Transition bibliographique

# Les data sans aléas :

## connaître ses métadonnées pour FRBRiser son catalogue

Groupe systèmes et données

Mis à jour le 30 décembre 2017

# Argumentaire

En amont d'un chantier de FRBRisation, il est nécessaire de bien connaître ses données pour les faire évoluer efficacement (notamment : maîtriser et harmoniser les conditions de production, identifier des lots de notices afin de préparer des corrections efficaces et automatisées).
L'atelier s'attachera à présenter les enjeux d'une bonne connaissance des métadonnées et permettra d'échanger sur des exemples de méthodologie.

# Partie 1 :
# Connaître ses données, problématiques

# Préambule

**Précision : on parle de métadonnées** => expliquer la distinction entre connaître ses métadonnées et connaître la composition des fonds (i.e. ce que les métadonnées sont censées décrire)

# Enjeux

**Un chantier de FRBRisation, c'est quoi ?**

• pour la majeure partie des bibliothèques, il s'agit d'aligner ses données sur celles de grands réservoirs (BnF, ABES en particulier).

# Une connaissance parfois « brouillée » des métadonnées

- Deux grandes familles de facteurs concourent à ce que la nature et la structuration des métadonnées du catalogue soient mal connues :
  - Hétérogénéité des pratiques
  - Hétérogénéités des sources

# Des pratiques de catalogage hétérogènes

- Présence de différentes strates chronologiques
- Pratiques différentes selon les services ou les supports

# Des sources hétérogènes

- disparités des sources
- mauvais alignement des données
- données mal reprises lors d'un changement de SIGB

# Des imports parfois mal maîtrisés

- Difficulté à obtenir les évolutions des passerelles et filtres d'import de nos SIGB par les fournisseurs
- Veille parfois pas assurée sur l'évolution des formats et des référentiels de travail

# Identifier les lacunes des métadonnées

- Apprendre à repérer les zones clés pour la FRBRisation et à en dresser un état des lieux

# Partie 2 :
# Connaître et maîtriser le processus d'entrée des données dans notre SIGB

# Différentes étapes

- Identifier les sources d'entrée dans le catalogue
- Mettre à jour le format dans le SIGB
- Manipuler les données
- Paramétrer les imports

# Des notices descriptives structurées

- dans le respect de normes
- dans des formats Marc (Unimarc le plus souvent)

# Structuration des données dans le format

**010** .. **$a** 2-07-040312-2 $b br. $d 28 F

101 0. $a fre

| | |
|---|---|
| FI | Finlande |
| FJ | Fidji |
| FK | Falkland, Iles (Malouines) |
| FM | Micronésie, Etats fédérés de |
| FO | Féroé, Iles |
| FR | France |
| GA | Gabon |
| GB | Royaume-Uni |

200 1. $a Des journées entières dans les arbres $b Texte imprimé $f Marguerite Duras

700 . | $3 11901349 $a Duras $b Marguerite $f 1914-1996 **$4** 070

**Duras, Marguerite (1914-1996)** *pseudonyme forme internationale*

**Nationalité(s)** : France

**Langue(s)** : français

**Sexe** : Féminin

**Responsabilité(s) exercée(s) sur les documents** : Auteur, Interprète, Participant

**Naissance** : 1914-04-04, Gia Dinh (Vietnam)

**Mort** : 1996-03-03, Paris

Romancière, cinéaste et dramaturge. - Pseudonyme de Marguerite Donnadieu

| | |
|---|---|
| 040 | artiste |
| 050 | titulaire des droits |
| 060 | nom associé |
| 065 | commissaire-priseur |
| 070 | auteur |
| 072 | auteur des citations ou fragments textuels |
| 075 | auteur de la postface, du colophon, etc. |
| 080 | préfacier, etc. |

# 2.1 Sources d'entrée dans le catalogue

- Identifier les sources d'entrée possibles
- Mieux les contrôler

- 3 cas

Duplication et recyclage de notices
Catalogage ex nihilo
Import d'une source externe

# Duplication et recyclage de notices

- Éviter la duplication et le recyclage de notices
- Une publication = Une notice

# Catalogage ex nihilo

- Respecter le format et les normes
- Dans la perspective d'un enrichissement futur

# Import d'une source externe

- Encourager la récupération de notices
- Choisir des sources fiables

# 2.2 Mettre à jour le format dans le SIGB

- Formats Unimarc standards
- Formats de diffusion

- Qui peut le faire et comment ?

