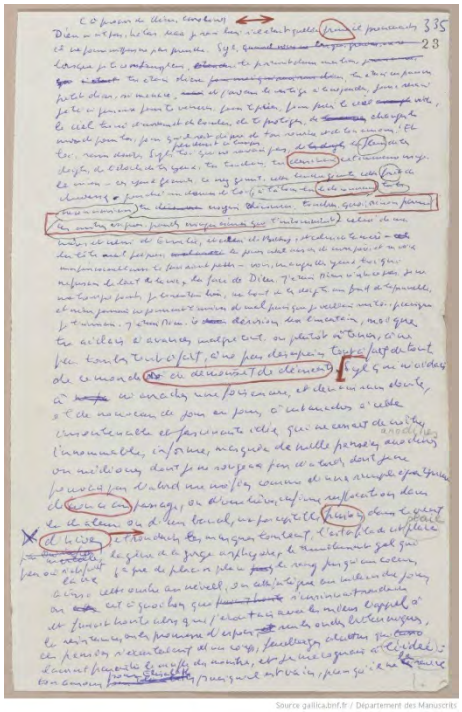
Référentiel d'enrichissement du texte



http://gallica.bnf.fr/ark:/12148/btv1b105097519/f46

Bibliothèque nationale de France direction des Services et des réseaux département de la Conservation service Numérisation	Date :le jeudi 30 avril 2015 Version :1 Référence BnF :BnF-ADM-2015-023247-01

TABLE DES MATIÈRES

1. INTRODUCTION

Ce document présente les différents modes d'enrichissement des contenus textuels pouvant être demandés par la BnF lors d'une prestation de numérisation de documents imprimés.

Ces enrichissements s'appliquent principalement à des documents de type périodique (revues, journaux).

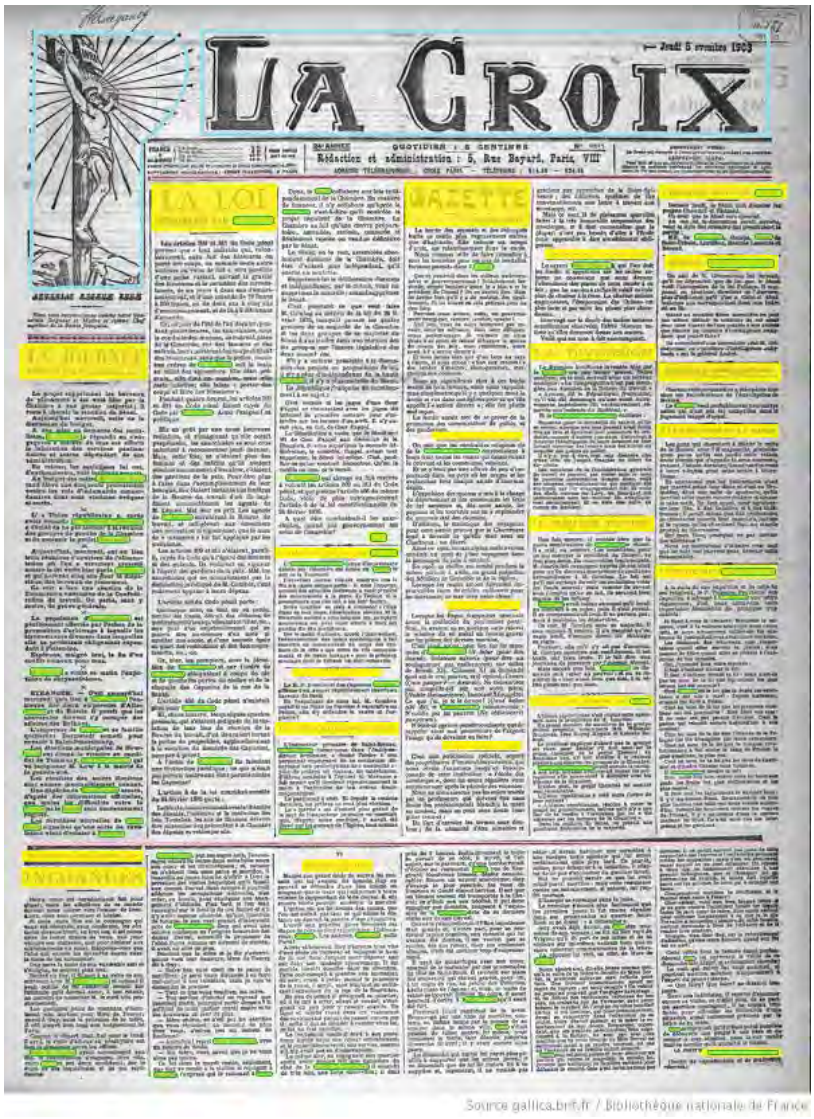
1.1 Documents applicables et de référence

Standards	
METS	http://www.loc.gov/standards/mets/
MODS	http://www.loc.gov/standards/mods/
PREMIS	http://www.loc.gov/standards/premis/
ALTO	http://www.loc.gov/standards/alto/
Format ENMAP	http://www.europeana-newspapers.eu/wp-content/uploads/2014/08/D5.2-Europeana-Newspapers-METS-ALTO-Profile-ENMAP-DRAFT.pdf
Référentiels BnF	http://www.bnf.fr/fr/professionnels/numerisation_boite_outils/a.numerisation_referentiels_bnf.html
Référentiel d'enrichissement des métadonnées METS	version 1 / BnF-ADM-2013-117422-01
Référentiel OCR	version 2 / BnF-ADM-2014-062821-02
Référentiel de livraison de document numérique	Version 3 / BnF-ADM-2013-077351-03

2. RECONNAISSANCE DES ENTITES NOMMEES

2.1 Principe

La reconnaissance d'entités nommées (REN) est une tâche d'extraction d'information dans un corpus documentaire. Elle consiste à rechercher et identifier des entités textuelles porteuses de sens dans un corpus, ces entités relevant de catégories sémantiques telles que noms de personnes, noms de lieux (toponymes, pays, villes, etc.), noms d'organisations, d'entreprises ou de marques, dates, quantités et montants, etc.



Exemple de REN (en vert) – http://gallica.bnf.fr/ark:/12148/bpt6k220169n

2.1.1 Catégories

Les catégories d'entités nommées sont les suivantes :

Catégories	Description	Exemples	Remarques
Personne	Nom de personne, patronyme, nom usuel, prénom, surnom, divinité…	Henri IV M. Dupont Franconi	Les noms d'auteur d'article seront si besoin distingués des autres

		Léa Jean Jaurès Uncle Sam	noms de personne (voir section 5)
Lieu	Nom de lieu géographique ou politique : ville, pays, Etat, province, département, toponyme, quartier, montagne, rivière, fleuve, océan, planète, monument, pont…	Paris Auvergne Amérique du sud Les Buttes Chaumont Le Pirée Loire	
Adresse	Adresse postale	10, passage du Panorama Quai de Seine, Paris	En option, selon prestation
Organisation	Administration, institution, gouvernement, ministère, entreprise, église, musée, université, hôpital, hôtel, marque, titre de journal/tv/radio, parti politique, syndicat	musée de la Marine ministère de la Défense Renault théâtre de l'Odéon ONU Louvre Titanic syndicat des cheminots du Mans	
Date	Toute date	11 novembre 10 avril 1912 9/10/1968 vendredi 8 1914	En option, selon prestation
Quantité	Toute quantité numérique, avec son éventuelle unité	100 % 125 euros 2 tonnes 325 km 125,48 £10	En option, selon prestation

Pour chaque prestation ou projet, une charte précisant l'identification attendue pour les différentes catégories d'entités nommées sera fournie par la BnF (règles syntaxiques, conventions, exceptions, ambiguïtés, etc.).

2.1.2 Référentiels d'autorités

La BnF pourra fournir ses référentiels d'autorités (cf. section 3.2) afin d'améliorer le processus de REN.

2.2 Corpus

Les corpus cible de la tâche de REN sont identifiés précisément par la BnF, en termes :

- de type documentaire : imprimés, périodiques, presse,
- de période historique : Ancien Régime, XIX^{e} siècle, XX^{e} siècle, etc.
- de genre : littérature, sciences humaines, presse quotidienne, presse spécialisée, etc.

2.3 Description

2.3.1 Dans le document numérique

Les entités nommées reconnues sont décrites dans les fichiers OCR du document numérique, sauf demande spécifique de la BnF. Le format ALTO BnF v2 est utilisé à cet effet, à l'aide du mécanisme de marquage décrit dans le « Référentiel OCR », section 5.10.1.

On utilise la syntaxe suivante pour décrire une entité nommée :

```
<NamedEntityTag ID="idx" TYPE="catégorie" LABEL="étiquette" />
```

Une étiquette est associée à un mot ou un groupe de mots (élément <String>) à l'aide de l'attribut TAGREFS et d'un identifiant d'étiquette :

```
< String ID="…"  HPOS="…"  VPOS="…" TAGREFS="idx"/>
```

Les identifiants d'entité nommée sont numérotés séquentiellement : TAG_NE001… TAG_NE*n*.

Les catégories d'entités nommées sont décrites avec l'attribut TYPE :

- Personne : PER
- Lieu : LOC
- Adresse : ADD
- Organisation : ORG
- Date : DATE
- Quantité : QTY

L'attribut LABEL contient la valeur textuelle de l'entité nommée (mot ou groupe de mots).


```
<Tags>
        <NamedEntityTag ID="TAG_NE001" TYPE="PER" LABEL="SÉRAPHIN"/>
        <NamedEntityTag ID="TAG_NE002" TYPE="ORG" LABEL="Conseil de l'Europe"/>
        …
</Tags>
<Layout> …
        <String ID="PAG_00000001_ST000071" … CONTENT="SÉRAPHIN."
            TAGREFS="TAG_NE001"/>
…
</Layout>
```

Pour une entité nommée composée de plusieurs mots, l'étiquette est associée à tous les mots :

```
<String ID="PAG_00000001_ST000080" … TAGREFS="TAG_NE002"
            CONTENT="Conseil" />
<String ID="PAG_00000001_ST000081" … TAGREFS="TAG_NE002"
            CONTENT="de" />
```

```
<String ID="PAG_00000001_ST000082" … TAGREFS="TAG_NE002"
                CONTENT="l'Europe" />
```

2.3.2 Dans le manifeste numérique

L'application d'un traitement REN est décrite dans le manifeste METS du document numérique par le biais d'un événement PREMIS de type « namedEntitiesRecognition » :

- type de traitement,
- agent ayant réalisé le traitement,
- logiciel utilisé, etc.

Le formalisme à utiliser est précisé dans le « Référentiel d'enrichissement des métadonnées, version METS ».


```
<digiprovMD ID="AMD.9" ADMID="AMD.11 AMD.12">
 <mdWrap MIMETYPE="text/xml" MDTYPE="PREMIS:EVENT">
  <xmlData>
        <premis:event>
                <premis:eventIdentifier>
                        <premis:eventIdentifierType>UUID</premis:eventIdentifierType>
                        <premis:eventIdentifierValue>9554c330-2f31-11e3-aa6e-
0800200c9a66</premis:eventIdentifierValue>
                </premis:eventIdentifier>
                <premis:eventType>namedEntitiesRecognition</premis:eventType>
                <premis:eventDateTime>2015-01-07T16:47:23+01:00</premis:eventDateTime>
                <premis:linkingAgentIdentifier>
                  <premis:linkingAgentIdentifierType>metsIdentifier</premis:linkingAgentIdentifierType>
                  <premis:linkingAgentIdentifierValue>agent.1</premis:linkingAgentIdentifierValue>
                  <premis:linkingAgentRole>implementer</premis:linkingAgentRole>
                </premis:linkingAgentIdentifier>

                <premis:linkingAgentIdentifier>
                  <premis:linkingAgentIdentifierType>metsIdentifier</premis:linkingAgentIdentifierType>
                  <premis:linkingAgentIdentifierValue>agent.2</premis:linkingAgentIdentifierValue>
                  <premis:linkingAgentRole>performer</premis:linkingAgentRole>
                </premis:linkingAgentIdentifier>
        </premis:event>
  </xmlData>
</mdWrap>
</digiprovMD>

<digiprovMD ID="AMD.11">
  <mdWrap MDTYPE="PREMIS:AGENT">
   <xmlData>
        <premis:agent>
          <premis:agentIdentifier>
                <premis:agentIdentifierType>producerIdentifier</premis:agentIdentifierType>
                <premis:agentIdentifierValue>agent.1</premis:agentIdentifierValue>
          </premis:agentIdentifier>
          <premis:agentName>SAFIG Madagascar</premis:agentName>
          <premis:agentType>organization</premis:agentType>
          <premis:agentNote>origine : groupement Safig</premis:agentType>
        </premis:agent>
        </xmlData>
  </mdWrap>
```

```
</digiprovMD>

<digiprovMD ID="AMD.12">
 <mdWrap MDTYPE="PREMIS:AGENT">
  <xmlData>
        <premis:agent>
         <premis:agentIdentifier>
          <premis:agentIdentifierType>producerIdentifier</premis:agentIdentifierType>
          <premis:agentIdentifierValue>agent.2</premis:agentIdentifierValue>
        </premis:agentIdentifier>
        <premis:agentName>NER</premis:agentName>
        <premis:agentType>software</premis:agentType>
        <premis:agentNote>version : 3.5.1</premis:agentNote>
        <premis:agnetNote>origine : Stanford</premis:agnetNote>
        </premis:agent>
  </xmlData>
 </mdWrap>
</digiprovMD>
```

2.4 Qualité

2.4.1 Mesure de la qualité

Le processus de REN est évalué lors de la phase de test à l'aide d'un corpus annoté (vérité terrain) représentatif des documents à traiter.

