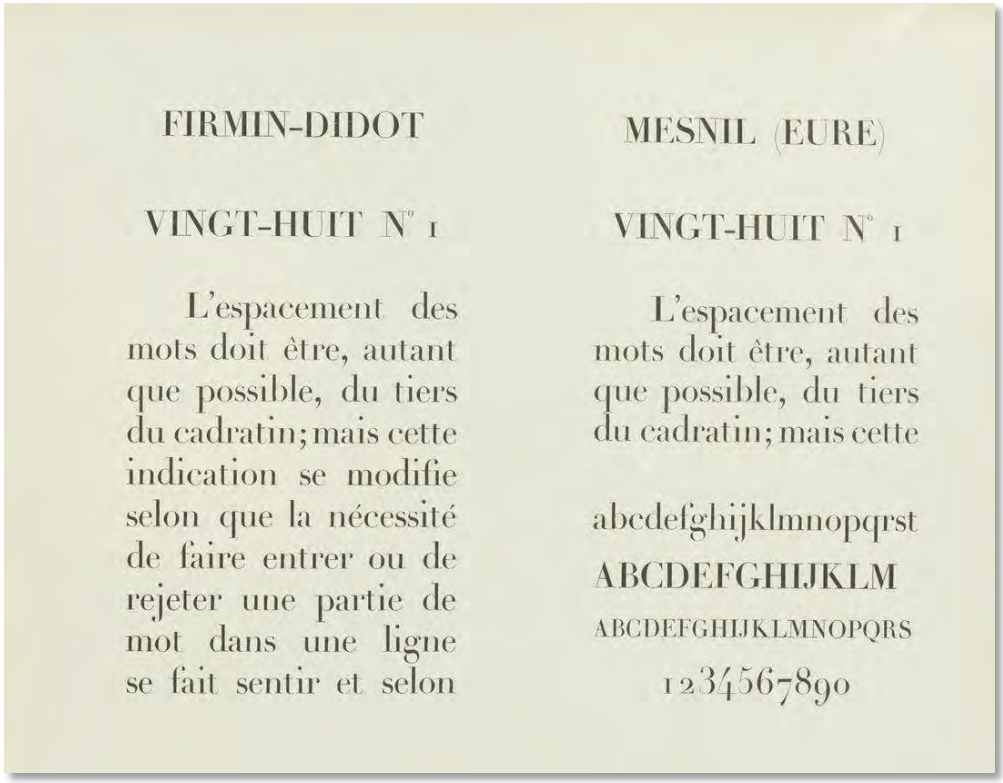
{BnF | Bibliothèque nationale de France

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 mardi 28 avril 2015 Version :2 Référence BnF :BnF-ADM-2014-062821-02

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 (version 3.0)
Schéma ALTO BnF	http://bibnum.bnf.fr/schema/alto_bnf-v2_0.xsd
BnF : Conversion en mode texte	http://www.bnf.fr/fr/professionnels/num_conversion_texte/s.num_conversion_texte_ocr.html
Référentiels BnF	http://www.bnf.fr/fr/professionnels/numerisation_boite_outils/a.numerisation_referentiels_bnf.html
Référentiel d'enrichissement des métadonnées	version 3 / BnF-ADM-2013-079381-04
Référentiel d'enrichissement des métadonnées METS	version 1 / BnF-ADM-2013-117422-01
Référentiel de livraison de document numérique	version 3 / BnF-ADM-2013-077351-03
Référentiel Tables	version 1 / BnF-ADM-2014-062888-01

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.

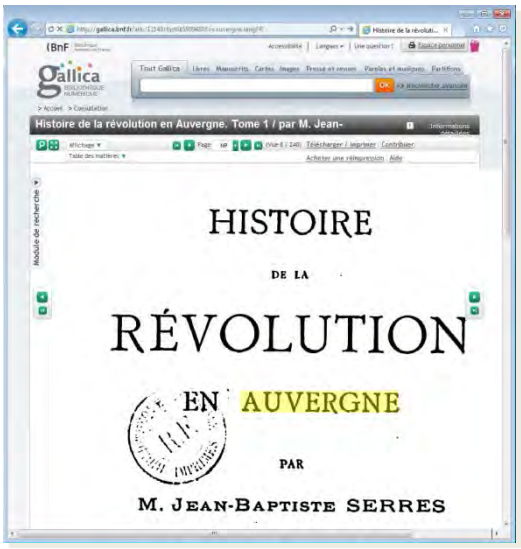
Ces systèmes OCR opèrent en plusieurs étapes :

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.

NOTES

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 restrictions dans le format ALTO originel et utilise donc une variante pour sa production. La nature de ces changements est documentée dans le schéma lui-même.

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 du format ALTO BnF

Le format est organisé en cinq grandes composantes :

- En-tête XML
- Description de la page et de son traitement
- Styles détectés dans la page
- Etiquettes présentes dans la page

- Segmentation de la page

Entête XML

L'élément <alto> donne les informations suivantes :

- Espaces de noms utilisés au sein du schéma ALTO BnF : xlink, ALTO LoC
- Version du schéma ALTO BnF : l'attribut SCHEMAVERSION donne la version du schéma utilisé, par exemple SCHEMAVERSION="alto_bnf-v2_0"

Description de la page et de son traitement

L'élément <Description> permet de fournir les informations suivantes :

- unité de mesure utilisée (<MeasurementUnit>) : pixel
- information sur l'image source océrisée :
    - nom du fichier image (<fileName>) : le motif imposé est \d{8}.(TIF|tif|JPG|jpg|jp2|JP2)
    - identifiant du fichier image dans la bibliothèque numérique, si cet identifiant est connu (dans le cas d'un traitement complémentaire) :

      <fileIdentifier>ark:/12148/bpt6k75043976/f1</fileIdentifier>
    - identifiant de production du document numérique : les deux catégories d'identifiants producteurs à utiliser sont « NUM » et « IFN » :

      <documentIdentifier documentIdentifierLocation="NUM">7504397</documentIdentifier>
- procédé de traitement OCR (<OCRProcessing>, <ocrProcessingStep>) :
    - date du traitement (<processingDateTime>) : format ISO AAAA-MM-JJ
    - information de production (<processingStepDescription>) : nombre de caractères, de chaînes (cf. section 7.2.3), types de traitement appliqués (par exemple correction des courbures de ligne), etc.
    - paramètres du traitement (<processingStepSettings>) : moteur OCR utilisé, version, réglages spécifiques, etc.

Styles détectés dans la page

L'élément <Styles> permet de décrire les styles détectés dans la page (cf. section 4.2).

Etiquettes présentes dans la page

L'élément <Tags> permet de décrire les étiquettes présentes dans la page (cf. section 5.10.1). Ces étiquettes caractérisent divers types de contenus : entités nommées, titres, etc.

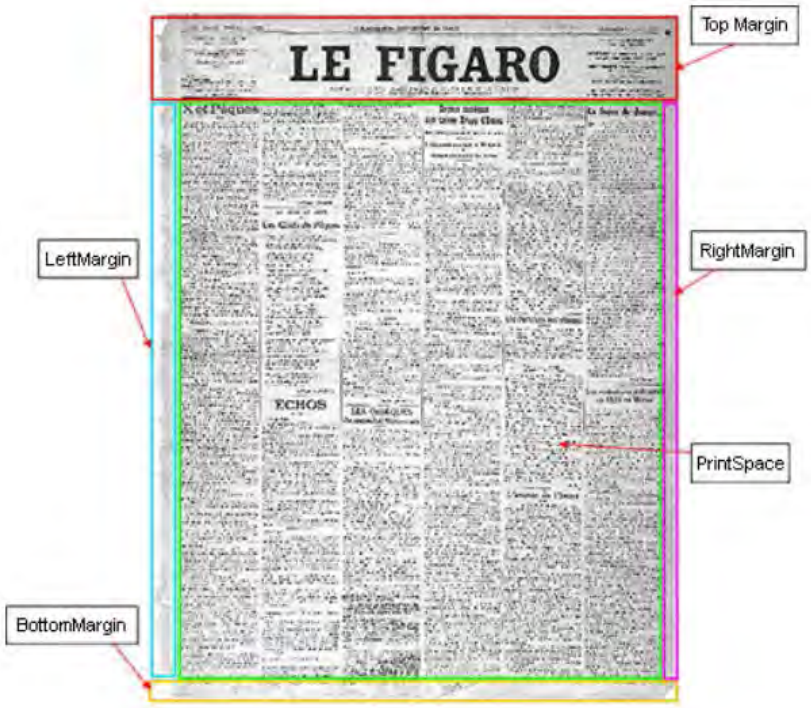
Segmentation de la page

Le format ALTO permet de décrire la segmentation d'une page (<Layout>) en ses différentes composantes. L'élément <page> peut ainsi contenir cinq composantes :

- 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 (*word dictionary,* optionnel) : appartenance ou non du mot 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.).

NOTES

Différents taux sont attendus selon les marchés (voir section 4.1.3) :

- OCR haute qualité : par exemple 99,9 %
- OCR taux qualité garantie : 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 évaluée 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 moteurs OCR, pour chaque mot traité par le logiciel.

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

ATTENTION

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. RECONNAISSANCE DU TEXTE

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.

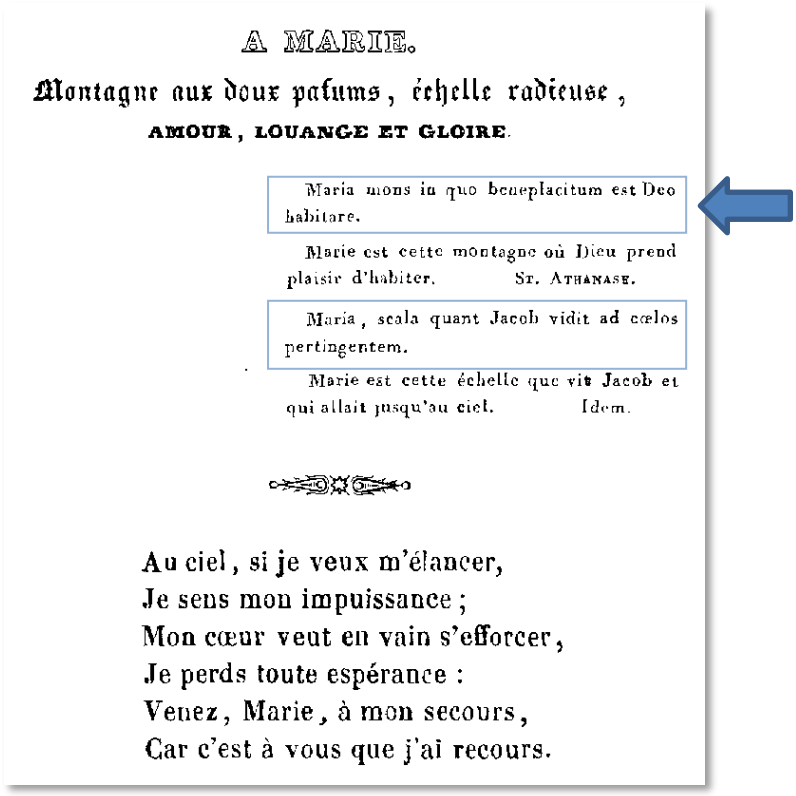
Tous les ouvrages sont traités par défaut comme étant en langue française, sauf détection par le moteur OCR d'autres langues.

