Ancient documents bleed through evaluation and its application for predicting OCR error rates

V. Rabeux, N. Journet, J.P. Domenger

LaBRI
Laboratoire **B**ordelais de **R**echerche en **I**nformatique
France (Bordeaux)

March 14, 2012

Research and work context

The project

Quality evaluation of very old document images by providing meta-data caracterizing a document's defects.

Why document image quality evaluation (instead of just restoration) ?

- **Avoid** (simplify) **manual analysis** of quality,
- **Drive** restoration algorithm, and **avoid restoration** of images that **don't** need it,
- **Predict** OCR error rates and other processes,

What this presentation aims to

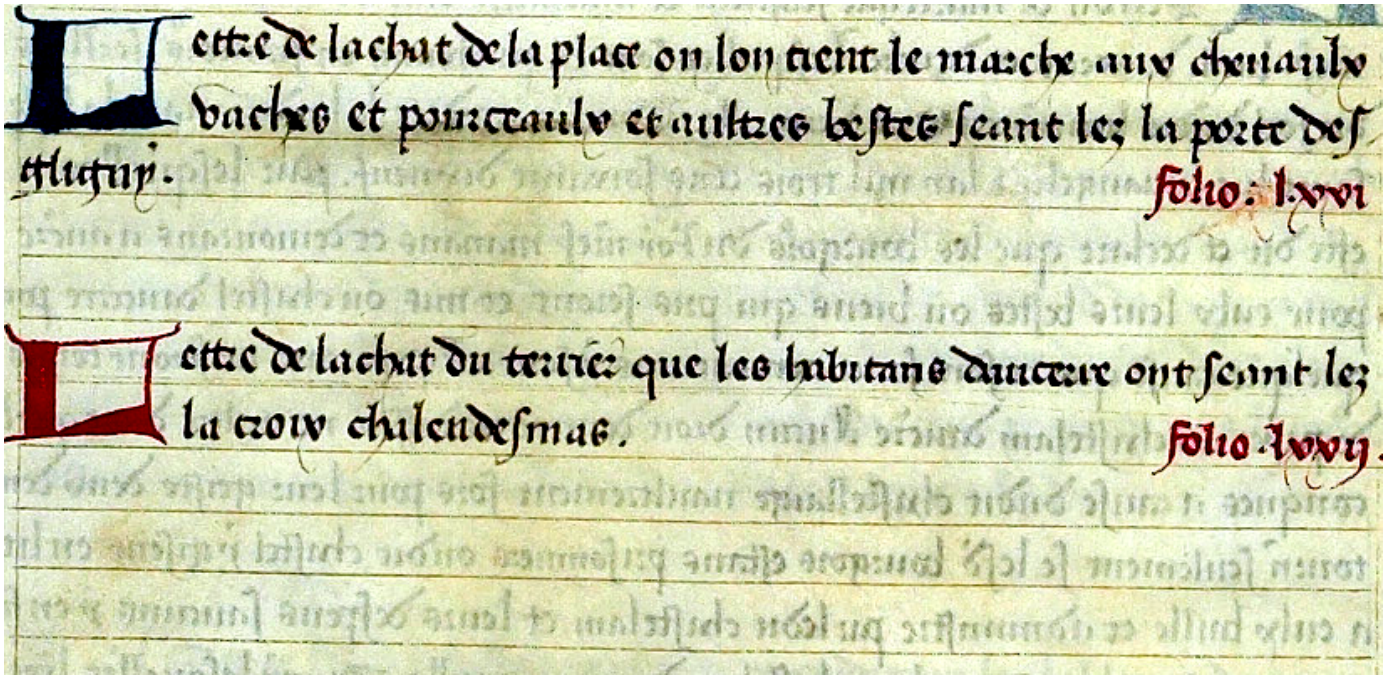


Figure: An old french document with bleed through

- Propose **measures** able to evaluate bleed through,
- Illustrate the measures accuracy by **predicting** the OCR error rate.

Measuring step by step

Our approach is composed of several steps :

