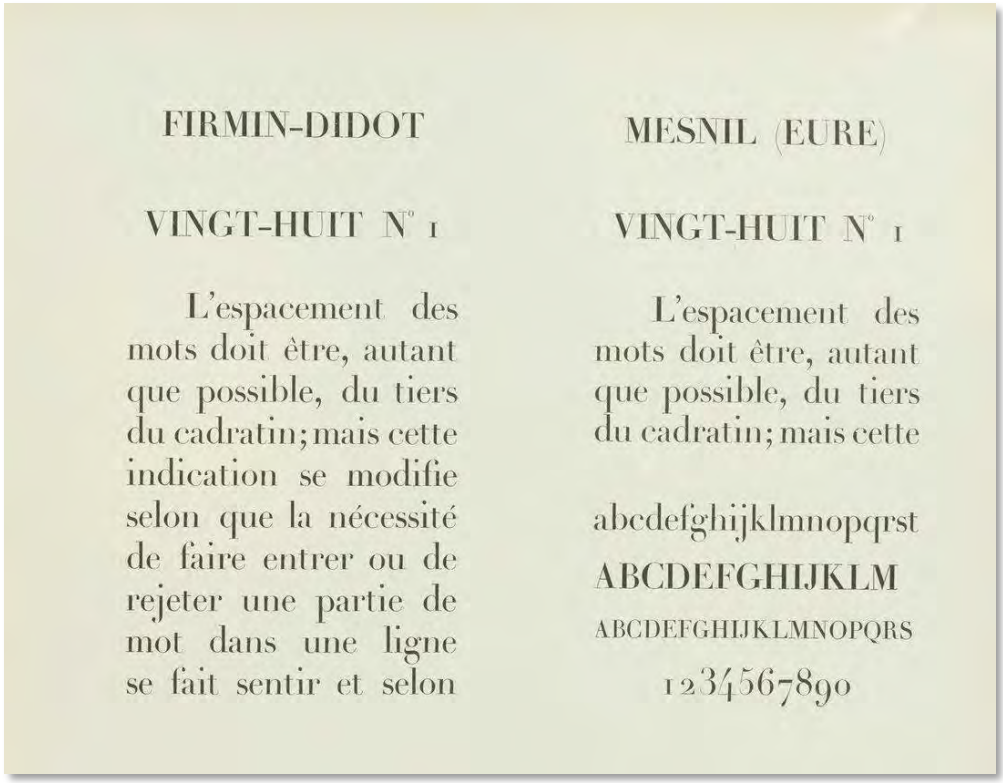
{BnF Bibliothèque nationale de France

Référentiel OCR



FIRMIN-DIDOT

VINGT-HUIT N° 1

L'espacement des mots doit être, autant que possible, du tiers du cadratin; mais cette indication se modifie selon que la nécessité de faire entrer ou de rejeter une partie de mot dans une ligne se fait sentir et selon

MESNIL (EURE)

VINGT-HUIT N° 1

L'espacement des mots doit être, autant que possible, du tiers du cadratin; mais cette

abcdefghijklmnopqrst

ABCDEFGHIJKLM

ABCDEFGHIJKLMNOPQRS

1234567890

Bibliothèque nationale de France direction des Services et des réseaux département de la Conservation service Numérisation	Date :le jeudi 5 décembre 2013 Version :1 Référence BnF :BnF-ADM-2013-081159-01

TABLE DES MATIERES

1. INTRODUCTION

1.1 Contexte

La Bibliothèque nationale de France a lancé divers programmes concourant à la constitution d'une bibliothèque numérique. Ces programmes s'appuient notamment sur des marchés de dématérialisation des collections de la BnF et de bibliothèques françaises partenaires.

En sus de la numérisation proprement dite des documents, un certain nombre d'autres prestations de dématérialisation sont demandées, dont la reconnaissance et la conversion des contenus textuels des documents numérisés.

En effet, la BnF désire également donner accès au contenu des documents grâce à la recherche à partir du texte des pages. La mise en place de ce procédé implique donc la conversion en mode texte de l'intégralité du contenu des pages afin de permettre la recherche plein texte, quelle que soit la partie consultée, puis l'affichage des images du document correspondant, avec possibilité d'accéder aux données en mode texte pour faire des sélections, des copies, des impressions.

Cette conversion en mode texte s'appuie principalement sur des techniques de reconnaissance optique de caractères (OCR, *optical character recognition*).

Ce document est organisé en plusieurs parties :

- Un chapitre « Généralités » qui expose les principes généraux de la reconnaissance optique de caractères.
- Un chapitre « OCR » qui traite de l'extraction des contenus.
- Un chapitre « Segmentation et structuration » qui traite de l'extraction de la structure du document.
- Un chapitre « Qualité » qui décrit les principes et méthodes de mesure de la qualité.

1.2 Objet

Le présent référentiel définit les caractéristiques attendues pour le traitement de reconnaissance optique de caractères appliqué aux documents des départements de Bibliothèque nationale de France et des bibliothèques partenaires. Il détaille les caractéristiques techniques des fichiers, les modalités de contrôle, etc.

1.3 Domaine d'application

Le présent référentiel s'applique aux prestations de numérisation d'ouvrages commandées par la Bibliothèque nationale de France, sur des marchés de numérisation ou de réfection, ainsi qu'aux éventuelles productions internes.

2. DOCUMENTS APPLICABLES ET DE REFERENCE

Schéma ALTO LoC	http://www.loc.gov/standards/alto
Schéma ALTO BnF	http://bibnum.bnf.fr/alto_prod/
BnF : Conversion en mode texte	http://www.bnf.fr/fr/professionnels/num_conversion_texte/s.num_conversion_texte_ocr.html
Référentiel d'enrichissement des métadonnées	version 2
Référentiel de livraison de document numérique	version 1
Référentiel tables	version 1

3. GENERALITES

3.1 OCR

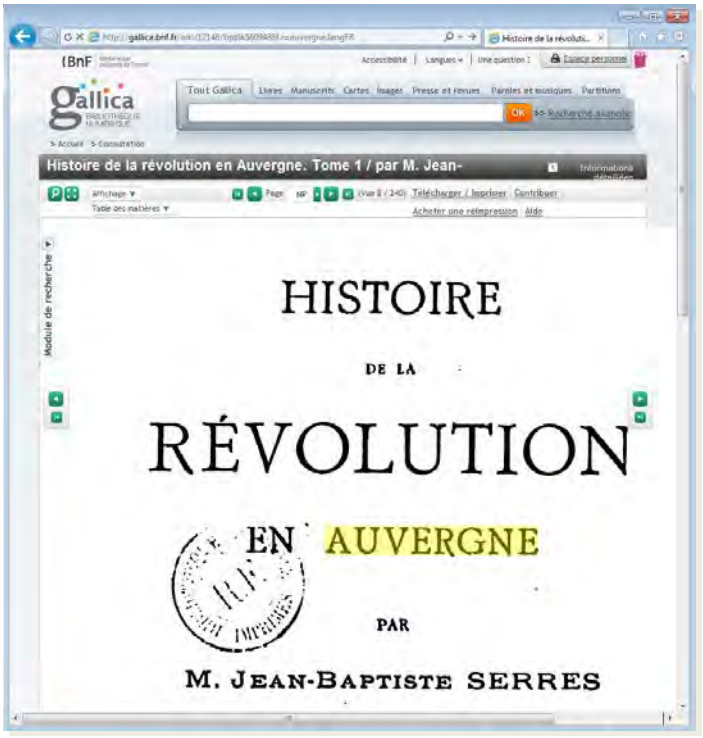
La reconnaissance optique de caractères désigne les procédés informatiques visant à extraire le texte présent dans l'image d'un texte imprimé. Un système OCR part donc de l'image numérique réalisée par un scanner optique ou une caméra numérique d'une page (document imprimé, feuillet dactylographié, documents transparents, etc.), et produit en sortie un fichier texte en divers formats.

Ces systèmes OCR opèrent en plusieurs étapes :

1. *Pré-analyse de l'image*, visant à améliorer la qualité de l'image en vue de faciliter la reconnaissance des caractères (redressement d'images inclinées, des corrections de contraste, binarisation de l'image).
2. *Segmentation des contenus*, permettant d'isoler dans l'image les différentes composantes (illustrations, blocs de texte, marges, etc.).
3. *Reconnaissance des caractères* : le texte contenu dans les blocs isolés à l'étape précédente est analysé et reconnu.

La segmentation physique de la page (étape 2) permet de mettre en correspondance le texte issu de l'OCR avec son emplacement dans l'image de la page. Cette correspondance est utilisée lors de la consultation des documents de la bibliothèque numérique Gallica, en permettant la mise en valeur, sur l'image du document, du ou des mots-clés recherchés.

Elle est également utilisée lors de la génération de documents PDF restituant le document numérisé (en mode image) mais aussi son contenu textuel (afin de rendre possible la recherche en texte intégral).



Recherche en texte intégral et mise en surbrillance du critère de recherche dans Gallica

3.1.1 Documents éligibles à la conversion OCR

Les collections de la BnF sont constituées d'une grande variété de typologies d'ouvrages. Du fait des limitations intrinsèques des procédés informatiques utilisés pour la conversion OCR, seuls un sous-ensemble des documents patrimoniaux sont éligibles à ce traitement (voir section 3.3).

La commande d'une conversion en mode texte pour un document donné a généralement lieu lors de l'état conjoint, pendant lequel les représentants de la BnF et le prestataire précisent la commande en fonction de la nature du document.

3.2 Format ALTO

3.2.1 Introduction

ALTO (*Analysed Layout and Text Object*) est un standard XML permettant de rendre compte de la mise en page physique et de la structure logique d'un texte reconnu par un système OCR. Ce format est issu du projet européen METAe1 et il est actuellement maintenu par un comité éditorial hébergé par la Bibliothèque du Congrès (http://www.loc.gov/standards/alto/).

ALTO est très utilisé pour la conversion en mode texte de documents patrimoniaux, en France et à l'étranger. Il est bien adapté à la conservation à long terme des données issues de la conversion et il permet une réutilisation ultérieure du mode texte, dans la mesure où il contient pour chaque mot et bloc de texte ses les coordonnées dans la page, le taux de confiance de reconnaissance, éventuellement des éléments de forme (styles de caractère, polices).

La Bibliothèque nationale de France a introduit des évolutions dans le format ALTO originel, et utilise désormais une variante pour sa production

(http://bibnum.bnf.fr/alto_prod). La nature de ces évolutions est documentée et elle peut être communiquée au prestataire.

Pourquoi le format ALTO et non le PDF ?

La BnF a préféré le format ALTO au format PDF comme support de l'action de dématérialisation en mode texte des documents patrimoniaux car il s'agit :

- d'un format XML, avec les qualités intrinsèques du monde XML : universalité, facilité de création, d'édition et d'archivage, compacité, etc.
- d'un format conçu expressément pour l'usage attendu (numérisation patrimoniale en modes image et texte) alors que le PDF répond à l'origine aux besoins liés à l'impression de documents numériques. Il est possible de générer le PDF à partir des images et des fichiers ALTO alors que l'inverse n'est pas possible.

3.2.2 Présentation

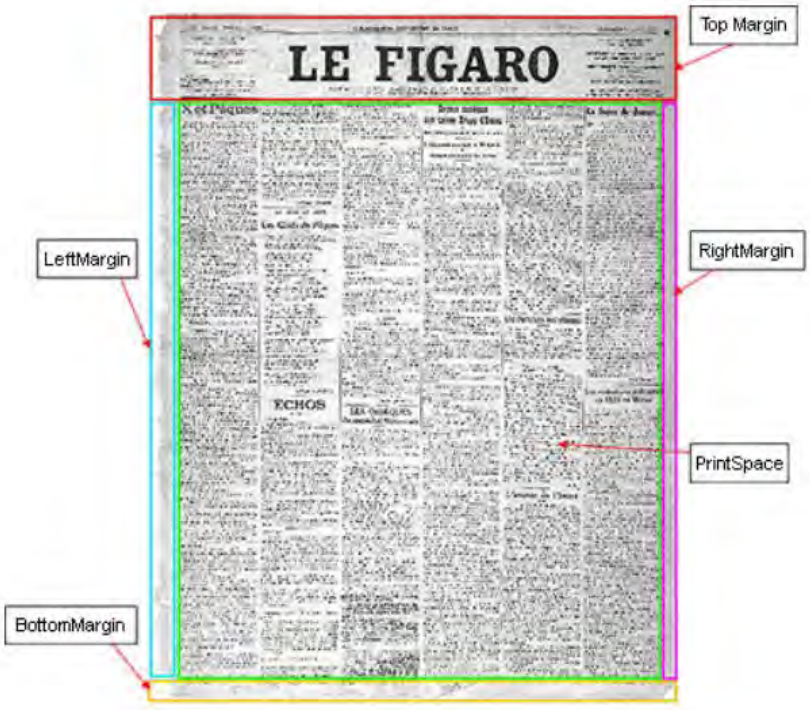
Éléments et sous-éléments composant le format ALTO

ALTO permet la segmentation d'une page en différents éléments composés de sous-éléments.

L'élément page peut contenir cinq éléments :

- TopMargin : désigne la zone supérieure de la page du bord gauche au bord droit hors zone de texte. Quand c'est possible, il s'agit de la zone contenant le titre, l'ours, etc.
- BottomMargin : désigne la zone inférieure de la page du bord gauche au bord droit hors zone de texte.
- LeftMargin : désigne la zone gauche de la page hors zone supérieure, zone inférieure et zone de texte.
- RightMargin : désigne la zone droite de la page hors zone supérieure, zone inférieure et zone de texte.
- PrintSpace : désigne la zone de texte. Cet élément est obligatoire. Il contient au moins un élément de type bloc.

EXEMPLE



Exemple de découpage d'une page de presse

Dès que l'un de ces éléments contient une information (texte, illustration etc.), cette information est décrite dans un ou plusieurs éléments de type bloc.

Les blocs peuvent être de quatre types différent :

- TextBlock : désigne le bloc de texte. Cet élément est utilisé pour regrouper les lignes de textes en un ensemble cohérent.
- Illustration : désigne une image ou une figure.
- GraphicalElement : désigne un élément graphique autre qu'une image ou une figure. Il peut être utilisé pour décrire un élément de séparation intertextuel ou un élément textuel non reconnu en tant que tel par l'OCR.
- ComposedBlock : est utilisé pour permettre l'imbrication d'éléments bloc.

A l'intérieur d'un TextBlock, l'élément Line décrit les lignes de texte et l'élément String rassemble les caractères en mots.

Les coordonnées des éléments sont définies à partir du point de repère le plus en haut à gauche de la page. Ainsi, chaque bloc, ligne ou chaîne de caractères reconnus est identifié dans l'ordre de présentation de l'original.

ALTO permet également de décrire des formes géométriques (cercle, polygone, ellipse), de gérer les césures... Les objets non textuels ont également leurs propres découpage et coordonnées.

Identification des mots et chaînes de caractères

Chaque chaîne de caractères composant un mot ou une partie de mot césuré (String) est identifiée avec les informations suivantes :

- content : le mot reconnu par le système OCR et/ou corrigé manuellement selon le niveau de qualité demandé ;
- wc (*word confidence*) : note de confiance de la reconnaissance de chaque mot, notée de 0 à 1 ;
- cc (*character confidence*) : note de confiance de la reconnaissance de chaque caractère du mot. Cette note est composée d'une liste de notes de 0 à 9, une note pour chaque caractère ;
- wd : appartenance du mot ou non à un dictionnaire.

3.3 Niveaux de qualité

Les niveaux de qualité attendus concernent tant la qualité de la segmentation que la qualité de la reconnaissance OCR.

3.3.1 Segmentation et structuration

La qualité de la segmentation et de la structuration des contenus du document d'origine lors de leur transcription au format ALTO concerne en particulier :

- l'ordre de lecture,
- le typage des blocs,
- le chevauchement des blocs.

La qualité de la segmentation et de la structuration est détaillée aux sections 7.1 et 7.2.1.

3.3.2 Reconnaissance OCR

Le résultat de la reconnaissance OCR est très variable, en fonction de la nature bibliographique et physique des documents océrisés :

- lisibilité du document : les défauts d'impression, le vieillissement du papier, les problèmes de migration d'encre ou de courbure de page, etc., influent sur la qualité de l'image numérisée et donc sur l'aptitude des systèmes OCR à en extraire le texte ;
- critères bibliographiques : la qualité de reconnaissance des systèmes OCR est particulièrement sensible à la nature des contenus, en termes de langue et d'alphabet notamment ;
- genres documentaires : les systèmes OCR s'appuient sur des dictionnaires pour affiner leur reconnaissance des mots. Idéalement, il faudrait donc disposer de dictionnaires contemporains de la date d'édition et adaptés au contenu de chaque ouvrage (littérature, sciences, philosophie, etc.).

Différents taux sont attendus selon les marchés :

- OCR haute qualité : par exemple 99,9 %
- OCR taux garanti : par exemple 98,5 %
- OCR brut : pour les ouvrages datant de 1750 ou antérieur, ou pour les ouvrages sans date.

Ces taux sont précisés dans le contexte de chaque marché de numérisation.

Cette qualité de reconnaissance OCR est mesurée à l'aide d'un taux de confiance (WC, voir section 3.2.2), et non d'un taux effectif. En effet, pour connaître le taux de qualité effectif pour un document océrisé, il faudrait avoir connaissance de sa vérité terrain (le texte exact du document) afin de la comparer avec le texte produit par le logiciel OCR, ce qui est bien sûr impossible dans le cadre d'une numérisation de masse. Ce taux de confiance est fourni par les logiciels d'OCR, pour chaque mot traité par le logiciel.

Les taux de qualité OCR sont calculés sur la base du mot, en moyennant les taux de confiance des mots présents dans un document.

Le taux de qualité d'un document n'est pas obtenu en moyennant les taux de qualité de chaque page, mais bien en moyennant les taux de confiance de tous les mots du document.

Un document peut avoir des parties dont le taux de reconnaissance est supérieur ou inférieur au taux qualité admissible, mais la moyenne doit correspondre à la qualité exigée.

La qualité de la reconnaissance OCR est détaillée à la section 6.1.

Le contrôle de la qualité de la reconnaissance OCR est détaillé à la section 7.2.2.

4. OCR

Ce chapitre présente des informations sur l'encodage du texte et le traitement des différentes polices et langues ainsi que sur la gestion des césures

4.1 Langues et encodage de caractères

4.1.1 Détection de la langue

Le moteur OCR permet de définir la langue et les caractères pouvant faire partie du contenu (soit à l'ouvrage, soit à la page, soit au bloc texte). Ceci permet d'associer le dictionnaire de la langue du texte et de déterminer la fiabilité de la reconnaissance.

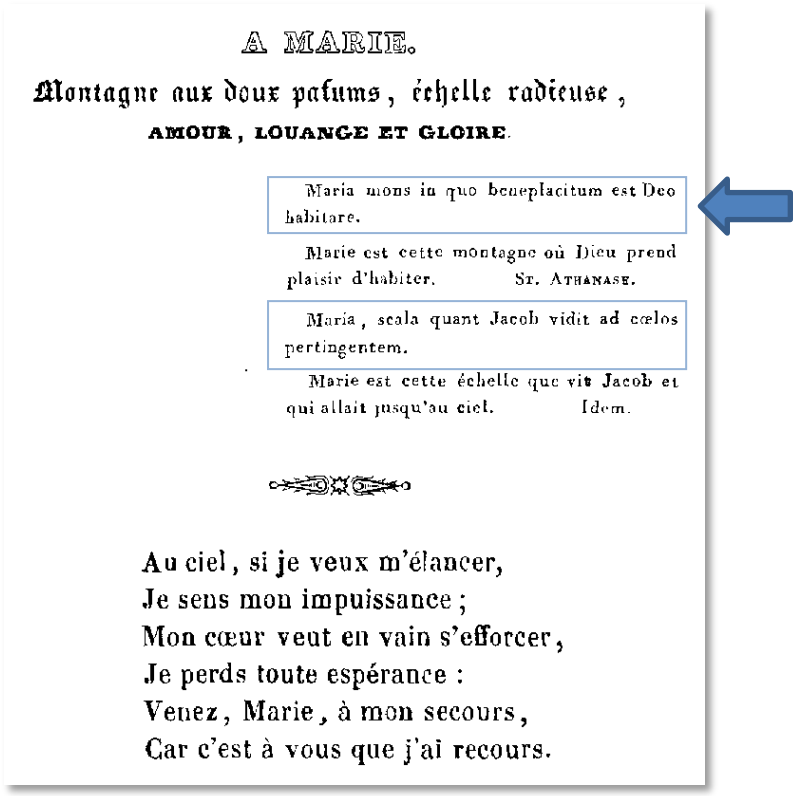
Tous les ouvrages sont traités par défaut comme étant en langue française, sauf détection par le moteur OCR d'autres langues.