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

```
LANG="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. Sr. Athanase.

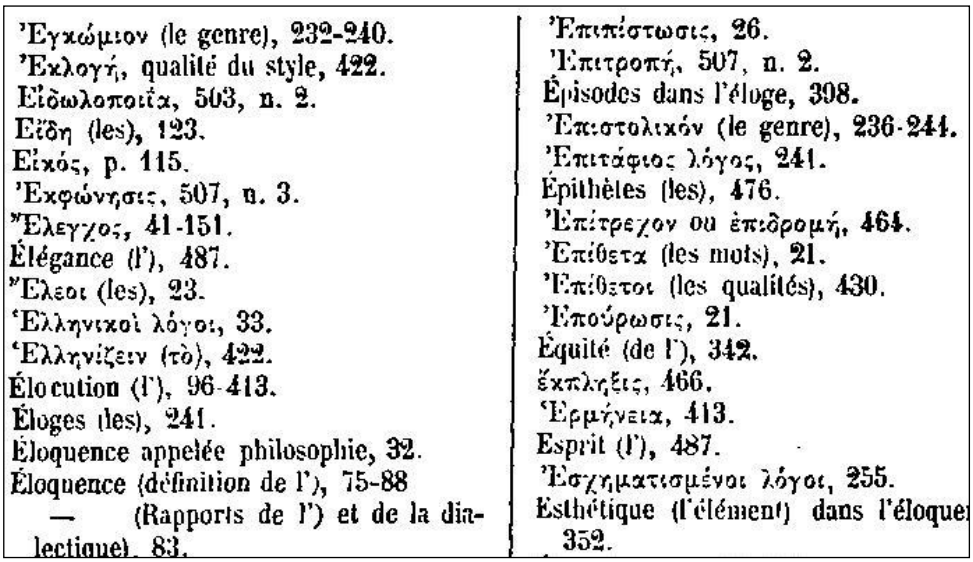
Maria, scala quant Jacob vidit ad cœlos pertingentem.

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

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 : 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'une étiquette LayoutTag LABEL="nonLatinScript" (cf. section 5.10).

EXEMPLE

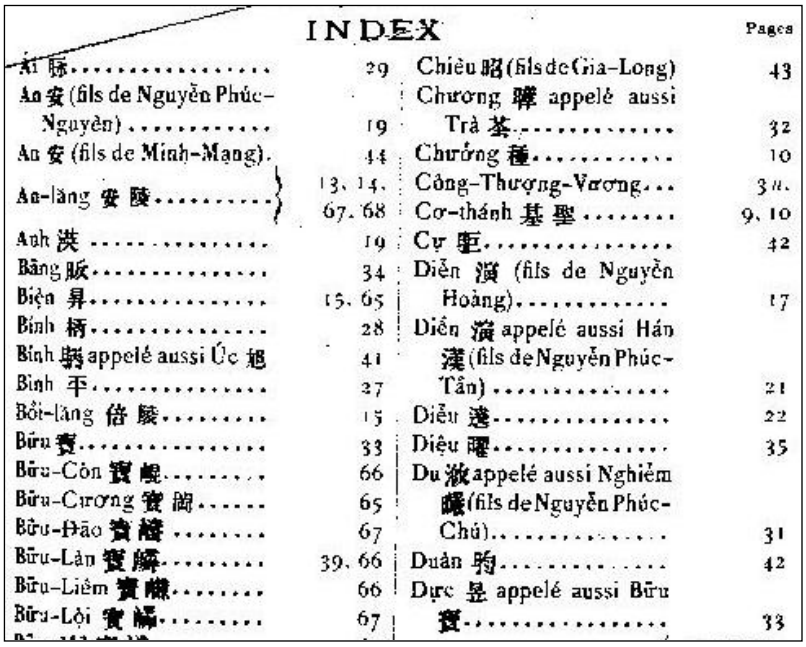
افرنجوى ٣٢, 17
افغورشاه ٣٣, 2 — ٣٩, 8
اكسيرخس ٣٩, 7
اكسيوثس ٣٩, 1
الثى ثودى ٣١٩, 11
القاء الحجارة ٣٢٨, 3. 8
امتلاء ١٧١, 9 — ١٧٣ — ١٧٥
املج ٨٣, 4
انحمرذكانيك ٣٣٧, 22
انوشيروان ٣٩, 11
الانيسان ٣٥١, 18
اهليلج ٨٣, 4
اوذرسانوس (?) ٢٥٨, 2

برتس بتارس ٣٩, 5
برخوشيا v. برخوشيا ٣٣١, 5
بركمومنس ٣٩, 1
بلاسوس ٣٩٢, 9
بلدة الثعلب ٣٥١, 17
بليليج ٨٣, 4
بليناس ٢٨٣, 18
بهارات ٣٠٩, 18
بوزنطيا ٣١٩, 5
بيت ١٣٨, 1 ff.
تابع النجم ٣٤٢, 15
تاسيس ٣٣٩, 22
التحتى ٣٢٢, 18 — ٣٥١, 8

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



INDEX

Pages

Ai 隘 29
An 安 (fils de Nguyễn Phúc-Nguyên) 19
An 安 (fils de Minh-Mạng). 44
An-lăng 安 陵 13, 14, 67, 68
Anh 洪 19
Băng 版 34
Biện 昇 15, 65
Bính 柄 28
Bính 炳 appelé aussi Úc 旭 41
Bình 平 27
Bồi-lăng 倍 陵 15
Bửu 寶 33
Bửu-Cồn 寶 崐 66
Bửu-Cương 寶 崗 65
Bửu-Đảo 寶 嶋 67
Bửu-Lân 寶 麟 39, 66
Bửu-Liêm 寶 廉 66
Bửu-Lội 寶 嵐 67

Chiêu 昭 (fils de Gia-Long) 43
Chương 曄 appelé aussi Trà 茶 32
Chưởng 種 10
Công-Thượng-Vương ... 3 n.
Cơ-thánh 基 聖 9, 10
Cự 聖 42
Diễn 演 (fils de Nguyễn Hoàng) 17
Diễn 演 appelé aussi Hán 漢 (fils de Nguyễn Phúc-Tần) 21
Diễu 遙 22
Diệu 曜 35
Du 淑 appelé aussi Nghiễm 嚴 (fils de Nguyễn Phúc-Chú) 31
Duân 駒 42
Dực 昱 appelé aussi Bửu 寶 33

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

4.1.3 Taux qualité OCR

Taux OCR brut

Le taux OCR brut est celui obtenu en sortie du moteur OCR. Il est calculé par le moteur OCR.

Ce taux est inscrit dans le manifeste du document numérique et dans le fichier ALTO chacune de ses pages (cf. section 7.2.3).

Ce taux peut s'appliquer à tout le document ou à des portions de document (dans le cas de zones déqualifiées, cf. section 6.3).

Taux OCR qualité garantie ou taux OCR haute qualité

Ces taux sont ceux attendus pour une prestation OCR avec des exigences particulières portant sur la qualité de la transcription du texte. Ces prestations impliquent une montée en qualité manuelle du texte pour les documents dont l'OCR brut est inférieur au taux qualité ciblé.

Périmètre

Tous les blocs de texte sont comptabilisés dans l'évaluation du taux OCR qualité garantie du document, exceptés :

- les blocs en langue autre que le français,
- les blocs déqualifiés (voir section 6.3)

Jeu de caractères

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. Ils constituent le « jeu de caractères hors 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								÷								

De plus, ne seront pas comptabilisées comme erreur :

- les erreurs concernant la reconnaissance des caractères accentués : substitutions ou omissions de diacritiques (« e » pour « é » par exemple) ;

- les erreurs concernant la casse : substitutions ou omissions de casse : « A » pour « a »
- les erreurs de reconnaissance sur les chiffres arabes : « 10 » pour « 16 ».

Par contre, sont considérées comme des erreurs :

- un caractère alphabétique transcrit erronément en un caractère hors garantie : « av1on » pour « avion »
- un caractère hors garantie collé à un mot transcrit erronément en un caractère alphabétique : « avioni » pour « avion; »
- un caractère hors garantie intramot non transcrit : « nétant » pour « n'étant »
- un caractère hors garantie intramot transcrit en un caractère alphabétique : « niétant » pour « n'étant »
- l'insertion parasite d'un caractère hors garantie dans un mot : « situe'e » pour « située »

NOTES

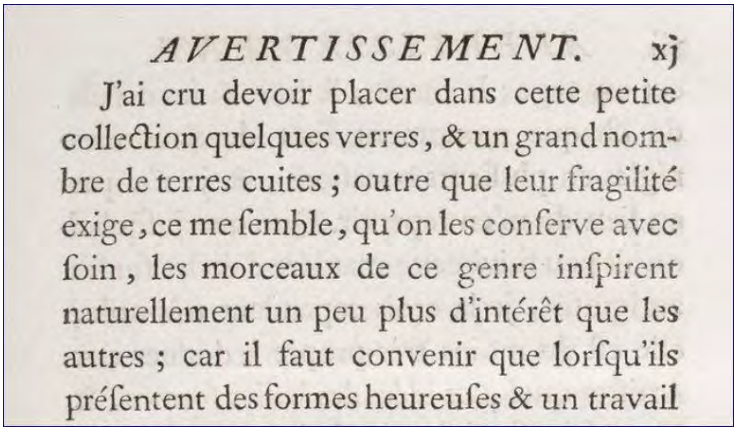
Une charte précisera ces règles en début de prestation OCR.

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.

Des taux OCR spécifiques peuvent s'appliquer.

EXEMPLE

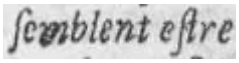


AVERTISSEMENT. xj

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.

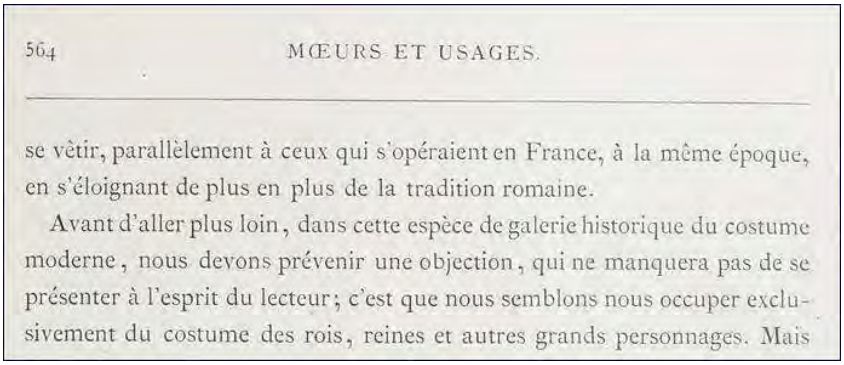
Glyphe	Exemple	Translittération
f/s		semblent estre
ff/ss		Palissades

NOTES

Il n'est pas demandé de dissimilation des lettres i/j et u/v.

Les autres ouvrages appartiennent à la catégorie générique « 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.

Taux OCR qualité éditoriale

Selon les prestations, ce taux pourra être exigé sur des zones spécifiques des documents (par exemples les titres). Il diffère des taux qualité garantie ou haute qualité par les règles suivantes :

- périmètre :
  - tous les blocs de texte en alphabet latin,
  - seuls les mots ou blocs illisibles sont exclus,

- jeu de caractères :
    - tous les caractères ASCII font partie du jeu de caractères garanti
    - la casse et les caractères accentués sont pris en compte.

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 –.

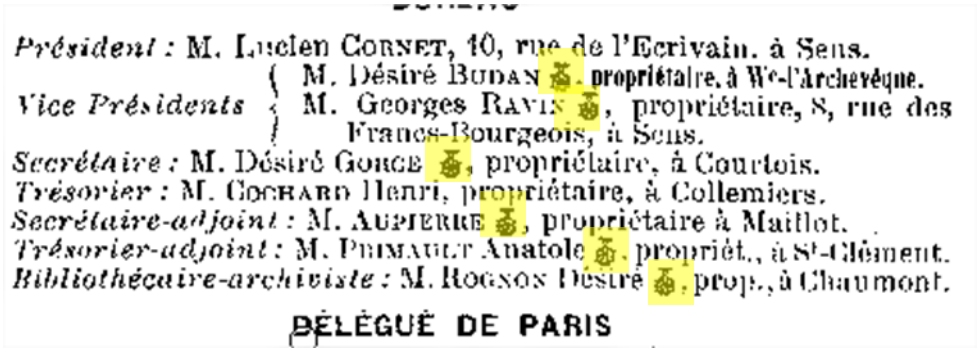
Les tirets de dialogue, les tirets longs (en général un tiret cadratin) sont à produire dans le flot de texte avec le caractère Unicode —.

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



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 (élément <Styles>).

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, l'éventuel 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.

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 Exposant, indice

Les mots composés en exposant ou indice seront identifiés par le mécanisme des styles.

4.2.3 Titres

Les titres ne sont pas identifiés en tant que tel, sauf si une tâche de reconnaissance des titres est demandée (cf. section 5.6.2).

4.2.4 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 5.6.9 et 6.2.2.

Ecriture manuscrite

Ce cas est décrit aux sections 5.7.5 et 5.8.5.

Polices script

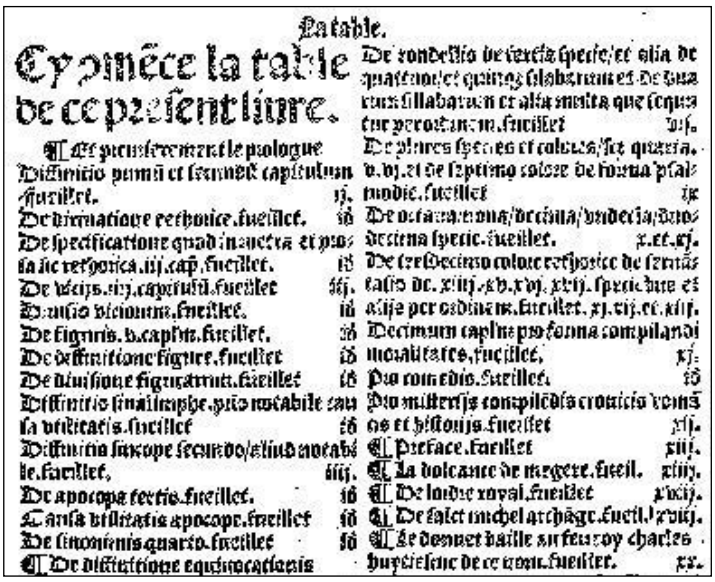
Les blocs de texte composés avec une police script sont décrits avec un élément TextBlock doté d'une étiquette LayoutTag LABEL="scriptFonts" (cf. section 5.10).

EXEMPLE

Autre cas de polices non exploitables

Ces blocs de texte sont décrits avec un élément TextBlock doté d'une étiquette LayoutTag LABEL="illegible" (cf. section 5.10).

EXEMPLE



La table.

Cy pmẽce la table de ce pzẽsent liure.

¶ Et premierement le prologue ... ffueillet. ij.
Diffinitio primi et secundi capitulum ... ffueillet. ij.
De diuinatione rethorice.fueillet. iõ
De specificatione quadrinarie et primi ... ta sic rethorica.iij.cap̃.fueillet. iõ
De vicijs.iij.capitulũ.fueillet iij.
Diuisio vitiorum.fueillet. iõ
De figuris.b.capm̃.fueillet. iõ
De diffinitione figure.fueillet iõ
De diuisione figurarum.fueillet iõ
Diffinitio similitudinis.pulcra notabile sau ... la vtilitatis.fueillet iõ
Diffinitio syncope secundo/aliud notabi ... le.fueillet. iiij.
De apocopa tertio.fueillet. iõ
¶ Ansa vtilitatis apocope.fueillet iõ
De synonimis quarto.fueillet iõ
¶ De diffinitione equiuocationis ...

De rondellio de tercia specie/et alia de quarta/et quinq; filiabarum et De tua rum filiabarum et alia multa que sequun 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/undecima/duo decima species.fueillet x.et.xj.
De tredecimo colore rethorice de senten tatis dc. xiiij.xv.xvj. xvij. species huius et alije per ordinem.fueillet. xj.et.xiij.
Decimum capm̃ pro forma compilandi moralitates,fueillet, xj.
Pro comedis.fueillet. iõ
Pro multeriis compilãdis cronicis roma nis et hystorijs.fueillet xij.
¶ Prefface.fueillet xiij.
¶ La dolcance de megere.fueil. xiij.
¶ De laude royal.fueillet xvij.
¶ De saict michel archãge.fueil.xvij.
¶ Le dosnet baille 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,"/>
```

1. Dans le cas d'une double césure (par ex. « a-bracada-bra »), on utilisera un SUBS_TYPE="HypPart1" suivi de deux SUBS_TYPE="HypPart2".

2. Il n'est pas demandé de distinguer entre tiret de césure et trait d'union. Un mot composé placé sur deux lignes pourra donc être traité avec une césure.

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).

5. SEGMENTATION ET STRUCTURATION

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

Un taux qualité spécifique au marché (cf. CCTP) 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

5.1.1 Numérotation des pages

L'attribut PHYSICAL_IMG_NR (obligatoire) de l'élément Page permet de renseigner le numéro de la page dans l'ordre séquentiel des pages du document.

L'attribut ID de l'élément Page représente le numéro d'ordre dans la numérotation des fichiers ALTO :

```
<Layout>
  <Page ID="PAG_00000002" HEIGHT="3353" WIDTH="2065"
      PHYSICAL_IMG_NR="2" > …
  </Page>
</Layout>
```

CET ATTRIBUT EST OBLIGATOIRE.

L'attribut PRINTED_IMG_NR (optionnel) de l'élément Page permet de renseigner les informations de page/folio ou les éléments descriptifs de couverture/couvrure, tel qu'imprimé sur la page (cf. « Référentiel d'enrichissement des métadonnées », section 3.3 pour les règles de valorisation de cet attribut).

```
<Layout>
  <Page ID="PAG_00000135" HEIGHT="3353" WIDTH="2065"
      PHYSICAL_IMG_NR="135" PRINTED_IMG_NR="129" > …
  </Page>
</Layout>
```

Quand le fichier ALTO représente une double page, on utilisera la syntaxe *x-y* pour ces deux attributs. Pour une double page de folio 8 et 9, on aura donc :

```
<Layout>
```

```
  <Page ID="PAG_000008" HEIGHT="3353" WIDTH="4050"
      PHYSICAL_IMG_NR="8-9" PRINTED_IMG_NR="8-9" > …
  </Page>
</Layout>
```

5.1.2 Qualité de numérisation des pages

L'attribut QUALITY (obligatoire) de l'élément Page permet d'indiquer au lecteur l'état du document d'origine et en conséquence son résultat numérique. Cette information correspond aux commentaires de type « usager » portés au niveau du feuillet ou de la page concernée dans l'élément <vueObjet> du refNum (cf. « Référentiel d'enrichissement des métadonnées », section 4.1).

Sa valeur sera "OK" dans le cas courant (rien à signaler) ou "Damaged" : tous les autres cas tel que décrit dans le « Référentiel d'enrichissement des métadonnées ».

Dans ce dernier cas, l'attribut QUALITY_DETAIL de l'élément Page permet de préciser le type de défaut. Il contiendra la même valeur que l'élément <commentaire> de l'élément <vueObjet> du refNum :

ALTO :

```
<Layout>
  <Page ID="PAG_00000101" HEIGHT="3353" WIDTH="2065"
      PHYSICAL_IMG_NR="101" PRINTED_IMG_NR="94"
      QUALITY="damaged"  QUALITY_DETAIL="Report d'encre"> …
  </Page>
</Layout>
```

refNum :

```
<commentaire type="USAGER" date="2013-09-30T15:00:43Z">Report d'encre
</commentaire>
```

5.1.3 Qualité de l'océrisation des pages

L'attribut ACCURACY de l'élément Page permet d'indiquer le taux OCR estimé pour la page, exprimé en %. Cette valeur doit être comprise entre 0.0 et 100.0.

```
<Layout>
  <Page ID="PAG_00000135" HEIGHT="3353" WIDTH="2065"
      PHYSICAL_IMG_NR="135" PRINTED_IMG_NR="129"
      ACCURACY="98.380"> …
```

5.1.4 Pages vides

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

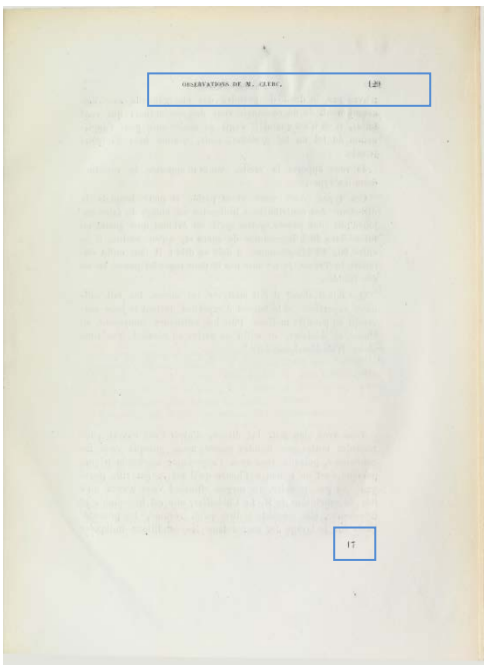
Les pages vides (sans aucun contenu) sont décrites avec PAGECLASS="BlankPage".