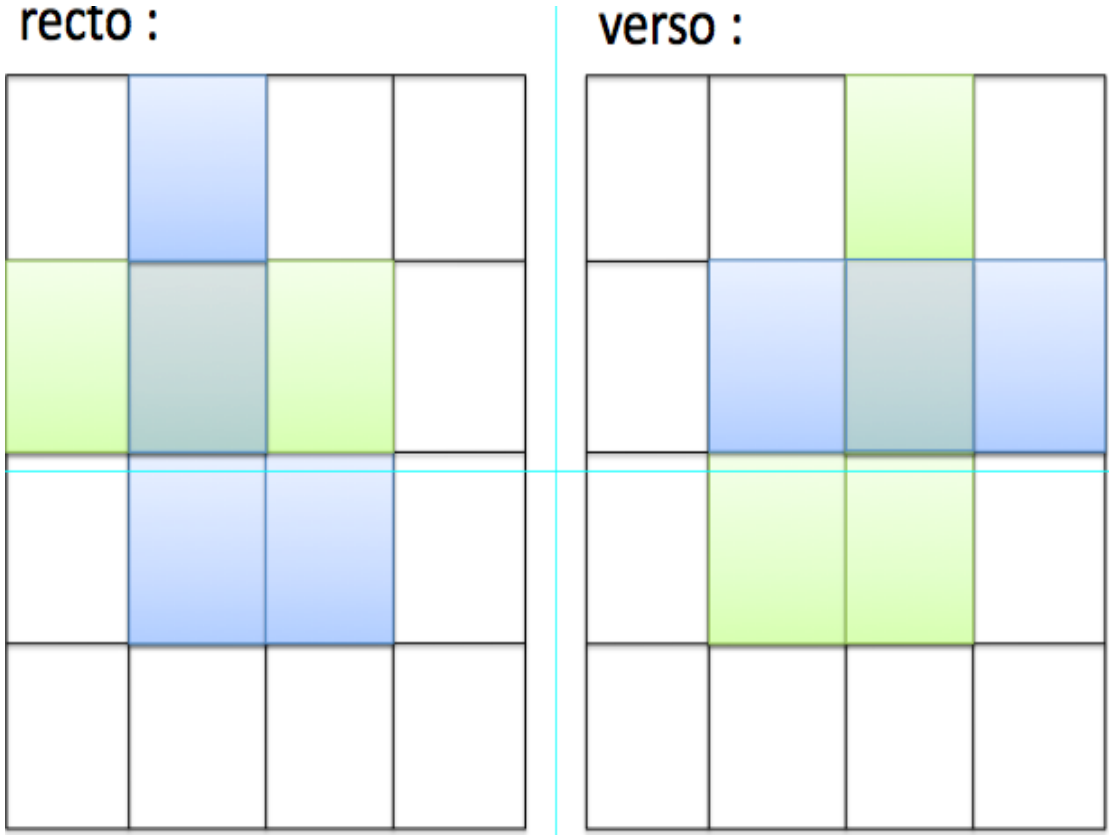
1. recto and verso **registration**,
2. **identification** of ink and bleed through pixels,
3. measures **computation**.

After the computation :

The bleed through's page is characterized by six measures.

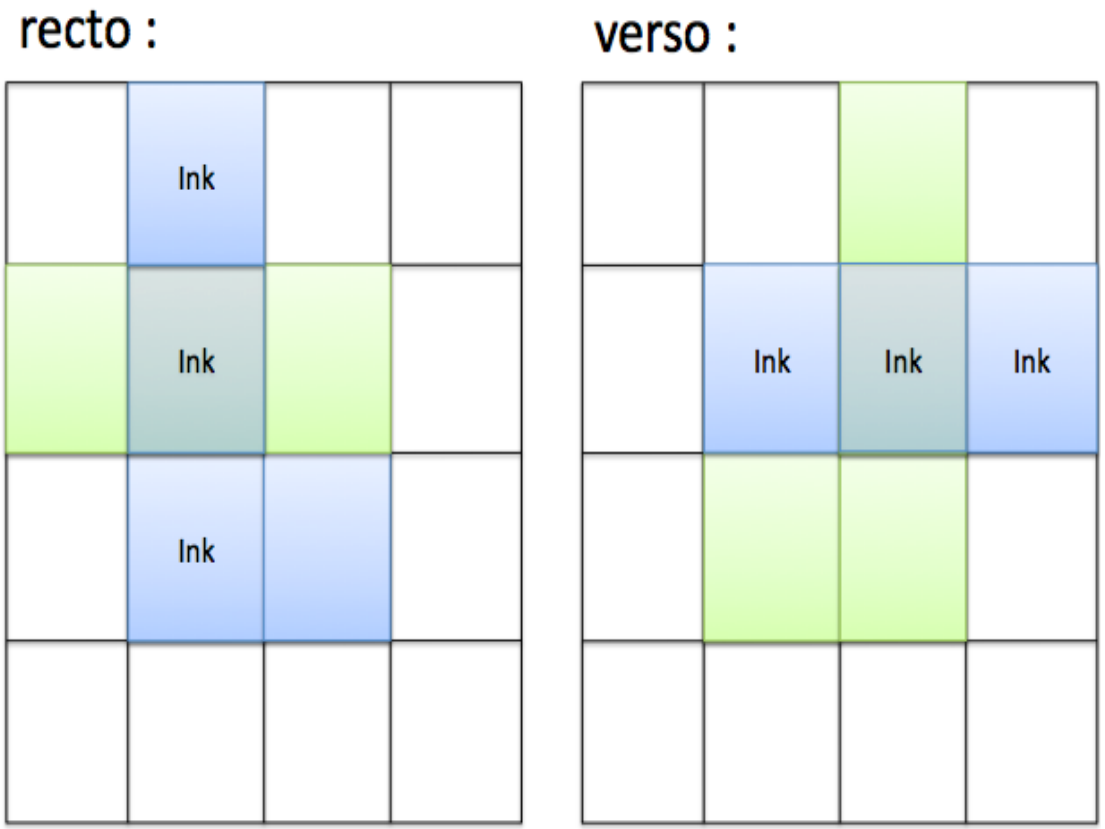
Bleed through identification

The verso and recto images are binarized and registered :



