Alignement-donnees-bnf : spécifications techniques

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

|  |  |
| --- | --- |
| **Version** | 1.0 |
| **Date de dernière mise à jour** | 12/04/2018 |

Table des matières

[Module central (blanc) de Alignements-données-BnF : alignements – notices bibliographiques 2](#_Toc510698128)

[Fichier en entrée : format 2](#_Toc510698129)

[Monographies imprimées 2](#_Toc510698130)

[CD, DVD (son, vidéo) 3](#_Toc510698131)

[Périodiques 3](#_Toc510698132)

[Processus d’alignements 3](#_Toc510698133)

[Erreurs les plus fréquentes 3](#_Toc510698134)

[Alignement par ARK 4](#_Toc510698135)

[Alignement par FRBNF 4](#_Toc510698136)

[Alignement par ISBN 4](#_Toc510698137)

[Alignement par EAN 4](#_Toc510698138)

[Alignement par n° commercial (CD-DVD) 5](#_Toc510698139)

[Alignement par ISSN (périodiques) 5](#_Toc510698140)

[Alignement par mots (Titre-Auteur-Date-Editeur-Lieu de publication) 5](#_Toc510698141)

[Mécanismes de contrôle 5](#_Toc510698142)

[Fichier en entrée 6](#_Toc510698143)

[Fichiers en sortie 6](#_Toc510698144)

[Les fichiers de résultats 6](#_Toc510698145)

[Module central (blanc) de Alignements-donnees-BnF : alignements – notices d’autorité 6](#_Toc510698146)

[Processus d’alignement 6](#_Toc510698147)

[Fichiers en sortie 7](#_Toc510698148)

[Modules périphériques de Alignements-donnees-BnF : fonctionnalités facilitatrices pour l’usager 7](#_Toc510698149)

[Module (bleu) de conversion de fichier de notices en fichiers tabulés 7](#_Toc510698150)

[Extraction de notices bibliographiques - Principes 7](#_Toc510698151)

[Extraction de notices bibliographiques - Zones utilisées pour générer chaque élément d’information 7](#_Toc510698152)

[Extraction de notices bibliographiques - Zones exportées dans les fichiers en sortie 8](#_Toc510698153)

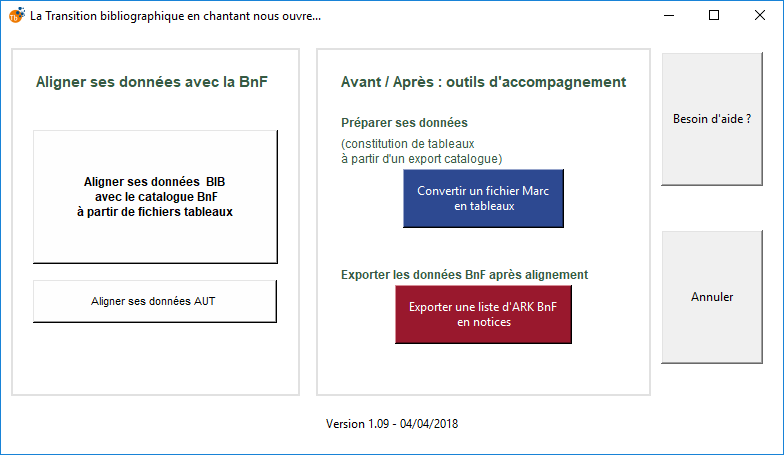
[Extraction de notices d’autorité – Principes 9](#_Toc510698154)

[Module (rouge) d’export de notices BnF 9](#_Toc510698155)

[Option « Données d’autorité liées » 9](#_Toc510698156)

[Option iso2709 / XML 9](#_Toc510698157)

[Option Unimarc / Intermarc 9](#_Toc510698158)



# Module central (blanc) de Alignements-données-BnF : 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.

La préparation d’un tel fichier peut parfois s’avérer plus délicate que prévue, en fonction du SIGB et de la base de données sources. Le logiciel intègre donc des modules facilitateurs décrits dans le chapitre 2 : Modules facilitateurs pour l’usager.

