Alignement-donnees-bnf : spécifications techniques

Description technique des mécanismes en œuvre dans les différents modules du logiciel **alignements-donnees-bnf**

# Module d’alignements – notices bibliographiques

Ce module prend en entrée un fichier tabulé, encodé en UTF-8, contenant pour chaque notice à aligner les métadonnées qui seront exploitées par le programme pour trouver la notice BnF correspondante.

## Fichier en entrée : format

* Fichier à colonnes : une ligne par notice, une colonne par élément d’information.
* Séparateur : tabulation
* Encodage : UTF-8 sans BOM

Selon le type de document, les métadonnées exploitables ne sont pas les mêmes. Le programme distingue donc plusieurs types de documents, et il n’est pas possible de mettre dans un même fichier des livres et des DVD, ou des livres et des périodiques

### Monographies imprimées

Numéro de notice

FRBNF

ARK initial

ISBN

EAN

Titre

Auteurs

Date

Tome

Editeur

### CD, DVD (son, vidéo)

Numéro de notice

FRBNF

ARK initial

EAN

Numéro commercial

Titre

Auteur

Date

Editeur

### Périodiques

Numéro de notice

FRBNF

ARK initial

ISSN

Titre

Auteur

Date

Lieu

## Processus d’alignements

### Erreurs les plus fréquentes

**Rappel : le programme est tributaire de la qualité des données contenues dans le fichier fourni, et des données du catalogue de la BnF.** Des erreurs dans la manière d’alimenter les zones ou les colonnes génèreront des erreurs dans la réponse du programme.

* Les colonnes attendues ne correspondent pas :
  + Il manque une colonne dans le fichier en entrée, si bien que le programme utilise la colonne Titre comme s’il s’agissait d’un EAN, par exemple
  + Oubli de sélectionner le bon type de document pour le fichier mis en entrée  
    
* Les zones codées Type de document / Type de contenu ont été mal renseignées  
  🡪 un DVD de film se retrouve dans le fichier des imprimés, et l’alignement proposé pointe sur le roman qui a inspiré le film…
* L’encodage du tableur n’est pas de l’UTF8 sans BOM, ce qui fait planter le programme
* Les données sont passées par Excel, qui a transformé les ISBN ou EAN en données du type « 9,78E15711 ».  
  Idem sur les colonnes Dates, notamment

## Fichiers en sortie

# Module de conversion de fichier de notices en fichiers tabulés

## Principes

Chacun peut directement le module d’alignement, en alimentant chaque colonne selon les sources d’information qu’il estime pertinentes (en fonction de ce qu’il connaît de ses données et de ses objectifs).

Mais pour faciliter cet alignement, le logiciel propose un module qui, à partir d’un fichier exporté d’un catalogue au format Unimarc (iso2709 ou XML), propose des tableaux à utiliser dans le module d’alignements

Pour chaque notice, extraction

* Du type de notice (zone codée)
* Du type de document (zone codée)
* De différentes métadonnées définies ci-dessous :
  + Numéro FRBNF
  + Numéro ARK BnF
  + Numéro ISBN
  + Numéro EAN
  + Numéro ISSN
  + Numéro commercial éditeur (documents sonores et vidéos)
  + Titre
  + Auteur
  + Editeur
  + Lieu de publication
  + Date de publication

Puis, selon le type de notice/type de document, certaines seulement de ces métadonnées seront conservées dans le fichier en sortie, pour être utilisés comme critères d’alignements

## Zones utilisées pour générer chaque élément d’information

La description du contenu de chaque zone peut être trouvée notamment sur le site de l’Abes : <http://documentation.abes.fr/>

### Numéro FRBNF

035$a (uniquement s’il contient les lettres "frbn")

### Numéro ARK BnF

033$a (uniquement si contient "ark:/12148/")

### Numéro ISBN

010$a

### Numéro EAN

073$a

### Numéro ISSN

011$a

### Numéro commercial éditeur (documents sonores et vidéos)

071$a

### Titre

200$a + 200$e

La notion de titre clé est définie en parallèle, avec 530$a, ou à défaut 200$a+200$e

### Auteur

700$a + 700$b + 710$a + 710$b + 701$a + 701$b + 711$a + 711$b + 702$a + 702$b + 712$a + 712$b

Si toutes ces zones sont vides, l’auteur est récupéré tel que transcrit en 200$f

Les accents et signes de ponctuation sont nettoyés ; les mots sont passés en minuscules, leurs diacritiques (accents, cédilles, etc.) supprimés, et ils sont ensuite dédoublonnés

### Editeur

210$c

### Lieu de publication

210$a

### Date de publication

Zone codée 100$a pos.9-13

Si celle-ci n’est pas renseignée : zone 210$d ou à défaut zones 219$d+219$i+219$p

### Numéro de tome/volume

200$h, ou à défaut 461$v

## Zones exportées dans les fichiers en sortie

### Monographies – texte imprimé

FRBNF

ARK

ISBN

EAN

Titre

Auteurs

Date de publication

Numéro de tome

Editeur

### Monographies – son, vidéo

FRBNF

ARK

EAN

ID commercial audio

Titre

Auteurs

Date de publication

Editeur

### Périodiques

FRBNF

ARK

ISSN

Titre clé, ou à défaut Titre

Auteurs

Date de publication

Lieu de publication

### Autres types de notice-type de document

FRBNF

ARK

EAN

Titre

Auteurs

Date de publication

# Module d’export de notices BnF

A partir d’un fichier donnant une liste d’ARK (un ARK par ligne), ce module permet d’extraire des notices Bnf en interrogeant le SRU de la BnF.

La liste des ARK peut contenir des ARK de notices bibliographiques, ou des ARK de notices BnF (mais pas mélanger les 2).

## Option « Données d’autorité liées »

Si l’utilisateur met en entrée une liste d’ARK de notices bibliographiques, et demande à récupérer aussi le fichier de notices d’autorité liées, le programme ouvre chaque notice bibliographique et regarde les zones suivantes :

100, 141, 143, 144, 145, 600, 603, 606, 607, 609, 610, 616, 617, 700, 702, 703, 709, 710, 712, 713, 719, 731

Il identifie le $3 (qui fournit l’identifiant de la notice d’autorité) et va chercher la notice correspondante pour l’agréger dans le fichier d’autorités en sortie.

Une même autorité n’est exportée qu’une seule fois.