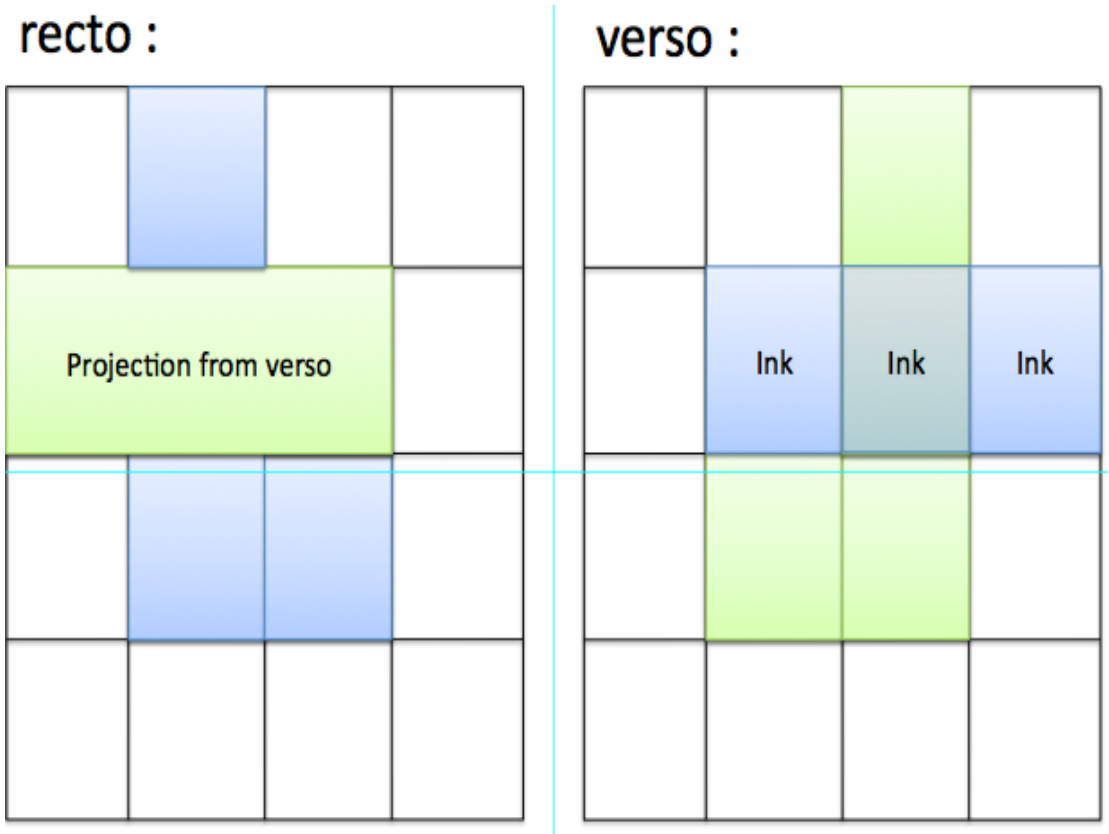
Bleed through identification

The verso and recto images are binarized and registered :