# Le format en intégralité

- Déclarer toutes les zones et sous zones standards
- Mettre à jour les référentiels embarqués

# Veille sur les formats

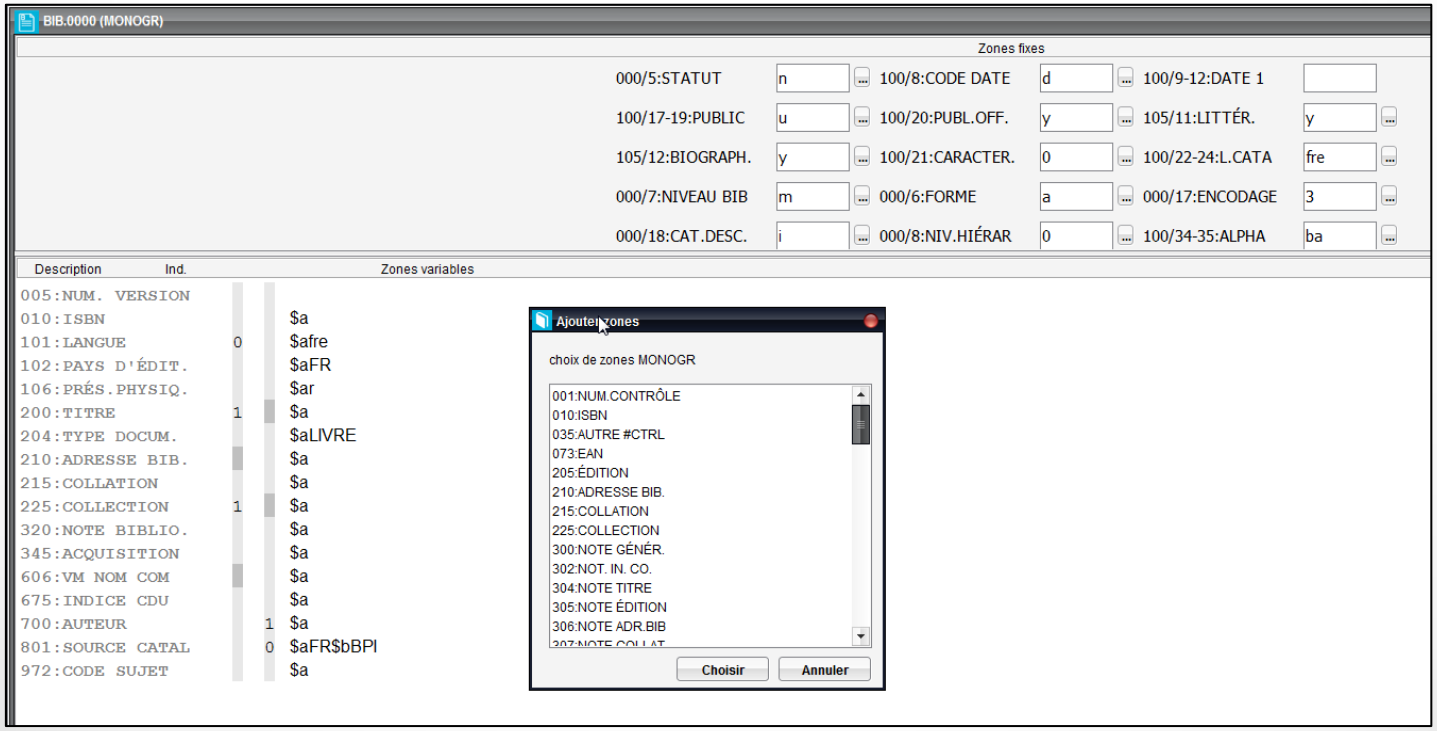
- Assurer une veille sur l'évolution des formats
- Mettre à jour le format intégré au SIGB

# 2.3 manipulation des données

- production courante :
  saisie des données dans le format
- traitement du rétrospectif :
  interventions sur les données saisies

# La production courante

- recommandations pour le courant
- paramétrage de contrôles de saisie
- révision du paramétrage des bordereaux de saisie



