# Implantation de la procédure pour la vérification des liens brisés

## Problématique

Mise en contexte

Avec l’arrivée du projet OSDP, des demandes grandissantes en terme de qualité de métadonnées sont reçues par l’équipe de la PGF. L’une d’elles concerne la fonctionnalité des URL inscrites dans les ressources des jeux de données. De plus en plus de jeux de données sont ajoutées dans la PGF mais aucun processus n’est en place pour assurer le cycle d’opération et de maintenance des métadonnées à travers le temps. Ainsi, les URL peuvent être valides lors de l’ajout des métadonnées dans la FGP mais rien nous permet d’assurer aux utilisateurs que les URL restent valides après cela.

Besoin

Au niveau de l’équipe de données de la FGP, nous avons besoin d’un outil automatique pour vérifier la validité des URL définies dans la section ressources dans les jeux de données contenues dans le catalogue de la FGP. L’équipe des données n’ayant pas autorité sur le contenu des métadonnées, le besoin est principalement ciblé sur l’identification des métadonnées en problème et la notification des problèmes à l’autorité concernée.

Out of Scope

* Correction des erreurs détectées par la solution
* Les URL validées sont uniquement celles présentes dans les ressources. La solution ne valide pas les URL pouvant se trouver ailleurs dans les métadonnées (contact, description, ...)
* Aucune validation au niveau de la synchronisation avec le RCS ne sera faite.
* Validation de conformité des métadonnées envers le HNAP.

## Analyse des besoins

Objectif

Mettre en place une solution validant automatiquement les URL et identifiant les jeux de données comportant des URL non fonctionnelles. La solution en place comportera le module de validation préalablement présenté ainsi qu’on module de gestion des erreurs et de notification par courriel.

Les étapes suivantes montrent une vue haut niveau de la séquence d’opération

1. Lecture des métadonnées et extraction des URL (données, service Web, lien html, etc) à partir des métadonnées
2. Identification du type d’URL et validation de manière appropriée
3. Gestion des codes de retour et identification des problèmes au besoin
4. Mise à jour du registre des erreurs identifiées
5. Envoie des notifications à l’autorité concernée

## Analyse détaillée

Sources de données

La solution développée aura comme source de données un point d’accès CSW diffusant des métadonnées sous le format HNAP. Pour l’instant, étant donné le firewall entre le GCNET et l’environnement AWS où est hébergé le FME Server de la FGP, la solution se connectera sur le catalogue externe de la FGP. Si un besoin se fait sentir pour effectuer une validation des métadonnées contenues dans le catalogue interne, il faudra déployer la solution vers un FME Server à l’intérieur du GCNET ou avoir un proxy à travers le firewall permettant l’accès au catalogue interne de la FGP à l’extérieur du GCNET.

Gestion des erreurs

AJOUTER TEXTE EXPLIQUANT QUE L’ON VEUT EFFECTUER UNE GESTION DES ERREURS AFIN DE NE PAS SURCHARGER LES CONTRIBUTEURS DE NOTIFICATIONS. AJOUTER LE MODÈLE PRÉSENTÉ À L’ÉTAPE 5. (ATTENTION, IL N’EST PLUS À JOUR)

Envoie des notifications

AJOUTER TEXTE EXPLIQUANT QU’ON VA ENVOYER LES NOTIFICATIONS VIA DES COURRIELS EN UTILISANT L’INFORMATION CONTENUE DANS LES MÉTADONNÉES Courriel :

* Nous statuerons d’un contact principal pour chaque représentant PGF par minitère, province ou territoire. Ces informations sont présentement accessibles sur la PGF interne et il faudra trouver un moyen de les obtenir lors de nos requêtes au niveau externe. La meilleure façon de trouver automatiquement ces informations (nom et adresse courriel) et est encore à être défini dans nos tests.
* Le Format et contenu du message seront standardisés et nous y mettrons quelques informations permettant d’identifier facilement quel jeu de donnée demande des correctifs (UUID, titre)
* L’information fournie par FME concernant le lien en problème et le type d’erreur (code) seront également ajouté au courriel afin que les contributeurs sachent ce qui n’a pas fonctionné.

Fréquence d’exécution

Choix technologique

Plusieurs technologies s’offrent à nous pour mettre en place la solution. Étant donné que la PGF dispose d’une infrastructure FME Server et que cette technologie nous offre toutes les capacités nécessaires autant au niveau de la validation que de la notification des erreurs, nous avons choisi cette technologie pour l’implantation de notre solution.

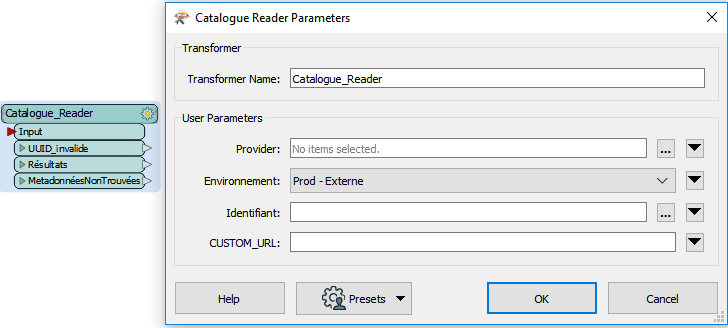
Au niveau de la gestion des erreurs et le suivi dans le temps des erreurs, nous préconisons une base de données SQL Lite. Le format de fichier offre des capacités similaires aux fonctionnalités de bases des SGDB tel que PostGreSQL ou Oracle. Pour les besoins actuels, ce type de base de données amplement suffisant.