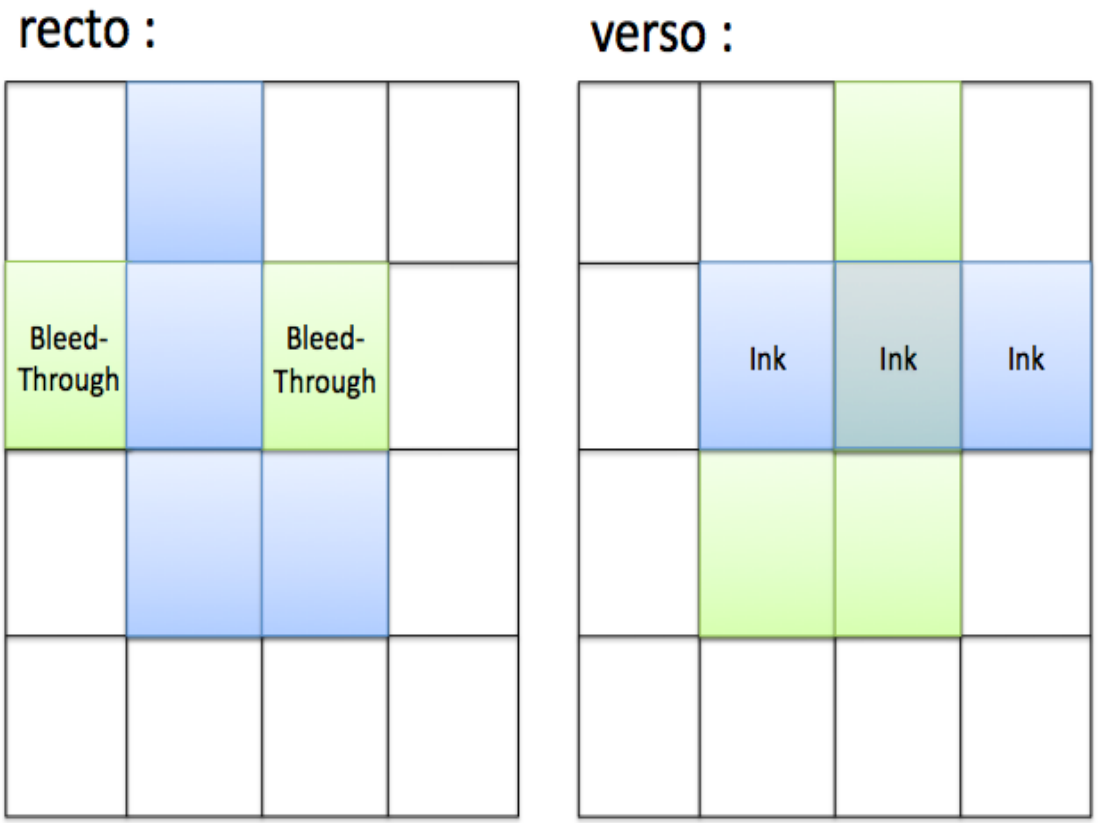
Bleed through identification

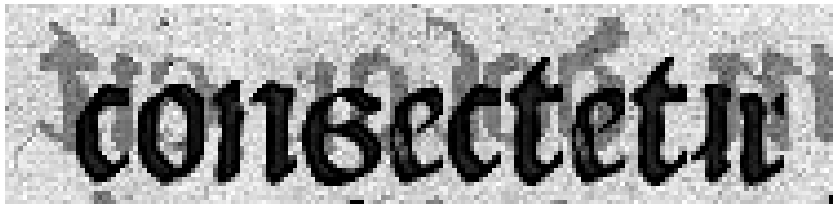
The verso and recto images are binarized and registered :



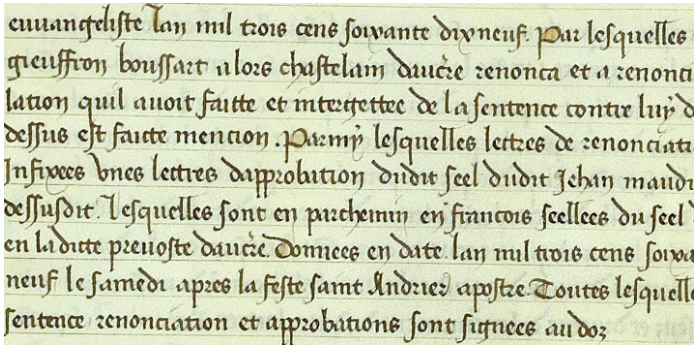
Bleed through identification

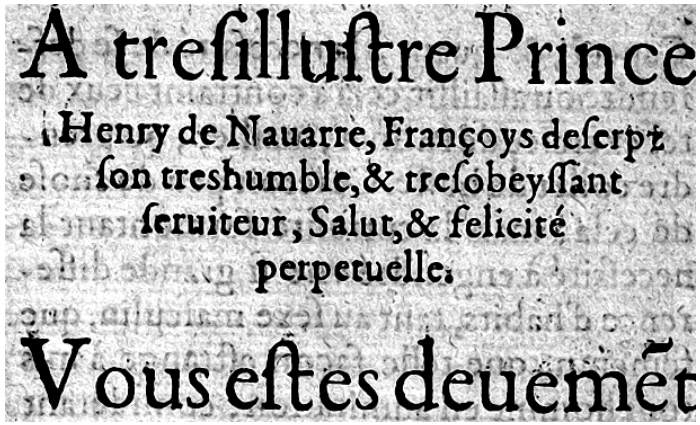
The verso and recto images are binarized and registered :



- Recto : 
- Verso : 
- Bleed through : 

Bleed through characteristics

a. 

b. 

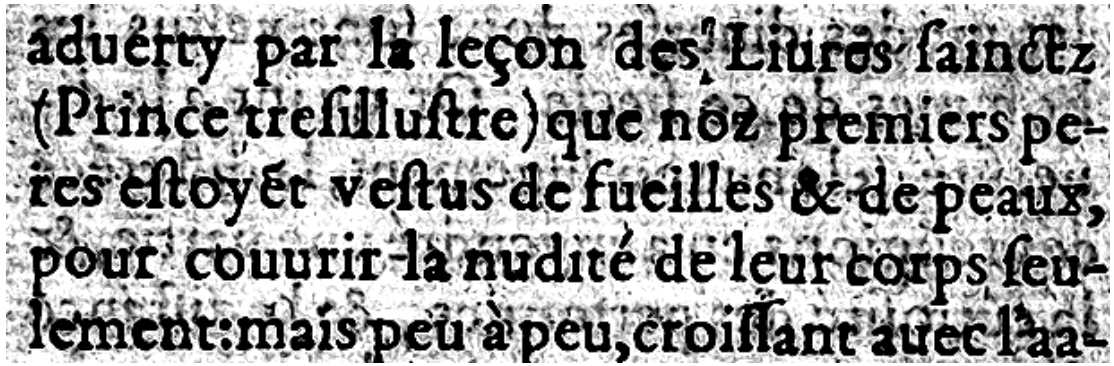
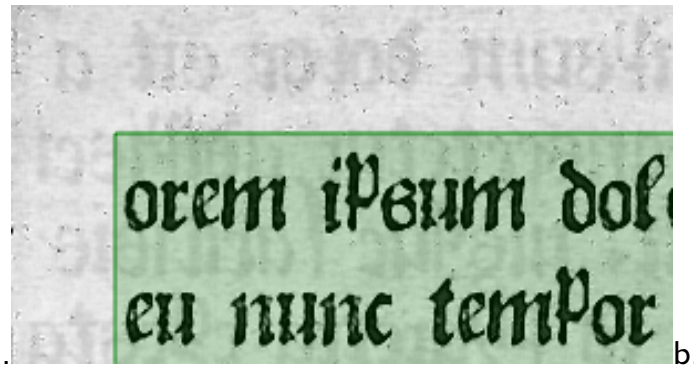
c. 