## 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

### Alignement par ARK

Recherche sur l’ARK. Quelle que soit la notice trouvée, son ARK (identique ou différent à l’ARK en entrée) est retourné

### Alignement par FRBNF

1. Recherche sur le nombre complet avec préfixe sur le critère de recherche bib.otherid du SRU BnF.
2. Si pas de résultat, recherche des 8 premiers chiffres comme numéro de notice actuel + contrôle sur ISBN
   * Si le contrôle sur ISBN échoue (ou si pas d’ISBN), contrôle sur le titre  
     Si le contrôle sur le titre est valide et qu’il y a un numéro de volume, contrôle sur le numéro de volume
3. Si pas de résultat, recherche sur les 9 premiers chiffres comme ancien numéro de notice (zone 9XX) + contrôle sur ISBN
   * Si le contrôle sur ISBN échoue (ou si pas d’ISBN), contrôle sur le titre + volume
4. Si pas de résultat, recherche sur les 8 premiers chiffres comme ancien numéro de notice (zone 9XX) + contrôle sur ISBN
   * Si le contrôle sur ISBN échoue (ou si pas d’ISBN), contrôle sur le titre + volume

### Alignement par ISBN

1. Recherche sur l’ISBN tel que présent dans le fichier dans le critère SRU bib.isbn  
   + contrôle sur le titre et la date
2. Si échec : recherche ISBN + Auteur (mot d’auteur le plus long dans la zone Auteur du fichier en entrée)   
   + contrôle sur la date
3. Si échec : recherche sur l’ISBN nettoyé : uniquement le premier ISBN (s’il y a en entrée plusieurs ISBN séparés par des ponctuation), en supprimant les espaces, les tirets, etc.  
   + contrôle sur le titre et la date
4. Si échec : recherche sur ISBN nettoyé + Auteur  
   + contrôle sur la date
5. Si échec, conversion de l’ISBN en entrée : ISBN10 en ISBN13 ou ISBN13 en ISBN10  
   + mêmes contrôles
6. Si échec, on refait les mêmes recherches, mais dans toute la notice (critère de recherche SRU bib.anywhere)
7. Si échec : recherche de l’ISBN10 + ISBN13 dans le Sudoc (web service isbn2ppn)
   * Pour chaque notice trouvée, on regarde s’il y en a une avec un FRBNF
   * Si au moins une notice avec FRBNF est trouvée, on cherche l’ARK et on renvoie l’ARK
   * Sinon, on renvoie le(s) PPN

### Alignement par EAN

Mêmes principes que pour l’ISBN :

* Recherche sur bib.ean + contrôle sur titre, volumaison et date
* Si échec, recherche dans toute la notice
* Si échec, recherche dans le Sudoc (web service ean2ppn)

### Alignement par n° commercial (CD-DVD)

Recherche du numéro commercial dans le critère bib.comref :

* Contrôle sur la 071$a de la notice BnF : si le numéro commercial en entrée est présent dans la 071$a de la notice BnF, ou si la notice BnF contient le numéro commercial initial
* + contrôle sur titre, volumaison et date

### Alignement par ISSN (périodiques)

* Recherche de l’ISSN dans le critère de recherche SRU bib.issn  
  + contrôle sur le type de notice pour chaque notice trouvée
* Si échec : recherche dans le Sudoc (web service issn2ppn)

### Alignement par mots (Titre-Auteur-Date-Editeur-Lieu de publication)

Titre nettoyé (suppression des accents et ponctuations, tout passé en minuscules) + conservation uniquement des mots des plus de 1 lettre

Recherche sur Titre + Auteur + Date + Editeur + Lieu de publication

+ Contrôle sur le titre

+ Contrôle sur la date

+ Contrôle sur la volumaison

**Spécifique pour les périodiques :**

Recherche sur Titre – Auteur – Date – Lieu de publication

Si aucune réponse 🡪 on modifie la date à -3 ans et +3 ans

### Mécanismes de contrôle

#### Sur l’ISBN

Récupération et nettoyage de la racine de l’ISBN (9 premiers chiffres d’un ISBN10, ou 4e-12e chiffres d’un ISBN13) du fichier en entrée

Recherche de l’ISBN dans la notice BnF (010$a, première occurrence uniquement) et nettoyage de sa racine.

