Processus de transformation des données en RDF

Analyse des best practices

Mars 2015

Sommaire

1 Workflows utilisés en bibliothèques

Plusieurs bibliothèques ont fait le pas du web sémantique et ont documenté leur méthodologie et leur processus de transformation des données bibliographiques. Ce chapitre décrit les processus des quatre bibliothèques suivantes :

1. La Biblioteca nacional de España
2. Le réseau Libris en Suède
3. Le réseau RERO en Suisse
4. Le réseau HBZ en Allemagne

Une synthèse de ces workflows est ensuite réalisée sous la forme d'un processus englobant tous les autres.

1.1 Biblioteca Nacional de España (BNE)

La Bibliothèque nationale espagnole a développé, dans le cadre de son projet Linked Data, un outil de transformation des données MARC21 en RDF destiné aux bibliothécaires lambda. Cet outil, nommé MARiMbA (détails au chapitre 2.1.1), permet de gérer une partie du processus de conversion expliqué ci-dessous.

La méthode utilisée pour le projet espagnol se décompose en sept étapes (1) :

1. Spécifications

- identifier et analyser les données de base
- choisir les URIs (y compris choix des vocabulaires/ontologies)
- définir la licence et les informations de provenance

2. Curation des données (en tant que tâche parallèle)

- Nettoyage des données d'origine et des données RDF. Cette étape intervient à plusieurs moment du processus)
  - identification des erreurs dans les données. Ex : utilisation interdite d'un sous-champ.
  - création de rapport automatiser. MARiMbA fournit automatiquement des rapports sur les erreurs découvertes dans les notices bibliographiques et d'autorités.
  - correction des erreurs

3. Modélisation

- analyser et choisir les ontologies du domaine
- développer son propre modèle
- ontologie pour les données de provenance

4. Génération

- sélectionner, développer, compléter des technologies de production de RDF
- établir les correspondances entre le format d'origine et RDF
  - Création du **mapping no 1 (classification)** pour attribuer des classes (Personne, Collectivité, Œuvre, Expression) aux entités présentes dans les registres, selon les

ontologies choisies, et génération de ressources RDF pour ces entités
Résultat : 1 feuille de correspondances

- Création d'un **mapping no 2 (annotations)** pour décrire ces ressources RDF au moyen de propriétés, selon les données présentes dans les champs et sous-champs du format MARC21.
  Résultat : 1 feuille de correspondances par classe identifiée
- Création d'un **mapping no 3 (relations)** pour générer des relations entre les ressources RDF nouvellement créées.
  Résultat : 1 fichier par type de relation (Person-Work/Person-Person/Work-Expression/...) contenant tous les identifiants de datos.bne.es constituant ces relations. Les relations sont identifiées au moyens de comparaison de chaînes de caractères entre les diverses entités.

- transformer les données. L'outil MARiMbA effectue la transformation des données, au moyen des inputs suivants :
  - les données sources MARC21
  - les URIs des vocabulaires utilisées (FRBR, etc.) + les spécifications pour la formes des URIs de datos.bne.es
  - les mappings 1, 2 et 3 établis par les bibliothécaires spécialistes

  Ajout (génération) des données de provenance (grâce aux Named graphs)
  - date et auteur de la requête des données : prv:retrievedBy et prv:completedAt
  - fournisseur et créateur des données : dcterms:publisher et dcterms:creator
  - licence : dcterms:license

5. Liens

- sélectionner les jeux de données que l'on veut connecter, créer les liens, valider les liens. Possible utilisation de Silk (détails au chapitre 2.3.2) pour générer ces liens automatiquement (mais à la BNE, tous les liens ont été créés grâce à VIAF).

6. Publication

- publication sur le web
  - via un SPARQL-endpoint (utilisation de Virtuoso[1] comme Triple store)
  - via une interface front-end de négociation de contenu utilisant Pubby (détails au chapitre 2.3.1)
  - via une API utilisant Puelia[2]
- publication de métadonnées sur le jeu de données (en utilisant une description voID)
- promotion (grâce à l'inscription du jeu de données sur http://datahub.io/)

7. Exploitation

- création d'applications

Ce processus, relativement long, s'est déroulé de manière itérative :

- 1ère itération avec toutes les notices (y compris d'autorité) liées d'une quelconque manière à Miguel de Cervantes :
- 2e itération : l'ensemble des notices de la BNE, à l'exception de certains types.
- 3e itération : tout.

