abes
agence bibliographique de l'enseignement supérieur

Etude sur la faisabilité et le positionnement d'un hub de métadonnées ABES

Rapport final

ABES – juillet 2013

Sommaire

Introduction

L'idée d'un hub

Le nouveau projet d'établissement (2012-2015) introduit la notion et l'ambition d'un hub de métadonnées :

> ***Dans le cadre du hub de métadonnées, l'ABES offrira le service suivant à tous les établissements: redistribution dans n'importe quel format des métadonnées enrichies récupérées dans n'importe quel format auprès des éditeurs. Outre la conversion de format, l'ABES apportera une plus-value aux métadonnées en termes de structuration, de richesse et d'interconnexion de l'information avec d'autres bases.***[1]

Dès sa conception et sa mise en service, le Sudoc a lui-même fonctionné comme un hub de métadonnées bibliographiques : le défi était d'agréger en une seule base des sources de données hétérogènes et de les redistribuer sous différentes formes (notices UNIMARC et MARC21, produits bibliographiques).

C'est précisément en gardant à l'esprit cet objectif partagé par le Sudoc et le hub de métadonnées qu'on peut mettre en évidence les spécificités de ce dernier :

- Les sources de données à agréger et à traiter sont de plus en plus hétérogènes : au-delà des données MARC issues principalement de catalogues de bibliothèque, les données à traiter proviendront surtout de divers éditeurs et prendront des formes très variées.
- Les sources de données à agréger et à traiter sont de plus en plus massives : au-delà des millions de notices de monographie et de périodique, il s'agit de traiter des millions d'articles et de chapitres de monographie (papier ou électronique).
- Le hub de métadonnées n'est pas une infrastructure de catalogage en réseau : il devra avant tout miser sur des traitements automatiques, pilotés de manière centralisée par l'ABES. Puisque les traitements automatiques ne pourront jamais suffire, il est évident que l'activité centralisée du hub et l'activité déconcentrée du réseau Sudoc sont complémentaires. Cette question, qui mériterait une étude spécifique, sera à peine effleurée par ce rapport.
- Les technologies actuelles offrent de nouvelles perspectives face aux défis de l'agrégation des données et de l'interopérabilité.

[1] http://www.abes.fr/Connaitre-l-ABES/Projet-d-etablissement

L'ABES a jugé raisonnable de ne pas se lancer dans un projet aussi ambitieux que le hub sans évaluer concrètement sa faisabilité et son positionnement précis à travers une étude, sous la forme d'un prototype.

L'étude

Cette étude a pour objectif de construire un prototype de traitement d'un certain nombre de corpus de métadonnées, qui permette de :

- Agréger ces données
- Evaluer leur qualité et leur complétude
- Les corriger automatiquement
- Les enrichir (normaliser, compléter, lier, etc.)
- Les rendre accessibles, interrogeables et réutilisables par des tiers

Cette étude doit :

1. Imaginer quels sont les traitements qui apporteront une réelle plus-value aux données de départ
2. Trouver des solutions techniques pour effectuer en masse ces traitements automatiques

C'est en travaillant sur des données réelles et en se donnant des objectifs opérationnels que l'étude pourra espérer tirer des conclusions fiables et formuler des recommandations utiles.

Les corpus étudiés sont les suivants :

- Springer revues et articles (licence nationale)
- Springer ebooks (licence nationale)
- Thèses avant 1985 (Sudoc)
- Ebooks Dalloz (documentation électronique sous abonnement)
- Revues.org (Open Access)

L'étude s'est déroulée entre septembre 2012 et avril 2013.

L’équipe était constituée de :

- Ilhem Addoun
- Benjamin Bober
- Christophe Bonnefond
- Michael Jeulin
- Thomas Michaux
- Yann Nicolas
- Christophe Parraud

Approche

Miser sur les technologies du web sémantique pour gérer l'hétérogénéité des métadonnées à traiter

Par définition, le hub doit accueillir et traiter des métadonnées hétérogènes, à divers titres :

- **Par leur origine**. Certains corpus viennent du monde de l'édition, d'autres du monde des bibliothèques.
- **Par leur format**. Certains corpus sont en MARC, d'autres en XML. On peut encore imaginer d'autres formats à supporter : exports CSV de bases de données, fichiers TXT *ad hoc*, etc. Quand bien même tous les corpus seraient en XML, le défi de l'hétérogénéité demeurerait car XML n'est qu'une syntaxe générique : toute l'intelligence et la spécificité d'un corpus en XML tiennent au *vocabulaire* employé (quelle arborescence ? quels noms d'éléments et d'attributs ? quels espaces de nom ?). Le même argument peut être avancé en ce qui concerne MARC, qui a ses propres dialectes (UNIMARC, MARC21, DanMARC[2], etc.).
- **Par leur qualité**. Il existe bien des manières de pécher :
    - Assertions inexactes (ex : mauvais ISSN)
    - Non-conformité au standard annoncé (ex : utilisation erronée de champs MARC)
    - Absence de certaines informations, jugées nécessaires
    - Etc.
- **Par leur volume**. Le corpus Dalloz comprend moins de 1 000 ebooks tandis que le corpus des revues Springer comprend près de deux millions d'articles.

Traditionnellement, face à cette coûteuse hétérogénéité, l'approche est la suivante :

1. On incite les fournisseurs à livrer des métadonnées conformes à certains standards.
2. Quand l'ensemble des métadonnées ne converge pas vers un standard unique, on les convertit en un format pivot, quitte à sacrifier une partie de la richesse d'origine.

La présente étude veut tenter et évaluer une approche opposée :

1. Prendre les données telles qu'elles sont
2. Les convertir en RDF sans *rien* perdre de leur richesse d'origine

[2] http://www.kat-format.dk/danMARC2/Danmarc2.4.htm

RDF est un modèle de données universel et très souple, qui permet d'exprimer n'importe quel ensemble de données. Une base de données RDF n'impose pas de schéma figé aux données à modéliser. Quand un élément d'information des données d'origine correspond à une notion présente dans un vocabulaire RDF établi, il n'y a aucune raison de ne pas privilégier cette solution standard. Dans les autres cas, voire dans le doute, on peut forger ses propres propriétés et ses propres classes RDF, c'est-à-dire inventer un vocabulaire *ad hoc*, exprimé en RDF.

Les métadonnées RDF standard et les métadonnées RDF *ad hoc* pourront parfaitement coexister dans la même base RDF, c'est-à-dire être interrogées ou affichées ensemble, au sein d'une même requête. Qui plus est, au-delà de cette simple coexistence, ces deux sous-ensembles pourront être interconnectés au moyen des mécanismes offerts par les schémas RDF (RDFS). En effet, si, pour ne rien perdre de la richesse d'origine du corpus Springer, on a dû forger une propriété *ad hoc* comme *springerJ:hasSubjectSecondary1*, il est utile et facile de la rapprocher explicitement de la propriété Dublin Core standard *dcterms:subject*. Il suffit d'ajouter une information qui affirme formellement que *springerJ:hasSubjectSecondary1* est un cas particulier de *dcterms:subject*. Cette information est elle-même exprimée en RDF :

springerJ:hasSubjectSecondary1 rdfs:subPropertyOf dcterms:subject

Elle va donc coexister avec le reste des métadonnées RDF et même les enrichir. En effet, grâce à cette nouvelle information, un tiers qui ne comprendrait pas la notion *ad hoc* de *springerJ:hasSubjectSecondary1* pourrait s'en sortir en lui substituant automatiquement la notion plus large et standard de *dcterms:subject*. D'apparence modeste, c'est ce type de mécanisme logique qui fait la force de RDF, et facilite l'interopérabilité de données hétérogènes.

Dans le contexte de cette étude, l'approche RDF offre deux atouts majeurs :

- Sa **souplesse** offre une grande liberté pour modéliser tous les corpus que le hub devrait accueillir, sans perte d'information. Ceci est d'autant plus important ici que le parti-pris du hub est de **ne rien écarter des données d'entrée**. Toute information de départ doit trouver sa place dans le modèle RDF d'arrivée.
- Sa **capacité à intégrer ensemble** différents corpus permet, non pas d'intégrer entre eux les différents corpus à traiter par le hub[3], mais d'intégrer chacun de ces corpus

[3] *A priori*, par exemple, il n'y a pas grand chose à attendre d'une fusion des métadonnées Springer et des métadonnées Dalloz. Par ailleurs, le but du hub n'est pas de proposer un moteur de recherche multibases, ce qu'il est désormais convenu d'appeler « outils de découverte » - jusqu'à contre-ordre.

avec des corpus auxiliaires, qu'il s'agisse de corpus décrivant les mêmes documents[4] ou de référentiels[5].

Les mécanismes logiques au cœur de RDF promettent davantage en théorie[6], à savoir la possibilité de formuler en RDF, au sein même des données, des contraintes sémantiques[7]. Deux avantages principaux découlent du fait que les données et les contraintes sur ces données puissent être exprimées dans le même langage et le même environnement:

- **Portabilité** : les contraintes voyagent avec les données ; elles ne sont pas enfermées dans le logiciel qui gère la base de données, ni dépendantes d'un langage de programmation spécifique.
- **Autonomie de l'expert métier**[8] : l'expert métier est aussi celui qui déclare les contraintes qui portent sur les données

Architecture technique

Une base de données RDF

Au cœur du prototype développé dans le cadre de cette étude, on trouve donc les métadonnées en RDF.

Ces métadonnées sont gérées dans une base de données RDF. Une telle base offre le luxe de pouvoir manipuler les données telles qu'elles ont été modélisées, sans se préoccuper de la structure physique de la base ou des index. Cette gestion immédiate des données donne une grande autonomie à l'expert données, qui peut, sans l'assistance d'un informaticien, accéder aux données, naviguer dans leur structure, lancer des requêtes SPARQL[9] voire les modifier. Il est sans aucun doute plus facile à un expert données d'acquérir les compétences et l'expérience suffisantes en RDF et en SPARQL qu'à un informaticien de devoir maîtriser la

[4] Ainsi, on verra comment les données Worldcat, chargées dans la même base RDF que les données Springer, viennent tout naturellement les compléter, donc les enrichir.
[5] Ainsi, on verra comment les données VIAF ou LCSH, une fois exprimées en RDF, s'associent tout naturellement aux données Springer de la base RDF du hub.
[6] « en théorie » seulement, car aujourd'hui, l'utilisation de ces mécanismes logiques demeure rare. Ces opérations de raisonnement sont coûteuses en termes de performances et délicates à comprendre et à maîtriser par les non-spécialistes : il n'y a pas (encore) de module de logique à l'enssib.
[7] Exemples : « une thèse ne peut pas avoir deux auteurs », « une université ne peut être l'auteur d'une thèse », « un document électronique et un document imprimé ne peuvent avoir le même ISBN », etc.
[8] Dans notre contexte, l'expert métier est un spécialiste des données bibliographiques ou des données de gestion documentaire, qu'il travaille à l'ABES ou dans ses réseaux : Sudoc, IdRef, Calames et Thèses.
[9] Voir le glossaire

logique des données bibliographiques, de surcroît modélisées en RDF. Il va sans dire que le recours à un informaticien s'impose dès qu'il s'agit de construire une workflow complet, de concevoir et écrire un algorithme complexe ou d'optimiser un traitement.

Dans le cadre de cette étude, c'est la version open source de la base Virtuoso qui a été utilisée, après analyse des benchmarks internationaux et prise en compte des priorités du hub : gros volumes, interface web d'administration, environnement de programmation, etc. Ce choix n'a rien de définitif.

Quoi qu'il en soit, la base RDF ne peut être l'Alpha et l'Omega du hub. Quand il s'agit d'acquérir les données RDF, de les traiter en masse ou de les redistribuer, SPARQL ne suffit plus ; le recours à d'autres outils et d'autres compétences s'impose.

Ingestion des données RDF

Le plus souvent, les données d'origine sont formatées en XML. A défaut, elles seront bien souvent en MARC – d'où il est aisé de les convertir en MARCXML, sans perte de richesse. La question de l'ingestion des données du hub revient donc à leur transformation d'XML vers RDF.

Pour ce faire, l'approche qui a été retenue en amont de l'étude et validée en cours consiste à transformer les données XML en RDF/XML au moyen d'un script XSLT. Cette approche présentait les avantages suivants, confirmés en cours de route :

- C'est une solution classique, qui demande des compétences répandues à l'ABES.
- C'est une solution portable, car un script XSLT peut être exécuté dans tous les environnements informatiques.
- C'est une solution qui permet l'intervention de l'expert données frotté d'XML comme du développeur familiarisé avec les données bibliographiques. Comme pour SPARQL, le fait qu'XSLT soit si proche des données est un atout important du point de vue de la gestion des compétences.

Le langage XSLT possède un autre atout, précieux dans le contexte du hub : il facilite le traitement des données à l'aveugle. En effet, les données XML à traiter en amont peuvent être très riches et mal documentées, ou encore irrégulières. De ce fait, il est difficile pour un script XSLT de prévoir tous les cas qu'il peut rencontrer. Or, en XSLT, le mécanisme des templates suit un raisonnement par défaut : si le script n'a pas explicitement prévu un template pour traiter tel élément ou tel attribut XML de départ, alors c'est un template par défaut qui est déclenché. Le script peut lui-même déterminer ce comportement par défaut. En l'occurrence,

un template par défaut génère des triplets RDF de gestion qui vont documenter le nom et la valeur des éléments et attributs XML imprévus et rattacher ces informations à l'entité principale décrite par la notice. Ces triplets seront ensuite chargés dans la base RDF au côté des triplets de description bibliographique. Il sera alors facile de les interroger pour lister toutes les informations présentes dans les données de départ mais oubliées par le travail de modélisation et donc non converties en RDF.

En résumé, une fois converties en RDF/XML, les métadonnées sont chargées dans la base RDF et, sans autre forme de procès, elles deviennent interrogeables en RDF, à l'aide du langage de requête SPARQL.

On verra plus loin qu'à la marge, il a fallu charger dans la base certaines données qui n'étaient pas disponibles sous forme XML :

- Un fichier Excel[10] ;
- Un fichier TXT ;
- Des données enfouies dans des pages HTML[11].

Dans ces cas-là, on est condamné au sur-mesure, ce qui n'est pas forcément très coûteux (en temps de travail). Là encore, ce qui compte n'est pas tant le format des données que leur cohérence et leur intelligibilité (y compris à l'aide d'une documentation associée).

Traitement en masse

Après chargement puis analyse[12] des données dans la base RDF, vient le moment de modifier ces données.

Dans le contexte d'une base de données RDF, où toute l'information est découpée en affirmations élémentaires (appelées « triplets »), modifier les données peut signifier :

- Ajouter de nouveaux triplets[13]
- Remplacer certains triplets par d'autres[14]

[10] Voir plus loin
[11] Voir plus loin
[12] On ne détaille pas ici cette phase d'analyse, car elle peut prendre bien des formes, automatiques ou non, qui sont détaillées dans les parties centrales de l'étude.
[13] Par exemple, il suffit de charger le corpus auxiliaire IdRef à côté des métadonnées de thèse pour *ipso facto* enrichir et donc modifier le corpus des thèses.

Dans certains cas simples, qui n'impliquent ni un algorithme trop complexe, ni le recours à des ressources non-RDF, ni une volumétrie importante, il est possible de modifier le corpus RDF au moyen d'une requête SPARQL[15]. Mais c'est en fait rarement le cas.

Il est donc nécessaire de recourir à des programmes qui vont travailler autour des requêtes SPARQL qui accèdent aux données,

- soit pour implémenter un algorithme trop complexe pour une requête SPARQL,
- soit pour faire appel à des sources d'information extérieures (comme des web services),
- soit pour gérer la volumétrie en exploitant les curseurs de la base de données,
- etc.

Au début de l'étude, le langage de programmation Python a été retenu, pour les raisons suivantes :

- C'est un langage riche .
- Il gère bien le RDF.
- Il est facile à lancer voire à modifier par un expert métier.

Au cours de l'étude, il s'est avéré plus aisé d'utiliser Virtuoso/PL[16], le langage de programmation procédurale interne à la base de données choisie pour le prototype, Virtuoso. En étant ainsi au cœur de la base, on gagne en performance et en fiabilité, mais la portabilité est menacée. Cette question reste donc ouverte.

Redistribution des données

Le hub n'est pas un cul de sac. C'est un lieu de transit qui doit redistribuer les données par des canaux qui vont leur ouvrir une nouvelle carrière.