Ce corpus annoté est traité par le processus de REN et les taux qualité *rappel* et *précision* décrits ci-après sont calculés automatiquement.

Définitions de rappel et précision

Rappel : Le rappel est défini par le nombre d'entités pertinentes retrouvées au regard du nombre d'entités pertinentes que possède le corpus analysé. Un système doté d'un mauvais taux de rappel est dit « silencieux ».

Précision : La précision est le nombre d'entités pertinentes retrouvées rapporté au nombre total d'entités retrouvées. Un système doté d'un mauvais taux de précision est dit « bruyant ».

Ces métriques sont calculées pour chacune des catégories d'entités nommées à traiter :

- **Rappel** : le rappel exprime le rapport entre les entités nommées correctement identifiées pour une catégorie *i*, et les entités nommées de catégorie *i* présentes dans le corpus :

  Rappel*i* = nombre d'EN*i* correctement identifiées pour la catégorie *i* /
  nombre d'EN*i* de la catégorie *i*

- **Précision** : la précision exprime le rapport entre les entités nommées correctement identifiées pour une catégorie *i*, et les entités nommées identifiées pour une catégorie *i* :

  Précision*i* = nombre d'EN*i* correctement identifiées pour la catégorie *i* /
  nombre d'EN*i* identifiées pour la catégorie *i*

Ces métriques peuvent être calculées pour un document ou pour un corpus de documents.

POUR UN CORPUS DE DOCUMENTS, ON NE CALCULERA PAS LES METRIQUES A L'AIDE D'UNE MOYENNE DES METRIQUES DES DOCUMENTS.

On pourra également calculer les moyennes globales du rappel et de la précision sur l'ensemble des catégories (*n*) :

- **Rappel** : $\sum_{i=1}^{n}$ rappel$i$ / n
- **Précision** : $\sum_{i=1}^{n}$ précision$i$ / n

2.4.2 Corpus d'évaluation

Le corpus annoté est fourni par la BnF (environ 200 000 mots).

2.4.3 Evaluation de la qualité

Le niveau de qualité acceptable concernant la reconnaissance des entités nommées est caractérisé par des valeurs seuils pour les métriques de rappel et de précision pour chaque catégorie d'entités :

	Taux de rappel	Taux de precision	Remarque
Personne	$seuil_{PERr}$	$seuil_{PERp}$	
Lieu	$seuil_{LOCr}$	$seuil_{LOCp}$	
Organisation	$seuil_{ORGr}$	$seuil_{ORGp}$	
…			

Lors de l'évaluation en phase de test, les valeurs constatées dans le corpus à contrôler sont comparées aux valeurs attendues, qui sont spécifiques à chaque marché (cf. CCTP).

POUR LA TACHE DE REN, LA BNF N'ATTEND PAS DE MONTEE EN QUALITE MANUELLE DANS LE CAS OU UN PROCESSUS AUTOMATIQUE NE PERMETTRAIT PAS D'ATTEINDRE LES TAUX DE RAPPEL ET DE PRECISION ATTENDUS.

3. ALIGNEMENT DES ENTITES NOMMEES SUR LES REFERENTIELS BNF

3.1 Principe

L'alignement des entités nommées identifiées lors du processus de REN consiste à les mettre en relation avec les référentiels d'autorités de la BnF :

- Pour la catégorie **Personne**, référentiel d'autorités BnF Personnes : 1,5 million de notices d'autorités,

  http://www.bnf.fr/fr/professionnels/autorites_bnf/s.personnes.html

  Pour la catégorie **Lieu**, référentiel d'autorités BnF Noms géographiques : 115 000 notices d'autorités, http://www.bnf.fr/fr/professionnels/ autorites_ bnf/s.noms_geographiques_bnf.html

- Pour la catégorie **Organisation**, référentiels d'autorités BnF Collectivités : 350 000 notices d'autorités,

  http://www.bnf.fr/fr/professionnels/autorites_bnf/s.collectivites.html

Ce traitement est automatique. Il produit en sortie, pour chaque entité nommée, un des résultats suivants :

- Alignement non réalisé : l'entité nommée n'est pas présente dans le référentiel d'autorités ou elle n'a pu y être détectée.

Pour ce cas, l'alignement de l'entité nommée avec un autre référentiel public, par exemple GEONAMES pour les lieux, pourra être envisagé.

- Alignement réalisé : l'entité nommée est présente dans un référentiel d'autorités BnF et elle est liée à la notice d'autorité à l'aide de l'identifiant de cette dernière (identifiant pérenne Ark).

3.2 Formats et mise à disposition

Ces données d'autorités sont disponibles selon les formats suivants :

- aux formats Intermarc ou Unimarc (http://www.bnf.fr/fr/professionnels/format_ intermarc/s.intermarc_presentation.html),
- au format InterXmarc (xmélisation des notices Intermarc),
- en RDF (*Resource description Framework* ; rdf-nt, rdf-xml, rdf-n3), via le projet data.bnf.fr. La documentation et le modèle RDF sont disponibles à l'adresse : http://data.bnf.fr/semanticweb. Ces notices sont décrites en RDF en utilisant les vocabulaires existants :
  - SKOS pour décrire les éléments principaux de toutes les notices d'autorité (personnes, œuvres, lieux) : formes préférées, formes rejetées, lien à la notice du catalogue, liens aux thèmes associés, par exemple.
  - FOAF , RDA et BIO pour décrire les informations spécifiques aux personnes et organisations ;
  - GEO et Geonames pour les informations spécifiques aux lieux.

EXEMPLE

InterXmarc

```
…
<datafield tag="045" ind1=" " ind2=" ">
    <subfield code="a" Sens="Auteur">a</subfield>
</datafield>
<datafield tag="100" ind1=" " ind2=" ">
    <subfield code="w"> 0  b     </subfield>
    <subfield code="a">Michel</subfield>
    <subfield code="m">Wilhelm</subfield>
    <subfield code="d">1877-1942</subfield>
</datafield>
<datafield tag="600" ind1=" " ind2=" ">
    <subfield code="a">Ecrivain</subfield>
</datafield>
<datafield tag="603" ind1=" " ind2=" ">
    <subfield code="a">Metz</subfield>
    <subfield code="b">Darmstadt (Allemagne)</subfield>
</datafield>
<datafield tag="610" ind1=" " ind2=" ">
    <subfield code="a">Das Leben Friedrich Hölderlins / Wilhelm Michel, 1963</subfield>
</datafield> …
```

RDF

http://data.bnf.fr/11915247/francois_mauriac/rdf.xml

…

```
<rdf:Description rdf:about="http://data.bnf.fr/ark:/12148/cb11915247g#foaf:Person">
   <bnf-onto:firstYear rdf:datatype="http://www.w3.org/2001/XMLSchema#integer">1885</bnf-
onto:firstYear>
<rdagroup2elements:fieldOfActivityOfThePerson>Littératures</rdagroup2elements:fieldOfActivityOfThePe
rson>
   <rdagroup2elements:fieldOfActivityOfThePerson rdf:resource="http://dewey.info/class/800/"/>
   <rdagroup2elements:dateOfBirth rdf:resource="http://data.bnf.fr/date/1885/"/>
   <rdagroup2elements:dateOfDeath rdf:resource="http://data.bnf.fr/date/1970/"/>
   <rdf:type rdf:resource="http://xmlns.com/foaf/0.1/Person"/>
   <bnf-onto:lastYear rdf:datatype="http://www.w3.org/2001/XMLSchema#integer">1970</bnf-
onto:lastYear>
   <rdagroup2elements:placeOfBirth>Bordeaux</rdagroup2elements:placeOfBirth>
   <rdagroup2elements:countryAssociatedWithThePerson
rdf:resource="http://id.loc.gov/vocabulary/countries/fr#"/>
   <rdagroup2elements:languageOfThePerson rdf:resource="http://id.loc.gov/vocabulary/iso639-2/fre"/>
   <owl:sameAs rdf:resource="http://www.idref.fr/027018741"/>
   <owl:sameAs rdf:resource="http://dbpedia.org/resource/François_Mauriac"/>
   <owl:sameAs rdf:resource="http://viaf.org/viaf/9850407"/>
   <foaf:familyName>Mauriac</foaf:familyName>
   <foaf:givenName>François</foaf:givenName>
   <rdagroup2elements:placeOfDeath>Paris</rdagroup2elements:placeOfDeath>
   <foaf:depiction rdf:resource="http://gallica.bnf.fr/ark:/12148/bpt6k8815216x/f3.item.thumbnail"/>
   <foaf:depiction
rdf:resource="http://commons.wikimedia.org/wiki/Special:FilePath/François_Mauriac_(1932).jpg?width=30
0"/>
   <foaf:gender>male</foaf:gender>
   <foaf:birthday>10-11</foaf:birthday>
   <bio:birth>1885-10-11</bio:birth>
   <rdagroup2elements:biographicalInformation>Membre de l'institut, Académie française (1933-1970),
prix Nobel de littérature 1952. - A écrit du 14 mars au 11 juillet 1914, douze articles dans le "Journal de
Clichy" sous le pseudonyme "François Sturel". - A signé pendant la Résistance sous le pseudonyme
"Forez"</rdagroup2elements:biographicalInformation>
   <foaf:page rdf:resource="http://data.bnf.fr/11915247/francois_mauriac/"/>
   <foaf:name>François Mauriac</foaf:name>
   <bio:death>1970-09-01</bio:death>
 </rdf:Description> …
```

Ces référentiels sont disponibles sous différentes formes :

- Unimarc ou InterMarc : fichier cumulatif des notices,
- InterXmarc : un fichier XML par notice,
- RDF
    - négociation de contenu en RDF et en JSON,
    - dumps RDF,
    - SPARQL endpoint : http://data.bnf.fr/sparql.

Voir aussi : http://www.bnf.fr/fr/professionnels/recuperation_donnees_bnf.html

3.3 Description

3.3.1 Dans le document numérique

Le résultat de la tâche d'alignement est décrit dans le fichier ALTO (cf. section 2.3 et section 5.10.1 du « Référentiel OCR »). On utilise les attributs URI et DESCRIPTION :

- URI : identifiant de l'entité nommée dans le référentiel utilisé (identifiant ark pour un référentiel BnF, URL pour un référentiel externe),
- DESCRIPTION : forme de référence de l'entité nommée dans le référentiel utilisé (forme vedette pour un référentiel BnF, forme en français pour un référentiel externe multilingue par exemple).

Référentiel d'autorités BnF

```
<NamedEntityTag ID="TAG_NE003" TYPE="ORG" LABEL="Louvre"
DESCRIPTION="Musée du Louvre (Paris)" URI="ark:/12148/cb11865019j" />
```

Autres référentiels :

```
<NamedEntityTag ID="TAG_NE003" TYPE="LOC" LABEL="Europe"
DESCRIPTION="Europe"  URI="http://www.geonames.org/2988507" />
```

3.3.2 Dans le manifeste numérique

L'application d'un traitement d'alignement est décrite dans le manifeste METS du document numérique.

3.4 Qualité

3.4.1 Mesure de la qualité

Le processus d'alignement est évalué lors de la phase de test à l'aide d'un corpus annoté (vérité terrain) représentatif des documents à traiter.

Ce corpus annoté est traité et les métriques qualité *rappel* et *précision* décrites ci-après sont calculées automatiquement, pour chacune des catégories d'entités nommées à aligner :

- **Rappel** : le rappel exprime le rapport entre les entités nommées présentes dans le corpus et correctement alignées pour une catégorie *i*, et les entités nommées de catégorie *i* présentes dans le corpus et dans le référentiel :

  Rappel*i* = nombre d'EN*i* correctement alignées pour la catégorie *i* /
  nombre d'EN*i* de la catégorie *i* dans le référentiel

- **Précision** : la précision exprime le rapport entre les entités nommées nommées présentes dans le corpus et correctement alignées pour une catégorie *i*, et les entités nommées alignées pour une catégorie *i* :

  Précision*i* = nombre d'EN*i* correctement alignées pour la catégorie *i* /
  nombre d'EN*i* alignées pour la catégorie *i*

Ces métriques peuvent être calculées pour un document complet ou pour un corpus de documents.

On pourra également calculer les moyennes globales du rappel et de la précision sur l'ensemble des catégories (*n*) :

- **Rappel** : $\sum^{n}_{i=1}$ rappel*i* / n **Précision** : $\sum^{n}_{i=1}$ précision*i* / n

3.4.2 Evaluation de la qualité

Le niveau de qualité acceptable concernant l'alignement des entités nommées est caractérisé par des valeurs seuils pour les métriques de rappel et de précision pour chaque catégorie d'entités :

	Taux de rappel	Taux de précision	Remarque
Personne	$seuil_{PERr}$	$seuil_{PERp}$	
Lieu	$seuil_{LOCr}$	$seuil_{LOCp}$	
Organisation	$seuil_{ORGr}$	$seuil_{ORGp}$	
…			

Lors du contrôle, les valeurs constatées dans le corpus à contrôler sont comparées aux valeurs attendues, qui sont spécifiques à chaque marché (cf. CCTP).