La langue de chaque bloc de texte sera consignée dans l'attribut language du schéma ALTO, selon la norme ISO 639-2 :

```
language="fr"
```

NOTES

Si deux blocs de langues différentes se suivent, leur langue sera détectée et consignée.



A MARIE.

Montagne aux doux parfums, échelle radieuse,
AMOUR, LOUANGE ET GLOIRE.

Maria mons in quo beneplacitum est Deo habitare.

Marie est cette montagne où Dieu prend plaisir d'habiter. St. Athanase.

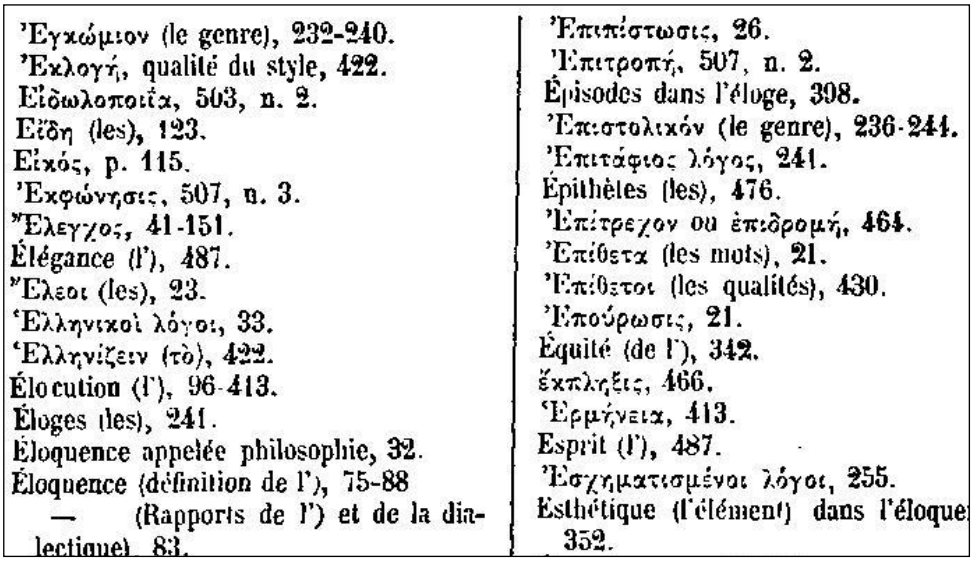
Maria, scala quant Jacob vidit ad cœlos pertingentem.

Marie est cette échelle que vit Jacob et qui allait jusqu'au ciel. Idem.

Au ciel, si je veux m'élancer,
Je sens mon impuissance ;
Mon cœur veut en vain s'efforcer,
Je perds toute espérance :
Venez, Marie, à mon secours,
Car c'est à vous que j'ai recours.

Texte en latin dans une page majoritairement en français

Si deux langues sont présentes dans le même bloc, la langue majoritaire sera consignée au niveau du bloc et la langue minoritaire au niveau des mots.



Blocs en français, en grec et mixtes

4.1.2 Traitement OCR

Les langues romanes et les langues écrites avec l'alphabet latin (y compris le latin) seront traitées en OCR, ainsi que le grec.

Pour le français, le prestataire décrira son aptitude à mettre en œuvre des dictionnaires d'orthographe adapté à la date de publication des documents : ancien français, moyen français, français classique, français moderne. Il exposera en outre sa capacité à traiter les formes anciennes de la lettre *s* minuscule : *s* long, (ſ), eszett (ß).

Pour le traitement des langues suivantes, le prestataire décrira son aptitude à mettre en œuvre des dictionnaires d'orthographe (ciblant la langue dans sa forme moderne) : allemand, anglais, italien, espagnol, grec, grec ancien, latin, portugais. Le taux qualité attendu pour ces langues sera l'OCR brut.

Les langues écrites avec l'alphabet grec (autre que le grec) et les langues non romanes écrites avec l'alphabet latin (albanais, croate, estonien, finnois, hongrois, lithuanien, roumain, slovaque, slovène, norvégien, polonais, tchèque, turc moderne, vietnamien en écriture latinisé, etc.), seront également traités en OCR, mais le taux qualité attendu sera l'OCR brut, sauf demande particulière de la BnF et selon les propositions techniques du prestataire.

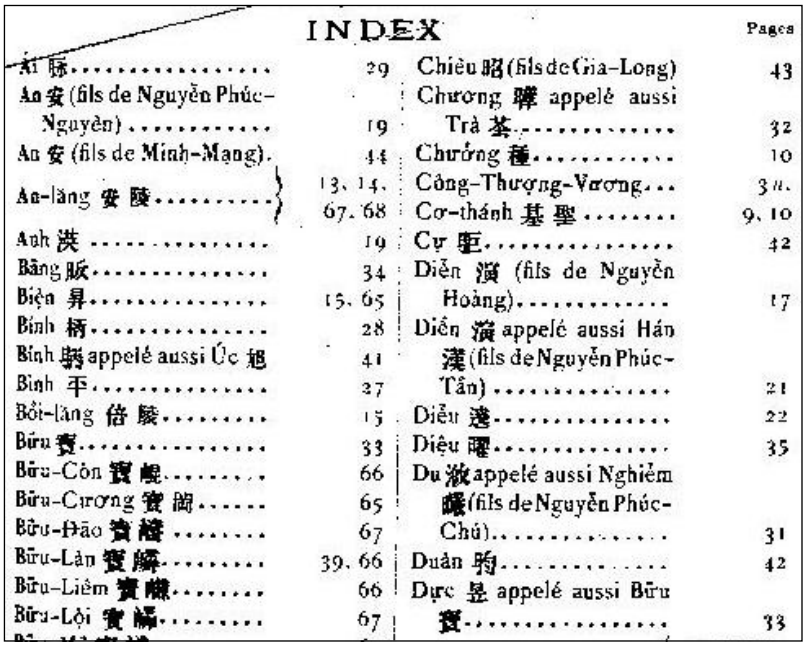
Les autres alphabets et types d'écriture (cyrillique, langues asiatiques, arabe, hébreu, etc.) ne seront pas traités en OCR. Les blocs de texte où ces alphabets ou systèmes d'idéogrammes sont présents seront décrits sous la forme de blocs image. Ces blocs seront des éléments Illustration doté d'un attribut TYPE="nonLatinScript".

EXEMPLE

Bloc de texte à traiter en mode image

Les blocs de texte où sont présents des langues en alphabet latin et non latin (cas des dictionnaires de langues par exemple) seront traités en OCR afin de transcrire les mots en alphabet latin ; la langue des mots en alphabet non latin sera consignée au niveau des mots.

EXEMPLE



Blocs de texte mixte (français, vietnamien latinisé et écriture en sinogrammes)

4.1.3 Taux qualité

Français

Tous les blocs de texte en langue française sont comptabilisés dans l'évaluation du taux OCR qualité garantie du document.

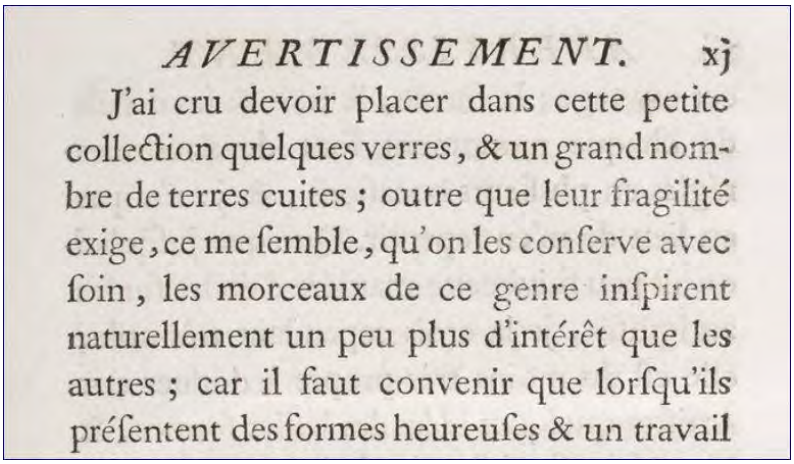
Cependant, les chiffres ainsi que les caractères non alphanumériques mis en évidence dans le tableau suivant ne sont pas comptabilisés dans l'évaluation du taux OCR qualité garantie.

	x0	x1	x2	x3	x4	x5	x6	x7	x8	x9	xA	xB	xC	xD	xE	xF
2x		!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3x											:	;	<	=	>	?
4x	@															
5x												[	\	]	^	_
7x												{	|	}	~	
Ax		¡	¢	£	¤	¥	¦	§	¨	©		«	¬		®	¯
Bx	°	±			´	µ	¶	·	¸			»				¿
Fx								÷								

Typographie ancienne

Les ouvrages composés avec des caractères d'imprimerie qui ne sont plus (ou peu) présents dans les imprimés contemporains relèvent de la catégorie « typographie ancienne » : *s* long (ſ), eszett (ß), ligatures, etc.

EXEMPLE

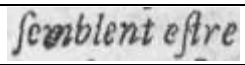
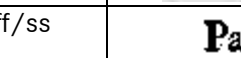


AVERTISSEMENT. xj

J'ai cru devoir placer dans cette petite collection quelques verres, & un grand nombre de terres cuites ; outre que leur fragilité exige, ce me ſemble, qu'on les conſerve avec ſoin, les morceaux de ce genre inſpirent naturellement un peu plus d'intérêt que les autres ; car il faut convenir que lorſqu'ils préſentent des formes heureuſes & un travail

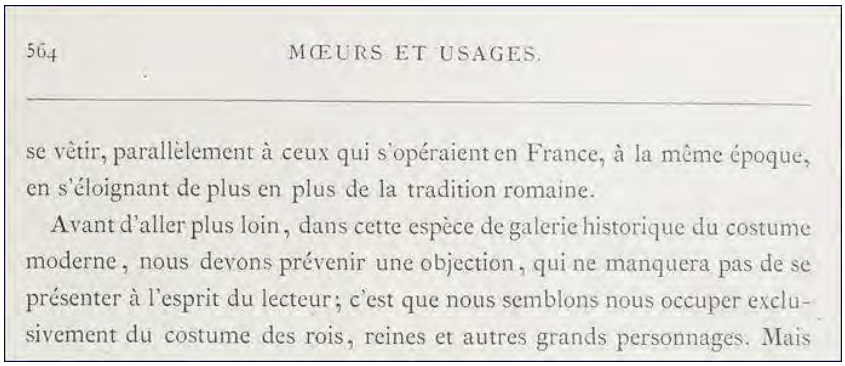
http://gallica.bnf.fr/ark:/12148/btv1b8626613n

Une translittération est exigée pour certains caractères, notamment les formes anciennes de la lettre *s* minuscule : *s* long, (ſ), eszett (ß). Cette translittération fait partie du périmètre du taux OCR qualité garantie.

Glyphe	Exemple	Translittération
f/s		semblent estre
ff/ss		Paliſſades
v/u		DUEL une Femme

Les autres ouvrages appartiennent à la catégorie « typographie moderne ».

EXEMPLE



564 MŒURS ET USAGES.

se vêtir, parallèlement à ceux qui s'opéraient en France, à la même époque, en s'éloignant de plus en plus de la tradition romaine.

Avant d'aller plus loin, dans cette espèce de galerie historique du costume moderne, nous devons prévenir une objection, qui ne manquera pas de se présenter à l'esprit du lecteur ; c'est que nous semblons nous occuper exclusivement du costume des rois, reines et autres grands personnages. Mais

http://gallica.bnf.fr/ark:/12148/bpt6k6547544k

Autres langues romanes et langues écrites avec les alphabets grec et latin

Comme il a été dit section 4.1.2, ces contenus sont à traiter en OCR brut.

Mélange de langues et/ou d'alphabets

Seuls les contenus en français présents dans des blocs mixtes (mélangeant langues et/ou alphabets) seront pris en compte dans l'évaluation du taux OCR qualité garantie.

4.1.4 Encodage

La transcription du texte dans les fichiers ALTO se fera avec l'encodage Unicode UTF-8 restreint à l'ensemble des blocs de caractères Unicode nécessaires à la transcription des langues romanes plus les blocs nécessaires à la transcription des caractères grecs.

4.1.5 Signes typographiques, caractères et symboles spéciaux, etc.

Signes typographiques

Les tirets d'incise, d'énumération et de liste (en général un tiret demi-cadratin) sont à produire dans le flot de texte avec le caractère Unicode –.

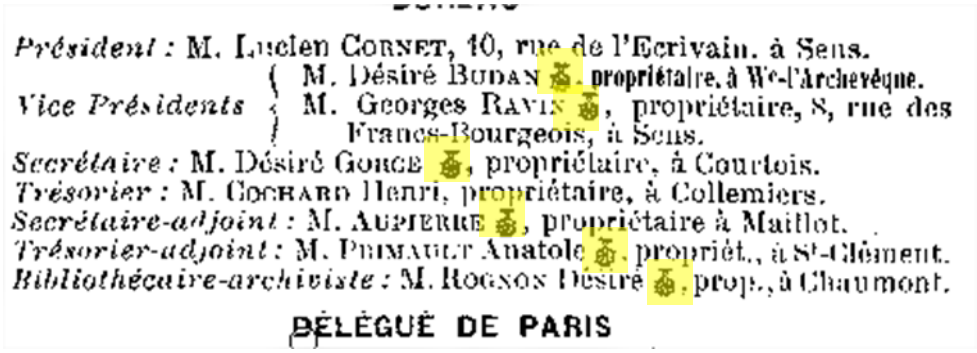
Les tirets de dialogue, les tirets longs (en général un tiret cadratin) sont à produire dans le flot de texte avec le caractère Unicode —.

Les puces de forme standard (•) sont à produire dans le flot de texte avec le code •.

Caractères spéciaux

Les caractères spéciaux sont traités en OCR et intégrés dans le flux texte sans traitement particulier. Ils peuvent entrainer la génération de mots illisibles.

EXEMPLE



Président : M. Lucien CORNET, 10, rue de l'Ecrivain, à Sens.
Vice Présidents { M. Désiré BUDAN, propriétaire, à Wé-l'Archevêque.
M. Georges RAVIX, propriétaire, 8, rue des Francs-Bourgeois, à Sens.
Secrétaire : M. Désiré GORGE, propriétaire, à Courtois.
Trésorier : M. COCHARD Henri, propriétaire, à Collemiers.
Secrétaire-adjoint : M. AUPIERRE, propriétaire à Maillot.
Trésorier-adjoint : M. PRIMAULT Anatole, propriét., à St-Clément.
Bibliothécaire-archiviste : M. ROGNON Désiré, prop., à Chaumont.
DÉLÉGUÉ DE PARIS

NOTES

Comme mentionné section 4.1.3, ces caractères ne sont pas comptabilisés dans l'évaluation du taux OCR qualité garantie.

Points de suite

Les points de suite, notamment dans les tables des matières et les index, peuvent être absents du flux de texte produit.

4.2 Styles typographiques

Par page, la liste des styles de paragraphes et des styles de caractères utilisés est donnée dans l'en-tête du fichier ALTO.

Au niveau de chaque bloc de texte, seul le style majoritaire est indiqué.

Au niveau de chaque ligne, seul le style majoritaire est indiqué.

Au niveau de chaque mot, le style utilisé est indiqué. Un seul style est possible par mot.

4.2.1 Niveau de qualité

Les styles sont obtenus par un traitement purement automatique lors de l'OCR brut, de la même manière que l'est la reconnaissance de chaque caractère.

Alors que les exigences de validité et de synchronisation des balises de styles sont respectées scrupuleusement (taux qualité garantie), l'exactitude des styles (police, taille, enrichissements) en regard du document d'origine ne peut pas l'être. Ainsi, la reconnaissance des styles n'est pas soumise au taux qualité garantie.

4.2.2 Rubriques et titres d'articles

Dans le cas où une détection de titres est demandée, le niveau de titre est stocké dans l'attribut TYPE du TextBlock ou du TextLine concerné (cf. section 5.6.2).

Les niveaux de titres sont codés ainsi :

- titre de niveau 1 (partie, chapitre, etc.) : titre1
- titre de niveau 2 (chapitre, section) : titre2
- titre de niveau3 (section, sous-section) : titre3
- etc.

4.2.3 Typographies mal reconnues par l'OCR

Différentes typographies et écritures sont mal reconnues par l'OCR :

- écriture manuscrite,
- police fantaisie, script, avec relief, double traits, etc.
- police italique avec un fort degré d'inclinaison (supérieur à 30°),
- police de type Fraktur, gothique, qui requiert des licences OCR spécifiques,
- alphabet non latin (russe, asiatique, etc.).

NOTES

Ces portions de texte mal reconnues par l'OCR seront considérées comme étant « illisibles » (*illegibles*) et seront traitées en OCR brut. Ce processus est détaillé aux sections 5.6.10 et 6.2.2.

Ecriture manuscrite

Ce cas est décrit aux sections 5.7.5 et 5.8.5.

Polices script

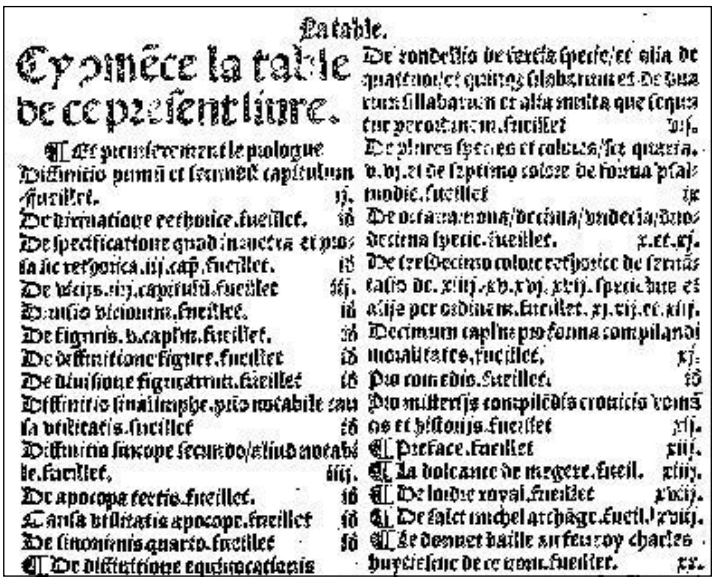
Les blocs de texte composés avec une police script sont décrits avec un élément TextBlock doté d'un attribut TYPE="scriptFonts".

EXEMPLE

Autre cas de polices non exploitables

Ces blocs de texte sont décrits avec un élément TextBlock doté d'un attribut TYPE="illegible".

EXEMPLE



La table.

Cy mẽce la table de ce present liure.