# Le rétrospectif

- inventaire des pratiques
- identification de données dans des zones inappropriées
- ménage préalable dans les notices
- déplacements de données par lots

# Exemple d'extraction via pgAdmin

- voir comment est renseignée la zone 305
- restitution dans sous forme de tableur

notices305a_sauf_rev.ods

Fichier Édition Affichage Insertion Format Données Outils Modules complémentaires Aide Dernière

100% € % .0 .00 123 Arial 10

fx seq_no

| | A | B | C |
|---|---|---|---|
| 13875 | 1317740 | Publié pour la 1ère fois aux Presses universitaires de France en 2007 dans la collection "Que sais-je ?" | LIVRE |
| 13876 | 796407 | Paru précédemment aux éd. de la Découvrance | LIVRE |
| 13877 | 1100342 | Précédemment paru sous le titre : La France en mutations : quand l'incertitude fait société | LIVRE |
| 13878 | 1160249 | Fait suite à : Le Premier ministre | LIVRE |
| 13879 | 1126854 | En espagnol avec encarts en français | LIVRE |
| 13880 | 262982 | Bibliogr. | LIVRE |
| 13881 | 723648 | Fac-sim. de l'éd. G. Baillière, 1875 | LIVRE |
| 13882 | 333778 | No spécial: Champs visuels, 12-13 | LIVRE |
| 13883 | 245813 | Index | LIVRE |
| 13884 | 548080 | No spécial: Europe, hors-série, | LIVRE |
| 13885 | 250282 | Bibliogr. p. 76-77 | LIVRE |
| 13886 | 861408 | Numéro de :Cahier - Groupe Afrique noire = ISSN 0981-339X, 24 (2008) | LIVRE |
| 13887 | 1203877 | Fait suite à : Ward : Ier-IIe siècle | LIVRE |
| 13888 | 700026 | No spécial: Beaux-arts magazine, hors série, | LIVRE |
| 13889 | 1052854 | Fac-sim. de l'ed. de Paris : Flammarion, 1916 | LIVRE |
| 13890 | 1086945 | Fac-similés des numéros 18 à 27 (1913-1914) | LIVRE |

# 2.4 Paramétrage et raffinement des imports

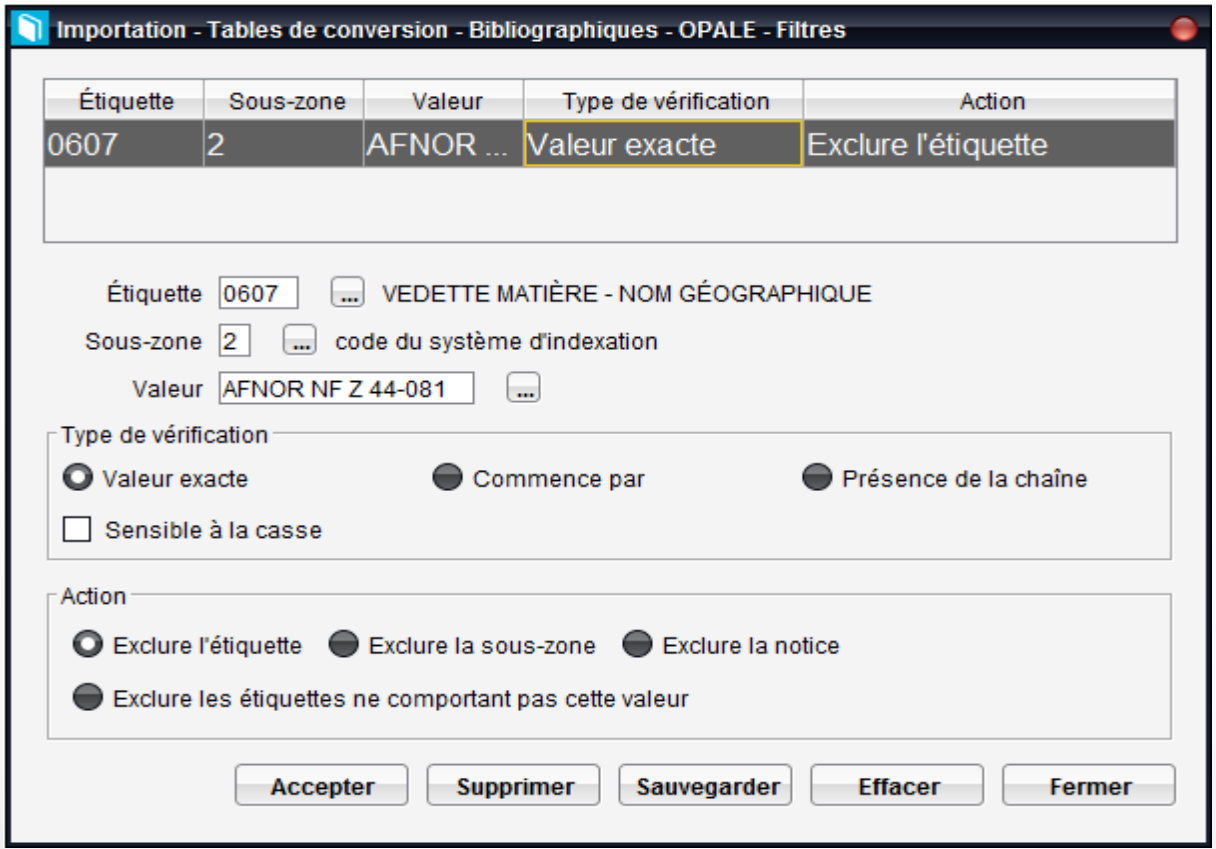
- les clefs de mise à jour
- les filtres
- raffinements des imports