ATTENTION

POUR LES TACHES D'ALIGNEMENT DES ENTITES NOMMEES, LA BNF N'ATTEND PAS DE MONTEE EN QUALITE MANUELLE DANS LE CAS OU UN PROCESSUS AUTOMATIQUE NE PERMETTRAIT PAS D'ATTEINDRE LES TAUX DE RAPPEL ET DE PRECISION ATTENDUS.

4. RECONNAISSANCE DES TITRES

4.1 Principe

La tâche de reconnaissance des titres consiste à identifier les titres, sous-titres et titres intérieurs des contenus textuels d'un fascicule de presse.

Cette tâche consiste également à corriger si nécessaire le texte de ces titres. La qualité attendue pour cette correction est le taux OCR qualité éditoriale (cf. « Référentiel OCR », section 4.1.3).

La tâche est limitée à l'identification de deux niveaux de titre (titres d'article et titres de section), sauf demande explicite de la BnF pour certains titres de presse pour lesquels un seul niveau sera demandé (titres d'article).

Pour chaque titre de presse à traiter, la BnF fournira une charte explicitant la mise en forme graphique des titres à reconnaître. Cette charte sera basée sur un fascicule annoté représentatif de l'ensemble du corpus à numériser (ou plusieurs si le principe de maquette du titre a évolué au fil du temps).

EXEMPLE

Rouge : titre – Orange : sous-titre – Bleu : titre intérieur

Exemple 1 : L'Œuvre


L'ŒUVRE

DERNIÈRE ÉDITION — 5 H. DU MATIN — UN FRANC

L'émigration à l'extérieur ou à l'intérieur est la plus mauvaise action qu'on puisse commettre.
ERNEST RENAN.

Plusieurs divisions soviétiques encerclées ou anéanties au sud-ouest de Narva

Entre le Boug, le Dniestr et le Prout les Bolcheviks ont été repoussés avec de lourdes pertes

ÊTRE MARIN c'est avoir un métier

Les dispositions adoptées pour la diminution de la consommation électrique sont maintenues cette semaine

Notre agriculture doit produire plus et mieux

Ce que nous dit M. Marcel Braibant

45 valises retrouvées au domicile d'un ami de Maurice PETIOT

UN TERRORISTE condamné à mort par une cour martiale française est exécuté

Vers un Conseil supérieur de la solidarité nationale

L'enquête du F.N.T. sur le travail des femmes

Les bons patrons et les mauvais...

141 bombardiers détruits au cours d'un raid sur Nuremberg

DOMINIQUE SORDET ET LA DÉMOCRATIE

par ★★★

24 HEURES EN TERRE D'ISLAM

Derrière les fenêtres grillagées une vie patriarcale

Une « Frégoli » du téléphone

Alerte à Londres hier matin

LA RÉCEPTION par Marcel Déat des représentants ouvriers et patronaux de l'industrie métallurgique de Marseille

Exemple 2 : La Presse Illustrée

LA PRESSE ILLUSTRÉE

JOURNAL QUOTIDIEN

LES COSAQUES A PARIS

PONSON DU TERRAIL

Exemple 3 : Le Petit Parisien

Le Petit Parisien

ABONNEMENTS

Direction : 18, rue d'Enghien

ANNONCES

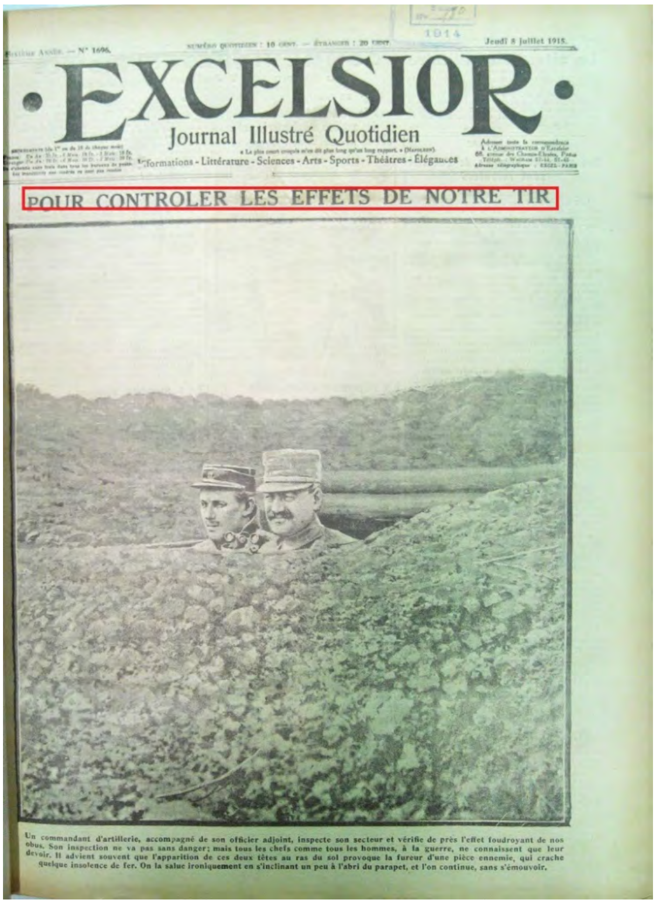
LE MONT-DE-PIÉTÉ

LE MASQUE ÔTÉ

LA LIBERTÉ INDIVIDUELLE

BABIOLE

Exemple 4 : L'Excelsior



EXCELSIOR

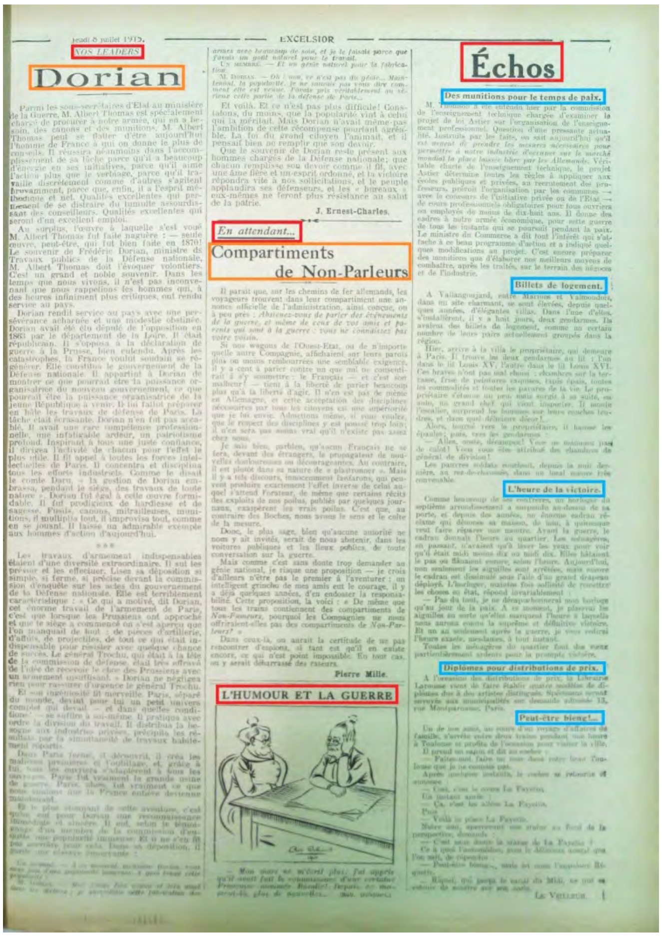
Journal Illustré Quotidien

Informations - Littérature - Sciences - Arts - Sports - Théâtres - Élégances

Jeudi 8 juillet 1915.

POUR CONTROLER LES EFFETS DE NOTRE TIR

Un commandant d'artillerie, accompagné de son officier adjoint, inspecte son secteur et vérifie de près l'effet foudroyant de nos obus. Son inspection ne va pas sans danger; mais tous les chefs comme tous les hommes, à la guerre, ne connaissent que leur devoir. Il advient souvent que l'apparition de ces deux têtes au ras du sol provoque la fureur d'une pièce ennemie, qui crache quelque insolence de fer. On la salue ironiquement en s'inclinant un peu à l'abri du parapet, et l'on continue, sans s'émouvoir.



Jeudi 8 juillet 1915. — EXCELSIOR

NOS LEADERS

Dorian

J. Ernest-Charles.

En attendant...

Compartiments de Non-Parleurs

Pierre Mille.

L'HUMOUR ET LA GUERRE

Échos

Des munitions pour le temps de paix.

Billets de logement.

L'heure de la victoire.

Diplômes pour distributions de prix.

Peut-être bien!...

Le Veilleur.

NOTES

La titraille de presse est caractérisée par l'utilisation d'une grande variété de polices de caractères et de mises en forme graphiques : polices fantaisies, corps des caractères, enrichissements typographiques (gras, italique) ou graphiques (blanc sur fond noir, noir sur iconographie, etc.). Cette caractéristique doit être prise en compte par la tâche de reconnaissance des titres.

EXEMPLE

4.2 Description

4.2.1 Dans le document numérique

Le mécanisme de marquage décrit dans le « Référentiel OCR », sections 4.2.3 et 5.10.1 est utilisé. Ce mécanisme nécessite la version 2 du format ALTO BnF.

On utilise la syntaxe suivante pour identifier une étiquette de titre :

```
<StructureTag ID="TAG_ST001" TYPE="Structural" LABEL="title"
DESCRIPTION="CHRONIQUE"/>
```

- Le titre complet est stocké dans l'attribut DESCRIPTION.

- Le niveau hiérarchique du titre est identifié à l'aide de l'attribut LABEL :
  - Pour les monographies, les niveaux de titres sont codés ainsi :
    - titre de niveau 1 (partie, chapitre, etc.) : title1
    - titre de niveau 2 (chapitre, section) : title2
    - titre de niveau 3 (section, sous-section) : title3
    - etc.
  - Pour les documents presse ou périodique, les niveaux de titres sont codés ainsi :
    - titre d'article : title, avec éventuellement son sous-titre : subtitle
    - titre intérieur : insidetitle

La distinction entre titre et sous-titre étant parfois difficile à déterminer sans ambiguïté, elle ne sera pas contrôlée.

Une étiquette de titre est associée à un bloc de texte ou à une ligne de texte à l'aide de l'attribut TAGREFS et d'un identifiant d'étiquette :

```
<TextBlock ID="PAG_00000001_TB000010"… TAGREFS="TAG_ST001" />
```

Pour un titre composé sur plusieurs blocs ALTO, l'étiquette de titre est associée tous les blocs :

```
<TextBlock ID="PAG_00000001_TB000010"… TAGREFS="TAG_ST001" />
<TextBlock ID="PAG_00000001_TB000011"… TAGREFS="TAG_ST001" />
```

L'ordre d'apparition des balises XML <StructureTag> dans le fichier ALTO doit être identique à celui des titres dans la page, lui-même lié à l'ordre de lecture. Les identifiants des balises ("TAG_ST001", ("TAG_ST001", etc.) sont numérotés en séquence, selon ce même ordre.

4.2.2 Dans le manifeste numérique

Table de navigation

L'ensemble des titres du fascicule sont reproduits dans le manifeste METS du fascicule, à l'aide d'une carte de structuration logique :

```
<mets:structMap LABEL="Structure logique" TYPE="LOGICAL">
```

Chaque titre est identifié par une balise   :

```
<div ID="…" TYPE="…" ORDER="…" LABEL="…">
    …
</div>
```

avec les attributs suivants :

- titre complet : LABEL,

- ordre d'apparition du titre dans la page (voir note ci-dessus) : ORDER,
- nature du titre : TYPE,
    - titre d'article : TYPE="HEADING"
    - titre intérieur : TYPE="INSIDEHEADING"

Le contenu du titre est identifié par une ou plusieurs sous-balise(s)   :

```
<div ID="…" TYPE="… " LABEL="…">
    <div ID="…" TYPE="TITLE">

    </div>
    …
</div>
```

avec les attributs suivants :

- nature du titre : TYPE,
    - titre d'article ou de section : TYPE="TITLE"
    - sous-titre : TYPE="SUBTITLE"

Les références du fichier et du bloc ALTO contenant le titre sont décrites dans des balises <fptr> </fptr> et <area></area> :

```
<fptr>
        <area BETYPE="IDREF" FILEID="ALTO00001" BEGIN="P1_TB00021"/>
</fptr>
```

Si plusieurs blocs ALTO sont concernés, on utilisera une balise <seq> </seq> :

```
<seq>
        <area BETYPE="IDREF" FILEID="ALTO00001" BEGIN="P1_TB00019"/>
        <area BETYPE="IDREF" FILEID="ALTO00001" BEGIN="P1_TB00020"/>
</seq>
```

Exemple 1 ci-avant : L'Œuvre

```
<structMap LABEL="Structure logique" TYPE="LOGICAL">
<div ID="DIVL1" TYPE="NEWSPAPER" LABEL="Le Petit Parisien"> -- niveau journal
  <div ID="DIVL2" TYPE="VOLUME" DMDID="…" LABEL="Le Petit Parisien">  -- niveau volume (option.)
    <div ID="DIVL3" TYPE="ISSUE" DMDID="…" LABEL="Le Petit Parisien n° 2643 22.01.1884"> -- niveau fascicule
        <div ID="DIVL4" TYPE="TITLE_SECTION"> -- contenu de l'ours
        <div ID="…" TYPE="CONTENT"> -- contenu du fascicule
            <div ID="…" TYPE="HEADING" ORDER="1" LABEL="Plusieurs divisions soviétiques encerclées ou
    anéanties" >  -- titre d'article avec sous-titre
                <div ID="…" TYPE="TITLE">
                  <fptr>
                    <seq> -- titre composé sur deux blocs
                        <area BETYPE="IDREF" FILEID="ALTO00001" BEGIN="P1_TB00019"/>
                        <area BETYPE="IDREF" FILEID="ALTO00001" BEGIN="P1_TB00020"/>
                    </seq>
                  </fptr>
                </div>
                <div ID="…" TYPE="SUBTITLE" LABEL="Entre le Boug, le Dniestr et le Prout les Bolcheviks
    ont été repoussés avec de lourdes pertes"> -- sous-titre
                  <fptr>
                        <area BETYPE="IDREF" FILEID="ALTO00001" BEGIN="P1_TB00021"/>
                  </fptr>
                </div>
```

```
                </div>
                <div ID="…" TYPE="INSIDEHEADING " ORDER="2" LABEL="Nouveaux échecs des Anglo-
        Américains en Italie">  -- titre de section
                        <div ID="…" TYPE="TITLE">
                          <fptr>
                                <area BETYPE="IDREF" FILEID="ALTO00001" BEGIN="P1_TB00032"/>
                          </fptr>
                        </div>
                </div>


                <div ID="…" TYPE="HEADING" ORDER="3"  LABEL="141 bombardiers détruits au cours d'un raid sur
        Nuremberg" > -- titre d'article
                        <div ID="…" TYPE="TITLE">
                          <fptr>
                                <area BETYPE="IDREF" FILEID="ALTO00001" BEGIN="P1_TB00035"/>
                          </fptr>
                        </div>
                </div>
…
   </div>  -- content
…
```

LES TITRES DE SECTION N'ONT PAS A ETRE INCLUS DANS LA DIVISION CORRESPONDANT AU TITRE DE L'ARTICLE. UNE TELLE STRUCTURATION IMPLIQUERAIT UNE RECONNAISSANCE DE LA STRUCTURE DES ARTICLES (VOIR SECTION 6).

