Cours 1

Introduction

A. Belaïd – LORIA - Nancy

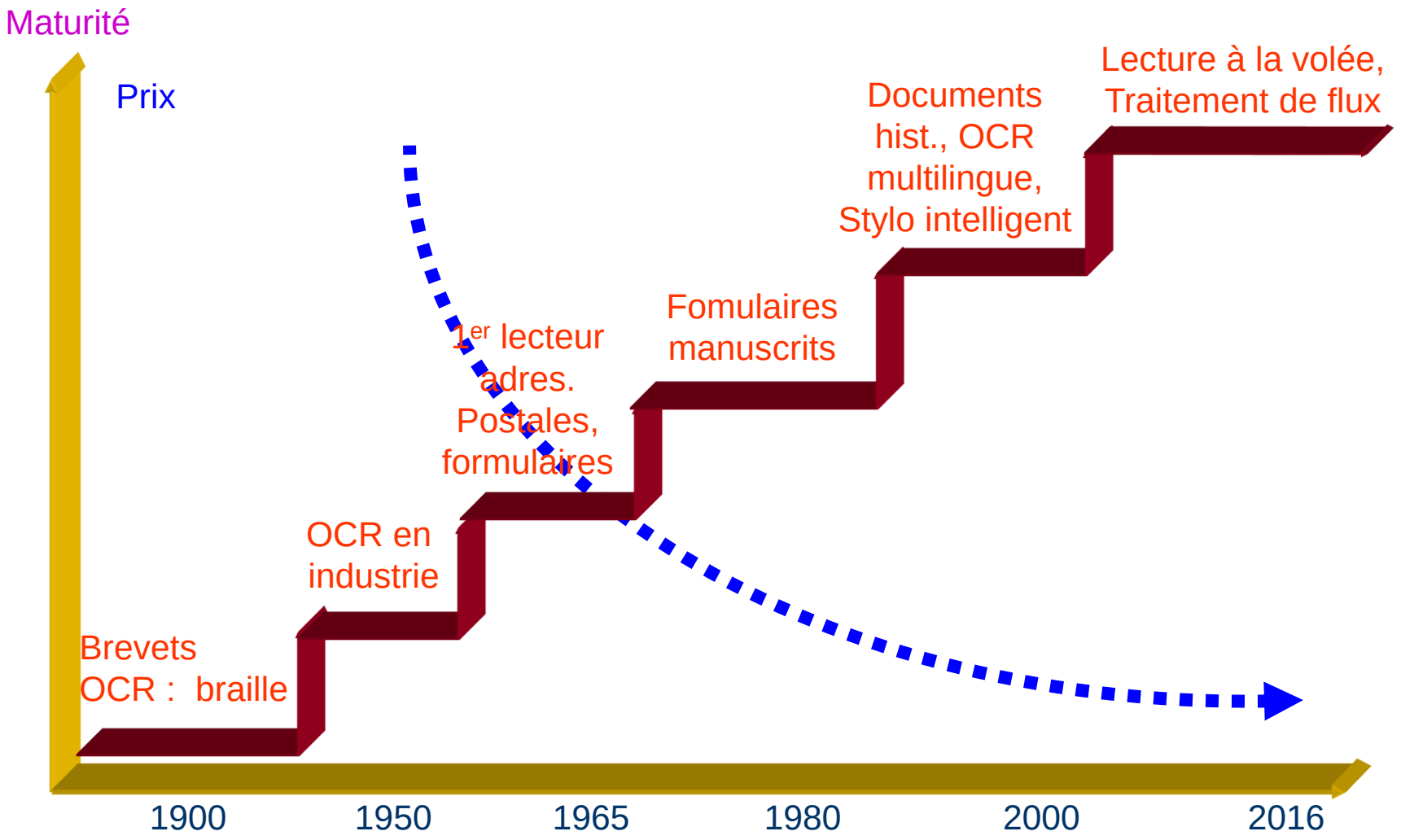
Objectif du cours

- Vous aider à comprendre
  - Pourquoi on a besoin de reconnaître l'écriture ?
  - Comment doit-on procéder pour la décoder ?
  - Quelles précautions doit-on prendre pour bien reconnaître ?
  - Pourquoi ce n'est pas toujours facile de le faire ?
  - Quel résultat doit-on attendre ?
  - Les sources à utiliser ?
- Discuter les quelques problèmes que l'on rencontre quand on traite ce genre de données

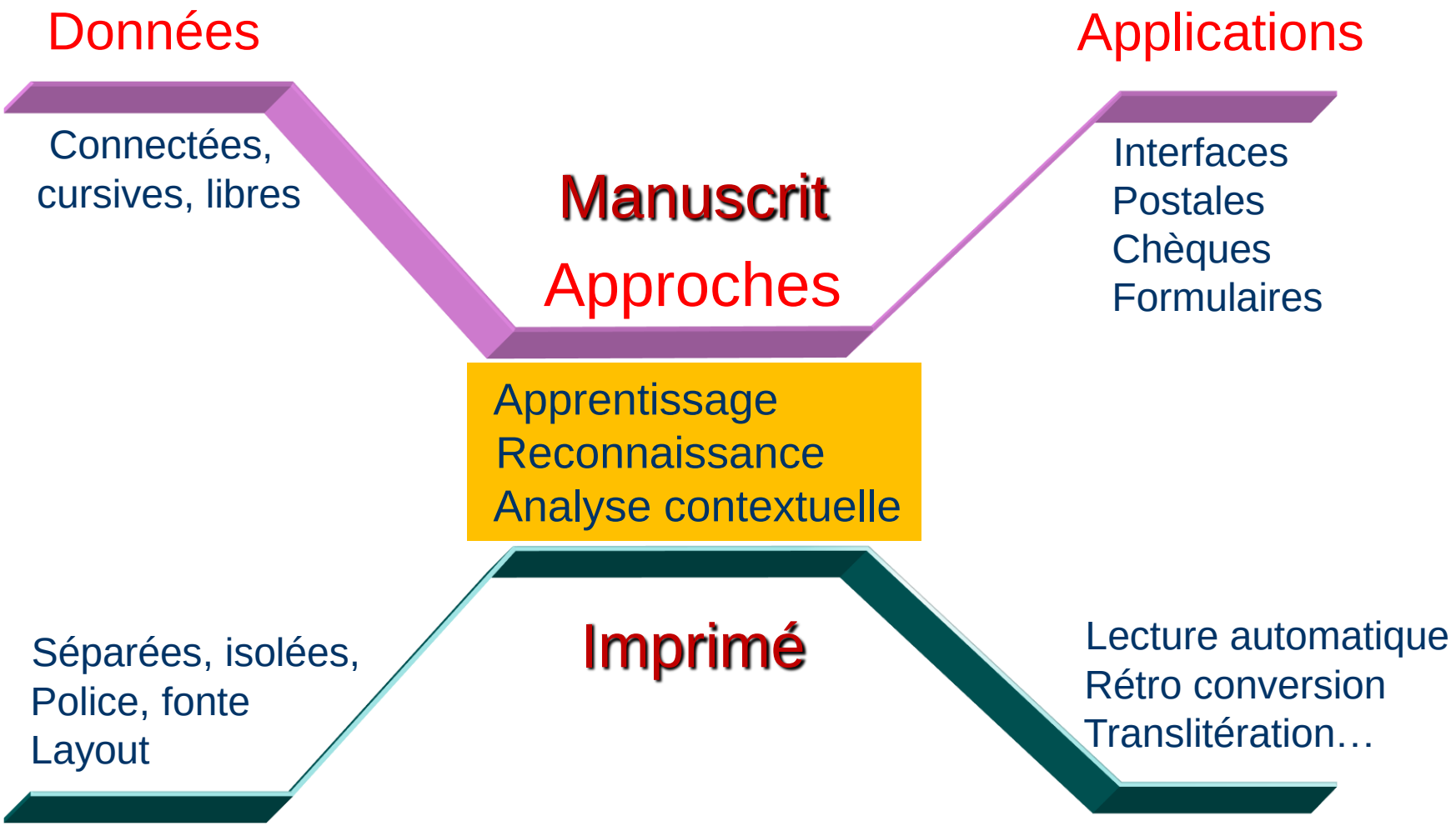
Reconnaissance de l'écriture

- Deux supports
  - Ecriture manuelle
    - Reconnaissance du script
      - Tri postal, reconnaissance automatique de montants de chèques
      - Identification/vérification de la signature (forensics)
      - Identification/vérification du scripteur
  - Document
    - Reconnaissance de la forme et du contenu
      - Indexation de flux de documents entrants : courriers, faxes, commandes…
      - Archivage, datation de documents historiques
      - Analyse de plans cadastraux, schémas mécaniques, formulaires, etc.
- Applications différentes, mais
  - méthodes et difficultés communes : connaissances métier

Historique : un siècle d'évolution



Manuscrit vs Imprimé

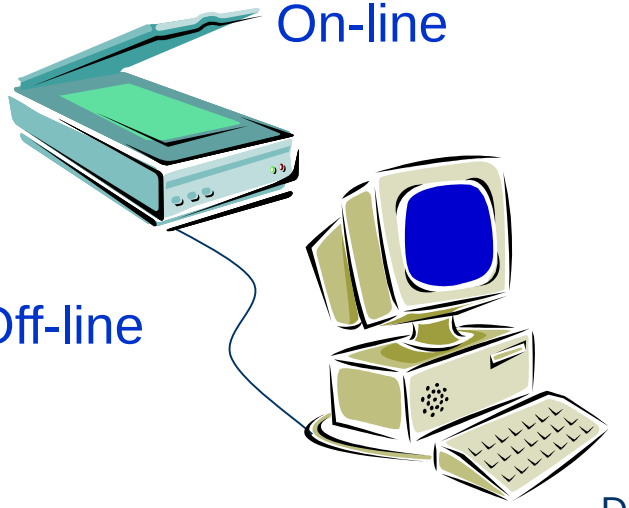


Manuscrit

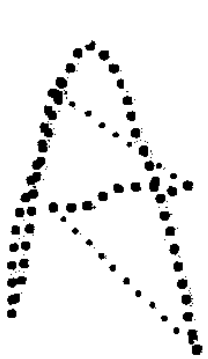
- Différents facteurs

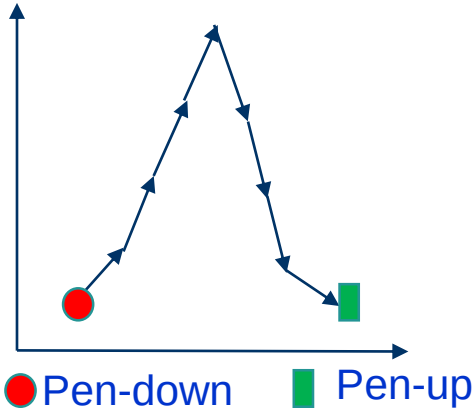
Mode d'acquisition

On-line

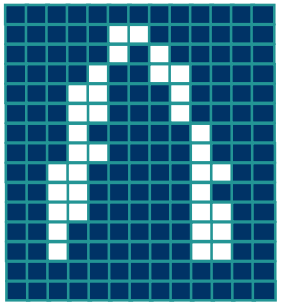


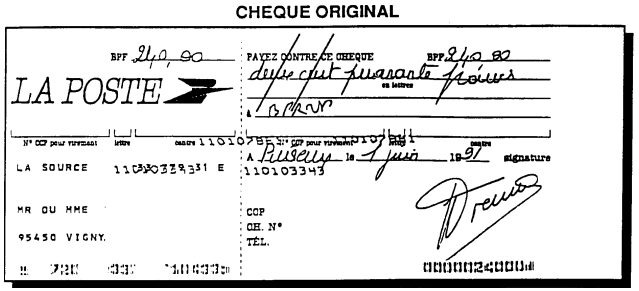
Off-line





Applications stylo
Interfaces, mobiles, PDA



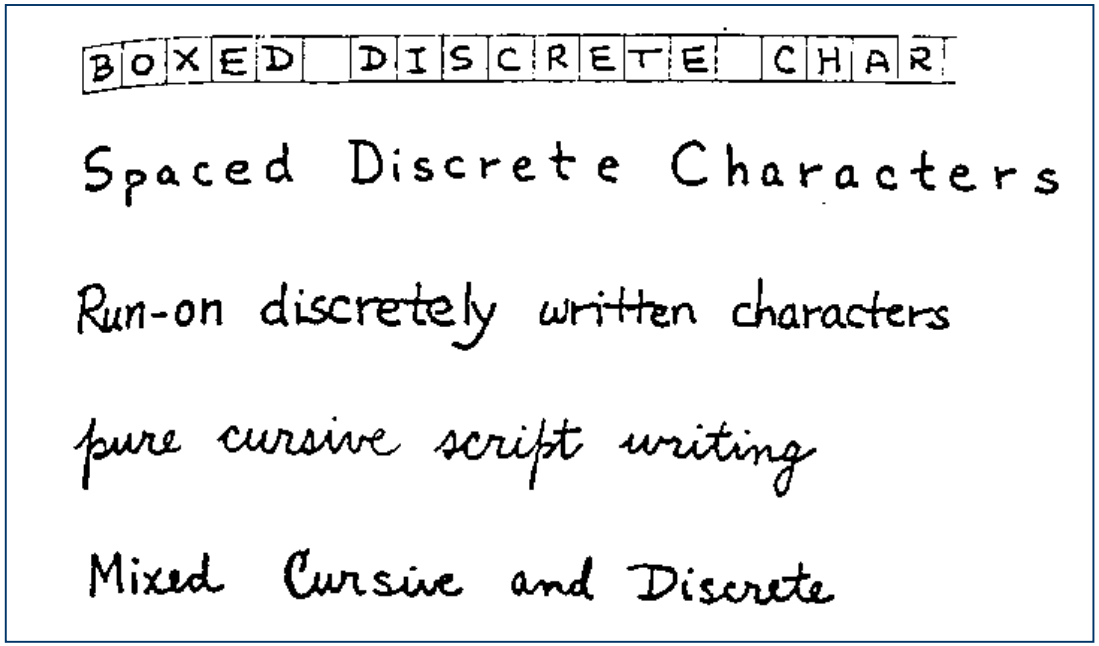


Courier, Chèques,
Formulaires

Manuscrit

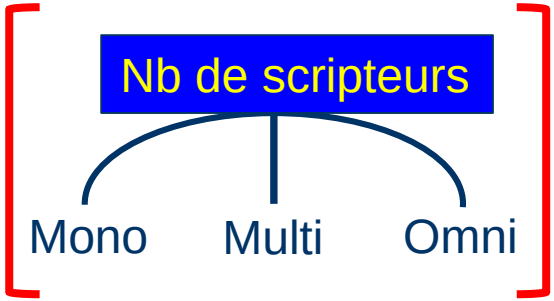
- D’autres facteurs

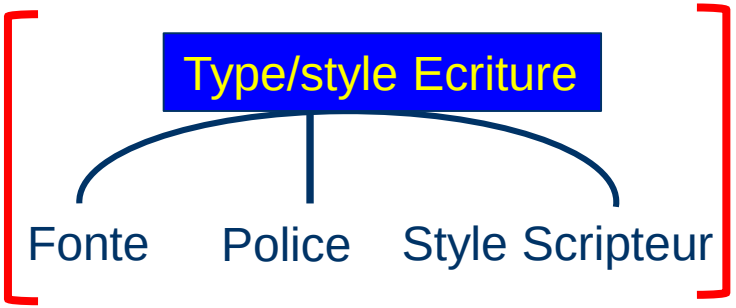
  Disposition spatiale des caractères

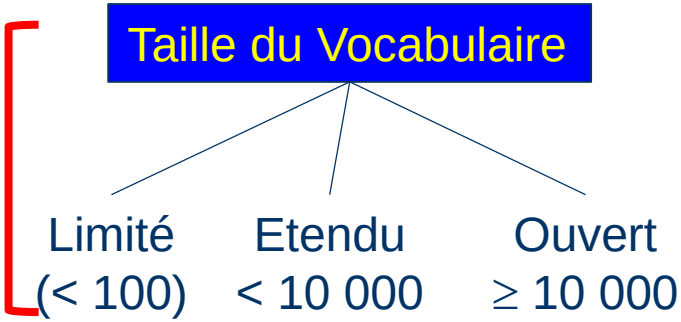


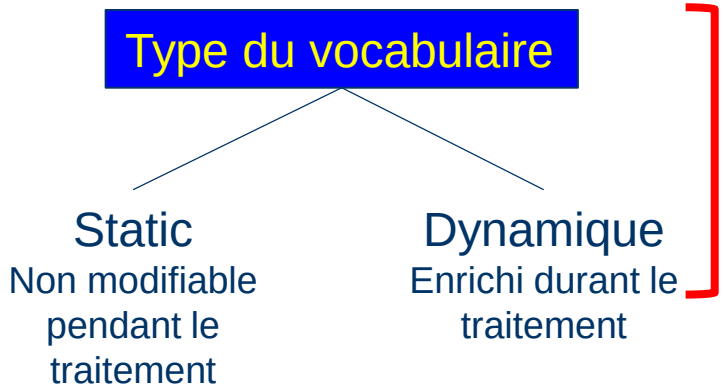
Manuscrit

- D’autres facteurs



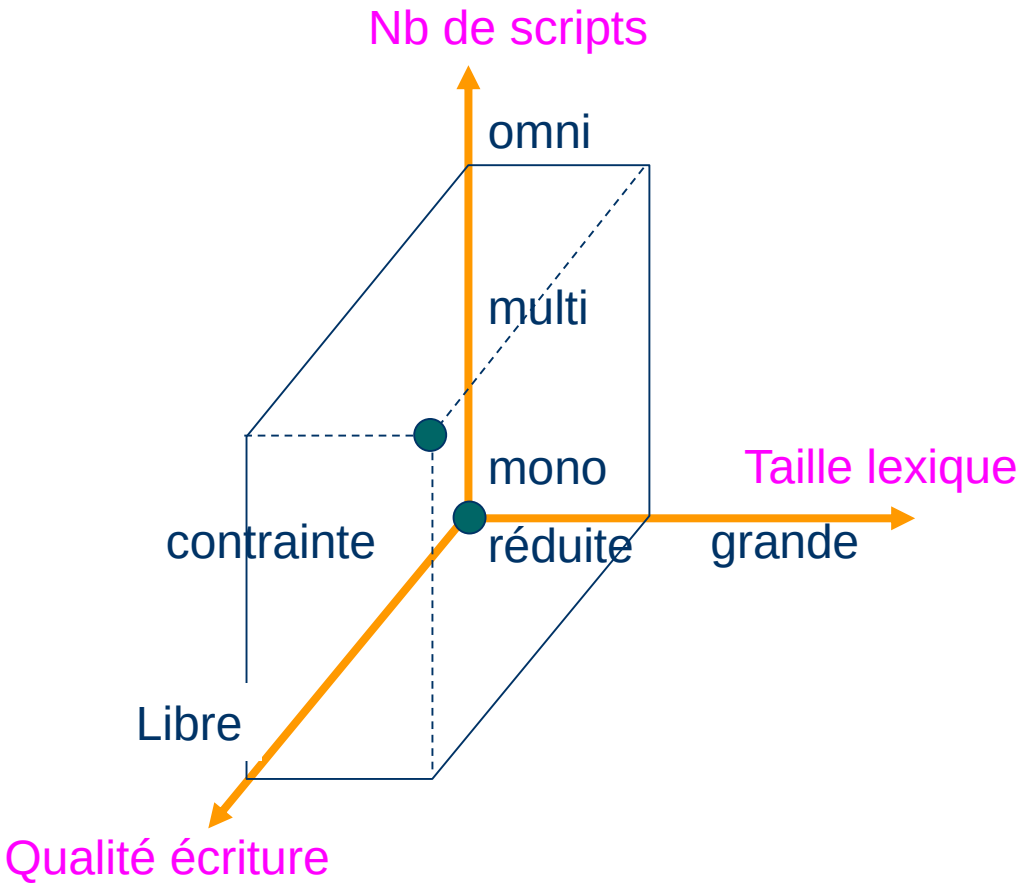






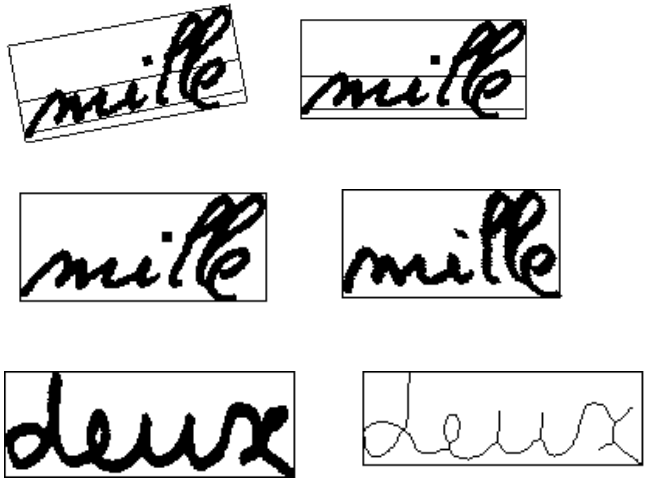
Introduction

Performances : critères influençant la qualité



Manuscrit : méthodes

Capture & Prétraitement



Extraction de caractéristiques

Haut niveau: boucles, hampes, jambages, etc.
Bas niveau: densité, contours, pixels