Ces canaux de sortie sont les suivants :

- Extractions (dumps)

[14] Par exemple, après analyse et alignement sur un référentiel des établissements habilités, il peut être nécessaire de remplacer la valeur « Rennes I » par la valeur « Rennes 1 » dans tous les triplets introduisant l'établissement de soutenance. Parfois, on jugera nécessaire de conserver l'information d'origine.
[15] La norme W3C SPARQL Update en l'occurrence.
[16] http://docs.openlinksw.com/virtuoso/sqlprocedures.html

- Echanges avec les bases de connaissance du marché, ou du moins avec la base de connaissance nationale recommandée par l'étude Pleiade[17]
- Google Scholar
- Sudoc
- Web de données
- Interface professionnelles de récupération de données. Ainsi, on peut imaginer que le site www.licencesnationales.fr propose aux professionnels un moteur de recherche expert sur chacun des corpus ISTEX traités par le hub. Grâce à ce moteur, les professionnels pourraient sélectionner des sous-ensembles sur mesure[18], correspondant à une requête aussi fine et sophistiquée que le permettra la structuration des métadonnées par le hub. Cette sélection opérée, les professionnels pourraient décider de récupérer ces lots sous tel ou tel format, ou encore de déclencher une exemplarisation automatique dans le Sudoc pour ce lot. Un tel dispositif permettrait d'offrir aux bibliothèques un (self-)service personnalisé.[19]

Cette étude ne s'est pas penchée en détail sur le développement de ces canaux de redistribution, car ces chantiers présentent moins d'inconnues. En effet, ces dernières années, l'ABES a été amenée à se confronter aux défis suivants :

- Génération mensuelle d'un dump de tout le Sudoc en RDF
- SELF Sudoc : interface professionnelle de récupération de gros volumes de données
- Maîtrise du moteur de recherche Apache Solr
- Exposition du Sudoc, d'IdRef, de Calames et de theses.fr sur le web de données
- Offre de web services pour le Sudoc, IdRef et theses.fr
- Exemplarisation automatique

Le développement de ces sorties représenterait une quantité de travail importante, mais ne présenterait pas de difficultés inédites.

[17] http://fil.abes.fr/2013/03/29/etudes-sgbm-et-dispositif-de-decouverte-publiees/

[18] Par exemple, une petite bibliothèque municipale pourrait « picorer » dans les grandes masses des données des licences nationales, et seulement récupérer les quelques dizaines de notices qui correspondent à sa politique documentaire.

[19] A terme, il serait sans doute souhaitable d'étendre ce dispositif au-delà des seules données des licences nationales, y compris aux données du Sudoc elles-mêmes.

Analyse et traitement des corpus

Chacune des cinq sections à suivre s'intéresse à un corpus et un seul. Néanmoins, en analysant un corpus, on peut être amené à introduire une notion qui servira à l'analyse des autres corpus. Il est donc préférable de lire ce chapitre dans l'ordre qui suit.

Les revues Springer

Acquisition, modélisation et conversion

Le programme de licences nationales dans lequel est impliquée l'ABES (achats au moyen de l'enveloppe dite « d'impulsion » ou dans le cadre du projet ISTEX) entraîne le brassage d'énormément de métadonnées. Contractuellement, chaque éditeur doit livrer les métadonnées décrivant l'intégralité du contenu acquis par l'ABES. En réalité, sans même avoir à appliquer des analyses et sondages poussés, on s'aperçoit rapidement qu'il est très difficile pour les éditeurs, particulièrement pour ceux de périodiques scientifiques, de se conformer à cette obligation. Métadonnées lacunaires et incorrectes risquent d'être le lot commun de corpus qui ne sont pas nés digitaux mais qui ont été numérisés. Pire, il ne semble pas totalement improbable d'envisager que, pour des raisons uniquement techniques, l'éditeur ne livre pas l'intégralité des contenus qu'il est censé fournir contractuellement.

Dans ce cadre, une des premières missions du hub sera d'analyser la complétude de tels corpus en mettant en regard la licence signée par l'éditeur avec les métadonnées des documents livrés. Il s'agira ensuite d'améliorer les données fournies et de les exposer au moyen des outils généralement utilisés dans la gestion des ressources électroniques.

Dans le cadre de l'accord de licence nationale signé le 13 juillet 2011, l'éditeur Springer a livré sur disque dur données (texte intégral) et métadonnées d'archives de périodiques de plus d'un millier de périodiques (des origines à 1996). Ces données brutes se présentent sous la forme d'arborescence de dossiers (revues/volume/fascicule/article). Dans le dernier dossier se trouve le texte de l'article en PDF ainsi qu'un fichier de métadonnées.

Les fichiers de métadonnées sont tous en XML et associés à une DTD propre à Springer appelée A++. Cette DTD, très riche, maintenue et bien documentée[20], a permis une modélisation en RDF au plus proche de ce qui existait.

Comme pour les autres corpus, le principe a été de modéliser tout ce qui existe. La documentation a permis de modéliser un grand nombre d'éléments avec des vocabulaires courants (FOAF, RDA Relationships, DCTerms, Bibo). Plusieurs propriétés (numéro de volume ou de fascicule, identifiant interne, indexation sujet) ont quant à elles dû être forgées, faute de vocabulaire propre suffisamment précis (par exemple n° de volume de début et n° de volume de fin, dans le cas de volumes multiples). Les informations qui n'ont pu être modélisées ont été identifiées et pourront être retraitées si elles s'avèrent nécessaires.

[20] http://production-customer.springer.com/Section_APlusPlus.html (accès contrôlé)

Le passage du XML A++ au RDF-XML s'est fait par transformation XSLT. Plusieurs essais ont dû être menés avant d'aboutir à une modélisation complètement satisfaisante.

En tout ont été modélisés 1 965 510 articles, soit 71 millions de triplets représentant environ 29 Go de données.

Analyse

Détection des anomalies

L'analyse globale du corpus est effectuée en dehors de la base RDF.
Une première analyse rapide consiste à relever toutes les valeurs que peut prendre une propriété via un script Python qui interroge la base RDF en HTTP. Ce « Sparql Tour[21] » permet d'avoir une bonne vue d'ensemble en détectant rapidement une anomalie (valeur textuelle à la place d'une valeur numérique par exemple). On trouve ainsi 194 mentions de date de copyright parmi lesquelles 22 sont erronées (1, 11981, « ugie »,…). Il convient toutefois de vérifier que l'anomalie constatée provient bien des données telles quelles, et non de leur modélisation, ce qui nécessite une intervention humaine.

Une deuxième approche consiste à exporter le résultat d'une requête lancée dans la base RDF dans Open Refine[22] et à appliquer une méthode de clusterisation par rapprochement (par exemple en s'appuyant sur la distance de Levenshtein). Cette méthode permet de repérer des variantes non pertinentes (fautes de frappe, erreurs d'encodage,…). Ainsi, sur les 1318 titres que l'on trouve dans le corpus, 54 sont des formes incorrectes devant être remplacées. Dans de nombreux cas, la différence constatée provient uniquement d'une différence de casse (*Bulletin volcanologique* / *Bulletin Volcanologique*). En revanche, on trouve quelques variantes pouvant avoir un impact sur une recherche (*Transaction* / *Trans-action*, *ou Audiovisual communication review* / *Audio Visual communication review*)
Si toutes les données peuvent être ainsi analysées, il convient de se focaliser sur les données qui seront directement utilisées pour les traitements (titre, ISSN, date, etc.).

[21] Ce SPARQL Tour est un script qui lance une série de requêtes SPARQL sur un corpus de la base RDF. En générant toutes sortes de statistiques et d'échantillons concernant les données de ce corpus, SPARQL Tour permet de s'en faire une idée générale et de détecter de premières anomalies (exemples : propriétés essentielles absentes pour les entités de telle classe, valeurs inattendues pour telle propriété, etc.).
[22] « *OpenRefine (ex-Google Refine) is a powerful tool for working with messy data, cleaning it, transforming it from one format into another, extending it with web services, and linking it to databases like Freebase.* » selon http://openrefine.org/.

Détection des lacunes (défaut de livraison)

Il convient de distinguer deux cas :

- Le corpus se suffit à lui-même pour déceler une ou plusieurs lacunes (interruption dans une séquence de volumes ou fascicules par exemple)
- Le corpus doit être confronté à une autre source de données, en premier lieu la licence (absence d'une revue entière ou lacunes en début ou fin d'état de collection)

Dans le premier cas, l'opération consiste à lister l'ensemble des fascicules disponibles pour une revue donnée et à les classer par date croissante.
Par la suite on peut :

- Comparer le numéro du dernier fascicule de chaque volume avec la valeur de la propriété « VolumeIssueCount » afin de voir s'il manque un fascicule au sein d'un volume.
- Tester s'il y a des lacunes dans l'état de collection au niveau des volumes en s'appuyant sur leur numérotation.

Dans le cadre de la présente étude la chaîne de traitement n'a pas été automatisée. La lourdeur de la requête SPARQL utilisée ne permettant pas de lister tous les fascicules de toutes les revues, on a procédé uniquement par échantillon. Les tests ont été effectués au moyen de scripts Python.
Sur 38 titres testés, soit 1396 volumes, on a trouvé par cette méthode 10 volumes manquants. Cela ne prête à aucune extrapolation possible mais montre que ces lacunes sont bien présentes, parfois dans des proportions importantes au sein d'un même titre.

Dans le deuxième cas il s'agit de comparer une liste de référence, en l'occurrence celle disponible dans la licence signée avec Springer, avec les données disponibles. La difficulté de cette opération vient du fait que la liste fournie par Springer dans la licence n'était pas bibliographiquement correcte (mauvaises attributions d'ISSN par exemple). Il a donc fallu faire un important travail manuel de correction de la liste de référence. De la même façon, la majeure partie de l'analyse a dû être faite à la main. Elle a révélé que manquaient 141 volumes répartis sur 22 revues et – plus grave encore – qu'étaient absents des données livrées 70 titres complets.

Les méthodes utilisées ici peuvent être largement optimisées.

Cas des ISSN électroniques

Springer n'a pas suivi les règles d'attributions d'un ISSN électronique : le même identifiant est donné à toute une lignée de périodique.

Une requête SPARQL permet de relever les cas où une revue et son lignage ont un ISSN électronique identique mais des ISSN papier différents. 240 « revues » sont dans ce cas, ce qui représente 476 ISSN électronique erronés.

Traitements

Corrections

Deux cas de figure sont à envisager :

- La correction de coquilles ou d'erreurs d'encodage : la suppression de la valeur originelle n'entraîne aucune perte d'information. Une requête SPARQL UPDATE devrait permettre de remplacer les valeurs erronées par des valeurs correctes.
- La correction d'ISSN : l'ISSN électronique a beau être faux, il s'agit néanmoins d'un identifiant interne utilisé par Springer. Dans ce cas il convient de garder cet identifiant au moyen d'une propriété forgée et d'affecter le bon ISSN à la propriété de base.

Dans le cadre de l'étude les corrections n'ont pas été appliquées. Elles nécessiteront une intervention humaine puisqu'il s'agira dans le meilleur des cas de choisir une forme correcte parmi plusieurs variantes, de chercher la vraie valeur d'une propriété dans les autres métadonnées disponibles (par exemple une date), voire de s'intégrer dans un workflow extérieur (demande de numérotation ISSN ?).

KBART

La recommandation KBART[23] est le fruit d'une collaboration NISO-UKSG. Elle vise à normaliser les fichiers qui seront utilisés par les bases de connaissance articulées avec des produits de découverte ou de gestion des ressources électroniques.

[23] http://www.uksg.org/kbart

Le hub se doit de redistribuer les métadonnées de document électronique sous ce format, en direction de ces bases de connaissance. Les données KBART ainsi générées aideront à rendre accessibles les revues disponibles en licence nationale dans de meilleures conditions, en s'affranchissant complètement des interventions des opérateurs commerciaux qui ne disposent pas toujours de listes censées refléter le contenu des bouquets, et encore moins de listes correctes.

La structure des données de la base RDF a rendu complexe la construction de la requête SPARQL permettant de faire apparaître les données nécessaires. Il semblerait cependant là encore que l'on puisse aller vers une meilleure optimisation des ressources.
Une fois la requête créée, elle peut être rejouée au fur et à mesure des corrections que l'on pourrait apporter aux données initiales.

Si l'on se réfère à la recommandation KBART, une information importante manque aux données fournies par Springer : l'URL d'accès au niveau de chaque titre. Cette absence est compréhensible : c'est une donnée qui est rarement pérenne, à moins d'utiliser un DOI par exemple. Cela est d'autant moins problématique que les données d'accès seront différentes suivant la plateforme utilisée (Springer ou ISTEX). Il faut donc les reconstituer pour chaque contexte.
Dans le cas de Springer deux méthodes sont à disposition :

- Détermination de la syntaxe propre utilisée par Springer :
  - Exemple : http://link.springer.com/journal/10463 . Elle s'appuie sur l'identifiant interne attribué par Springer à chaque revue
- Utilisation de la syntaxe OpenURL :
  - Exemple : http://link.springer.com/openurl?genre=journal&issn= 1572-9052 . Elle s'appuie sur l'ISSN électronique

Les alternatives supposent de connaître de toutes façons une syntaxe propre à Springer, au moins pour l'URL de base. On peut imaginer que la première solution pourra s'appliquer à l'ensemble des corpus à venir puisqu'elle ne nécessite pas de pré-requis techniques particuliers. La deuxième suppose au contraire que l'éditeur propose une plateforme qui est compatible OpenURL .

Au moment de générer les fichiers KBART, il ne semble pas y avoir de raison de privilégier une solution plutôt que l'autre.

Les ebooks Springer

Acquisition, modélisation et conversion

L'acquisition, la modélisation et la conversion en RDF du corpus des ebooks Springer font écho à celles des revues Springer. Là encore, il s'agit de documents diffusés par Springer, acquis en licence nationale et décrits en XML avec la DTD A++ de Springer.

Identifier le contenu du corpus

Le base RDF des ebooks Springer contient 11 642 ebooks, pour 11 636 ISBN électroniques uniques.

39,3 % de ces ebooks sont postérieurs à 2004 et, de ce fait, ne devraient pas faire partie de ce corpus de licence nationale. L'ABES a reçu les fichiers PDF[24] et les métadonnées XML de ces 4 576 ebooks postérieurs à 2004, mais également les accès aux documents sur la plateforme Springer : un script de simulation d'accès aux documents n'a échoué que dans 218 cas (moins de 0,05 % des ebooks postérieurs à 2004).

D'un point de vue contractuel, le défi du département Adèle[25] est de restaurer une cohérence entre le contrat signé et la réalité des documents acquis et accessibles, au bénéfice des usagers des bibliothèques.

Du point de vue du signalement, on constate que 98,6 % des ebooks de la base RDF sont d'ores et déjà signalés dans le Sudoc : sur 11 636 isbnelec, 11 473 correspondent à au moins une notice.[26] Seuls 162 ebooks n'ont pas de notice dans le Sudoc.[27]

Or, dans le cadre de la licence nationale, Springer n'a livré que 7 360 notices MARC. Comme le prouve une simple requête SolrTotal[28], aucune de ces 7 360 notices ne décrit un document postérieur à 2004.

---

[24] qui, par ailleurs, possèdent des métadonnées internes qu'il pourrait être intéressant d'extraire et d'exploiter.
[25] Adèle est un département de l'ABES dédié à l'« Achat de documentation électronique ».
[26] 42 ISBN correspondent à deux notices Sudoc.
[27] Parmi ces 162 ebooks du hub sans notice dans le Sudoc :
- 93 correspondent à la période 1922-2004 (dont 28 pour la seule année 2004)
- 69 sont postérieurs à 2004 (dont 63 de 2012)

[28] http://www.sudoc.fr/[redacted]?q=035-a_t:springerln%2A&version=2.2&start=0&rows=50&indent=on&facet=true&facet.method=fc&facet.limit=50&f

Inversement, on a pu lister les notices Sudoc qui correspondent à des ebooks Springer du hub mais ne sont pas marquées « SpringerLN »[29]. Là encore, SolrTotal nous apprend que tous les documents décrits par ces notices sont postérieurs à 2004. Ces 4 500 (environ) notices ont été livrées à l'ABES pour permettre aux établissements *abonnés* à titre individuel de se localiser dessous dans le Sudoc. De fait, aujourd'hui, elles décrivent également ces ebooks bonus, rendus disponibles pour tous à l'occasion du contrat licence nationale – disponibles de fait, mais non de droit.

Il reste à élucider le cas de 370 notices Sudoc marquées « SpringerLN » qui sont absentes de la base RDF.

Ces analyses ont été menées grâce à la base RDF du hub, à Solr Total, au web service isbn2ppn[30] et à Open Refine. Ces outils sont désormais assez bien maîtrisés pour pouvoir être mobilisés rapidement, notamment *pendant* les négociations ISTEX. Dans ce contexte, ils aideraient à comparer le corpus électronique en négociation et l'ensemble des documents imprimés correspondants signalés dans le Sudoc. Le nombre de documents imprimés et le nombre d'exemplaires associés peuvent constituer des éléments d'appréciation intéressants.

Lier aux autorités personne physique d'IdRef

Les notices XML d'ebooks sont riches d'information sur les contributeurs (auteurs et éditeurs intellectuels) des chapitres, des livres et même des collections. Outre les noms de ces personnes, avec titres et pléthore d'initiales, on trouve souvent l'email et l'affiliation très détaillée. Par contre, ces personnes ne sont associées à aucun identifiant, ni à un identifiant interne à Springer, ni à un identifiant international comme VIAF.

L'email et l'affiliation sont néanmoins des informations très précieuses pour identifier une personne, c'est-à-dire la ré-identifier d'un contexte à l'autre : si les auteurs de deux livres différents ont un nom ressemblant et le même email, on doit en déduire qu'il s'agit de la même personne. En RDF, il est possible de déclarer que deux entités sont identiques, même si on est incapable de leur attribuer un identifiant unique. En MARC, au contraire, à moins de rattacher deux noms à une même autorité, il n'est pas possible de dire que ces deux noms se rattachent à la même personne.