Opération de traitement

L'application d'un traitement de reconnaissance des titres (headlinesDetection) est décrite dans le manifeste METS du document numérique.

On fera le distinguo entre un processus simple d'extraction d'information (un seul atelier à l'aide d'un seul outil) et un processus complexe (plusieurs événements à décrire). Cf. « Référentiel d'enrichissement des métadonnées, version METS », section 8.6.2.

Relation avec la reconnaissance des articles

Lorsque qu'une tâche de reconnaissance des articles est également demandée (cf. section 6), les titres sont décrits en tant que métadonnées descriptives des articles (élément <mods:title>), dans une section <dmdSec>.

Exemple 1 : La Croix

http://gallica.bnf.fr/ark:/12148/bpt6k220169n

Métadonnées descriptives

```
                <mets:dmdSec ID="MODSMD_ARTICLE2">
                <mets:mdWrap MIMETYPE="text/xml" MDTYPE="MODS" LABEL="Métadonnées bibliographiques de
l'article">
                  <mets:xmlData>
                        <mods:mods>
                                <mods:titleInfo ID="MODSMD_ARTICLE1_TI1" xml:lang="fr">
```

```
                <mods:title>LA LOI INTERPRETEE PAR M. COMBES</mods:title>
            </mods:titleInfo>
            <mods:language>
                <mods:languageTerm type="code"
                    authority="rfc3066">fr</mods:languageTerm>
            </mods:language>
        </mods:mods>
    </mets:xmlData>
</mets:mdWrap>
</mets:dmdSec>
```

Carte de structure logique

```
<mets:div ID="DIVL11" TYPE="ARTICLE" DMDID="MODSMD_ARTICLE2"
    LABEL="LA LOI INTERPRETEE PAR M. COMBES">
```

4.3 Qualité

4.3.1 Qualité de la reconnaissance

Mesure de la qualité

Le processus d'alignement est évalué lors de la phase de test (à l'aide d'un corpus annoté représentatif des documents à traiter) et en production.

La mesure de la qualité de la tâche de reconnaissance des titres est réalisée à l'aide de deux métriques, qui sont calculées pour chacune des deux catégories de titres à identifier (titres + sous-titres ; titres intérieurs) dans un fascicule donné :

- **Rappel** : le rappel exprime le rapport entre les titres correctement reconnus dans le fascicule pour une catégorie de titre *i*, et les titres de catégorie *i* présents dans le fascicule :

  Rappel*i* = nombre de titres correctement reconnus pour la catégorie *i* / nombre de titres de la catégorie *i*

- **Précision** : la précision exprime le rapport entre les titres correctement reconnus dans le fascicule pour une catégorie de titre *i*, et les titres reconnus pour une catégorie *i* :

  Précision*i* = nombre de titres correctement reconnus pour la catégorie *i* / nombre de titres reconnus de la catégorie *i*

Ces mesures s'appliquent à un document complet (fascicule).

ON NE CALCULERA PAS LES METRIQUES DU DOCUMENT PAR UNE MOYENNE DES METRIQUES DES PAGES.

Evaluation de la qualité

Le niveau de qualité acceptable concernant la reconnaissance des titres d'un fascicule est caractérisé par des valeurs seuils pour les deux métriques et pour les deux catégories de titres :

	Taux de rappel	Taux de précision	Remarque

Titre d'article (titre, sous-titre)	$seuil_{ar}$	$seuil_{ap}$	La confusion entre titre et sous-titre n'est pas un motif d'erreur.
Titre intérieur	$seuil_{ir}$	$seuil_{ip}$	

Lors du contrôle, les valeurs constatées dans le fascicule à contrôler sont comparées aux valeurs attendues, qui sont spécifiques à chaque marché (cf. CCTP). Les valeurs constatées doivent être supérieures ou égales aux valeurs attendues.

POUR LA TACHE DE RECONNAISSANCE DES TITRES, LA BNF ATTEND UNE MONTEE EN QUALITE MANUELLE DANS LE CAS OU UN PROCESSUS AUTOMATIQUE NE PERMETTRAIT PAS D'ATTEINDRE LES TAUX DE RAPPEL ET DE PRECISION ATTENDUS.

4.3.2 Qualité de la transcription

Mesure de la qualité

La mesure de la qualité de la tâche de transcription des titres est réalisée à l'aide d'un taux qualité OCR calculé au mot.

Le taux qualité OCR sera calculé par le rapport des mots de titre erronés relativement au nombre total de mots composant les titres du document :

Taux qualité OCR (%) = (1 - nombre de mots erronés / nombre de mots des titres) * 100

Cette mesure s'applique à un document complet (fascicule).

Evaluation de la qualité

Lors du contrôle de qualité, le taux qualité OCR des titres constaté dans le fascicule à contrôler est comparé à un taux OCR qualité éditoriale (cf. « Référentiel OCR », section 4.1.3), qui est spécifique à chaque marché (cf. CCTP). Le taux qualité OCR des titres doit être supérieur ou égal à cette valeur attendue.

5. RECONNAISSANCE DES SIGNATURES D'ARTICLE

5.1 Principe

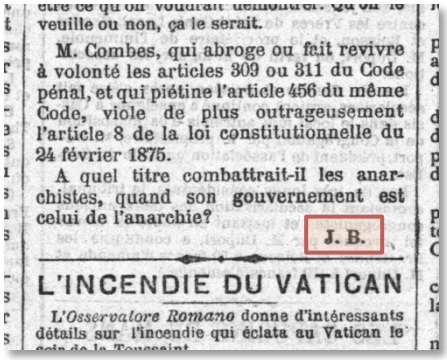
La tâche de reconnaissance des signatures consiste à identifier les noms des auteurs des contenus textuels (principalement les articles et feuilletons) d'un fascicule de presse.

Cette tâche consiste également à corriger si nécessaire le texte de ces signatures. La qualité attendue pour cette correction est le taux OCR qualité éditoriale (cf. « Référentiel OCR », section 4.1.3).

Pour chaque titre de presse à traiter, la BnF fournira une charte explicitant la mise en forme graphique des signatures à reconnaître. Cette charte sera basée sur un fascicule annoté représentatif de l'ensemble du corpus à numériser (ou plusieurs si le principe de maquette du titre a évolué au fil du temps).

EXEMPLE

Exemple 1 : La Croix



veuille ou non, ça le serait.

M. Combes, qui abroge ou fait revivre à volonté les articles 309 ou 311 du Code pénal, et qui piétine l'article 456 du même Code, viole de plus outrageusement l'article 8 de la loi constitutionnelle du 24 février 1875.

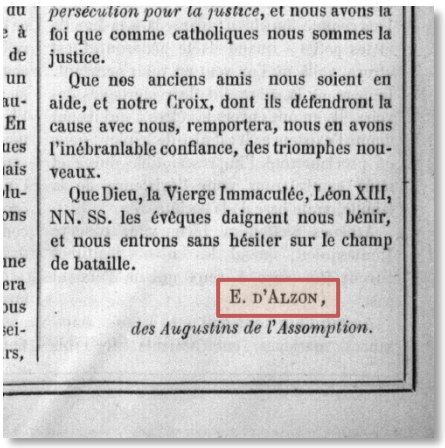
A quel titre combattrait-il les anarchistes, quand son gouvernement est celui de l'anarchie?

J. B.

L'INCENDIE DU VATICAN

L'*Osservatore Romano* donne d'intéressants détails sur l'incendie qui éclata au Vatican le

http://gallica.bnf.fr/ark:/12148/bpt6k220169n



persécution pour la justice, et nous avons la foi que comme catholiques nous sommes la justice.

Que nos anciens amis nous soient en aide, et notre Croix, dont ils défendront la cause avec nous, remportera, nous en avons l'inébranlable confiance, des triomphes nouveaux.

Que Dieu, la Vierge Immaculée, Léon XIII, NN. SS. les évêques daignent nous bénir, et nous entrons sans hésiter sur le champ de bataille.

E. D'ALZON,

des Augustins de l'Assomption.

http://gallica.bnf.fr/ark:/12148/bpt6k503572j/f9.image

Exemple 2 : L'Humanité

ploté contre l'agent du Ministère des Finances au *Matin* aussi. C'est sans aucune importance.

Votre très obéissant serviteur,

A. RAFFALOVITCH.

Après cela, on peut tourner la page.

Boris SOUVARINE.

Une lettre de M. Finot

Nous avons reçu hier une lettre de M. Finot, fils de M. Jean Finot, mis en cause dans une lettre de M. Raffalovitch.

http://gallica.bnf.fr/ark:/12148/bpt6k4009700

Exemple 3 : Le Petit Parisien

Le P

ABONNEMENTS

PARIS ET DÉPARTEMENTS : Trois mois, 5 fr. Six mois

AUJOURD'HUI, à notre troisième page, dans *les Mille et Une Nouvelles*, nous publions :

HISTOIRE D'UN NOYÉ

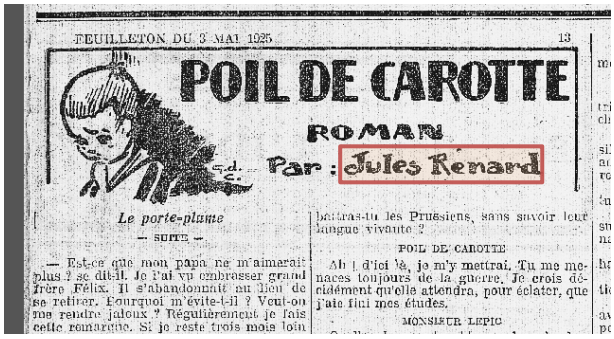
Par Alexis Bouvier

MONACO EN DEUIL

Au milieu des louangeuses nécrologies que les journaux boulevardiers publient dans leurs Echos, après la nouvelle de la mort du prince Auguste de Saxe-Cobourg-

http://gallica.bnf.fr/ark:/12148/bpt6k4718308

Exemple 4 : L'Humanité



FEUILLETON DU 3 MAI 1925 13

POIL DE CAROTTE

ROMAN

Par : Jules Renard

Le porte-plume

— SUITE —

— Est-ce que mon papa ne m'aimerait plus ? se dit-il. Je l'ai vu embrasser grand frère Félix. Il s'abandonnait au lieu de se retirer. Pourquoi m'évite-t-il ? Veut-on me rendre jaloux ? Régulièrement je fais cette remarque. Si je reste trois mois loin

battras-tu les Prussiens, sans savoir leur langue vivante ?

POIL DE CAROTTE

Ah ! d'ici là, je m'y mettrai. Tu me menaces toujours de la guerre. Je crois décidément qu'elle attendra, pour éclater, que j'aie fini mes études.

MONSIEUR LEPIC

5.2 Description

5.2.1 Dans le document numérique

Le mécanisme de marquage décrit dans le « Référentiel OCR », sections 4.2.3 et 5.10.1 est utilisé. Ce mécanisme nécessite la version 2 du format ALTO BnF.

On utilisera la syntaxe suivante pour identifier une étiquette de signature :

```
<RoleTag ID="TAG_author" LABEL="author"/>
```

Une étiquette de signature est associée à un mot ou à un groupe de mots à l'aide de l'attribut TAGREFS et d'un identifiant d'étiquette :

```
<String ID="…"  … CONTENT="nom" TAGREFS="TAG_author" />
```

Pour une signature composée de plusieurs mots, l'étiquette est associée à tous les mots :

```
<String ID="…" CONTENT="prénom" TAGREFS="TAG_author"/>
<String ID="…" CONTENT="nom" TAGREFS="TAG_author"/>
```

Relation avec la reconnaissance des entités nommées

Lorsque qu'une tâche de REN est également demandée (cf. section 2), les signatures sont décrites en tant qu'entité nommée de type « personne » (à l'aide d'un élément <NamedEntityTag>) et en tant qu'auteur (avec l'étiquette <RoleTag>).