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

5.1.5 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



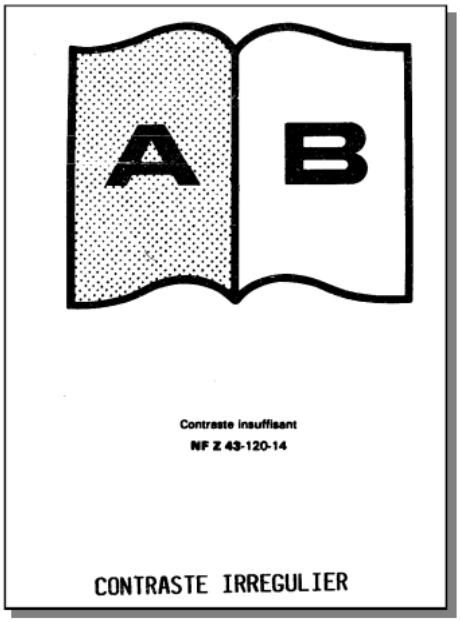
```
<Layout>
  <Page ID="PAG_00000135" PAGECLASS="BlankPrintSpace" HEIGHT="4153"
      WIDTH="3049" PHYSICAL_IMG_NR="135" PRINTED_IMG_NR="129"
      QUALITY="OK" > …

 <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" LANG="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" LANG="fr">
    <TextLine … </TextLine>
  </TextBlock>
  </BottomMargin>
  </Page>
 </Layout>
```

5.1.6 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.

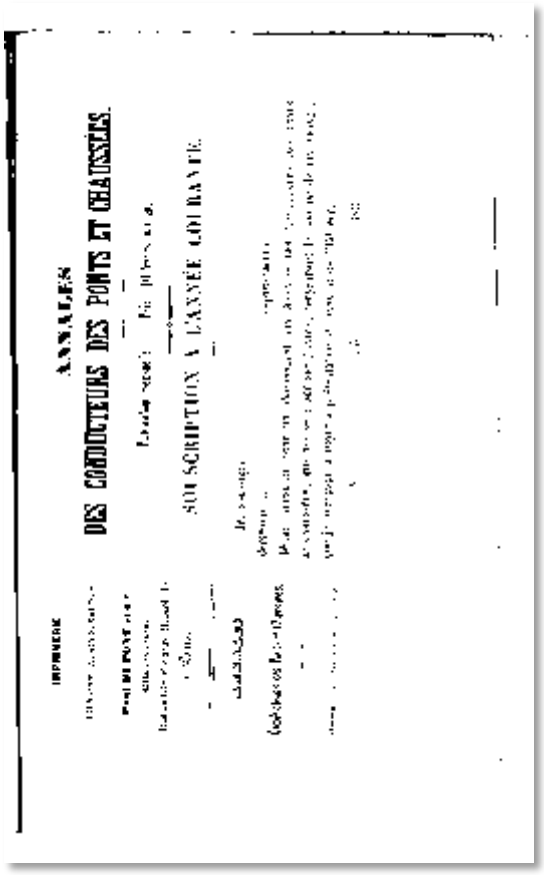
Si l'orientation de l'original (c'est-à-dire celle de la reliure) est différente de l'orientation de la lecture, l'angle de rotation des éléments de la page doit être spécifié dans le fichier ALTO à l'aide de l'attribut Rotation, selon un angle exprimé en degrés, dans le sens inverse des aiguilles d'une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

Deux cas peuvent se présenter :

- Si tous les éléments de la page sont dans une orientation de lecture donnée, alors en restitution, ces éléments comporteront un attribut Rotation. Dans l'exemple suivant, les composants de la page sont produits avec un attribut rotation à 90° :

EXEMPLE

```
<TextBlock ID="PAG_248_TB000001" STYLEREFS="TXT_1" HPOS="1996"
VPOS="1019" HEIGHT="1557"  WIDTH="66" LANG="fr" ROTATION="90">
```



Sens de la page originale : 90°

- 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 Ileifultz, 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 Coisance sort des grottes de Planchet-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 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.

Fig. 4.

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'Eblis. Prosterné le matin, le soir et durant les nuits, il récitait dévotement les versets les plus saints de l'Evidence, 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'Eternel.

» 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 recevoir un angle de rotation (90°)

ATTENTION

LE TRAITEMENT OCR DOIT ETRE APPLIQUE SUR TOUS LES BLOCS DE TEXTE, Y COMPRIS CEUX AVEC ROTATION, SAUF SI CELLE-CI N'EST PAS UN MULTIPLE DE 90°. LE BLOC CI-DESSOUS (ROTATION=45 °) PEUT NE PAS ETRE OCERISE.

CŒUR D'EUROPE

Un grand reportage d'actualité d'Andrée VIOLLIS

"Midi moins cinq, disent les Sudètes ! Au mois de juin il sera trop tard pour être avec nous"

SUITE DE LA PREMIÈRE PAGE

Leur insolence s'accroît de mois en mois, de jour en jour. Heuleja, comme son émule le

lui explique par la mimique la plus féroce quel sort l'attend : égorgement ou pendaison. Tel maire a reçu jusqu'à soixante lettres, enjolivées de caricatures ou de dessins de potences, lui enjoignant de démissionner. A telle institutrice juive, on annonce que ses écoliers feront grève jusqu'à ce qu'elle se décide à abandonner son poste. Et les menaces font leur effet. Un seul exemple : la veille du 1er mai, par une nuit très noire, 4.000 travailleurs suivaient la colonne aux flambeaux. Le lendemain, à la manifestation de jour, il n'en restait plus que 170 dont il faut admirer le courage. Encore était-ce l'an dernier. Depuis l'Anschluss, depuis que les deux partis activistes, les chrétiens-sociaux après les agraires, ont rejoint celui de Henlein, la situation ne cesse d'empirer.

nir à nous. Il y a déjà sept comités, des instructeurs militaires, des recrues. Mais ce qu'il faut, ce que nous essayons de constituer, dans toute la Tchécoslovaquie, comme ici, c'est un bloc des forces démocratiques qui ne veulent point pactiser avec l'ennemi. Notre front populaire...

— N'allez pas croire non plus, reprit un social-démocrate, qu'aucun de nous autres, antifascistes, soit partisan d'une politique de provocation, pouvant entraîner la guerre avec le troisième Reich. Nous savons trop quelles catastrophes cette guerre entraînerait... D'autre part, une politique de concessions excessives, de capitulation ne nous paraît pas moins dangereuse vis-à-vis de gens qui ne comprennent et n'admirent que

5.3 Structuration de la page

5.3.1 PrintSpace

Le contenu spécifique à la page doit être inclus dans le PrintSpace. 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.

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

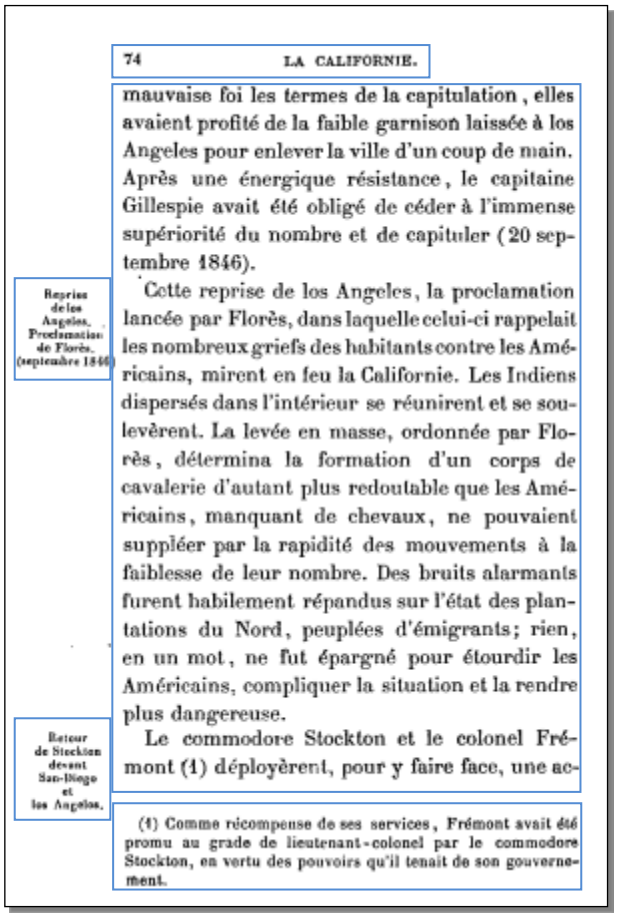
5.3.2 XxxMargin

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

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

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

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 de texte 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 une étiquette LayoutTag LABEL="missing" (cf. section 5.10).

NOTES

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.

Dans les fichiers XML ALTO, l'ordre de lecture est indiqué par la numérotation séquentielle des blocs, via leur attribut ID, ainsi que par leur ordre de présence dans le fichier XML.

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.

L'ordre de lecture au sein des zones de marges suit les mêmes règles.

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 verticaux de gouttière ne sont pas segmentés.

EXEMPLE

L'AMOUR
A LA MARÉCHALE,

COMÉDIE EN DEUX ACTES ET EN PROSE,

PAR Mme A. ROGER DE BEAUVOIR (Mlle DOZE),

Représentée pour la première fois, à Paris, sur le théâtre de la Montansier, le 14 août 1852.

DISTRIBUTION DE LA PIÈCE.

LE MARQUIS DE RIANCOURT. . . . . M. DERVAL.
LA MARQUISE DE RIANCOURT. . . . . Mlle PAULINE.
LE CHEVALIER D'AUBIGNY . . . . . M. VALAIRE.
MARIETTE. . . . . Mlle GALLOIS.
UN DOMESTIQUE DU CHEVALIER, UN DOMESTIQUE DU MARQUIS, PORTEURS.

ACTE PREMIER.

Le théâtre représente un boudoir rococo. — A droite, un canapé; à gauche, une toilette, porte au fond et porte latérale.

SCÈNE I.

LA MARQUISE, à sa toilette; MARIETTE.

LA MARQUISE, assise.
Que tiens-tu là, Mariette?
MARIETTE.
Une lettre pour madame la marquise.
LA MARQUISE.
Ah! la lettre quotidienne! toujours les mêmes phrases, les mêmes serments!... (Elle l'ouvre.) Pauvre chevalier! il tourne au pastoral, ce matin. Mariette, brûle cette lettre.
MARIETTE
Il y a quinze jours, madame la marquise m'aurait dit : « Mariette, va me chercher ma cassette », et elle y eût placé précieusement la lettre qu'elle jette aux flammes ce matin.
LA MARQUISE.
C'est vrai; mais si le présent ressemblait en tout au passé, la vie serait d'une mono-

La Marquise, Mariette.

tonie insupportable. N'est-ce pas assez de se lever, de se coucher, de déjeûner, de souper tous les jours aux mêmes heures, et ne faut-il pas varier ses idées et ses sentiments?
MARIETTE.
Madame la marquise me permet-elle de dire ce que je pense?
LA MARQUISE.
Oui certes, ne fût-ce que pour la rareté du fait.
MARIETTE.
Eh bien! madame la marquise s'ennuie.
LA MARQUISE.
C'est vrai.
MARIETTE.
Et cela, parce que madame n'aime plus.
LA MARQUISE.
Que veux-tu? l'amour est si difficile à rencontrer!... Est-ce que tu crois, Mariette, qu'il court les rues?
MARIETTE.
Non pas! quoiqu'il y ait terriblement d'a-

Double colonne type imprimé

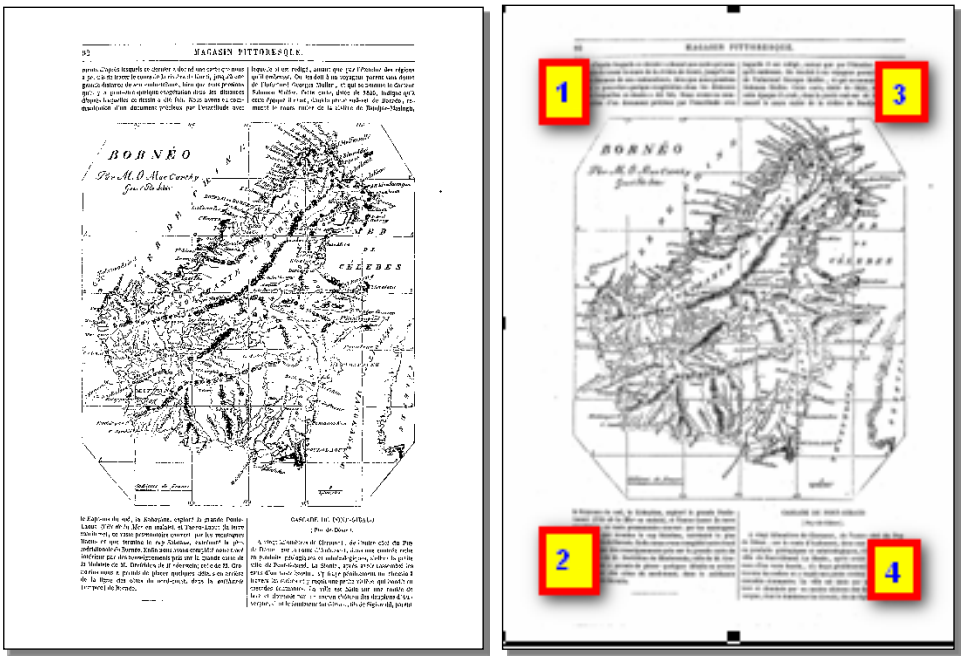
5.5.2 Mise en page en colonnes avec des éléments centraux non textuels

Un élément central non textuel est un élément non textuel (illustration, carte etc.) qui occupe (pleinement ou partiellement) 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

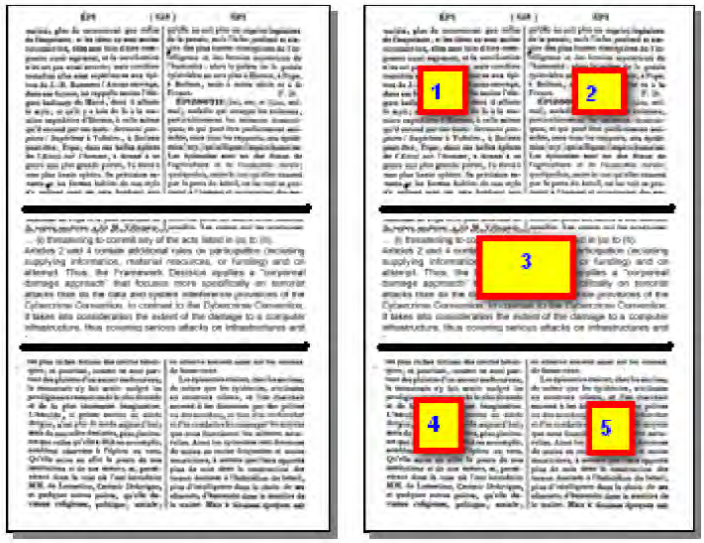
5.5.3 Mise en page en colonnes avec des éléments centraux textuels

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

En règle générale :

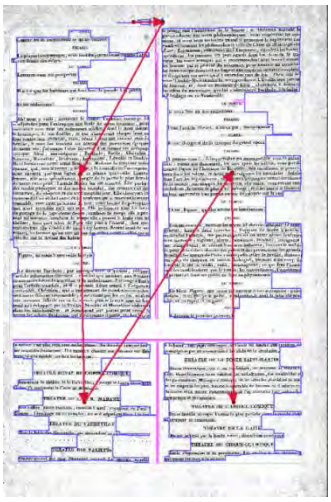
- En plus des règles sur les colonnes, les blocs dont la position verticale est avant un bloc textuel central, figurent avant ce bloc textuel central.

EXEMPLE



Ce cas est particulièrement courant pour les contenus presse et peut conduire à un ordre de lecture erroné.