---

acet.field=100-a-pos9-12_s&fl=001_s,035-a_s,&solrService=SolrTotal (cette notice appelle toutes les notices dont le numéro source (035$a) commence par « SpringerLN » et affiche une facette sur la date de publication) [NB : *SolrTotal est une instance du moteur de recherche Solr qui indexe chaque sous-zone et indicateur UNIMARC. Cet outil d'analyse du Sudoc n'est accessible qu'en interne, pour des raisons techniques*]

[29] Numéro source (035$a) commençant par « SpringerLN »

[30] A condition de chercher sur les deux types d'ISBN : 10 et 13

Pourtant, on aimerait pouvoir rattacher les contributeurs des ebooks Springer à des autorités. On peut imaginer différentes pistes :

- Exploiter le programme SudocAD[31] de liage automatique aux autorités IdRef, ce qui suppose que les notices à lier utilisent l'indexation Dewey
- Exploiter les notices Sudoc des manifestations imprimées correspondant aux ebooks Springer, pour lesquelles les catalogueurs du réseau Sudoc ont établi des liens vers les autorités IdRef[32]
- Exploiter les notices Worldcat des manifestations imprimées correspondant aux ebooks Springer, pour lesquelles les catalogueurs du réseau Worldcat ont établi des liens vers les autorités Viaf ou LC

D'ailleurs, ces différentes voies peuvent coexister. Reste à savoir dans quel ordre il faudrait les enchaîner.

Quelle que soit l'approche, le fait d'avoir au préalable (ré-)identifié une personne au sein du corpus Springer (grâce aux emails notamment) va porter ses fruits : si on a pu établir que A, B, C et D sont en fait la même personne, alors il suffit de lier l'une des quatre mentions à une autorité pour propager ce liage à travers les quatre autorités. Les technologies sémantiques se prêtent très bien à ce genre de raisonnement.

Pour les besoins de l'étude, la piste Worldcat a été explorée. La démarche est la suivante :

1. On part d'un ebook Springer de la base RDF.
2. On prend l'ISBN de la manifestation imprimée correspondant à cet ebook, que les données Springer mentionnent à côté de l'ISBN électronique.
3. On passe cet ISBN papier au web service de Worldcat xISBN[33].
4. Ce web service renvoie la liste des notices Worldcat qui décrivent une manifestation de l'œuvre incarnée par la manifestation de départ[34]. Chaque notice est identifiée par un OCN.
5. Pour chaque OCN, on appelle la page HTML de la notice dans Worldcat[35].
6. La plupart de ces notices de Worldcat contiennent des métadonnées schema.org, c'est-à-dire interprétables en RDF.

[31] http://www.abes.fr/Sudoc/Projets-en-cours/SudocAD

[32] Le web service Sudoc *isbn2ppn* permet de découvrir les notices Sudoc possédant tel ISBN - à condition de chercher à la fois l'ISBN10 et l'ISBN13, même quand les métadonnées de départ ne contiennent qu'une des deux formes. Sans cet effort de conversion entre les deux ISBN, on passe à côté de beaucoup de notices.

[33] http://xisbn.worldcat.org/webservices/xid/isbn/{$isbnpapier}?method=getEditions&format=xml&fl=oclcnum
L'utilisation de ce web service est très restreinte, même pour les « *affiliates* » aux API Worldcat que sont les clients de CBS comme l'ABES.

[34] On tire ainsi profit de la FRBRisation de Worldcat, qui, par ailleurs, n'est pas sans défaut.

[35] http://www.worldcat.org/oclc/{$ocn}

7. On analyse ces pages HTML pour en extraire les triplets RDF qu'on charge dans un nouveau graphe[36] de la base RDF.

A ce stade, la base RDF contient :

- La version RDF des notices XML fournies par Springer
- La version RDF des notices MARC de Worldcat fournies via l'HTML de Worldcat

Dans les données RDF de Worldcat, on découvre des triplets qui pointent vers les autorités de VIAF ou celles de la bibliothèque du Congrès – qui elles-mêmes pointent vers VIAF.[37] Et de VIAF à IdRef, il n'y a qu'un pas, qu'on peut franchir en chargeant dans la base RDF un dump de VIAF – ce qui a été fait.

Mais comment enrichir les données Springer avec ces précieuses informations Worldcat ? Deux solutions ont été envisagées :

- On pourrait accorder une confiance aveugle aux données Worldcat et décider d'écraser les données d'origine, mais :
  - on a pu constater que les notices Worldcat mentionnent souvent moins de contributeurs que les données Springer ;
  - les données Springer comprennent d'autres informations qu'il serait regrettable d'effacer (emails, affiliations, etc.).
- Il est préférable d'adopter une approche plus modeste qui consiste à comparer les noms mentionnés de part et d'autre. Le postulat est ici le suivant : si Springer affirme que A est l'auteur de ce livre, si Worldcat affirme que A' est l'auteur de ce même livre et que le nom de A et le nom de A' sont identiques à 95%, A et A' ne sont qu'une seule et même personne.

Cette dernière étape n'a pas été implémentée jusqu'au bout dans le cadre de cette étude, faute de temps. Deux techniques ont été essayées :

- Utiliser SILK[38], un outil spécialement dédié à la découverte de liens entre corpus RDF
- Utiliser SPARQL, après avoir ajouté au langage SPARQL une fonction spéciale qui permet de mesurer la distance entre deux chaînes de caractères

Cette dernière étape serait tout autant incontournable si on avait cherché à puiser des triplets d'autorité dans le Sudoc, plutôt que dans Worldcat. Par contre, la solution SudocAD

[36] Voir le glossaire
[37] Il est à noter que ces liens aux autorités sont bien présents dans la base Worldcat et visibles et modifiables à partir du client de catalogage Connexion, mais ils disparaissent de tous les formats de sortie : z39.50, web services, etc. Ici, paradoxalement, les modestes métadonnées schema.org contiennent quelque chose de très précieux qui est absent du touffu MARC.
[38] http://wifo5-03.informatik.uni-mannheim.de/bizer/silk/

dispenserait de cette étape puisque qu'il s'agit d'associer directement une autorité à un nom. Par contre, SudocAD exige que les données de départ contiennent un indice Dewey.

Lier à RAMEAU et à Dewey

Pour rattacher les données Springer à Rameau ou Dewey, on peut adopter la même approche que pour les autorités de contributeurs, car les données schema.org de Worldcat contiennent également des triplets qui pointent vers les Library of Congress Subject Headings (LCSH), qui eux-mêmes sont alignés avec RAMEAU. Et de RAMEAU à IdRef…

Mais, ici, on peut ignorer la dernière étape : on n'a pas à chercher quel mot-clé de Springer correspond à quel terme LCSH. Contrairement aux attributions d'auteur, les sujets d'un livre peuvent se cumuler. On peut perdre en pertinence et lisibilité, mais pas en véracité.

Si on préfère éviter la prolifération des descripteurs, il faut partir des descripteurs fournis par Springer. Il ne s'agit pas de mots-clés libres mais d'une indexation précise selon un vocabulaire contrôlé propriétaire[39]. Ce vocabulaire contient près de 1 000 termes, répartis sur quatre niveaux. En voici un extrait :

	Y	**Psychology**
1	Y00007	**Psychology, general**
2	Y12005	**Clinical Psychology**
3	Y12010	Psychotherapy and Counseling
3	Y12020	Health Psychology

On pourrait imaginer des traitements automatiques plus ou moins sophistiqués pour mettre en correspondance les termes de ce vocabulaire et des termes LCSH. Mais le résultat restera toujours incertain et incomplet. Etant donné la taille modeste de ce vocabulaire et l'ampleur des corpus Springer qu'il indexe, il est préférable d'effectuer cet alignement de manière entièrement manuelle, avant de l'exprimer en RDF (SKOS) et de le charger dans la base RDF. Il devient alors possible d'explorer et d'interroger les données Springer comme si elles avaient été indexées en LCSH. Le hub pourrait alors exposer des données Springer avec une indexation RAMEAU fiable, ce qui serait utile à des catalogues comme le Sudoc mais aussi à des moteurs de recherche qui manquent souvent d'indexation-matière en français à exploiter. Une des missions du hub serait également de publier cette table de correspondance entre les

39 http://www.springer.com/cda/content/document/cda_downloaddocument/product_market_codes_current_issue_01_2013_english.xls

vocabulaires Springer et LCSH, y compris en RDF pour enrichir le web de données. Il faut tirer profit du web de données mais également y contribuer, à la mesure de nos possibilités.

Enfin, l'étude s'est également penchée sur la question de la classification, voisine de la question de l'indexation matière. A beaucoup d'égards, il serait précieux de pouvoir appliquer la classification Dewey à tous les corpus du hub. Cela permettrait de pouvoir découper et redistribuer chaque corpus selon les catégories bien connues de Dewey. Cela permettrait aussi de déclencher d'autres enrichissements en cascade : ainsi, un service comme SudocAD s'appuie sur un indice Dewey donné pour déduire d'autres informations.

Là encore, l'étude a exploité un web service de Worldcat pour injecter de nouvelles informations dans la base RDF. En l'occurrence, il s'agit du web service expérimental Classify[40] qui prend un ISBN en entrée et fournit un indice Dewey en sortie. Pour ce faire, Classify analyse les indices Dewey associés à toutes les manifestations de l'œuvre qu'incarne la manifestation de départ, et en fait la synthèse. Dans le cas des ebooks Springer, cette méthode a permis de générer un indice Dewey dans un tiers des cas, ce qui n'est pas négligeable.

Comparer les notices MARC de Springer, du hub et du réseau Sudoc

Afin de mesurer l'apport potentiel du hub par rapport aux pratiques actuelles, on a comparé trois variétés de notices MARC décrivant les ebooks Springer :

- Les notices MARC livrées par Springer
- Les notices MARC générées par le hub à partir des données XML livrées par Springer
- Les notices MARC créées *ex nihilo* par un catalogueur scrupuleux

Le comparatif complet et précis se trouve en annexe 1, sous la forme d'un tableur.

Les notices MARC livrées par Springer :

- respectent *a minima* les consignes données par l'ABES aux éditeurs ;
- contiennent des erreurs factuelles, tout en respectant la forme de MARC21:
    - Les zones censées être descriptives (2XX) ne sont pas des transcriptions fidèles du document.
    - Il y a des erreurs dans les codes de fonction :
        - Seul le premier contributeur possède un code de fonction

[40] http://classify.oclc.org

        - Le code de fonction est le code « auteur » même quand il s'agit d'un éditeur scientifique
    - Les auteurs secondaires n'apparaissent pas.
    - La date est incorrecte car une confusion existe entre la date de l'original imprimé et la date du document numérique.
    - La collection mentionnée est incorrecte car il s'agit de la collection papier.
- L'indexation matière en LCSH est bienvenue mais fait double emploi avec les mots-clés libres. Même au sein de chaque type d'indexation, on constate une grande redondance : ainsi, le mot-clé libre très général « Computer Science » coexiste avec le mot-clé plus spécifique « data Encryption ».

Ces notices MARC livrées par Springer ont probablement été générées à partir d'une base de données qui est également derrière les notices XML. Leur conformité à MARC21 apparaît comme une conformité de surface, résultat d'un effort scolaire mal maîtrisé. Les erreurs introduites dans le MARC sont difficiles à détecter, sauf livre en main ou ebook à l'écran.

Les notices MARC que le hub serait aujourd'hui capable de générer à partir du traitement et de l'enrichissement des notices XML livrées par Springer présentent les caractéristiques suivantes :

- Les métadonnées Springer contiennent des métadonnées purement descriptives, qui semblent provenir du document lui-même via un processus de numérisation.
- On peut distinguer entre les auteurs et les éditeurs intellectuels.
- Il y a suffisamment d'informations sur les contributeurs pour pouvoir créer ou enrichir des autorités, notamment en exploitant les affiliations, non pas à des fins biographiques mais à des fins d'identification de la bonne personne et de différenciation des homonymes.
- Au-delà de la description du livre, chaque sous-partie et chaque chapitre sont décrits, ce qui permet d'inclure dans la notice du livre une table des matières détaillée et exacte ou bien la liste des œuvres contenues avec leur date de copyright.
- L'indexation matière fournie en XML n'est pas qualifiée selon les LCSH. Par contre, le hub pourrait introduire un mapping systématique entre le vocabulaire d'indexation de Springer et les LCSH – et donc RAMEAU, ce qui apporterait une indexation en français. Par ailleurs, le hub peut s'appuyer sur la hiérarchie du vocabulaire Springer pour alléger l'indexation : ne laisser que les descripteurs les plus précis, et oublier les descripteurs plus généraux qui les englobent.
- Les données XML ont de quoi établir un lien vers la bonne notice de collection et enrichir considérablement cette dernière.
- Le hub peut créer et modifier en masse les exemplaires des notices. Il pourrait également déclencher une alerte au moment où les documents des licences nationales

deviendront également consultables sur la plateforme ISTEX, et plus seulement chez l'éditeur.

- Le hub est capable de créer des liens symétriques entre le document imprimé et sa contrepartie électronique, car les données XML comprennent l'ISBN imprimé de l'ebook décrit. Il suffit alors d'aller récupérer le PPN de l'imprimé au moyen du web service isbn2ppn pour pouvoir créer ce lien, puis le lien inverse.

Les notices créées *ex nihilo* par un catalogueur scrupuleux seront toujours plus riches et précises que des notices générées. Cela tient surtout à l'observation directe du document par le catalogueur et à sa capacité à créer de nouvelles autorités si besoin est.

Le scénario le plus prometteur sera celui qui saura combiner le travail du hub et le travail du réseau de catalogage :

1. Le hub effectue ses traitements en série
2. Le hub identifie les éléments d'informations et les notices pour lesquels il n'a pas su prendre une décision,
    a. soit parce qu'il n'avait rien à proposer,
    b. soit parce qu'il avait plusieurs propositions entre lesquelles hésiter.
3. Le catalogueur intervient sur les données pour les compléter ou arbitrer en cas de doute,
    a. soit dans WinIBW, après chargement des notices dans le Sudoc[41],
    b. soit directement dans le hub.

Cette situation 3b suppose que l'ABES possède une interface d'édition des données par les professionnels qui soit branchée directement sur la base RDF du hub. Cette hypothèse n'a pas été envisagée dans le cadre de cette étude, mais il s'agit sans aucun doute d'un point essentiel pour l'avenir.

[41] Même dans ce contexte MARC, le hub peut fournir une aide à la décision : on pourrait imaginer que le hub, cherchant à lier une notice bibliographique à IdRef mais hésitant entre deux autorités, injecte dans une zone spéciale des notices MARC les différents PPN candidats pour que le catalogueur puisse trancher à la main, dans WinIBW, grâce à l'aide d'un script qui lui soumettrait ces PPN.

Les thèses avant 1985

Le hub au service de theses.fr

theses.fr a vocation à signaler toutes les thèses de doctorat françaises, mais, pour des raisons pratiques, dans un premier temps, il s'est consacré aux thèses postérieures à 1984. En effet, en deçà de cette date, les notices de thèse présentent trop d'anomalies et, en l'état, leur intégration dans theses.fr serait nuisible à la qualité du service, voire impossible. Ces thèses antérieures à 1985 exigent donc un traitement spécial avant d'être chargées dans theses.fr et la présente étude est apparue comme une bonne occasion d'y travailler.

Que peut faire le hub pour ces métadonnées de thèse ? On peut distinguer trois niveaux de priorité :

1. Ce qui bloque l'entrée dans theses.fr
2. Ce qui perturbe la recherche et/ou l'affichage dans theses.fr
3. Ce qui perturbe la recherche et/ou l'affichage dans le Sudoc

Ce qui bloque l'entrée dans theses.fr (niveau 1), c'est essentiellement l'absence de numéro national de thèse (NNT). Cette absence est compréhensible puisqu'avant 1985, cette information n'était pas exigée. Il revient alors à l'ABES de générer ce NNT, qui est obligatoire dans theses.fr.

Ce qui perturbe la recherche et/ou l'affichage dans theses.fr (niveau 2), c'est :

- L'absence de lien aux autorités pour les auteurs et autres contributeurs (y compris les collectivités)
- L'absence de date
- L'absence de langue
- L'absence d'indexation matière
- Un nom d'établissement de soutenance non conforme
- Etc.

Toutes ces anomalies n'ont d'impact sur theses.fr que si elles figurent dans les notices de thèse originales. En effet, la procédure de chargement des notices Sudoc dans theses.fr cherche à regrouper toutes les notices décrivant différentes expressions[42] ou différentes

[42] au sens du modèle bibliographique FRBR

manifestations[43] de la même thèse, mais ne prend en compte que les métadonnées de la manifestation originale (celle que le jury a lue). Sous cet angle, si la notice de l'original contient les bonnes infos mais pas les notices de reproduction liées, tant pis – à savoir, tant pis pour le Sudoc. Mais si la notice de l'original ne possède pas ces informations tandis qu'une notice de reproduction les possède, tant pis cette fois pour theses.fr – la thèse ne peut être chargée.

Ce qui perturbe la recherche et/ou l'affichage dans le Sudoc (niveau 3), ce sont à peu près les mêmes anomalies que celles du niveau 2, mais indépendamment du type de notice de thèse dans lesquelles elles apparaissent : notices d'original ou pas.