Représentation & classification

Structurelle (primitives)
Statistique / stochastique

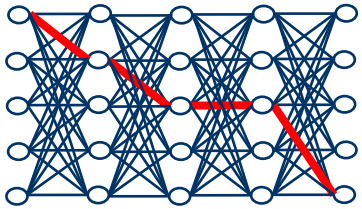
Post-traitement

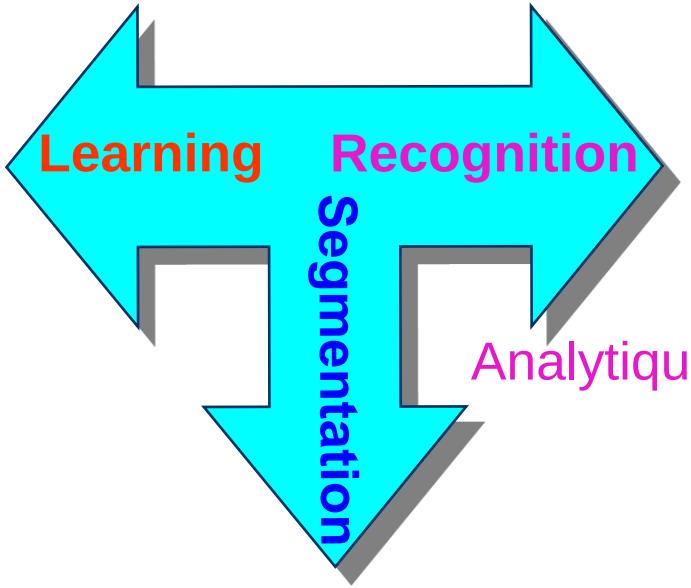
Contraintes lexicales, règles syntaxiques, langage

Manuscrit : méthodes

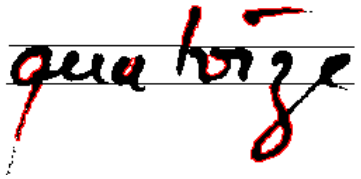
Modèle discriminant

Chemin discriminant

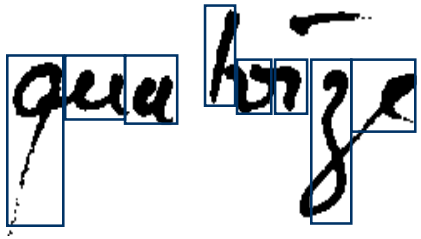




Globale

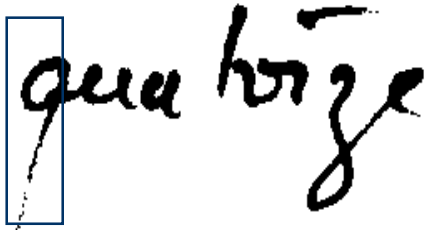


Analytique



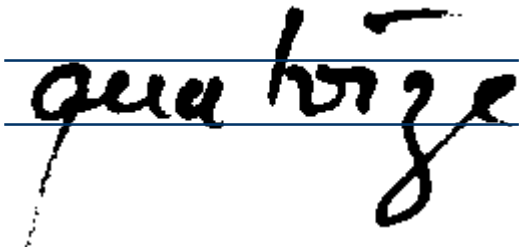
Pré-segmentation

Interne



Fenêtre glissante

Manuscrit : codage



Topologique

Zonale

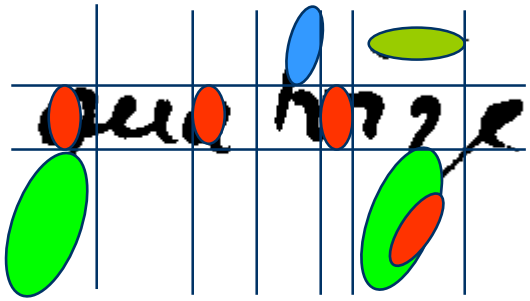
Géométrique

parametrisation continue

Modèle continu

VQ

Modèle discret



Hampe

Jambage

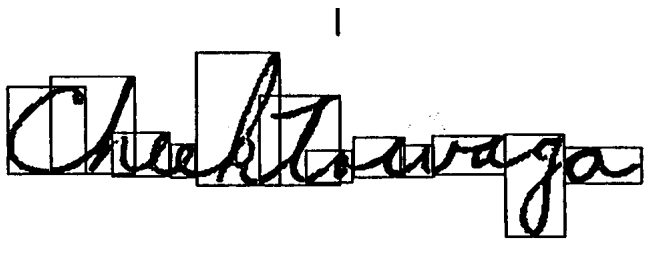
Barre T

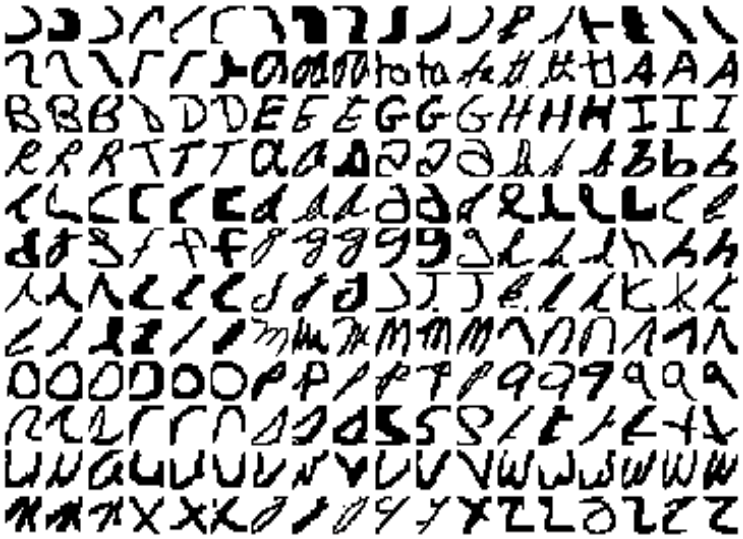
Boucle

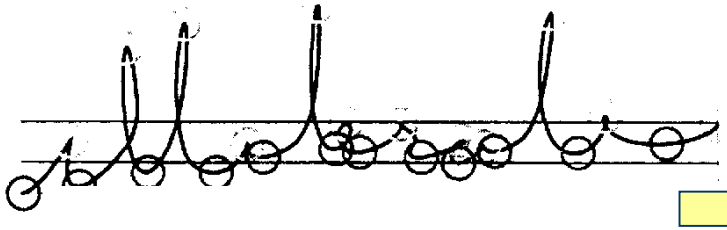
$$\begin{bmatrix}0\\1\\0\\1\end{bmatrix}\begin{bmatrix}0\\0\\0\\0\end{bmatrix}\begin{bmatrix}0\\0\\0\\1\end{bmatrix}\begin{bmatrix}1\\0\\0\\0\end{bmatrix}\begin{bmatrix}0\\0\\0\\1\end{bmatrix}\begin{bmatrix}0\\1\\1\\1\end{bmatrix}\begin{bmatrix}0\\0\\0\\0\end{bmatrix}$$

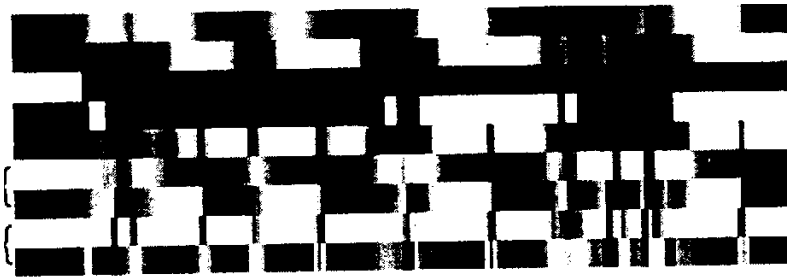
VQ

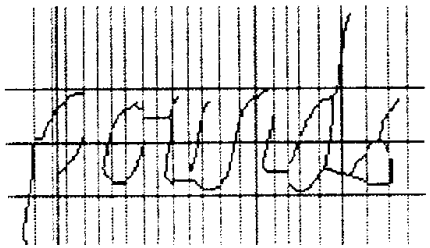
Manuscrit : stratégies de segmentation



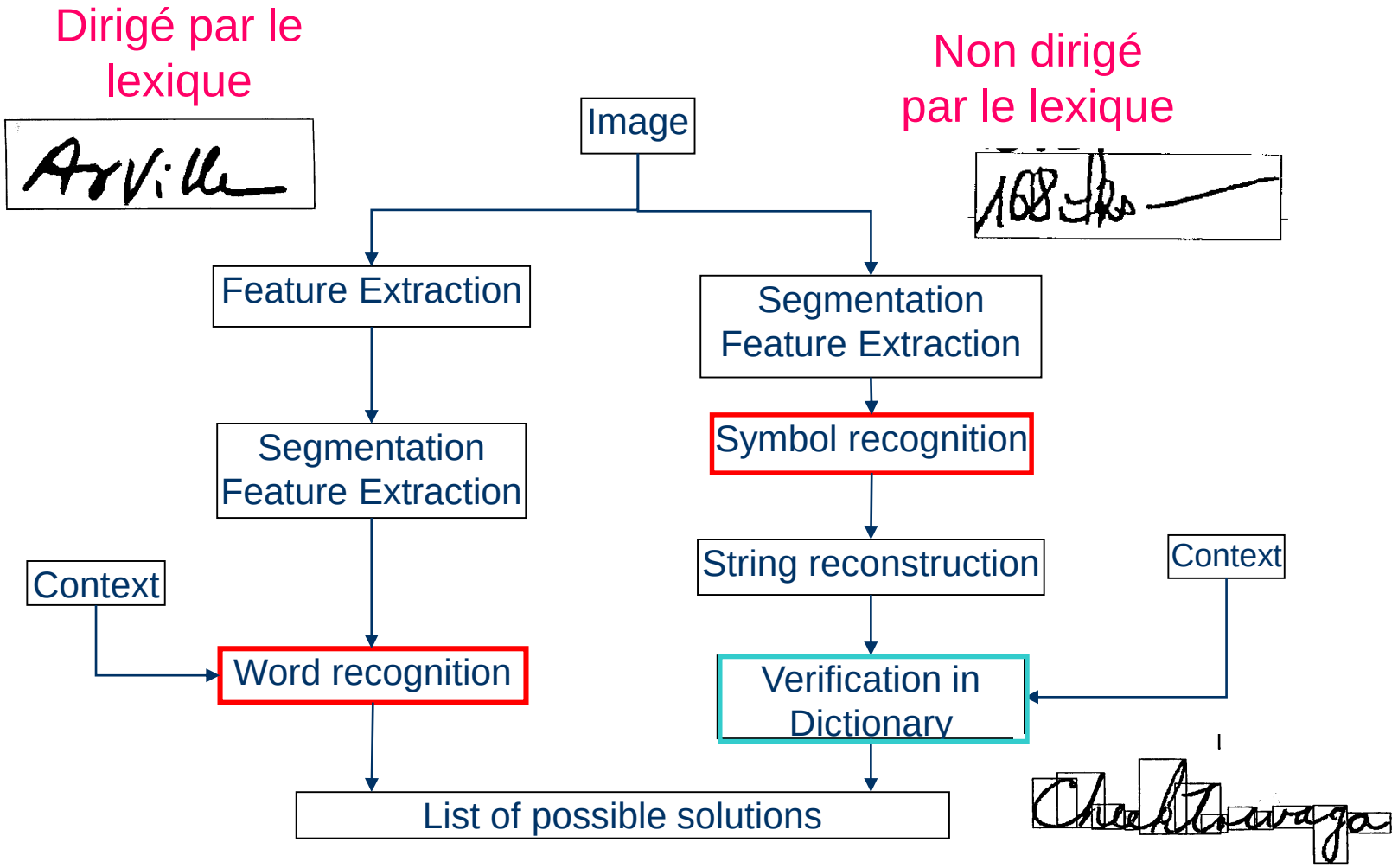








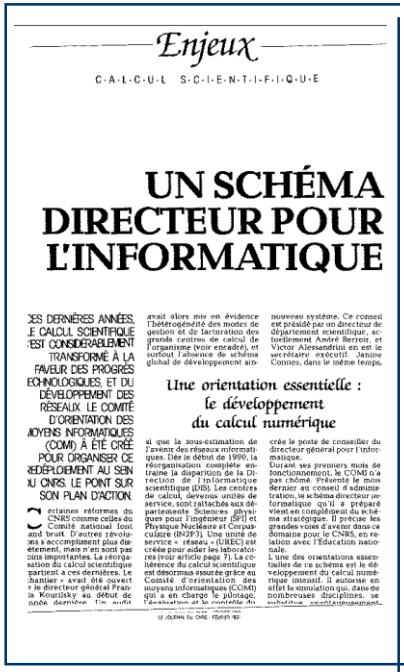
Manuscrit : stratégies de reconnaissance



Document

- Mode d'acquisition : toujours off-line : scanner/camera
- Layout : variable dépendant de la classe

Manhattan


Enjeux

C·A·L·C·U·L S·C·I·E·N·T·I·F·I·Q·U·E

UN SCHÉMA DIRECTEUR POUR L'INFORMATIQUE

DES DERNIÈRES ANNÉES, LE CALCUL SCIENTIFIQUE EST CONSIDÉRABLEMENT TRANSFORMÉ À LA FAVEUR DES PROGRÈS TECHNOLOGIQUES, ET DU DÉVELOPPEMENT DES RÉSEAUX. LE COMITÉ D'ORIENTATION DES MOYENS INFORMATIQUES (COMI) À ÉTÉ CRÉÉ POUR ORGANISER CE REDÉPLOIEMENT AU SEIN DU CNRS. LE POINT SUR SON PLAN D'ACTION.

Certaines réformes du CNRS comme celles du Comité national font aussi bruit. D'autres révolutions s'accomplissent plus discrètement, mais n'en sont pas moins importantes. La réorganisation du calcul scientifique partient à ces dernières. Le chantier » avait été ouvert par le directeur général Francis Kourilsky au début de cette dernière. Un audit avait alors mis en évidence l'hétérogénéité des modes de gestion et de facturation des grands centres de calcul de l'organisme (voir encadré), et surtout l'absence de schéma global de développement ambitieux. Un nouveau système. Ce conseil est présidé par un directeur de département scientifique, actuellement André Berroir, et Victor Alessandrini en est le secrétaire exécutif. Jaume Comes, dans le même temps,

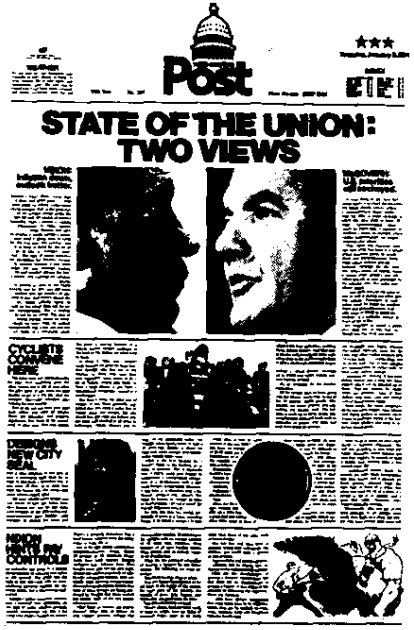
Une orientation essentielle : le développement du calcul numérique

si que la sous-estimation de l'avenir des réseaux informatiques. Dès le début de 1999, la réorganisation complète entraîne la disparition de la Direction de l'informatique scientifique (DIS). Les centres de calcul, devenus unités de service, sont rattachés aux départements Sciences physiques pour l'ingénieur (SPI) et Physique Nucléaire et Corpusculaire (IN2P3). Une unité de service « réseau » (URECI) est créée pour aider les laboratoires (voir article page 7). La cohérence du calcul scientifique est désormais assurée grâce au Comité d'orientation des moyens informatiques (COMI) qui a en charge le pilotage, l'évaluation et le contrôle du crée le poste de conseiller du directeur général pour l'informatique.

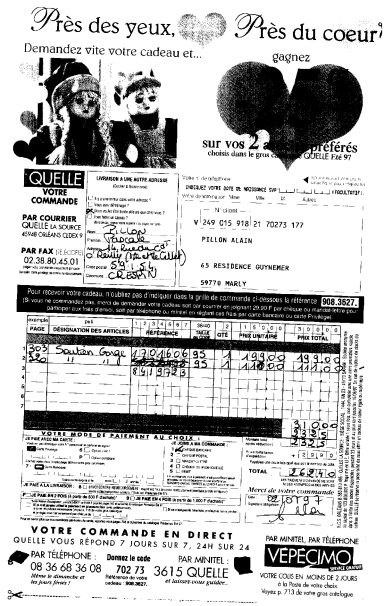
Durant ses premiers mois de fonctionnement, le COMI n'a pas chômé. Présenté le mois dernier au conseil d'administration, le schéma directeur informatique qu'il a préparé vient en complément du schéma stratégique. Il précise les grandes voies d'avenir dans ce domaine pour le CNRS, en relation avec l'Éducation nationale.