# Les clés de mise à jour

- les identifiants : pivots des imports
- ISBN = zone 010
- numéros FRBNF, PPN = zone 001 vers zone 035
- identifiants pérennes ARK = zone 003 vers zone 033

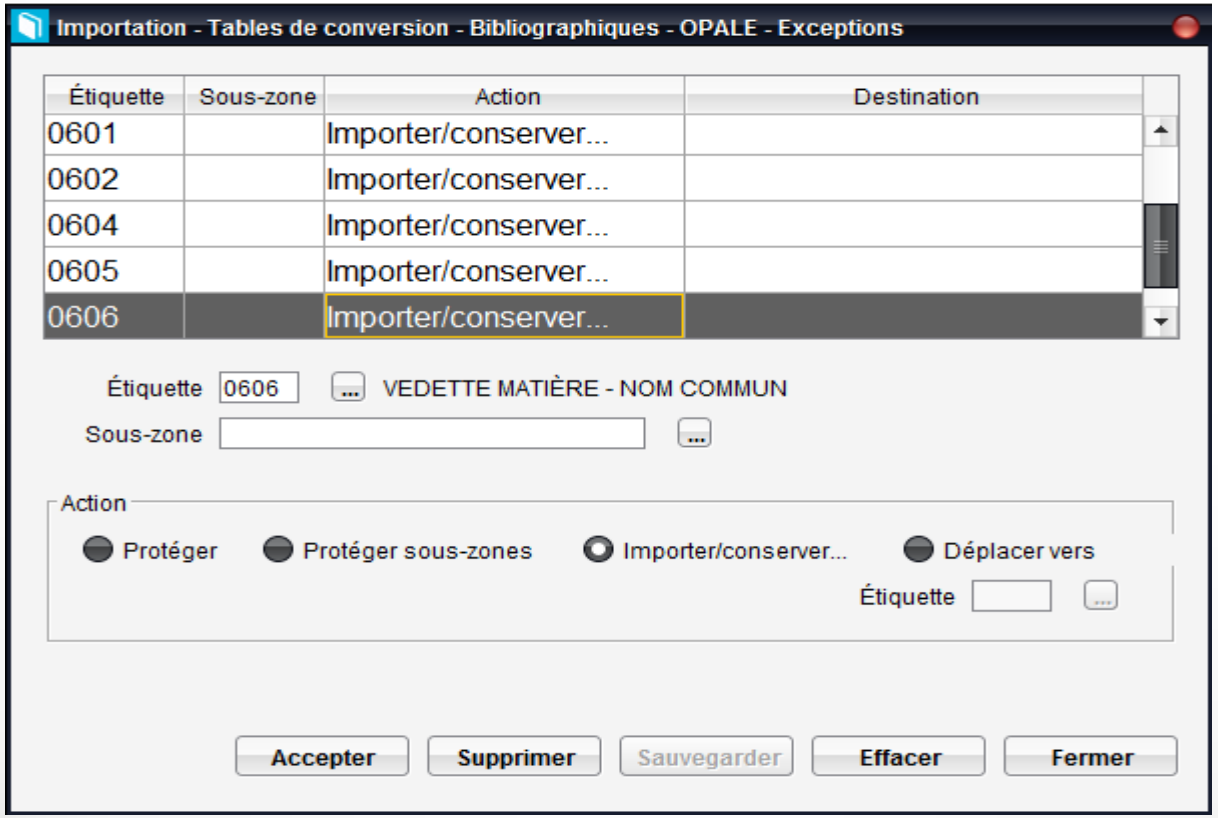
# Les filtres

- outils de contrôle qualité des données entrantes
- exclure certaines données non pertinentes



# Raffinement des imports

- protection de zones ne devant pas être mises à jour lors de l'import
- préservation des valeurs ajoutées propres



# Partie 3

## Un outil pour analyser des données

## Marc : Catmandu

- présentation à consulter à l'adresse suivante : https://github.com/medrbx/dsa
  - le fichier README.md constitue la trame détaillée,
  - les répertoires input/ et output/ contiennent les données d'entrées et de sortie
    => de sorte que quelqu'un ne disposant pas de Catmandu puisse examiner le résultat

- ## Problématique

Une fois que l'on sait ce que l'on doit chercher / vérifier au sein de nos notices, concrètement comment faire ? quels outils utiliser, sachant que les SIGB ne proposent pas nécessairement les fonctionnalités adéquates ? => on va partir de l'hypothèse qu'on est en mesure d'extraire du SIGB un fichier marc (ISO ou MARCXML) propre, auquel on va faire subir quelques transformations grâce à Catmandu pour effectuer ensuite une analyse avec des outils statistiques (tableur, etc...)

- ## Catmandu ?

Un ETL spécifique aux formats et protocoles utilisés en bibliothèques

Principe :

- on donne des données en entrée (par exemple sous forme de fichier csv, xml, marc (iso ou xml), ...),
- on indique éventuellement quels transformations appliquer,
- on récupère des données en sortie, que l'on va pouvoir intégrer dans un autre outil