Traitements possibles

Sur le papier, le hub pourrait s'attaquer à toutes les anomalies qui viennent d'être mentionnées.

La démarche générale consisterait à charger toutes les notices de toutes les thèses de doctorat et, dans un premier temps, à introduire une cohérence FRBR entre toutes les notices d'une même œuvre-thèse :

a. Identifier les notices d'une même œuvre-thèse via la procédure de chargement des notices de thèses – procédure extérieure au hub
b. Identifier les incohérences :
    - Exemple : indexation matière différente d'une notice à l'autre
    - Exemple : lien à l'autorité IdRef de l'auteur présente seulement dans la notice de l'original
c. Lever les incohérences faciles :
    - Exemple : fusionner l'indexation matière provenant de chaque notice et injecter cette fusion dans chaque notice
    - Exemple : ajouter le lien à l'autorité Sudoc dans chaque notice

Après cette mise en cohérence, qui d'elle-même permet de résoudre certaines des anomalies, le hub chercherait à :

- lever les incohérences difficiles
    - Exemple : que faire quand l'auteur est différent d'une notice à l'autre ?
- combler les lacunes

[43] au sens du modèle bibliographique FRBR

- Exemple : générer le lien à l'autorité quand il manque à chaque notice, grâce au prototype SudocAD[44]
- Exemple : générer un NNT

Sur le conseil de l'équipe qui développe et gère theses.fr, l'étude s'est consacrée à la tâche suivante : générer un NNT quand il manque. Cette tâche est essentielle mais très ponctuelle : il s'agit seulement d'ajouter un nouveau triplet dans la base RDF, puis une zone UNIMARC dans le Sudoc. Si les traitements effectués avaient été plus nombreux, la question de la mise à jour du Sudoc en sortie du hub se serait posée avec plus d'acuité. Cette question sera approfondie plus loin.

Acquisition, modélisation et conversion des données

Ce corpus de métadonnées de thèses est un sous-ensemble du Sudoc, ce qui présente au moins deux intérêts pour l'étude :

- Travailler sur des données en MARC
- Travailler sur des données déjà présentes dans une base – en l'occurrence le Sudoc, qu'il s'agit donc de mettre à jour

Sur le papier, les étapes d'acquisition, de modélisation et de conversion des données promettaient d'être grandement facilitées par le fait que l'ensemble du Sudoc est déjà exporté et exposé en RDF. Hélas, cette sortie du Sudoc en RDF est assez pauvre en comparaison de la richesse du MARC natif. En attendant les progrès apportés par le chantier SudocRDF[45], il a fallu enrichir cette sortie pour les besoins de l'étude :

- Création de la propriété th:NNT
- Création des classes th:Originale, th:Reproduction, th:Remaniement
- Modélisation spécifique de la note de thèse (répétable): th:aPourValidationActe, th:degree, th:discipline, th:date328, th:etabSout328e
- Prise en compte de toutes les dates : th:issued210d, th:issued100, th:issued100_tout, th:date328
- Prise en compte de tous les points d'accès 7XX, y compris en l'absence de lien à une autorité IdRef
- Création de th:oaiset

[44] L'algorithme SudocAD exige que la notice de départ contienne un indice Dewey, ce qui est le cas pour les notices de thèse, depuis un important travail effectué en 2012 avec l'aide d'OpenRefine.
[45] Voir le glossaire

On le voit, l'enrichissement de la modélisation a essentiellement consisté à forger de nouvelles propriétés et classes dans un espace de noms lui-même forgé : « http://www.hub.abes.fr/namespaces/theses/ », abrégé en « th ». Ce qui compte, dans ce contexte interne, c'est la finesse et l'exactitude de la modélisation, et non l'interopérabilité en direction de l'extérieur, qui incite à réutiliser des vocabulaires existants.

Par ailleurs, cette étude ne s'intéresse qu'aux notices de thèses de doctorat françaises antérieures à 1985. Ce périmètre n'est pas si facile à délimiter :

- En MARC, il existe une donnée codée identifiant les thèses et mémoires (105 $a pos.4-7, valeur m ou v), mais pas spécifiquement les thèses de doctorat. Faute d'une telle précision, il a fallu s'appuyer sur une liste de libellés de diplôme censés correspondre à cette notion : « th. Doct », « thèse de docotorat » (sic), etc.
- En MARC, il existe plusieurs zones pour exprimer la date du document. On a retenu les positions 9-12 (ou 10-13) de 100$a. Mais cette information, censée ne contenir que des données de type année, contient toutes sortes d'anomalies qui empêchent de traiter ces quatre caractères comme une année AAAA (et donc d'y appliquer des opérateurs de comparaison de date par exemple) : " X ", "196X", "19XX", NULL (si la 100$a ne comprend que quatre caractères, il n'y a pas de données en pos. 9-13)... Il a donc fallu traiter ces quatre caractères comme une simple chaîne de caractères, à savoir sélectionner toutes les notices dont la date est strictement inférieure à '1985', au sens lexicographique (ASCII) et non au sens chronologique. On a également sélectionné les thèses dont la date est NULL. Hélas, on a laissé de côté un certain nombre d'anomalies qui, théoriquement, devraient être prises en charge[46].

On observe donc le paradoxe suivant : certaines notices qui présentent des anomalies ne peuvent être importées et traitées par le hub en raison même de ces anomalies. Ce paradoxe n'aurait pas lieu d'être si on avait la capacité de traiter tout le Sudoc dans le hub : tout le Sudoc serait dans le hub, y compris ces notices en anomalie.

En théorie, le nombre de notices de thèse de doctorat françaises à traiter s'élève à environ 300 000. Dans le cadre de cette étude, le travail n'a porté que sur un échantillon de 20 000 notices.

[46] Leur nombre est relativement négligeable.

Générer un numéro national de thèse (NNT)

Le NNT (numéro national de thèse) est le numéro d'identification de la thèse, attribué par la bibliothèque de l'établissement de soutenance et composé de 12 caractères (10 caractères autrefois)

- les 4 premiers indiquent l'année de soutenance,
- les 4 suivants constituent le code de l'université ou de l'établissement,
- les 4 derniers correspondent à une séquence numérique ou alphanumérique d'enregistrement

Exemple :

Année de soutenance	Code court de l'établissement de soutenance	Chaîne de caractères alphanumérique
2012	*TOUR*	*4001*

Deux thèses ne peuvent avoir le même NNT.

On constate qu'un NNT n'est pas un numéro arbitraire. Il est chargé en informations qui, par ailleurs, sont des métadonnées de thèse essentielles :

- L'année de soutenance
- L'établissement de soutenance

En travaillant à générer les NNT qui manquent, on travaille aussi sur ces deux informations qui sont à la fois importantes et souvent mal renseignées dans les notices. S'il suffisait seulement de trouver dans la notice le code de l'établissement d'un côté, l'année de soutenance de l'autre et de les concaténer avec une chaîne alphanumérique, la tâche serait triviale. Or l'essentiel du travail va consister à identifier le bon établissement et la bonne date à partir d'informations présentes, mais lacunaires, non conformes à une convention, mal structurées voire inexactes.

L'effort s'est concentré sur les notices d'original : assigner un NNT aux notices d'original sans NNT, à partir des informations contenues dans ces seules notices. On aurait eu intérêt à exploiter également les informations présentes dans les autres notices de la même œuvre-thèse, comme évoqué plus haut.

L'année de soutenance

L'année de soutenance de la thèse peut, en théorie, apparaître en trois endroits de la notice MARC d'origine :

- Date de publication en 100
- Date de publication en 210$d
- Année de soutenance en 328$d

Pour être en mesure d'exploiter toutes ces sources d'information, la conversion de MARC vers RDF s'est attachée à prendre en compte toutes ces zones, quitte à forger des propriétés *ad hoc*, pour les besoins de l'analyse interne au hub :

MARC	RDF
Date de publication en 100	th:issued100
Date de publication en 210$d	th:issued210d
Année de soutenance en 328$d	th:date328

Une fois les notices MARC modélisées en RDF, converties puis chargées dans la base RDF du hub, il devient très facile de rechercher les anomalies et des moyens d'y remédier.

Une première méthode consiste à extraire toutes les valeurs possibles de chacune de ces propriétés RDF. C'est le travail de SparqlTour, qui lance autant de requêtes Sparql qu'il y a de propriétés dans un corpus RDF de la base. Comme une propriété peut contenir un nombre très important de valeurs distinctes, qu'il serait difficile de passer en revue à l'œil nu, SparqlTour peut ne s'intéresser qu'à des sélections de 1 000 valeurs : les 1 000 valeurs les plus fréquentes, les 1 000 valeurs les moins fréquentes, les 1 000 premières valeurs selon l'ordre alphabétique (croissant et décroissant), etc. Très vite, en parcourant ces listes de valeurs, on a la confirmation que la zone 210 est celle qui contient le plus de chaînes de caractère différentes de la date elle-même :

```
Impr.1974

1973, cop. 1969

1973-1974

c. 1975

1976, DL 1975

imp. 1976
```

```
1974, achevé d'impr. 1976
1973, achevé d'impr. 1975
1973, achevé d'impr. 1974
1975, achevé d'impr. 1976
1978, cop. 1971
[circa 1975]
DL1978
impr. 1978, cop. 1977
1977, impr. 1978
cop.1977
[19..]
```

SparqlTour permet également de comprendre certaines spécificités dans la zone 328$d de l'année de soutenance :

```
[1954]
1879-1880
1904-1905
30 cm
1913-1914
1884-1885
[1976 ?]
198?
1974 ; 28
1983 ; 123
. 1969.  Ì±
```

On **voit** qu'on ne peut prendre pour argent comptant l'information trouvée en 328$d. Il faut notamment :

- Séparer les années (format AAAA) des autres informations (ex : « 1974 ; 28 », « [1954] »)
- Choisir une année parmi toutes celles d'une période (ex : « 1913-1914 »)
- Ignorer les erreurs irrécupérables (ex : « 30 cm »)

Une autre approche permet de compléter et d'affiner ces analyses. En effet, étant donné que toutes ces dates sont exprimées dans la base RDF, des requêtes SPARQL plus ou moins sophistiquées permettent de comparer les valeurs de ces différentes zones pour chaque notice, tout en extrayant au passage l'information de type Année :

these	100debut	100fin	328debut	328fin	date328brute	210debut	210fin	date210brute
http://www.sudoc.fr/005283612/id	1896		1895	1986	1895-1986	1896		1896
http://www.sudoc.fr/008831270/id	1977		1977	1982	1977 ; 1982	1977		1977
http://www.sudoc.fr/00472982X/id	1974		1974	1974	1974 1974 ; 47	1974		[1974
http://www.sudoc.fr/010023488/id	1972	1973	1972	1973	1972-1973	1972	1973	1972-1973
http://www.sudoc.fr/010322434/id	1959	1960	1959	1960	1959-1960	1959	1960	1959-1960
http://www.sudoc.fr/005474639/id	1955		1952	1953	1952-1953	1955		1955
http://www.sudoc.fr/00319261X/id	1945	1949	1942	1949	1942-1949	1945	1949	1945-1949

Ce tableau aide à mesurer l'ampleur des anomalies et à concevoir un algorithme qui identifie la bonne année de soutenance quand elle n'est pas donnée.

Autre exemple : la requête suivante permet de sortir toutes les valeurs de 328$e qui ne sont pas une année AAAA :

```
PREFIX xsd:  <http://www.w3.org/2001/XMLSchema#>
select str(?o) {
  ?s <http://www.hub.abes.fr/namespaces/springerB/theses/date328> ?o
  filter ( datatype(?o) = xsd:gYear)
  filter ( coalesce(xsd:datetime(str(?o)), '!') = '!')
}
```

Ce qui donne :

```
1970, soutenue sous le titre : "Essai sur le mode de développement dualiste"
[1972 ?]
[19..]
[1932]
19XX
19XX
[s.d.]
[s.d.]
[s.d.]
[s.d.]
1893-1894
```

Grâce à de nombreuses requêtes de ce genre, un algorithme a pu être mis au point et implémenté :

- Si 328$d = 100, on retient la 328$d sans chercher plus loin (cas où une seule année dans chaque champ)
- Si pas de zone 328$d, on va la chercher dans la notice (liée) d'une autre version/reproduction[47], sinon on retient la zone 100,
    - Si zone 100 est une période (AAAA-AAAA) on retient la seconde année AAAA
- Si présence d'une zone 328 : Signaler une erreur (bloquante ou non) si elle est différente dans une notice liée (autre version / reproduction) ?
  Sinon :
    - Si une seule année en zone 328$d, on la retient
    - Si la zone 328$d est une période (AAAA-AAAA), on retient la seconde année AAAA

Les cas qui soulèvent une erreur ne seront pas traités par le hub, mais listés et signalés aux établissements concernés.

L'établissement de soutenance

En l'absence de NNT, pour forger le code court, le hub peut s'appuyer sur deux autres zones MARC censées désigner l'établissement de soutenance :

- Dans la note de thèse, la zone 328$e, qui doit désigner l'établissement de soutenance conformément à une liste fermée documentée dans le GM (= table des libellés).
- La zone 712 avec « 295 » (= établissement de soutenance) pour code de fonction et un lien à l'autorité IdRef .

En théorie, connaissant le libellé 328$e autorisé ou le PPN de l'établissement de soutenance, on devrait pouvoir en déduire le code court, qui doit être conforme à une liste des codes courts autorisés. Cette déduction serait possible s'il existait un document qui mette en correspondance ces trois informations :

Code court	Libellé 328$e	PPN

[47] L'idée d'aller chercher une information dans les autres notices de l'œuvre-thèse quand elle est absente de la notice de l'originale n'a pas été implémentée dans les requêtes SPARQL écrites pour l'étude, mais on retrouve cette approche FRBR dans le traitement d'autres corpus : Dalloz, ebooks Springer.

C'est ce document que l'ABES devrait créer et maintenir, sous le nom de **référentiel des établissements habilités**.

Pour amorcer la pompe, l'équipe du hub a commencé par fusionner les deux listes du GM, de façon à obtenir ce premier tableau :

Code court	Libellé 328$e

Dans un second temps, l'équipe STAR a exporté un tableau qui a cette structure :

Code court des établissements STAR	PPN

Après fusion de cet export avec le tableau précédent, on obtient le tableau recherché, même incomplet :

Code court	Libellé 328$e	PPN

Mais on sait que ces données ne sont pas figées et évoluent avec le temps :

- les organismes vont et viennent, se scindent ou fusionnent ;
- les habilitations à délivrer le doctorat peuvent aller d'un organisme à un autre (ex : PRES) ;
- les noms des organismes peuvent changer ;
- etc.

Il faut donc ajouter une colonne au référentiel qui précise la période de validité de cette mise en correspondance { code court / libellé de 328$e / PPN }.

Code court	Libellé 328$e	PPN	Validité[48]

En l'état, le référentiel est en chantier :

- Il manque tous les établissements de soutenance historiques (ex : scission de l'Université de Paris après 1968, etc.)
- Il manque quelques établissements actuels
- Il manque la plupart des périodes de validité

[48] La période de validité doit être formatée de manière à être utilisée par des humains et des programmes.

- Il manque des PPN
- Certains cas complexes sont à débrouiller (ex : ENS Lyon)
- Etc.

Pour les besoins immédiats de l'étude sur le hub, en l'état, le tableau permet de travailler. Mais il serait souhaitable de profiter de cette ébauche pour aller plus loin en complétant, en maintenant durablement et en publiant ce référentiel, par exemple dans un tableur. Cet objectif ne relève pas du hub. Mais le hub peut y aider (voir plus loin).

L'architecture RDF du hub va grandement faciliter l'exploitation de ce référentiel. La démarche est la suivante :

- Le fichier .xls du référentiel est chargé dans Open Refine
- Grâce au plugin RDF développé par DERI pour Open Refine, le référentiel est très facilement exporté en RDF (dans un vocabulaire *ad hoc* forgé pour l'occasion)
- La version RDF du référentiel est chargé dans la base RDF du hub, dans un graphe à part

A ce stade, le référentiel des établissements habilités et les données qu'il est censé normaliser sont intégrés dans la même environnement de travail. Une même requête SPARQL peut alors travailler sur ces deux graphes en même temps. Par exemple, la requêtes suivante permet de lister les thèses originales sans NNT dont ni le libellé (328$e), ni l'autorité IdRef (712$3) de l'établissement soutenance ne sont conformes au référentiel :

```
select *
FROM <http://www.hub.abes.fr/theses/referentiel/etabsout>
FROM <http://www.hub.abes.fr/theses85_jan13>
where {
?th1 dc:title ?ti1.
?th1 a <http://www.hub.abes.fr/namespaces/theses/Originale>.
?th1 <http://www.hub.abes.fr/namespaces/springerB/theses/aPourValidationActe>
?valacte.
?valacte <http://www.hub.abes.fr/namespaces/springerB/theses/etabSout328e>
?label328defacto.
?th1 <http://www.loc.gov/loc.terms/relators/dgg> ?dgg.

# sans 328e conforme au referentiel
#FILTER NOT EXISTS {?etab <http://www.hub.abes.fr/theses/etab/label>
?label328defacto}

# sans NNT
FILTER NOT EXISTS {?th1 <http://www.hub.abes.fr/namespaces/springerB/theses/nnt>
?nnt1}

# sans idref pour l'etab de sout (dgg) conforme au referentiel
FILTER NOT EXISTS {?etab2 <http://www.w3.org/2002/07/owl#sameAs> ?dgg}

}
```

Ce genre de requête révèle toute l'utilité de l'approche RDF : avec des efforts de modélisation modestes et quasiment aucune intervention d'un informaticien, une requête SPARQL assez simple permet d'obtenir un résultat immédiat qui demanderait beaucoup plus d'efforts et de technicité avec une approche classique. Cette aisance donne une grande liberté à l'expert bibliographique pour manipuler les données sous différents angles, faire des hypothèses et les tester dans la foulée, sans délai ni intermédiaire.