L'une des orientations essentielles de ce schéma est le développement du calcul numérique intensif. Il autorise en effet la simulation qui, dans de nombreuses disciplines, se substitue, avantageusement,

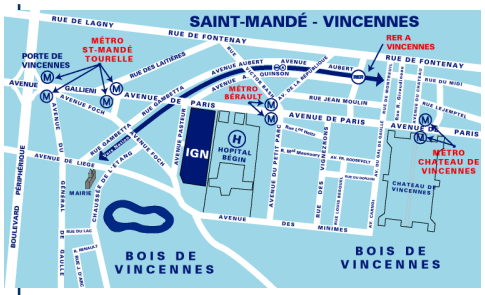
Mosaic

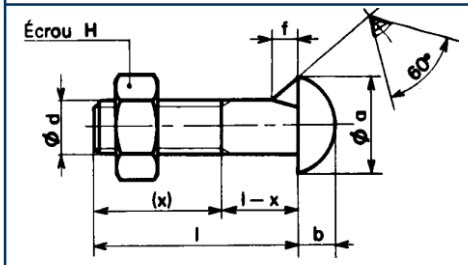


Zonal



Non linear





Font style Number

Mono, Multi or Omni-font

Document

- Objectifs de reconnaissance
  - Nous avons besoin d'accéder à un ensemble de "documents"
    - Archivage?
    - Indexation?
  - Souvent, on a des collections massives et hétérogènes
  - Différentes langues
  - Différentes présentations
  - Différentes sources
  - Nous avons de la chance si les métadonnées sont consistantes et uniformes

Document

- Pourquoi acquérir des documents image ?
  - Solution sans papier
    - Transfert efficace
    - Organisation
    - Commodité
  - Accès à une variété de contenus
    - Lecteur universel - courrier électronique, pièces jointes, feuilles de calcul
    - Pas besoin d'applications originales
  - Détecter la falsification : changement du contenu ?
    - Plus facile à certifier?
    - Authentifier le document

Contexte – Objectifs de recherche

Le passé récent

Applications ciblées

Données limitées

Classes connues

Classes figées

Volumétrie faible/moyenne

Aujourd'hui

Problèmes ouverts : courrier libre

Flux continu

Classes évolutives

Pas de connaissance a priori

Volumétrie importante

Modèles statiques

Beaucoup de paramètres

Bases de données locales

Evaluation subjective

Modèles incrémentaux

Paramétrage adaptatif

Bases de données publiques

Métriques pour l'évaluation

Document : comment acquérir ?

- Balayage?
  - Haute vitesse, automatisé, multiformes - livres, etc.
- Photocopieurs numériques
  - Mémoire d'entreprise
- Sortie d'application
  - Imprimer comme l'image
  - Conversion de masse
- Appareils photo?
  - Téléphones portables?
  - QipIt, ScanR, Hotcard

Tous ont des implications sur l'utilisation

Document : qu’est ce qu’une image ?

- Représentation pixel de la carte d'intensité
- Pas de "contenu" explicite, que des chiffres les uns à côté des autres
- L'analyse d'image
  - Tente d'imiter le comportement visuel humain
  - Tire des conclusions, formule des hypothèses et vérifie

10	27	33	29
27	34	33	54
54	47	89	60
25	35	43	9

Analyse d’image

Utiliser les techniques adéquates pour représenter le contenu

Transformer les requêtes sémantiques en "caractéristiques d'image"

- Couleur, forme, texture ...
- Relations spatiales

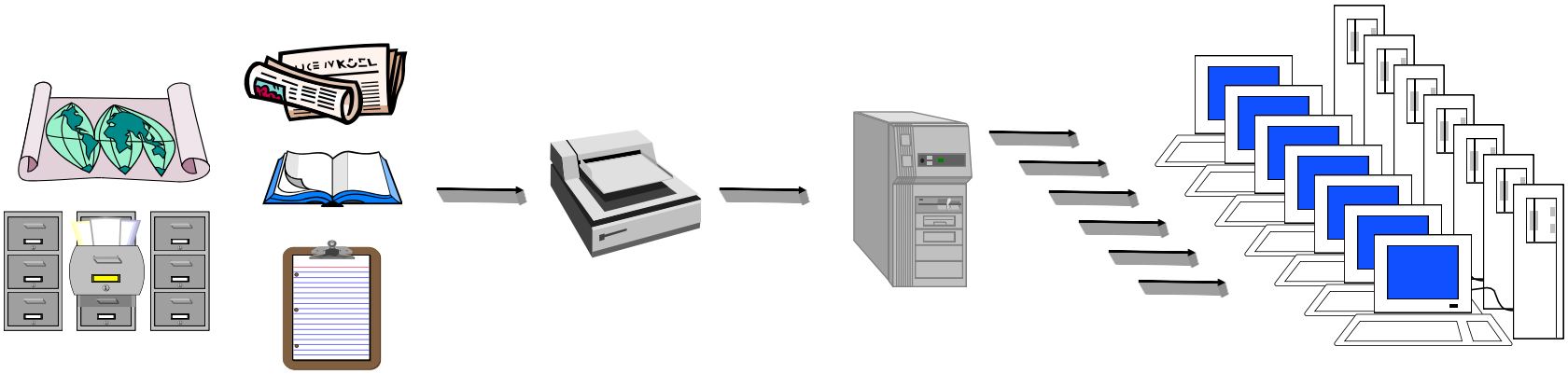
Image de document

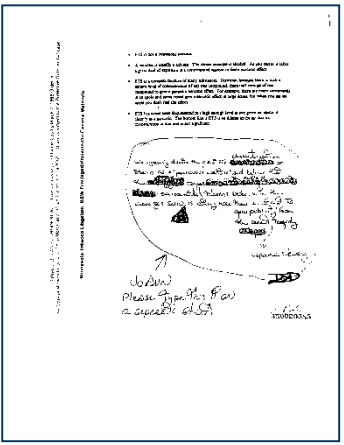
- Une collection de points appelés "pixels"
  - Disposés dans une grille appelée "bitmap"
- Pixels souvent binaires (noir, blanc)
  - Mais l'échelle de gris ou la couleur est parfois nécessaire
- 300 points par pouce (ppp) donne les meilleurs résultats
  - Mais les images sont assez grandes (1 Mo par page)
  - Les télécopies sont normalement de 100 à 200 ppp
- Habituellement stocké en format TIFF ou PDF

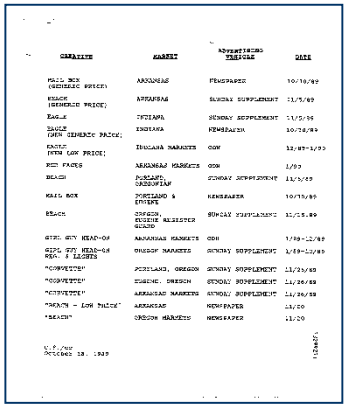
Pourtant, nous voulons pouvoir les traiter comme des fichiers texte!

Base, dataset…

- Collecte d'images numérisées
- Besoin d'être disponible pour l'indexation et la récupération, l'abstraction, le routage, l'édition, la diffusion, l'interprétation…

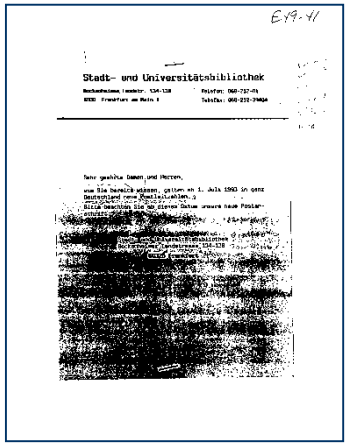


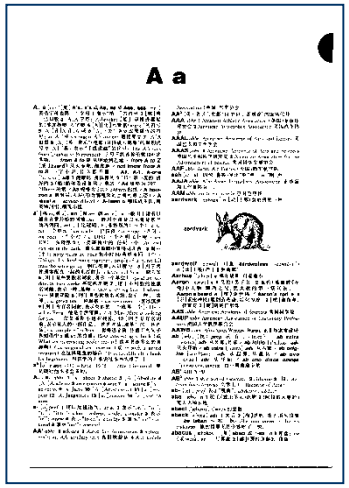


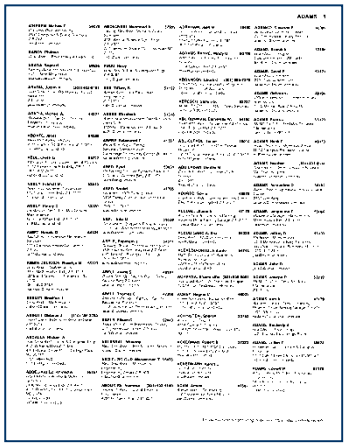


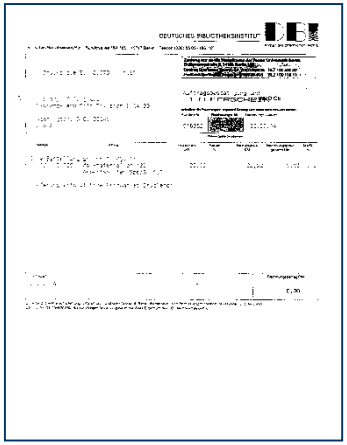
CREATIVE	AGENCY	ADVERTISING VEHICLE	DATE
MAIL BOX (GENERIC BRICK)	ARKOMEDIA	NEWSPAPER	10/10/89
BEACH (ENHANCE PRICE)	ARKANSAS	SUNDAY SUPPLEMENT	11/5/89
EAGLE	INDIANA	SUNDAY SUPPLEMENT	11/5/89
EAGLE (NEW LOW PRICE)	INDIANA MARKETS	COM	12/29-1/90
RED FACES	ARKANSAS MARKETS	COM	1/90
BEACH	PORTLAND, OREGONIAN	SUNDAY SUPPLEMENT	11/5/89
MAIL BOX	PORTLAND & ENVIRS	NEWSPAPER	10/15/89
BEACH	OREGON, EUGENE REGISTER GUARD	SUNDAY SUPPLEMENT	11/15.89
GIRL GUY HEAD-ON	ARKANSAS MARKETS	COM	1/89-12/89
GIRL GY HEAD-ON MEN & LIGHTS	OREGON MARKETS	SUNDAY SUPPLEMENT	1/89-12/89
"CORVETTE"	PORTLAND, OREGON	SUNDAY SUPPLEMENT	11/25/89
"CORVETTE"	EUGENE, OREGON	SUNDAY SUPPLEMENT	11/26/89
"CORVETTE"	ARKANSAS MARKETS	SUNDAY SUPPLEMENT	11/26/89
"BEACH - LOW PRICE"	ARKANSAS	NEWSPAPER	11/20
"BEACH"	OREGON MARKETS	NEWSPAPER	11/20

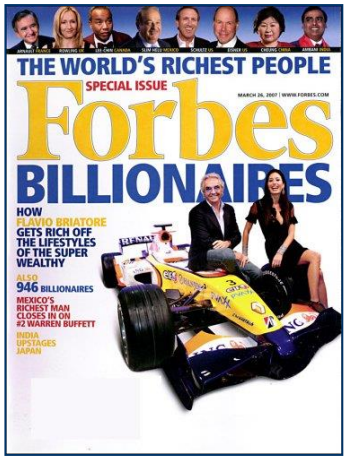
October 18, 1989

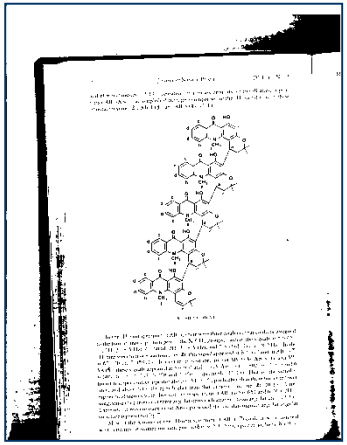




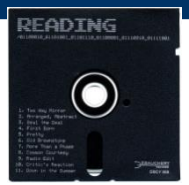


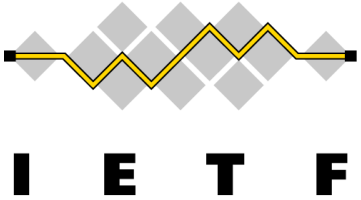


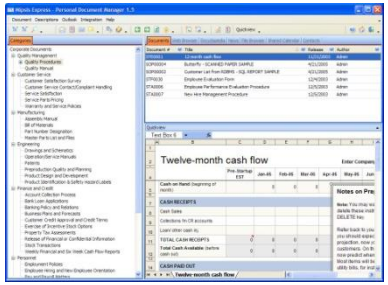


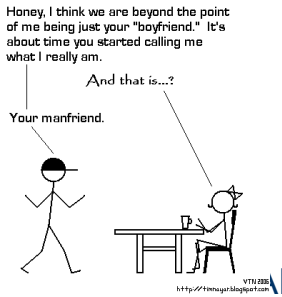


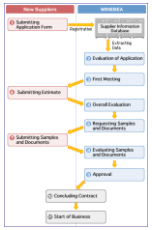
D’autres “documents”