# Trois exemples, à travers trois questions simples

- Exemple 1 : Comment puis-je savoir si mes notices comportent des identifiants qui me permettront d'effectuer un alignement avec les données de la BnF ?

- Exemple 2 : Comment connaître la composition d'un fichier autorités ? quelle part de notices récupérées auprès d'une agence comme la BnF ? quelle répartition selon les types d'autorités (nom de personne, nom de collectivité, sujet nom commun, ... ) ?

- Exemple 3 : Comment mettre en place un tableau de bord pour effectuer du contrôle qualité ?

## Exemple 1 : Comment puis-je savoir si mes notices comportent des identifiants qui me permettront d'effectuer un alignement avec les données de la BnF ?

ce que l'on va faire : produire à partir d'un fichier marc un fichier csv comportant quelques colonnes essentielles qu'on analysera via un tableur.

Etapes :

1. On a exporté au préalable du SIGB les notices bibliographiques sous forme de fichier unimarc ISO 2709 encodé en UTF-8, que l'on nomme input/biblio.mrc. Ici, on utilisera un fichier représentant seulement 5 % des notices bibliographiques de Roubaix, pour accélérer les temps de traitement.

2. On crée un fichier fix pour Catmandu, que l'on nommera fix/biblio.fix.

3. On exécute la commande suivante :

```
$ catmandu convert -v MARC --fix fix/biblio.fix to CSV < input/biblio.mrc > output/biblio.csv
```

## Exemple 2 : Comment connaître la composition d'un fichier autorités ? quelle part de notices récupérées auprès d'une agence comme la BnF ? quelle répartition selon les types d'autorités (nom de personne, nom de collectivité, sujet nom commun, ... ) ?

ce que l'on va faire : produire à partir d'un fichier marc un fichier csv comportant quelques colonnes essentielles qu'on analysera via un tableur.