EXEMPLE



Ordre de lecture erroné

Mise en page avec imbrication de colonnes

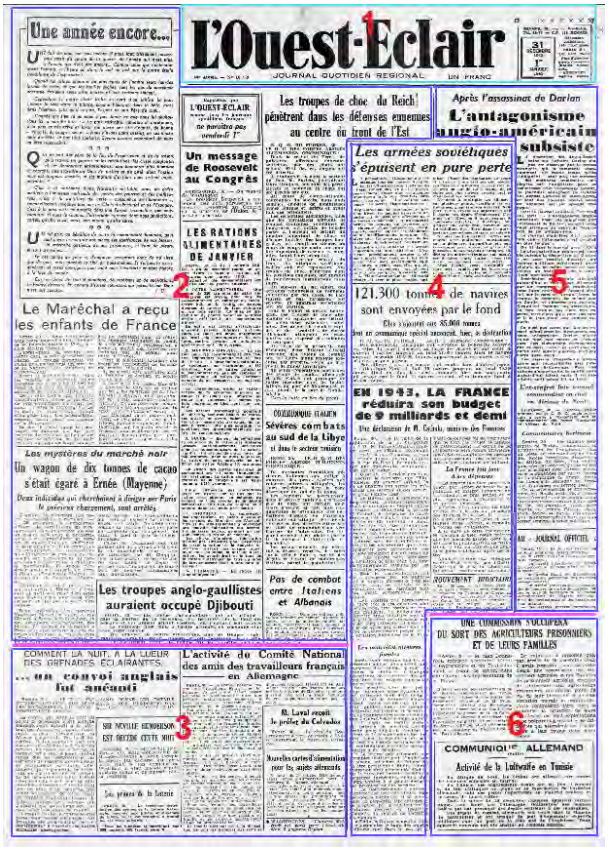
Les mises en page de type presse présentent des particularités :

- La zone de l'ours sera toujours placée en premier dans l'ordre de lecture, même si elle ne court pas sur toute la largeur de la page.

EXEMPLE

- L'ordre de lecture sera déterminé en appliquant les règles générales d'abord sur les zones de plus haut niveau, puis à l'intérieur de chaque zone.

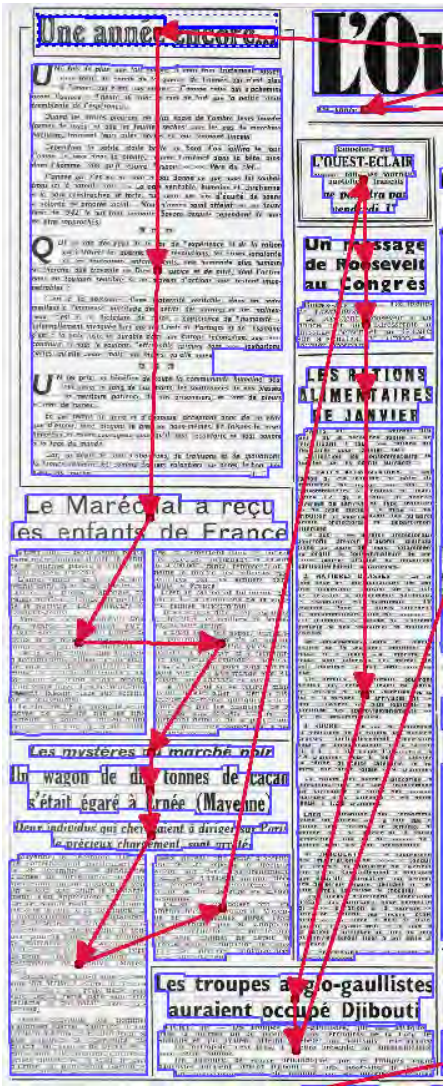
EXEMPLE



Identification des zones de plus haut niveau

http://gallica.bnf.fr/ark:/12148/bpt6k502978s

EXEMPLE



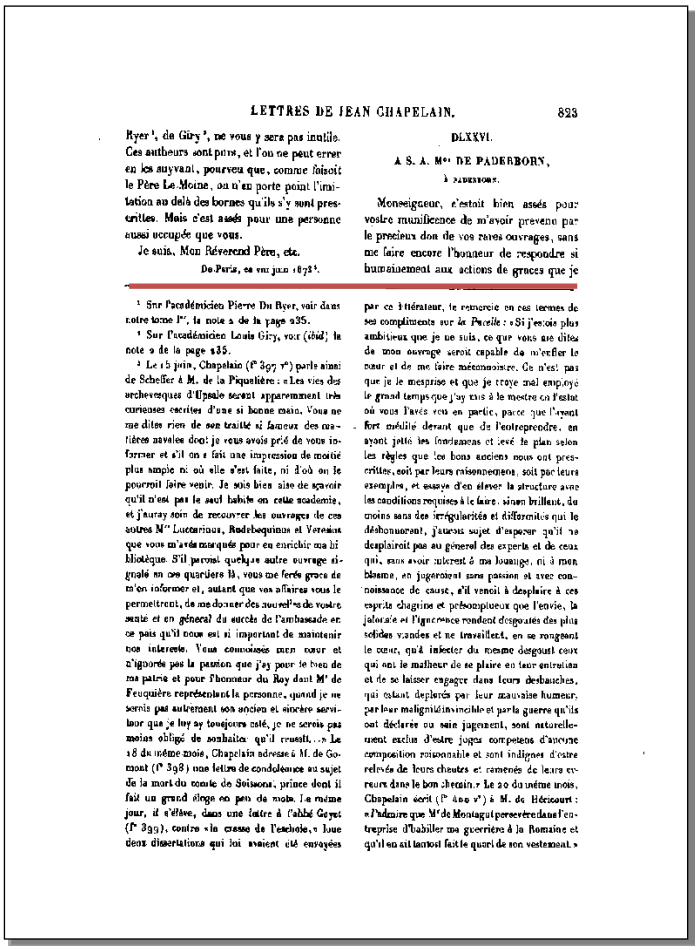
Ordre de lecture interne dans la zone n° 2

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 Révérend Père, etc.

De Paris, ce vnt 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 (Ms 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 Lucanius, Rudebequius 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 et de 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 représentoit la personne, quand je me servis par autrefois son ancien et sincère serviteur que je luy ay toujours esté, je ne serois pas moins obligé de souhaitter qu'il cressût, ..» Le 18 du même mois, Chapelain adresse à M. de Gomont (Ms 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 (Ms 399), contre « la cresse de l'eschole, » loue deux dissertations qui lui avoient 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 mécognoistre. 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'espérer qu'il ne desplairoit pas au général des experts et de ceux qui, sans avoir intérest 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 despoutes 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, tantost 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 jugés compétens 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 (Ms 400 vo) à M. de Héricourt : « J'admire que Mr de Montagut persévère dans l'entreprise d'habiller ses guerriers 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.

NOTES

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 Titres

Titres

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 le cas où une détection de titres est demandée, le niveau de titre est stocké par une étiquette StructureTag sur le TextBlock ou le TextLine concerné (cf. section 5.6.2 et 5.10).

La tâche de reconnaissance des titres est décrite en détail dans le « Référentiel d'enrichissement du texte ».

Titres de presse

Les titres de presse ainsi que les ours pourront être déqualifiés en OCR brut, du fait des polices ou des tailles de caractères utilisées.

EXEMPLE

5.6.3 Tableaux

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

NOTES

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 une étiquette LayoutTag LABEL="table" sur

l'élément TextBlock (cf. section 5.10). L'ordre de lecture à l'intérieur des différentes parties des tableaux est celui obtenu par le moteur OCR, 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 une étiquette LayoutTag LABEL="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.

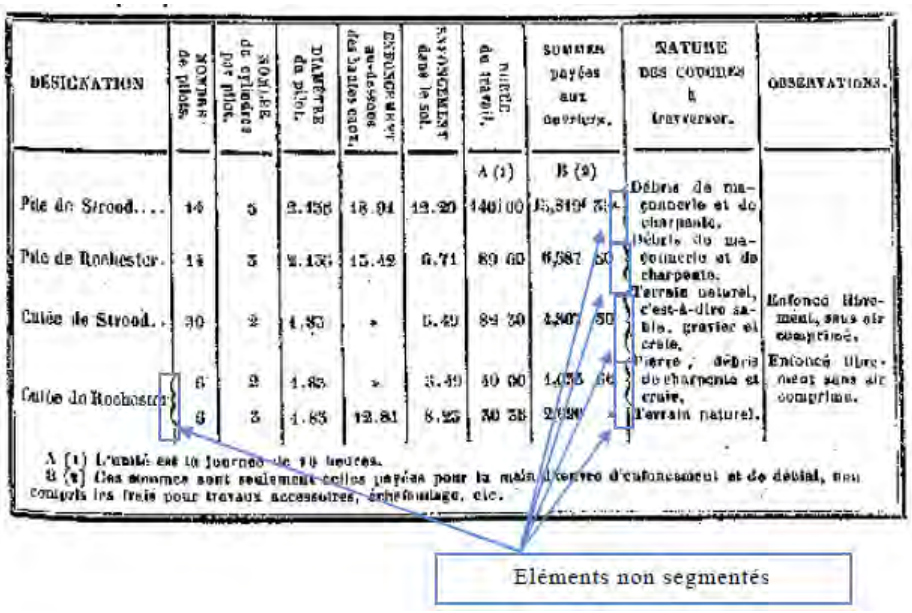
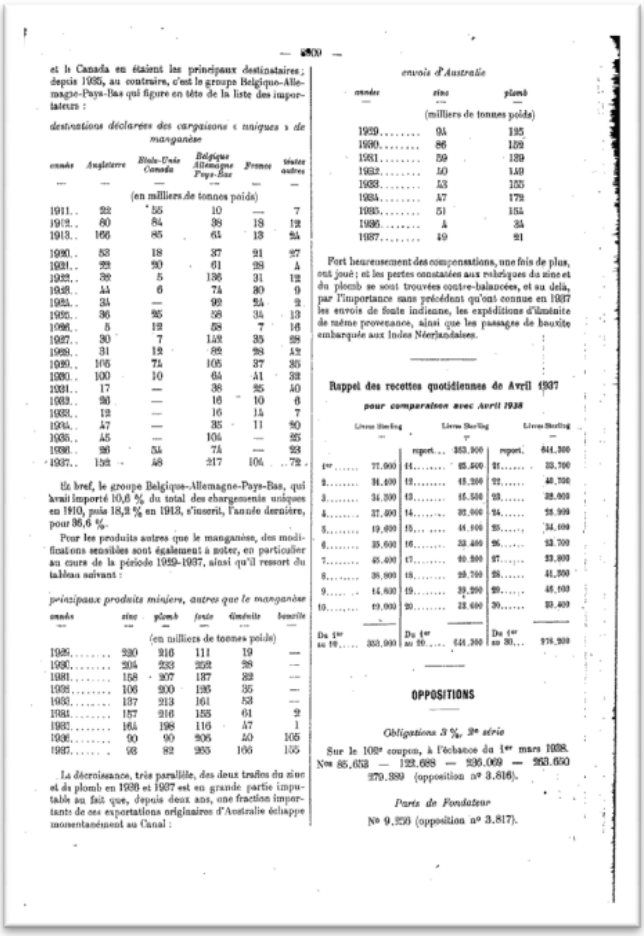


Tableau avec filet de séparation et accolades



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

année	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
1930	83	18	37	21	27
1931	22	20	61	28	4
1932	32	5	136	31	12
1933	14	6	74	30	9
1934	34	—	99	24	2
1935	36	25	58	34	13
1936	5	12	58	7	16
1937	30	7	142	35	28
1938	31	12	88	28	12
1939	105	74	105	37	25
1930	100	10	64	41	32
1931	17	—	36	25	10
1932	20	—	16	10	6
1933	12	—	16	14	7
1934	47	—	35	11	20
1935	45	—	104	—	25
1936	20	54	74	—	23
1937	128	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

année	zinc	plomb	fonte	tétrafère	bauxite
		(en milliers de tonnes poids)			
1928	220	216	111	19	—
1930	204	233	252	28	—
1931	158	207	187	22	—
1932	106	200	126	35	—
1933	137	213	161	53	—
1934	157	216	155	61	2
1935	164	198	116	47	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 ses expéditions originaires d'Australie échappe momentanément au Canal :

envois d'Australie

année	zinc	plomb
	(milliers de tonnes poids)	
1929	94	195
1930	86	192
1931	59	139
1932	50	140
1933	53	155
1934	57	172
1935	51	155
1936	4	34
1937	49	91

Fort heureusement des compensations, une fois de plus, ont joué ; et 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	11	65.800	21	33.700
2	34.400	12	43.200	22	40.700
3	34.300	13	45.500	23	28.900
4	37.200	14	32.000	24	55.900
5	19.600	15	41.100	25	34.100
6	35.600	16	33.400	26	32.700
7	45.400	17	40.200	27	33.900
8	36.800	18	29.700	28	41.300
9	14.600	19	32.200	29	46.100
10	19.000	20	23.600	30	23.400
Du 1er au 10	353.900	Du 1er au 20	641.300	Du 1er au 30	976.200

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

No 9.256 (opposition n° 3.817).

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

Taux qualité des tableaux

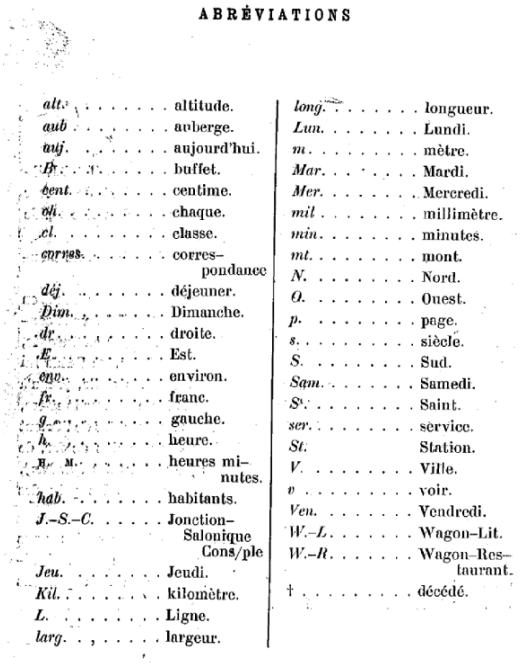
Le taux qualité des contenus tableau sera traité différemment pour les deux types de tableau suivants :

- 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**. Ils seront balisés :

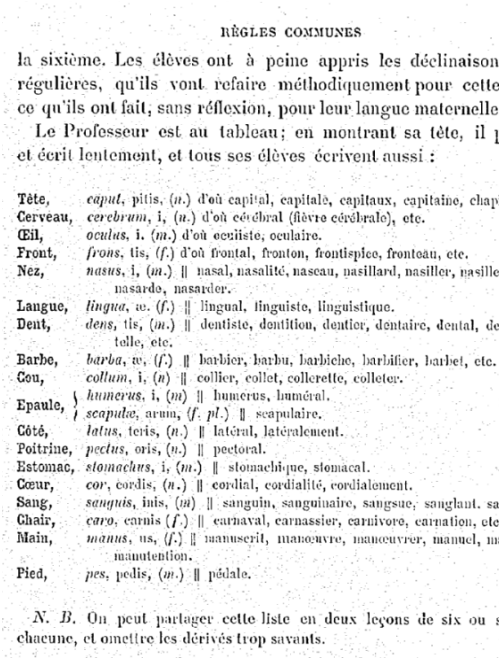
  TAGREFS="TAG_table"

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.



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, frontcau, 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.

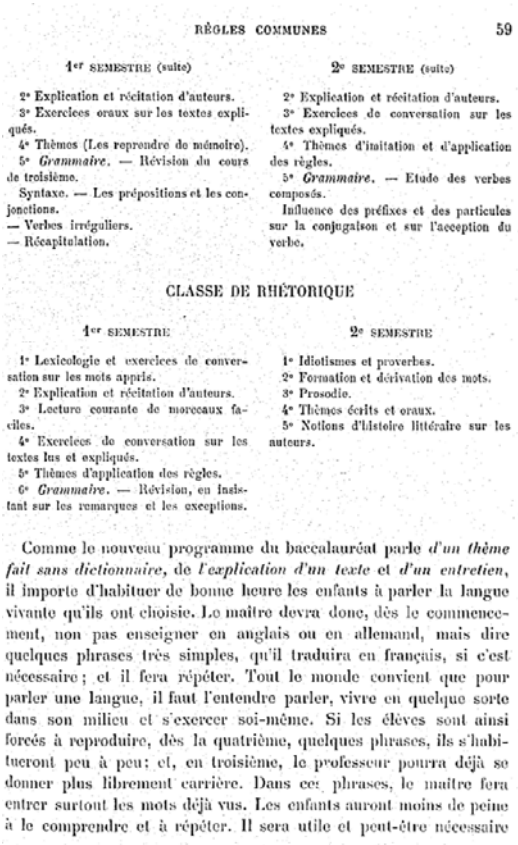
N. B. On peut partager cette liste en deux leçons de six ou chacune, et omettre les dérivés trop savants.

