Réseau des bibliothèques de Suisse occidentale
Westschweizer Bibliotheksverbund
Rete delle biblioteche della Svizzera occidentale
Library Network of Western Switzerland

Évaluation de Bibliostratus en vue de l'alignement des notices bibliographiques RERO avec la BnF

Sujet	Cette analyse, produite à des fins internes et rendue publique, documente le processus d'évaluation du logiciel Bibliostratus sur les données RERO et s'intéresse notamment à la qualité des résultats. Le logiciel Bibliostratus, développé par la BnF dans le contexte du projet de transition bibliographique, permet d'aligner des notices bibliographiques (manifestation) avec celles de la BnF ou du SUDOC. Des recommandations concluent ce document en vue d'une application de Bibliostratus sur l'ensemble des données RERO.
Auteurs	Kathia Darbellay, Nicolas Prongué
Date	24.06.19
Distribution	Publique

Table des matières

Introduction

Cette analyse documente le processus d'évaluation du logiciel Bibliostratus sur les données RERO en vue de l'intégration d'identifiants BnF et SUDOC dans la base de production de RERO.

L'évaluation de Bibliostratus a été réalisée en trois phases :

1. un premier test avec la **version 1.25** sur l'échantillon complet de 9'820 notices
2. un second test avec la **version 1.26 bêta** sur le sous-ensemble « texte-monographie » de ce même échantillon (5928 notices, cf. tableau 1). Cette version est le résultat d'adaptations suite à notre retour du premier test pour inclure notamment un meilleur signalement des problèmes dans le champ « Méthode d'alignement ».
3. un troisième test de confirmation avec la **version 1.26**.

Échantillonnage (notices utilisées)

Il s'agit d'un lot de 9'820 notices (le même celui utilisé pour RERO ILS), comprenant environ l'assortiment suivant :

- livre (5935)
- article (3200)
- journal (300)
- partition (200)
- son (100)
- vidéo (84)

Cet assortiment a été généré de manière artificielle, mais correspond plus ou moins à la proportion de ces types de document dans l'ensemble des notices du catalogue collectif RERO.

Types de document pris en charge par Bibliostratus

L'étape blanche a généré 13 fichiers selon les types de documents, résumés dans le tableau 1. Selon ces résultats, environ 65 % des notices RERO sont candidates à l'alignement par Bibliostratus. A noter que l'étape blanche a fourni exactement les mêmes résultats pour les versions 1.25, 1.26 bêta et 1.26.

No	Type de document	Nombre de notices	Prise en charge par Bibliostratus
1	texte-monographie	5928	✔
2	texte-analytique	3200	
3	texte-périodiques	300	✔
4	partition-monographie	163	
5	video-monographie	90	✔
6	son - musique-monographie	76	✔
7	partition-analytique	25	
8	son - non musical-monographie	24	✔
9	partition manuscrite-collection	7	
10	partition manuscrite-monographie	3	
11	partition-collection	2	
12	video-analytique	1	
13	video-collection	1	
		9820	**6418 (65%)**

Tableau 1 : Résultats de la 1ère étape (blanche)

Les types de document identifiés par Bibliostratus ne correspondent pas exactement aux types de documents présents dans RERO ILS (tableau 2). Les différences sont marquées en rouge.

Type de doc RERO ILS	Type de document Bibliostratus	Nombre de notices
livre (5935)	texte-monographie	5928
article (3200)	texte-analytique	3200
journal (300)	texte-périodiques	300
partition (200)	partition-monographie	163
son (100)	son - musique-monographie	76
vidéo (84)	video-monographie	90
	video-analytique	1
	video-collection	1

Tableau 2: Comparaison des types de document de Bibliostratus et de RERO ILS

Performance

1ère étape – étape bleue (génération des fichiers tabulaires sur la base des notices MARC ISO) :

- ensemble des notices traitées (9820)
- temps de traitement : < 1 min

2e étape – étape blanche (alignement avec BnF/SUDOC) :

- Nombre de notices traitées : 6418 (65 % des notices de l'étape 1)
  - « texte-monographie » : 5928
  - « texte-périodique » : 300
  - « vidéo-monographie » : 90
  - « son - musique-monographie » : 76
  - « son - non musical-monographie » : 24

Temps de traitement estimé : 2,7 sec. par notice

Processus ralenti par la mise en veille de l'ordinateur lors du test.

Résultats de l'alignement

La figure 1 illustre les résultats d'alignement pour chacune des catégories prises en charge par Bibliostratus (sauf les cartes, qui n'étaient pas présentes dans l'échantillon). Pour les monographies textuelles, le processus a été lancé une seconde fois avec la version 1.26 bêta. Le paramétrage de Bibliostratus a été fait de manière à prioriser d'abord un alignement avec la BnF, et à défaut avec SUDOC.