Figure: Bleed through characteristics : a.**intensity**, b.**quantity** and c.**location**.

Bleed through characteristics

Metrics 1 and 2 : Bleed through intensity

Measures the bleed through intensity in relation to the background or ink.

a. 

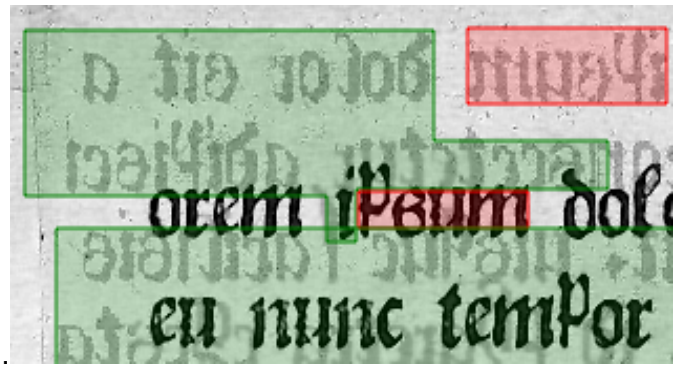
b. 

Figure: OCR errors due to the bleed through intensity (green zones correspond to recognized text; red zones correspond to unrecognized text)

Bleed through characteristics

Metrics 1 and 2 : Bleed through intensity

Measures the bleed through intensity in relation to the background or ink.

- Distance to the ink :

$$\mathcal{MI}_i = \frac{\mu_{T_r} - \mu_{I_r}}{255}$$

- bleed through is close to ink : 0,
- bleed through is far from ink : 1.

Bleed through characteristics

Metrics 1 and 2 : Bleed through intensity

Measures the bleed through intensity in relation to the background or ink.

- Distance to the ink :

$$\mathcal{MI}_i = \frac{\mu_{T_r} - \mu_{I_r}}{255}$$

  - bleed through is close to ink : 0,
  - bleed through is far from ink : 1.

- Distance to the background :

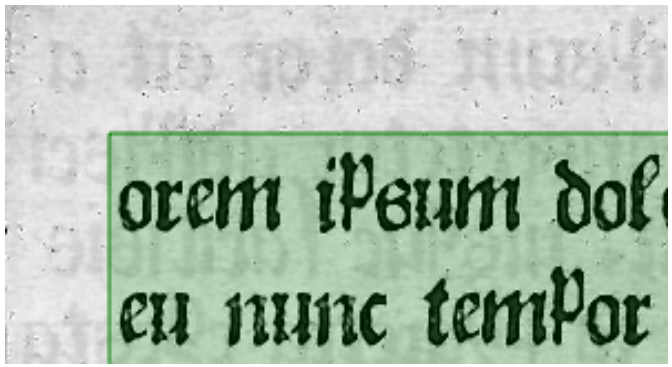
$$\mathcal{MI}_b = \frac{\mu_{B_r} - \mu_{T_r}}{255}$$

  - bleed through is close to background : 0,
  - bleed through is far from background : 1.

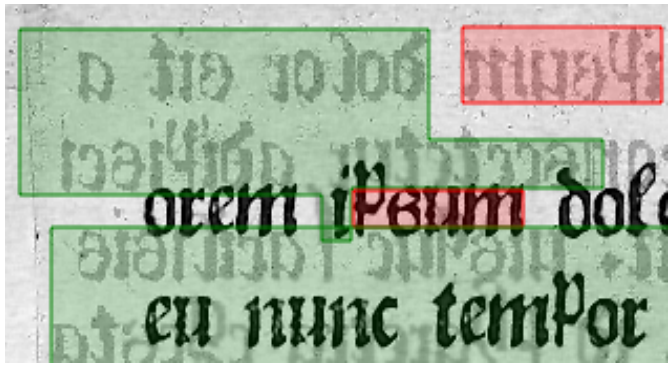
Bleed through characteristics

Metrics 1 and 2 : Bleed through intensity

Measures the bleed through intensity in relation to the background or ink.