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

Glossaire, liste d'abréviations

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 LES 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**. Ils seront balisés :

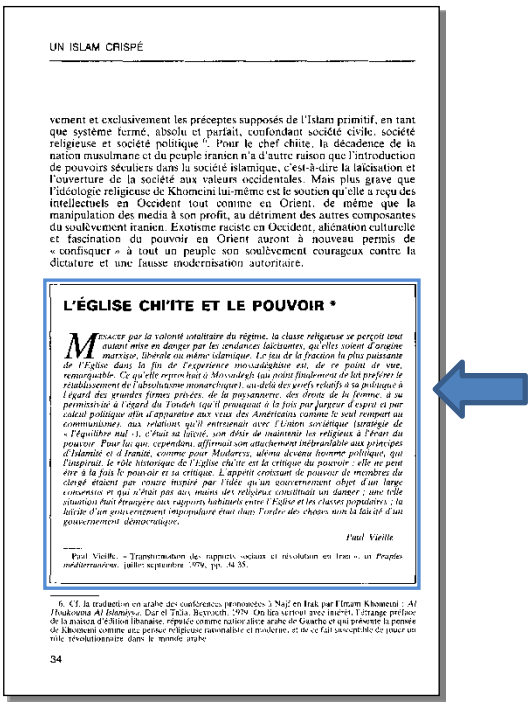
  TAGREFS="TAG_table TAG_raw"

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'Église 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, que, 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'Église 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 contre 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'Église 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.

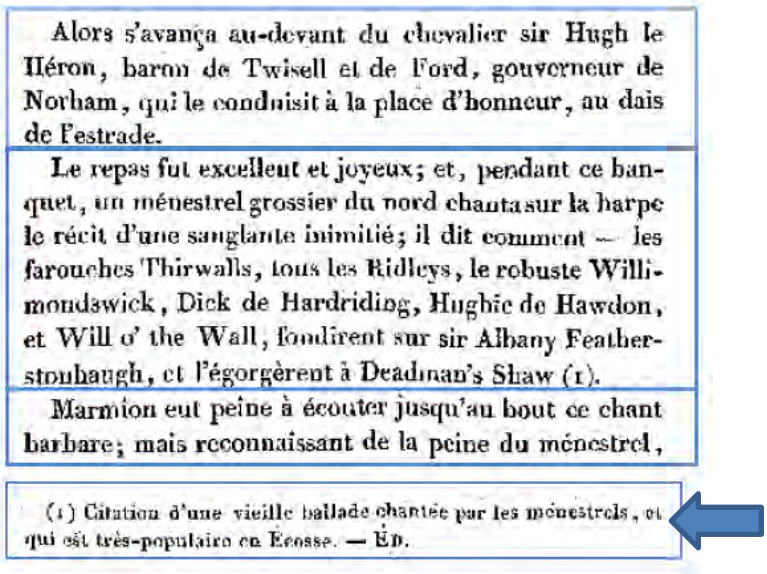
6. Cf. la traduction en arabe des conférences prononcées à Najf en Irak par Ibrahim Khomeini : *Al Thaqafa Al Islamiya*, Dar el Tali'a, Beyrouth, 1979. On lira surtout avec intérêt, l'étrange préface de la maison d'édition libanaise, réputée comme nation alice arabe de Gauche et qui présente la pensée de Khomeini comme une pensée religieuse rationaliste et moderne, et de 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).

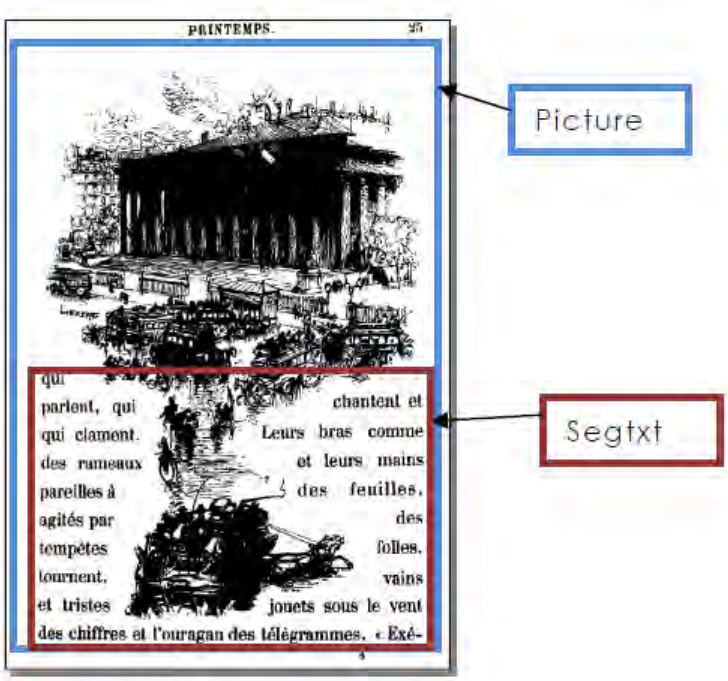
Marmion eut peine à écouter jusqu'au bout ce chant barbare; mais reconnaissant de la peine du ménestrel,

(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 section 5.9.2).

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.

J. H. ED. HEITZ, IMPRIMEUR-ÉDITEUR, STRASBOURG.

BIBLIOTHECA ROMANICA.

DIRECTION | ED. SCHNEEGANS, Strasbourg | PAUL HEITZ, Strasbourg

LISTE DES COLLABORATEURS:

C. Appel, Breslau
C. Bartsch, Vienne
F. Beck, Bamberg
Aug. Becker, Leipsick
A. Couret, Charroux
J. Cuervo, Paris
S. Debenedetti, Turin
F. Dondin, Rembas
J. Frohnleitner, Strasbourg
Th. Gerold, Strasbourg
G. Gigli, Viglietti
H. Gillot, Strasbourg
P. G. Gruber, Strasbourg
H. Hauvette, Paris
B. Heller, Budapest
E. Hupffer, Strasbourg
P. F. Holle, Berlin
L. Jordan, Munich
F. Kohler, Strasbourg
M. Liepelmann, Berlin
F. Lutz, Düsseldorf
C. Michaëlis de Vasconcellos, Porto
P. Naffi, Catania
F. Neri, Siracuse
J. J. Olivier, Venise
C. Orlando, Rome
R. Palmarochi, Florence
A. Paste, Metz
P. Savi-Lopez, Naples
B. Schmidhauser, Augsbourg
Ed. Schneegans, Strasbourg
E. Sicardi, Catania
L. Serrano, Catania
G. Teichio, Spezia
C. This, Strasbourg
H. Vagassy, Lyon
B. Wiese, Halle
W. Wurzbach, Vienne

Bibliothèque française

Guérin, Maurice de. Journal, Lettres, Poèmes, Fragments. — Intr. par F. Ed. Schneegans. 132 136.

La Heryerie. Caractères. — Intr. par F. Ed. Schneegans. 102 107.

Lamartine. Méditations. — Intr. par F. Ed. Schneegans. 70 77.

La petite Bourgeoisie. Poème satirique de l'an 1840. — Intr. par M. Liepelmann. 208.

Marot, Clément, Psautier Huguenot avec mélodies. — Intr. par Th. Gérold. 252/254.

Molière. Le Misanthrope. — Intr. par G. Gruber. 1.

—, Les Femmes savantes. — Intr. par G. Gruber. 2.

—, L'Avare. — Intr. par C. This. 46.

—, Tartuffe. — Intr. par G. Gruber. 114.

—, L'école des femmes. — La critique de l'école des femmes. — L'impromptu de Versailles. — Remerciement au roi. — Intr. par F. Ed. Schneegans. 255/257.

—, Le Malade imaginaire. — Intr. par F. Dondin. 225/228.

—, Les Fâcheux. — Intr. par F. Ed. Schneegans. 231.

—, Le Bourgeois gentilhomme. — Intr. par C. This. 249 250.

—, Monsieur de Pourceaugnac. Intr. par F. Ed. Schneegans. 252.

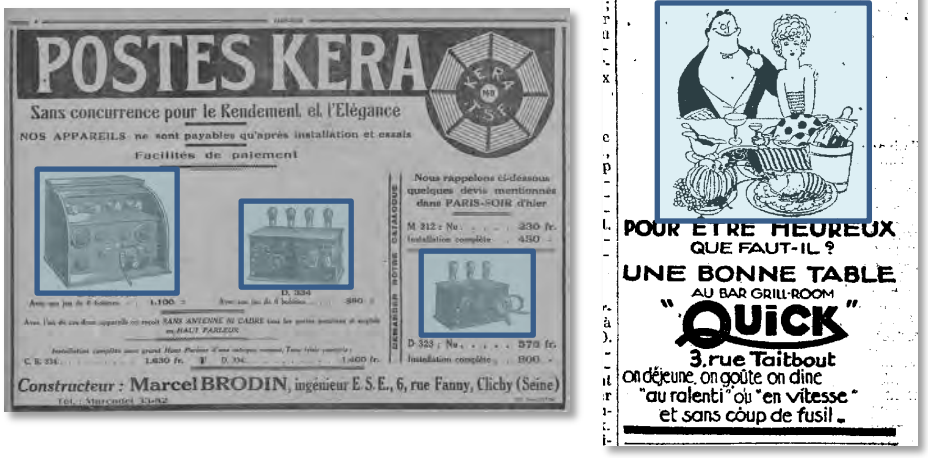
—, Femme médecin. Intr. par F. Ed. Schneegans. 256.

NOTES

Si un bloc de publicité est composé sous forme d'un tableau, un double étiquetage LayoutTag "advertisement" et LayoutTag "table" sera utilisé.

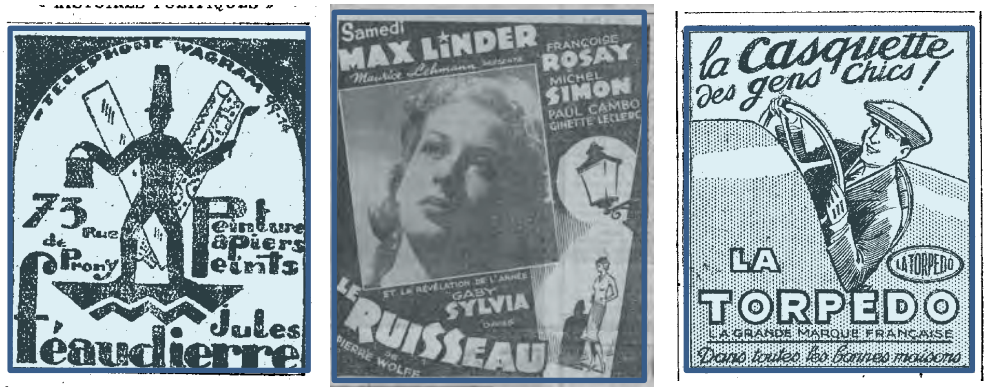
Si une zone de publicité comporte du texte et des illustrations, les illustrations seront segmentées en bloc Illustration et ces blocs seront également étiquetés LayoutTag LABEL="advertisement".

EXEMPLE



Si une zone de publicité comporte du texte et des illustrations mêlées, la totalité de la zone sera segmentée en bloc Illustration.

EXEMPLE



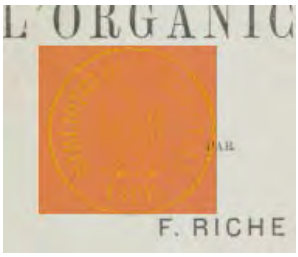
5.6.8 Texte sous tampon

Un bloc de texte placé sous un tampon sera marqué comme un TextBlock avec une étiquette LayoutTag LABEL="textStamped" (cf. section 5.10).

Le bloc de texte sera groupé avec le bloc tampon à l'aide d'un bloc composé.

Le texte contenu sera traité en OCR brut.

EXEMPLE

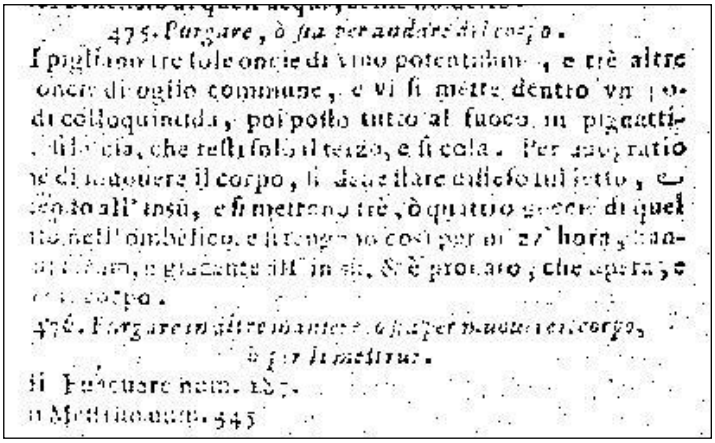


Ici, le bloc "PAR"

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 une étiquette LayoutTag LABEL ="illegible" (cf. section 5.10).

EXEMPLE



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

ATTENTION

LES ELEMENTS CONTENUS DANS UN BLOC ILLISIBLE (LIGNES ET MOTS) NE SERONT PAS EUX-MEMES ETIQUETES EN « ILLISIBLE ».

5.6.10 Texte manqué

Un élément textuel non reconnu en tant que tel par l'OCR sera segmenté en bloc texte de contenu "[texte manquant]", et identifié par une étiquette LayoutTag ayant pour valeur "missing" (cf. section 5.4).

EXEMPLE

Bloc avec texte en rotation non reconnu par l'OCR

5.6.11 Texte en OCR brut

Un bloc de texte traité en OCR brut sera marqué en tant que TextBlock avec une étiquette LayoutTag LABEL="raw". Ce typage est à réaliser uniquement :

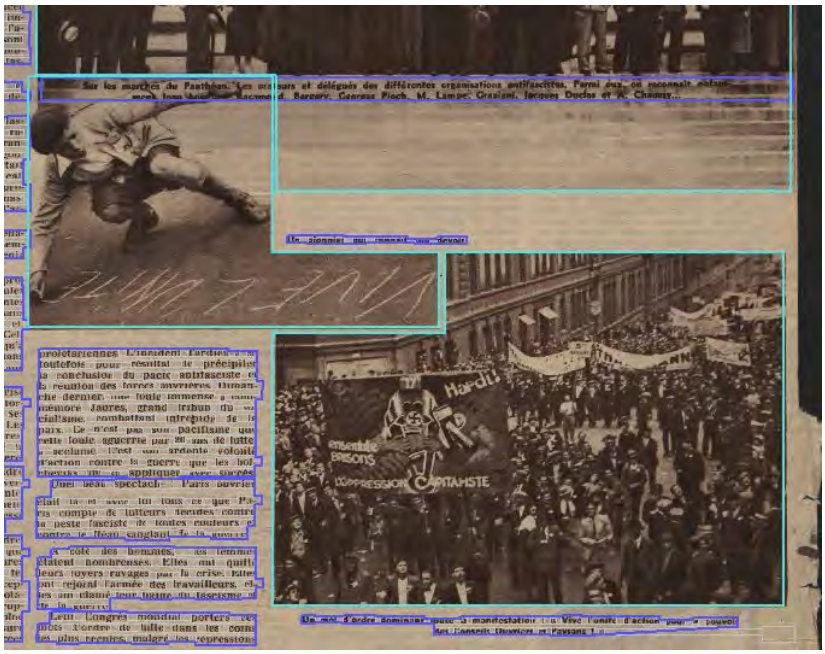
- pour certaines natures de contenus, en fonction des directives de la BnF,
- pour des prestations en qualité OCR garantie.

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.

Par défaut, les illustrations seront segmentées sous forme de rectangle. Pour certains types documentaires, des formes géométriques autres que le rectangle (notamment le polygone) pourront être utilisées. A cet effet, on emploiera un élément ALTO Shape.