¶ Et premierement le prologue
Diffinitio primi et secundi capitulum
ffueillet. ij.
De diuisione rethorice.fueillet. iij
De specificatione quadripartite et pro-
pria ipsi rethorice.iiij.cap̃.fueillet. iij
De vsu.v.capitulis.fueillet iiij.
Diuisio vicesimi.fueillet. iij
De signis.v.capm̃.fueillet. iij
De diffinitione figure.fueillet iij
De diuisione figurarum.fueillet iij
Diffinitio similiuphe.pro notabile cau
sa utilitatis.fueillet iij
Diffinitio syncope secundo/aliud notabi
le.fueillet. iiij.
De apocope tertio.fueillet. iij
Causa utilitatis apocope.fueillet iij
De simonimis quarto.fueillet iij
¶ De diffinitione equiuocationis
De rondellio de tercia specie/et alia de
quarta/et quinq; filabarum et De bua
rum syllabarum et alia multa que sequ
tur per ordinem.fueillet vij.
De plures species et colores/sic quarta.
v.vj.et de septimo colore de forma pfal
modie.fueillet ix
De octava/nona/decima/undecia/duo
decima specie.fueillet. x.et.xj.
De tredecimo colore rethorice de senten
cia de. xiij.xiv.xv.xvij.specie huc et
alijs per ordinem.fueillet. xj.cij.et.xiij.
Decimum capm̃ pro forma compilandi
moralitates,fueillet, xj.
Pro comedio.fueillet. iij
Pro materijs compilãdis cronicis rom̃
os et hystorijs.fueillet xij.
¶ Preface.fueillet xiij.
¶ La volcanne de mezgere.fueil. xiij.
¶ De laodre royal.fueillet xvij.
¶ De falet michel archãge.fueil.xvij.
¶ Le dosnet bataille au feu roy charles
huytiesme de ce nom.fueillet. xx.

Alphabets non latins

Ce cas est décrit à la section 4.1.2.

4.3 Césures

La partie reconnue est dans l'attribut content, complétée par les attributs :

- subs_type : précise la partie concernée, HypPart1 pour la première partie du mot césuré, HypPart2 pour la seconde ;
- subs_content : restitue le mot complet non césuré.

L'élément HYP représente le tiret de césure.

EXEMPLE

…
<String ID="PAG_00000001_ST000100" STYLEREFS="TXT_1" HPOS="1285" VPOS="1910" HEIGHT="47" WIDTH="76" WC="0.99" **CONTENT="va-" SUBS_TYPE="HypPart1" SUBS_CONTENT="variées,"**/>
<HYP HPOS="1361" VPOS="1957" WIDTH="30" CONTENT="-"/>
</TextLine>

<TextLine ID="PAG_00000001_TL000028" STYLEREFS="TXT_1" HPOS="336" VPOS="1955" HEIGHT="66" WIDTH="1024">
<String ID="PAG_00000001_ST000101" STYLEREFS="TXT_1" HPOS="336" VPOS="1984" HEIGHT="37" WIDTH="127" WC="0.99" **CONTENT="riées," SUBS_TYPE="HypPart2" SUBS_CONTENT="variées,"**/>

4.3.1 Répartition entre TextBlock

Le prestataire veillera à éviter de répartir un mot césuré sur deux TextBlock (sauf dans le cas où la césure intervient sur un mot à cheval sur deux colonnes ou deux pages).

SI UN GROUPE DE MOTS COMPORTANT DES TRAITS D'UNION (COMME DANS « A-T-IL », « AVONS-NOUS ») SE TROUVE SUR DEUX LIGNES, IL PEUT ARRIVER QU'IL SOIT TRAITE COMME UNE CESURE. CE CAS N'EST PAS SOUMIS AU TAUX QUALITE GARANTIE.

5. SEGMENTATION ET STRUCTURATION

Ce chapitre décrit de quelle manière la structure du document sera représentée dans le fichier ALTO produit en sortie.

Le taux qualité spécifique au marché s'applique à cette segmentation, indépendamment du taux qualité OCR (cf. section 6).

SAUF MENTION CONTRAIRE, TOUS LES ELEMENTS DECRITS DANS CE CHAPITRE SONT SOUMIS AU TAUX QUALITE GARANTIE.

5.1 Description des pages

L'attribut pageclass de l'élément Page permet de décrire certains cas particuliers.

5.1.1 Pages vides

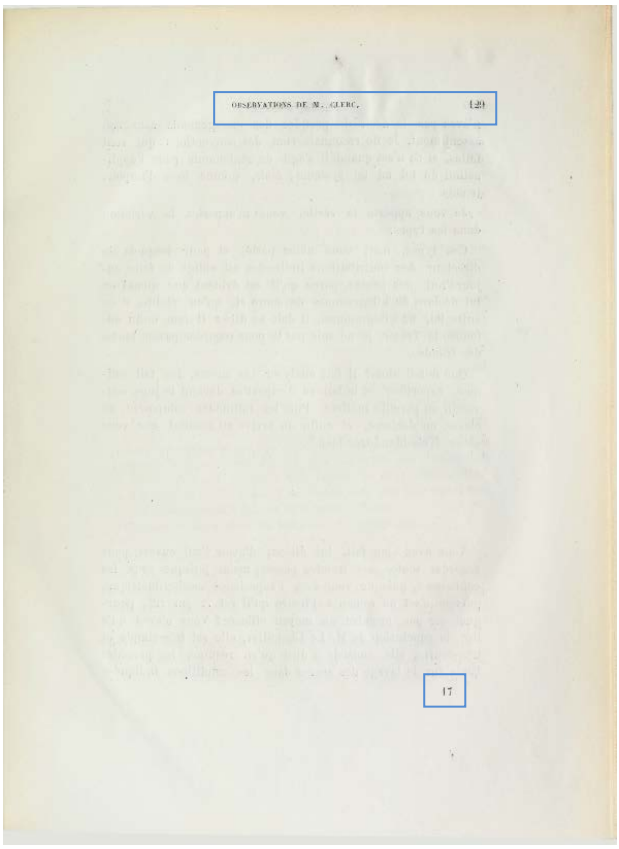
Les pages vides (sans aucun contenu) sont décrites avec pageclass="BlankPage".

```
<Layout>
  <Page ID="PAG_00000002" HEIGHT="3353" WIDTH="2065"
      PHYSICAL_IMG_NR="1" QUALITY="OK" PAGECLASS="BlankPage">
  </Page>
 </Layout>
```

5.1.2 Pages avec contenu en marges mais sans contenu principal

Si une page présente un contenu à placer en XxxMargin, mais que le reste de la page est vide, il faut utiliser la description pageclass="BlankPrintSpace".

EXEMPLE



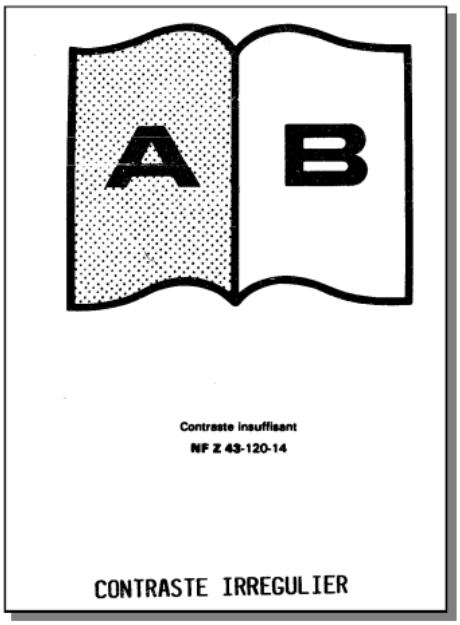
```
<Layout>
   <Page ID="PAG_00000135" pageclass="BlankPrintSpace" HEIGHT="4153"
      WIDTH="3049" PHYSICAL_IMG_NR="1" QUALITY="OK" >

  <TopMargin ID="PAG_00000135_TopMargin" HPOS="0" VPOS="0" HEIGHT="562"
WIDTH="3049">
   <TextBlock ID="PAG_00000135_TB000001" STYLEREFS="TXT_1" HPOS="1124"
VPOS="517" HEIGHT="44" WIDTH="1224" language="fr">
     <TextLine … </TextLine>
   </TextBlock>
   </TopMargin>

   <BottomMargin ID="PAG_00000135_BottomMargin" HPOS="0" VPOS="3363"
HEIGHT="790" WIDTH="3049">
   <TextBlock ID="PAG_00000135_TB000002" STYLEREFS="TXT_1" HPOS="2142"
VPOS="3363" HEIGHT="40" WIDTH="46" language="fr">
     <TextLine … </TextLine>
   </TextBlock>
   </BottomMargin>
   </Page>
  </Layout>
```

5.1.3 Pages de logo

Les pages ne contenant qu'un logo (page de type « L » dans le manifeste numérique refNum) sont traitées comme des pages blanches.

Les éventuelles légendes textuelles en plus de la représentation graphique à proprement parler ne seront pas traitées.

EXEMPLE



Logo + légende + n° NF du logo + légende

5.2 Orientation de la page

Les images numérisées seront toujours fournies dans le sens de l'original.

L'orientation de la lecture se détermine sur la totalité de la page et non pas seulement sur la partie textuelle.

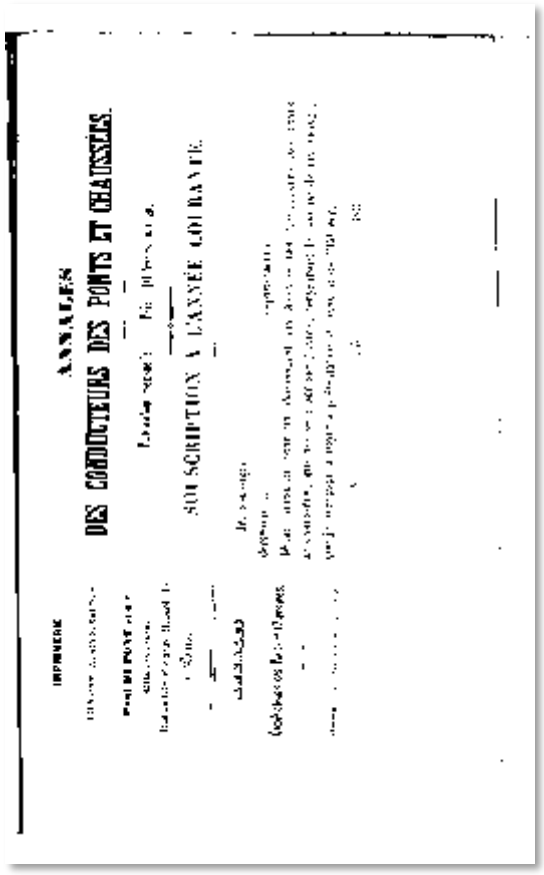
Si l'orientation de l'original (c'est-à-dire celle de la reliure) est différente de l'orientation de la lecture, les éléments de la page doivent être tournés dans le fichier ALTO à l'aide de l'attribut Rotation, selon un angle exprimé en degrés, dans le sens inverse des aiguilles d'une montre (sens anti-trigonométrique).

Deux cas peuvent se présenter :

- Si toutes les zones sont dans une orientation de lecture donnée, alors en restitution, ces zones comporteront un attribut Rotation. Dans l'exemple suivant, les TextBlock sont produits avec un attribut rotation à 270° (mesuré dans le sens anti-trigonométrique) :

```
<TextBlock ID="PAG_248_TB000001" STYLEREFS="TXT_1" HPOS="1996"
VPOS="1019" HEIGHT="1557"  WIDTH="66" TYPE="table" language="fr"
ROTATION="270">
```

EXEMPLE



Sens de la page originale

- Si la page comprend des zones ayant plusieurs orientations de lecture, on détermine l'orientation de lecture majoritaire selon l'ordre de priorité suivant :
  - L'entête et le numéro de page ne sont pas pris en compte (ils sont structurés en TopMargin).
  - Les éléments GraphicalElement ne sont pas pris en compte.
  - Pour déterminer le sens de lecture, le « corps du texte » prime sur le reste (les illustrations, les tableaux avec leurs éventuels contenus textuels et les éléments graphiques, etc.) s'il occupe plus d'un tiers de la page.
  - De la même manière, le « corps du texte » prime sur les légendes des illustrations et des tableaux, même quand ceux-ci sont en dehors de l'objet Illustration.

  Une fois l'orientation de lecture majoritaire déterminée, on applique l'attribut Rotation aux zones dont l'orientation de lecture est différente de l'orientation de lecture majoritaire.

EXEMPLE

12 MAGASIN PITTORESQUE.

tres. Ainsi, par exemple, il n'est pas rare de voir des cavernes parcourues par des cours d'eau réaliser sous nos yeux la supposition que nous avons faite pour les grottes d'Arcy. La Laibach, en Carinthie, s'engouffre dans la grotte d'Adelsberg, puis reparaît, pour disparaître de nouveau et se perdre enfin dans la caverne de Reifnitz, près de la ville de Laibach. Aux portes de Trieste il existe un cours d'eau souterrain que l'on a cherché à utiliser pour la ville. Dans le département du Jura, la Loisance sort des grottes de Plancher-sur-Arbois; la Seille, de celle de Baume-les-Messieurs. Dans celui de l'Isère, la Sassenage s'échappe des grottes du même nom, et la grotte de Balme est parcourue par un ruisseau. On ne peut pénétrer qu'en bateau dans la caverne de Frédéric, en Wurtemberg; et dans celle de Dunohl (Lancashire), en Angleterre, une cascade tombe du plafond et en forme d'autres avant de sortir de la grotte.

On le voit, les cavernes en général, et celles d'Arcy en particulier, sont des canaux souterrains qui ne sont plus parcourus par les eaux qui les ont creusés; et il serait facile de montrer qu'on trouve tous les passages, toutes les nuances entre une simple cavité creusée par une rivière dans les roches qui la bordent, et les systèmes de grottes et de cavernes les plus compliqués. L'action est la même; elle est lente, insensible, mais tous les faits géologiques sont d'accord pour nous prouver ce que peuvent les agents les plus faibles lorsque leur action se continue pendant les milliers de siècles qui correspondent aux âges géologiques de notre planète. En effet, c'est bien avant l'époque historique que les grottes d'Arcy formaient un canal souterrain à la Cure. Il est aisé de le démontrer. Le plafond et le sol sont couverts de stalactites et de stalagmites énormes qui se sont formées avec une extrême

Fig. 4.

lenteur, car elles sont l'œuvre des gouttes d'eau qui suintent de la voûte et s'évaporent en déposant la faible proportion de carbonate de chaux qu'elles tenaient en dissolution. La grosseur et la hauteur de ces stalactites dénotent donc une action prolongée pendant des centaines de siècles; or il ne se forme pas de stalactites dans un canal traversé par un cours d'eau, et, en effet, le pavé de stalagmite recouvre partout le limon et les cailloux roulés. Il faut donc se reporter bien au delà des temps historiques pour arriver à la période où les grottes d'Arcy étaient remplies par une rivière souterraine. Mais si l'on se demande à quelle époque ses eaux ont commencé à dissoudre et à désagréger lentement la pierre calcaire, l'imagination trouve encore des centaines, peut-être des milliers de siècles, entre le moment où la rivière attaquait le rocher et celui où elle remplissait les vastes cavités qu'elle a délaissées depuis.

PRIÈRE D'UNE FEMME ARABE

AU TOMBEAU DE SON ÉPOUX.

(Voy., sur les Funérailles des musulmans, la Table des dix premières années.)

Les Arabes récitent, devant les tombeaux, des prières consacrées par d'anciennes traditions; mais ils expriment aussi leurs souhaits pour les êtres qu'ils ont perdus dans des improvisations dont le caractère varie suivant leur sensibilité ou leur imagination. Rarement ils laissent éclater leur douleur; ils semblent plutôt s'étudier à la contenir: le sentiment qui domine dans ces épanchements de leur âme est une confiance absolue en la volonté divine. Voici, comme exemple, quelques passages d'une prière que l'on a entendu prononcer à une jeune femme.

« O Dieu puissant qui as créé la terre, les montagnes qui lui servent d'appui, et les sept cieux qui la couvrent; Dieu éternel qui as placé au firmament l'astre du jour et le flambeau de la nuit, qui as posé entre les deux océans d'eau douce et d'eau amère des barrières insurmontables; Dieu miséricordieux qui as créé l'homme avec l'eau, et qui, pour sa nourriture, fais couler la pluie des nuages, verdir l'herbe, germer le grain, croître la vigne et le palmier, mûrir la figue, l'olive et la grenade, prends pitié de ma douleur, ne permets pas que je blasphème. Louange à toi, Dieu unique et infini. Tu avais facilité à celui que je pleure le chemin qui conduit à la vie; tu lui avais donné une forme agréable, une taille fine, un corps délié, le recueillement de l'esprit et la sobriété de la parole. Tu lui avais donné l'ouïe et la vue, et, bien qu'il vécût au milieu des pervers, la doctrine divine ne l'a trouvé ni aveugle ni incrédule. Il a goûté la parole du prophète et les dogmes du Coran, merveilleux écrit sur la table gardée. Fidèle musulman, il n'a pas vécu avec faste au milieu de sa famille, il n'a pas transgressé le divin précepte qui défend le meurtre et l'infidélité; croyant vertueux, il n'a pas nié la résurrection et détourné ses regards de la vie future; serviteur du Miséricordieux, il suivait les inspirations de l'esprit, et résistait aux séductions d'Éblis. Prosterné le matin, le soir et durant les nuits, il récitait dévotement les versets les plus saints de l'Évidence, dont la lecture procure l'indulgence et les faveurs du Seigneur. Il a désiré des enfants qui lui inspirassent la crainte de Dieu; il a secouru ses proches; il a protégé l'orphelin, répandu l'aumône sur le voyageur et sur le pauvre; il s'est interdit les délassements défendus durant les mois sacrés; il a observé l'abstinence pendant le jeûne du Ramadan; il a visité les saints lieux; il a mérité la récompense de sa persévérance et l'accomplissement des promesses de l'Éternel.

» O Dieu, tu as fait passer le juste de la vie à la mort; que la paix soit avec lui. Rends-lui si frais et si doux le tombeau où tu lui as commandé de descendre, qu'au jour de la sépa-

L'orientation du texte prime sur celle de l'illustration : seul le bloc de l'illustration doit être tourné

5.3 Structuration de la page

5.3.1 PrintSpace

Le contenu récurrent et répétitif (titre, intitulé d'une section, sous-titre, nom d'auteur et numéro de page) doit appartenir à un XxxMargin. En cas de doute, il est possible d'inclure tout le texte d'une page dans le PrintSpace.

Le PrintSpace peut être serré autour du texte (c'est-à-dire ne pas englober l'ensemble de la page) à condition que tout le texte de la page (hors marge) soit compris dans celui-ci.

ATTENTION

LES PAGES BLANCHES DOIVENT ETRE DECRITES SELON LES REGLES EXPOSEES A LA SECTION 5.1.1.

5.3.2 XxxMargin

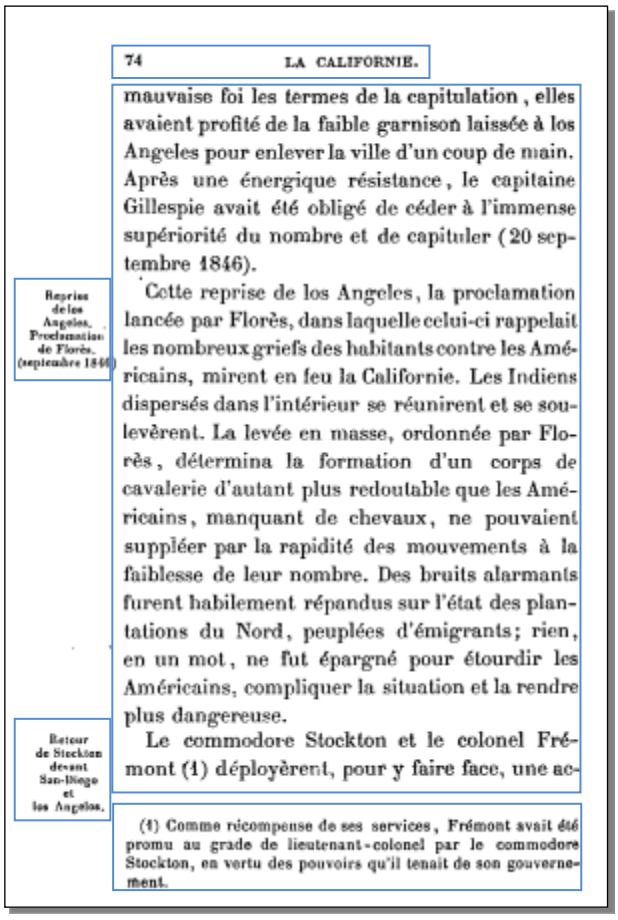
Le contenu récurrent et répétitif (titre, intitulé d'une section, sous-titre, nom d'auteur et numéro de page) doit appartenir à un xxxMargin.