$\mathcal{MI}_i = 0.6,\ \mathcal{MI}_b = 0.07$



$\mathcal{MI}_i = 0.4,\ \mathcal{MI}_b = 0.3$

Bleed through characteristics

Third metric : the bleed through quantity

Measure the quantity of bleed through in relation to the quantity of ink.

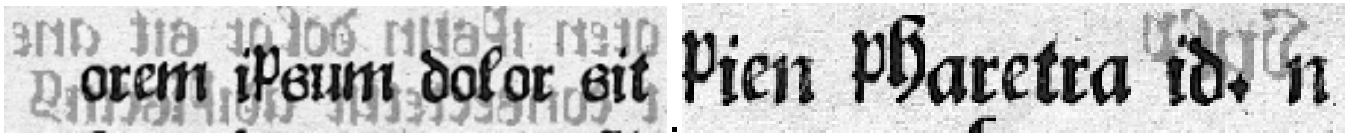


Figure: Variable bleed through quantity

Bleed through characteristics

Third metric : the bleed through quantity

Measure the quantity of bleed through in relation to the quantity of ink.

- $\mathcal{MQ}$: bleed through quantity ratio,

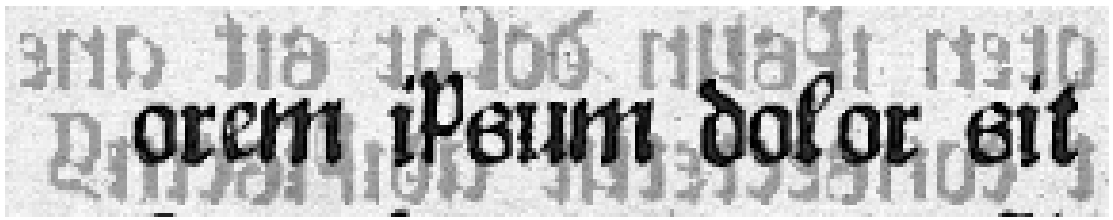
$$\mathcal{MQ} = \frac{\| T_r \|}{\| I_r \|}$$

- $\mathcal{MQ} > 1$: more bleed through than text.

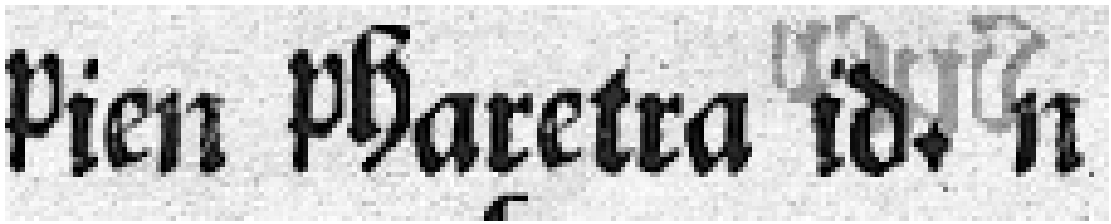
Bleed through characteristics

Third metric : the bleed through quantity

Measure the quantity of bleed through in relation to the quantity of ink.



$$\mathcal{MQ} = 1.73$$



$$\mathcal{MQ} = 0.3$$

Bleed through characteristics

Metrics 4, 5 and 6 : the bleed through location

Measure the bleed through location impact on letters.

a. b.

Figure: Locations of a bleed through component.

Bleed through characteristics

Metrics 4, 5 and 6 : the bleed through location

Measure the bleed through location impact on letters.

- $\mathcal{MA}$: components added by bleed through,

$$\mathcal{MA} = \frac{\|\overline{TC}\|}{\|\underline{I_v}\|}$$

- $\mathcal{MA} = 1$: No verso component overlap recto's components.

Bleed through characteristics

Metrics 4, 5 and 6 : the bleed through location

Measure the bleed through location impact on letters.

- $\mathcal{MS}$: letters having their shape modified by a bleed through component,

$$\mathcal{MS} = \frac{\|TC\|}{\|\underline{I_r}\|}$$

- $\mathcal{MS} = 1$: All recto's letters overlap with a bleed through component.

$\mathcal{MA}$ and $\mathcal{MS}$ example :

$\mathcal{MA} = 1$, $\mathcal{MS} = 0$ (no overlaps).

$\mathcal{MA} = 0$, $\mathcal{MS} = 1$ (E and K does overlap).