EXEMPLE



Illustrations de presse polygonales

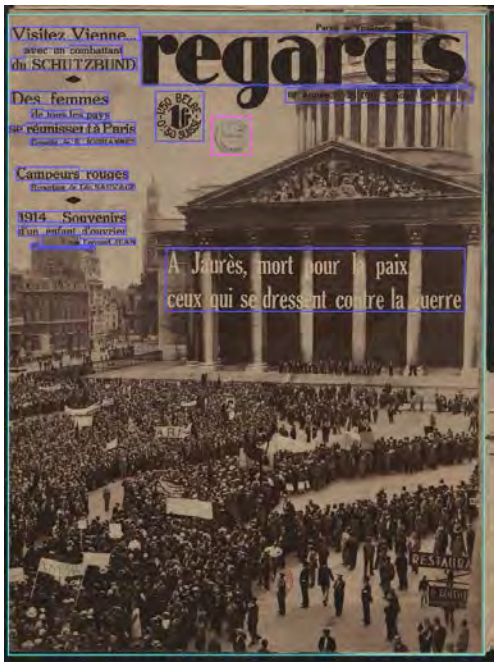
Dans le cas d'utilisation de la seule forme rectangle, la couverture d'un bloc illustration ou élément graphique pourra être répartie en plusieurs segments rectangulaires, à condition que l'ensemble des blocs couvre la totalité de la zone illustrée.

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.

En cas de mise en page complexe où textes et images sont mêlés, le bloc illustration aura l'emprise nécessaire et les blocs texte seront placés au-dessus.

EXEMPLE



5.7.1 Typage « illustration »

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

5.7.2 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.

La légende peut être placée sur l'illustration (cf. exemple du bas ci-dessous).

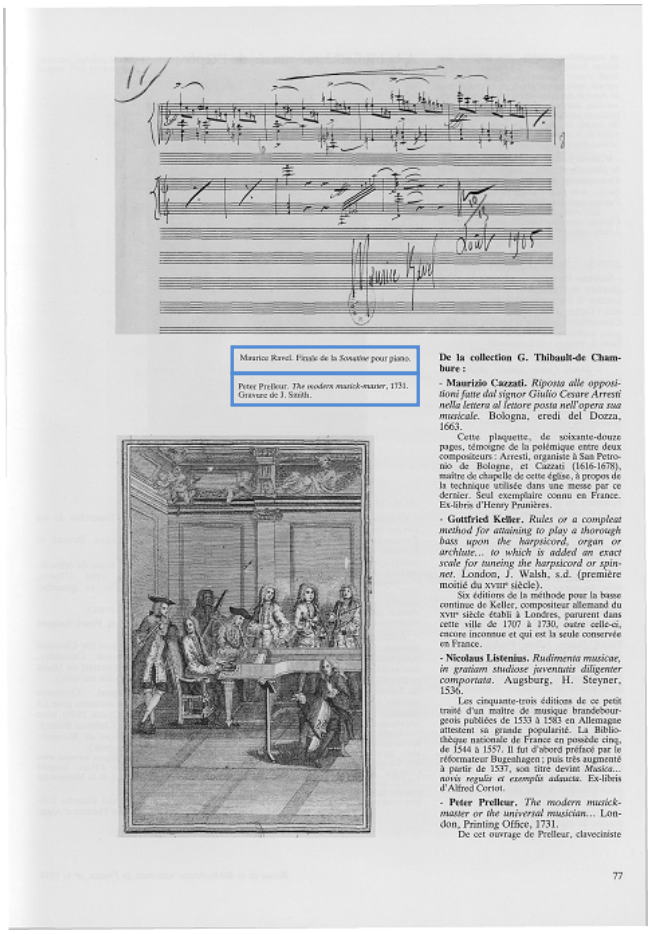
EXEMPLE

Bloc composé englobant illustration et légende

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. Finale 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.** *Riposta alle opposizioni 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 adaucta.* 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

5.7.3 Formules chimiques, mathématiques

La structuration des formules est réalisée en mode image, avec un bloc Illustration doté d'une étiquette LayoutTag égale à "formula" (cf. section 5.10).

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

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

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

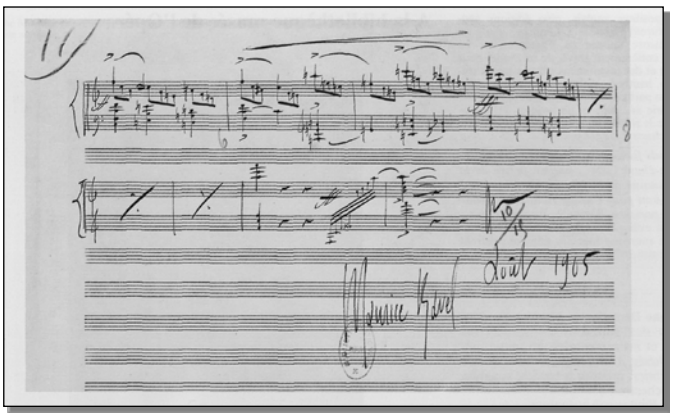
NOTES

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.

5.7.4 Partitions

Les extraits de partition de musique sont encodés en bloc Illustration dotés d'un étiquette LayoutTag LABEL="musicScore" (cf. section 5.10).

EXEMPLE

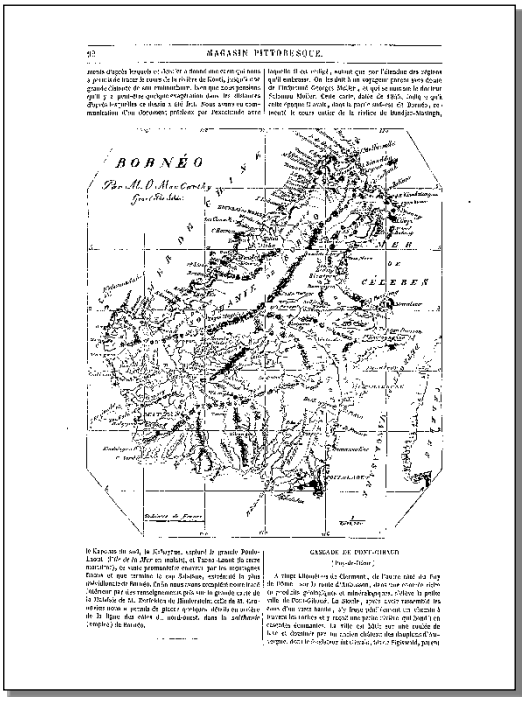


Exemple de partition

5.7.5 Cartes

Les illustrations reconnues comme étant des cartes géographiques seront typées spécifiquement en utilisant une étiquette LayoutTag égale à "map" (cf. section 5.10).

EXEMPLE



Exemple de carte

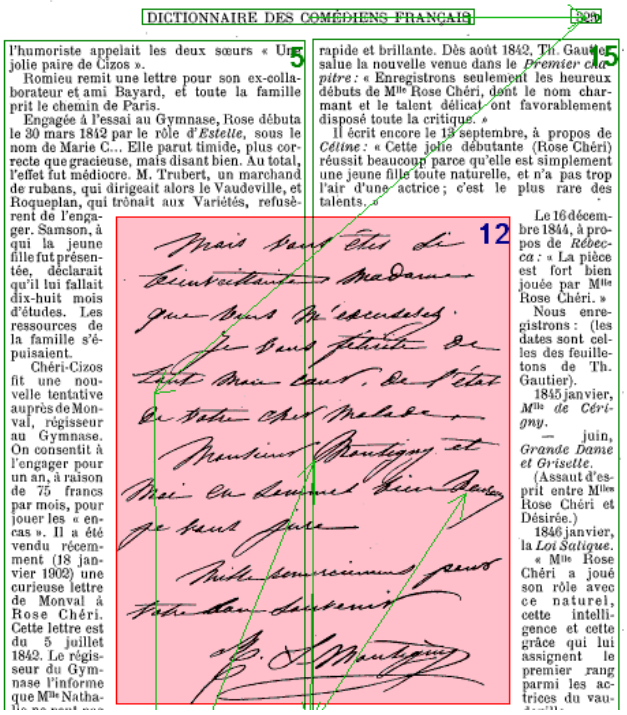
5.7.6 Alphabets non latins

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'une étiquette LayoutTag LABEL="nonLatinScript" (cf. section 5.10).

5.7.7 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



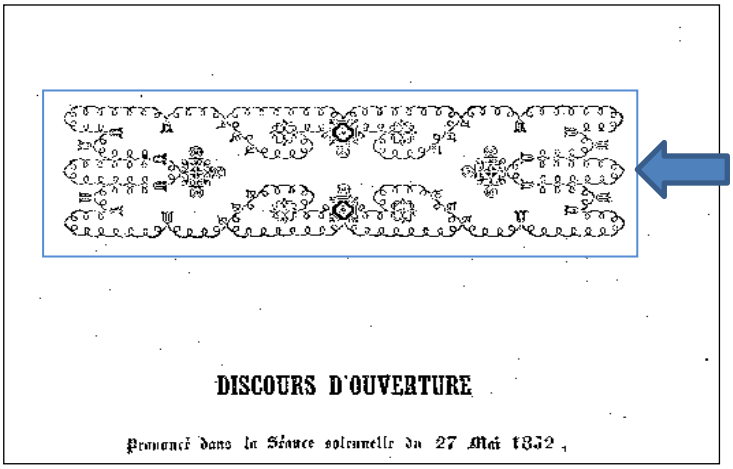
Exemple d'illustration sous forme d'écriture manuscrite

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'une étiquette LayoutTag LABEL="stamp" (cf. section 5.10).

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'une étiquette LayoutTag égale à "dropCap" (cf. section 5.10).

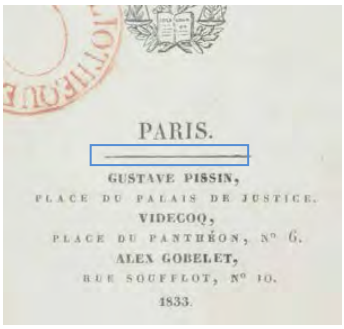
EXEMPLE

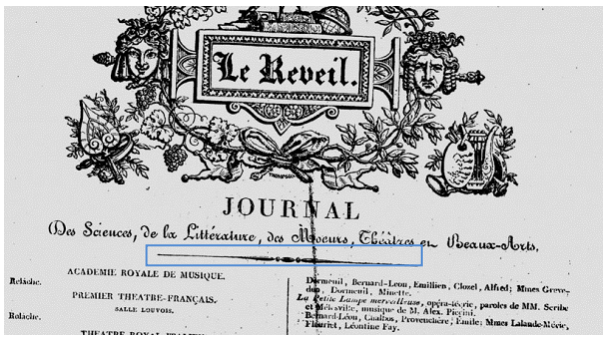
Exemple de lettrine (lettre R)

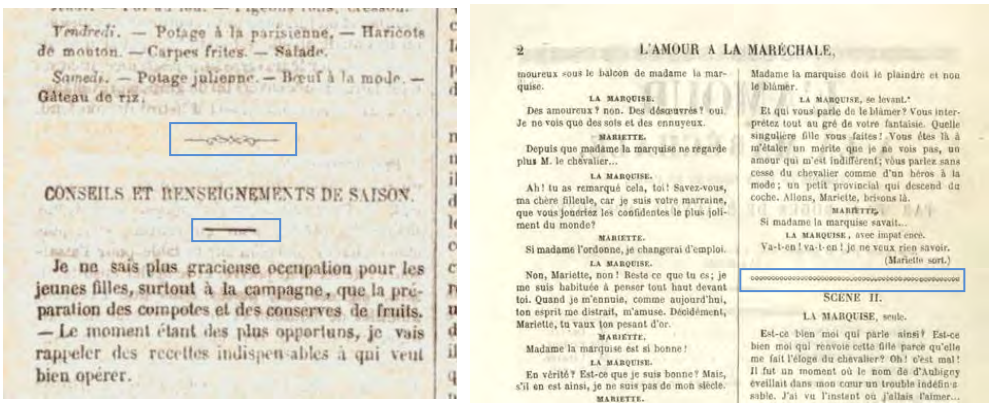
5.8.4 Traits de séparation

Les traits de séparation et les culs-de-lampe placés entre deux paragraphes doivent être décrits par un bloc GraphicalElement doté d'une étiquette LayoutTag LABEL ="transition" (cf. section 5.10).

EXEMPLE







Vendredi. — Potage à la parisienne. — Haricots de mouton. — Carpes frites. — Salade.

Samedi. — Potage julienne. — Bœuf à la mode. — Gâteau de riz.

CONSEILS ET RENSEIGNEMENTS DE SAISON.

Je ne sais plus gracieuse occupation pour les jeunes filles, surtout à la campagne, que la préparation des compotes et des conserves de fruits. — Le moment étant des plus opportuns, je vais rappeler des recettes indispensables à qui veut bien opérer.

2 L'AMOUR A LA MARÉCHALE.

moureux sous le balcon de madame la marquise.

LA MARQUISE.

Des amoureux ? non. Des désœuvrés ? oui. Je ne vois que des sots et des ennuyeux.

MARIETTE.

Depuis que madame la marquise ne regarde plus M. le chevalier...

LA MARQUISE.

Ah ! tu as remarqué cela, toi ! Savez-vous, ma chère filleule, car je suis votre marraine, que vous pourriez les confidentes le plus joliment du monde?

MARIETTE.

Si madame l'ordonne, je changerai d'emploi.

LA MARQUISE.

Non, Mariette, non ! Reste ce que tu es ; je me suis habituée à penser tout haut devant toi. Quand je m'ennuie, comme aujourd'hui, ton esprit me distrait, m'amuse. Décidément, Mariette, tu vaux ton pesant d'or.

MARIETTE.

Madame la marquise est si bonne !

LA MARQUISE.

En vérité ? Est-ce que je suis bonne ? Mais, s'il en est ainsi, je ne suis pas de mon siècle.

MARIETTE.

Madame la marquise doit le plaindre et non le blâmer.

LA MARQUISE, se levant.

Et qui vous parle de le blâmer ? Vous interprétez tout au gré de votre fantaisie. Quelle singulière fille vous faites ! Vous êtes là à m'étaler un mérite que je ne vois pas, un amour qui m'est indifférent; vous parlez sans cesse du chevalier comme d'un héros à la mode ; un petit provincial qui descend du coche. Allons, Mariette, brisons là.

MARIETTE.

Si madame la marquise savait...

LA MARQUISE, avec impatience.

Va-t-en ! va-t-en ! je ne veux rien savoir. (Mariette sort.)

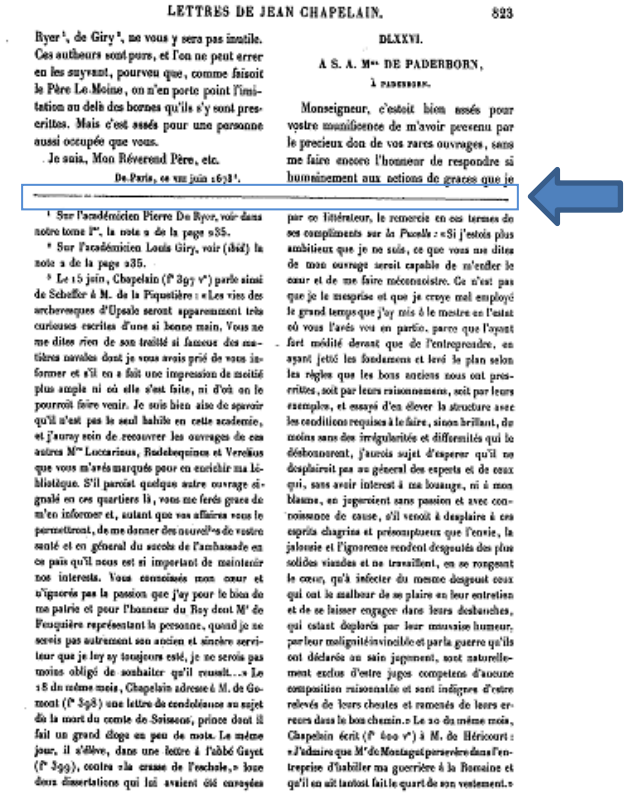
SCENE II.

LA MARQUISE, seule.

Est-ce bien moi qui parle ainsi ? Est-ce bien moi qui renvoie cette fille parce qu'elle me fait l'éloge du chevalier? Oh ! c'est mal ! Il fut un moment où le nom de d'Aubigny éveillait dans mon cœur un trouble indéfinissable. J'ai vu l'instant où j'allais l'aimer...

Les traits de séparation entre corps du texte et notes de bas de page sont à décrire en GraphicalElement avec une étiquette LayoutTag LABEL="footnoteSep".

