Wikibase pour FNE : fin de partie (rapport technique)

Abes – BnF, mai 2023

Table des matières

Le projet Pilote FNE a été lancé en novembre 2022 avec comme objectif fonctionnel de donner à voir un lot important de données modélisées conformément au code RDA-FR dans un outil susceptible d'être retenu pour devenir la base de coproduction du FNE. Il a été mené selon la méthode agile, en collaboration étroite entre la Bibliothèque nationale de France et l'Abes.

La première phase de ce projet, d'une durée de 6 mois, a eu pour but de tester la solution Wikibase en faisant suite aux conclusions du PoC Wikibase de 2019. Cette phase a été découpée en 5 sprints au cours desquels la solution Wikibase a été explorée :

- mise en place de l'infrastructure technique Wikibase
- opérations de chargement de données en masse pour vérifier leur reproductibilité
- tests de performance de Wikibase par rapport aux attendus du projet (vitesse de chargement en masse : 200 entités/seconde au minimum)
- tests sur la fusion des données alignées

Les données utilisées pour le test sont les données des notices d'autorité Personne produites en Unimarc dans CBS, socle technique de l'Abes, et actuellement visibles sur IdRef. Le dump complet rassemble 3 780 189 notices.

1. Formatage des données

Production automatique des données pour les tests de performance

La commande pour cette première phase du Pilote FNE a été de travailler en implémentant dans Wikibase une modélisation très sommaire.
Pour ce faire, dans un premier temps, les données en format Unimarc ont été chargées avec quelques propriétés jugées essentielles : nom et prénom, identifiants (ISNI, PPN, ark BnF), dates biographiques.
Le reste des données présentes dans les notices a été inséré automatiquement par une boucle générique transformant toutes les autres zones : une zone Unimarc = une propriété Wikibase pour faire masse et permettre d'avoir des tests de performance réalistes.

Construction d'un tableau de correspondance entre formats

En prévision du futur chargement du dump des données en Intermarc-NG (le nouveau format de catalogage de la BnF), un tableau de correspondance entre les propriétés RDA-FR des entités Personne et Identité publique de Personne, les zones Unimarc/A et les zones Intermarc-

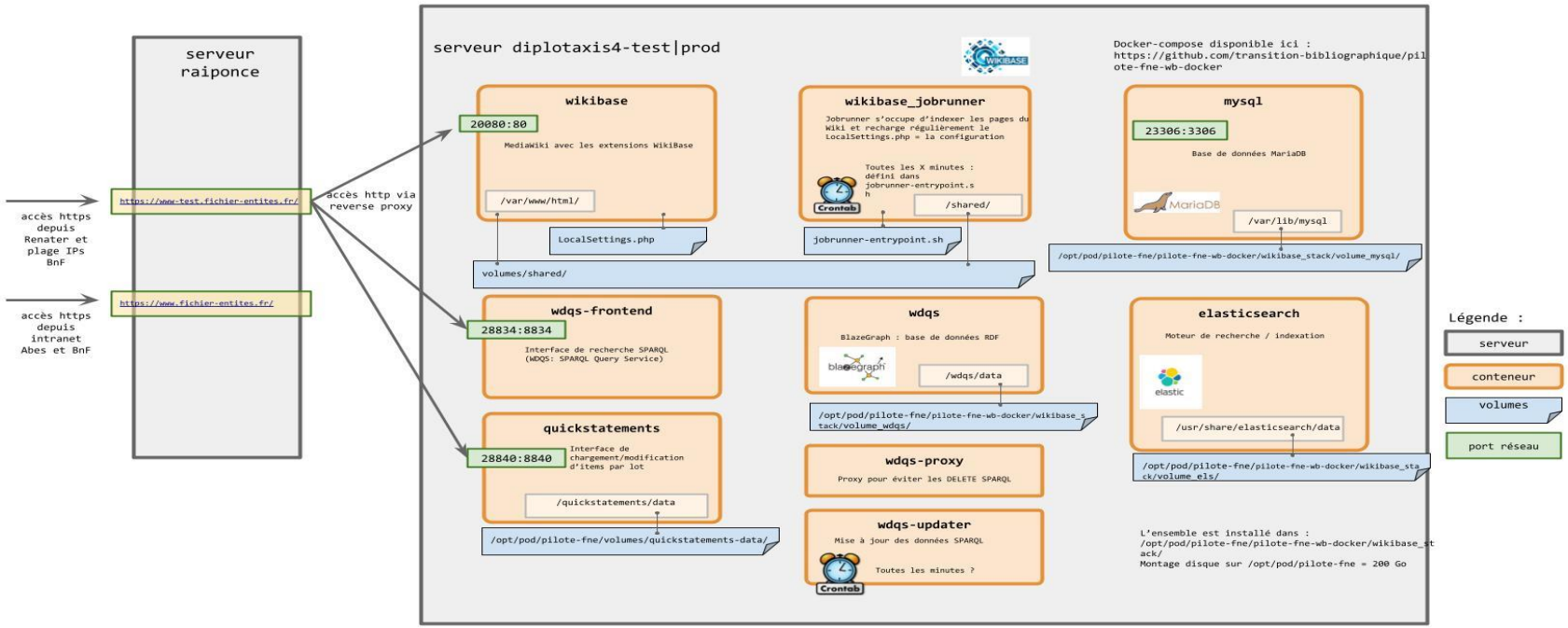
NG, a été réalisé, exploitable quel que soit le socle logiciel retenu. Ce mapping est à compléter avec les propriétés de l'ontologie RDA-FR, la partie sur ces deux entités étant désormais publiée.