1 Disponible à cette adresse : http://virtuoso.openlinksw.com/ (consulté le 18 mars 2015)
2 Disponible à cette adresse : https://code.google.com/p/puelia-php/ (consulté le 18 mars 2015)

1.2 Libris

Libris est le catalogue collectif suédois, géré par la Bibliothèque royale de Suède. Ce fut le premier acteur à publier l'ensemble des données d'un catalogue de bibliothèque en RDF, via une application web transformant les notices à la volée. Le format d'export est déterminé selon la négociation de contenu.

En résumé, le processus de transformation suivi chez Libris se déroule en quatre grandes étapes (2) :

1. Extraction des notices et de leurs relations du SIGB
   Utilisation d'une application web exportant les notices en MARC/XML et leurs relations en N-Triples, lors d'une requête par URL.
2. Choix d'un modèle d'URI
3. Etablissement des correspondances avec RDF
   Etape incluant le choix des ontologies à utiliser
   Seules deux relations existent entre notices bibliographiques et notices d'autorités : *dc:subject* et *dc:creator*. Entre les notices d'autorité sujets, seuls *skos:broader* et *skos:narrower* ont été utilisés.
4. Lien avec des ressources externes
5. Implémentation de la négociation de contenu et de la publication (SPARQL-endpoint)
   Le Triple store utilisé chez Libris est Sesame.

1.3 RERO

Le projet de publication des métadonnées de RERO en RDF a suivi les étapes suivantes (3, p. 19) :

1. Revue de la littérature
2. Analyse des données
3. Modélisation : choix d'un type de modèle, identification des types d'entités, ajout de données de provenance, choix des IRIs
4. Mapping : choix des ontologies, rédaction des règles de conversion
5. Transformation
6. Ajout de liens externes
7. Contrôle qualité
8. Publication

1.4 LOBID

LOBID (Linking Open Bibliographic Data) est le projet du Hochschulbibliothekszentrum (HBZ) Nordrhein-Westfalen pour la publication des données bibliographiques en LOD.

Six phases ont été planifiées pour la transformation des notices bibliographiques (4) :

1. Transformation des données MAB (une dérivation du format MARC utilisée en Allemagne) en une sérialisation RDF (RDFMAB).
2. Étude des ontologies pertinentes, notamment BIBO et RDA.
3. Établissement des correspondances entre RDFMAB et BIBO.
4. Conversion des données RDFMAB en données RDF utilisant BIBO.
5. Établissement des correspondances entre RDFMAB et RDA.
6. Conversion données RDFMAB (+BIBO) en données RDF utilisant RDA.

HBZ a réutilisé Metafacture, une technologie développée dans le cadre du projet Culturegraph, pour transformer ses données.

1.4.1 Les organisations LOBID

LOBID n'a pas publié que les notices bibliographiques en LOD. Le projet a également abouti à la publication de données sur les organisations (les bibliothèques) possédant les ressources, afin de les intégrer pleinement au web des données. Ainsi une ontologie décrivant les bibliothèques, leurs services et leurs collections a été développée.

Afin de ne pas devoir mettre à jour la base de données de toutes les institutions manuellement, les ingénieurs de LOBID ont établi la procédure suivante :

- Formations aux diverses bibliothèques afin qu'elles créent, sur leurs sites web respectifs, une description d'elles-mêmes en RDFa
- Mise en place à LOBID d'un système moissonnant automatiquement ces descriptions, afin qu'elles soient le plus actuelles possibles. Ainsi, les données ne doivent être mises à jour qu'en un seul endroit.

1.5 Synthèse

Le W3C, et plus précisément son Government Linked Data Working Group, s'est penché de manière approfondie sur la création de meilleures pratiques pour la publication de Linked Open Data (5). Il analyse et synthétise les workflows utilisés lors de divers projets de transformation de données, tels que celui de la Bibliothèque nationale d'Espagne. Si l'on additionne cette analyse avec celle de ce chapitre, il en résulte le workflow suivant, censé être exhaustif :

1. Préparer les acteurs (responsables, organisations)
2. Préparer les données de base
    - Sélectionner le jeu de données
    - Extraire les données
    - Analyser les données
    - Nettoyer/améliorer les données
3. Préparer la conversion
    - Modéliser le jeu de données (y compris données de provenance)
    - Choisir des IRIs
    - Choisir les vocabulaires ou en créer un
    - Établir les correspondances
4. Transformer les données en RDF
5. Relier les données

6. Publier les données
    - Choisir une licence appropriée
    - Mettre à disposition en ligne
    - Annoncer la nouvelle
7. Maintenir les données
    - Mettre à jour les données
    - Faire évoluer le modèle
8. Exploiter les données

2 Technologies utilisées

2.1 Outils de transformation de données

Deux solutions sont possibles pour convertir ses données en RDF:

1. Réutiliser un outil existant

   L'avantage est d'utiliser un outil qui deviendra peut-être un standard compris par une communauté de personnes. L'existence d'une documentation permet également de pérenniser l'utilisation d'un outil au sein de sa propre institution, notamment en cas de changement de personnel. Le désavantage est le temps d'apprentissage de l'outil et son adaptation aux données locales.