Exemple 1 : La Croix

http://gallica.bnf.fr/ark:/12148/bpt6k220169n

```
<Tags>
        <NamedEntityTag ID="TAG_NE0010" TYPE="PER" LABEL="J. B."/>
        <RoleTag ID="TAG_author" LABEL="author"/>
        …
</Tags>
<Layout> …
  <TextBlock ID="PAG_00000001_TB000055" … >
        <TextLine ID="PAG_00000001_TL000230" …">
                <String ID="PAG_00000001_ST001401" CONTENT="J. B."
                        TAGREFS="TAG_NE0010 TAG_author"/>
        </TextLine>
  </TextBlock>
…
</Layout>
```

Exemple 2 : L'Humanité

http://gallica.bnf.fr/ark:/12148/bpt6k220169n

```
<Tags>
        <NamedEntityTag ID="TAG_NE0020" TYPE="PER" LABEL="Boris Souvarine"/>
        <RoleTag ID="TAG_author" LABEL="author"/>
        …
</Tags>
<Layout> …
  <TextBlock ID="PAG_00000001_TB000058" … >
        <TextLine ID="PAG_00000001_TL0000120" …">
                <String ID="PAG_00000001_ST00141" CONTENT="Boris"
                        TAGREFS="TAG_NE0020 TAG_author"/>
                <String ID="PAG_00000001_ST00142" CONTENT="Souvarine."
                        TAGREFS="TAG_NE0020 TAG_author"/>
        </TextLine>
  </TextBlock>
```

...

</Layout>

5.2.2 Dans le manifeste numérique

Opération de traitement

L'application du traitement de reconnaissance des signatures (authorsDetection) est décrite dans le manifeste METS du document numérique (cf. section 2.3.2).

On fera le distinguo entre un processus simple d'extraction d'information (un seul atelier à l'aide d'un seul outil) et un processus complexe (plusieurs événements à décrire). Cf. « Référentiel d'enrichissement des métadonnées, version METS », section 8.6.2.

Relation avec la reconnaissance des articles

Lorsque qu'une tâche de reconnaissance des articles est également demandée (cf. section 6), les signatures sont décrites en tant que métadonnées descriptives des articles (élément <mods:name>), dans une section <dmdSec>.

Exemple 1 : La Croix

http://gallica.bnf.fr/ark:/12148/bpt6k220169n

Métadonnées descriptives

```
<mets:dmdSec ID="MODSMD_ARTICLE2">
<mets:mdWrap MIMETYPE="text/xml" MDTYPE="MODS" LABEL="Métadonnées bibliographiques de
l'article">
 <mets:xmlData>
        <mods:mods>
                <mods:titleInfo ID="MODSMD_ARTICLE1_TI1" xml:lang="fr">
                        <mods:title>LA LOI INTERPRETEE PAR M. COMBES</mods:title>
                </mods:titleInfo>
                <mods:language>
                        <mods:languageTerm type="code"
                                authority="rfc3066">fr</mods:languageTerm>
                </mods:language>
                <mods:name type="personal" >
                        <mods:namePart>J. B.</mods:namePart>
                        <mods:role>
                          <mods:roleTerm type="text"
                                authority="marcrelator">Author</ mods:roleTerm>
                        </mods:role>
                </mods:name>
        </mods:mods>
 </mets:xmlData>
</mets:mdWrap>
</mets:dmdSec>
```

Carte de structure logique

```
<mets:div ID="DIVL11" TYPE="ARTICLE" DMDID="MODSMD_ARTICLE2"
      LABEL="LA LOI INTERPRETEE PAR M. COMBES">
```

Lorsque qu'une tâche d'alignement des entités nommées est également demandée (cf. section 3), l'alignement est décrit dans l'élément <mods:name> (attribut valueURI).

Exemple 2 : L'Humanité

http://gallica.bnf.fr/ark:/12148/bpt6k4009700

```
<mets:dmdSec ID="MODSMD_ARTICLE1">
<mets:mdWrap MIMETYPE="text/xml" MDTYPE="MODS" LABEL=" Métadonnées bibliographiques de
l'article ">
 <mets:xmlData>
        <mods:mods>
                <mods:titleInfo ID="MODSMD_ARTICLE1_TI1" xml:lang="fr">
                        <mods:title> M. BUNAU-VARILLA ET SA COMPTABILITE</mods:title>
                </mods:titleInfo>
                <mods:language>
                        <mods:languageTerm type="code"
                                authority="rfc3066">fr</mods:languageTerm>
                </mods:language>
                <mods:name type="personal"
                                authority="bnf"
                                authorityURI="http://catalogue.bnf.fr"
                                valueURI="ark:/12148/cb119252826">
                        <mods:namePart>Boris SOUVARINE</mods:namePart>
                        <mods:role>
                           <mods:roleTerm type="text"
                                authority="marcrelator">Author</ mods:roleTerm>
                        </mods:role>
                </mods:name>
        </mods:mods>
 </mets:xmlData>
</mets:mdWrap>
</mets:dmdSec>
```

5.3 Qualité

5.3.1 Qualité de la reconnaissance

Mesure de la qualité

La mesure de la qualité de la tâche de reconnaissance des signatures est réalisée selon les mêmes modalités que celles de la reconnaissance des titres (cf. section 4.3.1, taux de rappel, taux de précision).

Evaluation de la qualité

Le niveau de qualité acceptable concernant la reconnaissance des signatures d'un fascicule est caractérisé par des valeurs seuils pour les deux métriques :

	Taux de rappel	Taux de précision	Remarque

Signature	$seuil_{sr}$	$seuil_{sp}$	

Lors du contrôle, les valeurs constatées dans le fascicule à contrôler sont comparées aux valeurs attendues, qui sont spécifiques à chaque marché (cf. CCTP).

ATTENTION

POUR LA TACHE DE RECONNAISSANCE DES SIGNATURES, LA BNF ATTEND UNE MONTEE EN QUALITE MANUELLE DANS LE CAS OU UN PROCESSUS AUTOMATIQUE NE PERMETTRAIT PAS D'ATTEINDRE LES TAUX DE RAPPEL ET DE PRECISION ATTENDUS.

5.3.2 Qualité de la transcription

La mesure de la qualité de la transcription des signatures est réalisée selon les mêmes modalités que celles de la reconnaissance des titres (cf. section 4.3.3).

6. RECONNAISSANCE DES ARTICLES

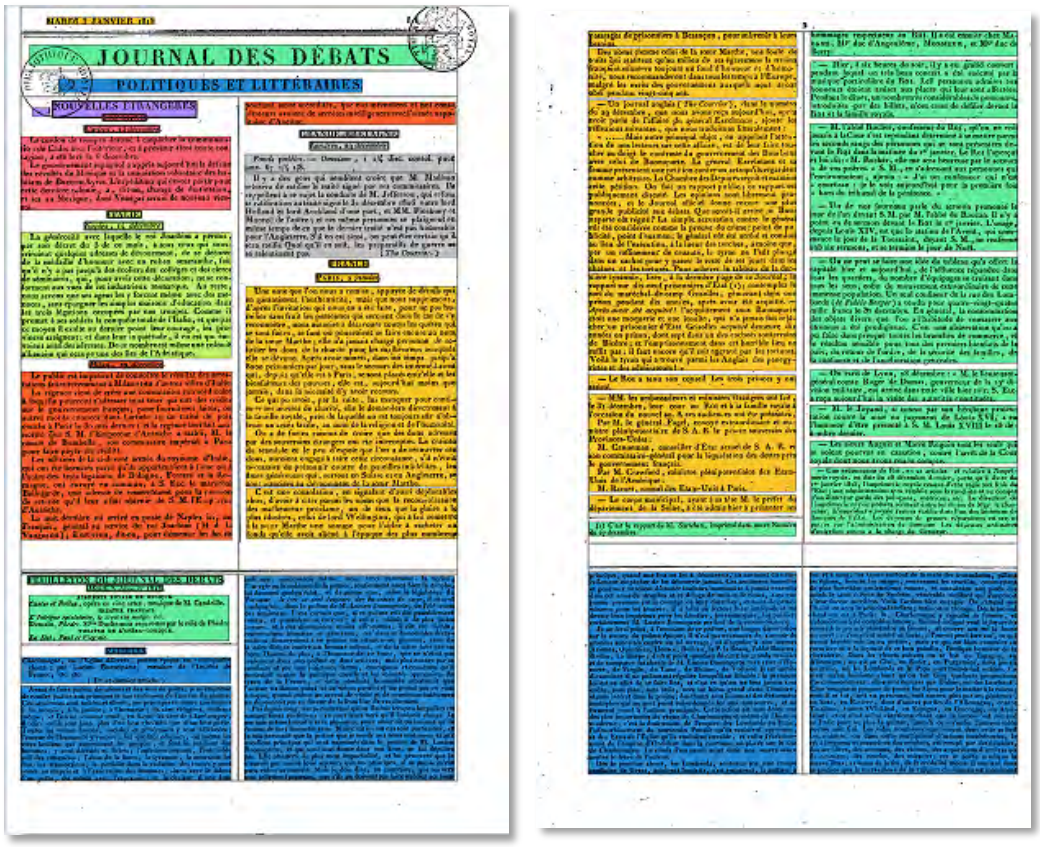
6.1 Principe

La tâche de reconnaissance des articles (ou OLR, *optical layout recognition*) consiste à identifier les éléments de contenus que sont les articles d'un fascicule de presse.

Cette tâche consiste également à identifier les titres de ces articles (cf. section 4).

EXEMPLE

Exemple : Le Journal des débats politiques et littéraires

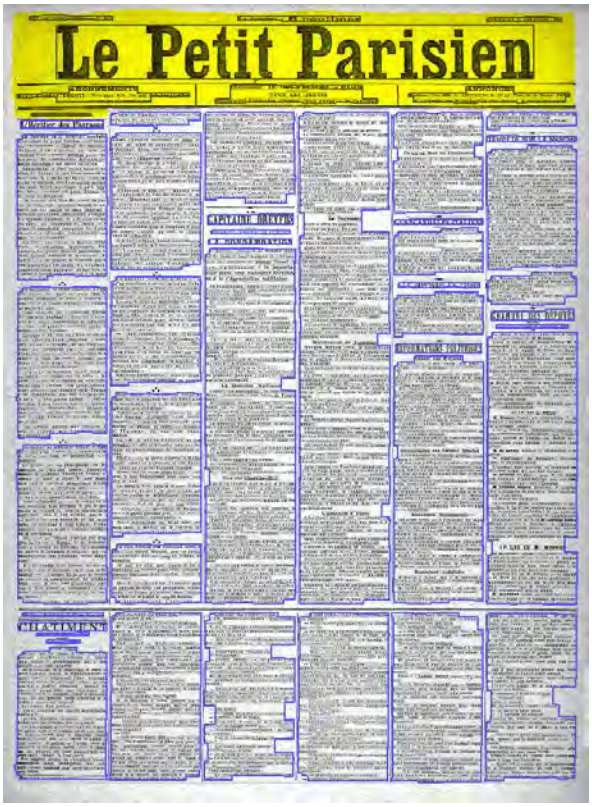


Au sein d'un fascicule de presse traité en reconnaissance des articles, la structure logique du fascicule est identifiée :

- zone des métadonnées (titre du journal, ours, achevé d'imprimer, n° édition, etc.),
- zone de contenu.

EXEMPLE

Exemple : Le Petit Parisien



Identification de la zone des métadonnées

http://gallica.bnf.fr/ark:/12148/bpt6k517311t/f1.image

La zone de contenu est décrite par :

- une liste ordonnée des articles selon l'ordre de lecture du fascicule,
- pour chaque article, liste ordonnée des contenus de l'article selon l'ordre de lecture :
  - titre (éventuel)
  - sous-titre (éventuel)
  - corps de l'article : paragraphes de texte, illustrations (éventuelles), titres intérieurs.

EXEMPLE

Zonage d'un article (orange) et identification des titres (vert)

http://gallica.bnf.fr/ark:/12148/bpt6k517311t/f1.image

NOTES

Le zonage est réalisé à l'échelle du fascicule. L'emprise d'un article peut donc courir sur plusieurs pages.

Pour chaque titre de presse à traiter, la BnF fournira une charte explicitant la mise en forme graphique des articles à reconnaître. Cette charte sera basée sur un fascicule annoté représentatif de l'ensemble du corpus à numériser (ou plusieurs si le principe de maquette du titre a évolué au fil du temps).

6.2 Description

6.2.1 Dans le document numérique

La description des articles nécessite un manifeste de document numérique METS. Dans les documents numériques au format ALTO, ne seront donc identifiés que les titres des articles, conformément aux règles décrites section 4.2.1.

6.2.2 Dans le manifeste numérique

Métadonnées

Chaque article est décrit dans une section <dmdSec>, où sont précisées ses métadonnées descriptives : titre, sous-titre, langue.