Cependant afin d'affiner les premiers tests, de nouvelles propriétés (ou *statements* dans Wikibase) spécifiées par le travail de modélisation des experts données ont pu être déclarées. Ces *statements* ont été intégrés dans le programme de création de format (variantes de nom, points d'accès privilégiés, notes biographiques…). La faisabilité technique de relation entre entités (Personne et Identité publique) a également été testée.

2. Installation de Wikibase

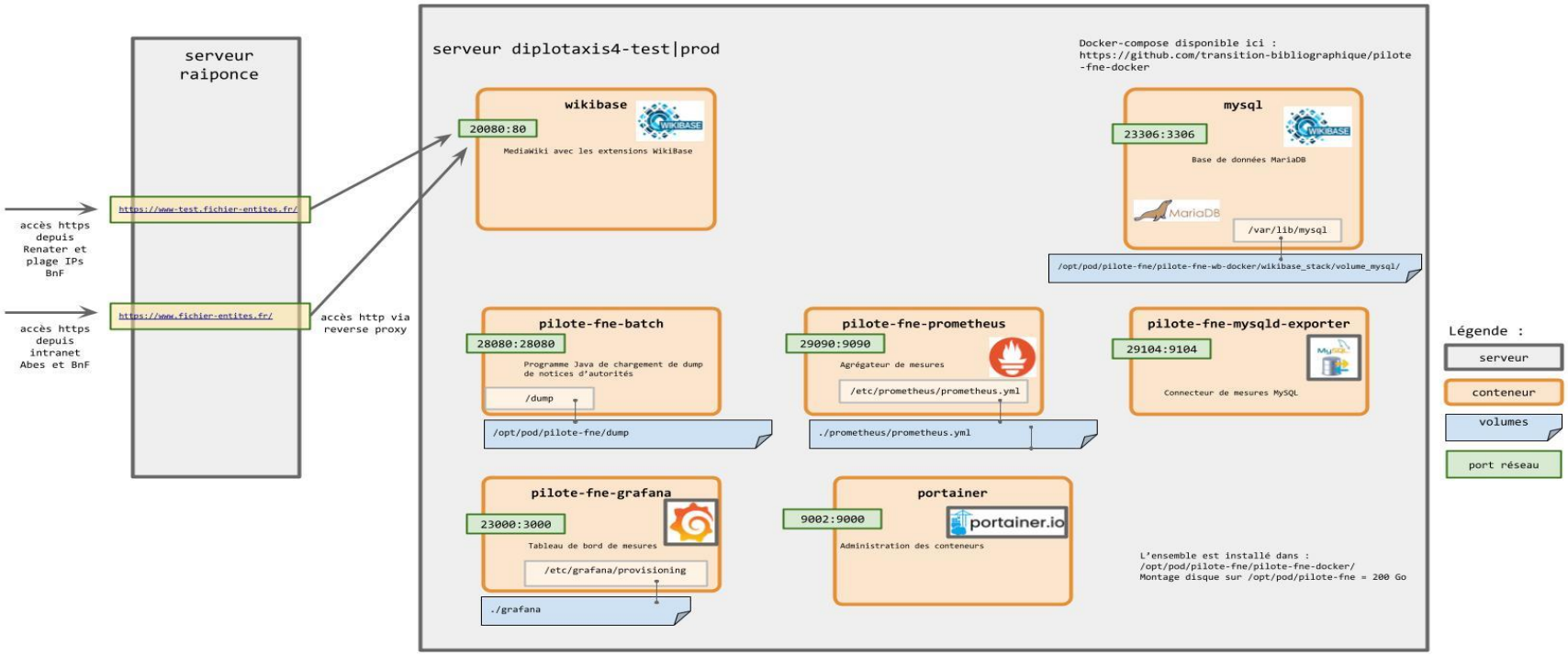
Des containers Docker Wikibase ont été mis en place sur 3 environnements conçus pour le projet : développement, test et production, sur les serveurs de l'Abes. Ces serveurs ont un stockage SSD, 24 Go de RAM et 10 CPU. Une ouverture en VPN a été mise en place pour le travail en commun avec la BnF.