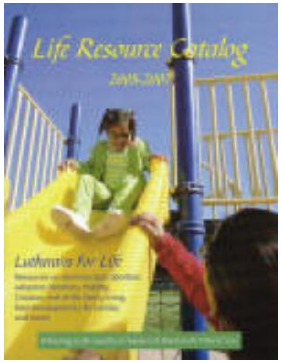


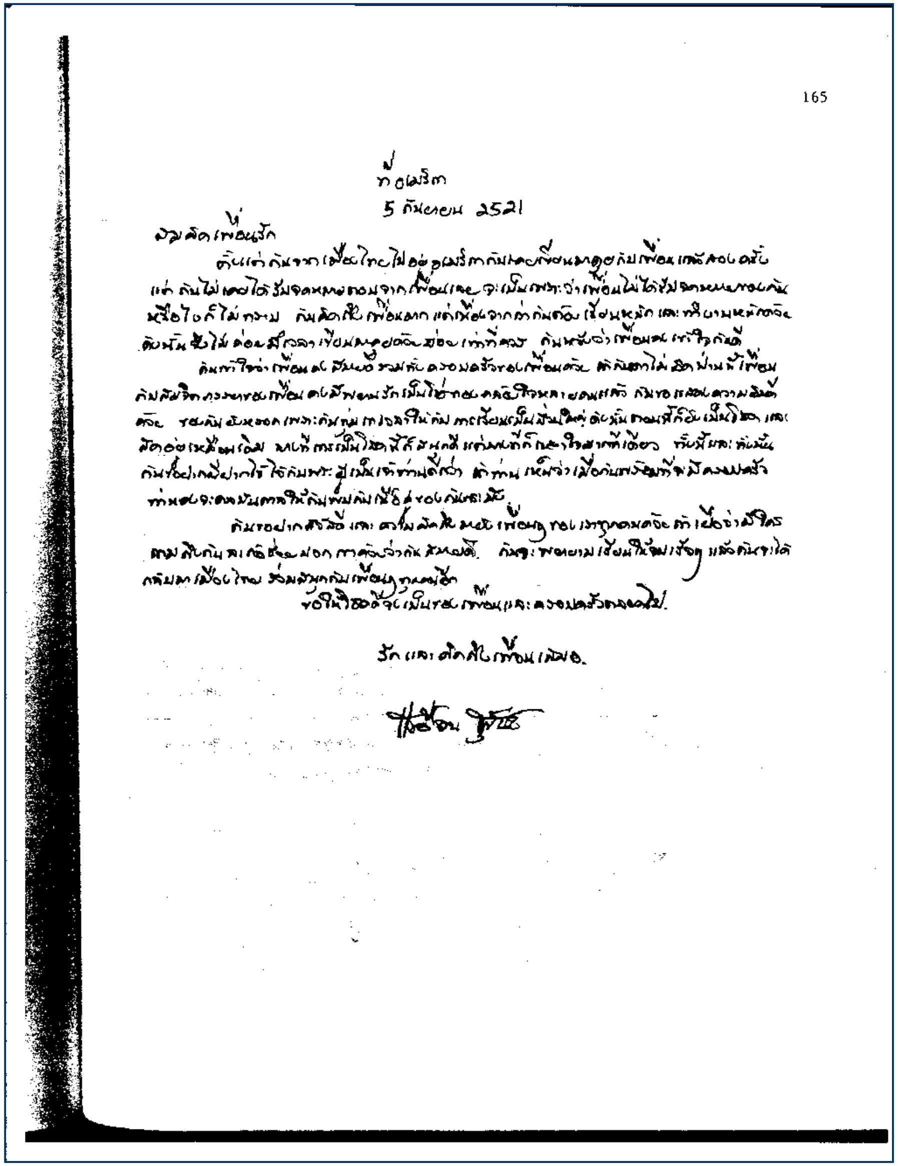


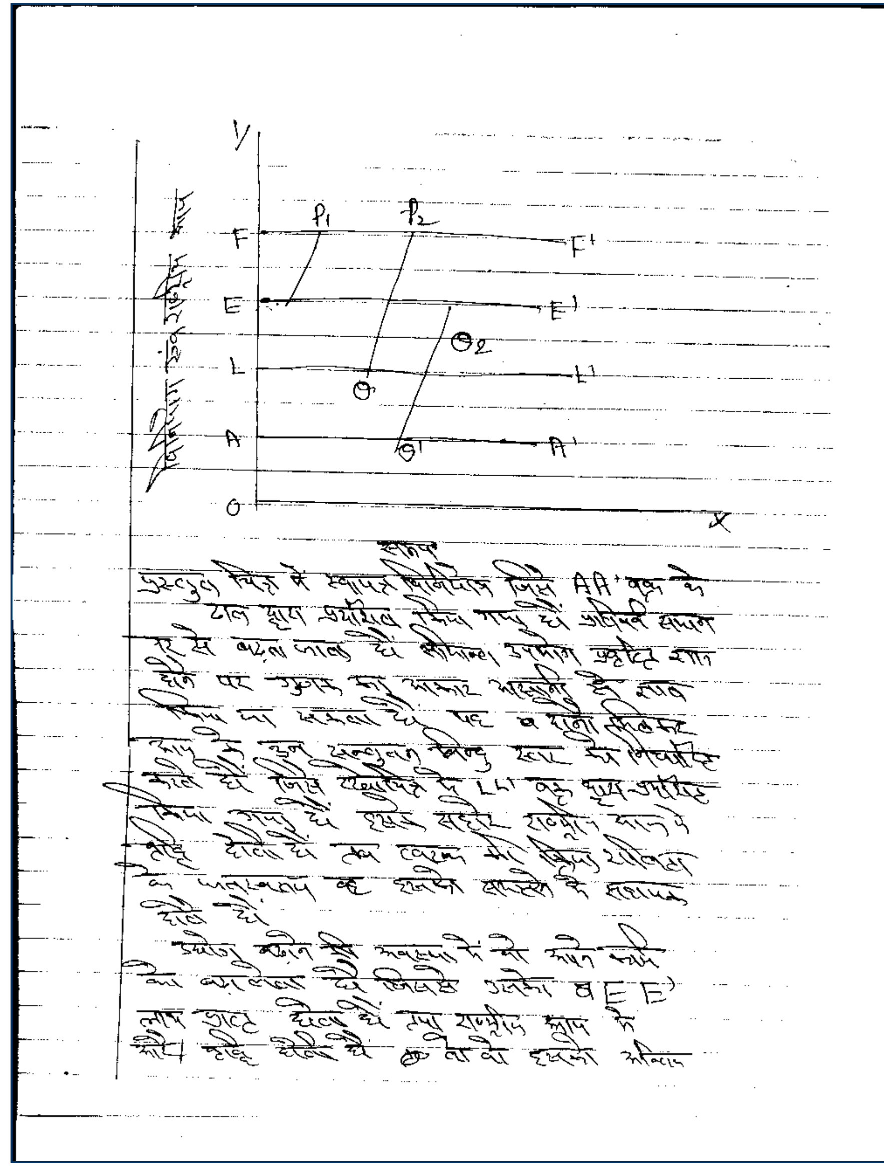






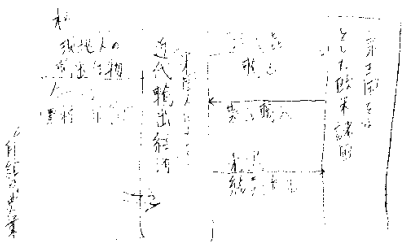






Ⅱ. 遼上國の社会経済的特質

1. 基本的特徴 ～ 牧民型の重経済



6. The Three Kingdoms and the Six Dynasty

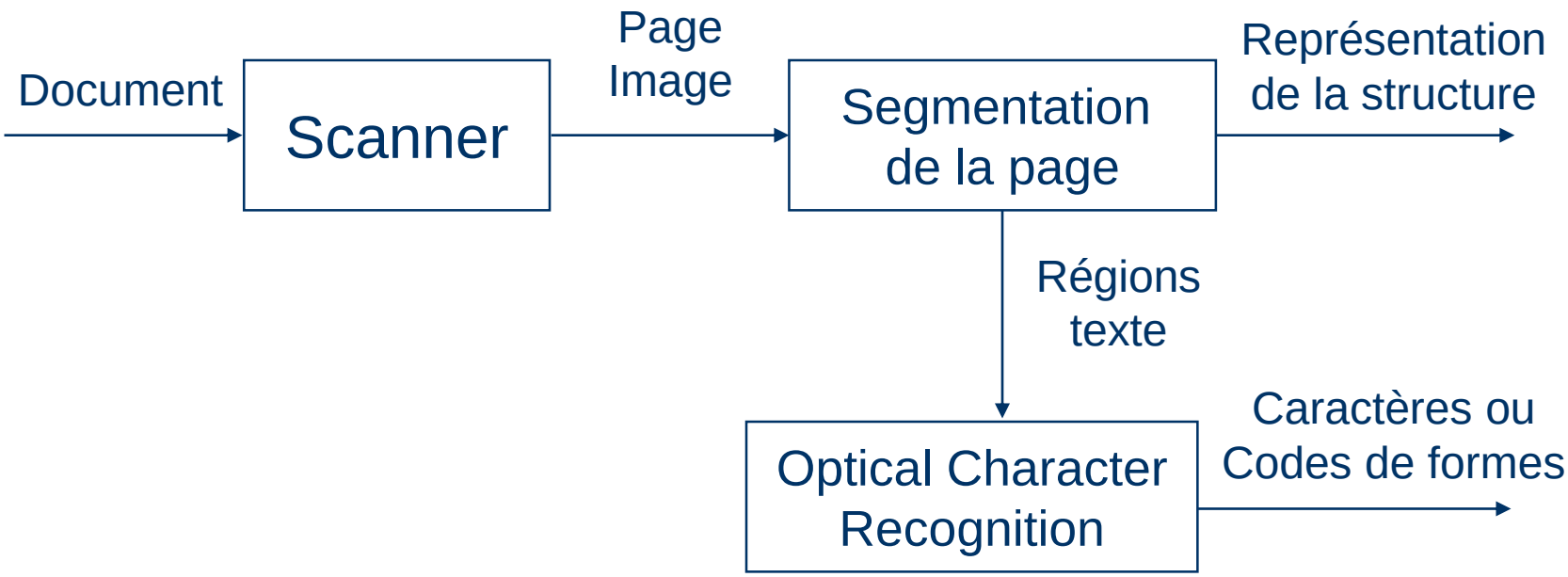
p85 ferment n. 불안, 동요, 정치적 동란
usurp v. 강탈하다
incessant a. 끊임없는
shrink v. 움츠러들다, 움켜들다, 기가 죽다 (p.p. shrunken)
rashly ad. 무모하게, 성급하게, 성급하게
massacre v. 학살하다
flee v. 달아나다, 도피하다 (p. fled)
tribe n. 부족, 종족
abandon v. 버리다, 포기하다
nomadic a. 유목인의, 유목 생활을 하는
* sinicize v. 중국화하다, 중국식으로 만들다
cavalry n. 기병대, 기마부대
refugee n. 망명자, 피난자, 난민
perpetual a. 영구한, 부단한
turmoil n. 혼란, 동요, 불안
p89 undermind v. 약화시키다, 서서히 허물어지다
* monastery n. 수도원, 사원 '사찰(절)'
vast a. 광대한, 거대한, 엄청난
* proportion n. 크기, 비례, 비율
realm /relm/ n. 왕국, 영역
bureaucracy n. 관료제, 관료정치, 관료주의
* exert v. (힘, 권력 등을) 행사하다, (영향을) 미치다

〈 Taoism 〉 도교

Taoism n. 도교
*calligraphy n. 서예, 서법, 글씨, 필적
conglomeration n. 응집, 집합

LONG BANG MOO FANCY

Indexation d'images de pages
(Schéma de conversion traditionnel)

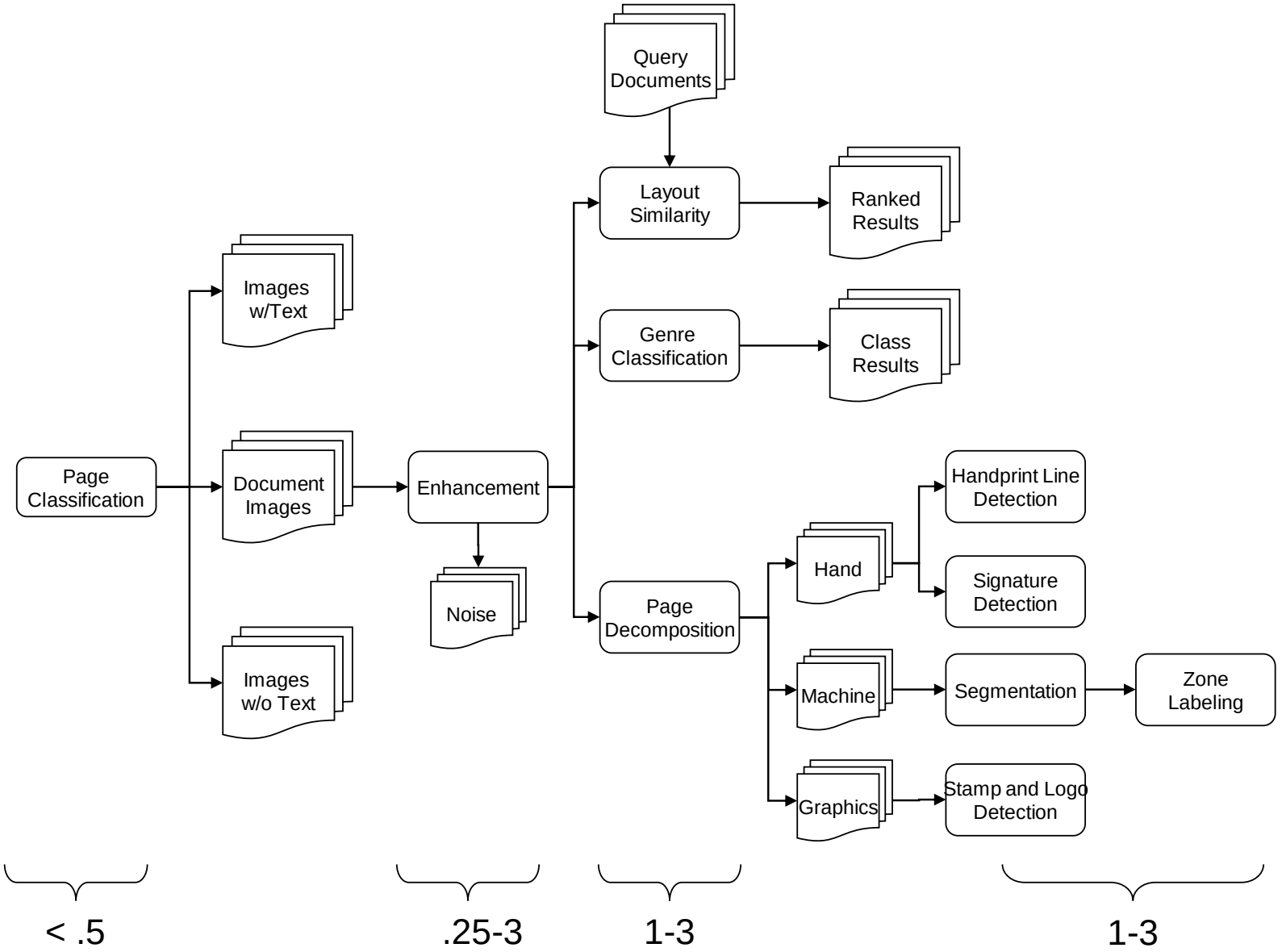


Analyse de l’image du document

- Schéma général
  - Acquisition de l’image – digitalisation
  - Prétraitement
  - Extraction de paramètres
  - Classification
- Tâches spécifiques
  - Analyse de la structure physique et logique de la page
  - Classification de zones
  - Identification de la langue
  - Traitement spécifique d’une zone
    - Reconnaissance
    - Vectorisation

Analyse de l’image du document

- Ce que vous devez faire avant de pouvoir traiter les images sous forme de "documents électroniques" ....
  - Analyse de l'image du document
  - Décomposition de la page
  - Reconnaissance optique de caractères
  - Indexation traditionnelle avec conversion
  - Matrice de confusion
  - Codes de forme
  - Faire des choses sans conversion
  - Dépistage, classification, résumé,
  - Repérage des mots clés, etc.



Temps de traitement necessaire en secondes

Pourquoi l'analyse de document est difficile ?

- Plusieurs raisons
  - Tableau 2D de "valeurs"
  - Représente un langage symbolique
  - Beaucoup de variations dans les symboles

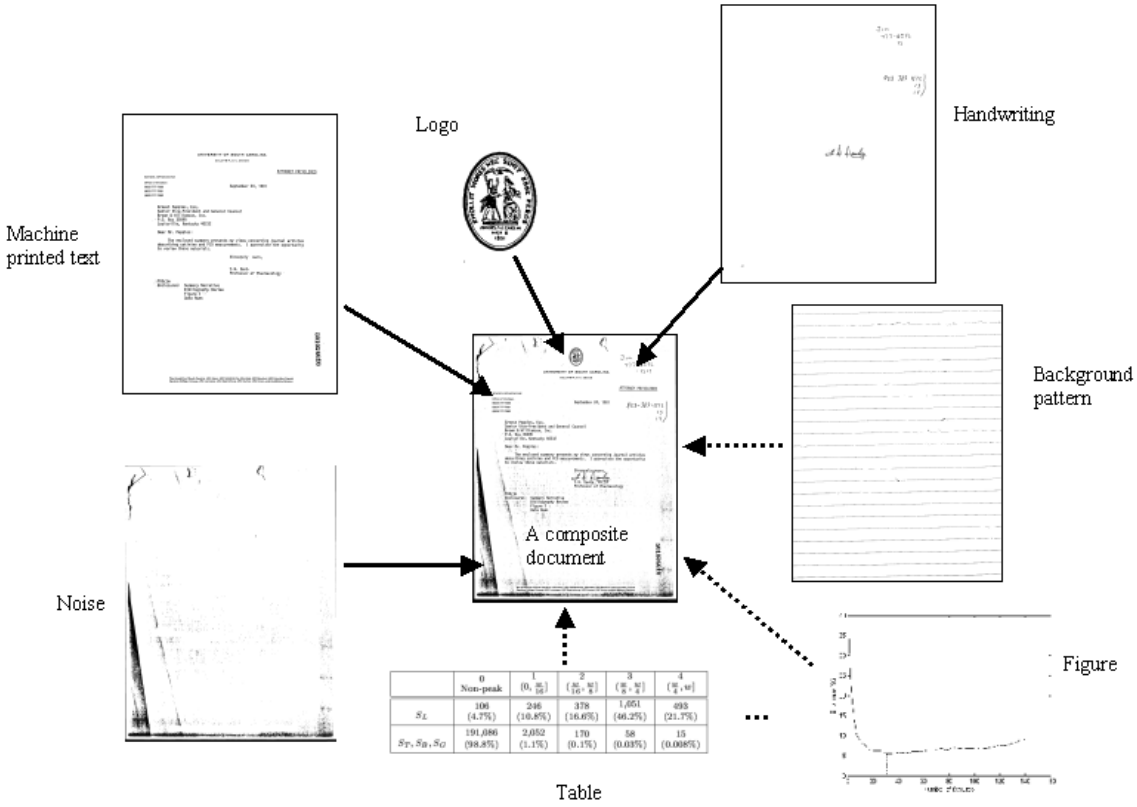
  - 3-4 fois plus grand que des images normales
  - et ça c'est uniquement le cas de documents texte imprimé latin !

Analyse de page

- En supposant regarder que le texte
  - Correction d'obliquité
    - Basé sur la recherche de l'orientation principale des lignes
  - Détection d'une zone d'image et de texte
    - Basé sur la texture et l'orientation dominante
  - Classification structurelle
    - Infirmer la structure logique de l'agencement physique
  - Classification de la région textuelle
    - Titre, auteur, en-tête, bloc de signature, etc.

Segmentation physique de la page

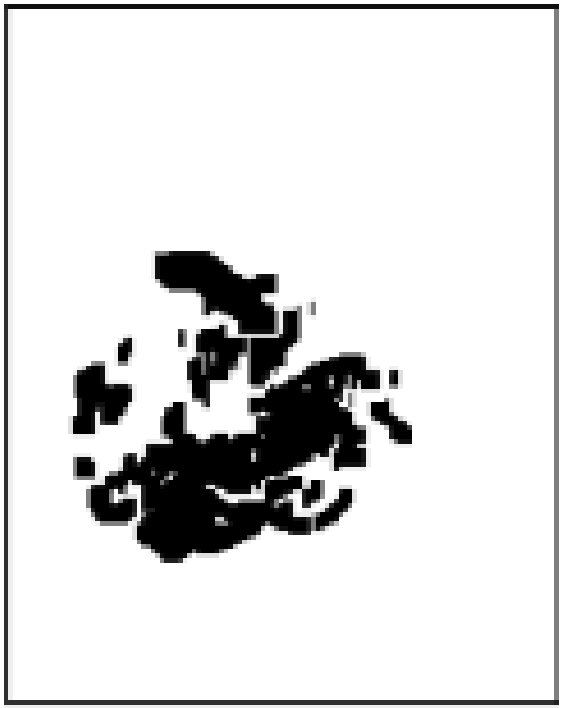
- Couches d'information
  - Un document se compose de plusieurs couches, telles que l'écriture manuscrite, le texte imprimé par machine, les motifs d'arrière-plan, les tableaux, les chiffres, le bruit, etc.

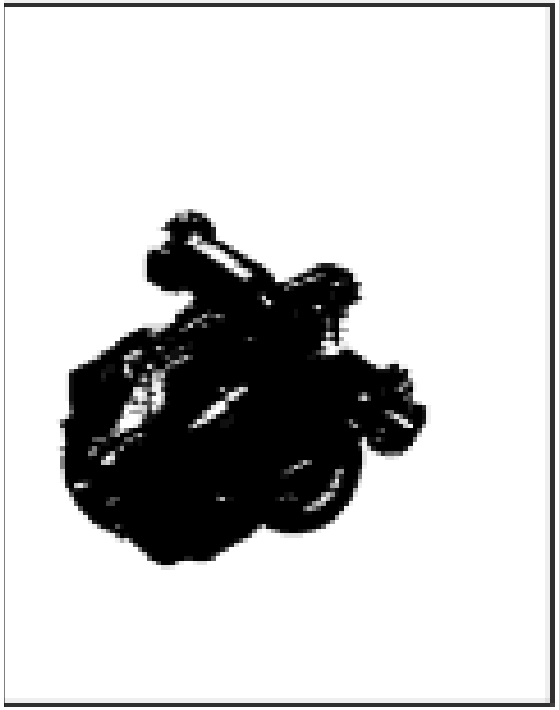


Segmentation de la page

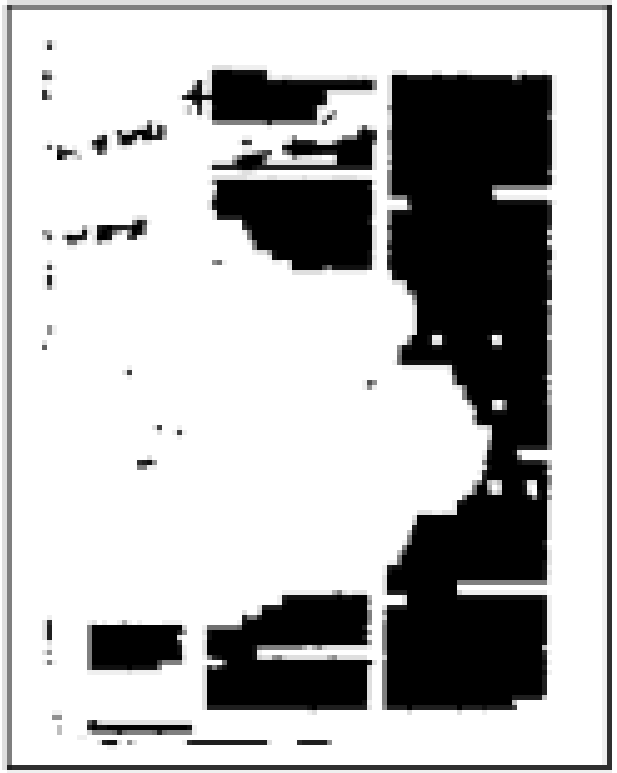
- Généralement basée sur la proximité spatiale
  - Espaces blancs
  - Marges
  - Différences de type de contenu
- Peut être très sensible au bruit
- Distinguer entre
  - Top Down
    - On découvre en décomposant
  - Bottom up
    - On rassemble ce qu'on connaît