Après avoir analysé la situation aux moyens de ce genre de requêtes SPARQL SELECT, l'expert bibliographique peut lui-même tester d'autres requêtes SPARQL qui vont corriger les données. Tout en restant dans la base de production, il peut effectuer ces corrections « en vrai » et en masse, mais dans un graphe à part, qu'il peut supprimer une fois ses tests terminés. Ainsi, la requête suivante effectue un certain nombre d'opérations :

- Elle s'intéresse seulement aux thèses originales sans NNT
- Elle identifie les thèses dont le libellé d'établissement de soutenance (328$e) est conforme au référentiel
- Pour chacune de ces thèses, elle sélectionne dans le référentiel le code correspondant au libellé
- A partir de ce code court, elle crée un nouveau triplet contenant ce code, et ce dans un graphe à part appelé « http://www.hub.abes.fr/theses85_jan13_TEST »

```
INSERT INTO <http://www.hub.abes.fr/theses85_jan13_TEST>
{ ?th1 <http://www.hub.abes.fr/namespaces/springerB/theses/nntOK> ?nntOK .
}
WHERE
{ GRAPH <http://www.hub.abes.fr/theses85_jan13>
{
?th1 dc:title ?ti1.
?th1 a <http://www.hub.abes.fr/namespaces/theses/Originale>.
?th1 <http://www.hub.abes.fr/namespaces/springerB/theses/aPourValidationActe>
?valacte.
?valacte <http://www.hub.abes.fr/namespaces/springerB/theses/etabSout328e>
?label328defacto.
}

GRAPH <http://www.hub.abes.fr/theses/referentiel/etabsout>
{
?etab <http://www.hub.abes.fr/theses/etab/label> ?label328defacto.
?etab <http://www.hub.abes.fr/theses/etab/codeNNT> ?nntOK
}

}
```

Grâce à cette requête SPARQL UPDATE, l'expert bibliographique peut effectuer d'autres requêtes pour évaluer cette modification en masse. Que cette évaluation soit positive ou non,

il peut à tout moment effacer ce graphe de test pour repartir à zéro ou lancer la mise à jour en vrai, dans un graphe de production.

Dans les faits, étant donné que l'algorithme global qui doit générer les NNT manquants ne peut tenir en une seule requête, c'est un programme en Virtuoso/PL qui l'a implémenté. L'algorithme est le suivant :

1. S'il existe un libellé d'établissement de soutenance (328$e)
    a. Si ce libellé est conforme au référentiel, ce dernier nous donne le code recherché
    b. Sinon, on s'intéresse à l'autorité IdRef du point d'accès établissement de soutenance (712$e avec $4=295) :
        i. Si l'autorité est conforme au référentiel, ce dernier nous donne le code recherché
        ii. Sinon, le programme s'arrête en laissant une trace qui dit qu'il faut aligner à la main avec l'aide d'Open Refine
2. Sinon (= pas de libellé 328$e), on passe directement à 1b.

Quand le programme trouve le bon code, en 1a et 1bi, cela permet non seulement de commencer à construire le nouveau NNT, mais également d'améliorer les données de manière systématique en imposant une cohérence entre les différentes manières d'exprimer une même information, en l'occurrence l'établissement de soutenance :

- Si un élément au moins manque dans le trio {code court dans le NNT, libellé 328$e, point d'accès}, il est facile de compléter la notice.
- Si les trois éléments sont présents mais en contradiction avec ce que prévoit le référentiel, il faut gérer une exception : comprendre d'où vient l'erreur et la corriger, soit dans le hub, soit en signalant le problème à la bibliothèque concernée.

En 1bii, le programme s'arrête et passe le relais à une procédure manuelle : il s'agit de mettre en correspondance les libellés présents dans les notices et les libellés du référentiel. Souvent, les deux libellés sont proches mais pas identiques : casse différente, chiffres romains au lieu des chiffres arabes (« Rennes 1 » / « Rennes I »), coquilles, etc. Il serait possible d'essayer de régler ces problèmes de manière entièrement automatique, mais Open Refine présente l'avantage de permettre des opérations automatiques tout en laissant l'expert vérifier ce qui se passe et trancher à la main s'il le faut.

Il est intéressant de noter comment Open Refine et la base RDF du hub peuvent s'articuler pour effectuer le travail d'alignement :

1. Une requête SPARQL extrait de la base RDF les libellés non conformes, sous la forme d'un fichier .csv

2. Ce fichier .csv est chargé dans Open Refine
3. On normalise la forme des libellés (exemple : remplacement des chiffres romains par des arabes)
4. On lance une opération de « réconciliation » qui compare les valeurs de ce tableau et (via SPARQL) celle du référentiel stocké dans la base RDF
5. On peut décider d'accepter automatiquement les propositions d'alignement qui sont supérieures à un certain seuil et arbitrer les autres à la main
6. On réinjecte enfin le résultat du travail effectué dans Open Refine dans la base RDF (via SPARQL)

En fait, la situation est plus complexe car il faut prendre en compte la dimension historique du référentiel des établissements de soutenance. En effet, le libellé « Paris » ne peut être aligné sur le même établissement selon qu'il est trouvé dans une thèse de 1950 (avant la scission de l'Université de Paris) ou dans thèse de 1980. Pour prendre en compte cette dimension, au moins deux stratégies sont envisageables :

- Appliquer la procédure décrite à l'instant mais ne pas prendre en compte le résultat pour les établissements qui posent problème. Mais avec la récente vague de fusion des universités, les cas problématiques risquent de devenir majoritaires.
- Avoir recours à une solution d'alignement plus complexe, qui sache traiter des données en RDF et effectuer des comparaisons en mixant différents critères. Il s'agirait d'aligner l'occurrence d'un libellé dans telle notice de thèse en prenant en compte à la fois le libellé et la date de soutenance. Il est possible d'effectuer ce genre d'opération dans SPARQL, à condition d'ajouter au langage SPARQL une fonction qui permet de mesurer la proximité de deux chaînes de caractère. C'est aussi la vocation d'un outil comme SILK, qui a été testé pendant l'étude.

Quelle que soit la procédure d'alignement entre les données et le référentiel, elle restera dépendante de la qualité du référentiel. Or, on sait que le référentiel des établissements habilités doit être complété pour les établissements historiques, qui n'existent plus ou ne sont plus habilités depuis des décennies parfois. Ce n'est pas au hub de compléter puis de maintenir ce référentiel, mais il peut aider. En effet, il peut mettre à la disposition du gestionnaire de référentiel des moyens d'analyse et de visualisation des données telles qu'elles sont.

Ainsi, la requête SPARQL suivante regroupe les thèses par établissement de soutenance, en s'appuyant sur l'autorité liée à la notice avec le rôle « établissement de soutenance » (712$e avec $4=295). Pour chaque établissement, elle indique le nombre de thèses par an et le nombre de thèses total (sur l'échantillon de 20 000). L'ensemble des établissements est trié par le nombre de thèses total. Mais, pour chaque thèse, la distribution annuelle des thèses est triée dans l'ordre chronologique.

```
SELECT  ?etab ?name(sql:GROUP_CONCAT(CONCAT(?date,"x",str(?timesOneYear)), ', ') AS
?howManyTimesEachYear) SUM(?timesOneYear) as ?cbn
WHERE {
{ SELECT ?etab ?name ?date (COUNT(?th) AS ?timesOneYear)
WHERE {
?th <http://www.loc.gov/loc.terms/relators/dgg> ?etab.
?etab foaf:name ?name.
?th dc:date ?date .
}
GROUP BY ?etab ?name ?date
ORDER BY ?date  }
}
GROUP BY ?etab ?name
ORDER BY DESC(?cbn)
```

Etablissement	nom	Combien de thèses / an	Combien en tout
http://www.idref.fr/034526110/id	Université de Paris (1896-1968)	1928x1, 1937x1, 1947x1, 1948x2, 1951x1, 1958x1, 1962x4, 1963x2, 1964x5, 1965x13, 1966x27, 1967x68, 1968x87, 1969x181, 1970x164, 1971x21, 1972x2	581
http://www.idref.fr/026404796/id	Université Joseph Fourier (Grenoble)	1969x1, 1970x9, 1971x34, 1972x41, 1973x62, 1974x53, 1975x47, 1976x20, 1977x18, 1978x29, 1979x36, 197Xx1, 1980x35, 1981x42, 1982x33, 1983x31, 1984x34	526
http://www.idref.fr/027542084/id	Université Paris Diderot - Paris 7	1971x19, 1972x27, 1973x31, 1974x40, 1974-1975x1, 1975x100, 1976x51, 1977x51, 1978x41, 1979x11, 1980x17, 1981x12, 1982x11, 1983x13, 1984x5	430
http://www.idref.fr/026388820/id	Institut national polytechnique (Toulouse)	1975x7, 1976x4, 1977x23, 1978x48, 1979x46, 1980x49, 1981x60, 1982x52, 1983x81, 1984x47	417
http://www.idref.fr/026404184/id	Université Lille 1 - Sciences et technologies	1971x12, 1972x32, 1973x43, 1974x12, 1975x17, 1976x22, 1977x13, 1978x3, 1979x6, 1980x39, 1981x66, 1982x71, 1983x26, 1984x22	384

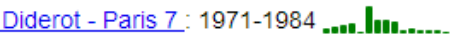
A partir d'un tel tableau, servi par le serveur SPARQL en XML, JSON ou d'autres formats faciles à exploiter, on peut générer ce genre de représentations, qui peuvent aider à la décision :

abes
agence bibliographique de l'enseignement supérieur

Histogramme sur un échantillon de 20 000 thèses avant 1985

581 thèses Université de Paris (1896-1968) : 1928-1972 

526 thèses Université Joseph Fourier (Grenoble) : 1969-1984 

430 thèses Université Paris Diderot - Paris 7 : 1971-1984 

417 thèses Institut national polytechnique (Toulouse) : 1975-1984 

384 thèses Université Lille 1 - Sciences et technologies : 1971-1984 

Les ebooks Dalloz

Acquisition, modélisation et conversion

Les 872 ebooks Dalloz étudiés correspondent à la collection d'ebooks diffusés sur la plateforme Bibliothèque numérique Dalloz (BND)[49]. Via le groupement de commandes ABES/Couperin, s'abonner à cette plateforme, c'est s'abonner à tous les ebooks qu'elle diffuse, y compris les nouveautés mises en ligne après la signature du marché – nouveautés qui sont donc absentes de la liste de titres jointe au document du marché ABES-Dalloz 2012-09[50]. De ce fait, la comparaison des titres de cette liste et de la liste des documents effectivement accessibles ne revêt pas la même importance que pour la licence nationale Springer, qui porte sur une liste fermée de titres - qui sont définitivement acquis.

Néanmoins, il est important de pouvoir à tout moment obtenir une liste à jour des titres de la Bibliothèque Dalloz, ce que devrait permettre le serveur OAI-PMH de la plateforme[51]. D'ailleurs, il serait possible d'automatiser la procédure de moissonnage et de chargement des métadonnées dans la base RDF, de rejouer cette procédure chaque mois et d'identifier ainsi les nouveautés et les retraits sur la BND.

Les métadonnées XML moissonnées en OAI-PMH combinent des éléments Dublin Core et des éléments ONIX[52], selon une DTD hybride hébergée par la BnF[53]. Le vocabulaire Dublin Core possède une interprétation RDF officielle, mais pas encore le vocabulaire ONIX. La modélisation et la conversion des éléments du vocabulaire ONIX se sont donc avérées plus délicates, d'autant plus que Dalloz en fait un usage très approximatif. Enfin, étant donné les nombreuses redondances entre les éléments Dublin Core et les éléments ONIX, on s'est souvent contenté de convertir une seule des deux occurrences de la même information.

Analyse

Une analyse superficielle permet rapidement de conclure que ces métadonnées fournies par l'éditeur sont non seulement pauvres, mais également fautives :

[49] http://www.dalloz-bibliotheque.fr/
[50] http://www.couperin.org/negociations/liste-des-negociations/download/2516/258/15
[51] https://logistic.book-vision.com/services/oai/act68.php (Il resterait à s'assurer de ce que cette base OAI-PMH reflète bien le contenu de la base dalloz)
[52] Voir le glossaire
[53] http://bibnum.bnf.fr/ns/onix_dc.xsd

- L'URL d'accès est inutilisable, en raison d'un mauvais encodage des caractères.
- ONIX est parfois mal interprété par Dalloz :
    - Les hyperliens ne sont pas associés au bon élément ONIX
    - L'ISBN électronique et l'ISBN papier ont été inversés
- Les sujets sont concaténés dans un même élément, comme s'il provenait d'une page web d'affichage.

Face à cette source médiocre, on est tenté de partir à la recherche d'une autre source, décrivant les mêmes documents.

Dériver des triplets du Sudoc, au niveau expression

Les ebooks Dalloz en question sont probablement des éditions électroniques d'œuvres qui ont préexisté sous forme imprimée. Et ces éditions imprimées ont probablement déjà été cataloguées dans le Sudoc. En conséquence, la démarche la plus naturelle et la plus efficace serait d'identifier les notices Sudoc de ces éditions imprimées et d'en faire des notices d'édition électronique, quitte à puiser à la marge dans les notices d'ebooks récupérées chez Dalloz pour compléter certaines informations.

Or, on dispose du tableau qui liste les ebooks Dalloz[54] en précisant leur ISBN imprimé et leur ISBN électronique. Grâce au web service Sudoc isbn2ppn, il est aisé d'identifier la notice Sudoc du livre imprimé correspondant à l'ebook qui nous intéresse.[55]

Hélas, si on s'en tient à ces informations, il est impossible de déterminer avec certitude la relation entre l'ebook et l'édition imprimée. S'agit-il d'une différence de contenu ou de forme ? S'agit-il de deux manifestations de la même expression ou bien de deux expressions de la même œuvre ? S'il s'agit de deux manifestations de la même expression, un grand nombre de métadonnées sont communes aux deux notices : mêmes contributeurs, même langue, même date de copyright. S'il s'agit de deux expressions de la même œuvre, les métadonnées en commun sont moins nombreuses.

[54] http://www.couperin.org/negociations/liste-des-negociations/download/2516/258/15 (il s'agit d'un document contractuel)

[55] A condition de chercher à la fois l'ISBN10 et l'ISBN13, même quand les métadonnées de départ ne contiennent qu'une des deux formes. Sans cet effort de conversion entre les deux ISBN, on passe à côté de beaucoup de notices.

Dans le doute, le hub doit raisonner prudemment, et se contenter de postuler qu'il s'agit de deux expressions de la même œuvre, dans la même langue. On ne peut donc conclure que ce qui suit :

- Les deux livres doivent avoir la même indexation matière.
- Ils ont probablement certains auteurs en commun (mais ils n'ont pas forcément exactement les mêmes auteurs).
- Ils ont la même langue.

La stratégie de « dérivation » de ces informations de la notice de l'imprimé dans les données de l'ebook est la suivante :

- Charger dans la base RDF les métadonnées RDF des notices Sudoc des livres imprimés mentionnés dans la liste Couperin
- Construire à la volée des notices d'ebooks enrichies, qui puisent à la fois dans les données d'ebook moissonnées chez Dalloz et dans les notices créées par le réseau Sudoc pour décrire les imprimés

La *dérivation des triplets* de langue ne pose pas de problème, si ce n'est la nécessité de gérer dans le hub un référentiel des langues qui s'accommode des codes sur deux caractères et des codes sur trois caractères.

La dérivation des triplets d'indexation pose problème en ce qui concerne les vedettes RAMEAU. Si le vocabulaire RAMEAU est depuis plusieurs années disponible en RDF (dans IdRef et surtout désormais dans data.bnf.fr), on ne peut que déplorer l'absence de bonnes pratiques pour exprimer en RDF les vedettes elles-mêmes, avec leur tête de vedette et leurs subdivisions qui s'enchainent dans un ordre précis. Aujourd'hui, on est incapable de traduire en RDF toute la structure de l'information complexe contenue dans une vedette construite. *A fortiori*, on ne sait pas faire une conversion aller-retour de type MARC➔RDF➔MARC sans perte. Face à cette difficulté, les options suivantes sont sur la table :

- Renoncer à la fidélité à l'information RAMEAU de départ en ne s'intéressant, par exemple, qu'à la tête de vedette. Cela voudrait dire que les notices Sudoc des ebooks de l'œuvre O contiendraient une indexation RAMEAU simpliste et bien plus pauvre que l'indexation RAMEAU de l'imprimé de la même œuvre.
- Renoncer à charger les vedettes dans la base RDF et manipuler directement les données de la base MARC pour recopier les zones RAMEAU des notices d'imprimé vers les notices d'ebook. Cette solution résoudrait le problème pour la base Sudoc, mais pas pour les autres débouchés.
- Forger une modélisation de l'UNIMARC en RDF qui ne s'intéresse qu'à la syntaxe MARC et non à son interprétation. Voir plus loin.