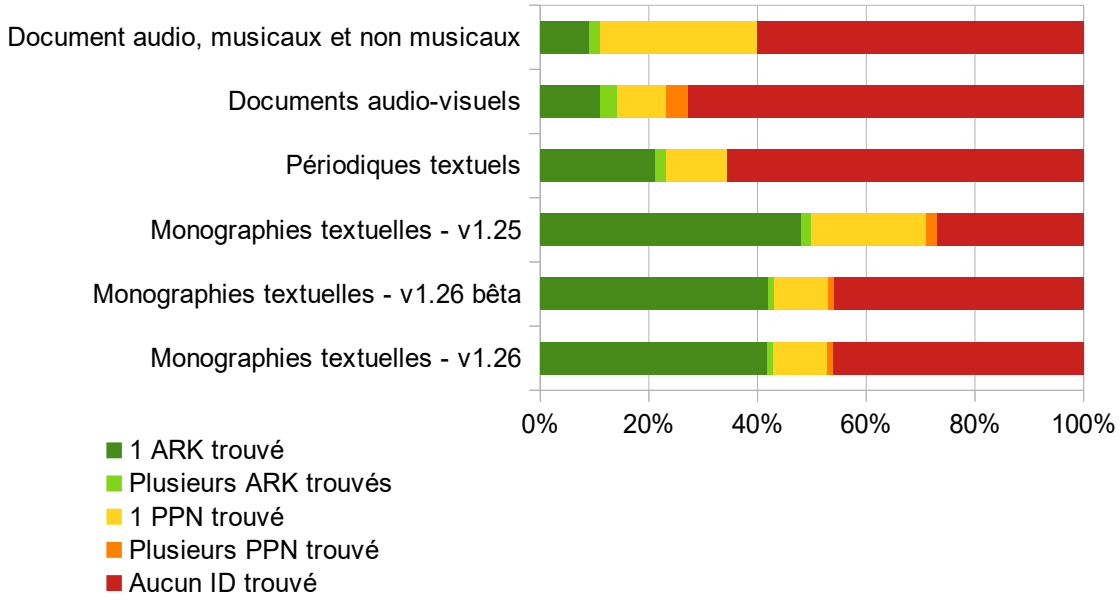

Figure 1: Résultats de l'alignement

La synthèse des résultats de l'alignement se trouve dans le tableau 3. Ces chiffre se réfèrent uniquement aux notices prises en charge par Bibliostratus, qui correspondent environ à 65 % de l'ensemble des données. Le recoupement avec un et un seul identifiant ARK est d'environ 40-50 % pour les livres (monographies), 20 % pour les périodiques et 10 % pour les documents audio-visuels.

Type de document	Total	1 ARK trouvé	Plusieurs ARK trouvés	1 PPN trouvé	Plusieurs PPN trouvés	Aucun ID trouvé
texte-monographie (v1.25)	100.0%	47.8%	1.5%	21.3%	2.3%	27.1%
texte-périodiques	100.0%	21.3%	2.0%	11.3%	0.0%	65.3%
video-monographie	100.0%	11.1%	8.9%	3.3%	4.4%	72.2%
son – musique-monographie + son - non musical-monographie	100.0%	9.0%	29.0%	2.0%	0.0%	60.0%
Ensemble des données testées	**100.0%**	**45.5%**	**2.1%**	**20.3%**	**2.2%**	**30.0%**

Tableau 3 : Synthèse de l'alignement par types de document (pourcentages par rapport à l'ensemble des éléments testés)

Les versions 1.26 bêta et 1.26 ont donné les mêmes résultats à quelques unités près (tableau 4).

Type de document	Total	1 ARK trouvé	Plusieurs ARK trouvés	1 PPN trouvé	Plusieurs PPN trouvés	Aucun ID trouvé
Texte-monographie v1.25	100.0%	47.8%	1.5%	21.3%	2.3%	27.1%
Texte-monographie v1.26 bêta	100.0%	41.73%	1.13%	10.02%	1.00%	46.12%
Texte-monographie v1.26	100.0%	41.77%	1.13%	10.05%	0.98%	46.07%

Tableau 4 : Synthèse de l'alignement par types de document (pourcentages par rapport à l'ensemble des éléments testés)

Méthodes d'alignement

Dans le fichier des résultats de l'alignement, Bibliostratus génère une note concernant la méthode (log de toutes les étapes réalisées par le logiciel) utilisée pour chaque alignement.