Un contenu non répétitif et spécifique à la page (notes ou en-têtes de section en marge) fait partie du printSpace.

Si un XxxMargin est indiqué, il peut couvrir tout l'espace entre le bord de la page concernée et le(s) printSpace concerné(s).

ATTENTION

UNE MISE EN PAGE AVEC DES BLOCS EN MARGE A DROITE OU EN MARGE A GAUCHE NE SIGNIFIE PAS NECESSAIREMENT QUE CES BLOCS DOIVENT ETRE PLACES EN LEFTMARGIN OU EN RIGHTMARGIN.



74 LA CALIFORNIE.

mauvaise foi les termes de la capitulation, elles avaient profité de la faible garnison laissée à los Angeles pour enlever la ville d'un coup de main. Après une énergique résistance, le capitaine Gillespie avait été obligé de céder à l'immense supériorité du nombre et de capituler (20 septembre 1846).

Reprise de los Angeles. Proclamation de Florès. (septembre 1846)

Cette reprise de los Angeles, la proclamation lancée par Florès, dans laquelle celui-ci rappelait les nombreux griefs des habitants contre les Américains, mirent en feu la Californie. Les Indiens dispersés dans l'intérieur se réunirent et se soulevèrent. La levée en masse, ordonnée par Florès, détermina la formation d'un corps de cavalerie d'autant plus redoutable que les Américains, manquant de chevaux, ne pouvaient suppléer par la rapidité des mouvements à la faiblesse de leur nombre. Des bruits alarmants furent habilement répandus sur l'état des plantations du Nord, peuplées d'émigrants; rien, en un mot, ne fut épargné pour étourdir les Américains, compliquer la situation et la rendre plus dangereuse.

Retour de Stockton devant San-Diego et los Angeles.

Le commodore Stockton et le colonel Frémont (1) déployèrent, pour y faire face, une ac-

(1) Comme récompense de ses services, Frémont avait été promu au grade de lieutenant-colonel par le commodore Stockton, en vertu des pouvoirs qu'il tenait de son gouvernement.

Le titre courant (« 74 LA CALIFORNIE. ») doit constituer un topMargin.

PrintSpace : le contenu des deux notes marginales à gauche est lié au corps du texte de la page et il lui est spécifique. Il en ressort qu'ils ne font pas partie d'un leftMargin mais doivent être séparés afin que des parties de leur contenu ne soient pas mélangées avec le texte de la colonne principale.

Dans cet exemple, les notes marginales apparaitront dans le flux ALTO après les notes de bas de page.

5.4 Blocs manqués

Les blocs non identifiés par la segmentation automatique seront identifiés visuellement et décrits manuellement : position, taille, type.

Les blocs texte ainsi identifiés seront ensuite traités en OCR. Si le moteur OCR ne détecte aucun texte à l'intérieur d'un tel bloc, le bloc texte sera rempli avec une seule ligne composée de la mention "[texte manquant]", et il sera typé avec l'attribut TYPE="missingText".

L'identification visuelle des blocs manqués n'est pas soumise au taux qualité garantie.

5.5 Ordre de lecture et ordre des segments

L'ordre de lecture est également déterminé lors de la reprise de la segmentation automatique, avec application des règles qui suivent :

- Le texte se lit en colonne, de haut en bas et de gauche à droite.
- Les illustrations sont segmentées dans l'ordre où elles apparaissent, de haut en bas et de gauche à droite.
- Quand une illustration s'accompagne d'une légende, utiliser un ComposedBlock qui inclut l'image et la zone de texte.
- L'entête et le numéro de page ne sont pas pris en compte.
- Les éléments GraphicalElement ne sont pas pris en compte.

Dans le fichier ALTO, l'ordre de lecture est indiqué par la numérotation séquentielle des blocs, via leur attribut ID.

Cette numération séquentielle est propre à chaque type de bloc. Par exemple, les éléments String auront une numérotation de PAG_00000001_SP000001, PAG_00000001_SP000002 à PAG_00000001_SP00000n, les éléments TextBlock de PAG_00000001_TB000001, PAG_00000001_TB000002 à PAG_00000001_TB00000n, etc.

Dans chaque identifiant, le premier numéro est le numéro de page, et le second l'ordre de l'élément dans sa séquence.

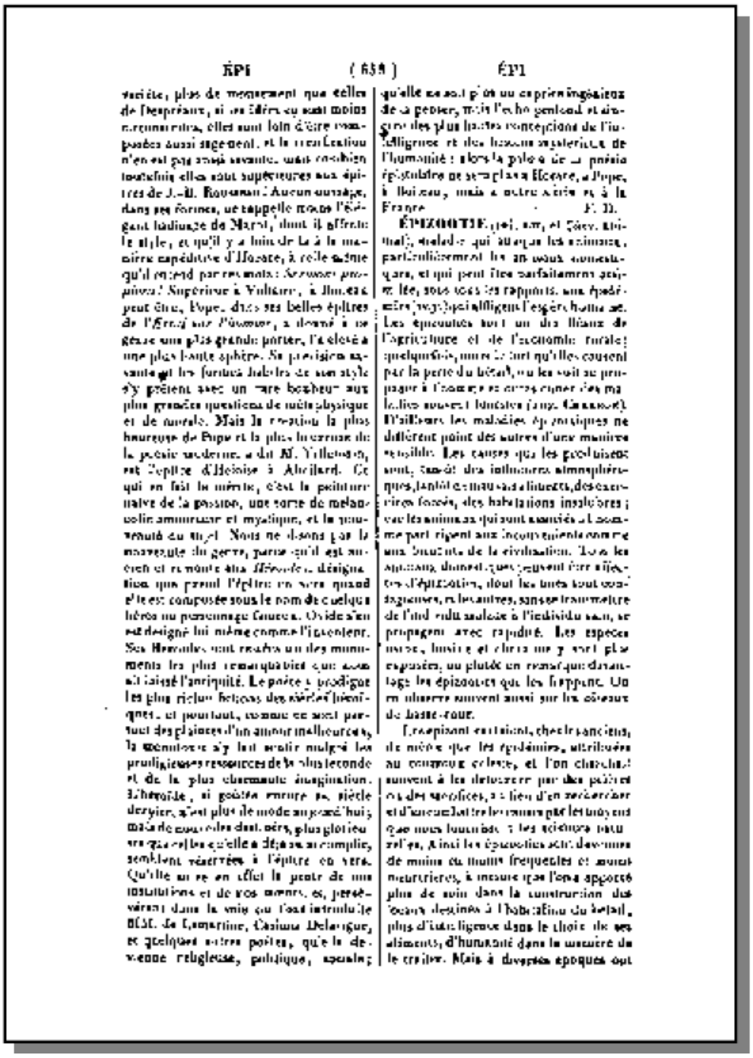
5.5.1 Mise en page en colonnes

En règle générale :

- Les entêtes des colonnes, si elles ne sont pas dans le topMargin, figurent en haut de la colonne la plus proche.
- Les bas des colonnes, si elles ne sont pas dans le bottomMargin, figurent en bas de la colonne la plus proche.
- Les colonnes sont traitées dans le sens de lecture.
- La totalité des segments de la colonne N sont traités avant un segment qui appartient à une colonne N+1.
- Les filets de gouttière ne sont pas segmentés.

Même si les entêtes de colonnes peuvent figurer dans le printSpace, il faut respecter les points suivants :

- Ne pas couper un mot.
- Indépendamment de l'ordre de lecture, ne pas « supprimer » le passage à la ligne selon la hauteur dans la page.



La position de (658) dans l'ordre de lecture est ambiguë. L'ordre de lecture peut être

ÉPI ; [col. 1] ; (658) ÉPI ; [col. 2]

ou

ÉPI (658) ; [col. 1] ; ÉPI ; [col. 2]

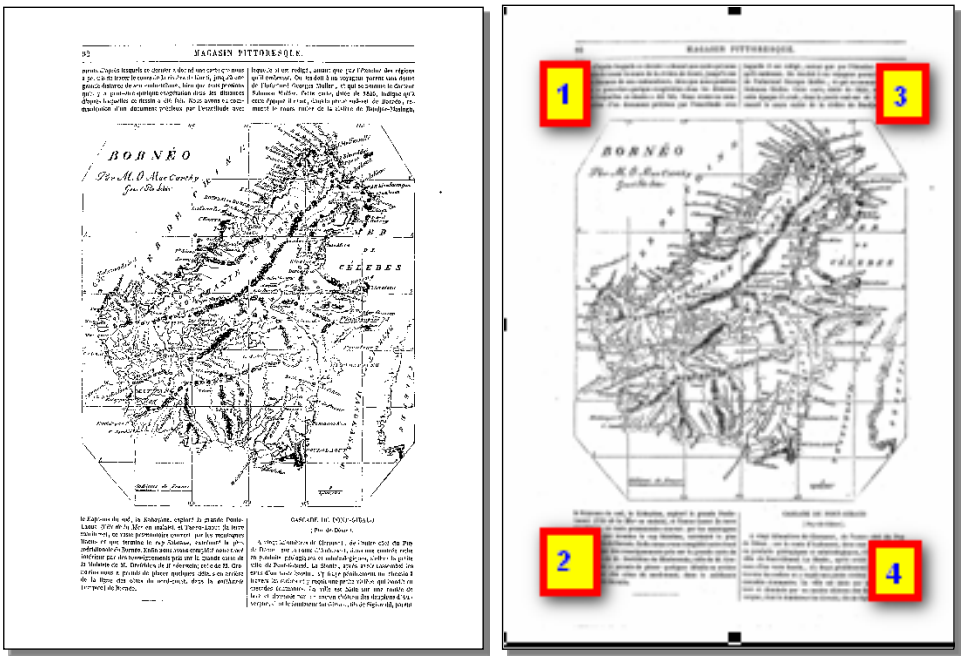
5.5.2 Mise en page en colonnes avec des éléments centraux non textuels

Un élément central non textuel est un élément non textuel (illustration, carte etc.) qui s'étale (pleinement ou partiellement) sur plusieurs colonnes.

En règle générale :

- L'ordre de lecture suit les règles générales sur les colonnes.
- Le bloc central non textuel peut être mis à n'importe que point entre les TextBlock ainsi ordonnés.

EXEMPLE



La légende de l'illustration n'est pas centralisée, elle sera incluse dans « sa » colonne

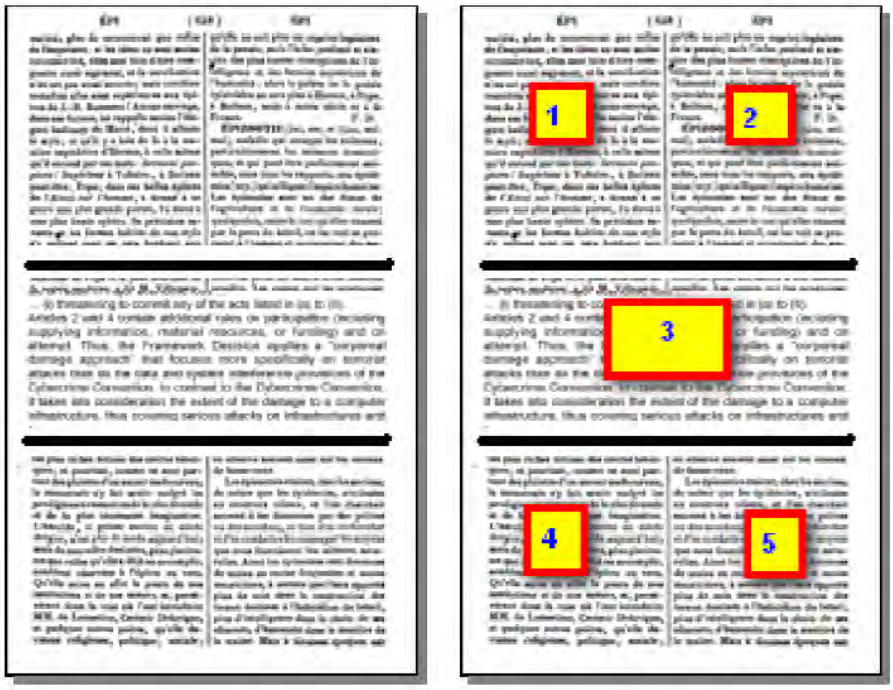
5.5.3 Mise en page en colonnes avec des éléments centraux textuels

Un élément central textuel est un élément textuel (simple texte, encadré avec texte, table, etc.) qui s'étale (pleinement ou partiellement) sur plusieurs colonnes.

En règle générale :

- En plus des règles sur les colonnes, les blocs dont la position verticale est avant un bloc textuel central, figurent avant ce bloc textuel central.

EXEMPLE

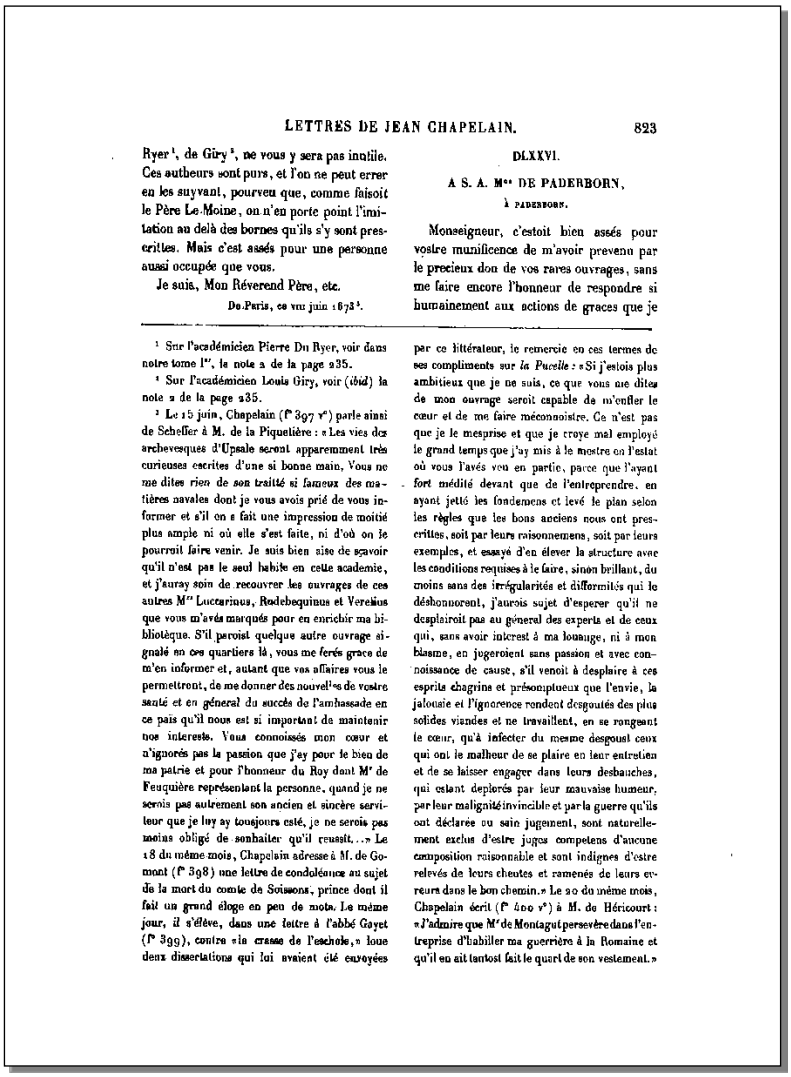


5.5.4 Corps du texte et notes séparés par un trait

La séparation entre le corps du texte et les notes sera contrôlée.

Si aucun trait séparateur ou indice de séparation n'est apparent, alors le traitement sera basé sur l'exploitation de deux colonnes « simples ».

EXEMPLE



LETTRES DE JEAN CHAPELAIN. 823

Ryer[1], de Giry[2], ne vous y sera pas inutile. Ces autheurs sont purs, et l'on ne peut errer en les suyvant, pourveu que, comme faisoit le Père Le Moine, on n'en porte point l'imitation au delà des bornes qu'ils s'y sont prescrittes. Mais c'est assés pour une personne aussi occupée que vous.

Je suis, Mon Réverend Père, etc.

De Paris, ce VIIe juin 1673[3].

DLXXVI.

A S. A. Mgr DE PADERBORN,

À PADERBORN.

Monseigneur, c'estoit bien assés pour vostre munificence de m'avoir prevenu par le precieux don de vos rares ouvrages, sans me faire encore l'honneur de respondre si humainement aux actions de graces que je

[1] Sur l'académicien Pierre Du Ryer, voir dans notre tome Ier, la note a de la page 235.

[2] Sur l'académicien Louis Giry, voir (*ibid*) la note a de la page 235.

[3] Le 15 juin, Chapelain (f° 397 v°) parle ainsi de Scheffer à M. de la Piquetière : « Les vies des archevesques d'Upsale seront apparemment très curieuses escrites d'une si bonne main. Vous ne me dites rien de son traitté si fameux des matières navales dont je vous avois prié de vous informer et s'il en a fait une impression de moitié plus ample ni où elle s'est faite, ni d'où on le pourroit faire venir. Je suis bien aise de scavoir qu'il n'est pas le seul habile en cette academie, et j'auray soin de recouvrer les ouvrages de ces autres Mrs Loccarinus, Rodebequinus et Verelius que vous m'avés marqués pour en enrichir ma bibliotèque. S'il paroist quelque autre ouvrage signalé en ces quartiers là, vous me ferés grace de m'en informer et, autant que vos affaires vous le permettront, de me donner des nouvelles de vostre santé et en général du succès de l'ambassade en ce pais qu'il nous est si important de maintenir nos interests. Vous connoissés mon cœur et n'ignorés pas la passion que j'ay pour le bien de ma patrie et pour l'honneur du Roy dont Mr de Feuquière représentant la personne, quand je ne serois pas autrement son ancien et sincère serviteur que je luy ay tousjours esté, je ne serois pas moins obligé de sonhaiter qu'il reussit…» Le 18 du même mois, Chapelain adresse à M. de Gomont (f° 398) une lettre de condoléance au sujet de la mort du comte de Soissons, prince dont il fait un grand éloge en peu de mots. Le même jour, il s'élève, dans une lettre à l'abbé Goyet (f° 399), contre « la crasse de l'eschole, » loue deux dissertations qui lui avaient été envoyées par ce littérateur, le remercie en ces termes de ses compliments sur *la Pucelle* : « Si j'estois plus ambitieux que je ne suis, ce que vous me dites de mon ouvrage seroit capable de m'enfler le cœur et de me faire méconnoistre. Ce n'est pas que je le mesprise et que je croye mal employé le grand temps que j'ay mis à le mestre en l'estat où vous l'avés veu en partie, parce que l'ayant fort médité devant que de l'entreprendre, en ayant jetté les fondemens et levé le plan selon les règles que les bons anciens nous ont prescrittes, soit par leurs raisonnemens, soit par leurs exemples, et essayé d'en élever la structure avec les conditions requises à le faire, sinon brillant, du moins sans des irrégularités et difformités qui le déshonnorent, j'aurois sujet d'esperer qu'il ne despleiroit pas au géneral des experts et de ceux qui, sans avoir interest à ma louange, ni à mon blasme, en jugeroient sans passion et avec connoissance de cause, s'il venoit à despleire à ces esprits chagrins et présomptueux que l'envie, la jalousie et l'ignorence rendent desgoutés des plus solides viandes et ne travaillent, en se rongeant le cœur, qu'à infecter du mesme desgoust ceux qui ont le malheur de se plaire en leur entretien et de se laisser engager dans leurs desbauches, qui estant deplorés par leur mauvaise humeur, par leur malignité invincible et par la guerre qu'ils ont déclarée au sain jugement, sont naturellement exclus d'estre juges competens d'aucune composition raisonnable et sont indignes d'estre relevés de leurs cheutes et ramenés de leurs erreurs dans le bon chemin.» Le 20 du même mois, Chapelain écrit (f° 400 v°) à M. de Héricourt : « J'admire que Mr de Montagut persevère dans l'entreprise d'habiller ma guerrière à la Romaine et qu'il en ait tantost fait le quart de son vestement. »

Exemple de notes de bas de page composées sur deux colonnes

5.6 Texte

5.6.1 Paragraphes

Lors de l'étape de segmentation des ouvrages, les blocs textes faisant partie du PrintSpace sont détectés sans séparation entre les paragraphes. Ils sont ensuite envoyés dans l'étape d'OCR.