Détection d’objets



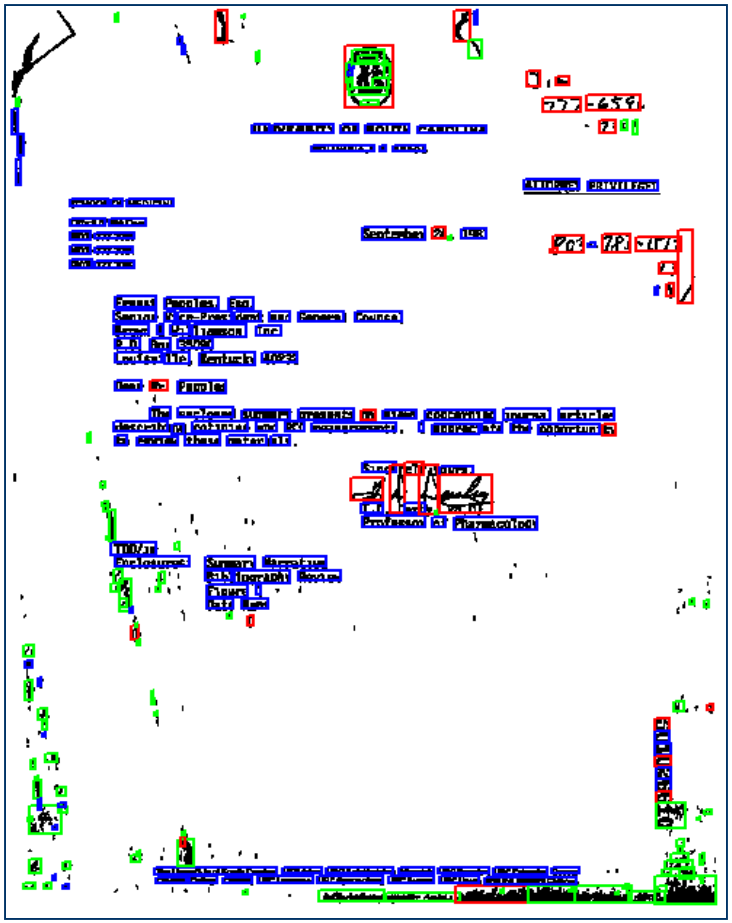


Detection de régions de texte

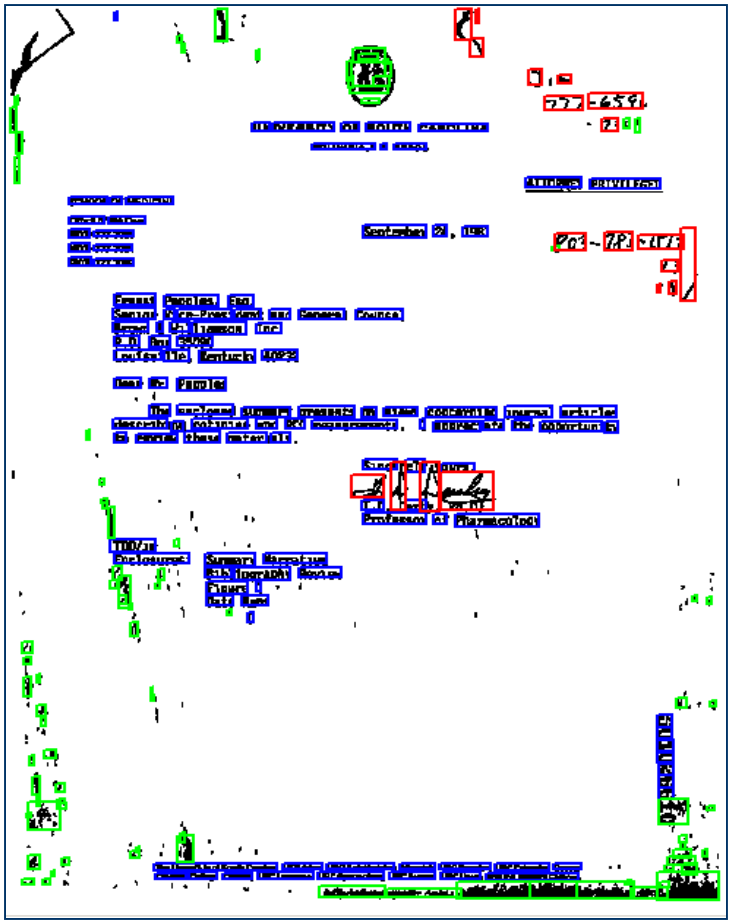


Printed text
Handwriting
Noise

Un exemple plus complexe



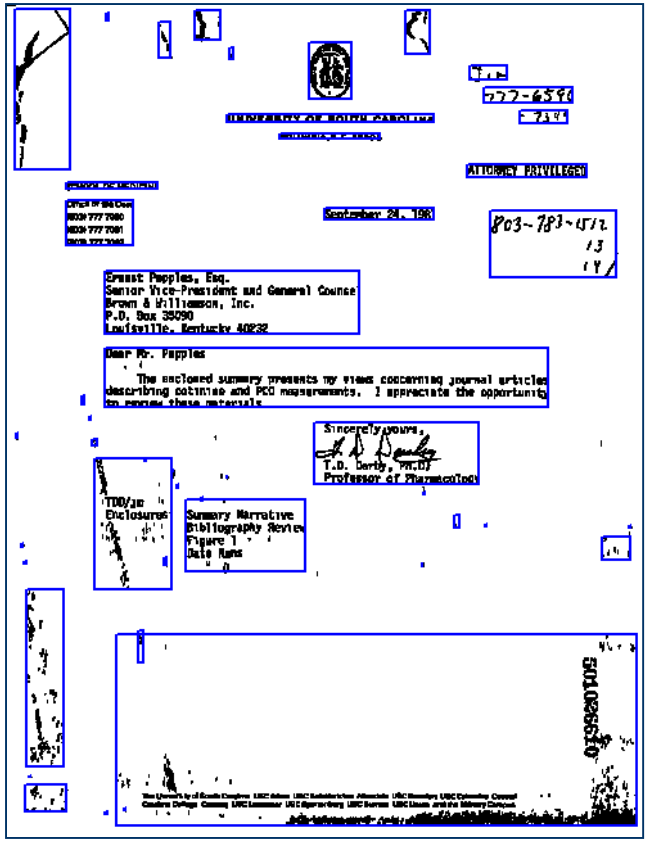
Avant post-traitement par MRF



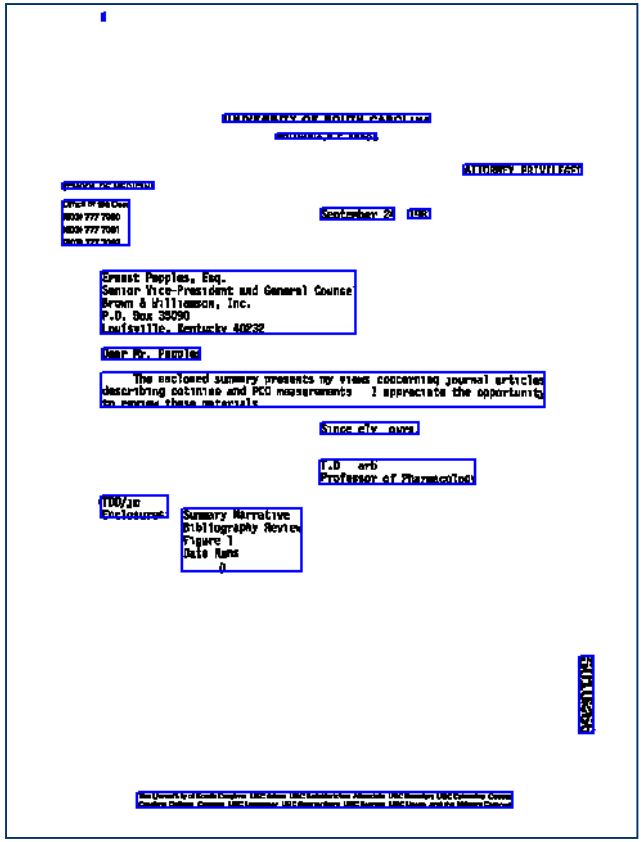
Après post-traitement par MRF

MRF: Markov Random Field: prend en compte le contexte

Application à la segmentation de la page



Avant réhaussement



Après réhaussement

Identification de la langue

- Détection de l’inclinaison indépendamment de la langue
- Accommoder écriture horizontale et verticale
- Reconnaissance de classe de script
- Scripts asiatiques ont des caractères sous forme de blocs
- Les scripts connectés ne peuvent pas être segmentés facilement
- Identification des langues
- Les statistiques de forme fonctionnent bien pour les langues occidentales
- Les classificateurs concurrents fonctionnent pour les langues asiatiques

Qu’en est-il du manuscrit ?

Optical Character Recognition

- Approche d'appariement de modèles
  - Approche standard dans les systèmes commerciaux
  - Segmenter les caractères individuels
  - Reconnaître en utilisant un classificateur de type réseau neuronal
- Approche du modèle de Markov caché
  - Approche expérimentale développée à BBN
  - Segmenter en tranches de sous-caractères
  - Lookahead limité pour trouver le meilleur choix de caractères
  - Utile pour les scripts connectés (par exemple, l'Arabe)

Problèmes de précision des OCR

- Erreurs de segmentation de caractères
  - En anglais, la segmentation change souvent "m" en "rn"
- Confusion de caractères
  - Les caractères avec des formes semblables sont souvent confondus
- OCR sur les copies est bien pire que sur les originaux
  - coupure de caractère, fission, pliage
- Les polices peu communes peuvent causer des problèmes
  - Si elles ne sont pas utilisées pour l'entrainement du réseau neuronal

Amélioration de la précision des OCRs

- Prétraitement de l'image
  - Morphologie mathématique pour la floraison et le fractionnement
  - Particulièrement important pour les images dégradées
- Le «vote» entre plusieurs moteurs OCR
  - Les systèmes individuels dépendent de données de formation spécifiques
- L'analyse linguistique peut corriger certaines erreurs
  - Utilisez les statistiques de confusion, les listes de mots, la syntaxe, ...
  - Mais des erreurs plus nocives pourraient être introduites

Vitesse des OCR

- Les réseaux neuronaux prennent environ 10 secondes par page
  - Les modèles de Markov cachés sont plus lents
- Le vote peut améliorer la précision
  - Mais à une vitesse de pénalité substantielle
- Facile à accélérer les choses avec plusieurs machines
  - Par exemple, par traitement par lots - utilisant des ordinateurs de bureau la nuit

Problème : analyse logique de la page (ordre de lecture)

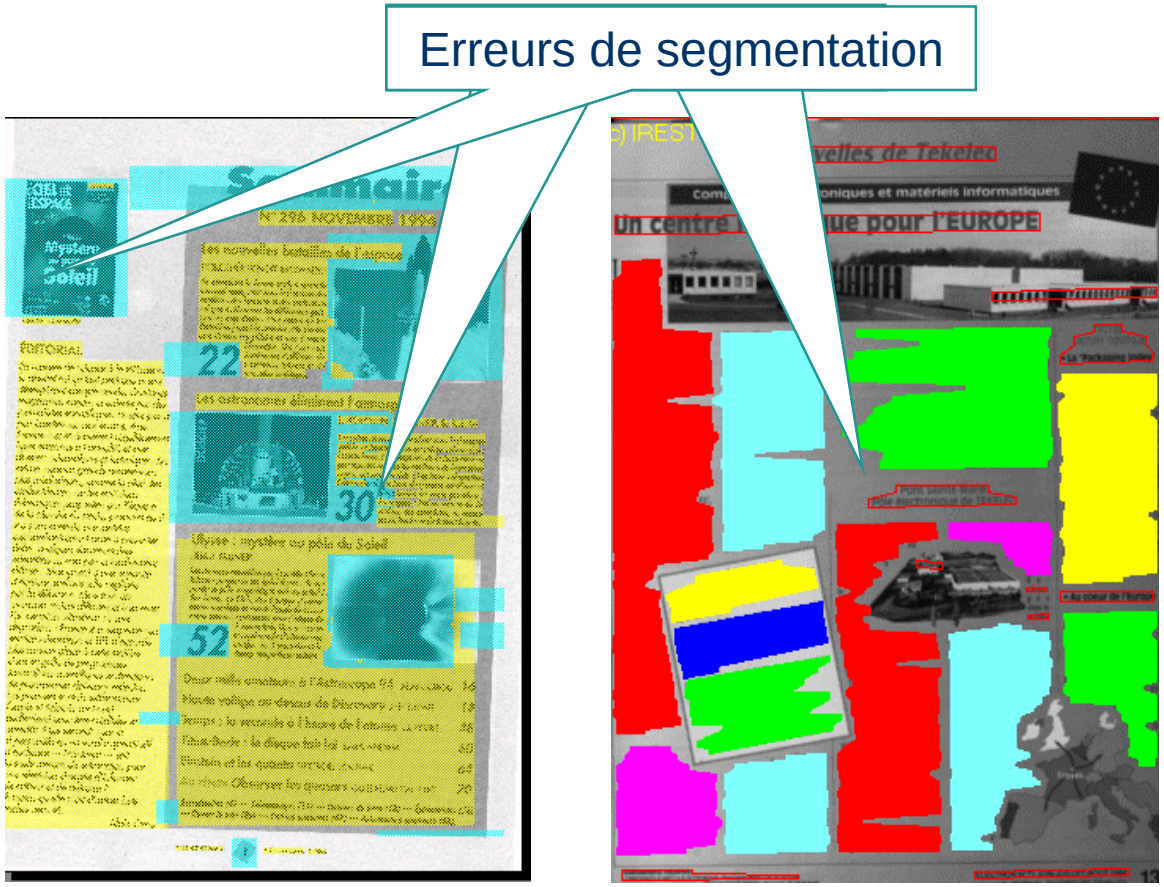
- Peut être difficile à deviner dans certains cas
  - Colonnes de journaux, légendes, appendices, ...
- Parfois, il existe des guides explicites
  - «Suite à la page 4» (mais la page 4 peut être grande!)
- Les repères structurels peuvent aider
  - La colonne 1 pourrait continuer à la colonne 2
- L'analyse de contenu est également utile
  - Statistiques de cooccurrence de mots, analyse de syntaxe

Document : traitement

- Intégration de la connaissance métier
  - Un document est destiné à véhiculer un message
  - Pour faciliter la compréhension, des efforts ont été faits pour uniformiser les documents avec un certain nombre de conventions :
    - Sur la langue, les formes de caractères, la typographie (style, police, mise en page ...), la structure des documents
    - Selon la fonction du document, il existe différentes conventions d'écriture et de présentation
    - Besoin de reconnaître le type de document
    - Peu de travaux sur l'identification du type de document pour appliquer une reconnaissance appropriée
  - Le critère de choix est important !

Exemple 1:
Choix d'un critère pour la séparation texte/image

- Critère : un texte est composé d'alignements de caractères



➔ Le manque de familiarité avec les difficultés réelles

Exemple 2 :
Séparation Texte / Formules

- Critère : une formule est composée de blocs isolés du texte

La deuxième solution est soit $> a$ si $a < 2$, soit < 0 si $a > 2$ *ce qui semble implicitement exclu*. Je ne retiens que la première solution. La formule trouvée convient aussi dans le cas où $a = 2$. Nous avons donc :

$$b = \varphi(a) = \frac{2a}{2+a} \quad r' = \frac{a^2}{4} \quad r'' = \frac{b^2}{4}.$$

b) L'intervalle $]0, \infty[$ est stable par la fonction φ ; la suite (a_n) est donc bien définie et à valeurs dans $]0, \infty[$. Comme sur cette intervalle $\varphi(x) < x$, pour tout n, $a_{n+1} = \varphi(a_n) < a_n$.

La suite (a_n) est donc strictement décroissante et minorée par 0, donc convergente vers un nombre ≥ 0. Ce nombre L doit vérifier, puisque φ est continu sur $[0, \infty[$, $\varphi(L) = L$, ce qui implique $L = 0$. La suite (a_n) est donc strictement positive, strictement décroissante de limite nulle.

Critère insuffisant

La deuxième solution est soit $> a$ si $a < 2$, soit < 0 si $a > 2$ *ce qui semble implicitement exclu*. Je ne retiens que la première solution. La formule trouvée convient aussi dans le cas où $a = 2$. Nous avons donc :

$$b = \varphi(a) = \frac{2a}{2+a} \quad r' = \frac{a^2}{4} \quad r'' = \frac{b^2}{4}.$$

b) L'intervalle $]0, \infty[$ est stable par la fonction φ ; la suite (a_n) est donc bien définie et à valeurs dans $]0, \infty[$. Comme sur cette intervalle $\varphi(x) < x$, pour tout n, $a_{n+1} = \varphi(a_n) < a_n$.

La suite (a_n) est donc strictement décroissante et minorée par 0, donc convergente vers un nombre ≥ 0. Ce nombre L doit vérifier, puisque φ est continu sur $[0, \infty[$, $\varphi(L) = L$, ce qui implique $L = 0$. La suite (a_n) est donc strictement positive, strictement décroissante de limite nulle.

Exemple 2 :
Séparation Texte / Formules

- Critère 2 : Une formule est située dans le texte parce qu'elle est constituée :
  - De blocs de texte isolés
  - Ou d'éléments du texte autour de marques comme ('=', '<', '[', ']', des chiffres, des lettres grecques, des mots clés comme 'série', 'fonction' ...)
  - Ou de composantes spécifiques généralement dans les grandes formules

Critère suffisant

La deuxième solution $> a$ si $a < 2$, soit < 0 si $a > 2$ *ce qui semble implicitement exclu*. Je ne retiens que la première solution. La formule trouvée convient aussi dans le cas où $a = 2$. Nous avons donc :

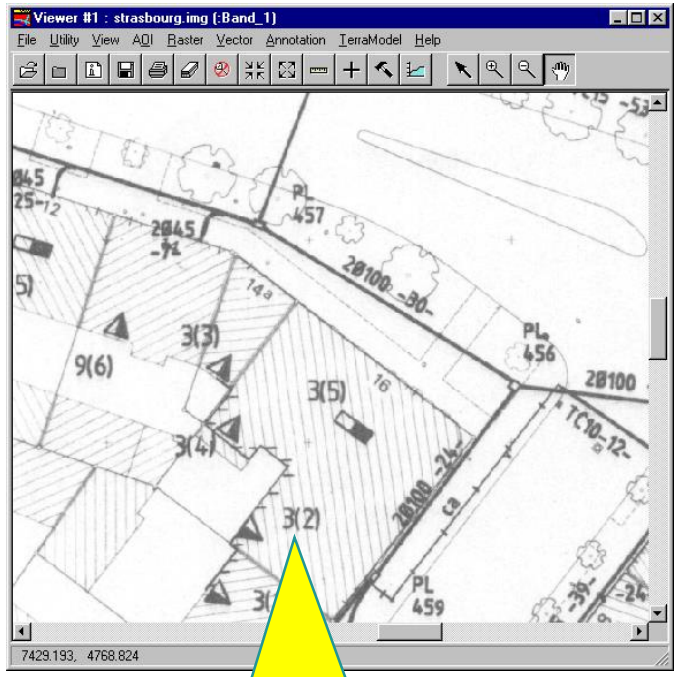
$$b = \varphi(a) = \frac{2a}{2+a} \quad r' = \frac{a^2}{4} \quad r'' = \frac{b^2}{4}.$$

b) L'intervalle $]0, \infty[$ est stable par la fonction φ ; la suite (a_n) est donc bien définie et à valeurs dans $]0, \infty[$. Comme sur cette intervalle $\varphi(x) < x$, pour tout n, $a_{n+1} = \varphi(a_n) < a_n$.

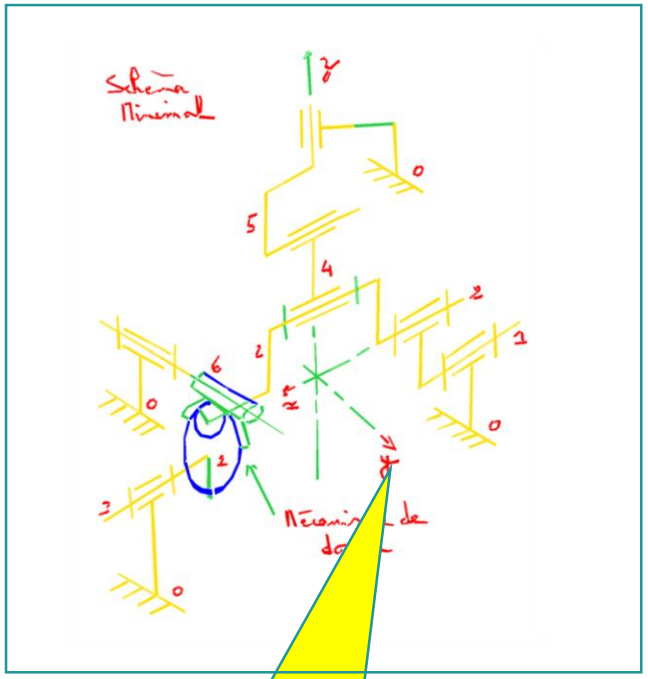
La suite (a_n) est donc strictement décroissante et minorée par 0, donc convergente vers un nombre ≥ 0. Ce nombre L doit vérifier, puisque φ est continu sur $[0, \infty[$, $\varphi(L) = L$, ce qui implique $L = 0$. La suite (a_n) est donc strictement positive, strictement décroissante de limite nulle.