La dérivation des triplets indiquant les auteurs doit être abordée de la même manière que la récupération des contributeurs Worldcat pour les ebooks Springer. Mais, en l'espèce, il serait peut-être plus judicieux de commencer par l'approche SudocAD, qui, appliquée aux données Dalloz, propose un lien trois fois sur quatre.

On aurait aimé pouvoir établir un lien Sudoc entre la notice de l'ebook et la notice de l'imprimé. Hélas, puisqu'on ignore la nature exacte de la relation entre les deux documents, il est impossible d'établir un quelconque lien 4XX entre eux. Par contre, beaucoup de vocabulaires non MARC, comme le Dublin Core, prévoient des propriétés dénotant une relation générique – ce qui est toujours mieux qu'une absence de relation ou qu'un lien erroné.

Enfin, une dernière piste d'enrichissement a été envisagée, mais elle est restée inexplorée. Chaque ebook Dalloz correspond à une page Web qui contient des informations utiles mais absentes des métadonnées exposées en OAI-PMH :

- La table des matières
- Le nom de la collection

Si on est capable d'observer une régularité dans le code HTML, le hub pourrait moissonner et analyser ces pages afin d'en extraire ces deux informations et de les injecter dans la base RDF. Mais cette nouvelle source d'information révèle surtout le fait que le système d'information de Dalloz contient des informations qui ne sont pas exportées dans les données qu'ils partagent. Là encore, c'est à cette source première que le hub doit pouvoir remonter, pour en savoir plus sur les documents, et sans le filtre d'une conversion qui ne fait qu'augmenter les risques d'erreur.

Les articles revues.org

L'étude était censée se pencher sur le corpus des articles du portail revues.org . Or, faute de pouvoir accéder aux données[56], aucune analyse n'a pu être menée sur ces articles avec les outils du hub.

C'est d'autant plus regrettable que revues.org est une plateforme publique, avec laquelle l'ABES pourrait envisager un partenariat autour des métadonnées, dans l'esprit du hub :

- Fourniture des métadonnées brutes de revues.org à l'ABES
- Traitement par le hub et, en complément, par le réseau Sudoc
- Redistribution des données enrichies et formatées, en direction des sorties habituelles du hub (dont le Sudoc) et de la plateforme revues.org elle-même.

Néanmoins, on a pu appliquer à un échantillon d'articles le programme SudocAD de liage aux autorités de personne physique IdRef.[57] Le résultat est sensiblement identique aux conclusions de l'évaluation menée en 2011 sur les données Persée. En ce qui concerne la *revue d'économie industrielle*,

- 51 % des auteurs sont liés automatiquement.
- 96 % de ces liens sont corrects.
- Les erreurs de liage ont été analysées et comprises.
- Les améliorations qui seront apportées par le projet en cours Qualinca[58] permettront d'augmenter le taux de liage et de diminuer le taux de liage erroné[59].

[56] en raison d'un bug bloquant du côté du serveur OAI-PMH de revues.org, comme l'a reconnu récemment un représentant du CLEO. A cette occasion, on a appris qu'un nouveau serveur OAI-PMH serait bientôt mis en service.
[57] Rapport final du projet SudocAD : http://www.abes.fr/Sudoc/Projets-en-cours/SudocAD
[58] http://www.lirmm.fr/qualinca
[59] Il faut noter que ce taux de liage erroné reste acceptable car il est comparable à celui des catalogueurs en situation ordinaire, comme l'a montré l'étude SudocAD. Mais l'objectif est bien de diminuer encore ce taux.

Conclusions

#1 Faisabilité et utilités du hub

L'étude conclut que le hub de données serait en mesure de rendre les services attendus :

- Acquisition des métadonnées, par toutes sortes de moyens
- Analyse des métadonnées
  - Exactitude
  - Complétude
- Normalisation
- Correction
- Enrichissement
  - Dérivation de triplets à partir d'autres sources - RDF ou non
  - Génération de nouveaux triplets à partir d'informations existantes ou d'une intervention humaine
- Redistribution, sous différentes formes et par toutes sortes de canaux

Par ailleurs, l'étude Pleiade sur la stratégie de découverte de la documentation électronique[60] a levé l'hypothèque sur le consentement des moteurs de recherche verticaux académiques (comme les *discovery tools* ou Google Scholar) à intégrer des métadonnées enrichies par d'autres :

> ***Enriched metadata from an open platform would be included in the centralised indexes of the discovery tools via match & merge mechanisms.***[61]

Le hub apparaît donc comme une stratégie pertinente pour :

- Faciliter le signalement et la découverte de corpus entiers de documents électroniques
- Diminuer le catalogage initial à la main tout en maintenant des exigences de qualité

Pour autant, il ne faut pas imaginer que le hub rendra caduc ou même marginal le catalogage au sein du réseau Sudoc.

D'une part, fatalement, les traitements automatiques ont leurs limites : ils ne peuvent pas tout, surtout quand on s'oblige à un certain niveau de qualité. Au-delà, les traitements manuels ou

[60] http://fil.abes.fr/2013/03/29/etudes-sgbm-et-dispositif-de-decouverte-publiees/
[61] p.21

semi-automatiques du hub eux-mêmes trouveront leur limites, car les ressources humaines internes à l'ABES sont faibles. Le recours aux catalogueurs du réseau Sudoc restera donc incontournable. Il faudra imaginer une manière d' articuler le travail centralisé du hub avec les efforts décentralisés des bibliothèques.

D'autre part, il s'agit d'une approche artisanale, au sens d'un traitement sur mesure qui demande du temps et de l'application. Tous les corpus de documents ne peuvent recevoir un tel soin de la part de l'ABES. Celle-ci doit sélectionner les corpus qui justifient ces efforts, en fonction de critères qui restent à fixer : documents achetés ou documents loués, taille du corpus, spécificités françaises, etc. La contribution de l'ABES à cette mission de signalement doit être envisagée dans le cadre très large d'une coopération nationale et même internationale.[62]

#2 Architecture technique basée sur les technologies du web sémantique

Plutôt que d'évaluer en surface plusieurs solutions techniques, l'équipe hub a préféré faire le pari d'étudier en profondeur une solution basée sur les technologies du web sémantique. Rétrospectivement, l'intuition était bonne :

- Dans une base RDF performante, la gestion des données est d'une grande aisance et d'une grande souplesse, ce qui permet de se concentrer sur les données elles-mêmes.
- Les bibliothécaires peuvent jouir d'une grande autonomie, car :
    - les données sont stockées telles quelles dans la base RDF : on manipule les données telles qu'elles ont été pensées ;
    - le langage SPARQL est puissant.
- Le web de données étant en forte croissance, de plus en plus de données sont disponibles en RDF et donc, telles quelles, disponibles pour venir interagir avec les données du hub.
- La banalisation des technologies du web sémantique se traduit par l'arrivée sur le marché de logiciels matures capables d'exploiter les données RDF du hub.

Par ailleurs, au commencement de l'étude, on pouvait craindre que les bases RDF ne puissent pas supporter facilement les volumétries en jeu. Dans l'ensemble, la volumétrie n'a jamais été un obstacle définitif. Pour autant, cette question de la volumétrie se poserait certainement si on envisageait de charger tout le Sudoc dans une base RDF.

[62] C'est aussi la stratégie de l'étude Pleiade, qui recommande de collaborer avec le JISC (GB) et au-delà.

Cependant, d'autres craintes initiales se sont vu confirmer :

- Quitter le monde MARC ou XML pour le monde RDF suppose un investissement humain important.
- Les standards W3C du web sémantique sont encore jeunes : la version 1.1 de SPARQL, indispensable, est une recommandation depuis mars 2013 seulement.
- Le fait d'analyser les données après leur conversion en RDF pose problème : certaines anomalies constatées peuvent tenir à la modélisation ou à la conversion, et non aux données d'origine. On ne peut donc pas se contenter d'outils de diagnostic en RDF. Il est nécessaire d'effectuer certaines tâches de diagnostic sur les données MARC ou XML elles-mêmes. Mais le reste des opérations doit porter sur les données en RDF.

L'étude a permis de valider l'hypothèse RDF et de construire un prototype, mais on ne peut pas s'en tenir là :

- Il faut mettre en place un véritable environnement de production, avec la même base de données ou pas.
- Il faut approfondir la maîtrise des technologies du web sémantique. En particulier, le prototype n'a pas du tout exploré l'aspect inférence[63], qui est pourtant l'une de leurs grandes forces.
- Il faut mettre en place des outils de diagnostic qui portent sur les données en XML et en MARC. Ne serait-ce que dans le cadre des achats en licence nationale, c'est dans leur format d'origine que les données livrées doivent être validées.
- Il faut choisir l'environnement dans lequel interagir avec la base RDF : langage de programmation intégré à la base (PLSQL), autre langage (Java), scripts (Python), etc. Aucune solution optimale n'a été trouvée dans le cadre de l'étude : les scripts Python sont trop simples et le PLSQL rend dépendant de la base.

#3 Automatisation et qualité des données

L'idée du hub est bien d'appliquer sur des métadonnées préexistantes des traitements de masse automatiques, mais il faut aussitôt préciser que :

- ces traitements automatiques sont conçus, développés et testés à la main ;
- il peut s'agir de traitements *ad hoc*, qu'on ne peut appliquer tels quels à chaque nouveau corpus ;

[63] Il s'agit de faire raisonner les données, au sens logique.

- l'automatisation a ses limites :
    - un traitement automatique ne parvient jamais à traiter avec succès 100% des cas ;
    - certains types d'interventions ne peuvent pas être automatisés.

S'il existait des méthodes automatiques, universelles et infaillibles pour normaliser et enrichir toutes les données bibliographiques, la question de ce hub de métadonnées ne se poserait pas : les métadonnées en circulation seraient déjà parfaites. Mais ce n'est pas le cas. D'où la pertinence du hub, foyer de production de métadonnées bibliographiques de qualité, dans le cadre d'un effort global en faveur d'un espace de données ouvertes et fiables. Depuis quelques années, cet espace s'est matérialisé sous la forme particulière du web de données. Même s'il existe d'autres modalités de partage des données, notamment dans notre contexte métier (les bases de connaissance, par exemple), l'alimentation du web de données sera l'une des missions principales du hub.

Or, qu'il s'agisse du web de données ou de la constitution de bases de connaissance, les problèmes à régler en priorité par les acteurs en place sont les mêmes. Après une phase primitive d'accumulation massive des données, il s'agit de passer de la quantité à la qualité :

- **Sélection des corpus**. On atteint aisément de grandes masses en s'intéressant aux bases de données majeures, peu nombreuses mais volumineuses. Pourtant, la longue traîne des petites bases de données représente au total une masse importante. La prise en compte de cette longue traîne ne peut être l'affaire de quelques grands opérateurs, mais plutôt le résultat de l'action (plus ou moins coordonnée) de dizaines d'acteurs moyens comme l'ABES.
- **Qualité des données**. Là encore, s'il est relativement aisé de traiter en masse des centaines de millions de notices à l'origine hétérogènes, c'est souvent au prix de l'appauvrissement des données initiales, voire d'un taux d'erreur ou d'approximation élevé. Or, dans le contexte globalisé des bases de connaissances et du web de données, où la réutilisation des données est la norme, les problèmes de qualité se propagent et même s'amplifient. Chacun doit balayer devant sa porte, au bénéfice de tous. Pour atteindre un niveau de qualité supérieur, il faut en passer par des efforts ciblés et souvent *ad hoc*.

Ces problématiques justifient l'approche du hub : le sur-mesure. Elles justifient également le recours à des traitements manuels. Ainsi, les métadonnées de périodiques, d'articles, de collections, de livres et de chapitres fournies par Springer s'appuient toutes sur un vocabulaire d'indexation propriétaire. Par des méthodes automatiques très simples, on parvient à mettre en correspondance des concepts de ce vocabulaire et les concepts des LCSH (ou de la classification Dewey). Avec des méthodes plus sophistiquées, le taux de mise en

correspondance augmenterait encore. Mais, par de telles méthodes automatiques, on n'atteindra jamais 100% de couverture avec 100% de fiabilité. Par ailleurs, la mise au point de ces traitements automatiques demande du temps. C'est pourquoi, dans des cas semblables, il semble tout à fait justifié de recourir à une méthode entièrement manuelle : traiter à la main quelques centaines de concepts permet d'améliorer la description de millions de documents.

Enfin, l'approche du hub est *ad hoc* en un dernier sens : de nombreuses propriétés et classes sont forgées en tant que de besoin. Au cours de cette étude, l'expérience a montré que chercher à tout prix à dénicher dans un vocabulaire établi la propriété ou la classe qui exprime parfaitement une notion de départ est un processus fastidieux, qui débouche trop souvent sur la frustration ou l'approximation. Face à une notion très spécifique qui ne correspond pas immédiatement à un élément de vocabulaire connu, on doit sans scrupule forger une propriété ou une classe *ad hoc*, dans un nouveau vocabulaire RDF propre au hub. A l'usage, si cet élément *ad hoc* s'avère réutilisable d'un corpus du hub à l'autre, c'est que la notion en question n'était pas si spécifique et on découvrira tôt ou tard un vocabulaire du marché qui l'exprime. Il sera alors aisé de remplacer en masse notre propriété *ad hoc* par cet élément standard[64]. On gagnera en interopérabilité, mais sans effort inutile.

On comprend donc que ce qui compte dans la modélisation, c'est l'exactitude de l'analyse conceptuelle et non le respect superficiel des étiquettes standard. C'est pour cette raison que le hub n'aura pas recours à des approches qui essaient d'automatiser le processus de modélisation (« alignement d'ontologies »).

#4 Le hub et l'ABES

Le hub ne signerait pas un changement de cap pour l'ABES. Au contraire, il s'inscrit parfaitement dans son histoire et son cœur de métier :

- réutiliser et standardiser des métadonnées préexistantes ;
- coordonner la production de métadonnées[65] ;
- à partir de ces métadonnées, fournir des services aux professionnels en premier lieu.

De plus, la démarche du hub s'intègre bien dans la stratégie actuelle de l'ABES, en cohérence avec les autres développements qui concourent à cette stratégie :

[64] ou bien on précisera (en RDF) la relation entre notre propriété *ad hoc* et la propriété standard.
[65] Ce qui ne signifie pas seulement produire de nouvelles notices : l'essentiel du travail consiste déjà à compléter ou reprendre des métadonnées préexistantes.

- **Evolution des standards métier**. Les nouveaux standards bibliographiques s'appuient sur RDF, soit comme modèle conceptuel, soit comme syntaxe, souvent les deux. C'est le cas de RDA et de Bibframe. Certes, la priorité du hub est de modéliser les données fournies dans toute leur richesse initiale, quitte à renoncer à l'emploi de vocabulaires RDF standards. Mais cette convergence autour de RDF simplifie le plan de formation des agents et favorise une culture professionnelle commune au sein de l'ABES et de ses réseaux. Ajoutons qu'avec le hub, la maîtrise des technologies du web sémantique ira un cran plus loin : il ne s'agit plus seulement de modéliser nos données en RDF pour les exposer ou les exporter, mais de les gérer en RDF (requêtes, mises à jour).
- **Exposition des données**. L'ABES est engagée dans une politique d'ouverture des données. Cette politique se traduit notamment par le chantier « SudocRDF », qui vise à exposer les données du Sudoc dans un RDF de plus en plus riche, qui, à terme, n'écarte aucune information contenue dans les données MARC initiales. Cet effort et la démarche du hub sont complémentaires : si l'ABES devenait capable d'exprimer toute la richesse des données Sudoc en RDF, ces données pourraient s'intégrer parfaitement dans l'infrastructure du hub, enrichir d'autres données et y être enrichies, avant d'être réinjectées dans le Sudoc de façon simple et efficace – ce qui n'est pas possible aujourd'hui (cf. paragraphe suivant).
- **ISTEX**. Dans le dispositif ISTEX, le hub devrait se positionner principalement entre le début de la négociation et la phase de diffusion au sein de la plateforme ISTEX. (cf. plus loin)
- **Stratégie de découverte**. Le rapport de Pleiade Consulting a montré comment le hub pourrait tenir une place importante dans la stratégie de découverte de la documentation électronique qu'il recommande.
- **Système de gestion des bibliothèques mutualisé (SGBM)**. Face aux risques de dépendance à l'égard d'un système mutualisé au sein d'un cloud privé, le hub fera partie de la palette d'outils par laquelle l'ABES et ses réseaux s'efforcent de conserver une réelle maîtrise sur les données et les workflows.
- **SudocAD et Qualinca**. Le liage automatique aux autorités est une des activités principales du hub.

#5 Un point difficile : synchroniser le hub et le Sudoc

Le hub et le Sudoc peuvent entretenir 3 types de relation :

- **Le Sudoc comme débouché**
- **Le Sudoc comme source d'enrichissement**
- **Le Sudoc comme corpus à traiter**

Une des vocations premières du hub est de servir d'intermédiaire entre des corpus de métadonnées livrés à l'ABES et des outils de signalement publics, dont le Sudoc, entre autres **débouchés**. Appliquée aux corpus ISTEX, cette idée signifie que les métadonnées livrées à l'ABES ne seront pas directement chargées dans le Sudoc – ce qui signifie aussi que le format MARC ne devrait pas être exigé des fournisseurs. Ces métadonnées seront traitées par le hub, puis redistribuées vers différents outils en aval, dont le Sudoc. C'est le hub qui aura la responsabilité de produire des données MARC correctes, et non les fournisseurs.