Un dépôt Github de la configuration Wikibase utilisée a été créé (il se base sur la version WDME 9) : https://github.com/transition-bibliographique/pilote-fne-wb-docker.



Dès les premiers tests, il a été constaté que le paramétrage standard (my.cnf) de MariaDB utilisé par Wikibase n'était pas suffisant pour atteindre les performances attendues. Il a donc été procédé à des adaptations pour certains certains éléments (buffer, allocateur mémoire, compression, taille paquets, pages…).

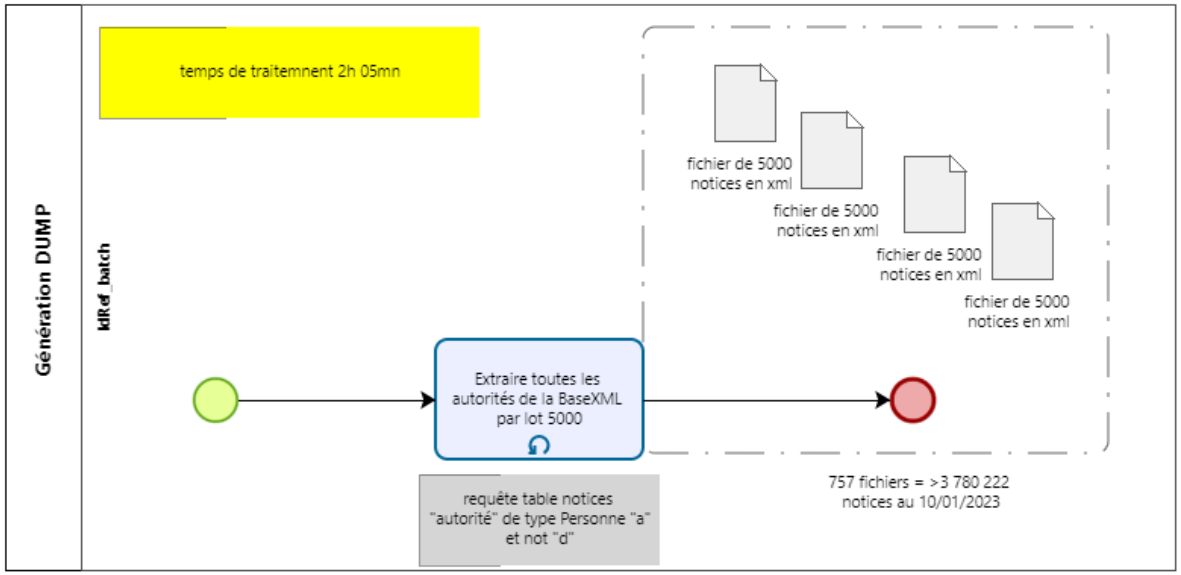
Un autre dépôt Github contenant la configuration du programme de chargement Java et des containers pour effectuer des mesures (Grafana et Prometheus) a été créé : https://github.com/transition-bibliographique/pilote-fne-docker



3. Description du programme de chargement

Génération du dump des données d'autorité en vue de son chargement dans Wikibase

Les données d'autorité Personne en format Unimarc produites dans le CBS sont actuellement stockées en Marc XML dans une base Oracle (base XML). C'est de cette base XML que les données ont été extraites.