EXEMPLE



LETTRES DE JEAN CHAPELAIN. 823

Ryer [1], de Giry [2], ne vous y sera pas inutile. Ces auteurs sont purs, et l'on ne peut errer en les suivant, 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 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 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 mettre 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 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 seroit 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 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 M^r de Montagut poursuive dans l'entrepreise d'habiller ma guerrière à la Romaine et qu'il en ait tantost fait le quart de son vestement. »

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

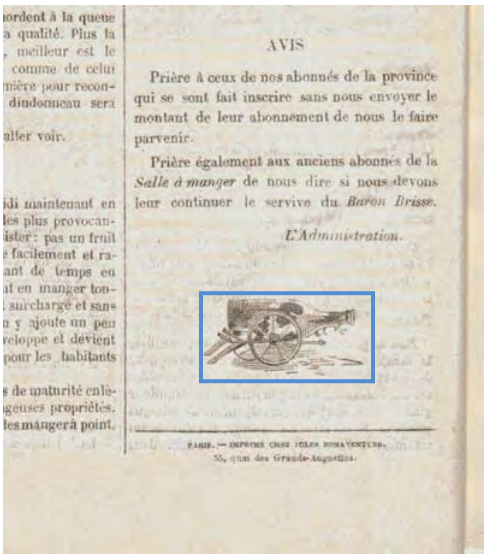
[2] Sur l'académicien Louis Giry, voir (ibid.) la note 2 de la page 235.

[3] Le 15 juin, Chapelain (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 sçavoir qu'il n'est pas le seul habille en cette academie, et j'auray soin de recouvrer les ouvrages de ces autres M^rs Loccenius, Rudebequius 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 nouvel^les 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 M^r de Fesquière représentant la personne, quand je ne serois pas autrement son ancien et sincère serviteur que je luy ay toujours esté, je ne serois pas moins obligé de souhaiter 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 Gayet (f° 399), contre « la crasse de l'escholle, » pour deux dissertations qui lui avoient été envoyées

Exemple de trait de séparation des notes de bas de page

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

EXEMPLE



ordent à la queue
a qualité. Plus la
, meilleur est le
comme de celui
mière pour recon-
dindonneau sera

aller voir.

idi maintenant en
les plus provocan-
ister : pas un fruit
facilement et ra-
ant de temps en
it en manger ton-
surchargé et sans
y ajoute un peu
veloppe et devient
pour les habitants

s de maturité enlè-
geuses propriétés.
les mangera point.

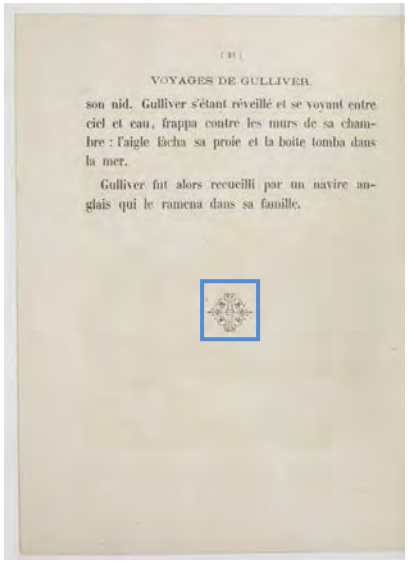
AVIS

Prière à ceux de nos abonnés de la province qui se sont fait inscrire sans nous envoyer le montant de leur abonnement de nous le faire parvenir.

Prière également aux anciens abonnés de la *Salle à manger* de nous dire si nous devons leur continuer le service du *Baron Brisse*.

L'Administration.

PARIS. — IMPRIMÉ CHEZ JULES BONAVENTURE,
55, quai des Grands-Augustins.



(31)

VOYAGES DE GULLIVER.

son nid. Gulliver s'étant réveillé et se voyant entre ciel et eau, frappa contre les murs de sa chambre : l'aigle fâcha sa proie et la boîte tomba dans la mer.

Gulliver fut alors recueilli par un navire anglais qui le ramena dans sa famille.

GRAMMAIRE

FRANÇAISE

DE L'ABBÉ GAULTIER,

ENTIÈREMENT REFONDUE ET CONSIDÉRABLEMENT AUGMENTÉE

PAR

MM. BLIGNIÈRES, DEMOYENCOURT, DUFROS (DE SIXT) ET LE CLERC AÎNÉ, SES ÉLÈVES.

PREMIÈRE ÉDITION.

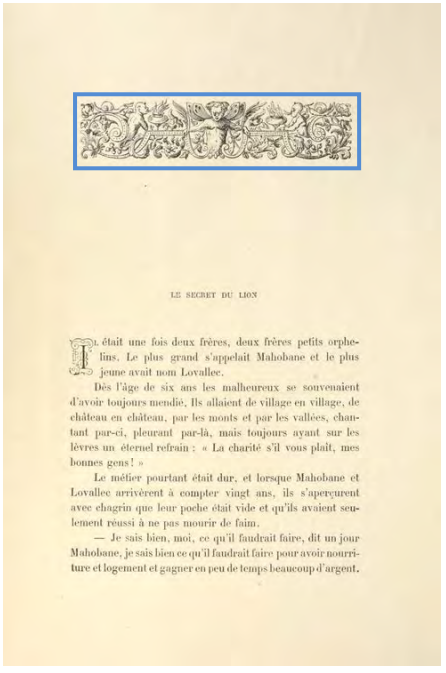
BIBLIOTHÈQUE ROYALE

PARIS,

JULES RENOUARD, LIBRAIRE,

RUE DE TOURNON, N° 6.

M DCCC XXIX.



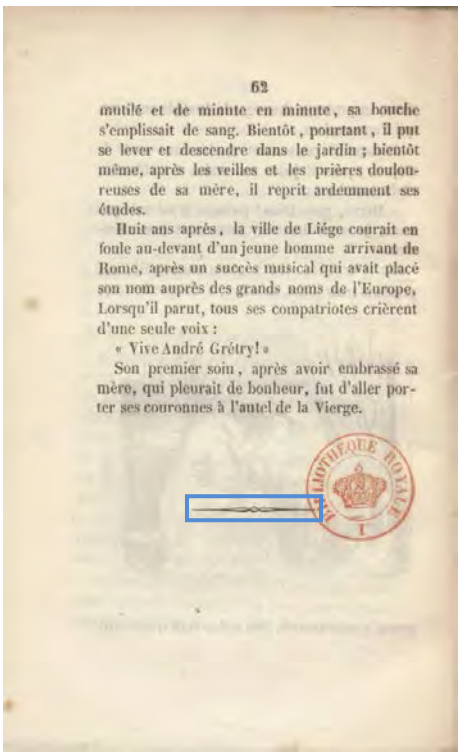
LE SECRET DU LION

Il était une fois deux frères, deux frères petits orphelins. Le plus grand s'appelait Mahobane et le plus jeune avait nom Lovallec.

Dès l'âge de six ans les malheureux se souvenaient d'avoir toujours mendié. Ils allaient de village en village, de château en château, par les monts et par les vallées, chantant par-ci, pleurant par-là, mais toujours ayant sur les lèvres un éternel refrain : « La charité s'il vous plaît, mes bonnes gens ! »

Le métier pourtant était dur, et lorsque Mahobane et Lovallec arrivèrent à compter vingt ans, ils s'aperçurent avec chagrin que leur poche était vide et qu'ils avaient seulement réussi à ne pas mourir de faim.

— Je sais bien, moi, ce qu'il faudrait faire, dit un jour Mahobane, je sais bien ce qu'il faudrait faire pour avoir nourriture et logement et gagner en peu de temps beaucoup d'argent.



62

mutilé et de minute en minute, sa bouche s'emplissait de sang. Bientôt, pourtant, il put se lever et descendre dans le jardin ; bientôt même, après les veilles et les prières douloureuses de sa mère, il reprit ardemment ses études.

Huit ans après, la ville de Liége courait en foule au-devant d'un jeune homme arrivant de Rome, après un succès musical qui avait placé son nom auprès des grands noms de l'Europe. Lorsqu'il parut, tous ses compatriotes crièrent d'une seule voix :

« Vive André Grétry! »

Son premier soin, après avoir embrassé sa mère, qui pleurait de bonheur, fut d'aller porter ses couronnes à l'autel de la Vierge.

BIBLIOTHÈQUE ROYALE

Exemple de cul-de-lampe et décorations

Contenu presse

Les décorations et traits de séparation ornés seront segmentés.

Les traits de séparation non ornés ne seront segmentés que s'ils font toute la largeur de la page ou s'ils sont disposés à cheval sur au moins deux colonnes.

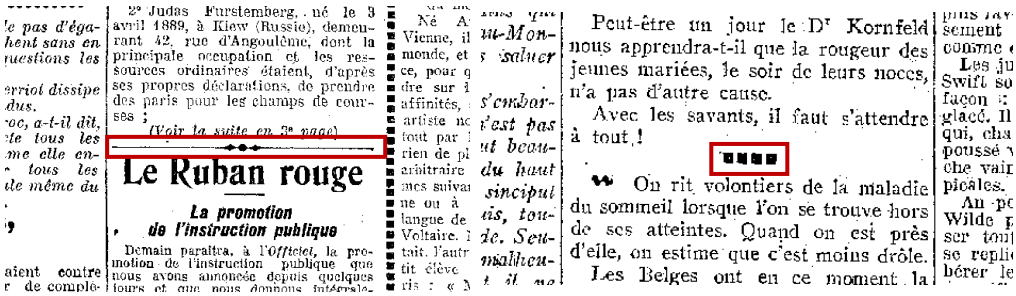
Les traits de séparation verticaux ne seront pas segmentés.

EXEMPLE

Exemples de traits de transition à segmenter

ATTENTION

LES AUTRES TYPES DE SEPARATEURS DE TRANSITION POURRONT EVENTUELLEMENT ETRE SEGMENTES, MAIS EN AUCUN CAS OCERISES, NOTAMMENT CEUX POUVANT ETRE CONFONDUS AVEC DU TEXTE.



5.8.5 Ecriture manuscrite

Les mentions manuscrites parfois présentes sur les pages des documents sont à capturer en GraphicalElement, avec une étiquette LayoutTag ayant pour valeur "manuscript" (cf. section 5.10).

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 (voir section suivante).

5.9 Blocs composés

L'élément ComposedBlock est utilisé pour permettre le regroupement ou le recouvrement d'éléments de type bloc :

- regroupement logique (par exemple image et légende, cf. section 5.7.2),
- recouvrement physique : cf sections suivantes.

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 (cf. section 5.7.2).
- les textes au sein même d'un dessin, d'une œuvre d'art, d'un ornement, etc. (cf. section 5.8).

EXEMPLE

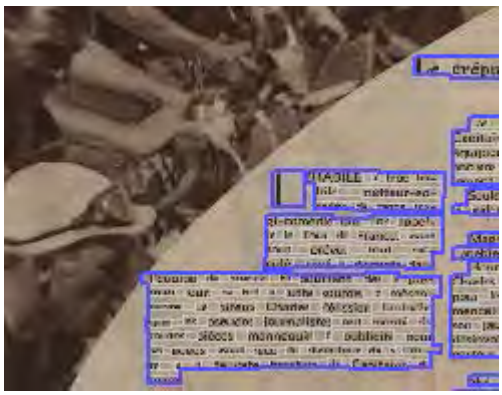
Ce cas sera traité ainsi :

- le ou les TextBlock seront serrés autour du texte,
- l'image sera couverte par un bloc Illustration/GraphicalElement,
- les TextBlock et le bloc Illustration/GraphicalElement seront englobés par un ComposedBlock.

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.

EXEMPLE



Ce cas sera traité ainsi :

- les TextBlock seront serrés autour du texte – jusqu'à un TextBlock distinct par ligne de texte,
- l'illustration sera couverte par un bloc Illustration, qui peut recouvrir entièrement ou partiellement les TextBlock ;

- l'ordre des blocs Illustration est libre, à condition que les TextBlock figurent dans l'ordre de la lecture. Les TextBlock peuvent être non contigus.

S'il y a recouvrement, les TextBlock et les GraphicalElement seront englobés par un ComposedBlock.

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

Le tableau suivant détaille les différents cas d'usage des éléments structurants du format ALTO.

TextBlock			
Cas	Etiquette	Section	Remarque
Paragraphe		5.6.1	
Titre	StructureTag : "titre1", "titre2"	4.2.3, 5.6.2	
Tableau	LayoutTag : "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	LayoutTag : "advertisement"	5.6.7	
Police “script”	LayoutTag : "scriptFonts"	4.2.3	
Texte de l'ouvrage sous un tampon	LayoutTag : "textStamped"	5.6.8	
Texte illisible	LayoutTag : "illegible"	4.2.3, 5.6.9	
Texte manquant	LayoutTag : "missing"	5.4	texte non détecté par l'OCR
Texte en OCR brut	LayoutTag : "raw"	5.6.10	texte traité en OCR brut
TextLine			
Cas	Etiquette	Section	Remarque
Ligne illisible	LayoutTag : "illegible"	4.2.3, 5.6.9	

String			
Cas	Etiquette	Section	Remarque
Mot illisible	LayoutTag : "illegible"	6.3.2	
Mot important	OtherTag : "important"	6.2.1	
Illustration			
Cas	Etiquette	Section	Remarque
Image, illustration, schéma, etc.		5.7	
Formule mathématique ou chimique	LayoutTag : "formula"	5.7.1	
Partition	LayoutTag : "musicScore"	5.7.2	
Carte, plan	LayoutTag : "map"	5.7.3	
Alphabet non latin	LayoutTag : "nonLatinScript"	4.1.2, 5.7.4	
Illustration avec légende		5.7.6	utiliser un ComposedBlock
GraphicalElement			
Cas	Etiquette	Section	Remarque
Décoration, cul de lampe		5.8.1	
Tampon	LayoutTag : "stamp"	5.8.2	
Lettrine	LayoutTag : "dropCap"	5.8.3	
Trait de séparation entre paragraphes	LayoutTag : "transition"	5.8.4	
Trait de séparation entre texte et notes de bas de page	LayoutTag : "footnoteSep"	5.8.4	
Ecriture manuscrite	LayoutTag : "manuscript"	5.8.5	

5.10.1 Etiquetage des éléments

On utilisera la syntaxe suivante :

<Tag ID="x" LABEL="étiquette" [DESCRIPTION="…"]/>

Une étiquette est associée à un élément à l'aide de l'attribut TAGREFS et d'un identifiant d'étiquette :

<Element ID="…" HPOS="…" VPOS="…" TAGREFS="id-x"/>

Une étiquette peut être *générique* (par exemple l'étiquette des « mots illisibles ») ou *instanciée*, c'est-à-dire associée à un élément particulier (par exemple l'étiquette de la « lettrine A du 2e bloc de la page 24 »).

Les identifiants des étiquettes génériques ne sont pas numérotés (la liste des étiquettes génériques est donnée à la fin de cette section) :

```
<LayoutTag ID="TAG_table" LABEL="table" />
```

Les identifiants des étiquettes instanciées sont numérotés (sur 3 digits, de 001 à 999) :

```
<LayoutTag ID="TAG_LT001" LABEL="table"
  DESCRIPTION="tableau de nombres "/>
```

L'attribut DESCRIPTION est optionnel. Lorsqu'il est présent, il informe du contenu concerné par l'étiquette (soit en le décrivant, soit par son contenu textuel lui-même).

Une étiquette doit référencer le plus haut niveau disponible, par exemple un bloc et non tous les mots du bloc.

Un contenu à cheval sur deux pages peut être étiqueté par la même étiquette.

Un contenu peut être étiqueté par plusieurs étiquettes, en séparant les identifiants d'étiquette par un espace : TAGREFS="id1 id2"

Layout

Les étiquettes LayoutTag servent à typer la nature des éléments de contenu (tableau, publicité, formule mathématique, carte, etc.).