Exemple 3 :
Séparation Texte / Graphique

- **Critère :** Le texte est composé de composantes connexes éloignées des longs traits



Critère suffisant



Critère insuffisant

Exemple 4 :
Séparation Texte /Vidéo

- Critère : Le texte est formé de composantes connexes alignées de la même couleur qui apparaissent sur plusieurs images consécutives

Critère suffisant

Critère insuffisant

Extraction d'information

The structure

Typographical context

INTERNATIONAL
Exclusive: The CIA's War in Iraq

Textual context

Exclusive: The CIA's War in Iraq **OCR**

Syntactical context

CIA	War	Iraq
THE	Warm	Iraq

Logical context

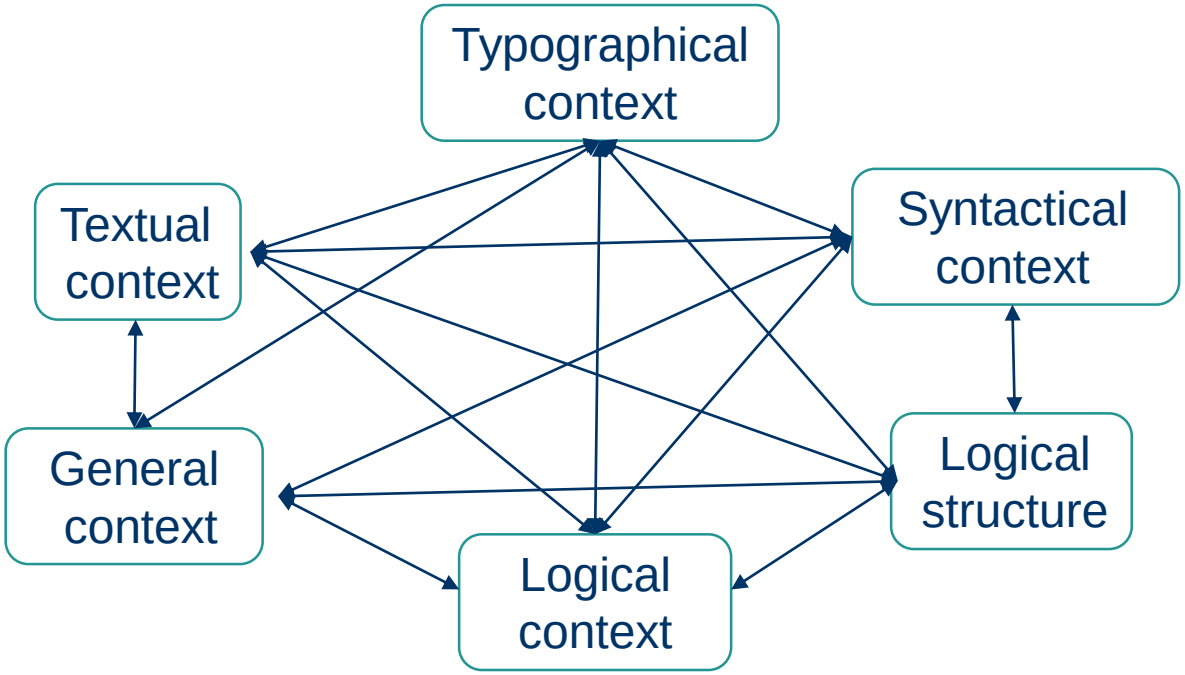
text,abstract,author,reference,n° page,photo,title...

General context

Summary? Letter? Book? Journal paper?

Processus d'extraction d'information

- Utiliser autant d'informations que possible à tous les niveaux
- Problème : l'information est inter-dépendante !

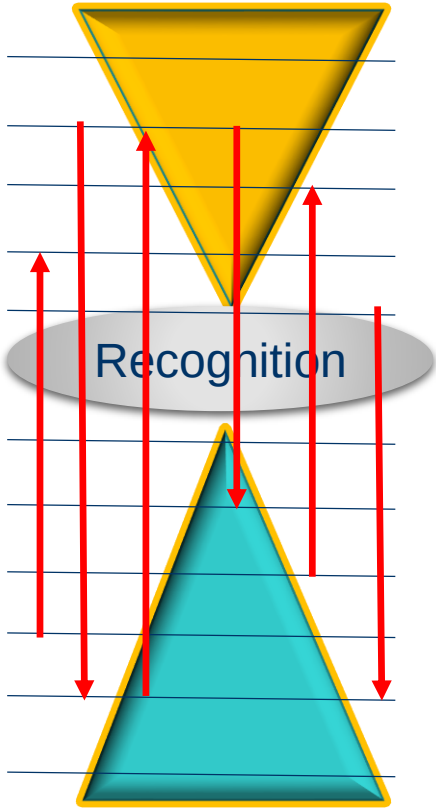


Approche dépendant du niveau sémantique

- Descendante : de la connaissance aux données
- Ascendante : des données à la connaissance
- Mixte: aller-retour entre niveaux

Approche descendante

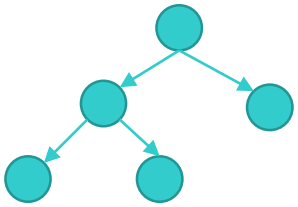
Culture générale
Connaissances particulières sur un document
Structure logique et fonctionnelle
Reconnaissance sémantique
Reconnaissance de texte adaptée



Syntaxe
Reconnaissance de caractères
Typographie
Structure physique
Segmentation
Pixels

Approcha ascendante

Approche utilisant des cycles perceptifs



Document model

Pattern grouping in objects

Hypothesis emission

Hypothesis validation

Coherence evaluation

Internal

External

Pattern recognition

Questionning of the segmentation

Segmentation

Images

Research of characteristic elements of an object

Retrieval du texte d'OCR

- Nécessite des méthodes d'indexation robustes
- Les méthodes statistiques avec de grands documents fonctionnent le mieux
- Évaluations clés
  - Le succès de la ROC de haute qualité (Croft et al 1994, Taghva 1994)
  - Succès limité pour un OCR de mauvaise qualité (1996 TREC, UNLV)
  - Clustering réussie pour une précision> 85% (Tsuda et al, 1995)

N-Grams

- Méthode statistique puissante et peu coûteuse pour caractériser les populations
- Approche
  - Diviser le document en paires de caractères n : échoue
  - Utiliser des représentations d'indexation traditionnelles pour effectuer des analyses
  - "DOCUMENT" -> DOC, OCU, CUM, UME, MEN, ENT
- Avantages
  - Statistiquement robustes à un petit nombre d'erreurs
  - Indexation rapide et récupération
  - Fonctionne entre 70% et 85% de précision des caractères lorsque l'IR traditionnelle échoue

Matching avec les erreurs OCR

- Au-dessus de 80% de précision des caractères, utilisez des mots
  - Avec correction linguistique
- Entre 75% et 80%, utiliser des n-grammes
  - Avec n un peu plus faible que d'habitude
  - Et peut-être avec des statistiques de confusion de caractère
- En dessous de 75%, utilisez des codes de forme de longueur de mot

Traitement d’images de texte

- Les caractéristiques
  - Ne nécessite pas de coûteuses opérations d’OCR / Conversion
  - Applicable aux applications de filtrage
  - Peut être plus résistant au bruit
- Inconvénients possibles
  - Le domaine d'application peut être très limité
  - Le temps de traitement peut être un problème si l'indexation est autrement requise

Keyword Spotting

Techniques:

- Work Shape/HMM - (Chen et al, 1995)
- Word Image Matching - (Trenkle and Vogt, 1993; Hull et al)
- Character Stroke Features - (Decurtins and Chen, 1995)
- Shape Coding - (Tanaka and Torii; Spitz 1995; Kia, 1996)

Applications:

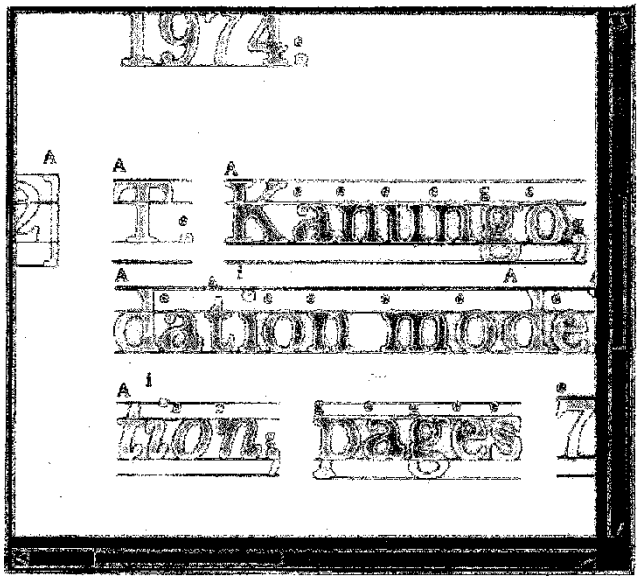
- Filing System (Spitz - SPAM, 1996)
- Numerous IR
- Processing handwritten documents

Evaluation formelle

- Scribble vs. OCR (DeCurtins, SDIUT 1997)

Codage de forme (shape coding)

- Approche
  - Utilisation de descripteurs génériques de caractères
  - Utiliser le pouvoir du langage pour résoudre l'ambiguïté
  - Carte Caractère basé sur les caractéristiques de la forme, y compris les ascendants, les descendants, la ponctuation et le caractère avec des trous



Applications supplémentaires

- Manuscrits d'archives manuscrits
  - (Manmatha, 1997)
- Classification de pages
  - (Decurtins et Chen, 1995)
- Correspondance des enregistrements manuscrits
  - (Ganzberger et coll., 1994)
- Extraction des titres
  - Compression d'images de documents (UMD, 1996-1998)

Une application industrielle

Traitement des formulaires

Le traitement des formulaires

- Aujourd'hui, un vrai marché
  - **Concerne toutes les administrations et les services**
    - qui manipulent de l'information de masse
      - salaires, factures…
      - plusieurs milliers par jour
  - **Il existe des bases métiers**
    - très riches et très spécifiques
      - plusieurs millions d'enregistrements
    - qui peuvent aider au traitement automatique

L'analyse du marché
les clients potentiels

Applications	Clients potentiels (France)	Nb de sites	Nb de clients potentiels
Assurance maladie	CPAM, MSA, …	700	700
Allocations familiales	CAF	100	100
Retraite	CRAM, Caisses de Retraites privées et publiques,…	500	200
Agences pour l'emploi	ANPE, APEC, ASSEDIC, et bureaux locaux	500	100
Services publics locaux (écoles, élections, sports, …)	Mairies	36000	1500
Services publics locaux (RMI, transports scolaires,…)	Conseils Généraux et Régionaux	100	100
Services publics locaux (création d'entreprise, …)	Chambres de Commerce, des Métiers, d'Agriculture	400	200
Services publics locaux (passeports, cartes d'identité, cartes grises, …)	Préfectures, Sous Préfectures	150	150
Taxes locales	Percepteur, Centres des Impôts, Douanes, …	300	100
Education et formation continue	Universités, centres de formation, …	500	100
Santé	Hôpitaux, cliniques, centres médicaux, …	500	100
Services de la relation clientèle	Grandes entreprises, Banques, Assurances, Opérateurs, sociétés de BtoC	10000	1000
	Total	**49750**	**4350**

Analyse du marché
Des spécialistes de la dématérialisation

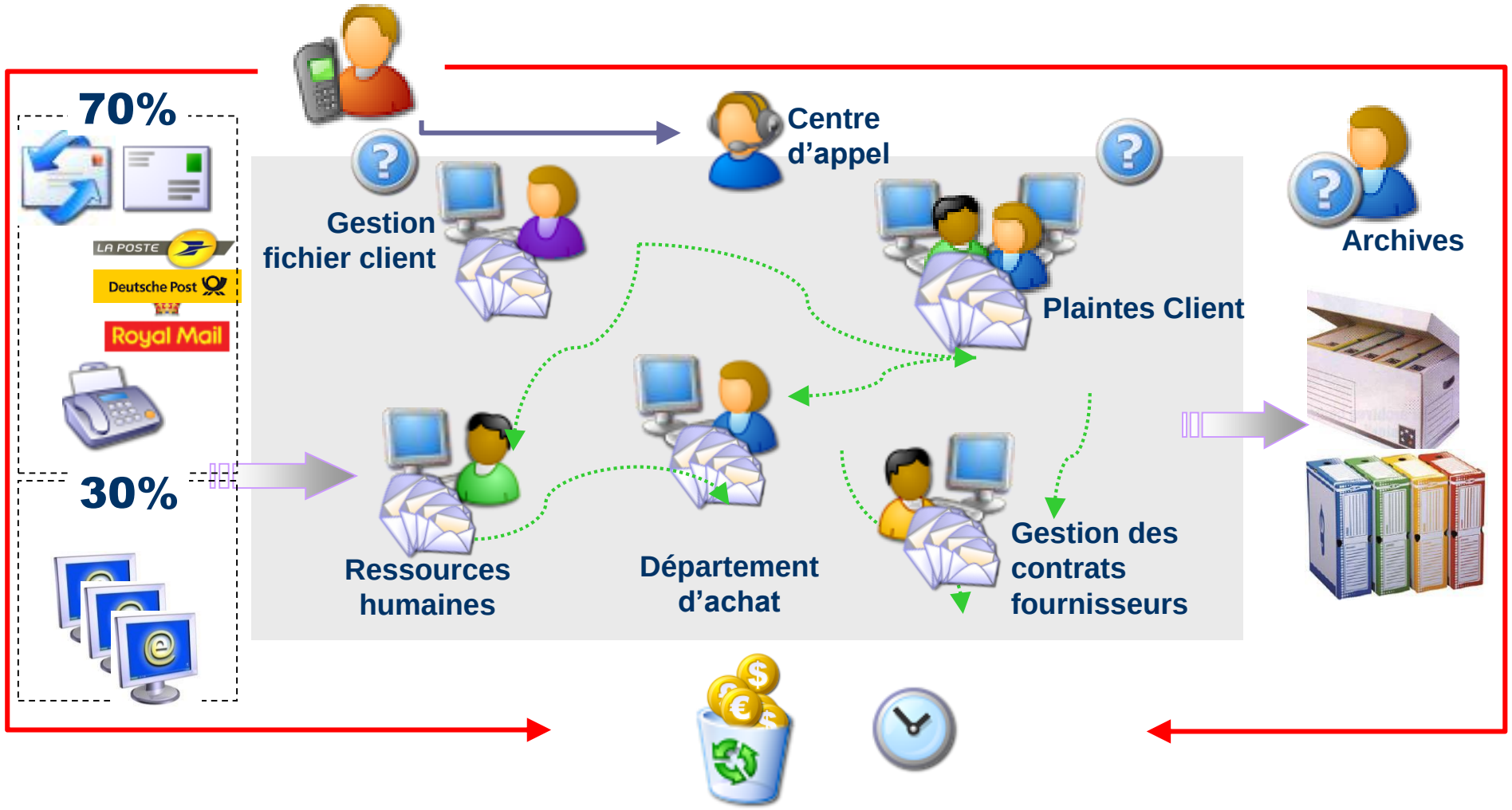
Acteur	Document	Produit	Traitement
Readsoft	Facture	«DOCUMENTS for Invoice»	Capture jusqu'à transformation
Captiva	Facture	« InputAccel for Invoices »	Tr. Automatique de facture – extraction d'information
Kofax	Document administratif	«Xtrata pro»	Classification document - extraction de données
BancTec	Chèque, remise, titre…	«eFIRST Clearing»	~ 5 milliards de chèques traités en France chaque année, 50% avec BancTec
IRIS	Courrier	«DocuTec»	Rétroconversion, Indexation
A2IA	Chèque	«CheckReader»	Lecture automatique de champs imprimés et manuscrits

- Plusieurs applications administratives
  - Dématérialisation des salles courriers : "Digital mailroom"
    - ➔ Traitement des flux entrants
  - Traitement automatique des moyens de paiement
    - ➔ Reconnaissance de montants : chèques, titres, traites…
  - Traitement de factures
    - ➔ Extraction d'identifiants
  - Gestion électronique de documents (contenus)
    - ➔ Circulation, diffusion, traçabilité, sécurisation, etc.

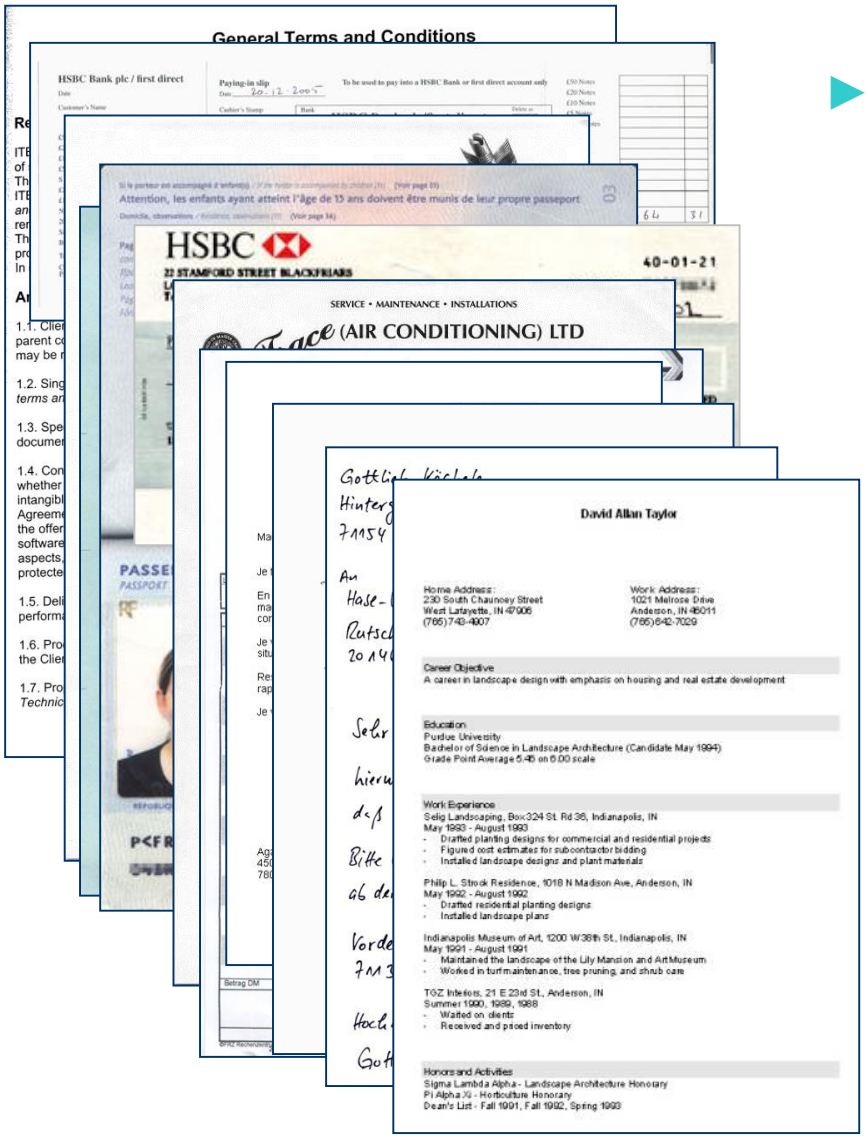
Problème du courrier

Le courrier met en interaction tous les départements de l'entreprise …
Le client envoie une commande, celle-ci fait un circuit…