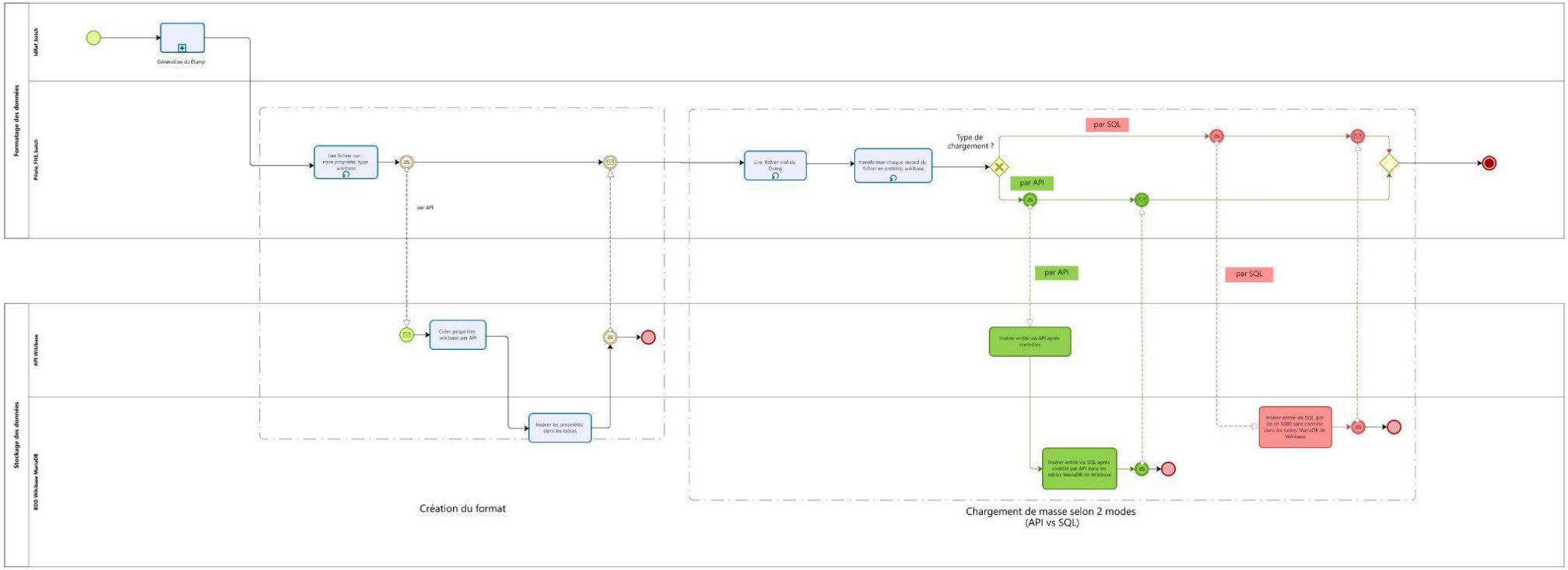
La première mesure a été le temps de génération du dump de ces notices Unimarc/XML du Sudoc. Le dump complet a été généré en 2 h 05.

Création du format dans Wikibase

Pour charger des données d'autorité dans une Wikibase, il faut au préalable créer les propriétés et les types d'entités qui serviront à décrire ces données, et ce, quel que soit le modèle de données choisi (cf. Construction d'un format dit "pivot"). Pour cela, une méthode Java, utilisant l'API Wikibase, a été développée :
https://github.com/transition-bibliographique/pilote-fne/blob/main/src/main/java/fr/fne/batch/tasklet/FormatTasklet.java

4. Insertion des données

Schéma BPMN (Business Process Model and Notation) du processus de chargement :



