}</code>	<code>/ {format}</code>
<code>/50,50,1500,1500</code>	<code>/full</code>	<code>/0</code>	<code>/bitonal</code>	<code>.jpg</code>

<https://gallica.bnf.fr/iiif/ark:/12148/bpt6k313644d/f15/50,50,1500,1500/full/0/bitonal.jpg>

pour extraire une image:

- commençant à 50 sur l'axe horizontal, 50 sur l'axe vertical, de 1500 pixels de largeur et 1500 pixels de hauteur
- en pleine résolution
- sans rotation
- en noir et blanc
- au format `.jpg`

pour extraire une image commençant à 0 sur l'horizontal, 1900 sur l'axe vertical, de 2400 pixels en largeur et 1200 pixels en hauteur, à laquelle on applique ensuite une rotation de 90°.

/ {region}	/ {size}	/ {rotation}	/ {quality}	/ {format}
/0,1900,2400,1200	/full	/90	/native	.tif

<https://gallica.bnf.fr/iiif/ark:/12148/bpt6k313644d/f15/0,1900,2400,1200/full/90/native.tif>

pour redimensionner l'image originale en une nouvelle image de 750 pixels en largeur, à laquelle on applique ensuite une rotation de 135°.

/ {size}	/ {region}	/ {rotation}	/ {quality}	/ {format}
/full	750,	/135	/gray	.pdf

<https://gallica.bnf.fr/iiif/ark:/12148/bpt6k313644d/f15/full/750,/135/bi-tonal.pdf>

Exercice:

1. utiliser IIIF: ark:/12148/bpt6k1280589b/
2. Installez l'add-on IIIF si vous avez Firefox:
<https://addons.mozilla.org/fr/firefox/addon/iiif-download/>
3. Créer un projet sur eScriptorium et importer le fichier via IIIF

Charger les documents

Dans le dossier `2_1_introduction_exercice` se trouvent dix lots contenant chacun une mazarinade au format pdf et un lien vers son *manifest* IIIF.

- Commencer par vous répartir les documents (un par personne) ;
- Essayer dans un premier temps de charger le document via IIIF (le `manifest` se trouve dans le [README.md](#)) ;
- Si ça ne fonctionne pas, charger le document au format pdf.