La commande peut arriver par téléphone ou en complément…



Traitement du courrier : un processus difficile

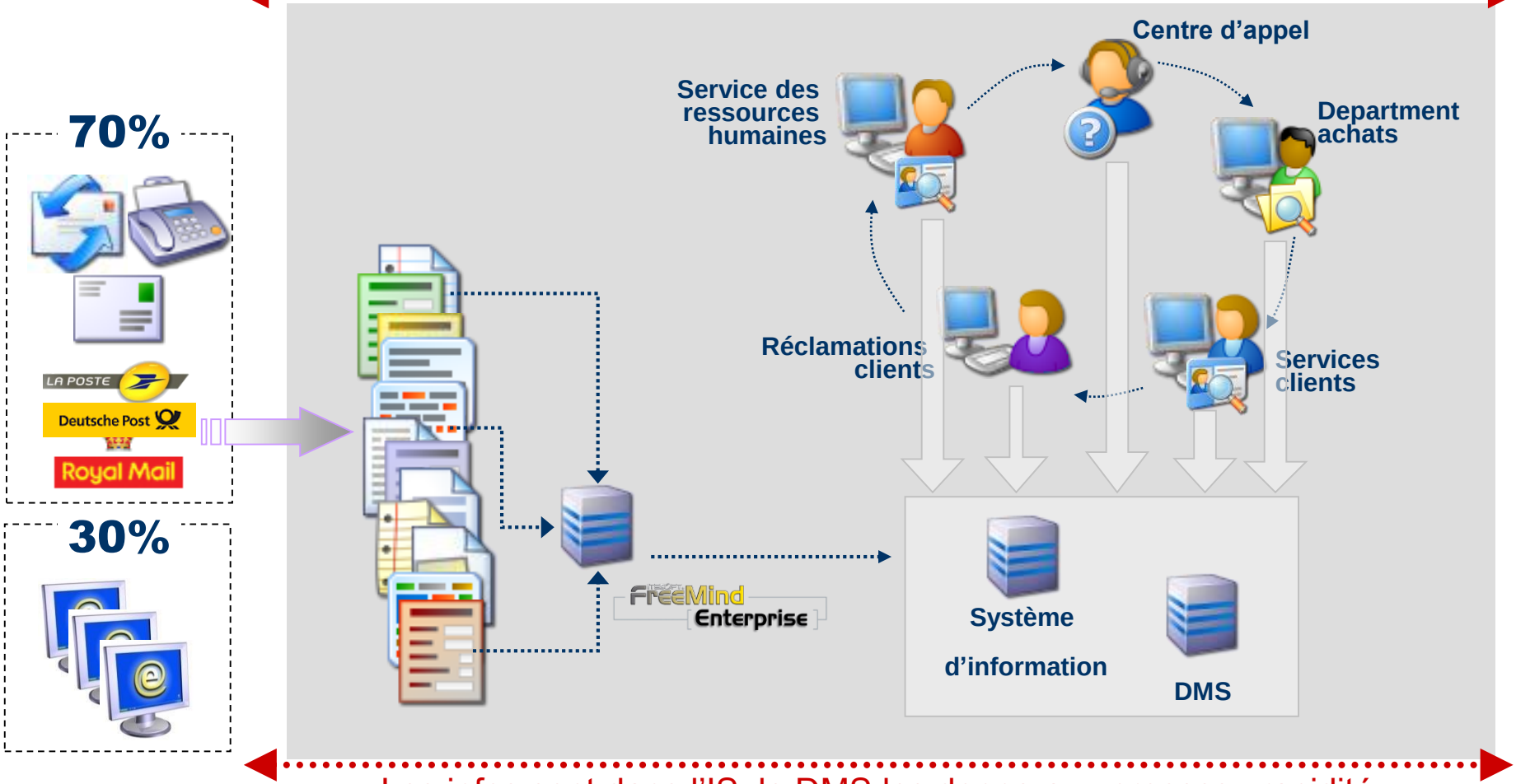


- **Le courrier n'est jamais le même, le format change … Ceci a une influence sur le process**
- **Il faut disposer d'une infrastructure d'exception et d'outils performants …**
- **L'efficacité des services dépend de la rapidité de la distribution du courrier**
- **La distribution sous forme papier n'est pas efficace, la traçabilité des documents est difficile**

Le traitement électronique du courrier :

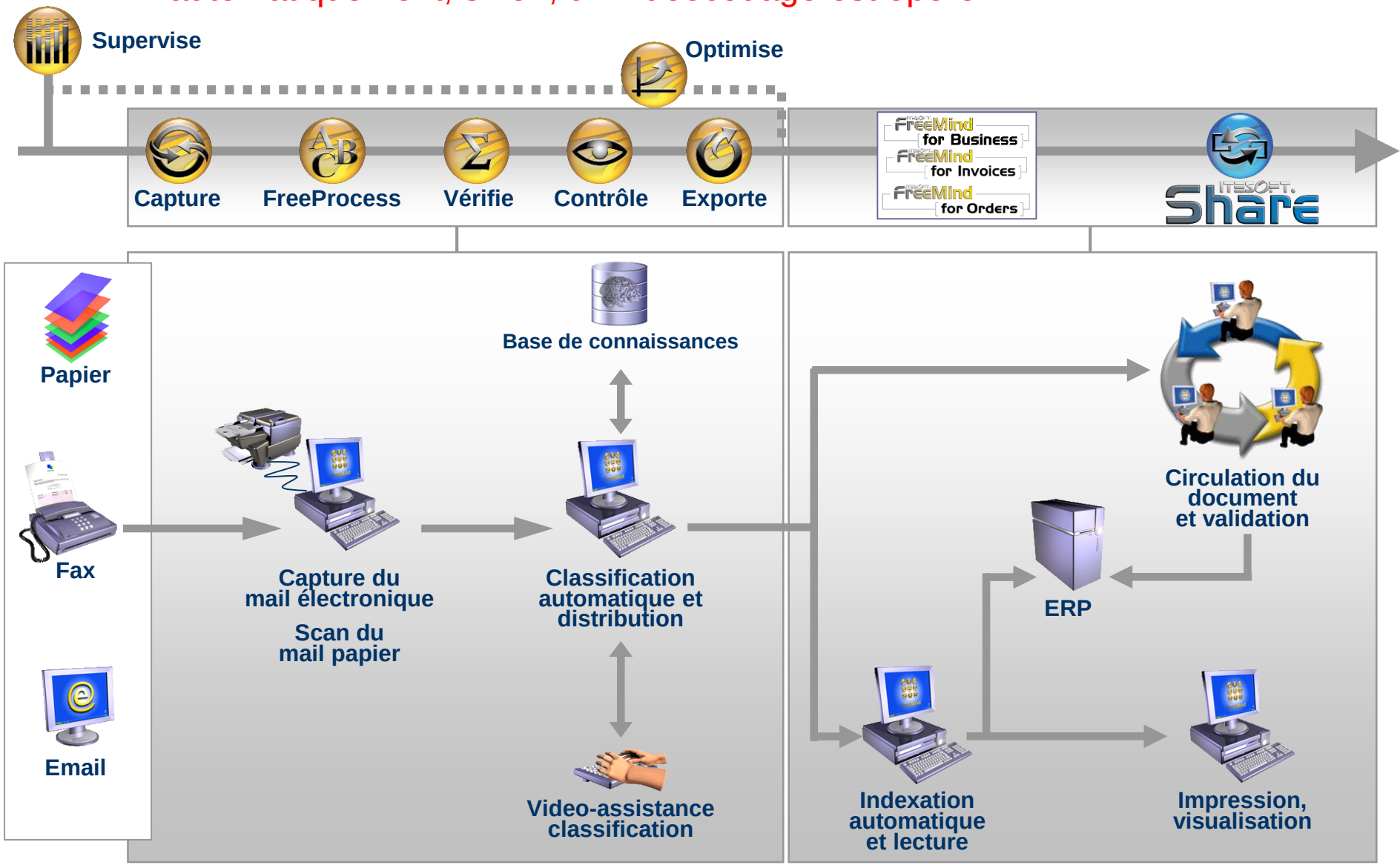
permet au système de l'insérer correctement dans la boucle

Plus clair pour sa distribution entre les services : coût, traçabilité



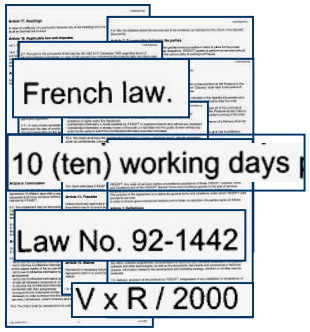
Les infos sont dans l'IS, le DMS les donne aux process : rapidité

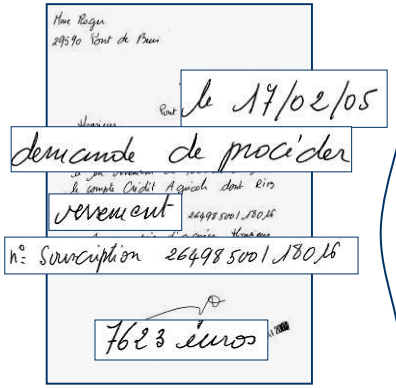
Le courrier est analysé en utilisant les connaissances de l'entreprise : elle essaie de le classer automatiquement, sinon, un vidéocodage est opéré

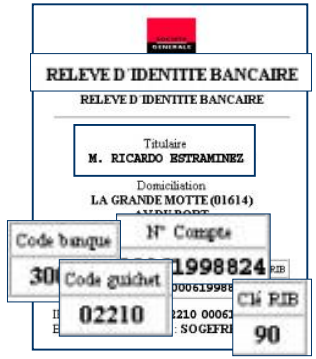


Traitement du texte : 2 phases :

>>> **Indexation pour repérer** **>>>** **Lecture ciblée** **>>>**

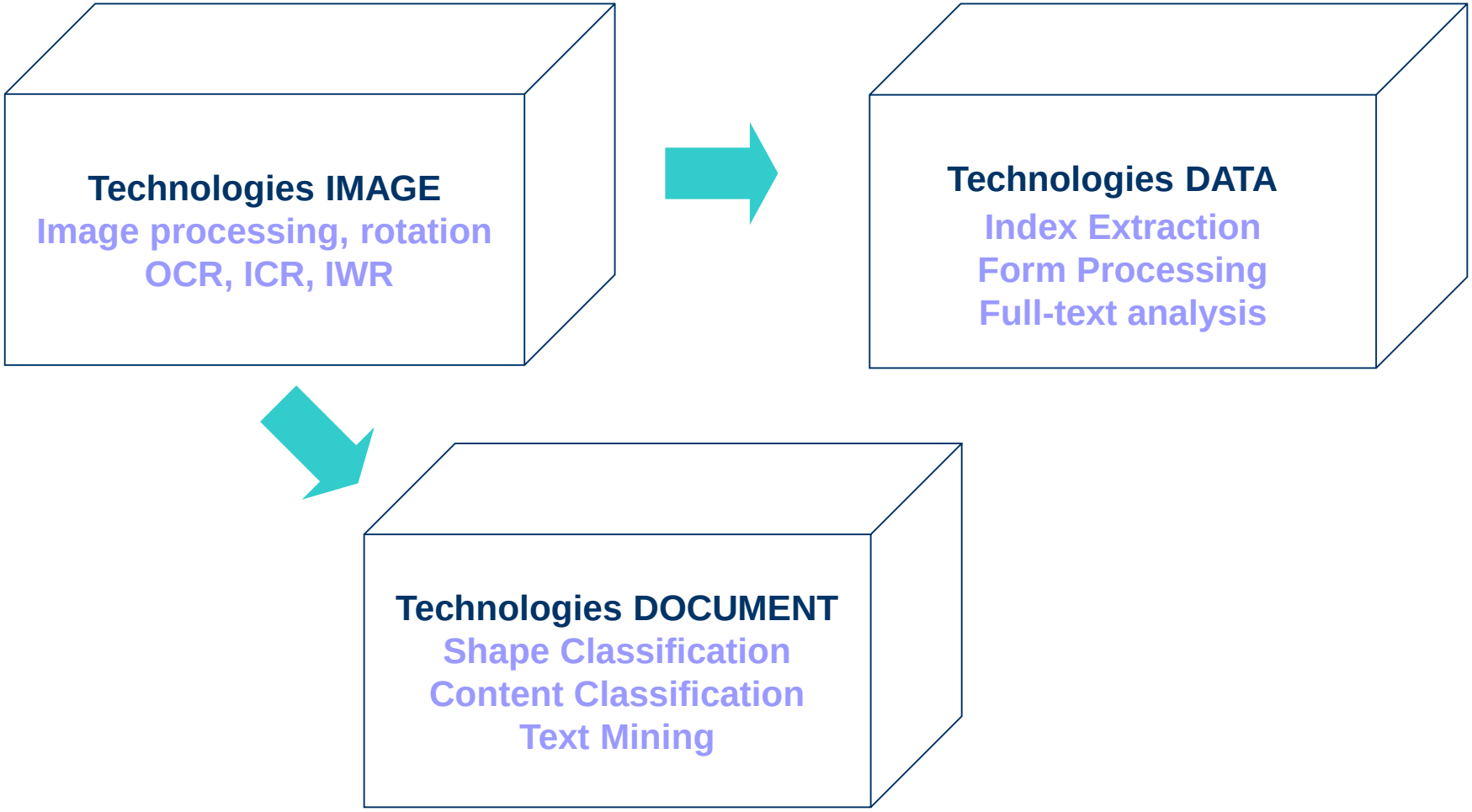






Pas d'intelligence, uniquement information métier

Les technologies utilisées

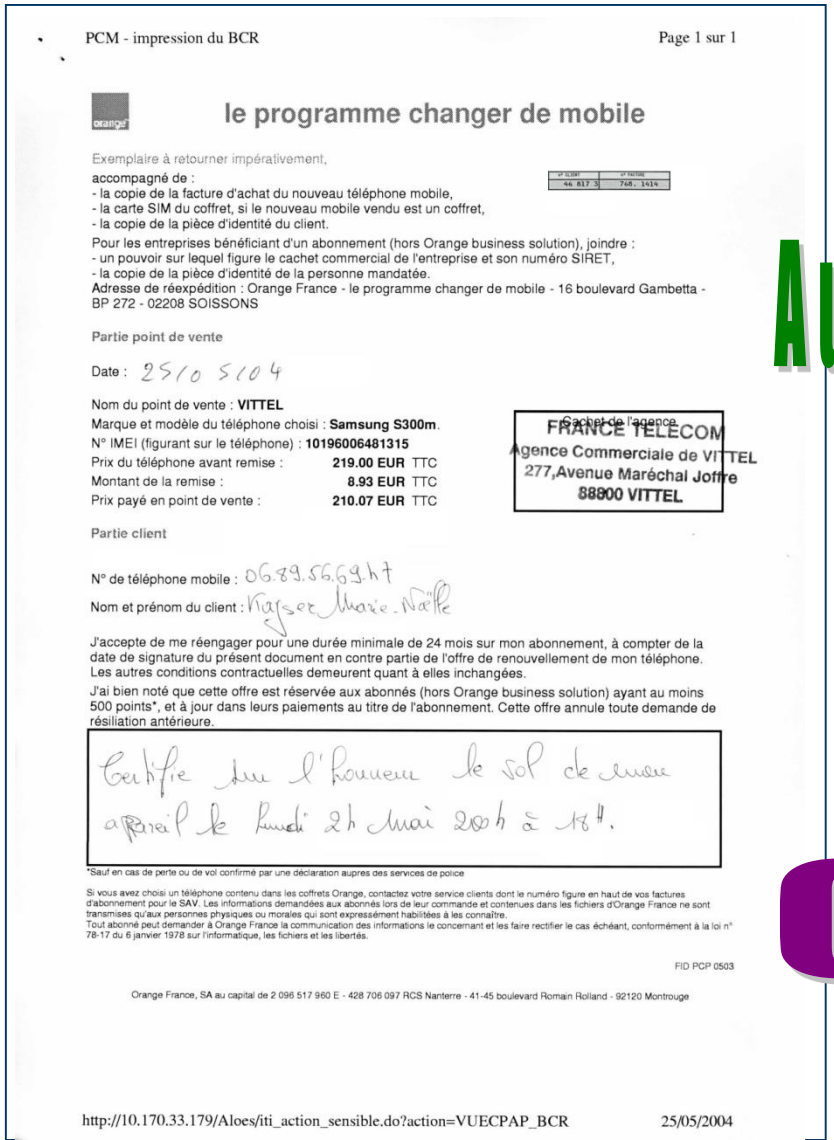


Combinaison de tecnhniques

Autorotation

IWR

Autodespeckle



PCM - impression du BCR Page 1 sur 1

le programme changer de mobile

Exemplaire à retourner impérativement,
accompagné de :
- la copie de la facture d'achat du nouveau téléphone mobile,
- la carte SIM du coffret, si le nouveau mobile vendu est un coffret,
- la copie de la pièce d'identité du client.

Pour les entreprises bénéficiant d'un abonnement (hors Orange business solution), joindre :
- un pouvoir sur lequel figure le cachet commercial de l'entreprise et son numéro SIRET,
- la copie de la pièce d'identité de la personne mandatée.

Adresse de réexpédition : Orange France - le programme changer de mobile - 16 boulevard Gambetta - BP 272 - 02208 SOISSONS

Partie point de vente

Date : 25/05/2004

Nom du point de vente : **VITTEL**
Marque et modèle du téléphone choisi : **Samsung S300m**.
N° IMEI (figurant sur le téléphone) : **10196006481315**
Prix du téléphone avant remise : **219.00 EUR** TTC
Montant de la remise : **8.93 EUR** TTC
Prix payé en point de vente : **210.07 EUR** TTC

Cachet de l'agence
FRANCE TELECOM
Agence Commerciale de VITTEL
277,Avenue Maréchal Joffre
88800 VITTEL

Partie client

N° de téléphone mobile : 06.89.56.69.47

Nom et prénom du client : Kaiser Marie-Noëlle

J'accepte de me réengager pour une durée minimale de 24 mois sur mon abonnement, à compter de la date de signature du présent document en contre partie de l'offre de renouvellement de mon téléphone. Les autres conditions contractuelles demeurent quant à elles inchangées.

J'ai bien noté que cette offre est réservée aux abonnés (hors Orange business solution) ayant au moins 500 points*, et à jour dans leurs paiements au titre de l'abonnement. Cette offre annule toute demande de résiliation antérieure.

Certifie sur l'honneur le vol de mon appareil le lundi 24 mai 2004 à 18 h.

*Sauf en cas de perte ou de vol confirmé par une déclaration auprès des services de police

Si vous avez choisi un téléphone contenu dans les coffrets Orange, contactez votre service clients dont le numéro figure en haut de vos factures d'abonnement pour le SAV. Les informations demandées aux abonnés lors de leur commande et contenues dans les fichiers d'Orange France ne sont transmises qu'aux personnes physiques ou morales qui sont expressément habilitées à les connaître.
Tout abonné peut demander à Orange France la communication des informations le concernant et les faire rectifier le cas échéant, conformément à la loi n° 78-17 du 6 janvier 1978 sur l'informatique, les fichiers et les libertés.

FID PCP 0503

Orange France, SA au capital de 2 096 517 960 E - 428 706 097 RCS Nanterre - 41-45 boulevard Romain Rolland - 92120 Montrouge

http://10.170.33.179/Aloes/iti_action_sensible.do?action=VUECPAP_BCR 25/05/2004

ICR

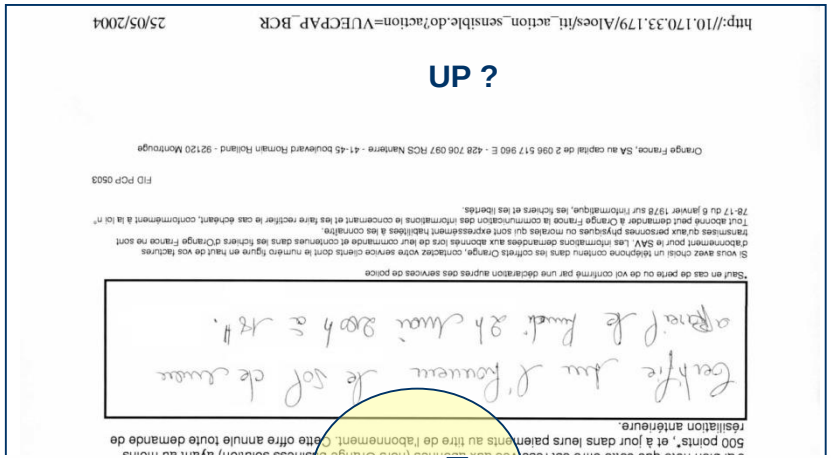
OCR

Autorotation :

Par l'orientation des lignes ou l'emplacement des hampes et jambages

Par l'orientation des lignes

Par l'emplacement des hampes et jambages



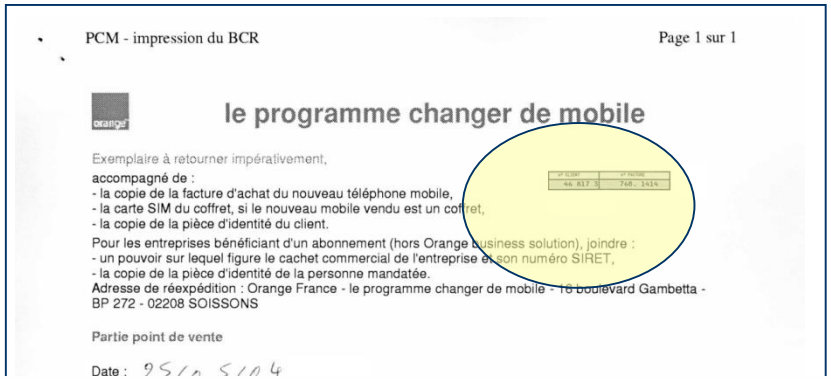
UP ?