A. Chargement en utilisant l'API de Wikibase (flux vert dans le schéma ci-dessus)

Une première approche a été d'utiliser le WDTK-Toolkit Wikibase, API Java, qui permet d'interagir avec une instance Wikibase : https://github.com/Wikidata/Wikidata-Toolkit.
Il se trouve que ce projet a initialement été conçu pour être utilisé avec WikiData. Le nombre de requêtes vers l'API Wikibase a été volontairement limité dès l'origine, pour ne pas le surcharger.
Pour contourner ces limites, l'équipe a décidé d'utiliser à la place la librairie Spring Boot RestTemplate, qui permet d'interagir avec l'API Wikibase par requêtes classiques HTTP REST (Json). Le code source est disponible ici :
https://github.com/transition-bibliographique/pilote-fne/blob/main/src/main/java/fr/fne/batch/writer/ItemDocumentWriterAPI.java
L'utilisation de l'API Wikibase s'est montrée lente, même en utilisant l'asynchrone* ou le multithreading** : des temps de chargement de l'ordre de **17 entités chargées par seconde** ont été mesurés.
*Asynchrone : requêtage sans attente de la réponse
**Multithread : plusieurs processus fonctionnant en parallèle

B. Chargement par insertion directe SQL (flux rouge dans le schéma ci-dessus)

Au vu des temps de chargement par l'API, et du temps de chargement attendu pour le projet (au moins 200 entités/seconde), l'équipe a décidé de tester une insertion des données directement dans la base de données MariaDB de Wikibase, en SQL.
Pour cela, l'équipe s'est largement inspirée et a repris des portions de code de 2 projets :

- Un projet Java d'insertion, fonctionnant avec une ancienne version de Wikibase : https://github.com/jze/wikibase-insert
- Un projet Python d'insertion, fonctionnant sur une version actuelle de Wikibase : https://github.com/UB-Mannheim/RaiseWikibase/

Ces codes d'insertion de données SQL ont l'avantage de la rapidité, mais ils ont aussi des défauts :

- Ils doivent être adaptés à la structure de la base de données utilisée par Wikibase. Or cette structure est sujette à évoluer dans de prochaines versions, imposant l'évolution de ces codes.
- Les contrôles inhérents à l'API de Wikibase sur les données insérées ne sont de fait pas réalisés, ce qui est susceptible de provoquer des dysfonctionnements lors de l'affichage ou l'édition des entités.
- Le code qui insère les données en SQL dans une quinzaine de tables peut comporter des problèmes non décelés au premier abord par manque d'expertise de Wikibase. Ce code source développé par l'équipe est disponible ici :
  https://github.com/transition-bibliographique/pilote-fne/blob/main/src/main/java/fr/fne/batch/writer/ItemDocumentWriterSQL.java

Le temps de chargement, avec cette méthode, a atteint **128 entités chargées par seconde**, soit une nette amélioration par rapport aux tests menés avec l'API Wikibase, mais loin encore des exigences requises pour le Fichier national d'entités.

Conclusion

Les besoins fonctionnels du FNE requièrent de pouvoir opérer des modifications de masse non seulement au moment du chargement initial des données dans la base, mais également plus tard, pendant tout le cycle de vie de l'application, avec un niveau de performance équivalent de 200 entités/secondes.
Si le mode d'insertion par SQL peut à la rigueur être envisagé lors de la migration des données, il ne saurait être question de maintenir son utilisation dans le temps alors qu'il présente des risques quant à la maintenabilité de l'application. Il est du reste déconseillé par la communauté au regard des effets de bord sur d'autres fonctionnalités de Wikibase. Dans le cas de modification lors d'une montée de version de Wikibase (en particulier sur son mécanisme de persistance des données), il y aurait ainsi de très grands risques de devoir modifier/re-développer certains aspects de l'insertion.
Ce programme n'assure donc pas la soutenabilité ni la maintenabilité informatique nécessaires au projet[1].

Aucune de ces deux méthodes ne permettant l'insertion ni la modification en masse rapide des données dans wikibase, il a donc été décidé de renoncer à cette infrastructure logicielle comme socle du Fichier national d'Entités.

[1] https://www.wikibase.consulting/fast-bulk-import-into-wikibase/
https://phabricator.wikimedia.org/T287164