Comparaison des 2 valeurs

#### Sur le titre

Le logiciel reprend d’abord la chaîne de caractères Titre en entrée et la normalise :

* Suppression des ponctuations et des espaces
* Suppression des diacritiques
* Passage de toutes les lettres en minuscules

Il compare la chaîne de caractères obtenue au titre de la notice BnF, cherchée successivement en :

* 200$a (titre)
* Si échec, successivement : 200$e (complément de titre), 200$i (titre de partie), 750$a (autre forme de titre), 753$a (autre forme de titre de périodique), 500$a (titre uniforme), 500$e (titre uniforme, complément), 503$a, 503$e (titre de forme), 540$a, 540$e (titre ajouté par le catalogueur), 410$a (collection), 225$a (collection)
* Si échec, il faut la comparaison sur la moitié de la longueur du titre

#### Sur la date

Récupération et nettoyage de la date du fichier en entrée (on ne conserve que les chiffres)

Récupération de la date de la notice BnF :

* 4 premiers caractères de la zone 100$a -> on prend aussi les années n-1 et n+1
* Zone 210$d : on ne garde que les chiffres

Si la date en entrée est présente dans les dates récupérées de la notice BnF, le contrôle est valide.

#### Sur le numéro de tome/volume

Tous les chiffres romains sont passés en chiffres arabes pour faciliter les comparaisons.

Récupération des chiffres présents dans les zones :

* 200$h
* 200$u
* 461$v
* (si ces zones n’existent pas, récupération des chiffres en 200$a)

Si le numéro de tome en entrée est présent dans un des numéros trouvés dans la notice BnF, le contrôle est valide.

## Fichier en entrée

Fichier tabulé, encodé en UTF-8, dont les colonnes correspondent à celles de l’option choisie (TEX, VID, PER)

## Fichiers en sortie

Le fichier renvoie :

* Un ou plusieurs fichiers donnant le résultat des alignements (selon l’option 1 fichier / plusieurs fichiers choisie dans le formulaire)
* Un fichier « statistiques » qui renvoie global de notices pour lequel 1 ARK a été trouvé ; ou 2 ARK ; ou 3 ARK, etc.  
  Ce fichier contient également (si besoin) la liste des problèmes rencontrés : indisponibilité d’un web service pendant une certaine durée, ou URL renvoyant ponctuellement un message d’erreur

### Les fichiers de résultats

Ces fichiers contiennent systématiquement les colonnes suivantes :

* Le numéro de notice en entrée
* Le nombre d’identifiants trouvés
* La liste des identifiants trouvés (ARK ou PPN, séparés par une virgule)
* La méthode qui a permis de trouvé le(s) identifiant(s)
* Toutes les métadonnées fournies en entrée
* Si la case « Métadonnées BnF » a été cochée dans le formulaire : les métadonnées BnF de l’ARK ou des ARK trouvés

# Module central (blanc) de Alignements-donnees-BnF : alignements – notices d’autorité

Ce module permet d’aligner des notices d’autorité avec les autorités BnF – uniquement des personnes physiques et organisations.

Deux types de fichier en entrée sont possibles :

* Soit on a exporté un fichier d’autorités : les métadonnées servant à l’alignement sont donc les points d’accès (nom + dates)
* Soit on a exporté un ensemble de notices bibliographiques : le logiciel trouvera d’abord le document, pour remonter ensuite à l’auteur

Ce module a été beaucoup moins développé que le module d’alignement sur les données bibliographiques.

## Processus d’alignement

Comme pour les notices bibliographiques : alignement d’abord sur les n° internes BnF (ARK ou FBRNF).

A défaut (si absents ou aucun résultat trouvé), utilisation de la colonne ISNI

A défaut (si ISNI absent ou aucun résultat trouvé), utilisation du point d’accès (combinaison Nom – Prénom – Date de naissance)

A défaut : recherche du point d’accès dans la base [isni.org](http://isni.org/)

Toutes les colonnes indiquées dans le formulaire doivent être présentes, même si elles ne sont jamais renseignées.

## Fichiers en sortie

Cf. la description des fichiers en sortie pour les alignements bibliographiques

# Modules périphériques de Alignements-donnees-BnF : fonctionnalités facilitatrices pour l’usager

## Module (bleu) de conversion de fichier de notices en fichiers tabulés

### Extraction de notices bibliographiques - 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

### Extraction de notices bibliographiques - 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

### Extraction de notices bibliographiques - 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

### Extraction de notices d’autorité – Principes

Les principes sont les mêmes que pour les notices bibliographiques : on met en entrée un fichier Unimarc de notices d’autorité, et le programme répartit ces notices par types (personnes physiques, organisations, mots-matières, oeuvres)

## Module (rouge) 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.

### Option iso2709 / XML

Possibilité que les fichiers en sortie soient en iso2709 (encodés en UTF-8) ou en XML (encodés en UTF-8 aussi)

### Option Unimarc / Intermarc

Possibilité de choisir le format MARC en sortie.