Ce circuit est très classique, mis à part le fait que cet intermédiaire qu'est le hub s'appuie sur RDF et qu'il prétend enrichir les données, et pas seulement les convertir. Et ce circuit classique rencontre une difficulté, elle aussi classique : la question des mises à jour. Au chargement initial, en général, tout est simple : le hub génère des notices MARC qui sont chargées dans la base centrale du Sudoc. Au-delà, tout est plus compliqué : il devient nécessaire de fusionner les nouvelles informations venant du hub et celles déjà présentes dans le Sudoc.[66] Il s'agit d'une question classique, mais dans un environnement nouveau puisqu'il ne s'agit plus de fusionner entre elles deux notices MARC. On peut imaginer au moins trois approches :

- **Approche classique**. Le hub génère à nouveau des notices MARC et c'est au Sudoc de fusionner la nouvelle notice et l'ancienne, en fonction de règles de fusion qui auront été définies par l'équipe Sudoc conjointement avec l'équipe hub.
- **Approche micro-update**. Si le hub prend soin de documenter tous les traitements qu'il applique sur les données concernées ainsi que l'impact de ces modifications sur les données du Sudoc, alors le hub pourrait tirer profit de l'API Sudoc développée par l'ABES pour mettre à jour le Sudoc de manière chirurgicale. Le fait d'effectuer ces

[66] Cette question se pose également entre le hub et n'importe quel autre débouché. Mais la question est particulièrement sensible dans le cas du Sudoc car, entre deux mises à jour, les notices peuvent être modifiées par n'importe quel catalogueur, *sans qu'on puisse tracer ce qui a été modifié*. En soi, le modèle RDF n'apporte pas de lui-même une amélioration sur ce point. Mais la normalisation des graphes nommés et de la notion de provenance par le W3C fera bientôt de cette traçabilité des données RDF une fonctionnalité essentielle des bases RDF. Mieux encore : cette traçabilité de la *vie des données RDF dans une base* sera elle-même un ensemble de données, qui pourra *suivre les données hors de leur base* d'origine !

modifications à petite dose permettrait de mieux contrôler les opérations et d'identifier les dysfonctionnements.

- **Approche tout-hub**. Tout se passe dans la base RDF du hub. Le temps du traitement, on verrouille les notices Sudoc à modifier, on les exporte en RDF, on les charge dans le hub, on les fusionne avec les nouvelles données et on génère à nouveau des notices MARC qui écrase les anciennes.

L'approche classique est envisageable, mais fait peser la charge de travail sur l'équipe Sudoc et perpétue la dépendance à l'égard du système CBS.

L'approche tout-hub a le mérite de la cohérence du point de vue de la démarche générale du hub, mais elle suppose d'avoir une conversion Sudoc➔hub qui soit exhaustive, de façon à permettre des allers-retours Sudoc⇔ hub sans perte d'informations. Cette conversion n'existe pas aujourd'hui et sera difficile à produire : les données Sudoc sont riches, complexes et, fatalement, en raison des strates historiques et des aléas humains, elles contiennent des irrégularités.

L'approche micro-update semble aujourd'hui être la stratégie la plus raisonnable, mais elle demande une étude approfondie, en particulier sur les points suivants :

- Cette approche est-elle compatible avec le fonctionnement actuel du Sudoc, qui permet aux établissements de récupérer dans leur système toutes les modifications ou seulement les leurs ?
- D'un point de vue technique, comment optimiser voire automatiser ces requêtes de mise à jour entre le hub et le Sudoc ? Et comment mettre en place un mécanisme qui soit assez générique pour survivre au remplacement de CBS par un autre système, y compris dans les nuages ? Au cours de l'étude, une nouvelle méthode a été esquissée, qui s'appuie sur le standard W3C XQuery.

On peut noter que cette question des mises à jour du Sudoc ne se pose pas seulement pour les notices de documents électroniques. En effet, à l'occasion du chargement de métadonnées décrivant des documents électroniques, on peut être amené à vouloir modifier les notices des éditions papier de ces mêmes documents. Il peut s'agir tout simplement d'ajouter un lien de la notice du papier vers la notice de l'électronique – ce qui est précieux. Il peut également s'agir d'enrichir la notice papier de manière plus substantielle, à partir d'informations nouvelles fournies par l'éditeur – comme l'étude l'a montré à propos du corpus Dalloz, ce qui peut aussi contribuer à renforcer une cohérence des données Sudoc dans un esprit FRBR. A ce stade, on n'est pas loin de l'idée que le Sudoc lui-même peut, en tout ou partie, devenir un **corpus à traiter** par le hub. Ce qui, encore une fois, serait facilité par le fait de disposer d'une conversion Sudoc➔hub qui soit exhaustive.

Enfin, l'exemple Dalloz a montré que le Sudoc devait également être conçu comme une **source d'enrichissement** pour les corpus à traiter par le hub. Pour ce faire, l'idéal serait de pouvoir disposer d'une version complète du Sudoc au sein même du hub. Cela supposerait une base RDF capable de supporter la masse du Sudoc et, une fois encore, une conversion Sudoc➔hub qui soit exhaustive.

#6 Tendre vers une conversion Sudoc➔hub qui soit exhaustive

On aura compris que le chantier SudocRDF, qui vise à exprimer chaque information contenue dans le MARC en RDF, est un effort difficile et indispensable, en général mais en particulier pour le hub.

En attendant les progrès de ce chantier, deux pistes alternatives méritent d'être étudiées.

ISO2709 en RDF

Convertir UNIMARC en RDF suppose d'analyser chaque élément d'information, d'imaginer une structure RDF qui rende compte de sa signification, de réutiliser ou de forger les vocabulaires RDF qui exprimeront cette structure et enfin de rédiger le script qui implémentera la conversion … avant de se rendre compte que les données d'origine ne respectaient pas toujours l'UNIMARC théorique, ce qui génère des erreurs de conversion.

Une autre approche consisterait à modéliser MARC (ISO2709) et non UNIMARC. Autrement dit, il s'agirait de modéliser les notion de zone, de sous-zone, d'indicateur et de position, et non la notion de 200$a ou de 100$a-positions0-7. Une telle modélisation aurait le seul mérite de permettre des aller-retour *sans perte d'information* entre la base Sudoc et la base RDF, ce qui n'est pas rien. Elle aurait seulement un usage interne : au moment d'exposer ou de partager des données à l'extérieur de l'ABES, le hub devrait modéliser l'UNIMARC sur le fond.

SolrTotal en RDF

Dans certaines circonstances, il n'est pas nécessaire de charger la totalité d'une notice Sudoc dans la base RDF. On peut n'avoir besoin que de certaines informations ponctuelles, mais à l'échelle de plusieurs millions de notice. Il serait à la fois inutile, difficile et coûteux de chercher à convertir la totalité de ces millions de notices en RDF.

Une alternative pragmatique consisterait à exploiter Solr Total, moteur de recherche Solr interne à l'ABES qui indexe en temps réel toutes les sous-zones et toutes les positions de toutes les notices bibliographiques Sudoc. Tout le Sudoc est donc reproduit dans Solr Total, sous la forme de plus de deux mille index. Par exemple, si le hub avait besoin d'exploiter le titre de toutes les notices Sudoc, il lui suffirait d'exploiter l'index 200$a_exact de SolrTotal. Il n'aurait plus qu'à charger quelques millions de triplets de la forme suivante :

```
http://www.sudoc.fr/126225141/id    solrt : 200$a_exact    « Bayard »
```

Ce détour par SolrTotal présente deux avantages principaux :

- Sélectionner les notices dont on veut extraire quelques zones en utilisant toute la force d'un moteur de recherche (croisement des index, opérateurs divers et variés)
- Sélectionner de manière très fine les zones à charger dans le hub sans passer par une extraction à partir de la notice entière, qui supposerait un traitement trop long et trop coûteux pour une opération ponctuelle et ciblée

Une telle solution peut rendre de grands services, tant qu'on garde à l'esprit les limitations suivantes :

- Solr est un moteur de recherche, pas une base de données. Il structure les données dans le seul objectif de faciliter la recherche. Ainsi, il aplatit les données en faisant éclater les couples de zones ou de sous-zones.
- Ces triplets issu de Solr Total ne servent que les besoins internes du hub : ils n'ont pas vocation à être réinjectés dans le Sudoc.

#7 Le hub ne sera pas seulement une base

Même si le cœur du hub est une base de données RDF, le hub est autant un ensemble de méthodes et de programmes pour interagir avec cette base :

- préparer des données non RDF pour qu'elles puissent être chargées dans la base ;
- analyser les données de la base ;
- modifier en masse les données de la base ;
- extraire les données de la base et les redistribuer

#8 Choisir une base RDF de production sans en devenir dépendant

C'est la version Open Source du système de base de données Virtuoso qui a été utilisée pendant l'étude. Ce produit s'est avéré robuste, commode et capable de rester performant avec de gros volumes. Mais cette expérience positive ne signifie pas que le hub utilisera Virtuoso en production. L'étude a permis à l'ABES de mieux comprendre ses propres besoins, et c'est à cette aune que la décision devra être prise, en s'appuyant sur les benchmarks publics disponibles. Quel que soit le choix, l'ABES devra veiller à ne pas dépendre étroitement du système retenu (langage de programmation propriétaire, extensions propriétaires à SPARQL).

#9 Le hub ne sera pas une seule base

Le hub traitera différents corpus, mais ces corpus peuvent rester étrangers les uns aux autres. Que pourrait apporter la fusion des données Dalloz et des données Springer ? Le hub n'est pas une grande base qui agrège tout, mais une chaîne de traitement avec des lignes de production parallèles.

Par contre, chacun des corpus a besoin de corpus complémentaires qui vont aider à les enrichir :

- des référentiels : référentiels de langues, de pays, d'auteurs, de revues, de vocabulaires d'indexation, etc. ;
- des bases bibliographiques : Sudoc, Worldcat, etc.

Aujourd'hui, pour les exploiter dans l'environnement du hub, il est préférable de les charger dans la base RDF – dans des sous-ensembles appelés « graphes ». Se posent alors les questions de la mise à jour de ces corpus extérieurs et de la volumétrie : jusqu'où aller ? faut-il charger tout le web de données dans le hub pour pouvoir en tirer partie ?

Il existe d'autres raisons de partitionner la base RDF du hub en sous-bases (en « graphes »). En effet, il semble prudent de ne pas mélanger les données reçues par le hub et les données qu'il génère lui-même. On peut même aller plus loin et dédier un graphe à chaque traitement qui génère de nouveaux triplets. Par exemple, on l'a vu au cours de l'étude, il peut y avoir

plusieurs manières de générer le lien à une autorité IdRef et il peut sembler judicieux d'isoler les triplets générés par chacun d'entre eux dans un graphe à part.[67]

#10 La base RDF s'auto-documente

La souplesse du modèle RDF permet d'utiliser la base de données non seulement pour manipuler les données mais également pour documenter ces manipulations elles-mêmes. Dans une telle base auto-documentée, il devient possible de :

- tracer plus facilement l'historique des traitements ;
- identifier plus facilement les données qui ont suivi tel traitement et déclencher des actions ciblées (mises à jour du Sudoc, envois de listings aux bibliothèques pour interventions manuelles irréductibles, etc.)

Aujourd'hui, ce souci de traçabilité de la vie des données RDF demande un effort particulier. A terme, cette traçabilité relèvera probablement du fonctionnement par défaut des bases RDF.

#11 Exploiter davantage les technologies du web sémantique

Dans le cadre de cette étude, les technologies du web sémantique ont été exploitées de manière modeste, mais efficace, en se concentrant sur :

- La possibilité de fusionner immédiatement des données exprimées en RDF
- Le langage de requêtes SPARQL

A l'avenir, il faudra aller plus loin et exploiter la capacité de raisonnement des bases RDF :

- pour améliorer la recherche ;
- pour valider l'intégrité des données au moyen de règles logiques

[67] Cette partition en graphes permet d'exporter une partie seulement des données générées, d'ajouter des métadonnées à un graphe (exemple : un degré de confiance, une date, etc.), de comparer le contenu de deux graphes l'un à l'autre, etc.

#12 Miser sur d'autres approches complémentaires

Au-delà des technologies du web sémantique et de l'exploitation des web services, au cœur du prototype, le hub devra élargir sa palette en ayant recours à des approches complémentaires comme :

- **Le traitement du langage naturel**. Il s'agit d'analyser le texte intégral afin d'en extraire des informations structurées.
  - Exemple : extraire des mots-clés à partir d'un résumé
  - Exemple : détecter la langue d'un titre ou d'un résumé[68]
- **La fouille de données**. Il s'agit d'appliquer des méthodes statistiques à des séries de données pour identifier des régularités.
  - Exemple : classer automatiquement des documents, en extrapolant à partir d'un échantillon classé à la main en Dewey
- **Le crowdsourcing**. Il s'agit de mobiliser un grand nombre de personnes pour effectuer des micro-tâches que les ordinateurs échouent à réaliser.[69]
  - Exemple : proposer à des personnes de choisir entre deux indices Dewey possibles quand l'ordinateur n'a pas réussi à décider lui-même

#13 Liage à IdRef

Rattacher des données brutes à des référentiels est une priorité du web. En particulier, lier des noms de contributeur à IdRef apporte une grande plus-value, d'autant plus que l'interconnexion d'IdRef avec VIAF permet de propager ce type d'enrichissements.

En attendant les avancées qu'on peut espérer du projet Qualinca (2012-2015), l'ABES dispose déjà d'un prototype, développé dans le cadre du projet SudocAD. Hélas, ce prototype n'a fait l'objet d'aucun effort d'optimisation systématique qui le rendrait utilisable sur de grands volumes et donc utile au hub. Cet objectif pourra difficilement être atteint avec les moyens du hub, mais l'équipe Qualinca de l'ABES bénéficie de ressources en développement depuis avril 2013.

[68] Les premiers tests effectués avec un web service d'AlchemyAPI sont très encourageants : http://www.alchemyapi.com/api/lang/

[69] Quelques essais techniques ont été effectués sur la plateforme de *crowdsourcing* http://crowdcrafting.org/ . Il va de soi que le défi n'est pas tant technique que stratégique et organisationnel : qui faut-il mobiliser pour intervenir sur de telles plateformes ? des amateurs de bonne volonté ? les professionnels de l'IST ? les chercheurs ? Et dans quelles interfaces ? les interfaces traditionnelles (interfaces de consultation, interfaces de catalogage) ou bien de nouvelles interfaces dédiées ? Et quels éléments d'information confier à ces arbitrages collectifs ?

Il existe deux voies pour avancer vers cet objectif :

1. Optimiser l'accès aux notices FRBRoo en effectuant a priori la transformation MARC➔RDF/FRBRoo et en mettant le résultat en cache
2. Réécrire l'application SudocAD en pré-calculant tout ce qui aujourd'hui est calculé à la volée par le programme CoGui du LIRMM

Recommandations

#1 S'engager dans le projet de construction du hub de métadonnées

L'étude sur le hub a montré la faisabilité et l'utilité d'un hub de métadonnées, tel que défini dans le projet d'établissement. Ce rapport final recommande d'engager le chantier du hub dès la fin de l'étude, afin d'éviter l'évaporation des compétences acquises et de suivre au plus près la progression des acquisitions ISTEX.

#2 Tout en construisant le hub, le faire tourner en production, de manière intégrée au fonctionnement régulier de l'ABES

Après l'étude, vient le temps du projet. L'étude a construit un prototype. L'étape suivante est de construire une application réelle, capable de traiter des données en production.

Pour autant, il n'est ni souhaitable ni nécessaire que le projet de hub soit achevé pour que des données réelles commencent à être traitées par le hub. Dès le début du projet, le hub devra être capable de fonctionner en production, même de manière partielle. Les premières données traitées devront donc l'être à nouveau au fur et à mesure du développement des possibilités du hub. Ce fonctionnement itératif est parfaitement cohérent avec la recommandation #3.

En termes d'organisation, le projet de hub devra tout de suite être intégré au fonctionnement régulier de l'ABES, notamment en ce qui concerne l'acquisition des ressources électroniques et de leurs métadonnées (Département Adèle), la modélisation des données (pôle META), le traitement des données et notamment des imports Sudoc (pôle PIT). Même si le hub aura ses propres méthodes, priorités et ressources, comme n'importe quelle application ABES, il est essentiel d'éviter d'en faire un projet à part.

#3 Toujours concevoir le hub comme une infrastructure à enrichir avec des approches et des outils innovants

Selon la recommandation #2, le hub produira des résultats effectifs dès la phase projet. Inversement, au-delà de cette phase projet, le hub devra demeurer en quête de nouvelles approches et de nouveaux outils pour améliorer son traitement des données.

Corollairement, un même corpus de métadonnées pourra être traité régulièrement par le hub, au fur et à mesure de l'amélioration de ses procédures. Ce fonctionnement itératif sera la règle, et non l'exception, même pour des données en principe figées comme les métadonnées d'archives.