$\mathcal{MSG}$ measure

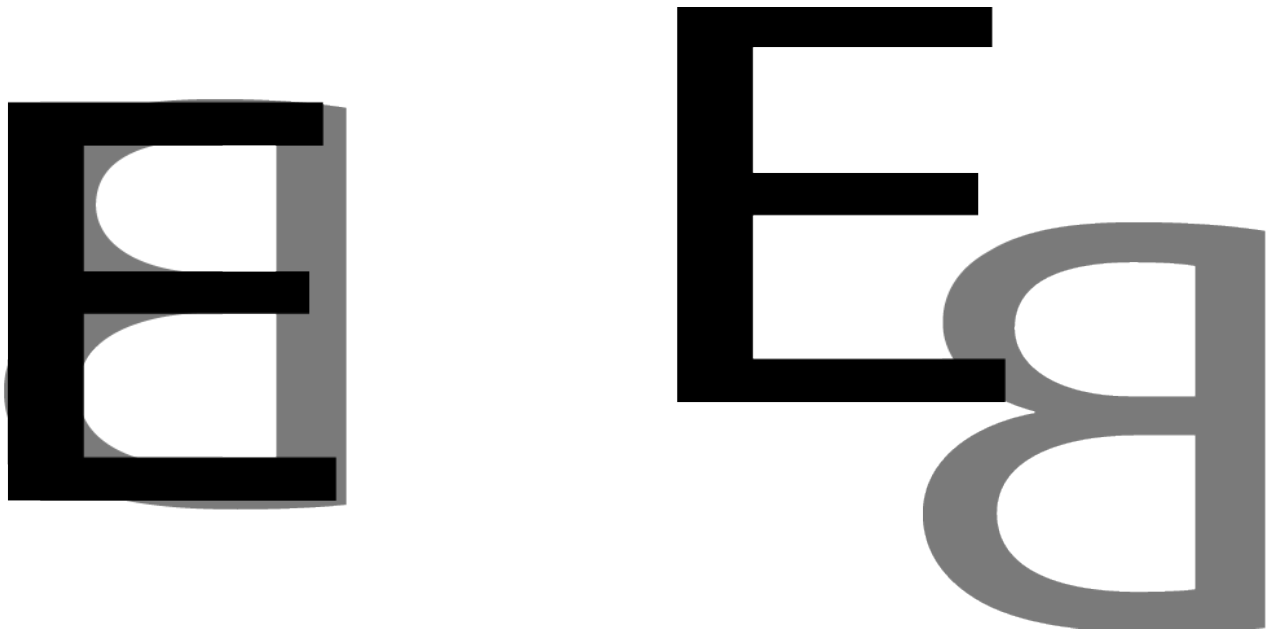


Figure: $\mathcal{MSG}$: measures the mean component expansion (in terms of component area).

$\mathcal{MSG}$ measure

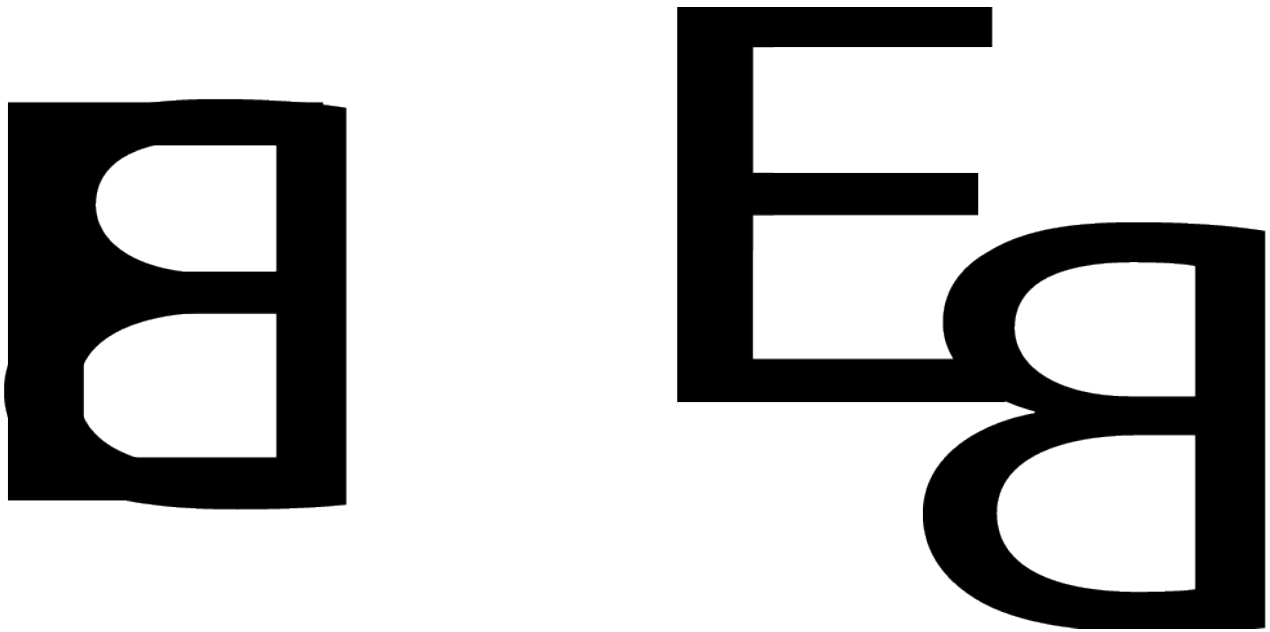


Figure: $\mathcal{MSG}$: measures the mean component expansion (in terms of component area).