Etiquettes génériques :

```
<LayoutTag ID="TAG_table" LABEL="table" />

<LayoutTag ID="TAG_textStamped"  LABEL="textStamped"/>

<LayoutTag ID="TAG_manuscript"  LABEL="manuscript"/>

<LayoutTag ID="TAG_scriptFonts"  LABEL="scriptFonts"/>

<LayoutTag ID="TAG_footnoteSep"  LABEL="footnoteSep"/>
```

Etiquettes instanciées :

```
<LayoutTag ID="TAG_LT001" LABEL="table" DESCRIPTION="tableau de nombres "/>

<LayoutTag ID="TAG_LT008"  LABEL="formula" DESCRIPTION="4 formules
d'algèbre"/>

<LayoutTag ID="TAG_LT009"  LABEL="dropCap " DESCRIPTION="A"/>
```

<LayoutTag ID="TAG_LT010" LABEL="nonLatinFont " DESCRIPTION="idéogrammes chinois"/>

Structure

Les étiquettes StructureTag servent à décrire la structure logique des contenus. On utilisera la syntaxe suivante :

<StructureTag ID="x" TYPE="Structural" LABEL="étiquette" [DESCRIPTION="…"]/>

Etiquettes génériques :

<StructureTag ID="TAG_title1" TYPE="Structural" LABEL="title1"/>

<StructureTag ID="TAG_title2" TYPE="Structural" LABEL="title2"/>

Etiquettes instanciées :

<StructureTag ID="TAG_ST008" TYPE="Structural" LABEL="title1" DESCRIPTION="Chroniques littéraires"/>

Autres cas

Les étiquettes OtherTag seront utilisées pour décrire les mots importants (cf. section 6.2.1) :

<OtherTag ID="x" [TYPE="…"] LABEL="étiquette" />

L'attribut TYPE pourra être utilisé pour préciser les types de mot important, par exemple les entités nommées (EN) ou les mots présents dans les titres.

Etiquettes génériques :

<OtherTag ID="TAG_important" LABEL="important"/>
<OtherTag ID="TAG_importantEN" TYPE="EN" LABEL="important"/>
<OtherTag ID="TAG_importantTL" TYPE="titre" LABEL="important"/>

Etiquettes instanciées :

<OtherTag ID="TAG_OT001" TYPE="EN" LABEL="important" DESCRIPTION="Charles Dickens"/>

Liste des étiquettes génériques

TextBlock		
Cas	Etiquette	Section
Titre	<Structure Tag ID="TAG_title1" TYPE="Structural" LABEL="title1" /> <Structure Tag ID="TAG_title2" TYPE="Structural" LABEL="title2" /> <Structure Tag ID="TAG_title3" TYPE="Structural" LABEL="title3" /> <Structure Tag ID="TAG_title4" TYPE="Structural" LABEL="title4" />	4.2.2, 5.6.2
Tableau	<LayoutTag ID="TAG_table" LABEL="table" />	5.6.3
Publicité	<LayoutTag ID="TAG_advertisement" LABEL="advertisement" />	5.6.7
Police script	<LayoutTag ID="TAG_scriptFonts" LABEL="scriptFonts" />	4.2.3
Texte de l'ouvrage sous un tampon	<LayoutTag ID="TAG_textStamped" LABEL="textStamped" />	5.6.8
Texte illisible	<LayoutTag ID="TAG_illegible" LABEL="illegible"/>	4.2.3, 5.6.9
Texte manqué	<LayoutTag ID="TAG_missing" LABEL="missing" />	5.4
Texte en OCR brut	<LayoutTag ID="TAG_raw" LABEL="raw" />	5.6.10
TextLine		
Cas	Etiquette	Section
Ligne illisible	<LayoutTag ID="TAG_illegible" LABEL="illegible"/>	4.2.3, 5.6.9
String		
Cas	Etiquette	Section
Mot illisible	<LayoutTag ID="TAG_illegible" LABEL="illegible"/>	6.3.2
Mot important	<OtherTag ID="TAG_important" LABEL="important"/>	6.2.1
Illustration		
Cas	Etiquette	Section
Formule mathématique ou chimique	<LayoutTag ID="TAG_formula" LABEL="formula" />	5.7.1
Partition	<LayoutTag ID="TAG_musicScore" LABEL="musicScore" />	5.7.2
Carte, plan	<LayoutTag ID="TAG_map" LABEL="map" />	5.7.3
Alphabet non latin	<LayoutTag ID="TAG_nonLatinScript" LABEL="nonLatinScript" />	4.1.2, 5.7.4
GraphicalElement		
Cas	Etiquette	Section

Tampon	<LayoutTag ID="TAG_stamp" LABEL="stamp" />	5.8.2
Lettrine	<LayoutTag ID="TAG_dropCap" LABEL="dropCap" />	5.8.3
Trait de séparation entre paragraphes	<LayoutTag ID="TAG_transition" LABEL="transition" />	5.8.4
Trait de séparation entre texte et notes de bas de page	<LayoutTag ID="TAG_footnoteSep" LABEL="footnoteSep" />	5.8.4
Ecriture manuscrite	<LayoutTag ID="TAG_manuscript" LABEL="manuscript" />	5.8.5

6. QUALITE DE LA RECONNAISSANCE OCR

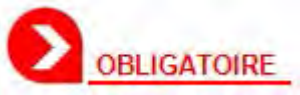
6.1 Qualité de la transcription du texte

La montée en qualité du texte, lorsqu'elle est demandée sur une prestation via un taux OCR qualité garantie ou haute qualité, a pour objectif d'atteindre un certain taux de reconnaissance des mots composant le document d'origine.

Ce taux est mesuré avec pour granularité le mot (élément ALTO <String>). Il est 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).

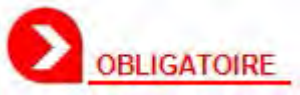


Chaque mot corrigé par un opérateur sera identifié dans le contenu ALTO (par exemple en positionnant à 1,0 le critère de confiance WC).

6.1.1 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.



Chaque mot identifié comme important sera repéré dans le contenu ALTO à l'aide d'une étiquette OtherTag de valeur "important" (cf. section 5.10).

6.2 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 nécessaire de corriger (structuration, typage, …) le résultat proposé par le moteur OCR.

La qualité de segmentation est mesurée à l'échelle du document numérique (ensemble des pages).

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 lui affectant une étiquette "illegible" (cf. section 5.10).

Lignes ou blocs

L'illisibilité d'une ligne ou d'un bloc est définie par une valeur seuil de la proportion de mots illisibles dans la ligne ou le bloc. Ce seuil est défini à 50 %.

La déqualification opérée à l'échelle d'une ligne ou d'un bloc de texte consiste à typer l'élément concerné (TextLine ou TextBlock) en lui donnant une étiquette "illegible" (cf. section 5.10).

Les mots inclus dans un bloc ou une ligne illisible n'ont pas à être étiquetés "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 (selon les critères décrits aux sections 6.3.1 et 6.3.2) 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 6.3.

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 ainsi que des avertissements.

Une erreur entraîne le rejet du document.

SI CE CONTROLE EMET UNE ERREUR 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.
- Codage du fichier ALTO : UNICODE UTF-8 sans BOM (*Byte Order Mark*).
- 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 blocs :
  - entre éléments de même niveau (ex: entre deux éléments String) ;
  - entre éléments de haut niveau (ex: entre un TopMargin et un PrintSpace) ;
  - entre éléments enfants et parents (ex: entre un TextLine et un TextBlock).

  La tolérance est par défaut de 5%. Elle pourra être adaptée selon les prestations.
- 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 reconnaissance du texte,
- la qualité de la segmentation/structuration.

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.

7.2.1 Qualité de la reconnaissance du texte

La qualité de la reconnaissance du texte est mesurée au mot.

Une erreur de reconnaissance correspond à tout mot transcrit erronément par rapport à l'image d'origine. Un mot est considéré comme erroné quel que soit le nombre de signes erronés qu'il contient.

Pour un document, la qualité se calcule sur la population des mots présents dans toutes les pages du document, en tenant compte des règles définies pour le taux OCR considéré (brut, qualité garantie, qualité éditoriale, cf. section 4.1.3) :

- prise en compte des parties déqualifiées,
- jeu de caractères.

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.

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 :

taux réel = *cardinal des mots justes parmi les mots en qualité garantie / 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). Cette évaluation sera réalisée sur des documents proches du taux qualité garantie du marché.

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) :

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

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.

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

7.2.2 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 (cf. CCTP). Il indique donc le nombre maximum de pages non conformes dans un document. Pour les documents à faible pagination (presse), la mesure sera réalisée sur le nombre total de pages de l'échantillon de contrôle.

Ce contrôle ne s'applique pas sur les documents traités en OCR brut.

Pour une page à contrôler, le contrôle de la segmentation/structuration se décompose en plusieurs contrôles réalisés selon des modalités spécifiques à chaque contrôle :

Contrôle	Modalité	Résultat
Typage des pages	contrôle unitaire	correct/incorrect
Identification du PrintSpace	contrôle unitaire	correct/incorrect
Identification des xxxMargin	contrôle unitaire	correct/incorrect
Ordre de lecture	contrôle unitaire	correct/incorrect
Typage des blocs	contrôle numérique	taux d'erreur ≤ (1 - taux qualité garantie) ?
Segmentation des blocs	contrôle numérique	taux d'erreur ≤ (1 - taux qualité garantie) ?

ATTENTION

CHACUN DE CES CONTROLES DOIT ETRE SATISFAIT, CE QUI IMPLIQUE :

- QU'UN SEUL CONTROLE UNITAIRE EN ECHEC INDUIT LA NON-CONFORMITE DE LA PAGE
- QU'UN SEUL CONTROLE NUMERIQUE EN ECHEC INDUIT LA NON-CONFORMITE DE LA PAGE.

Typage des pages

Le contrôle vérifie le typage des pages et de leur orientation (cf. sections 5.1 et 5.2).

Identification du PrintSpace et des xxxMargin

Le contrôle vérifie l'identification du PrintSpace et des xxxMargin (cf. section 5.3).

Ordre de lecture

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

Typage des blocs

Le contrôle vérifie le typage des blocs identifiés dans la page (cf. sections 5.6 à 5.8) :

- nature des blocs segmentés (par ex. confusion blocs texte/bloc illustration),
- typage des blocs (par ex. confusion blocs de texte/blocs de publicité),
- blocs déqualifiés (par ex. bloc illisible abusif).

Le taux d'erreur est calculé ainsi :

$$taux\ d'erreur\ typage = \sum erreurs\ typage\ /\ nombre\ de\ blocs$$

Il doit être inférieur au taux qualité attendu pour la segmentation (cf. CCTP).

Segmentation des blocs

Le contrôle vérifie la bonne segmentation des blocs :

- taille et position des blocs relativement aux contenus (sauf pour les blocs déqualifiés en illisible),

- découpage cohérent des paragraphes du contenu textuel en blocs de texte (sauf pour les blocs déqualifiés en illisible),
- blocs oubliés (par ex. blocs de texte non segmentés),
- blocs parasites (par ex. blocs de texte sans contenu textuel),
- blocs composés (application des règles).

Les erreurs de segmentation détectées lors du contrôle automatique de structure (cf. section 7.1) ne sont pas comptabilisées ici.

Le taux d'erreur est calculé ainsi :

$$taux\ d'erreur\ segm. = \sum erreurs\ segm. / nombre\ de\ blocs$$

Il doit être inférieur au taux qualité attendu pour la segmentation (cf. CCTP).

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.

7.2.3 Détail des métriques qualité

La BnF et le prestataire s'accorderont sur le format descriptif à utiliser pour les indicateurs statistiques et les taux qualité demandés par la BnF.

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 sur la totalité de la page
- le taux OCR corrigé : le taux de confiance après correction manuelle et prise en compte des éventuels mots et zones déqualifiés,
- le coefficient d'étalonnage utilisé (cf. section 7.2.2)
- le taux NQA (niveau de qualité acceptable) moyen, exprimé par : *taux OCR corrigé* × *coefficient d'étalonnage*. C'est ce taux qui fait foi relativement aux exigences qualité du marché.

Dans les fichiers ALTO

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

- nombre de blocs illisibles,
- nombre de mots illisibles,
- nombre de lignes dans les zones de courbures,
- nombre total de mots ALTO
- nombre total de caractères ALTO, par catégorie (lettres, chiffres, ponctuations)
- le taux OCR brut : le taux de confiance en sortie de l'OCR

- le taux OCR corrigé : le taux de confiance après correction manuelle (dernière ligne de l'exemple suivant). Ce taux sera identique au taux OCR brut dans le cas d'un document non corrigé ou d'une prestation sans correction.

  Ce taux sera inséré dans l'attribut ACCURACY de l'élément Page (cf. section 5.1.3).

D'autres indicateurs pourront être mis en place, après étude avec le prestataire

- nombre de lignes en courbure détectées et traitées
- difficulté de post-correction de la page
- etc.

Ces indicateurs seront décrits à l'aide d'instances de l'élément <processing StepDescription>. Par exemple :

```
<OCRProcessing ID="OCR_1">
 <ocrProcessingStep>
  <processingDateTime>2013-04-23</processingDateTime>
  <processingStepDescription>[001_OCR_BRUT]OCR BRUT 96.240 %</processingStepDescription>
 <processingStepDescription>[002_CHARS]{12187}</processingStepDescription>
 <processingStepDescription>[003_CHARS_USED]{11262}</processingStepDescription>
<processingStepDescription>[004_CHARS_SUSPECTS_USED]{0}</processingStepDescription>
 <processingStepDescription>[005_CHARS_UNUSED]{925}</processingStepDescription>
<processingStepDescription>[006_CHARS_SUSPECTS_UNUSED]{143}</processingStepDescription>
<processingStepDescription>[007_STRINGS]{2615}</processingStepDescription>
<processingStepDescription>[011_CHARS_PUNCTUATION]{781}</processingStepDescription>
<processingStepDescription>[012_CHARS_DIGITS]{52}</processingStepDescription>
<processingStepDescription>[013_CHARS_SUSPECT]{143}</processingStepDescription>
<processingStepDescription>[022_CHARS_UNUSED_BY_BLOCK]{0}</processingStepDescription>
<processingStepDescription>[111_LIGS_DETECTED_CURV]{0}</processingStepDescription> …
 </ocrProcessingStep>
</OCRProcessing>
```

Le tableau suivant présente la liste des indicateurs. Les indicateurs en gras doivent obligatoirement être inscrits dans les fichiers ALTO.

Code	Description	Exemple de valeur
001_OCR_BRUT	Taux OCR brut en sortie du moteur, calculé sur la totalité de la page	OCR BRUT 96.240 %
002_CHARS	Nombre total de caractères de la page	12187
003_CHARS_USED	Nombre total de caractères utilisés pour le calcul du taux OCR pondéré (cf. section 4.1.3 : hors caractères en OCR brut, illisibles, chiffres,	11262

	ponctuations, spéciaux)	
004_CHARS_SUSPECTS_USED	Nombre total de caractères comptabilisés marqués comme suspects par l'OCR. Toujours égal à 0	0
005_CHARS_UNUSED	Nombre total de caractères non comptabilisés pour le calcul du taux OCR pondéré. Rq : 005 = 002 - 003	925
006_CHARS_SUSPECTS_UNUSED	Nombre total de caractères non comptabilisés marqués comme suspects par l'OCR Rq : = tous les caractères des mots/blocs de type illisible	143
	Rq : Nombre total de caractères en OCR brut = 002 – 003 - 006	782
007_STRINGS	Nombre total de mots (String)	2615
008_STRINGS_USED	Nombre total de mots utilisés pour le calcul du taux OCR pondéré (hors mots en OCR brut et illisibles)	2544
009_STRINGS_SUSPECTS_UNUSED	Nombre total de mots non comptabilisés marqués comme suspects par l'OCR (illisibles isolés ou dans les blocs illisibles).	21
	Rq : Nombre total de mots en OCR brut = 007 – 008 – 009	50
010_CHARS_LETTERS	Nombre de caractères de type lettre	
011_CHARS_PUNCTUATION	Nombre de caractères de ponctuation	781
012_CHARS_DIGITS	Nombre de caractères de type chiffre arabe (0-9)	52
	Rq : Autres caractères = 002 - 010 + 011 + 012	
013_CHARS_SUSPECT	Nombre de caractères suspects (comptabilisés ou non). Tous les caractères suspects sont ceux des textes typés en illisible. Rq : 013 = 006	143
022_CHARS_UNUSED_BY_BLOCK	Tous les caractères des blocs suspects (illisibles) Rq : dans les blocs seuls, pas ceux des mots illisibles isolés	100
	Rq : Nombre de caractères suspects isolés = 013 - 022	43
023_BLOCKS_SUSPECTS	Nombre total de blocs marqués comme suspects (illisibles) par	1

	l'OCR.	
111_LIGS_DETECTED_CURV	Nombre de lignes courbes détectées sur la page.	0