Exemple 1 : La Croix

http://gallica.bnf.fr/ark:/12148/bpt6k220169n

```
<mets:dmdSec ID="MODSMD_ARTICLE2">
<mets:mdWrap MIMETYPE="text/xml" MDTYPE="MODS" LABEL=" Métadonnées bibliographiques de l'article ">
 <mets:xmlData>
        <mods:mods>
                <mods:titleInfo ID="MODSMD_ARTICLE1_TI1" xml:lang="fr">
                        <mods:title> LA LOI INTERPRETEE PAR M. COMBES</mods:title>
                        <mods:subTitle>Xxxx</mods:title>
                </mods:titleInfo>
                <mods:language>
                        <mods:languageTerm type="code"
                                authority="rfc3066">fr</mods:languageTerm>
                </mods:language>
        </mods:mods>
 </mets:xmlData>
</mets:mdWrap>
</mets:dmdSec>
```

Carte de structure

L'ensemble des contenus du fascicule sont reproduits dans le manifeste METS du fascicule, à l'aide d'une carte de structuration logique :

```
<mets:structMap LABEL="Structure logique" TYPE="LOGICAL">
```

Cette carte est hiérarchique :

```
<mets:structMap LABEL="Logical Structure" TYPE="LOGICAL">
  <mets:div ID="DIVL1" TYPE="Newspaper" LABEL="Le Petit Parisien no. 6631 23.12.1894">
    <mets:div ID="DIVL2" TYPE="VOLUME" DMDID="MODSMD_PRINT MODSMD_ELEC" LABEL="Le Petit Parisien no. 6631 23.12.1894">
      <mets:div ID="DIVL3" TYPE="ISSUE" DMDID="MODSMD_ISSUE1" LABEL="Le Petit Parisien no. 6631 23.12.1894">
        <mets:div ID="DIVL4" TYPE="TITLE_SECTION">
        <mets:div ID="DIVL15" TYPE="CONTENT">
      </mets:div>
    </mets:div>
  </mets:div>
</mets:structMap>
```

- Niveau 1 : titre de journal

- Niveau 2 : volume
- Niveau 3 : fascicule (*issue*)
    - Niveau 3a. : zone des métadonnées,
    - Niveau 3b. : zone des contenus.

Chaque niveau est identifié par une balise   :

```
<div ID="…" TYPE="…" DMDID="…" LABEL="…">
    …
</div>
```

avec les attributs suivants :

- titre (éventuel) : LABEL,
- référence vers une section de métadonnées descriptives (<dmdSec>) : DMDID,
- nature de la division : TYPE,
    - journal : TYPE="NEWSPAPER"
    - volume : TYPE="VOLUME"
    - fascicule : TYPE="ISSUE"
    - zone des métadonnées : TYPE="TITLE_SECTION"
    - zone des contenus : TYPE="CONTENT"

Description des éléments de contenu

A l'intérieur de chacune des deux zones de contenus, on trouve la liste ordonnée des éléments de contenus, identifiés par les références du fichier ALTO et du ou des blocs ALTO contenant l'élément, à l'aide de balises <fptr> </fptr> et <area></area> :

```
<fptr>
        <area BETYPE="IDREF" FILEID="ALTO00001" BEGIN="P1_TB00021"/>
 </fptr>
```

Si plusieurs blocs ALTO constituent un élément, on utilisera une balise <seq> </seq> pour les associer :

```
<seq>
        <area BETYPE="IDREF" FILEID="ALTO00001" BEGIN="P1_TB00019"/>
        <area BETYPE="IDREF" FILEID="ALTO00001" BEGIN="P1_TB00020"/>
 </seq>
```

L'ordre de lecture de ces éléments est défini à deux niveaux :

- ordre des articles dans la carte de structure : il est défini par l'ordre séquentiel des éléments XML,
- ordre des contenus à l'intérieur d'un article : il est défini par l'attribut ORDER porté par les éléments de contenu.

Zone des métadonnées

La section contient tous les blocs de l'ours, de l'achevé d'imprimé, de l'édition, etc., ordonnés selon l'ordre de lecture (attribut ORDER).

Pour un fascicule de presse, ces contenus sont généralement en haut de la première page, éventuellement en bas.

EXEMPLE


```
<div ID="DIVL4" TYPE="TITLE_SECTION">
        <div ID="DIVL5" TYPE="TEXTBLOCK" ORDER="1">
            <fptr>
                <area BETYPE="IDREF" FILEID="ALTO00001" BEGIN="P1_TB00001"/>
            </fptr>
        </div>
        <div ID="DIVL6" TYPE="TEXTBLOCK" ORDER="2">
            <fptr>
                <area BETYPE="IDREF" FILEID="ALTO00001" BEGIN="P1_TB00002"/>
            </fptr>
      </div>
```

…

Eléments de la zone des métadonnées

http://gallica.bnf.fr/ark:/12148/bpt6k517311t/f1.image

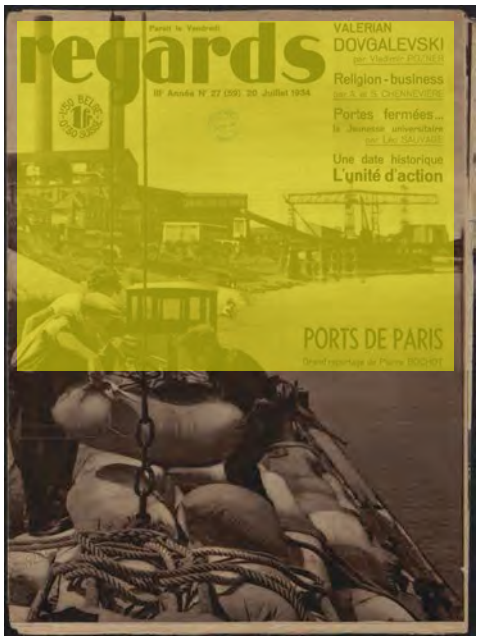
Parmi ces éléments, seul le titre du journal est identifié par un typage HEADLINE.

EXEMPLE


```
<div ID="DIVL7" TYPE="HEADLINE" ORDER="3">
        <fptr>
            <area BETYPE="IDREF" FILEID="ALTO00001" BEGIN="P1_TB00003"/>
        </fptr>
</div>
…
```

Ces contenus de métadonnées peuvent être répartis en plusieurs points du fascicule. Dans ce cas, la section des métadonnées contiendra plusieurs sous-sections :

EXEMPLE




```
<div ID="DIVL4" TYPE="TITLE_SECTION">
  <div ID="DIVL5" >
      <div ID="DIVL6" TYPE="TEXTBLOCK" ORDER="1">
          …
      </div>
      <div ID="DIVL7" TYPE="TEXTBLOCK" ORDER="2">
          …
      </div>
      …
  </div>
  <div ID="DIVL10" >
      <div ID="DIVL12" TYPE="TEXTBLOCK" ORDER="1">
          …
      </div>
      …
  </div>
</div>
```

http://gallica.bnf.fr/ark:/12148/bpt6k7635838f

Les éventuels titres courants des pages internes d'un fascicule (segmentés dans les éléments ALTO TopMargin et/ou BottomMargin) seront également décrits dans la section des métadonnées.

EXEMPLE

…

```
<div ID="DIVL7"  ORDER="3">
        <fptr>
          <area BETYPE="IDREF" FILEID="ALTO00003" BEGIN="P3_TB00001"/>
        </fptr>
        <fptr>
          <area BETYPE="IDREF" FILEID="ALTO00003" BEGIN="P3_TB00002"/>
        </fptr>

</div>
…
```

Titre courant de la page 3 : “Le Petit Parisien” “3”

Zone des contenus

Cette zone contient tous les articles du fascicule, ordonnés selon l’ordre de lecture logique du fascicule.

EXEMPLE

```
<mets:div ID="DIVL15" TYPE="CONTENT">
  <mets:div ID="DIVL16" TYPE="ARTICLE" DMDID="MODSMD_ARTICLE1" LABEL="L'Héritier des Pharaons">
  <mets:div ID="DIVL39" TYPE="ARTICLE" DMDID="MODSMD_ARTICLE2" LABEL="CAPITAINE DREYFUS">
  <mets:div ID="DIVL54" TYPE="ARTICLE" DMDID="MODSMD_ARTICLE3" LABEL="LES SCANDALES ITALIENS">
  <mets:div ID="DIVL61" TYPE="ARTICLE" DMDID="MODSMD_ARTICLE4" LABEL="LES JAPONAIS EN CHINE">
  <mets:div ID="DIVL68" TYPE="ARTICLE" DMDID="MODSMD_ARTICLE5" LABEL="INFORMATIONS POLITIQUES">
  <mets:div ID="DIVL77" TYPE="ARTICLE" DMDID="MODSMD_ARTICLE6" LABEL="TEMPÊTE SUR LA MANCHE">
  <mets:div ID="DIVL86" TYPE="ARTICLE" DMDID="MODSMD_ARTICLE7" LABEL="CHAMBRE DES DÉPUTÉS">
  <mets:div ID="DIVL108" TYPE="ARTICLE" DMDID="MODSMD_ARTICLE8" LABEL="N° 50 Feuilleton du PETIT PARISIEN">
  <mets:div ID="DIVL141" TYPE="ARTICLE" DMDID="MODSMD_ARTICLE10" LABEL="Dépêches de l'Étranger">
  <mets:div ID="DIVL148" TYPE="ARTICLE" DMDID="MODSMD_ARTICLE11" LABEL="A MADAGASCAR">
  <mets:div ID="DIVL161" TYPE="ARTICLE" DMDID="MODSMD_ARTICLE12" LABEL="NOUVELLES MILITAIRES">
  <mets:div ID="DIVL168" TYPE="ARTICLE" DMDID="MODSMD_ARTICLE13" LABEL="A L'ÉCOLE DE DROIT">
  <mets:div ID="DIVL175" TYPE="ARTICLE" DMDID="MODSMD_ARTICLE14" LABEL="LES LIVRES D'ÉTRENIVES">
  <mets:div ID="DIVL184" TYPE="ARTICLE" DMDID="MODSMD_ARTICLE15" LABEL="CONSEIL MUNICIPAL DE PARIS">
  <mets:div ID="DIVL193" TYPE="ARTICLE" DMDID="MODSMD_ARTICLE16" LABEL="LE VACCIN DU CROUP">
  <mets:div ID="DIVL202" TYPE="ARTICLE" DMDID="MODSMD_ARTICLE17" LABEL="AFFAIRES DE CHANTAGE">
  <mets:div ID="DIVL209" TYPE="ARTICLE" DMDID="MODSMD_ARTICLE18" LABEL="Calendrier du « Petit Parisien »">
  <mets:div ID="DIVL216" TYPE="ARTICLE" DMDID="MODSMD_ARTICLE19" LABEL="LES PROJETS POUR 1900">
  <mets:div ID="DIVL225" TYPE="ARTICLE" DMDID="MODSMD_ARTICLE20" LABEL="LES TRIBUNAUX">
  <mets:div ID="DIVL234" TYPE="ARTICLE" DMDID="MODSMD_ARTICLE21" LABEL="ÉCHOS ET NOUVELLES">

…
```

Chaque article est décrit par les attributs suivants :

- titre (éventuel) : LABEL,
- référence vers une section de métadonnées descriptives : DMDID,
- type : TYPE="ARTICLE"

et présente la structure suivante :

```
<div ID="…" TYPE="ARTICLE" DMDID="…" LABEL="…">
      <div ID="…" TYPE="HEADING" DMDID="…" LABEL="…">
         <div ID="…" TYPE="TITLE" DMDID="…" LABEL="…">
      </div>
      <div ID="…" TYPE="BODY" DMDID="…" LABEL="…">
         …
      </div>
</div>
```

EXEMPLE

```
<div ID="DIVL16" TYPE="ARTICLE" DMDID="MODSMD_ARTICLE1"
     LABEL="L'Héritier des Pharaons">
   <div ID="DIVL17" TYPE="HEADING">
     <div ID="DIVL18" TYPE="TITLE">
       …
     </div>
   </div>

   <div ID="DIVL19" TYPE="BODY">
     <div ID="DIVL20" TYPE="BODY_CONTENT">
       <div ID="DIVL21" TYPE="PARAGRAPH" ORDER="1">
         …
       </div>

       …
     </div>
   </div>
</div>
```

http://gallica.bnf.fr/ark:/12148/bpt6k517311t/f1.image

NOTES

Un article est un élément de contenu pris dans un sens générique. Il peut ne pas posséder de titre. Exemples : cotation de bourse, publicités.

Regroupement de contenus

Les contenus récurrents ou répétitifs seront regroupés à l'aide de sections (TYPE="SECTION") :

- petites annonces,
- cotations boursières,
- résultats sportifs,
- programme de spectacles,
- publicités.

Pour chaque titre de journal, la BnF précisera les typologies de contenus à placer dans des sections (cf. section 7).