Etapes :

1. On a exporté au préalable du SIGB les notices autorités sous forme de fichier unimarc ISO 2709 encodé en UTF-8, que l'on nomme input/auth.mrc. Ici, on utilisera un fichier représantant seulement 5 % des notices autorités de Roubaix, pour accélerer les temps de traitement.

2. On crée un fichier fix pour Catmandu, que l'on nommera fix/auth.fix .

3. On exécute la commande suivante :

```
$ catmandu convert -v MARC --fix fix/auth.fix to CSV < input/auth.mrc > output/auth.csv
```

## Exemple 3 : Comment mettre en place un tableau de bord pour effectuer du contrôle qualité ?

ce que l'on va faire : pour suivre les imports et les remplacements effectués au quotidien, il est nécessaire de mettre en place des tableaux de bord, régulièrement mis à jour. Montrer comment l'on peut faire cela simplement avec Elasticsearch / Kibana, grâce à des exports effectués via Catmandu.

Etapes :

1. On reprend les paires de fichiers biblio.mrc, biblio.fix et auth.mrc, auth.fix des deux exemples précédents.

2. On présuppose que l'on a au préalable installé et configuré Elasticsearch et Kibana. On exécute les deux commandes suivantes :

```
$ catmandu import -v MARC --fix fix/biblio.fix to ES --index-name 'catmandu_ex'
--bag 'biblio' < input/biblio.mrc
$ catmandu import -v MARC --fix fix/auth.fix to ES --index-name 'catmandu_ex'
--bag 'auth' < input/auth.mrc
```

## Pour aller plus loin : Catmandu comme outil de prototypage

Réaliser les opérations d'alignement peuvent être complexes à réaliser au sein d'un SIGB, on peut en revanche réaliser des prototypes à l'aide de Catmandu.

Exemple : pour chaque notice du catalogue disposant d'un identifiant type isbn / issn / ean, lancer une requête sur le service SRU de la BnF pour récupérer un identifiant ark et l'ajouter à la notice locale.

# Conclusion :
## des outils...

- Moyens techniques :
  - se doter d'outils de contrôle qualité
  - se doter d'outils d'interrogation des données
  - disposer d'un SIGB permettant un paramétrage souple et simple à administrer

# Echanges avec la salle

- Question 1 : Parmi les services SRU et Z39.50 de la BnF, lequel choisir ?

Si Z39.50 est le protocole "historique", mis en place et utilisé depuis plusieurs années, SRU est plus facile d'utilisation (les interrogations passent par des requêtes HTTP et propose des possibilités d'interrogation bien plus larges, en particulier sur les notices d'autorités.
Pour plus d'informations, voir les pages de la BnF consacrées aux services SRU et Z39.50 .

- Question 2 : Le groupe Systèmes et Données recommande-t-il l'usage d'un système d'information documentaire libre plutôt que propriétaire ?

Le groupe n'a pas vocation à s'aventurer sur un tel terrain. On insistera en revanche sur la nécessité d'avoir accès aux données, a minima pour effectuer un export dans un format d'échange (MARC ISO 2709 ou marcxml par exemple) ou, idéalement, pour réaliser des modifications / enrichissement dans le système, sans recourir à une prestation de l'éditeur.

# Echanges avec la salle (suite)

- Question 3 : Puis-je aussi analyser mes données avec MarcEdit plutôt qu'Open Refine ou Catmandu, voire avec des outils directement en ligne ?

A notre connaissance, il n'existe pas d'outils en ligne simple. Catmandu ou Open Refine peuvent faire peur mais sont accessibles à des non informaticiens.
MarcEdit est plus aisé à prendre en main, mais est davantage tourné vers le MARC21 et surtout n'est pas en mesure d'analyser des fichiers volumineux de données (plus de 100 000 notices).

# Contacts

Pour toute précision, n'hésitez pas à nous solliciter :

- Sylvain Franceschi, Médiathèques de Montpellier Méditerranée Métropole
- Annick Le Follic, Bibliothèque nationale de France
- Sylvie Lemaire, Bibliothèques de l'Université de Rennes 2
- Karine Meneghetti, Bibliothèque publique d'information
- François Pichenot, Médiathèque de Roubaix
- Claudine Rovis, Bibliothèque municipale de Nice

ou à contacter le groupe "Systèmes et Données" via le site de la Transition bibliographique.