Si les blocs textuels sont séparés par des éléments graphiques ou des tableaux, plusieurs blocs de texte sont segmentés.

Après l'OCR, c'est lors de l'étape de structuration (processus automatique avec validation manuelle) que le découpage du corps de texte en paragraphes a lieu et que les paragraphes sont identifiés.

La BnF attire l'attention du prestataire sur la nécessité d'un découpage en paragraphes le plus fidèle possible à l'original, afin de permettre un bon reformatage des textes lors de leur diffusion sur le Web ou sur supports nomades.

5.6.2 Rubriques et titres d'article

Selon la nature des contenus à numériser, la prestation pourra inclure une reconnaissance des titres, notamment dans le cas de documents de type presse (titres et rubriques d'article de journaux).

Dans ce cas, le prestataire et la BnF s'accorderont sur les titres à identifier, et sur les règles d'identification (par exemple avec l'attribut TYPE). Ces règles seront notamment spécifiées par type documentaire, la composition typographique des titres étant par nature variable selon les époques, les genres documentaires, etc.

Les taux qualité mesurant la réalisation de la tâche de reconnaissance des titres (taux de détection, taux de précision) feront l'objet d'un accord spécifique entre la BnF et le prestataire. Le taux qualité garantie générique du marché ne s'applique pas.

5.6.3 Tableaux

Les tableaux sont identifiés lors de l'étape de segmentation.

Sont définis comme tableaux les éléments multicolonnes qui font une rupture dans le flux de texte, ayant ou non un cadre externe et des traits de séparation entre les cellules.

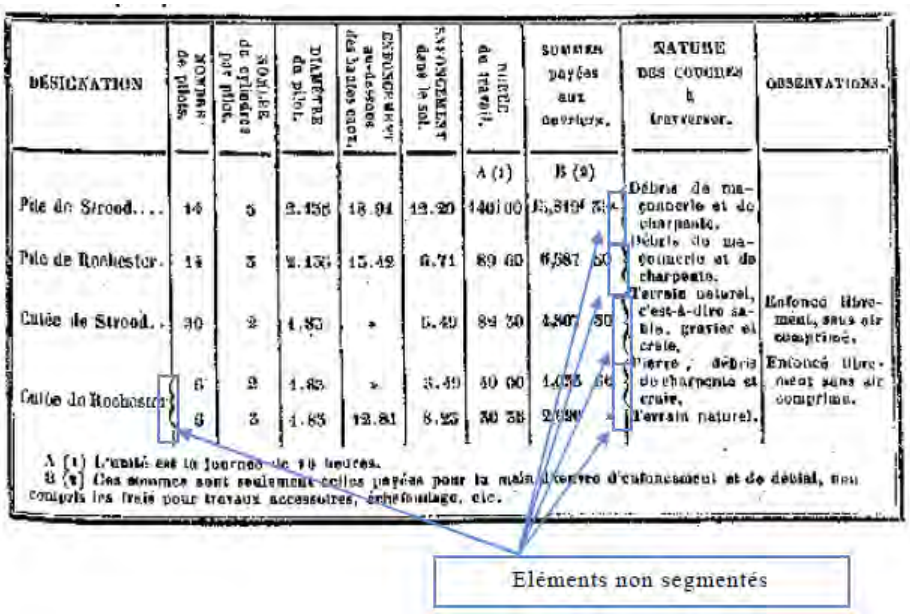
Les tableaux contenant des chaînes alphanumériques seront convertis en texte non structuré en tableau, en utilisant l'attribut TYPE de l'élément TextBlock avec la valeur "table". L'ordre de lecture à l'intérieur des différentes parties des tableaux est celui obtenu par le moteurOCR, c'est-à-dire en lignes d'abord, puis en colonnes.

Chaque bloc de texte contenu dans le tableau sera marqué comme un TextBlock avec l'attribut TYPE="table".

Il peut y avoir plusieurs blocs texte de type Table pour un même tableau physique sur une même page.

NOTES

Les éléments graphiques au sein d'un tableau qui font office de filets ou de traits de séparation ne seront pas segmentés.

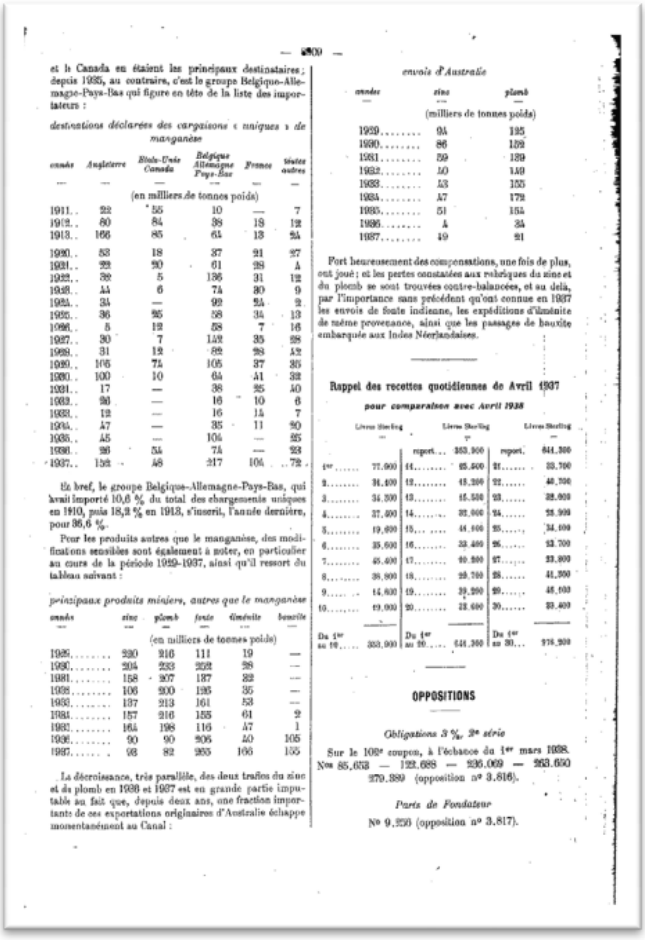


DÉSIGNATION	NOMBRE de pieux.	NOMBRE de cylindres par pieux.	DIAMÈTRE du pieu.	ENFONCEMENT au-dessous du terrain naturel.	ENFONCEMENT dans le sol.	NATURE du travail.	SOMMES payées aux ouvriers.	NATURE DES COUCHES à traverser.	OBSERVATIONS.
						A (1)	B (2)		
Pile de Strood....	14	3	2.156	18.94	12.20	140 00	5,319f 75c	Débris de maçonnerie et de charpente.	
Pile de Rochester.	14	3	2.135	15.42	6.71	89 60	6,587 50	Débris de maçonnerie et de charpente.	
Culée de Strood..	30	2	1.85	»	5.49	82 50	4,807 50	Terrain naturel, c'est-à-dire sable, gravier et craie.	Enfoncé librement, sans air comprimé.
Culée de Rochester	6	2	1.85	»	5.49	40 00	1,055 56	Pierre, débris de charpente et craie.	Enfoncé librement sans air comprimé.
	8	3	1.85	13.81	8.25	70 38	2,020	Terrain naturel.	

A (1) L'unité est la journée de 10 heures.
B (2) Ces sommes sont seulement celles payées pour la main d'œuvre d'enfoncement et de déblai, non compris les frais pour travaux accessoires, enrochement, etc.

Eléments non segmentés

Tableau avec filet de séparation



et le Canada en étaient les principaux destinataires ; depuis 1935, au contraire, c'est le groupe Belgique-Allemagne-Pays-Bas qui figure en tête de la liste des importateurs :

destinations déclarées des cargaisons « uniques » de manganèse

années	Angleterre	États-Unis Canada	Belgique Allemagne Pays-Bas	France	toutes autres
			(en milliers de tonnes poids)		
1911	22	55	10	—	7
1912	60	84	38	18	12
1913	166	85	61	13	24
1920	53	18	37	21	37
1921	22	20	61	28	4
1922	32	5	136	31	12
1923	44	6	74	30	9
1924	34	—	92	24	2
1925	36	25	58	34	13
1926	5	12	58	7	16
1927	30	7	142	35	28
1928	31	12	82	28	42
1929	105	74	105	37	35
1930	100	10	64	41	32
1931	17	—	38	25	10
1932	20	—	16	10	6
1933	12	—	16	14	7
1934	47	—	35	11	20
1935	45	—	104	—	25
1936	26	54	74	—	23
1937	152	48	217	104	72

En bref, le groupe Belgique-Allemagne-Pays-Bas, qui avait importé 10,6 % du total des chargements uniques en 1910, puis 18,2 % en 1913, s'inscrit, l'année dernière, pour 36,6 %.

Pour les produits autres que le manganèse, des modifications sensibles sont également à noter, en particulier au cours de la période 1929-1937, ainsi qu'il ressort du tableau suivant :

principaux produits miniers, autres que le manganèse

années	zinc	plomb	fonte	émeraude	bauxite
			(en milliers de tonnes poids)		
1925	320	216	111	19	—
1926	204	233	222	28	—
1927	158	207	187	22	—
1928	106	200	195	35	—
1929	137	213	161	53	—
1930	157	216	155	61	2
1931	164	198	116	57	1
1936	90	90	206	40	105
1937	98	82	255	106	155

La décroissance, très parallèle, des deux trafics du zinc et du plomb en 1936 et 1937 est en grande partie imputable au fait que, depuis deux ans, une fraction importante de ces exportations originaires d'Australie échappe momentanément au Canal :

envois d'Australie

années	zinc	plomb
	(milliers de tonnes poids)	
1929	94	195
1930	86	192
1931	59	189
1932	50	149
1933	53	155
1934	57	172
1935	51	154
1936	4	34
1937	19	91

Fort heureusement des compensations, une fois de plus, ont joué ; si les pertes constatées aux rubriques du zinc et du plomb se sont trouvées contre-balancées, et au delà, par l'importance sans précédent qu'ont connue en 1937 les envois de fonte indienne, les expéditions d'ilménite de même provenance, ainsi que les passages de bauxite embarquée aux Indes Néerlandaises.

Rappel des recettes quotidiennes de Avril 1937

pour comparaison avec Avril 1938

Livres Sterling	Livres Sterling	Livres Sterling
	report... 353.900	report. 644.300
1er..... 77.900	14..... 25.600	21..... 33.700
2..... 31.400	15..... 43.200	22..... 40.700
3..... 34.300	13..... 15.500	23..... 22.900
4..... 37.400	14..... 32.000	24..... 25.900
5..... 19.600	15..... 44.100	25..... 34.100
6..... 35.600	16..... 33.400	26..... 32.700
7..... 45.400	17..... 40.200	27..... 23.800
8..... 38.800	18..... 29.700	28..... 41.200
9..... 14.800	19..... 32.200	29..... 45.103
10..... 19.000	20..... 23.600	30..... 23.400
Du 1er au 10..... 353.900	Du 1er au 20..... 644.300	Du 1er au 30... 276.200

OPPOSITIONS

Obligations 3 %, 3e série

Sur le 109e coupon, à l'échance du 1er mars 1938. Nos 85.653 — 122.688 — 236.069 — 263.650 279.389 (opposition n° 3.816).

Parts de Fondateur

N° 9.250 (opposition n° 3.817).

Tableaux sans filet de séparation, dans une maquette en deux colonnes

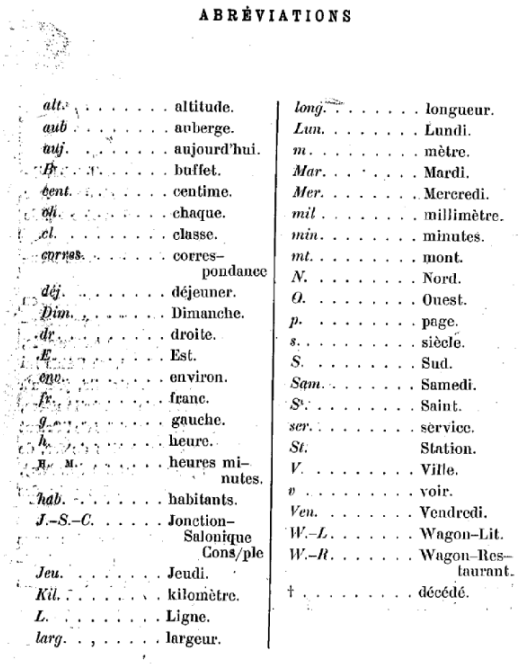
Taux qualité des tableaux

Le taux qualité des contenus tableau sera traité différemment pour les deux types de tableau suivants :

- tableau simple et court de deux colonnes, le contenu étant généralement :
  - composé avec le même corps que le texte courant,
  - présenté sans filet ni bordure,
  - les deux colonnes étant séparées par des tabulations ou des points de suite.

  Ces contenus seront traités en OCR **taux qualité garantie**.

EXEMPLE

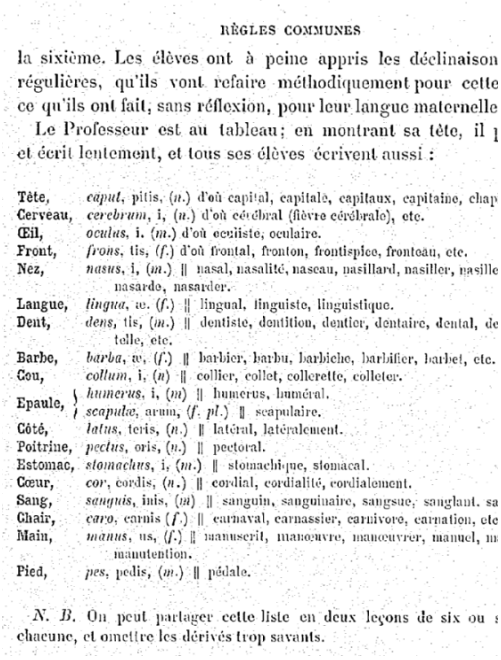


ABRÉVIATIONS

alt.	altitude.
aub	auberge.
auj.	aujourd'hui.
B.	buffet.
cent.	centime.
ch.	chaque.
cl.	classe.
corres.	correspondance
déj.	déjeuner.
Dim.	Dimanche.
dr.	droite.
E.	Est.
env.	environ.
fr.	franc.
g.	gauche.
h.	heure.
h. m.	heures minutes.
hab.	habitants.
J.-S.-C.	Jonction-Salonique Cons/ple
Jeu.	Jeudi.
Kil.	kilomètre.
L.	Ligne.
larg.	largeur.
long.	longueur.
Lun.	Lundi.
m.	mètre.
Mar.	Mardi.
Mer.	Mercredi.
mil	millimètre.
min.	minutes.
mt.	mont.
N.	Nord.
O.	Ouest.
p.	page.
s.	siècle.
S.	Sud.
Sam.	Samedi.
S.	Saint.
ser.	service.
St.	Station.
V.	Ville.
v	voir.
Ven.	Vendredi.
W.-L.	Wagon-Lit.
W.-R.	Wagon-Restaurant.
†	décédé.

http://gallica.bnf.fr/ark:/12148/bpt6k55584488/f19.image

Glossaire, liste d'abréviations



RÈGLES COMMUNES

la sixième. Les élèves ont à peine appris les déclinaison régulières, qu'ils vont refaire méthodiquement pour cette ce qu'ils ont fait, sans réflexion, pour leur langue maternelle

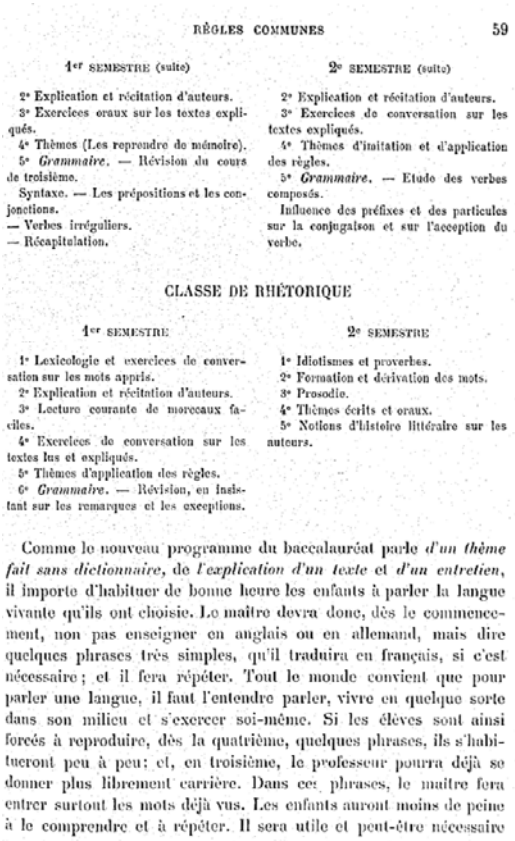
Le Professeur est au tableau; en montrant sa tête, il l et écrit lentement, et tous ses élèves écrivent aussi :

Tête,	*caput*, pitis, (*n.*) d'où capital, capitale, capitaux, capitaine, chap
Cerveau,	*cerebrum*, i, (*n.*) d'où cérébral (fièvre cérébrale), etc.
Œil,	*oculus*, i. (*m.*) d'où oculiste, oculaire.
Front,	*frons*, tis, (*f.*) d'où frontal, fronton, frontispice, fronteau, etc.
Nez,	*nasus*, i, (*m.*) ‖ nasal, nasalité, naseau, nasillard, nasiller, nasille nasarde, nasarder.
Langue,	*lingua*, æ. (*f.*) ‖ lingual, linguiste, linguistique.
Dent,	*dens*, tis, (*m.*) ‖ dentiste, dentition, dentier, dentaire, dental, de telle, etc.
Barbe,	*barba*, æ, (*f.*) ‖ barbier, barbu, barbiche, barbifier, barbet, etc.
Cou,	*collum*, i, (*n*) ‖ collier, collet, collerette, colleter.
Epaule,	*humerus*, i, (*m*) ‖ humerus, huméral. *scapulæ*, arum, (*f. pl.*) ‖ scapulaire.
Côté,	*latus*, teris, (*n.*) ‖ latéral, latéralement.
Poitrine,	*pectus*, oris, (*n.*) ‖ pectoral.
Estomac,	*stomachus*, i, (*m.*) ‖ stomachique, stomacal.
Cœur,	*cor*, cordis, (*n.*) ‖ cordial, cordialité, cordialement.
Sang,	*sanguis*, inis, (*m*) ‖ sanguin, sanguinaire, sangsue, sanglant. sa
Chair,	*caro*, carnis (*f.*) ‖ carnaval, carnassier, carnivore, carnation, etc
Main,	*manus*, us, (*f.*) ‖ manuscrit, manœuvre, manœuvrer, manuel, m manutention.
Pied,	*pes*, pedis, (*m.*) ‖ pédale.