PORTRAIT

Down ?

Autodespeckle

Localise les zones texturées et enlève les points : technique run length (transitions)



PCM - impression du BCR Page 1 sur 1

le programme changer de mobile

Exemplaire à retourner impérativement,
accompagné de :
- la copie de la facture d'achat du nouveau téléphone mobile,
- la carte SIM du coffret, si le nouveau mobile vendu est un coffret,
- la copie de la pièce d'identité du client.

Pour les entreprises bénéficiant d'un abonnement (hors Orange business solution), joindre :
- un pouvoir sur lequel figure le cachet commercial de l'entreprise et son numéro SIRET,
- la copie de la pièce d'identité de la personne mandatée.

Adresse de réexpédition : Orange France - le programme changer de mobile - 16 boulevard Gambetta - BP 272 - 02208 SOISSONS

Partie point de vente

Date : 25/0 5/04

n° CLIENT	n° FACTURE
46 817 3	768. 1414

e,

coffret,

nge business solution), joindre :

eprise et son numéro SIRET,

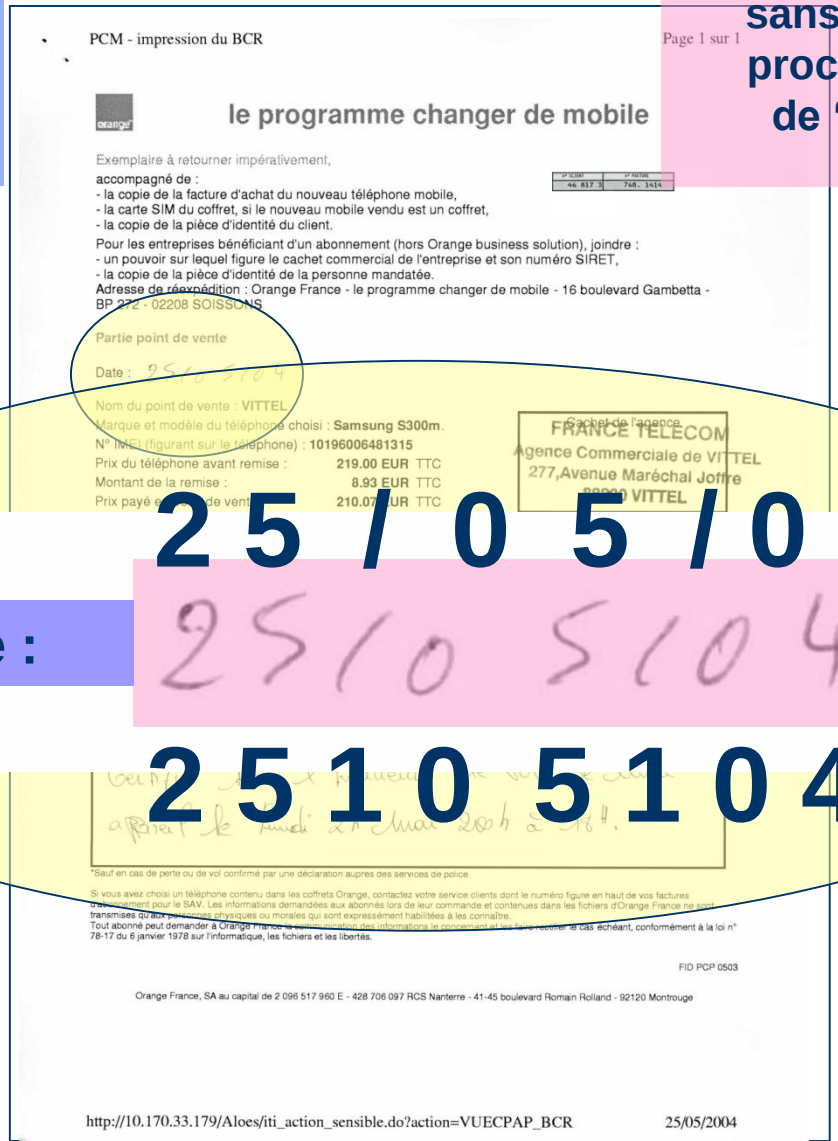
Différence ente OCR et ICR :

O C R

Permet de localiser le mot “date”, mais ne peut pas lire le reste

I C R

Lit la date mais se trompe sans contexte : 1 et / sont proches, la connaissance de “date” lui permet de corriger


PCM - impression du BCR

Page 1 sur 1

le programme changer de mobile

Exemplaire à retourner impérativement,
accompagné de :
- la copie de la facture d'achat du nouveau téléphone mobile,
- la carte SIM du coffret, si le nouveau mobile vendu est un coffret,
- la copie de la pièce d'identité du client.

Pour les entreprises bénéficiant d'un abonnement (hors Orange business solution), joindre :
- un pouvoir sur lequel figure le cachet commercial de l'entreprise et son numéro SIRET,
- la copie de la pièce d'identité de la personne mandatée.

Adresse de réexpédition : Orange France - le programme changer de mobile - 16 boulevard Gambetta - BP 272 - 02208 SOISSONS

Partie point de vente

Date : 25/05/04

Nom du point de vente : VITTEL
Marque et modèle du téléphone choisi : **Samsung S300m**.
N° IMEI (figurant sur le téléphone) : **10196006481315**
Prix du téléphone avant remise : **219.00 EUR** TTC
Montant de la remise : **8.93 EUR** TTC
Prix payé [illegible] de vente : **210.07 EUR** TTC

Cachet de l'agence
FRANCE TELECOM
Agence Commerciale de VITTEL
277,Avenue Maréchal Joffre
88800 VITTEL

*Sauf en cas de perte ou de vol confirmé par une déclaration auprès des services de police

Si vous avez choisi un téléphone contenu dans les coffrets Orange, contactez votre service clients dont le numéro figure en haut de vos factures pour tout règlement pour le SAV. Les informations demandées aux abonnés lors de leur commande et contenues dans les fichiers d'Orange France ne sont transmises qu'aux personnes physiques ou morales qui sont expressément habilitées à les connaître.
Tout abonné peut demander à Orange France la communication des informations le concernant et en faire rectifier le cas échéant, conformément à la loi n° 78-17 du 6 janvier 1978 sur l'informatique, les fichiers et les libertés.

FID PCP 0503

Orange France, SA au capital de 2 096 517 960 E - 428 706 097 RCS Nanterre - 41-45 boulevard Romain Rolland - 92120 Montrouge

http://10.170.33.179/Aloes/iti_action_sensible.do?action=VUECPAP_BCR 25/05/2004

Date :

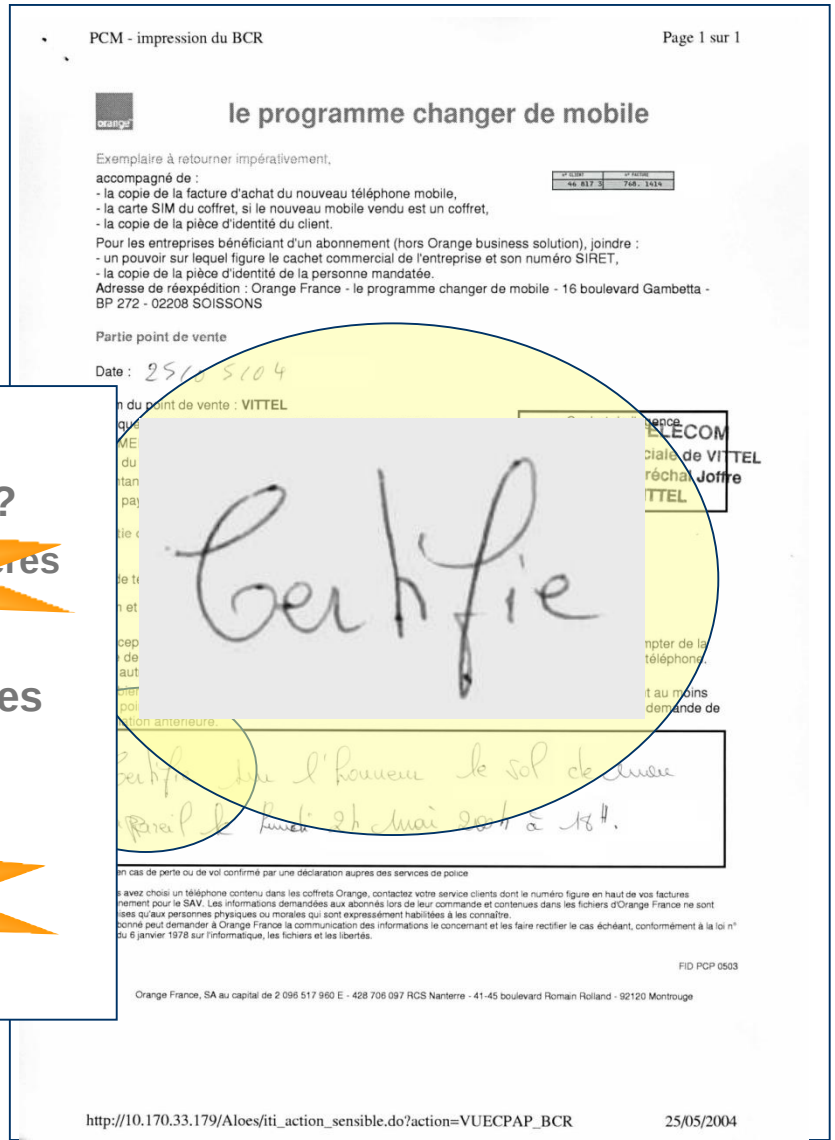
2 5 / 0 5 / 0 4

2510 5104

2 5 1 0 5 1 0 4

I W R

Français ou anglais ? Caractères collés. Le ti ressemble au h ?


PCM - impression du BCR

Page 1 sur 1

le programme changer de mobile

Exemplaire à retourner impérativement,

accompagné de :
- la copie de la facture d'achat du nouveau téléphone mobile,
- la carte SIM du coffret, si le nouveau mobile vendu est un coffret,
- la copie de la pièce d'identité du client.

Pour les entreprises bénéficiant d'un abonnement (hors Orange business solution), joindre :
- un pouvoir sur lequel figure le cachet commercial de l'entreprise et son numéro SIRET,
- la copie de la pièce d'identité de la personne mandatée.

Adresse de réexpédition : Orange France - le programme changer de mobile - 16 boulevard Gambetta - BP 272 - 02208 SOISSONS

Partie point de vente

Date : 25/05/2004

Orange France, SA au capital de 2 096 517 960 E - 428 706 097 RCS Nanterre - 41-45 boulevard Romain Rolland - 92120 Montrouge

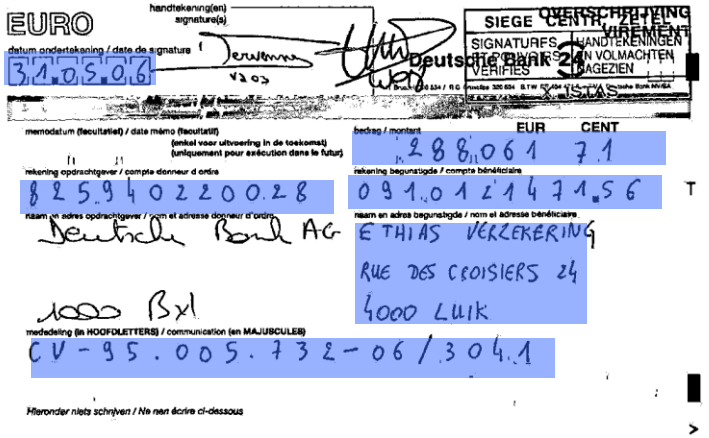
http://10.170.33.179/Aloes/iti_action_sensible.do?action=VUECPAP_BCR 25/05/2004

Problèmes

- Segmentation des mots ?
  - Pas de segmentation
- Lien avec des dictionnaires
  - Petit, moyen, large ?
  - Fermé ou ouvert ?
    - Moyen & Ouvert

Extraction de données

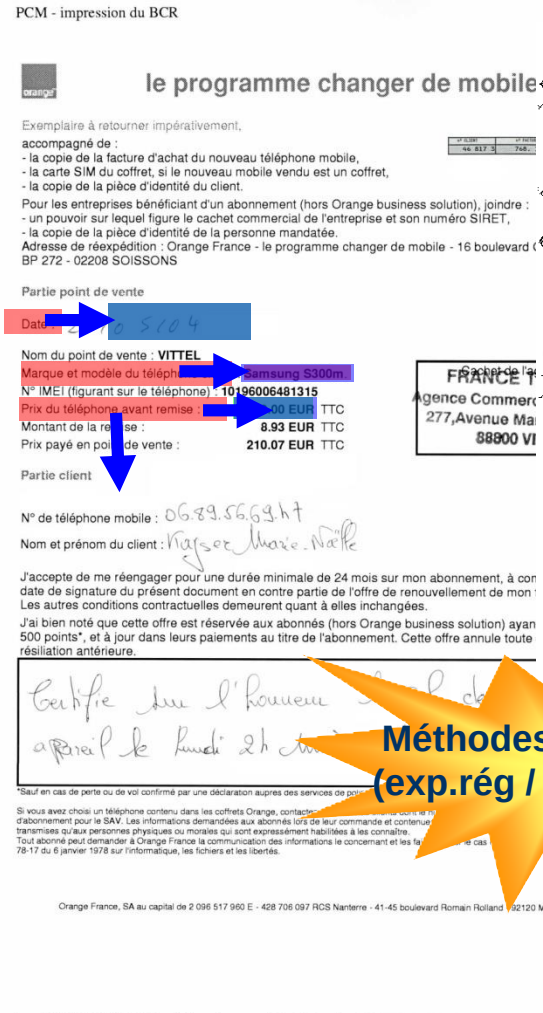
Structurées



1. Paper based domestic payment hand written

Méthode des masques

Semi-structurées

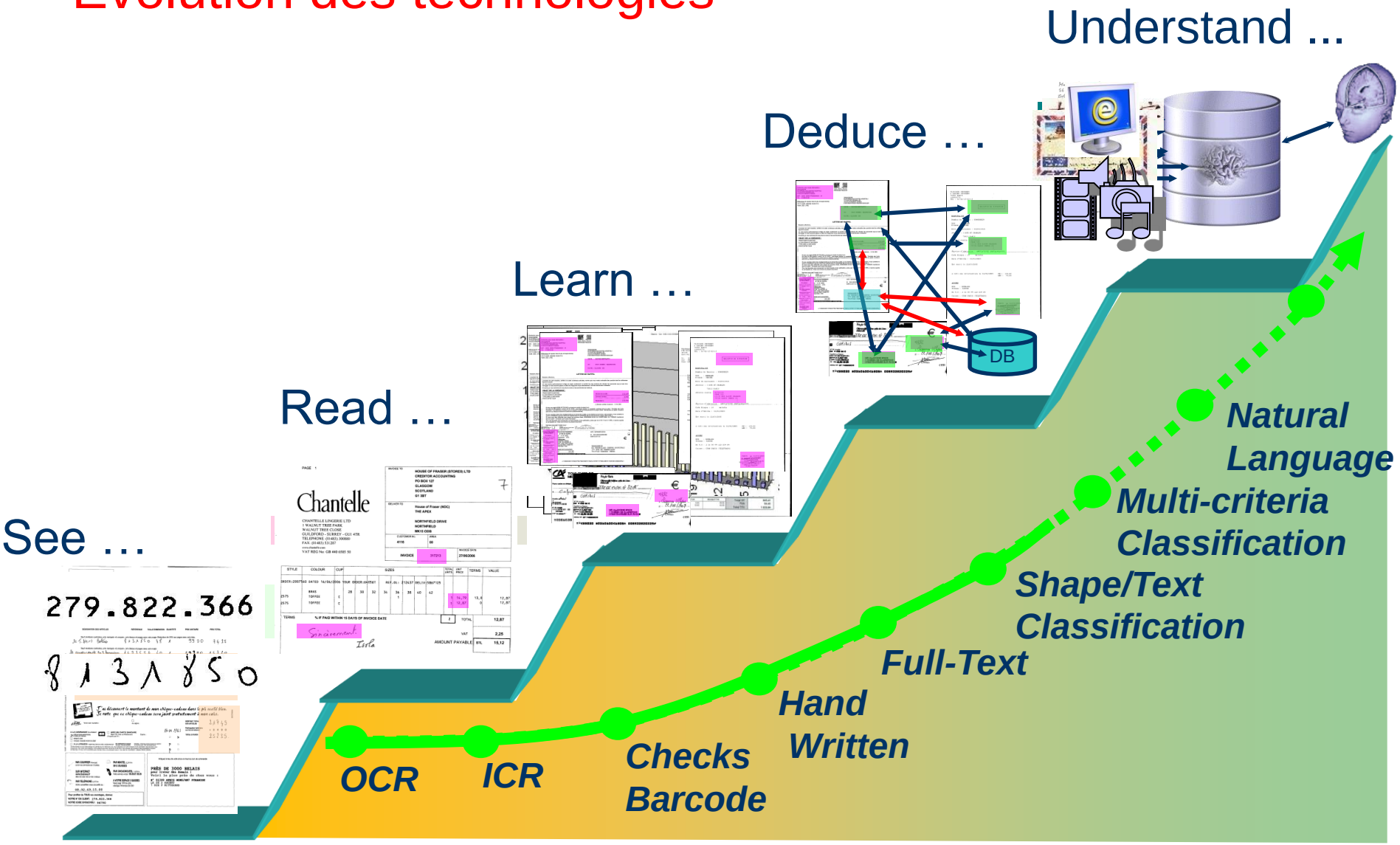


Non structurées

Méthodes plein texte (exp.rég / text mining)

Conclusion sur les formulaires

Évolution des technologies



Conclusion sur la READ

- 3 périodes
  - Un passé pessimiste
    - Rien ne fonctionnait
    - Les OCRs étaient voués à l'échec
    - Pas d'horizon pour le manuscrit
  - Un présent satisfaisant
    - Entrée dans le monde industriel
    - On peut faire de la masse rapidement
    - On a noué des liens avec le NLP, le Data Mining (ce n'est plus une thématique isolée, mais un élément d'un grand ensemble)
  - Un futur prometteur
    - Capable de répondre à beaucoup de défis dans des problématiques réelles : numérisation du patrimoine ancien
    - OCR : Élément essentiel des moteurs de recherche
    - Domaine de recherche ouvert...