2. Développer son propre outil

   L'avantage est la rapidité de formulation et d'adaptation des règles de transformation, car elles sont définies dans un langage connu. Ceci peut au final être moins coûteux qu'un développement maison.

Dans cette section sont décrits les outils qui ont été créés principalement par la communauté des bibliothèques afin d'être réutilisés pour une transformation de leurs données.

2.1.1 MARiMbA (MARC mappings and rdf generator)[3]

Outil permettant de gérer le processus complet de génération de RDF à partir de MARC21.

- Disponible sur demande pour des essais
- Permet d'utiliser les ontologies de son choix
- Permet la création de mapping au moyen de feuilles de calcul (accessible sans connaissances préalables de technologies complexes)
- Crée des analyses de données pour faciliter l'établissement du mapping

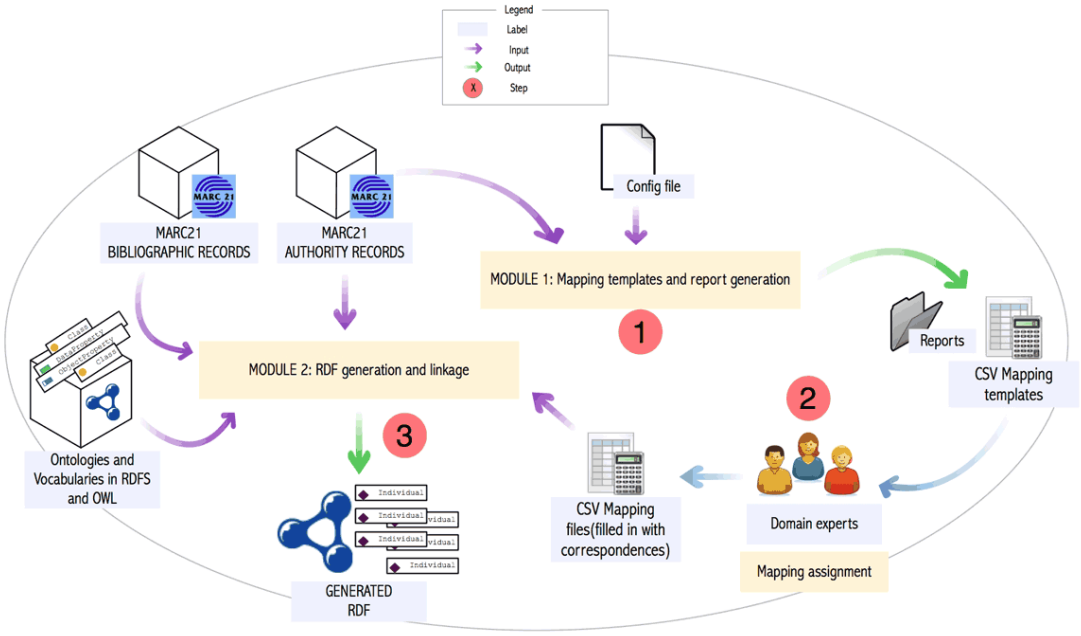
La particularité de MARiMbA est qu'il ne nécessite pas de connaissances spécifiques en programmation. Le processus géré le logiciel est schématisé dans l'illustration 1. Certaines tâches doivent néanmoins être effectuées par des humains ; MARiMbA les guide et les aide (6) :

1. Analyse des données MARC.
2. Classification des données : attribution d'une classe RDF pour des types de notice selon une combinaison de zones et de valeurs présentes.
   Ceci génère le mapping no 1 utilisé par MARiMbA.

3 Disponible à cette adresse : http://marimba4lib.com/ (consulté le 18 mars 2015)

3. Description des notices MARC, selon chaque classification décelée au point précédent, au moyen d'équivalences RDF.
   Ceci correspond le mapping no 2.
4. Génération ou extraction en fichiers des relations entre les notices (selon les classes identifiées)

Illustration 1: Fonctionnement de MARiMbA



La BNE a développé avec MARiMbA son propre outil de transformation car les technologies existantes (1) :

- étaient conçues pour être utilisées par des développeurs
- n'étaient pas appropriées pour le modèle FRBR
- n'étaient pas conçues pour gérer des notices bibliographiques et d'autorité en même temps.