N. B. On peut partager cette liste en deux leçons de six ou chacune, et omettre les dérivés trop savants.

http://gallica.bnf.fr/ark:/12148/bpt6k55125514/f40.image

Liste de définitions



RÈGLES COMMUNES 59

1er SEMESTRE (suite)

2e Explication et récitation d'auteurs.
3e Exercices oraux sur les textes expliqués.
4e Thèmes (Les reprendre de mémoire).
5e *Grammaire.* — Révision du cours de troisième.
Syntaxe. — Les prépositions et les conjonctions.
— Verbes irréguliers.
— Récapitulation.

2e SEMESTRE (suite)

2e Explication et récitation d'auteurs.
3e Exercices de conversation sur les textes expliqués.
4e Thèmes d'imitation et d'application des règles.
5e *Grammaire.* — Etude des verbes composés.
Influence des préfixes et des particules sur la conjugaison et sur l'acception du verbe.

CLASSE DE RHÉTORIQUE

1er SEMESTRE

1e Lexicologie et exercices de conversation sur les mots appris.
2e Explication et récitation d'auteurs.
3e Lecture courante de morceaux faciles.
4e Exercices de conversation sur les textes lus et expliqués.
5e Thèmes d'application des règles.
6e *Grammaire.* — Révision, en insistant sur les remarques et les exceptions.

2e SEMESTRE

1e Idiotismes et proverbes.
2e Formation et dérivation des mots.
3e Prosodie.
4e Thèmes écrits et oraux.
5e Notions d'histoire littéraire sur les auteurs.

Comme le nouveau programme du baccalauréat parle *d'un thème fait sans dictionnaire*, de *l'explication d'un texte* et *d'un entretien*, il importe d'habituer de bonne heure les enfants à parler la langue vivante qu'ils ont choisie. Le maître devra donc, dès le commencement, non pas enseigner en anglais ou en allemand, mais dire quelques phrases très simples, qu'il traduira en français, si c'est nécessaire ; et il fera répéter. Tout le monde convient que pour parler une langue, il faut l'entendre parler, vivre en quelque sorte dans son milieu et s'exercer soi-même. Si les élèves sont ainsi forcés à reproduire, dès la quatrième, quelques phrases, ils s'habitueront peu à peu ; et, en troisième, le professeur pourra déjà se donner plus librement carrière. Dans ces phrases, le maître fera entrer surtout les mots déjà vus. Les enfants auront moins de peine à le comprendre et à répéter. Il sera utile et peut-être nécessaire

http://gallica.bnf.fr/ark:/12148/bpt6k55125514/f80.image

Texte composé sur deux colonnes

ATTENTION

LES TABLES DES MATIERES ET INDEX, QUI ONT SOUVENT LA FORME D'UN TABLEAU ET QUI SONT TRANSCRIT DANS CERTAINS MARCHES NE SONT PAS CONCERNES PAR CETTE REGLE ; CES CONTENUS DOIVENT ETRE TRAITES SELON LES CONSIGNES DU « REFERENTIEL TABLES ». UN FICHIER ALTO OCR BRUT EST CEPENDANT FOURNI POUR CES PAGES DANS LE DOCUMENTS NUMERIQUES.

- tous les autres types de tableaux (mise en page complexe, tableaux de texte et nombres, petit corps de police, etc.).

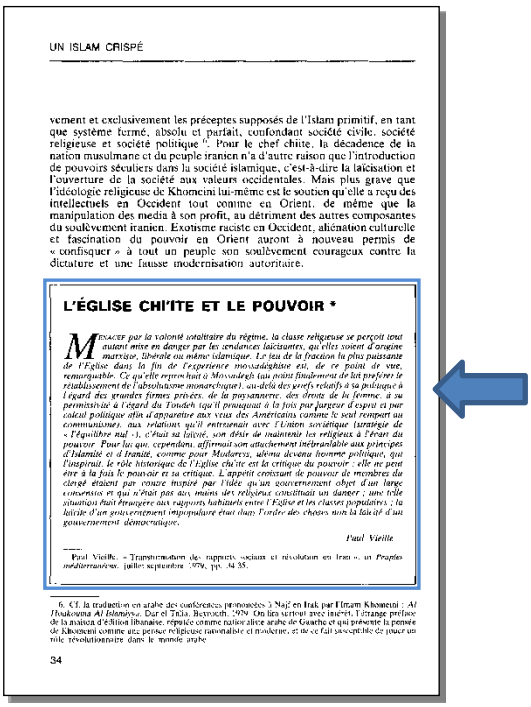
  Ces contenus seront traités en **OCR brut**.

5.6.4 Encadrés

Il s'agit d'un encadré au sein d'une colonne (c'est-à-dire qui n'est pas un élément central.)

En règle générale, il n'y a pas un balisage particulier de l'encadré.

EXEMPLE



UN ISLAM CRISPÉ

vement et exclusivement les préceptes supposés de l'Islam primitif, en tant que système fermé, absolu et parfait, confondant société civile, société religieuse et société politique [6]. Pour le chef chiite, la décadence de la nation musulmane et du peuple iranien n'a d'autre raison que l'introduction de pouvoirs séculiers dans la société islamique, c'est-à-dire la laïcisation et l'ouverture de la société aux valeurs occidentales. Mais plus grave que l'idéologie religieuse de Khomeini lui-même est le soutien qu'elle a reçu des intellectuels en Occident tout comme en Orient, de même que la manipulation des media à son profit, au détriment des autres composantes du soulèvement iranien. Exotisme raciste en Occident, aliénation culturelle et fascination du pouvoir en Orient auront à nouveau permis de « confisquer » à tout un peuple son soulèvement courageux contre la dictature et une fausse modernisation autoritaire.

L'ÉGLISE CHI'ITE ET LE POUVOIR *

Menacée par la volonté totalitaire du régime, la classe religieuse se perçoit tout autant mise en danger par les tendances laïcisantes, qu'elles soient d'origine marxiste, libérale ou même islamique. Le seul de la fraction la plus puissante de l'Eglise dans la fin de l'expérience mossadeghiste est, de ce point de vue, remarquable. Ce qu'elle reprochait à Mossadegh fut moins finalement de lui préférer le rétablissement de l'absolutisme monarchique), au-delà des griefs relatifs à sa politique à l'égard des grandes firmes privées, de la paysannerie, des droits de la femme, à sa permissivité à l'égard du Tudeh qu'il pratiquait à la fois par largeur d'esprit et par calcul politique afin d'apparaître aux yeux des Américains comme le seul rempart au communisme, aux relations qu'il entretenait avec l'Union soviétique (stratégie de « l'équilibre nul »), c'était sa laïcité, son désir de maintenir les religieux à l'écart du pouvoir. Pour lui, cependant, affirmait son attachement inébranlable aux principes d'Islamité et d'Iranité, comme pour Modarress, ulémа devenu homme politique, qui l'inspirait, le rôle historique de l'Eglise chi'ite est la critique du pouvoir : elle ne peut être à la fois le pouvoir et sa critique. L'appétit croissant de pouvoir de membres du clergé étaient par ailleurs inspiré par l'idée qu'un gouvernement objet d'un large consensus et qui n'était pas aux mains des religieux constituait un danger ; une telle situation était étrangère aux rapports habituels entre l'Eglise et les classes populaires ; la laïcité d'un gouvernement impopulaire était dans l'ordre des choses non la laïcité d'un gouvernement démocratique.

Paul Vieille

Paul Vieille, « Transformation des rapports sociaux et révolution en Iran », in *Peuples méditerranéens*, juillet-septembre 1979, pp. 34-35.

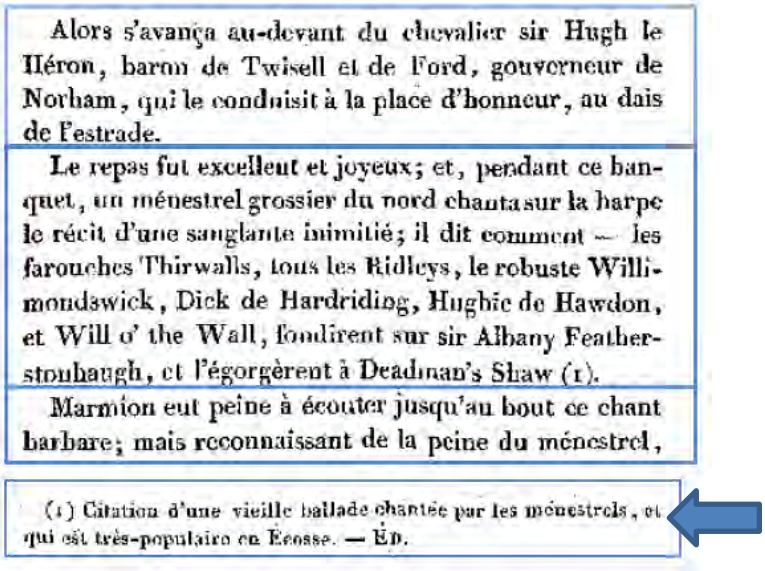
6. Cf. la traduction en arabe des conférences prononcées à Najf en Irak par Ilham Khomeini : *Al Thaksima Al Islamiyya*, Dar el Tela, Beyrouth, 1979. On lira surtout avec intérêt, l'étrange préface de la maison d'édition libanaise, réputée comme nation alice arabe de Gauche et qui présente la pensée de Khomeini comme une pensée religieuse rationaliste et moderne, et ne ce fait susceptible de jouer un rôle révolutionnaire dans le monde arabe.

34

5.6.5 Notes de bas de page

Les notes sont segmentées dans un bloc différent de celle du corps du texte. Chaque note doit être segmentée dans un bloc distinct (voir aussi section 5.5.4).

EXEMPLE



Alors s'avança au-devant du chevalier sir Hugh le Héron, baron de Twisell et de Ford, gouverneur de Norham, qui le conduisit à la place d'honneur, au dais de l'estrade.

Le repas fut excellent et joyeux; et, pendant ce banquet, un ménestrel grossier du nord chanta sur la harpe le récit d'une sanglante inimitié; il dit comment — les farouches Thirwalls, tous les Ridleys, le robuste Willimondswick, Dick de Hardriding, Hughie de Hawdon, et Will o' the Wall, fondirent sur sir Albany Featherstonhaugh, et l'égorgèrent à Deadman's Shaw (1).

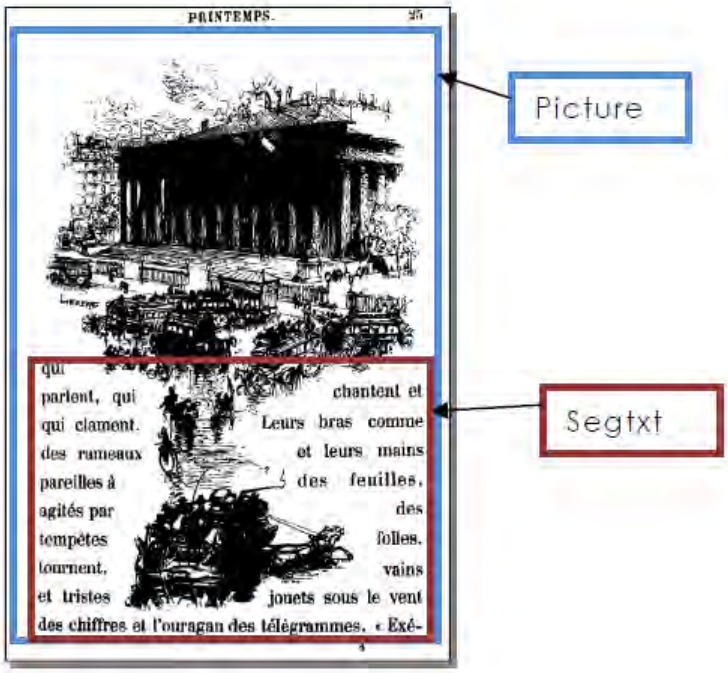
Marmion eut peine à écouter jusqu'au bout ce chant barbare; mais reconnaissant de la peine du ménestrel,

(1) Citation d'une vieille ballade chantée par les ménestrels, et qui est très-populaire en Écosse. — ÉD.

5.6.6 Illustrations avec habillage de texte traversant

Il s'agit des cas où l'ordre de la lecture « traverse » ligne par ligne l'image, comme dans l'exemple ci-dessous.

EXEMPLE



Ce type de mise en page sera décrit avec un composedBlock comprenant des blocs textes pour le texte et un bloc Illustration autour de l'image, ainsi la lecture ligne à ligne de l'original est respectée (voir aussi section 5.9).

5.6.7 Publicités et catalogues d'éditeur

Il s'agit de zones à caractère publicitaire qui présentent des éléments typographiques particuliers tels que polices curvilignes, logos, encadrés, etc. Elles seront traitées en OCR brut.

NOTES

Ces pages sont de type « A » dans le manifeste numérique refNum.

EXEMPLE


AUX
PHARES
DE LA
BASTILLE
LE SUCCÈS superbe habillement complet en drap nouveauté..... 26f
MARIAGE Habillement complet, redingote, pantalon, gilet, le tout pour..... 32f
L'INUSABLE magnifique pardessus pour hommes, drap nouveauté riche... 18f
1re COMMUNION Habillement complet tout en drap noir fin..... 14f
SOLIDITÉ, ÉLÉGANCE, BON MARCHÉ
Ne se trouvent qu'aux
PHARES DE LA BASTILLE
5 et 7, place de la Bastille
PARIS
Envoi *franco* du MAGNIFIQUE CATALOGUE ILLUSTRÉ à toute personne qui en fait la demande.

Si une zone de publicité comporte du texte et des éléments typographiques particuliers, l'ensemble de la zone est identifiée en publicité. Chaque bloc de texte contenu dans une publicité sera donc marqué comme un TextBlock avec l'attribut TYPE="advertisement".

Si plusieurs publicités sont adjacentes, elles ne sont pas groupées au sein d'un même bloc mais identifiées séparément.

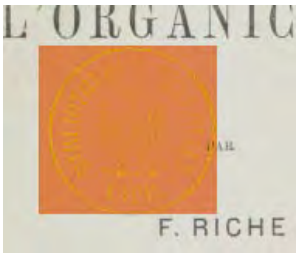
NOTES

Si un bloc de publicité est composé sous forme d'un tableau, le typage TYPE="advertisement" est prioritaire.

5.6.8 Texte sous tampon

Un bloc de texte placé sous un tampon sera marqué comme un TextBlock avec l'attribut TYPE="textStamped".

EXEMPLE

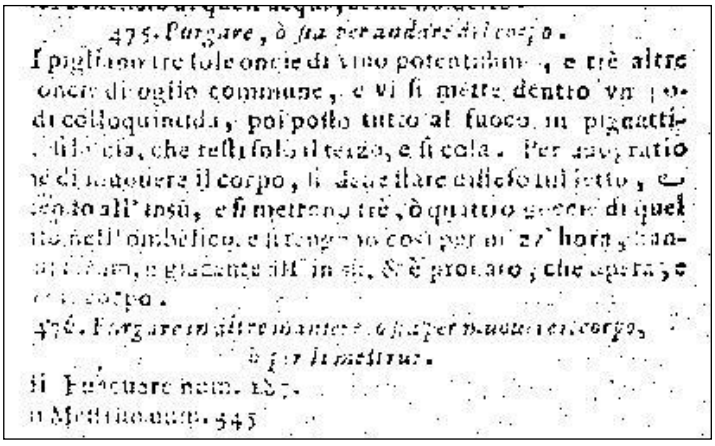


Ici, le bloc "PAR"

5.6.9 Texte illisible

Un bloc de texte illisible du fait de la dégradation physique du document, de problème de courbure ou de tout autre phénomène affectant la netteté de l'image, sera marqué en tant que TextBlock avec l'attribut TYPE="illegible".

EXEMPLE



La section 6.2 donne une définition précise de la notion de « texte illisible ».

5.7 Illustrations

Tout élément de nature visuelle qui n'est à l'évidence pas une décoration, une lettrine ou un tampon sera traité sous la forme d'un bloc Illustration, sans chercher à statuer sur le lien éditorial qu'il a ou non avec le contenu textuel : cela couvre entre autres les illustrations, figures, diagrammes, photos, dessins, schémas, reproductions d'art, plans et cartes.

NOTES

Le prestataire et la BnF évalueront durant la phase de test si des formes géométriques autres que le rectangle (polygone, cercle, etc.) peuvent être utilisées pour segmenter illustrations et éléments graphiques.

Dans le cas d'utilisation de la seule forme rectangle, la couverture d'un bloc illustration ou élément graphique pourra être répartie en plusieurs segments rectangulaires sans importance de l'ordre de ces blocs dans l'ALTO, mais à condition que l'ensemble des blocs couvre la totalité de la zone illustrée.

OBLIGATOIRE

On peut accepter que les illustrations combinées à d'autres éléments amènent à découper une zone « illustrée » en plusieurs blocs, à l'exception des partitions, formules et cartes qui doivent être découpées en un seul bloc

Le typage D (« dessin ») dans le fichier RefNum pourra être généré automatiquement à partir du typage ALTO : toute page dont le fichier ALTO contient un seul bloc Illustration (éventuellement accompagné d'un bloc de texte pour la légende de l'illustration) et qui n'a pas un autre typage prioritaire (P/E/T/I/R/L).

OBLIGATOIRE

Pour générer un typage D, l'illustration doit occuper la majorité de la surface de la page.

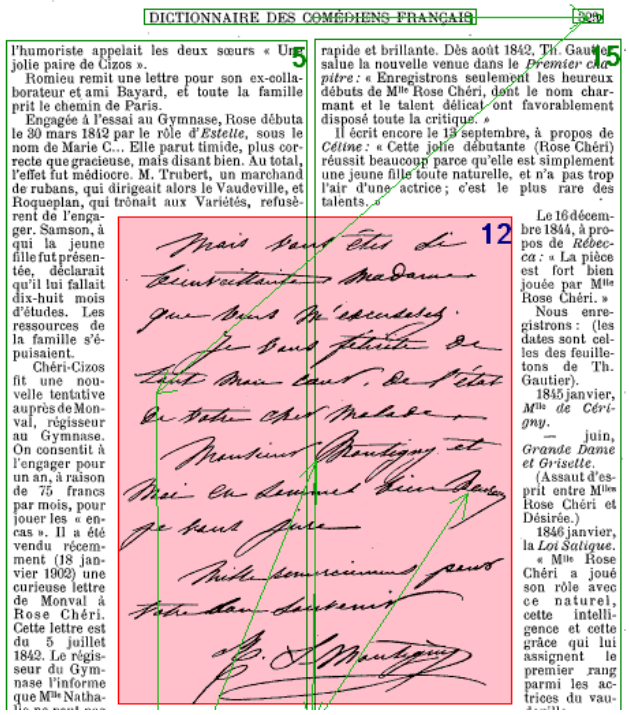
Exemple de cas de typage D

http://gallica.bnf.fr/ark:/12148/bpt6k65164532/f19.image

Illustration en écriture manuscrite

Certaines illustrations sont la reproduction d'un texte manuscrit (extraits de manuscrit d'auteur, par exemple). Ces blocs seront décrits sous forme d'éléments Illustration (et non avec un élément graphique, cf. section 5.8.5).

EXEMPLE


DICTIONNAIRE DES COMÉDIENS FRANÇAIS

l'humoriste appelait les deux sœurs « Une jolie paire de Cizos ».

Romieu remit une lettre pour son ex-collaborateur et ami Bayard, et toute la famille prit le chemin de Paris.