EXEMPLE

```
…

<div ID="DIVL16" TYPE="SECTION" DMDID="MODSMD_SECTION8"
     LABEL="Publicités">

   <div ID="DIVL17" TYPE="ARTICLE" DMDID="MODSMD_ARTICLE10"
     LABEL="Pub. 1 page 2">
     <div ID="DIVL19" TYPE="BODY">
                …
```

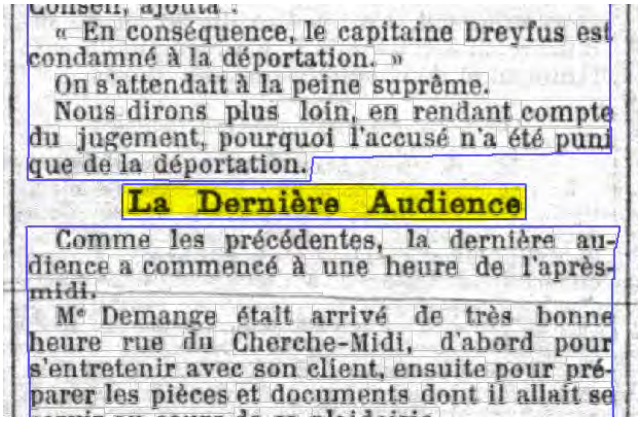
```
      </div>
    </div>
    <div ID="DIVL24" TYPE="ARTICLE" DMDID="MODSMD_ARTICLE11"
     LABEL="Pub. 1 page 3">
      <div ID="DIVL25" TYPE="BODY">
                ...
      </div>
    </div>

</div>
...
```

Publicités d'un fascicule regroupées dans une section

Titres intérieurs

Les titres intérieurs d'article sont identifiés (TYPE="INSIDEHEADING"):

EXEMPLE



« En conséquence, le capitaine Dreyfus est condamné à la déportation. »

On s'attendait à la peine suprême.

Nous dirons plus loin, en rendant compte du jugement, pourquoi l'accusé n'a été puni que de la déportation.

La Dernière Audience

Comme les précédentes, la dernière audience a commencé à une heure de l'après-midi.

Me Demange était arrivé de très bonne heure rue du Cherche-Midi, d'abord pour s'entretenir avec son client, ensuite pour préparer les pièces et documents dont il allait se

```
...
<mets:div ID="DIVL58" TYPE="INSIDEHEADING" ORDER="4"
     LABEL="La dernière Audience">
  <mets:div ID="DIVL59" TYPE="TITLE">
    <mets:fptr>
      <mets:area BETYPE="IDREF" FILEID="ALTO00001" BEGIN="P1_TB00028"/>
    </mets:fptr>
  </mets:div>
</mets:div>
```

Titre intérieur d'un article

Illustrations

Les illustrations (cf. « Référentiel OCR », section 5.7) sont identifiées (TYPE="ILLUSTRATION"). Le type de l'illustration référencée pourra être de différente nature :

- illustration générique, écriture manuscrite (cf. « Référentiel OCR », section 5.7.7) : TYPE="IMAGE",
- carte : TYPE="MAP",
- partition : TYPE="MUSICSCORE",

EXEMPLE

```
<div ID="DIVL1200" TYPE="ILLUSTRATION" ORDER="2" DMDID="MODSMD_PICT4">
  <div ID="DIVL1201" TYPE="IMAGE">
    <fptr>
      <area BETYPE="IDREF" FILEID="ALTO00006" BEGIN="P6_ILL00001"/>
    </fptr>
  </div>
</div>
```

Description d'une illustration

EXEMPLE

```
<div ID="DIVL120" TYPE="ILLUSTRATION" ORDER="3" DMDID="MODSMD_PICT5">
  <div ID="DIVL121" TYPE="MAP">
    <fptr>
      <area BETYPE="IDREF" FILEID="ALTO00005" BEGIN="P5_ILL00001"/>
    </fptr>
  </div>
</div>
```

Description d'une carte

Dans le cas d'une illustration accompagnée d'une légende, c'est le bloc composé ALTO regroupant illustration et légende qui sera référencé dans la carte de structure :

EXEMPLE

```
<div ID="DIVL1300" TYPE="ILLUSTRATION" ORDER="2" DMDID="MODSMD_PICT1">
  <div ID="DIVL1301" TYPE="IMAGE">
    <fptr>
      <area BETYPE="IDREF" FILEID="ALTO00001" BEGIN="P1_CB00001"/>
    </fptr>
  </div>
</div>
```

Description d'une illustration avec légende

Eléments graphiques

Les éléments graphiques (cf. « Référentiel OCR », section 5.8) suivants ne seront pas identifiés dans la carte de structure :

- tampon,
- traits de séparation,
- lettre ornée (le bloc composé incluant lettre ornée et paragraphe est référencé),

- écriture manuscrite, sauf s'il s'agit d'une illustration (cf. section précédente).

Seuls les éléments suivants seront identifiés (TYPE="ILLUSTRATION") :

- décoration : TYPE="ORNAMENT".

Récapitulatif du vocabulaire METS pour la description des contenus

Carte logique			
Eléments	TYPE	Section	Remarque
journal	NEWSPAPER		
volume	VOLUME		
fascicule	ISSUE		
zone des métadonnées	TITLE_SECTION		
zone des contenus	CONTENT		
Zone des métadonnées			
Eléments	TYPE	Section	Remarque
bloc de texte générique	TEXTBLOCK		
titre du journal	HEADLINE		
illustration	ILLUSTRATION		
Zone des contenus			
Eléments	TYPE	Section	Remarque
article	ARTICLE		
groupement d'articles	SECTION		
entête d'article	HEADING		
titre d'article	TITLE		
sous-titre d'article	SUBTITLE		
contenu d'article	BODY		
paragraphe	PARAGRAPH		
illustration	ILLUSTRATION		
image	IMAGE		
carte	MAP		
partition	MUSICSCORE		
décoration	ORNAMENT		
tableau	TABLE		
titre intérieur	INSIDEHEADING		
titre	TITLE		

NOTES

Des exemples de cartes de structure METS complètes peuvent être fournies.

Opération de traitement

L'application des traitements OLR (olrSegmentation) et de reconnaissance des titres (headlines Detection) est décrite dans le manifeste METS du document numérique.

On fera le distinguo entre un processus simple d'extraction d'information (un seul atelier à l'aide d'un seul outil) et un processus complexe (plusieurs événements à décrire). Cf. « Référentiel d'enrichissement des métadonnées, version METS », section 8.6.2.

6.3 Qualité

6.3.1 Mesure de la qualité

La qualité de la reconnaissance est évaluée lors de la phase de test (à l'aide d'un corpus annoté représentatif des documents à traiter) et en production.

La mesure de la qualité de la tâche de reconnaissance des articles recouvre :

- la reconnaissance des titres (cf. section 4.3),
- la reconnaissance des articles.

La mesure de la qualité de la tâche de reconnaissance des articles est réalisée à l'aide de la métrique suivante :

Qualité = nombre pondéré des articles reconnus / nombre d'articles présent

Cette mesure s'applique à un document complet.

Pour obtenir la pondération d'un article reconnu, on utilisera la méthode suivante :

- la valeur 1 (article parfaitement reconnu) est pondérée selon les éventuels défauts constatés :

	Pondération	**Remarque**
Un ou plusieurs éléments de contenu (blocs ALTO) de l'article sont erronés	-0,5	éléments en trop ou éléments absents
La structure logique de l'article dans la table de navigation est erronée	-0,10	ordre interne des contenus de l'article erroné, défaut de structuration
Ordre de lecture de l'article dans la table de navigation	-0,20	ordre de l'article dans la table de navigation erroné

Exemples :

	nbre articles présents	articles corr. reconnus	articles non reconnus	articles partiel. reconnus	structure logique	ordre erroné	**Qualité**
1 article non reconnu	**100**	99	1	0	0	0	99,0%
1 article reconnu en 2 articles (sur-découpage)	**100**	99	0	1	0	0	99,5%

2 articles reconnus en 1 (sous-découpage)	**100**	98	1	1	0	0	98,5%

6.3.2 Evaluation de la qualité

Le niveau de qualité acceptable concernant la reconnaissance des articles d'un fascicule est caractérisé par une valeur seuil :

	Qualité	Remarque
Articles	$seuil_{art}$	

Lors du contrôle, les valeurs constatées dans le fascicule à contrôler sont comparées aux valeurs attendues, qui sont spécifiques à chaque marché (cf. CCTP).

POUR LA TACHE DE RECONNAISSANCE DES ARTICLES, LA BNF ATTEND UNE MONTEE EN QUALITE MANUELLE DANS LE CAS OU UN PROCESSUS AUTOMATIQUE NE PERMETTRAIT PAS D'ATTEINDRE LES TAUX DE RAPPEL ET DE PRECISION ATTENDUS.

7. RUBRIQUAGE

7.1 Principe

Le rubriquage consiste à classifier les contenus d'un fascicule de presse en fonction de la nature de ces contenus. Les catégories généralement retenues sont les suivantes :

- information,
- éditorial,
- rubrique judiciaire,
- loisirs (feuilleton littéraire, poésie, humour, mots-croisés, etc.),
- publicités commerciales,
- annonces (emplois, ventes, décès, etc.)

- Par convention, on considèrera que la catégorie Information est définie par exclusion des autres catégories. Elle n'a pas à être identifiée en tant que telle.
- Pour certains titres de presse, certaines catégories pourront ne pas être représentées (par exemple « Editorial » ou « Loisirs »). Symétriquement, on pourra définir des catégories particulières pour certains titres (par exemple « Bourse »), mais le nombre de catégories à identifier pour un titre de presse donné ne pourra dépasser un maximum défini dans le CCTP.

Pour chaque titre de presse à traiter, la BnF fournira une charte listant les catégories à reconnaître ainsi que leur éventuelle mise en forme graphique, notamment pour les catégories sujettes à ambiguïté (éditorial/information).

Cette identification sera basée sur un fascicule annoté représentatif de l'ensemble du corpus à numériser (ou plusieurs si le rubriquage ou le principe de maquette du titre a évolué au fil du temps).

Rouge : éditorial – Orange : information – Vert : loisirs – Violet : publicité

Exemple 1 : La Presse Illustrée


LA PRESSE ILLUSTRÉE

JOURNAL QUOTIDIEN

COURRIER DE PARIS

LES COSAQUES A PARIS

PONSON DU TERRAIL

Exemple 2 : Excelsior


EXCELSIOR

Dorian

Échos

Des munitions pour le temps de paix.

Billets de logement.

L'heure de la victoire.

Diplômes pour distributions de prix.

Peut-être bien!...

En attendant...

Compartiments de Non-Parleurs

J. Ernest-Charles.

L'HUMOUR ET LA GUERRE

Exemple 3 : La Croix


LA CROIX

GUÉRISON DE LA TUBERCULOSE

SON NOUVEAU TRAITEMENT par la MÉTHODE VANADIÉE HÉLOUIS

LE **Vanadiol Hélouis** GUÉRIT LA **TUBERCULOSE** à tous les degrés.

Les effets du **VANADIOL** *sur les malades se résument ainsi :*

1° Après une semaine de traitement, l'état général est remarquablement amélioré, la toux cesse, les expectorations diminuent ;

2° Le **VANADIOL** excite l'appétit d'une manière extraordinaire en supprimant d'abord la répulsion, parfois invincible, qu'ont les tuberculeux pour toute nourriture. Ce résultat est obtenu par la merveilleuse propriété que possède seul le Vanadiol, de détruire par oxydation les microbes et les toxines (poisons microbiens) de l'appareil digestif.

En vingt jours, on obtient une augmentation de poids de deux kilogrammes ;

3° Le **VANADIOL** augmente rapidement le nombre des globules sanguins. Il favorise l'hématose (phénomène respiratoire) par l'oxydation rapide de l'hémoglobine. Il accélère la nutrition d'une façon remarquable.

Applications dans les Hôpitaux à Paris et à l'Etranger. Résultats extraordinaires. — Retour certain de l'appétit et augmentation considérable du poids du corps chez tous les malades. Opinion des Médecins.

LE **Vanadiol Hélouis** *grâce à ses propriétés oxygénantes, constitue également le remède souverain :*

1° Des maladies dites par ralentissement de la nutrition (Rhumatismes, Diabète, Albuminurie, Goutte, Gravelle, etc.)

2° De toutes les maladies de l'estomac (Dyspepsie, Gastralgie, etc.) ;

3° Des maladies des organes respiratoires (Tuberculose, Bronchites chroniques, etc.) par l'oxydation des toxines microbiennes et par l'énorme appétit qu'il donne aux malades, résultats qui font disparaître rapidement tous les phénomènes alarmants en rendant la vigueur et la force, grâce à l'oxygène actif que la médication vanadiée répand dans tout l'organisme.

L'action oxygénante et tonique du **VANADIOL HÉLOUIS** *est considérablement supérieure à celle du Fer.*

*Les effets en sont immédiats dans tous les cas d'***Anémie**, *de* **Chlorose** *et dans toutes les Convalescences.*

De tout ce qui précède on peut conclure que :

Le VANADIOL HÉLOUIS constitue le traitement le moins cher et le plus efficace de la **TUBERCULOSE** à tous les degrés.

Son pouvoir de guérir réside dans sa propriété de détruire les microbes, non seulement de la Tuberculose, mais encore de toutes les maladies des voies respiratoires et des maladies infectieuses.

Ses propriétés oxydantes particulières en font la médication antiseptique interne, non caustique ni toxique, tant cherchée par les médecins ; elles permettent son application avec succès certain dans toutes les maladies qui amènent une diminution des forces de l'organisme.