2.1.2 Metafacture (projet Culturegraph)

Culturegraph regroupe un ensemble de projets traitant de l'analyse et l'interconnexion des données culturelles (7). L'un des aboutissements de Culturegraph consiste en un logiciel de traitement de données semi-structurées, et spécifiquement les données de bibliothèques, nommé Metafacture.

> *« Software written to perform transformations is often coded from scratch for each and every individual case; despite great potential for component reuse. » (8)*

L'idée générale de Metafacture, et en particulier de son langage Metamorph, est de partager un langage de transformation afin qu'il puisse être réutilisé par l'ensemble d'une communauté (bibliothèques, musées, etc.).

Metafacture est composé de trois parties :

- Framework : contient la base du logiciel (interface et classes abstraites), nécessaire si l'on souhaite utiliser Metafacture comme bibliothèque Java.
- Metamorph : bibliothèque Java, incluant un langage de programmation basé sur XML, conçu pour faciliter le traitement des métadonnées. Plus d'info : (8).
- Flux : gère les processus de transformation

Détails: https://github.com/culturegraph/metafacture-core/wiki

2.1.3 Catmandu (projet LibreCat)

Le projet LibreCat est une collaboration entre les universités de Gand (Belgique), de Bielefeld (Allemagne) et de Lund (Suède). Le but est de mettre à disposition des outils de programmation permettant de construire des bibliothèques digitales et des services de recherche adaptés aux besoins de chacun (9).

Pour ce faire, un outil de traitement des données Open Source nommé Catmandu a été créé. Celui-ci permet notamment, au moyen de commandes spécifiques, d'extraire et de transformer des données – en formats MARC, MAB et RDF entre autres – et de les exporter dans un moteur de recherche. Le language Fix a été développé pour effectuer des transformations d'un modèle de données à un autre.

2.1.4 OpenRefine (RDF extension)

OpenRefine[4] est un logiciel Open Source permettant de nettoyer et améliorer les données ainsi que de les transformer. Il possède une interface utilisateur accessible par un navigateur web et traite les données sous forme de tableaux, comme un tableur, avec des fonctionnalités complémentaires.

L'extension RDF[5] d'OpenRefine permet de générer, sur la base de données structurées en tables (lignes/colonnes), des données RDF. Pour ce faire, il faut paramétrer soi-même un squelette, établissant les correspondances entre le modèle de données sous forme de table et le modèle sous forme de triplets RDF.

2.1.5 RDF Mapping Language (RML)[6]

Cet outil permet de transformer des données structurées en RDF, par exemple depuis les formats CSV, JSON et XML. Les règles de transformation sont programmées au moyen du format Turtle.

Ce logiciel n'est pas conçu spécifiquement pour le format MARC utilisé en bibliothèque, et est limité dans son traitement des données.

2.2 Triplestores

Un triplestore est une base de données permettant d'enregistrer d'enregistrer des informations sous forme de triplets uniquement. Le triplet est la structure de base du modèle RDF ; il se compose de trois parties : un sujet, un prédicat, un objet. Les données d'un triplestore sont ensuite interrogées au moyen du langage de requête SPARQL.

Le site web http://www.w3.org/wiki/LargeTripleStores met à disposition une liste de triplestores, avec indication de leurs capacités, notamment la quantité de triplets supportée. Ces quantités se mesurent en millions (10^6) et milliards (10^9), voire en billions (10^{12}).

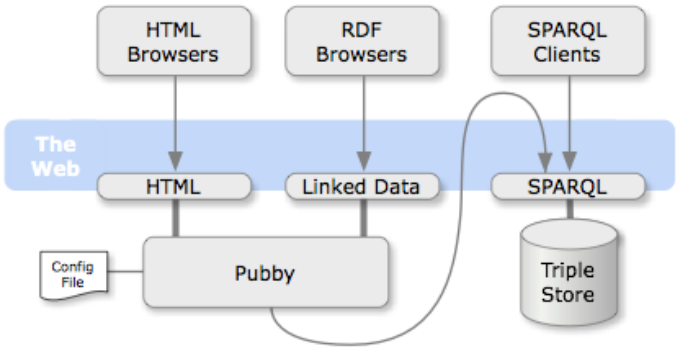
4 Disponible à cette adresse : http://openrefine.org/ (consulté le 18 mars 2015)
5 Disponible à cette adresse : http://refine.deri.ie/ (consulté le 5 mars 2015)
6 Disponible à cette adresse : http://rml.io/index.html (consulté le 5 mars 2015)

2.3 Autres

2.3.1 Pubby

Logiciel fournissant une interface Linked Data, qui permet l'accès aux données RDF du Triple store pour les navigateurs sémantiques et les navigateurs traditionnels (10). Ceci évite de devoir passer par les requêtes complexes du SPARQL-Endpoint. Son fonctionnement est résumé dans l'illustration 2.