Engagée à l'essai au Gymnase, Rose débuta le 30 mars 1842 par le rôle d'*Estelle*, sous le nom de Marie C... Elle parut timide, plus correcte que gracieuse, mais disant bien. Au total, l'effet fut médiocre. M. Trubert, un marchand de rubans, qui dirigeait alors le Vaudeville, et Roqueplan, qui trônait aux Variétés, refusèrent de l'engager. Samson, à qui la jeune fille fut présentée, déclarait qu'il lui fallait dix-huit mois d'études. Les ressources de la famille s'épuisaient.

Chéri-Cizos fit une nouvelle tentative auprès de Monval, régisseur au Gymnase. On consentit à l'engager pour un an, à raison de 75 francs par mois, pour jouer les « encas ». Il a été vendu récemment (18 janvier 1902) une curieuse lettre de Monval à Rose Chéri. Cette lettre est du 5 juillet 1842. Le régisseur du Gymnase l'informe que Mlle Natha-

rapide et brillante. Dès août 1842, Th. Gautier salue la nouvelle venue dans le *Premier chapitre* : « Enregistrons seulement les heureux débuts de Mlle Rose Chéri, dont le nom charmant et le talent délicat ont favorablement disposé toute la critique. »

Il écrit encore le 13 septembre, à propos de *Céline* : « Cette jolie débutante (Rose Chéri) réussit beaucoup parce qu'elle est simplement une jeune fille toute naturelle, et n'a pas trop l'air d'une actrice ; c'est le plus rare des talents. »

Le 16 décembre 1844, à propos de *Rébecca* : « La pièce est fort bien jouée par Mlle Rose Chéri. »

Nous enregistrons : (les dates sont celles des feuilletons de Th. Gautier).

1845 janvier, Mlle *de Cérigny*.

— juin, *Grande Dame et Grisette*.

(Assaut d'esprit entre Mlles Rose Chéri et Désirée.)

1846 janvier, la *Loi Salique*.

« Mlle Rose Chéri a joué son rôle avec ce naturel, cette intelligence et cette grâce qui lui assignent le premier rang parmi les actrices du vau-

Exemple d'illustration sous forme d'écriture manuscrite

5.7.1 Formules chimiques, mathématiques

La structuration des formules est réalisée en mode image, avec un bloc Illustration doté d'un attribut TYPE égal à "formula".

Il est permis de réunir plusieurs formules adjacentes (c'est-à-dire, quand il n'y a aucun autre bloc entre elles) en un seul bloc.

EXEMPLE

$$bO = AG - Am\ ;\ FO = FG - bm\ ;$$
$$bF = \sqrt{\overline{bO}^2 + \overline{FO}^2};\ \text{sin. } bFO = \frac{bO}{bF}.$$

$$K_A = \frac{[A^-]_{eq} \cdot [H_3O^+]_{eq}}{[AH]_{eq}}$$

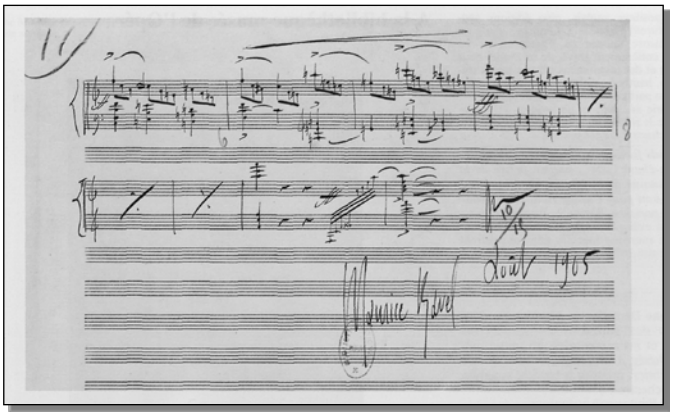
NOTES

Quand des formules sont incluses dans des paragraphes de texte, elles ne sont pas traitées spécifiquement et la conversion OCR pourra donner des résultats illisibles.

5.7.2 Partitions

Les extraits de partition de musique sont encodés en bloc Illustration dotés d'un attribut TYPE égal à "musicScore".

EXEMPLE

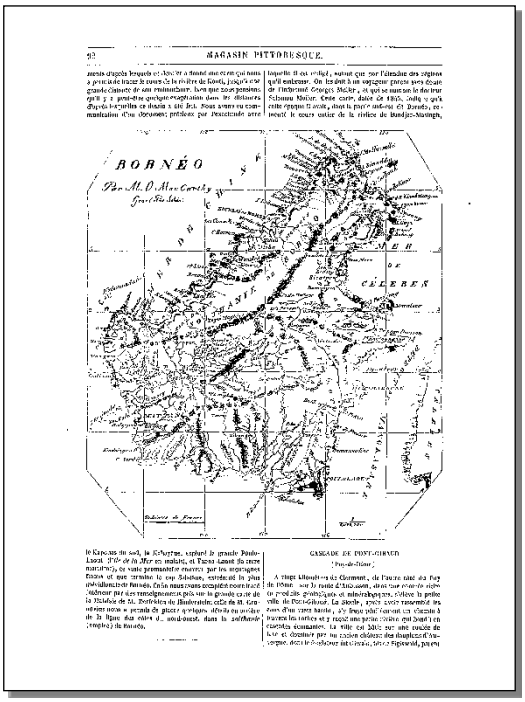


Exemple de partition

5.7.3 Cartes

Les illustrations reconnues comme étant des cartes géographiques seront typées spécifiquement en utilisant l'attribut TYPE égal à "map".

EXEMPLE



Exemple de carte

5.7.4 Alphabets non latins

Comme il a été dit section 4.1.2, les blocs de texte composés avec des alphabets non latins ou des systèmes d'idéogrammes seront décrits sous la forme de blocs image. Ces blocs seront des éléments Illustration dotés d'un attribut TYPE="nonLatinScript".

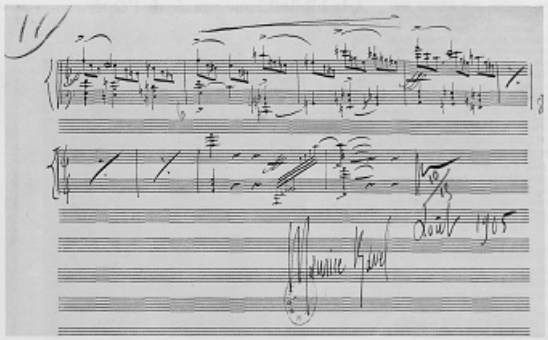
5.7.5 Imbrication de blocs Illustration et d'autres blocs, notamment des TextBlock

Quand une illustration s'accompagne d'une légende, il convient d'utiliser un ComposedBlock qui inclut l'image et la zone de texte.

NOTES

Selon les typologies de mise en page, cette règle ne sera pas toujours applicable et elle n'est pas soumise au taux qualité garantie.

EXEMPLE



Maurice Ravel. Pirnde de la *Sonatine* pour piano.

Peter Prelleur. *The modern musick-master*, 1731. Gravure de J. Smith.

De la collection G. Thibault-de Chambure :

- **Maurizio Cazzati.** *Risposta alle opositioni fatte dal signor Giulio Cesare Arresti nella lettera al lettore posta nell'opera sua musicale.* Bologna, eredi del Dozza, 1663.

Cette plaquette, de soixante-douze pages, témoigne de la polémique entre deux compositeurs : Arresti, organiste à San Petronio de Bologne, et Cazzati (1616-1678), maître de chapelle de cette église, à propos de la technique utilisée dans une messe par ce dernier. Seul exemplaire connu en France. Ex-libris d'Henry Prunières.

- **Gottfried Keller.** *Rules or a compleat method for attaining to play a thorough bass upon the harpsicord, organ or archlute... to which is added an exact scale for tuneing the harpsicord or spinnet.* London, J. Walsh, s.d. (première moitié du XVIII[e] siècle).

Six éditions de la méthode pour la basse continue de Keller, compositeur allemand du XVII[e] siècle établi à Londres, parurent dans cette ville de 1707 à 1730, outre celle-ci, encore inconnue et qui est la seule conservée en France.

- **Nicolaus Listenius.** *Rudimenta musicae, in gratiam studiose juventutis diligenter comportata.* Augsburg, H. Steyner, 1536.

Les cinquante-trois éditions de ce petit traité d'un maître de musique brandebourgeois publiées de 1533 à 1583 en Allemagne attestent sa grande popularité. La Bibliothèque nationale de France en possède cinq, de 1544 à 1557. Il fut d'abord préfacé par le réformateur Bugenhagen ; puis très augmenté à partir de 1537, son titre devint *Musica... novis regulis et exemplis adjuncta.* Ex-libris d'Alfred Cortot.

- **Peter Prelleur.** *The modern musick-master or the universal musician...* London, Printing Office, 1731.

De cet ouvrage de Prelleur, claveciniste

77

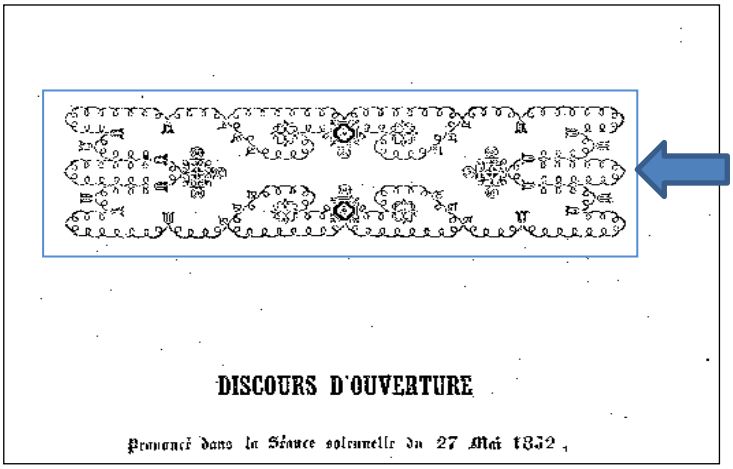
Mise en page de type presse, avec placement « souple » des légendes d'illustration

5.8 Eléments graphiques

5.8.1 Décorations et ornements

Les décorations et autres ornements sont à capturer en bloc GraphicalElement.

EXEMPLE



Exemple de décoration

5.8.2 Tampons

Les tampons sont à capturer en bloc GraphicalElement doté d'un attribut TYPE égal à "stamp".

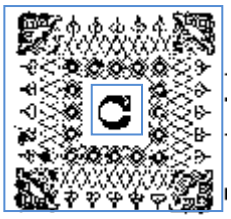
EXEMPLE

Exemple de tampon

5.8.3 Lettrines (lettres ornées)

Les lettres ornées seront capturées en ComposedBlock avec un bloc texte pour le texte et un bloc GraphicalElement autour des ornements, doté d'un attribut TYPE égal à "dropCap".

EXEMPLE

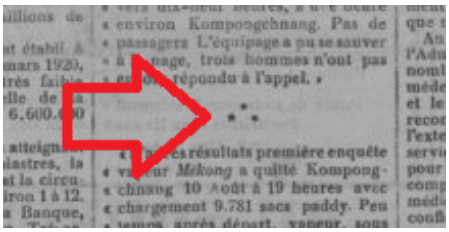


Exemple de lettrine (lettre C)

5.8.4 Traits de séparation

Les traits de séparation et les culs-de-lampe placés entre deux paragraphes doivent être décrits par un bloc GraphicalElement avec l'attribut TYPE="transition".

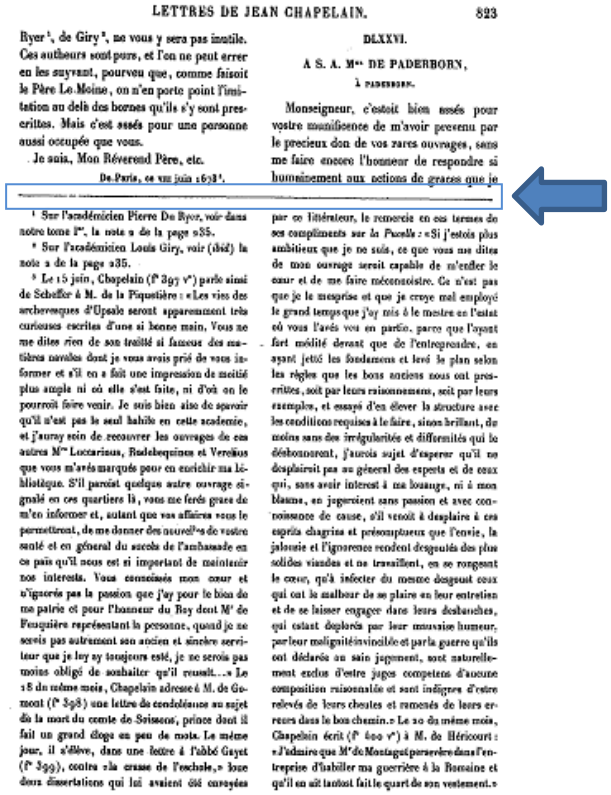
EXEMPLE



Exemples de cul de lampe et traits simples dits « de transition »

Les traits de séparation entre corps du texte et notes de bas de page sont à décrire en GraphicalElement avec l'attribut TYPE="note".

EXEMPLE



LETTRES DE JEAN CHAPELAIN. 823

Ryer[1], de Giry[2], ne vous y sera pas inutile. Ces autheurs sont purs, et l'on ne peut errer en les suyvant, pourveu que, comme faisoit le Père Le Moine, on n'en porte point l'imitation au delà des bornes qu'ils s'y sont prescrittes. Mais c'est assés pour une personne aussi occupée que vous.

Je suis, Mon Réverend Père, etc.

De Paris, ce vm juin 1638[3].

DLXXVI.

A S. A. Mgr DE PADERBORN,

à paderborn.

Monseigneur, c'estoit bien assés pour vostre munificence de m'avoir prevenu par le precieux don de vos rares ouvrages, sans me faire encore l'honneur de respondre si humainement aux actions de graces que je

[1] Sur l'académicien Pierre Du Ryer, voir dans notre tome Ier, la note 2 de la page 235.

[2] Sur l'académicien Louis Giry, voir (ibid) la note 2 de la page 235.

[3] Le 15 juin, Chapelain (fo 397 vo) parle ainsi de Scheffer à M. de la Piquetière : « Les vies des archevesques d'Upsale seront apparemment très curieuses escrites d'une si bonne main. Vous ne me dites rien de son traitté si fameux des matières navales dont je vous avois prié de vous informer et s'il en a fait une impression de moitié plus ample ni où elle s'est faite, ni d'où on le pourroit faire venir. Je suis bien aise de sçavoir qu'il n'est pas le seul habile en cette academie, et j'auray soin de recouvrer les ouvrages de ces autres Mrs Loccarinus, Rudebequinus et Verelius que vous m'avés marqués pour en enrichir ma bibliotèque. S'il paroist quelque autre ouvrage signalé en ces quartiers là, vous me ferés grace de m'en informer et, autant que vos affaires vous le permettront, de me donner des nouvell[es] de vostre santé et en général du succès de l'ambassade en ce pais qu'il nous est si important de maintenir nos interests. Vous connoissés mon cœur et n'ignorés pas la passion que j'ay pour le bien de ma patrie et pour l'honneur du Roy dont Mr de Fesquière représentant la personne, quand je ne sçavois pas autrement son ancien et sincère serviteur que je luy ay toujours esté, je ne serois pas moins obligé de souhaiter qu'il reussît... » Le 18 du même mois, Chapelain adresse à M. de Gomont (fo 398) une lettre de condoléance au sujet de la mort du comte de Soissons, prince dont il fait un grand éloge en peu de mots. Le même jour, il s'élève, dans une lettre à l'abbé Gayet (fo 399), contre « la cruauté de l'eschole, » pour deux dissertations qui lui avaient été envoyées par ce littérateur, le remercie en ces termes de ses compliments sur la *Pucelle* : « Si j'estois plus ambitieux que je ne suis, ce que vous me dites de mon ouvrage seroit capable de m'enfler le cœur et de me faire méconnoistre. Ce n'est pas que je le mesprise et que je croye mal employé le grand temps que j'y mis à le mestre en l'estat où vous l'avés veu en partie, parce que l'ayant fort médité devant que de l'entreprendre, en ayant jetté les fondemens et levé le plan selon les règles que les bons anciens nous ont prescrittes, soit par leurs raisonnemens, soit par leurs exemples, et essayé d'en élever la structure avec les conditions requises à le faire, sinon brillant, de moins sans des irrégularités et difformités qui le déshonorent, j'aurois sujet d'esperer qu'il ne desplairoit pas au général des experts et de ceux qui, sans avoir interest à ma louange, ni à mon blasme, en jugeroient sans passion et avec connoissance de cause, s'il venoit à desplaire à ces esprits chagrins et présomptueux que l'envie, la jalousie et l'ignorance rendent desgoutés des plus solides viandes et ne travaillent, en se rongeant le cœur, qu'à infecter du mesme desgoust ceux qui ont le malheur de se plaire en leur entretien et de se laisser engager dans leurs desbauches, qui estant déplorés par leur mauvaise humeur, par leur malignité invincible et par la guerre qu'ils ont déclarée au sain jugement, sont naturellement exclus d'estre juges competens d'aucune composition raisonnable et sont indignes d'estre relevés de leurs chutes et ramenés de leurs erreurs dans le bon chemin. » Le 20 du même mois, Chapelain écrit (fo 400 vo) à M. de Héricourt : « J'admire que Mr de Montagut persevère dans l'entreprise d'habiller ma guerrière à la Romaine et qu'il en ait tantost fait le quart de son vestement. »

Exemple de cul-de-lampe

Tous les autres cas de culs-de-lampe et de décorations (notamment ceux placés en fin de page ou en fin de chapitre) sont à décrire en GraphicalElement (sans attribut TYPE).

EXEMPLE

Exemple de cul-de-lampe

5.8.5 Ecriture manuscrite

Les mentions manuscrites parfois présentes sur les pages des documents sont à capturer en GraphicalElement, l'attribut TYPE ayant pour valeur "manuscript".

EXEMPLE

Exemples d'écriture manuscrite

ATTENTION

CERTAINES PORTIONS D'ECRITURE MANUSCRITE SONT DES ILLUSTRATIONS. ELLES DOIVENT ETRE TRAITEES SELON LES CONSIGNES ET EXEMPLE DE LA SECTION 5.7.

5.8.6 Imbrication de blocs GraphicalElement et d'autres blocs, notamment des TextBlock

Les blocs GraphicalElement peuvent englober (via un ComposedBlock) d'autres blocs, notamment des TextBlock.

La couverture d'un bloc GraphicalElement peut être répartie en plusieurs segments sans importance de l'ordre de ces blocs dans l'ALTO mais à condition que l'ensemble des blocs couvre la totalité de la zone « graphique ».

5.9 Blocs composés

L'élément ComposedBlock est utilisé pour permettre l'imbrication d'éléments de type bloc.

5.9.1 Texte au sein des illustrations ou des éléments graphiques

Ce cas concerne un texte pleinement intégré, c'est-à-dire que le texte fait partie intégrante de la zone graphique :

- de courts textes explicatifs au sein d'une illustration, notamment des légendes imbriquées dans le rectangle de l'illustration.
- les textes au sein même d'un dessin, d'une œuvre d'art, d'un ornement, etc. (cf. section 5.8.1).