Approach protocol

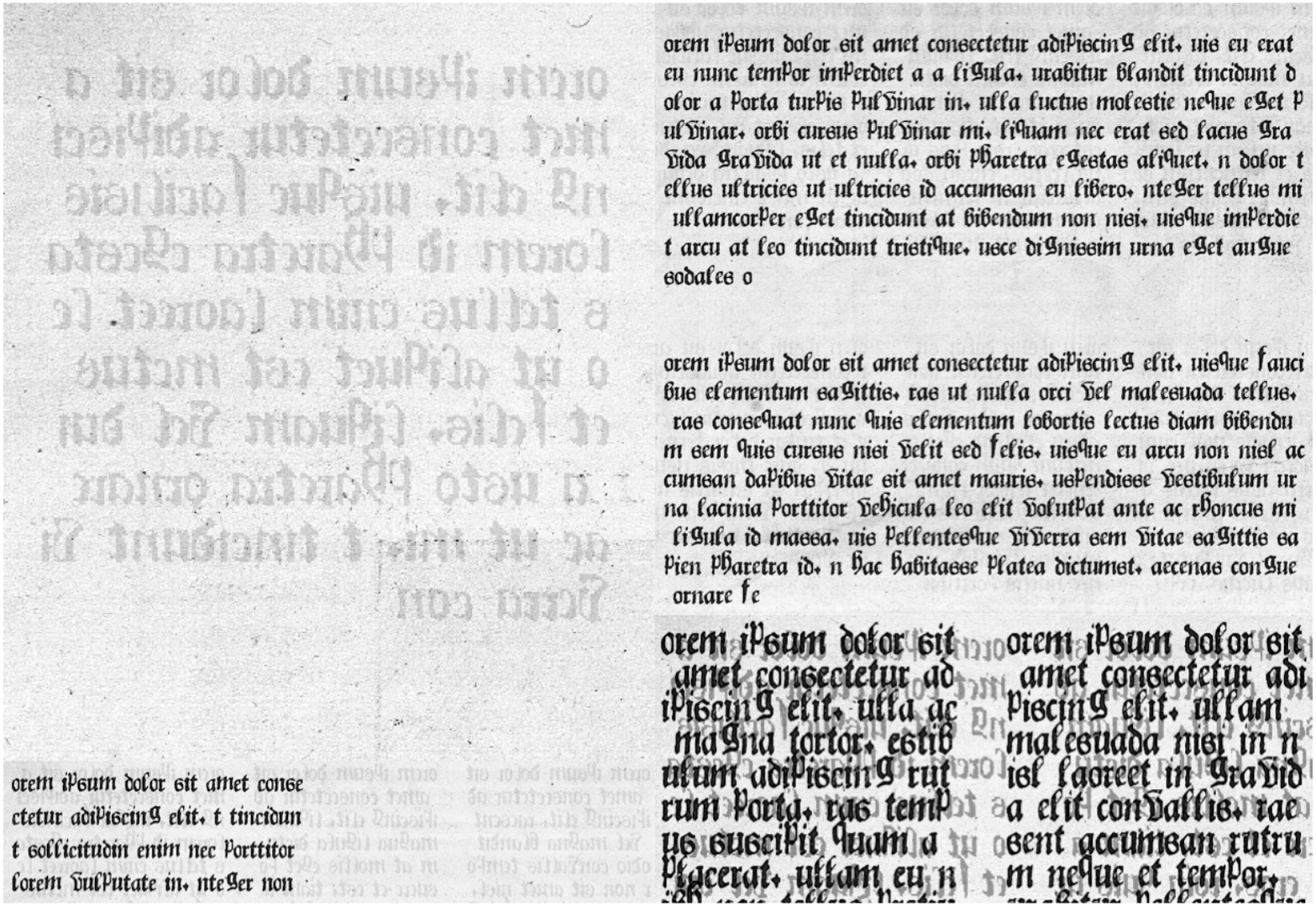
Metrics relevance :

Analyze bleed through metrics in correlation to the OCR error rate.

For a given OCR and dataset :

1. **Measures** computation (6 measures / page),
2. OCR runs and **error rates** computation,
3. **Linear regression** multivariate and sequential (leads to a prediction model).
4. Statistical **validation** of the model .

Example of generated documents



Results

Statistical model accuracy

OCROPUS: Coefficient of determination $R^2 = \mathbf{0.99}$,
Standard error (RMSE) = **7,5**

ABBYY: Coefficient of determination $R^2 = \mathbf{0.97}$,
Standard error (RMSE) = **12,77**

Statistical validation

Correlation coefficient between ground truth and the prediction :

OCROPUS **0.99** ABBYY **1.006**

Bleed through has a strong effect on the OCR process. High prediction accuracy if bleed through is the only defect.

Perspectives :

Conclusion :

- We proposed **measures** in order to evaluate **bleed through**,
- We demonstrated that the **bleed through has** a strong **effect on OCR** results.
- The model accuracy shows the **relevance** of our measures.

Perspectives :

- We **can not predict OCR results** with just bleed through measures.
- Characters quality and noise evaluation, on very old documents.

Thank you !