Illustration 2: Fonctionnement de Pubby



2.3.2 Silk[7]

Cet outil a été développé principalement par l'Université de Mannheim dans le but de créer des liens automatiquement entre un jeu de données et un référentiel du web sémantique. Le logiciel peut être utilisé soit au moyen d'une interface graphique ou par commandes.

Il contient notamment un éditeur de règles de liaison, permettant de choisir les éléments à relier et de définir des opérateurs de comparaison, de transformation et d'agrégation. Il utilise les points d'accès SPARQL pour accéder aux données externes.

Silk a été utilisé par la BNE.

7 Disponible à cette adresse : http://wifo5-03.informatik.uni-mannheim.de/bizer/silk/ (consulté le 5 mars 2015)

3 Bibliographie

1. VILA-SUERO, Daniel et GÓMEZ-PÉREZ, Asunción. datos.bne.es and MARiMbA: an insight into Library Linked Data. *Library hi tech*. 2013. Vol. 31, n° 4, pp. 575-601. DOI 10.1108/LHT-03-2013-0031.

2. MALMSTEN, Martin. Exposing library data as Linked Data. *Satellite meetings IFLA 2009: Emerging trends in technology, : libraries between web 2.0, semantic web and search technology* [en ligne]. Florence. 19 août 2009. [Consulté le 18 mars 2015]. Disponible à l'adresse : http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.181.860&rep=rep1&type=pdf

3. PRONGUÉ, Nicolas. *Modélisation et transformation des métadonnées de RERO en Linked Open Data*. Genève : Haute école de Gestion, 2014.

4. Offene HBZ-Titeldaten als Linked Data. *Wiki des Hochschulbibliothekszentrums des Landes Nordrhein-Westfalen* [en ligne]. 11 novembre 2011. [Consulté le 18 mars 2015]. Disponible à l'adresse : https://wiki1.hbz-nrw.de/display/SEM/Offene+hbz-Titeldaten+als+Linked+Data

5. GOVERNMENT LINKED DATA WORKING GROUP. GLD Life cycle. *World Wide Web Consortium - Government Linked Data Working Group wiki* [en ligne]. 29 août 2012. [Consulté le 18 mars 2015]. Disponible à l'adresse : http://www.w3.org/2011/gld/wiki/GLD_Life_cycle

6. BIBLIOTECA NACIONAL DE ESPAÑA. Datos enlazados en la BNE. *Biblioteca Nacional de España* [en ligne]. [Consulté le 18 mars 2015]. Disponible à l'adresse : http://www.bne.es/es/Inicio/Perfiles/Bibliotecarios/DatosEnlazados/index.html

7. BÖHME, Christoph, CHRISTOPH, Pascal, STEEG, Fabian, PFEFFER, Magnus et GEIPEL, Markus Michael. Culturegraph: Metafacture-Core. *GitHub* [en ligne]. février 2015. [Consulté le 18 mars 2015]. Disponible à l'adresse : https://github.com/culturegraph/metafacture-core

8. GEIPEL, Markus Michael. *Metamorph user guide* [en ligne]. 23 novembre 2012. Culturegraph. [Consulté le 18 mars 2015]. Disponible à l'adresse : http://sourceforge.net/p/culturegraph/code/1691/tree/metamorph/trunk/docs/user_guide/metamorph.pdf?format=raw

9. *LibreCat* [en ligne]. 2015. [Consulté le 18 mars 2015]. Disponible à l'adresse : http://librecat.org/

10. CYGANIAK, Richard et BIZER, Christian. Pubby: a Linked Data frontend for SPARQL endpoints. *Fakultät für Wirtschaftsinformatik & Wirtschaftsmathematik, Universität Mannheim* [en ligne]. [Consulté le 18 mars 2015]. Disponible à l'adresse : http://wifo5-03.informatik.uni-mannheim.de/pubby/