#4 Veiller à ne pas faire du hub une boîte noire fonctionnant en vase clos. Articuler le hub avec des services extérieurs, dans le périmètre de l'IST française et au-delà

Dès la phase de prototypage, le hub a montré qu'il devait s'appuyer sur des services existants, notamment pour enrichir les données. Ces services peuvent émaner de l'ABES elle-même, du monde des bibliothèques ou d'autres acteurs. Le hub devra continuer à faire feu de tout bois pour apporter une plus-value aux données.

L'ABES devrait être particulièrement attentive aux perspectives de collaboration avec d'autres acteurs de l'IST en France, notamment dans le cadre d'ISTEX. S'il est décidé que le hub s'intéressera en premier lieu aux métadonnées des documents acquis dans le cadre d'ISTEX, la collaboration entre le hub et la plateforme développée par l'INIST pourra se matérialiser sous la forme d'allers-retours fructueux :

1. L'éditeur livre les données à l'ABES
2. Le hub effectue de premiers traitements sur ces données
3. L'ABES livre à l'INIST les données traitées par le hub. Eventuellement, le hub continue à traiter ces mêmes données.
4. L'INIST exploite le texte intégral des documents ISTEX et apporte sa propre plus-value aux métadonnées
5. L'INIST livre à l'ABES les plus-values générées à partir du texte intégral et des métadonnées
6. L'ABES intègre ces plus-values dans le hub, ce qui permet de nouveaux traitements, de nouveaux enrichissements, des corrections, etc.
7. L'ABES livre à nouveau à l'INIST les données générées par le hub
8. Etc.

Il faut ajouter que, par le seul fait d'exposer son travail sur le web de données, le hub contribuera à faciliter l'interopérabilité et la coopération au sein de l'IST française. Le détour par le web est souvent une manière efficace d'échanger des données de gré à gré, en échappant à la viscosité administrative.

#5 Impliquer la communauté des catalogueurs des réseaux ABES dans le fonctionnement du hub

Si le hub parvient à produire certaines informations qui sont traditionnellement créées par les catalogueurs, il ne rendra pas caduque l'activité de catalogage. Il faut imaginer comment le hub et les catalogueurs cohabiteront voire collaboreront. Cette question mériterait une nouvelle étude, mais voici quelques pistes :

- Quand le hub échoue à analyser, corriger ou enrichir certains éléments d'information, il devra veiller à identifier avec précision ces éléments et, si on considère que le jeu en vaut la chandelle, à soumettre aux catalogueurs les indications précises qui leur permettront d'améliorer les données sur ce point, en priorité dans WinIBW.
- Le hub pourrait fournir aux catalogueurs de nouvelles interfaces de travail pour effectuer ces « micro-tâches » (*crowdsourcing* professionnel).
- On pourrait imaginer un rôle encore plus actif pour certains catalogueurs : alignement de vocabulaires contrôlés *spécifiques d*ans tel ou tel domaine scientifique par des experts du domaine ; participation à la modélisation RDF ; contribution aux spécifications générales des traitements et à la définition de leur priorité ; test des résultats des traitements etc.

#6 Définir les priorités du hub, en termes de corpus à traiter et de traitements

Le hub sera essentiellement un ensemble de traitements appliqués à un ensemble de corpus. C'est donc en deux sens qu'il épousera une géométrie variable :

- Il pourra embrasser plus ou moins de corpus.
- Il pourra effectuer des traitements plus ou moins nombreux, plus ou moins complexes.

Il appartiendra donc à l'ABES de définir le plan de travail du hub, en identifiant les corpus et les traitements prioritaires.

#7 Définir les acquisitions ISTEX comme corpus prioritaires

Ce rapport final recommande de traiter en priorité les données issues d'ISTEX, pour les raisons suivantes :

- Le projet ISTEX est un projet stratégique, piloté au niveau de BSN et donc à l'échelle de l'IST française, et pas seulement des bibliothèques universitaires.
- Les négociations ISTEX sont d'ores et déjà attentives à la question des métadonnées.
- Les calendriers du projet de hub et du projet ISTEX peuvent coïncider parfaitement, si les premiers échantillons de métadonnées ISTEX sont livrées en septembre 2013 et que l'équipe hub est constituée à cette date – hypothèses toutes deux très vraisemblables.
- La plateforme ISTEX construite par l'INIST pourra bénéficier des résultats du hub.

Si cette recommandation est suivie, l'activité du hub devra suivre au plus près les progrès des négociations ISTEX, dans chacune de leurs phases :

- Avant l'acquisition ferme : sur la base d'échantillons, le hub devra aider à identifier précisément le périmètre du corpus négocié (liste des titres, états de collection, etc.) et évaluer la qualité des métadonnées associées.
- Après l'acquisition : le hub devra manipuler et enrichir les métadonnées livrées avant de les redistribuer à ISTEX – et vers les autres débouchés (Sudoc, licencesnationales.fr, bases de connaissance, outils de découverte, web de données, etc.)

Ce couplage fort entre la construction du hub et les négociations ISTEX exige de l'équipe hub une grande disponibilité et une grande réactivité. (voir plus loin la recommandation #9)

Au-delà des corpus ISTEX, les principaux types de corpus susceptibles d'être traités par le hub sont :

- Les métadonnées de la documentation électronique accessible sur abonnement
- Les bases de données gérées par l'ABES : le Sudoc lui-même - en tout ou partie -, Calames, theses.fr

#8 Définir l'identification précise des ressources et l'interconnexion avec des référentiels comme les traitements prioritaires du hub

Parmi les divers traitements qu'il sera capable d'effectuer, le hub pourrait se concentrer en priorité sur les traitements suivants :

- **Identification précise des documents** : identification précise des œuvres, expressions et manifestations (au sens FRBR), états de collection des publications en série, mention d'identifiants internationaux (ISBN et ISSN papier et électroniques, DOI, etc.), etc.
- **Alignement sur des référentiels** : référentiels de contributeurs (IdRef, VIAF), référentiels d'indexation matière (RAMEAU, LCSH, etc.), classifications thématiques (Dewey), typologies des documents, etc. Le fait d'aligner des données sur un référentiel reconnu présente plusieurs avantages précieux :
    - On ne réinvente pas la roue
    - On favorise l'interopérabilité avec d'autres données qui s'appuient sur le même référentiel
    - On favorise le multilinguisme (par exemple, l'indexation matière en français d'un corpus à l'origine indexé en anglais)

Dans le contexte de la gestion de la documentation électronique, le premier objectif contribuera surtout à la qualité des bases de connaissance, dont la priorité est l'accès. Le second objectif contribuera surtout aux index centraux des outils de découverte. En fonction des priorités stratégiques décidées suite à l'étude Pleiade, le hub devra lui-même ajuster ses priorités d'analyse et de traitement.

#9 Doter le hub d'un noyau de ressources humaines minimal

La construction et le fonctionnement du hub demandent des compétences pointues, qui, pour être maintenues, doivent être employées et renouvelées en permanence. Le fonctionnement du hub ne peut être irrégulier. De surcroît, si le hub doit être synchronisé avec ISTEX, il devra être réactif : ses forces pourront être mobilisées de manière ponctuelle et intensive.

Ces exigences de fonctionnement continu et de sur-mobilisation occasionnelle justifient les recommandations générales de ce rapport en termes de ressources humaines :

- Définir une équipe minimale en-deça de laquelle l'investissement dans le hub ne serait plus pertinent

- Eviter un émiettement des compétences hub sur trop de personnes, ce qui rendrait impossible l'acquisition d'une expertise poussée et la sur-mobilisation occasionnelle

L'équipe minimale se composerait de la manière suivante :

- Responsable fonctionnel : 0.75 ETP sur 1 personne
- Responsable informatique : 0.25 ETP sur 1 personne
- Développeur : 1.5 ETP sur 2 personnes
- Développeur XSLT : 0.5 ETP sur 1 personne
- Expert en données bibliographiques et RDF : 2 ETP sur 3 personnes

Ce noyau équivaut à 5 ETP sur 8 personnes.[70]

Au-delà de ce noyau, l'équipe du hub pourrait grossir de manière assez linéaire en fonction des ambitions que l'ABES y placerait, en termes de nombre et de complexité des corpus de métadonnées, en termes de niveaux de qualité et d'enrichissement visés et en termes de délais de traitement.

Au-delà de cette équipe, le hub devra pouvoir compter sur d'autres équipes ABES, et notamment l'équipe Qualinca, afin de bénéficier dès maintenant du traitement de liage à IdRef effectué par le prototype SudocAD.

#10 Multiplier les canaux de redistribution des données traitées par le hub

Le hub n'est pas une fin en soi, mais un moyen pour rendre disponibles et réutilisables des métadonnées de qualité.

Le catalogue Sudoc sera l'un des débouchés majeurs du hub, mais pas le seul :

- Bases de connaissance[71]
- Index centraux des moteurs de recherche verticaux (« outils de découverte »)
- ISTEX
- Web de données

[70] Suite au présent rapport, les ressources humaines allouées officiellement par la direction de l'ABES sont proches de cette recommandation. Les ETP « expert bibliographique » seront de 1.5, au lieu de 2.
[71] Il est probable que le hub n'alimentera pas directement les bases de connaissances du marché : il alimentera plutôt la base de connaissance nationale prévue par l'étude Pleiade, qui elle-même échangera avec les bases de connaissances extérieures, dans le cadre de la collaboration internationale initiée par GoKB.

- Services de récupération de lots de données pour les professionnels – via licencesnationales. fr ?[72]
- Les éditeurs sensibilisés à la qualité des données, qu'il s'agisse de favoriser la description, la recherche ou l'accès

Etant à la fois l'opérateur du Sudoc et du hub, l'ABES doit veiller à ne pas privilégier les besoins du Sudoc au détriment des autres débouchés. En adoptant une approche très souple basée sur RDF et indépendante de l'infrastructure technique du Sudoc (CBS) et de son format natif (MARC), le hub sera capable de générer des données sous une forme qui convienne à ses différents débouchés. Dans le dernier cas du Sudoc, le hub pourra lui fournir des données MARC plus fiables que celles qui sont habituellement livrées par les éditeurs.

#11 Demander aux éditeurs les données les plus riches et natives possible, plutôt que des données standardisées au risque d'un appauvrissement et d'une qualité dégradée

Jusqu'à présent, l'ABES demande aux fournisseurs de données – y compris aux éditeurs ou diffuseurs – des métadonnées conformes à UNIMARC ou MARC21, standards des bibliothèques. Or, ce mode de livraison demande aux fournisseurs qui ne sont pas des bibliothèques un effort qui s'avère souvent contre-productif, pour eux comme pour nous :

- La livraison demande un dialogue long, fastidieux, qui n'aboutit pas toujours
- La conversion en MARC appauvrit les données gérées en interne par le fournisseur
- La conversion en MARC est souvent approximative, voire fautive : non seulement les données peuvent violer les règles UNIMARC, mais elles peuvent devenir inexactes

Ce rapport final recommande donc fortement de ne plus exiger exclusivement des fournisseurs des données MARC. Ce qu'il faut obtenir, ce sont les donnés les plus riches et les plus brutes possibles, c'est-à-dire les plus proches de la base de données interne du fournisseur – à condition qu'elles soient documentées. Ensuite, il revient à l'ABES d'analyser, de modéliser et de convertir ces données pour les intégrer au hub et, en bout de chaîne, les redistribuer sous différentes formes, en préservant le maximum de richesse.

Comme aujourd'hui, ce circuit de récupération de données demande une phase de dialogue avec le fournisseur. Mais, d'une part, le dialogue devrait être facilité dès lors que c'est l'ABES qui s'adapte au fournisseur, et non plus l'inverse. D'autre part, ce dialogue peut

[72] Cette idée est développée plus haut.

prendre un tour nouveau : il faut inciter le fournisseur de données brutes à incorporer dans sa propre base les enrichissements et surtout les corrections effectués dans le cadre du hub. En effet, si les erreurs relevées et corrigées par l'ABES ne sont pas prises en compte par le fournisseur, elles reviendront dans chaque nouvelle livraison de ses données et obligeront l'ABES à prendre à chaque fois les mêmes précautions. Par ailleurs, ces corrections seront profitables à tous les outils alimentés par ce fournisseur (outils de découverte, moteurs de recherche, résolveurs de lien, etc.).

D'une manière générale, ce changement de stratégie à l'égard des fournisseurs de métadonnées ne consiste pas à relâcher ses exigences, mais à les concentrer sur des éléments d'information jugés cruciaux (les identifiants internationaux, par exemple), indépendamment du format.

#12 Placer toutes les données traitées par le hub sous la *Licence Ouverte*

L'ABES a décidé[73] d'exposer les données qu'elle traite dans les conditions juridiques prévues par la Licence Ouverte[74]. Cette stratégie d'ouverture explicite se heurte parfois au refus des fournisseurs de placer les données qu'il livre à l'ABES sous ce régime juridique.

Dès la livraison des données par le fournisseur, l'ABES devra s'assurer qu'il sera en mesure de placer sous ce même régime les données redistribuées par le hub. Dans le cas contraire, l'ABES devrait envisager de renoncer à les traiter.

A terme, la Licence Ouverte pourrait s'avérer trop contraignante. En effet, certains partenaires de l'ABES ont fait le choix de placer leurs métadonnées relatives à la documentation électronique dans le domaine public. C'est le cas de KB+[75].

[73] http://fil.abes.fr/2012/11/29/dimportantes-decisions-prises-par-le-conseil-dadministration-de-labes/
[74] http://www.etalab.gouv.fr/pages/Licence_ouverte_Open_licence-5899923.html
[75] http://www.kbplus.ac.uk/kbplus/publicExport

Glossaire

Graphe	Dans le contexte de cette étude, un graphe est un ensemble de données RDF regroupées sous une URI. Par ce mécanisme, il devient possible de décrire ce graphe comme un tout, par exemple pour préciser sa provenance ou sa date de création. En toute rigueur, on devrait ici parler de *graphe nommé*. (En savoir davantage sur les graphes nommés)
Linked data	« Modèle pour établir des hyperliens entre des données exploitables par un ordinateur au moyen des technologies du web sémantique, en utilisant notamment RDF et des URIs » (définition du glossaire W3C sur le Linked data[76])
ONIX	ONIX est un ensemble de standards XML qui vise à encoder et échanger des métadonnées de documents publiés. ONIX est une initiative des éditeurs. (Plus d'information sur ONIX)
RDF	RDF (*Resource description Framework*) est un « ensemble de standards internationaux produits par le W3C, en vue de l'échange des données sur le Web. [RDF] repose sur l'idée d'identifier toute chose au moyen d'identifiants Web, à savoir des URIs, et de décrire ces ressources en utilisant de simples propriétés et valeurs associées. » (définition du glossaire W3C sur le Linked data[77])
SPARQL	« *SPARQL Protocol and RDF Query Language* (SPARQL) est un langage de requête pour les données en RDF, comparable à the *Structured Query Language* (SQL) pour les bases de données relationnelles » (définition du glossaire W3C sur le Linked data[78])
SudocRDF (chantier)	Chantier interne à l'ABES qui vise à enrichir la sortie RDF des données du Sudoc. Ce chantier a commencé en septembre 2012. Il fait suite à un premier chantier qui avait abouti, en 2011, à un export minimal des notices Sudoc en RDF (plus d'information sur le blog de l'ABES, Punktokomo).

[76] Voir le glossaire
[77] Voir le glossaire
[78] Voir le glossaire

Annexes

Tableau comparatif des notices d'ebooks Springer

Voir le fichier TableauComparatifEbooksSpringer.xlsx

abes
agence bibliographique de l'enseignement supérieur

Liste et volumétrie des corpus Licences Nationales cibles

Corpus	Nombre de documents	Type de documents
JAMA and the Archives Journals Backfiles collection	10 revues, 421 500 articles	périodiques
American Physical Society (APS)	7 revues, 355 000 articles	périodiques
Annual Reviews Electronic Back Volume Collection (EBVC)	33 revues, 25 000 articles	périodiques
Brepols Miscellanea Online – Essays in Medieval Studies: Archive & Collection 2011	313 volumes, 6 200 articles	ebooks
Brepols Periodica Online - Archives	23 revues, 14 000 articles	périodiques
Brill Journal Archive Online	288 revues, 110 000 articles	périodiques
New Pauly Online	40 volumes, 36 000 entrées	
Recueil des Cours de l'Académie de la Haye en ligne (RCADI)	358 volumes, 1264 items	ebooks
Cambridge Journals Digital Archive (CJDA) Collections	242 revues, 730 000 articles	périodiques
De Gruyter eJournal Archives	474 revues	périodiques
Duke Mathematical Journal	1 revue, 6000 articles	périodiques
Emerald Management eJournal Archive	230 revues, 123 000 articles	périodiques
IOP Publishing Archives historiques des revues électroniques	48 revues, 356 000 articles	périodiques
Wiley-Blackwell Backfiles	940 revues	périodiques
Oxford Journal Archives	200 revues,	périodiques
RSC Journals Archive	76 revues, 260 000 articles	périodiques
SAGE Deep Backfile	468 revues	périodiques
BMJ Journals Archives	32 revues, 612 000 articles	périodiques
Elsevier Backfiles	2200 revues	périodiques