Une brève analyse a été réalisée, uniquement pour les monographies textuelles, et a révélé qu'environ 200 processus différents ont été utilisés selon les cas, et que leurs intitulés sont difficilement exploitables dans le cadre d'un contrôle qualité.

Ce retour a été fait à la BnF, qui a produit la version 1.26 bêta et 1.26 de Bibliostratus, en améliorant le signalement des problèmes dans le champ « Méthode ». Le tableau 5 montre le top 10 des méthodes par nombre de notices concernées (uniquement les méthodes ayant généré un alignement avec un seul ARK). En jaune sont marquées les méthodes où un problème est signalé. Les résultats sont identiques pour les versions 1.26 bêta et 1.26, à quelques unités près.

Intitulé de la méthode	Notices concernées (incl. % des alignements avec 1 ARK)	
	v1.26 bêta	v1.26
ISBN + contrôle Titre 200$a	1640 (66%)	1643 (66%)
ISBN + contrôle Titre 200$a[demi-titre-Xcaractères]	379 (15%)	379 (15%)
ISBN | Problèmes dans métadonnées Titre-Auteur-Date-Volume	208 (8%)	209 (8%)
ISBN + contrôle Titre BnF contenu dans titre initial	49 (2%)	49 (2%)
Titre-Auteur-Date + contrôle Titre 200$a,Problème ISBN non reconnu	48 (2%)	49 (2%)
ISBN + Auteur > ARK	42 (2%)	41 (2%)
ISBN + contrôle Titre 225$a	13 (1%)	13 (1%)
[titre court],ISBN + contrôle Titre 200$a	11 (0%)	11 (0%)
ISBN + contrôle Titre 200$i	11 (0%)	11 (0%)
Titre-Auteur-Date + contrôle Titre 200$a[demi-titre-Xcaractères],Problème ISBN non reconnu	9 (0%)	9 (0%)

Tableau 5 : Top 10 des méthodes par nombre de notices concernées (uniquement les méthodes ayant généré un alignement avec un seul ARK), pour les méthodes v1.26 bêta et v1.26

Les trois méthodes du tableau 5 signalant un problème ont fait l'objet d'un contrôle qualité aléatoire de cinq alignements (cf. section « Contrôles qualité »).

Contrôles qualité

Version 1.25

Un rapide contrôle qualité a été effectué sur les alignements des monographies textuelles (version 1.25), en prenant chaque fois deux notices par méthode d'alignement :

- uniquement pour les méthodes utilisées pour plus de 40 notices de l'échantillon
- uniquement pour les méthodes aboutissant à un alignement avec un et un seul ARK

Les résultats sont bons, mais 10 faux positifs ont été identifiés. Ce sont tous des faux positifs où l'édition exacte ne correspond pas, mais l'expression est la même. Il semble que la date ne soit pas un critère déterminant pour l'établissement d'une correspondance.

Version 1.26 bêta

Parmi les 10 faux positifs identifiés avec la version 1.25, la version 1.26 bêta :

- fournit le même alignement dans 9 cas sur 10. Parmi ces 9 résultats similaires :
  - un problème est signalé dans 7 cas sur 9
  - aucun problème n'apparaît pour les 2 autres cas (deux alignements avec un ARK)
- 1 cas sur 10 est un résultat différent, non plus avec 1 ARK mais avec 2 PPN, donc sort du périmètre de notre analyse.

Les trois méthodes du tableau 5 signalant un problème ont fait l'objet d'un contrôle qualité aléatoire de cinq alignements :

- 8 faux positifs sur 15 ont été identifiés
- 5 faux positifs sur 5 concernent la méthode « ISBN | Problèmes dans métadonnées Titre-Auteur-Date-Volume ».

- Les faux positifs sont presque systématiquement dus à une date différente. Dans quelques cas, on soupçonne que les deux manifestations soient de fait les mêmes, mais que les données aient été saisies de manière différente. Cela peut s'expliquer par une divergence des règles de catalogage, une divergence d'analyse de la part du catalogueur ou une erreur humaine.

Version 1.26

Les contrôles ont révélé exactement les mêmes résultats que pour la version 1.26 bêta.

Recommandations

1. Lancer l'étape bleue de Bibliostratus (génération des fichiers tabulaires par type de document) sur l'ensemble des données RERO, mais procéder à l'étape blanche (alignement avec BnF/SUDOC) uniquement pour les monographies textuelles.
2. Sélectionner les 3-4 méthodes ne signalant aucun problème et ayant le plus d'occurrences, puis en évaluer au moins 10 alignements.
3. Si les évaluations sont positives, intégrer les résultats dans la base de production Virtua.