On peut donc affirmer que :

LE VANADIOL HÉLOUIS

est, dans l'état actuel de la science, le remède souverain de la Tuberculose pulmonaire à tous les degrés, il peut donc être considéré, dans la lutte contre ce terrible fléau, comme le véritable **REMÈDE D'UTILITÉ PUBLIQUE.**

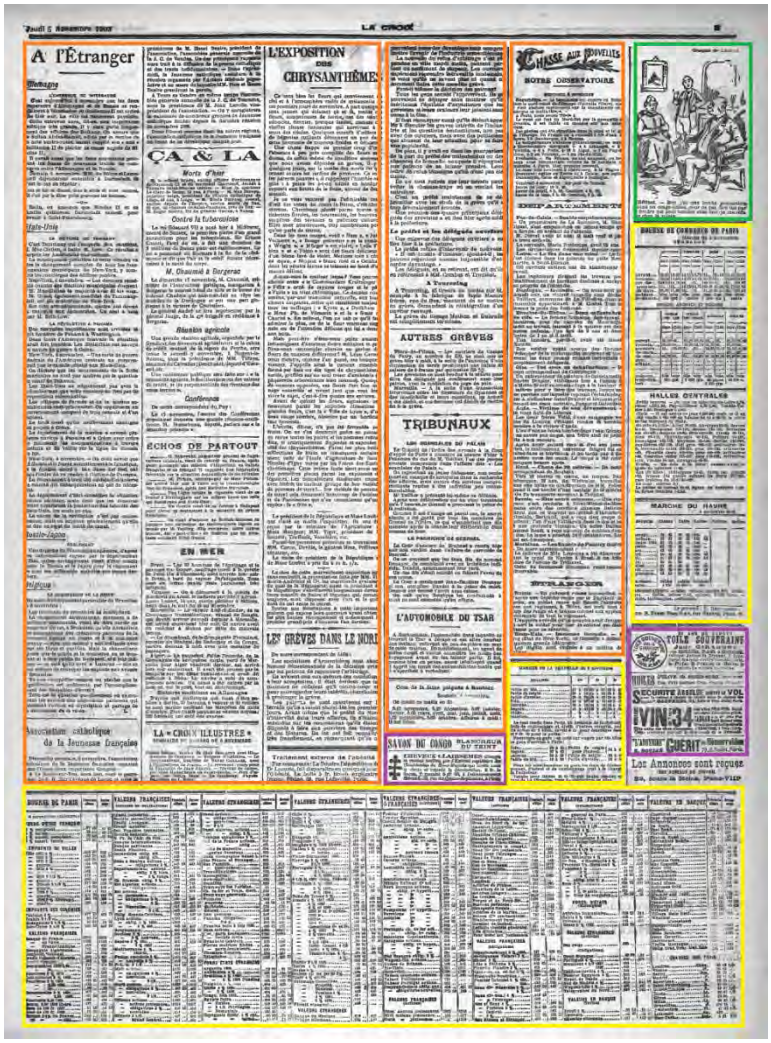
TRAITEMENT D'UN MOIS : Le Flacon **10** francs — Toutes bonnes Pharmacies.

Envoi franco contre 10f 60 adressés à la SOCIÉTÉ FRANÇAISE des COMPOSÉS du VANADIUM, 17, Rue des Bateliers, à Clichy (Seine)

BROCHURE ET RENSEIGNEMENTS GRATUITS SUR DEMANDE

EXEMPLE

Jaune : bourse et cotations – Exemple : La Croix



7.2 Description

7.2.1 Dans le document numérique

Le mécanisme de marquage décrit dans le « Référentiel OCR », sections 4.2.3 et 5.10.1 sera utilisé. Ce mécanisme nécessite la version 2 du format ALTO BnF.

On utilisera la syntaxe suivante pour identifier une étiquette de type de contenu :

```
<LayoutTag ID="…" TYPE="Content" LABEL="…"/>
```

Les catégories de contenu à identifier seront représentées par les étiquettes génériques suivantes :

- éditorial :

  ```
  <LayoutTag ID="TAG_opinion" TYPE="Content" LABEL="opinion"/>
  ```

- information :

  ```
  <LayoutTag ID="TAG_information" TYPE="Content" LABEL="information"/>
  ```

- loisirs (feuilleton littéraire, poésie, humour, mots-croisés, etc.) :

  ```
  <LayoutTag ID="TAG_entertainment" TYPE="Content"
  LABEL="entertainment"/>
  ```

- publicités commerciales :

<LayoutTag ID="TAG_advertisement" TYPE="Content"
LABEL="advertisement"/>

- annonces (emplois, ventes, décès, etc.) :

<LayoutTag ID="TAG_freead" TYPE="Content"
LABEL="freead"/>

Une étiquette de contenu est associée à un bloc de texte (élément <TextBlock>) ou à une illustration (élément <Illustration>) à l'aide de l'attribut TAGREFS et d'un identifiant d'étiquette :

<TextBlock ID="PAG_00000001_TB000010"… TAGREFS="TAG_opinion" />

ON N'ETIQUETTERA PAS LES LIGNES DE TEXTE, LES MOTS ET LES ELEMENTS GRAPHIQUES.

7.2.2 Dans le manifeste numérique

Opération de traitement

L'application du traitement (événement PREMIS de type contentClassification) est décrite dans le manifeste METS du document numérique.

On fera le distinguo entre un processus simple d'extraction d'information (un seul atelier à l'aide d'un seul outil) et un processus complexe (plusieurs événements à décrire). Cf. « Référentiel d'enrichissement des métadonnées, version METS », section 8.6.2.

Relation avec la reconnaissance des articles

Lorsque qu'une tâche de reconnaissance des articles est également demandée (cf. section 6), le rubriquage est décrit en tant que métadonnées descriptives des articles (élément <mods:classification>), dans une section <dmdSec> .

Exemple 1 : La Croix

Métadonnées descriptives

```
<mets:dmdSec ID="MODSMD_ARTICLE2">
<mets:mdWrap MIMETYPE="text/xml" MDTYPE="MODS" LABEL=" Métadonnées bibliographiques de
l'article ">
 <mets:xmlData>
        <mods:mods>
                <mods:titleInfo ID="MODSMD_ARTICLE1_TI1" xml:lang="fr">
                        <mods:title> LA LOI INTERPRETEE PAR M. COMBES</mods:title>
                </mods:titleInfo>
                <mods:language>
                        <mods:languageTerm type="code"
                                authority="rfc3066">fr</mods:languageTerm>
                </mods:language>
                <mods:classification>information</mods:classification>
        </mods:mods>
  </mets:xmlData>
 </mets:mdWrap>
```

```
</mets:dmdSec>

Carte de structure logique

<mets:div ID="DIVL11" TYPE="ARTICLE" DMDID="MODSMD_ARTICLE2"
      LABEL="LA LOI INTERPRETEE PAR M. COMBES">
```

7.3 Qualité

7.3.1 Mesure de la qualité

La qualité de la qualité de la tâche de rubriquage est évaluée lors de la phase de test (à l'aide d'un corpus annoté représentatif des documents à traiter) et en production.

La mesure de la qualité est réalisée à l'aide de deux métriques, qui sont calculées pour chacune des catégories de rubrique à identifier dans un fascicule donné :

- **Rappel** : le rappel exprime le rapport entre les contenus correctement reconnus dans le fascicule pour une catégorie de contenus *i*, et les contenus de catégorie *i* présents dans le fascicule :

  Rappel*i* = nombre de contenus correctement reconnus pour la catégorie *i* / nombre de contenus de la catégorie *i*

- **Précision** : la précision exprime le rapport entre les contenus correctement reconnus dans le fascicule pour une catégorie de contenus *i*, et les contenus reconnus pour une catégorie *i* :

  Précision*i* = nombre de contenus correctement reconnus pour la catégorie *i* / nombre de contenus reconnus de la catégorie *i*

Par « contenus », on entend blocs de texte et blocs d'illustration.

CES MESURES S'APPLIQUENT A UN DOCUMENT COMPLET (FASCICULE).

7.3.2 Evaluation de la qualité

Le niveau de qualité acceptable concernant la reconnaissance des contenus d'un fascicule est caractérisé par des valeurs seuils pour les deux métriques et pour les catégories de rubrique :

	Taux de rappel	Taux de précision	Remarque
Annonces	$seuil_{ann\_r}$	$seuil_{ann\_p}$	
Publicités	$seuil_{pub\_r}$	$seuil_{pub\_p}$	
Feuilletons	…	…	

Lors du contrôle, les valeurs constatées dans le fascicule à contrôler sont comparées aux valeurs attendues, qui sont spécifiques à chaque marché (cf. CCTP).

POUR LA TACHE DE RECONNAISSANCE DES RUBRIQUES, LA BNF ATTEND UNE MONTEE EN QUALITE MANUELLE DANS LE CAS OU UN PROCESSUS AUTOMATIQUE NE PERMETTRAIT PAS D'ATTEINDRE LES TAUX DE RAPPEL ET DE PRECISION ATTENDUS.

8. CONTROLE DE LA QUALITE

Le contrôle de la qualité est assuré par plusieurs moyens :

- des contrôles automatiques appliqués sur les fichiers livrés,
- un contrôle par échantillonnage visuel.

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

Cette section décrit les critères de rejets ou d'acceptation

8.1 Contrôle automatique

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

SI CE CONTROLE EMET UNE ERREUR MAJEURE SUR UN FICHIER NUMERIQUE D'UN DOCUMENT, LE DOCUMENT EST ECARTE ET NE PASSE PAS EN CONTROLE PAR ECHANTILLONNAGE VISUEL. LE DOCUMENT EST DONC REJETE DES CETTE ETAPE.

Ces contrôles automatiques sont de plusieurs natures :

- validation des fichiers ALTO relativement au schéma XML ALTO BnF v2,
- validation des fichiers METS relativement au schéma METS
- respect des consignes d'étiquetage des enrichissements telles que décrites dans le présent document et dans le « Référentiel OCR »,
- contrôles spécifiques à certaines tâches d'enrichissement. Par exemple pour la reconnaissance des articles, certains blocs ALTO doivent être référencés dans la carte de structure.

Les modalités de ce contrôle sont détaillées dans une charte de contrôle élaborée conjointement par la BnF et le prestataire.

8.2 Contrôle par échantillonnage visuel

Le contrôle visuel des documents sera effectué par échantillonnage.

La taille minimale de l'échantillon est définie par la norme « ISO 2859-1 : Règle d'échantillonnage pour les contrôles par attribut ».

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 les tâches d'enrichissement du document selon les critères qualité décrits dans les sections propres à chaque traitement.

8.2.1 Fréquence des contrôle

La constitution des échantillons de contrôle sera effectuée de manière régulière tout au long du marché ou projet.

Des opérations de contrôle ciblées pourront aussi être mise en place.

8.2.2 Règles d'échantillonnage

- Le nombre documents à contrôler est au moins égal à l'effectif de l'échantillon, lequel est déterminé par l'effectif du lot (un lot peut être par exemple l'ensemble d'une livraison).
- La population de l'échantillon de contrôle sera tirée de façon aléatoire sur l'effectif du lot contrôlé.
- La méthode de contrôle est par attribut, l'individu est le document.
- L'échantillonnage est effectué par défaut en plan simple (le nombre de documents contrôlé est égal à l'échantillon) ; toutefois dans certains cas particulier des plans échantillonnage doubles ou multiples pourront être mis en place.

En règle générale, l'échantillonnage sur une production courante est effectué en mode de contrôle normal et en niveau de contrôle général II (voir norme NF ISO 2859-1 chapitre 10.1).

En fonction des résultats de contrôle, les règles de passage en mode de contrôle renforcé ou réduit selon le cas reposent sur les principes décrits dans la norme NF ISO 2859-1 chapitre 9.3.

8.2.3 Modalité du contrôle

Reconnaissance des titres, des signatures, des articles, rubriquage

Le tableau suivant récapitule les critères de conformité appliqués sur les documents d'un échantillon, tâche par tâche.

	Taux de rappel	**Taux de précision**	**Taux de transcription du texte**	**Remarque**
Reconnaissance des titres				
Titre d'article (titre, sous-titre)	≥ $seuil_{ar}$	≥ $seuil_{ap}$	≥ taux OCR qualité éditoriale	
Titre intérieur	≥ $seuil_{ir}$	≥ $seuil_{ip}$	≥ taux OCR qualité éditoriale	
Reconnaissance des signatures	≥ $seuil_{sr}$	≥ $seuil_{sp}$	≥ taux OCR qualité éditoriale	
Reconnaissance des articles				
Reconnaissance des titres	Cf. ci-dessus		≥ taux OCR qualité éditoriale	
Reconnaissance des contenus	≥ $seuil_{art}$			

Rubriquage				
Annonces	≥ $seuil_{ann\_r}$	≥ $seuil_{ann\_p}$		
Publicités	≥ $seuil_{pub\_r}$	≥ $seuil_{pub\_p}$		
…				

⋆ Les valeurs des seuils et des taux sont données dans le CCTP de chaque marché ou projet.

La détection d'une seule non-conformité sur un document entraîne la non-conformité du document et donc son rejet.

Le rejet de l'ensemble de la population d'un échantillon (l'effectif de l'échantillon) est déterminé par le pourcentage d'individus non conformes dans cet effectif.

La détection d'une non-conformité structurelle sur tous les documents composant le lot de contrôle entraîne un audit du processus de production et éventuellement le rejet de tous les documents déjà produits.

Entités nommées

Le contrôle des tâches concernant les entités nommées (reconnaissance et alignement) sera précisé dans le CCTP de chaque marché.

9. LIVRAISON

Les documents enrichis seront livrés à la BnF selon les attendus du « Référentiel de livraison de document numérique », qui décrit le format physique à utiliser.

L'ensemble des fichiers sera livré sous la forme d'une archive compressée au format ZIP.