Note : Étant donné l’utilisation de FME, il sera très facile de changer la BD SQL-Lite si désiré durant le cycle d’opération et de maintenance de la solution mise en place.

## Conception détaillée

Lecture de la données source

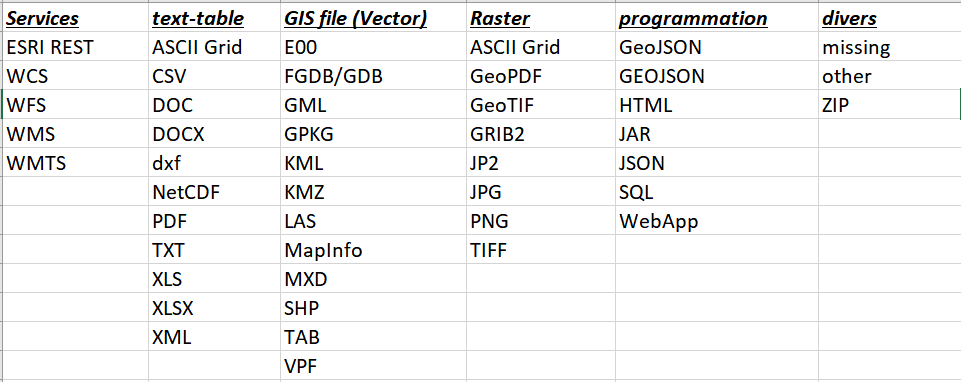
Utilisation du logiciel FME et du transformer CatalogueReader afin de lire le contenu du catalogue de notre choix. Comme mentionné, nous choisirons le catalogue externe au départ afin de questionner le contenu qui est accessible pour tout le monde et qui est utilisé par la plateforme de recherche de OSDP. Lors de vérification, nous pourrons utiliser les options du CatalogueReader permettant de valider le contenu d’un UUID ou d’une liste d’UUID.



Nous aurons ainsi accès au contenu du catalogue en format XML (HNAP) et nous pourrons utiliser les balises qui nous intéressent pour conserver les informations nécessaires à nos travaux.

Identification des URL et des formats

La section « ressources » des métadonnées sera celle dans laquelle nous obtiendrons le plus grand nombre d’informations nous intéressant. Le nom de la ressource sera utilisé pour l’identification du lien, le type de ressource ainsi que le format permettront d’orienter le type d’appel que nous devons faire pour obtenir une réponse du serveur quant à la validité du lien. Voici une liste des formats de fichiers regroupés par type au moment de notre analyse.



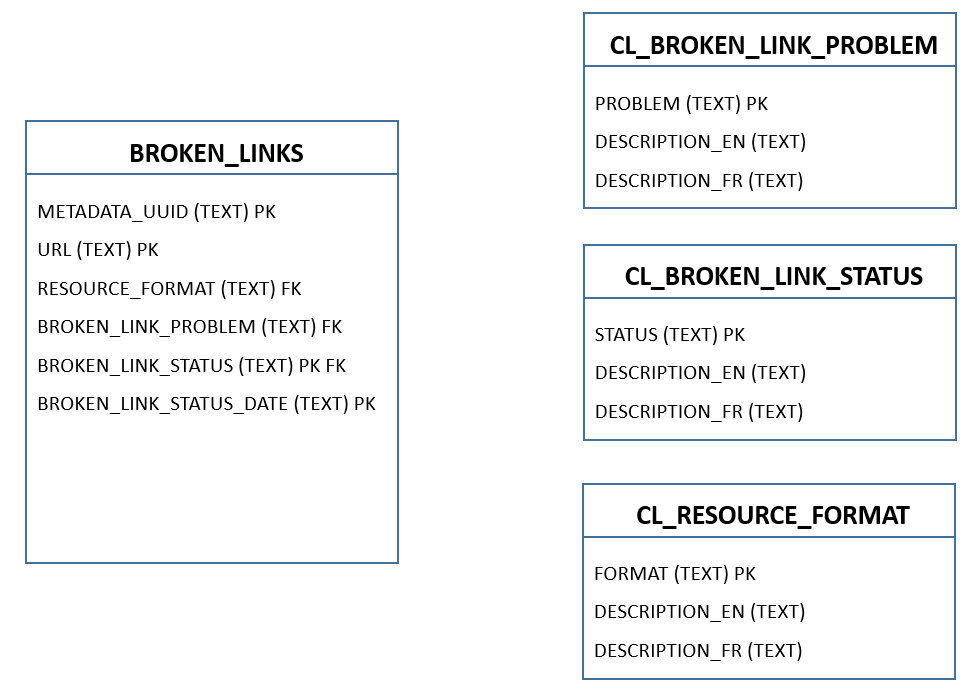
Validation des URL en fonction du type de ressources

Mise à jour du registre des problèmes

Envoie de notification aux contributeurs

~~La plupart des jeux de données comporte plusieurs ressources et nous devrons faire la vérification de tous ceux-ci et avoir des précisions dans nos communications mentionnant quels liens sont problématiques parmi tous ceux présents.~~

~~Étape 5 :~~

~~~~