EXEMPLE

Ce cas est typé soit par un ComposedBlock comme décrit ci-dessus, soit par des blocs séparés. Dans ce dernier cas, pour éviter un chevauchement des blocs au même niveau :

- *Les blocs illustrations ne touchent pas le bloc texte ; l'image est répartie en un ou plusieurs blocs Illustration, dont l'ordre n'est pas indiqué.*
- *Les blocs texte (ici, un seul) peuvent figurer avant, au milieu ou après les blocs Illustration/GraphicalElement à conditions qu'entre eux, ils respectent l'ordre de la lecture.*

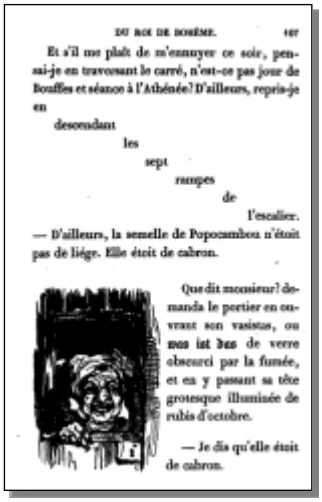
5.9.2 Imbrication d'illustrations ou d'éléments graphiques et de texte

Ce cas concerne du texte qui habille étroitement une illustration, en particulier, une illustration non rectangulaire :

- les TextBlock seront serrés autour du texte – jusqu'à un TextBlock distinct par ligne de texte (cf. exemple de la section 5.6.7).
- l'illustration sera couverte par un bloc Illustration/GraphicalElement, qui peut recouvrir entièrement ou partiellement les TextBlock ;
- l'ordre des blocs Illustration/GraphicalElement est libre, à condition que les TextBlock figurent dans l'ordre de la lecture. Les TextBlock peuvent être non contigus.

Lors d'un recouvrement au même niveau, les TextBlock et les GraphicalElement seront englobés par un ComposedBlock.

EXEMPLE


DU ROI DE BOHÊME. 197

Et s'il me plaît de m'ennuyer ce soir, pensai-je en traversant le carré, n'est-ce pas jour de Bouffes et séance à l'Athénée? D'ailleurs, repris-je en descendant les sept rampes de l'escalier.

— D'ailleurs, la semelle de Popocambou n'étoit pas de liége. Elle étoit de cabron.

Que dit monsieur? demanda le portier en ouvrant son vasistas, ou mon lit bas de verre obscurci par la fumée, et en y passant sa tête grotesque illuminée de rubis d'octobre.

— Je dis qu'elle étoit de cabron.

Ce cas est typé soit par un ComposedBlock comme décrit ci-dessus, soit par des blocs séparés. Les illustrations (ici, une seule) peuvent figurer avant, au milieu ou après les TextBlock.

5.9.3 Ordre de lecture entre texte et illustrations ou éléments graphiques

Quand un ou plusieurs TextBlock partagent la largeur de la page ou de la colonne avec un ou plusieurs blocs Illustration/graphicalElement, seul compte l'ordre de lecture des TextBlock. L'ordre des blocs Illustration/graphicalElement n'est pas indiqué, ni l'emplacement de chacun d'entre eux avant, après ou au milieu des TextBlock.

5.10 Tableau récapitulatif de la structuration ALTO

TextBlock			
Cas	Attribut TYPE	Section	Remarque
Paragraphe		5.6.1	
Titre	"titre1", "titre2"	4.2.2, 5.6.2	
Tableau	"table"	5.6.3	
Encadré		5.6.4	
Note (note de bas de page, note marginale)		5.6.5	
Texte habillé par une illustration		5.6.6	utiliser un ComposedBlock
Publicité	"advertisement"	5.6.7	
Police “script”	"scriptFonts"	4.2.3	
Texte de l'ouvrage sous un tampon	"textStamped"	5.6.8	

Texte illisible	"illegible"	4.2.3, 5.6.9	
Illustration			
Cas	Attribut TYPE	Section	
Image, illustration, schéma, etc.		5.7	
Formule mathématique ou chimique	"formula"	5.7.1	
Partition	"musicalNotation"	5.7.2	
Carte, plan	"map"	5.7.3	
Alphabet non latin	"nonLatinScript"	4.1.2, 5.7.4	
Illustration avec legende		5.7.6	utiliser un ComposedBlock
GraphicalElement			
Cas	Attribut TYPE	Section	
Décoration, cul de lampe		5.8.1	
Tampon	"stamp"	5.8.2	
Lettrine	"dropCap"	5.8.3	
Trait de séparation entre paragraphes	"transition"	5.8.4	
Trait de séparation entre texte et notes de bas de page	"footnote"	5.8.4	
Ecriture manuscrite	"manuscript"	5.8.5	

6. QUALITE DE LA RECONNAISSANCE OCR

6.1 Qualité de la segmentation

Le résultat de la segmentation doit permettre de faire correspondre le texte issu de l'OCR à l'image par transparence grâce au calcul des coordonnées de la position des éléments dans l'image.

Pour atteindre le taux qualité attendu, il est souvent nécessaire de corriger (structuration, typage, …) le résultat proposé par le moteur OCR.

Ce taux est mesuré et renseigné à l'échelle de chaque page ainsi qu'à l'échelle du document. C'est la valeur au document numérique qui fait foi sur les marchés.

6.2 Montée en qualité du texte

La montée en qualité du texte a pour objectif d'atteindre un certain taux de reconnaissance des mots composant le document d'origine.

Ce taux est mesuré et renseigné à l'échelle de chaque page ainsi qu'à l'échelle du document. C'est la valeur au document numérique qui fait foi sur les marchés.

Les mots reconnus par le moteur OCR sont marqués avec un critère de confiance calculé par le moteur OCR, selon le degré de reconnaissance du mot, en utilisant l'attribut wc de l'élément String.

Ce critère de confiance varie de 0 (mot non reconnu) à 0,99 (mot reconnu sans doute).

Chaque mot corrigé par un opérateur sera identifié dans le contenu ALTO (par exemple en positionnant à 1,0 le critère de confiance WC).

6.2.1 Correction ciblée

Pour arriver au taux qualité attendu, le prestataire devra traiter prioritairement les mots dit « importants ». Ces mots importants se caractérisent par des caractéristiques typographiques ou intellectuelles, notamment :

- les mots avec une majuscule à l'initiale,
- les mots composés en majuscules,
- les éléments en gras ou en italique,
- les mots présents dans les titres,
- les entités nommées : noms propres, noms de lieu, noms de personne ou d'institution, etc.

OBLIGATOIRE

Chaque mot identifié comme important sera repéré dans le contenu ALTO (par exemple à l'aide de l'attribut TYPE).

6.3 Déqualification de contenus dans un document

Il s'agit d'une opération d'identification des blocs de texte (TextBlock) ou des mots (String) difficilement océrisable, du fait de leurs caractéristiques propres. Ces zones de texte sont donc exclues de fait du périmètre du taux qualité garantie et elles sont traitées en OCR brut.

Cette identification se fait selon deux axes :

- par types de contenus,
- par lisibilité des contenus.

6.3.1 Déqualification par types de contenu

Cette déqualification s'appuie sur les résultats de la phase de segmentation/structuration, qui a conduit à typer certains contenus de manière objective, selon leur nature physique ou logique.

Ces critères discriminants quant aux types des contenus sont les suivants :

- textes dont la langue majoritaire n'est pas le français (cf. section 4.1.3),
- textes composés en polices manuscrites (cf. section 4.2.3),
- textes composés dans des polices non reconnues par le moteur OCR (cf. section 4.2.3),
- tableaux à traiter en OCR brut (cf. section 5.6.4),
- publicités (cf. section 5.6.8) : il s'agit de zones à caractère publicitaire qui présentent des éléments typographiques particuliers tels que polices curvilignes, logos, encadrés, etc.
- textes placés sous un tampon (cf. section 5.6.9).

Cette typologie peut faire l'objet d'un référentiel commun d'exemples type, qui sera alimenté d'un commun accord au fur et à mesure des cas spécifiques rencontrés.

Dans le cadre de l'amélioration continue des processus de production, cette opération pourra faire l'objet d'un processus plus automatisé. Dans ce cas, l'évolution du processus sera soumise à l'approbation de la BnF.

6.3.2 Déqualification des mots ou blocs illisibles

Les zones de texte illisibles (cf. section 5.6.10) sont définies non du fait de leur type logique mais à partir de leur nature physique :

- zones près de la reliure de l'ouvrage et présentant une forte courbure de l'image,
- zones affectées par une dégradation du support physique de l'œuvre, laquelle affecte la netteté, le contraste ou même l'intégrité des contenus

- zones affectées par un problème de transparence, de migration d'encre, etc.

Mots

Un mot est considéré comme illisible si un opérateur humain ne peut le déchiffrer à l'œil nu, ou s'il ne parvient à le faire qu'avec un fort degré de supposition et d'interprétation du contexte du texte.

La lisibilité est évaluée à partir de l'image binarisée ou de l'image source en cas de doute.

La déqualification opérée à l'échelle d'un mot consiste à typer le mot concerné (élément String) en donnant à l'attribut TYPE la valeur "illegible".

Blocs

L'illisibilité d'un bloc est définie par une valeur seuil du taux de confiance WC calculé à l'échelle du bloc (moyenne des WC de tous les mots du bloc). Cette valeur est fixée par le prestataire en fonction de son expérience du moteur OCR utilisé.

Un bloc intégralement composé de mots illisibles (au sens de la définition ci-dessus) sera également considéré comme illisible.

La déqualification opérée à l'échelle d'un bloc de texte consiste à typer le bloc concerné (élément TextBlock) en donnant à l'attribut TYPE la valeur "illegible".

Les mots inclus dans le bloc de texte n'ont pas à être typés "illegible".

6.3.3 Limites au principe de déqualification

Un document demandé par la BnF en taux qualité garantie et dont plus de 50 % du contenu textuel (relativement au nombre de mots) est déqualifié entre automatiquement dans le champ d'action du principe de déqualification du taux qualité, tel que décrit à la section 3.3.3. Il est alors livré par le prestataire en OCR brut.

6.4 Déqualification du taux qualité sur un document

Des demandes de déqualification en OCR brut peuvent être demandées (modalités à préciser pour chaque marché) a priori par le prestataire pour des documents dont les qualités physiques ne sont pas suffisantes pour atteindre la qualité garantie :

- les documents tâchés, bruités ou maculés (ex. des microfiches, microfilms) ;
- les documents à contraste insuffisant (fond foncé/tramé avec superposition de texte, ex : livre d'enfants) ;
- les documents avec forte transparence (ex : presse, dictionnaire, ouvrage à papier pelure) ;
- les documents dont la langue majoritaire n'est pas en taux garanti (latin, etc.).

Ces demandes de déqualification se font avant tout prétraitement OCR, sans utiliser le mécanisme de déqualification des contenus exposé à la section 6.2.

6.5 Déqualification ou refus de documents

Les documents connus pour être quasi intégralement inexploitables en OCR (documents manuscrits, composés en polices Fraktur ou gothique, avec un alphabet non latin, etc.) doivent être refusés ou déqualifiés en OCR Brut par le prestataire pour la prestation OCR en fonction du type de marché.

Ces documents se reconnaissent à ce qu'ils se transcrivent sous la forme de blocs de texte vides ou très fortement bruités.

7. CONTROLE DE LA QUALITE

Le contrôle de la qualité est assuré par plusieurs moyens :

- des contrôles automatiques appliqués sur les contenus au format ALTO,
- un contrôle par échantillonnage visuel.

Au terme du contrôle, la BnF prononce le rejet ou l'acceptation des documents livrés par le prestataire.

Cette section décrit les critères de rejets ou d'acceptation

7.1 Contrôle automatique ALTO

Un contrôle automatique exhaustif de format est appliqué sur tous les fichiers ALTO avant le passage en contrôle par échantillonnage visuel. Ce contrôle émet des erreurs (standard ou majeure) ainsi que des avertissements.

SI CE CONTROLE EMET UNE ERREUR MAJEURE SUR UNE PAGE ALTO D'UN DOCUMENT, L'ENSEMBLE DES PAGES ALTO DU DOCUMENT SONT ECARTEES ET NE PASSENT PAS EN CONTROLE PAR ECHANTILLONNAGE VISUEL. LE DOCUMENT EST DONC REJETE DES CETTE ETAPE.

Ces contrôles automatiques sont de plusieurs natures :

- nommage du fichier ALTO,
- validation des fichiers ALTO relativement au schéma XML ALTO,
- présence et format des identifiants et des différents attributs,
- positionnement des blocs relativement aux dimensions de la page
- chevauchement entre des blocs de même niveau,
- format des éléments de production : opérations, agents, résultats (schéma XML detailsOperation.xsd)

Les modalités de ce contrôle seront détaillées dans une charte de contrôle OCR/ALTO élaborée conjointement par la BnF et le prestataire.

7.2 Contrôle par échantillonnage visuel

Le contrôle par échantillonnage visuel opère sur des lots de documents constitués selon un plan d'échantillonnage adapté à chaque marché. Il vise à contrôler deux aspects du traitement OCR :

- la qualité de la segmentation/structuration,
- la qualité de la reconnaissance du texte.

7.2.1 Qualité de la segmentation/structuration

Le contrôle de la segmentation/structuration s'applique à contrôler dans une page ALTO les éléments PrintSpace, xxxMargin, TextBlock, TextLine String, Illustration, GraphicalElement et ComposedBlock.

La granularité du contrôle est la page. Le niveau de qualité de segmentation/ structuration acceptable est spécifique à chaque marché. Il indique donc le nombre maximum de pages non conformes dans un document.

ATTENTION

LA DETECTION D'UNE NON-CONFORMITE SUR UN DOCUMENT ENTRAINE LE REJET DES DOCUMENTS COMPOSANT LE LOT DE CONTROLE.

LA DETECTION D'UNE NON-CONFORMITE STRUCTURELLE SUR TOUS LES DOCUMENTS COMPOSANT LE LOT DE CONTROLE ENTRAINE UN AUDIT DU PROCESSUS DE PRODUCTION ET EVENTUELLEMENT LE REJET DE TOUS LES DOCUMENTS DEJA PRODUITS.

Le contrôle de la segmentation/structuration se décompose en plusieurs contrôles dédiés :

- le typage des pages,
- l'ordre de lecture,
- le typage des blocs,
- le chevauchement des blocs,
- l'oubli de blocs,
- la reconnaissance des titres (si une tâche de ce type est attendue, cf. section 5.6.2).

Chacun de ces contrôles doit satisfaire le taux qualité attendu, ce qui implique :

- que l'on ne procédera pas à une moyenne des contrôles ;
- qu'un seul contrôle en échec entraîne la non-conformité de la page.

Typage des pages

Le contrôle vérifie le typage des pages et de leur orientation, détaillé sections 5.1 et 5.2.

Ordre de lecture

Le contrôle vérifie que l'ordre de lecture logique de la page est bien respecté (cf. section 5.5).

Segmentation et typage des blocs

Le contrôle vérifie :

- la bonne segmentation des blocs :
  - répartition en XxxMargin ou en PrintSpace (par ex. un contenu récurrent ou répétitif hors xxxMargin),
  - taille et position,
  - règles de composition des blocs composés,

- découpage en paragraphes,

- leur typage (cf. sections 5.6 à 5.8) :
  - nature des blocs (par ex. confusion blocs texte/illustration),
  - déqualification des blocs (par ex. un bloc illisible abusif).

Chevauchement des blocs

Un contrôle visuel du chevauchement est réalisé.

Oubli de blocs

Un contrôle visuel identifie les blocs éventuellement oubliés lors du traitement de segmentation.

Reconnaissance des titres

Le niveau de qualité acceptable concernant l'extraction des titres (lorsque cette tâche est demandée) se caractérise par deux métriques spécifiques :

- taux de détection des titres,
- taux de précision de la détection.

Lors du contrôle, les valeurs constatées dans la page sont comparées aux valeurs attendues pour ces deux taux, valeurs qui sont spécifiques à chaque marché.

Le mode opératoire de ce contrôle et le calcul du taux qualité seront détaillés dans une charte de contrôle OCR/ALTO élaborée conjointement par la BnF et le prestataire.

7.2.2 Qualité de la reconnaissance du texte

La qualité de la reconnaissance du texte est mesurée au mot. Une erreur correspond à tout mot erroné (conformité par rapport à l'image d'origine), quel que soit le nombre de signes erronés qu'il contient.

Pour un document, elle se calcule sur la population des mots présents dans les parties lisibles et non déqualifiées de toutes les pages du document (déqualifiées par nature, par un opérateur ou par l'OCR).

LA DETECTION D'UN ECART ENTRE LA QUALITE ATTENDUE ET LA QUALITE CONSTATEE SUR UN DOCUMENT ENTRAINE LE REJET DES DOCUMENTS COMPOSANT LE LOT DE CONTROLE.

LA DETECTION D'UN ECART ENTRE LA QUALITE ATTENDUE ET LA QUALITE CONSTATEE SUR TOUS LES DOCUMENTS COMPOSANT LE LOT DE CONTROLE ENTRAINE UN AUDIT DU PROCESSUS DE PRODUCTION.

NOTES

Un document peut avoir des parties dont le taux de reconnaissance est supérieur ou inférieur à la qualité attendue, mais l'ensemble du document doit valider la qualité attendue.

De plus, ne seront pas comptabilisées comme erreur :

- les erreurs concernant les ponctuations et les caractères non alphanumériques (cf. section 4.1.3), sauf si ce sont des ponctuations générées à la place d'un caractère alphabétique ;
- les substitutions ou omissions de diacritiques (« e » au lieu de « é » par exemple), les substitutions ou omissions de casse (« A » au lieu de « a ») ;
- les erreurs sur les chiffres arabes.

Phase de test

En numérisation de masse, la qualité de l'OCR est calculée à l'aide d'un taux de confiance estimé (en l'absence de vérité terrain). Ce taux (en %) est donnée par la formule :

$$\textit{taux de confiance} = \sum (\text{WC } \textit{des mots en qualité garantie}) \,/\, \textit{cardinal des mots en qualité garantie}$$

Le taux de qualité réel (spécifié pour chaque marché de numérisation comme qualité garantie à atteindre) est donnée par la formule :

$$\textit{Taux réel} = \textit{cardinal des mots justes parmi les mots en qualité garantie} \,/\, \textit{cardinal des mots en qualité garantie}$$

Durant la phase de test, on évaluera l'écart moyen entre la qualité estimée (taux de confiance) et la qualité réelle (taux réel).

Cet écart sera utilisé en phase de production pour étalonner la valeur du taux de confiance (et donc l'effort de correction manuelle du texte nécessaire pour atteindre la qualité garantie) :

$$\textit{taux de confiance} \times \textit{coefficient d'étalonnage} \geq \textit{taux garanti}$$

OBLIGATOIRE

Les informations présentes dans les fichiers ALTO (cf. section 7.2.3) doivent permettre à la BnF de recalculer automatiquement les taux qualité fournis par le prestataire.

Phase de production

Le contrôle par échantillonnage visuel opère sur des lots de documents constitués selon un plan d'échantillonnage.

Le contrôle de la qualité de l'OCR porte sur l'écart entre la qualité attendue et la qualité constatée par la BnF.

NOTES

Les modalités de ce contrôle seront détaillées dans une charte de contrôle OCR/ALTO élaborée conjointement par la BnF et le prestataire.

7.2.3 Détail des métriques qualité

Dans le manifeste du document numérique (refNum ou METS)

Les indicateurs à fournir (à l'échelle du document) sont notamment :

- le taux OCR brut : le taux de confiance en sortie de l'OCR
- le taux OCR corrigé : le taux de confiance après correction manuelle
- le coefficient d'étalonnage utilisé (cf. section 7.2.2)
- le taux NQA moyen, exprimé par : *taux OCR corrigé* × *coefficient d'étalonnage*

Dans les fichiers ALTO

Les indicateurs à fournir sont notamment, pour chaque page ALTO :

- nombre de blocs déqualifiés (publicité, illisibles, etc.)
- nombre de mots déqualifiés (illisibles, courbures, etc.)
- nombre total de mots ALTO
- nombre total de caractères ALTO
- nombre de mots corrigés par un opérateur
- nombre de caractères contenus dans les mots corrigés
- nombre de mots décidés par un opérateur
- nombre de caractères contenus dans les mots décidés
- le taux OCR brut : le taux de confiance en sortie de l'OCR
- le taux OCR corrigé : le taux de confiance après correction manuelle