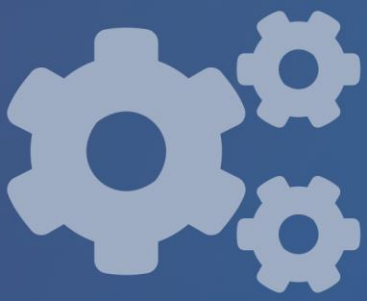
abes
agence bibliographique
de l'enseignement supérieur



Projet Sudoc 21

Rapport de fin de projet présenté
au Conseil Scientifique de l'Abes
Mars 2021

Ensemble, avec nos réseaux,
réinventons le service public des données

abes.fr

Sommaire

Introduction : genèse et circonstances du projet

Le projet Sudoc21 fait suite à une longue gestation : les premiers travaux sur un « Sudoc2 » pour prendre le relais du Sudoc existant datent de 2012. Le problème a été abordé de plusieurs manières au cours des années 2012-2018. Il a commencé par le souci de trouver un successeur au logiciel CBS que le fournisseur OCLC annonçait alors en fin de vie. La question a ensuite été couplée avec celle du SGBm : profiter du grand cycle de ré-informatisation des établissements pour faire basculer le Sudoc sur le même nouveau socle. Puis la publication du modèle IFLA-LRM en 2017 a mis un coup de projecteur sur ce qu'on a appelé « transition bibliographique » : le mouvement vers de nouvelles normes et pratiques de description pour rendre visible les contenus des catalogues sur le web de données et faciliter la recherche de l'usager. Parallèlement, le système d'information de l'Abes est devenu de plus en plus complexe : le nombre d'applications a augmenté ainsi que les synchronisations internes entre les bases de stockage hébergeant des données dans des formats différents : au côté du MARC du Sudoc se sont ajoutés l'EAD pour Calames, le TEF pour les thèses, le KBART pour les bouquets de documentation électronique, et le RDF pour le traitement du formidable afflux de métadonnées des licences nationales et d'ISTEX.

Le projet Sudoc21 a été lancé en avril 2019, selon une méthodologie Scrum Agile avec une équipe de 9 personnes mêlant informaticiens et bibliothécaires, dont deux responsables : Stéphane Rey, informaticien, jouant le *scrum master* ; Carole Melzac, bibliothécaire, jouant le *product owner*. Chaque membre de l'équipe a rejoint le projet avec une quotité de 40%, ce qui équivaut à 3,6 ETP. La durée prévue initialement était d'un an. Il est rapidement apparu que les trois preuves de concept envisagées pour tester différents outils nécessitaient une phase préliminaire qui s'est étendue jusqu'à l'été 2019. Elle a permis la cohésion de l'équipe, l'apprentissage d'une nouvelle méthode de travail, l'appropriation des éléments normatifs et de modélisation, et la définition des cas d'usage. La première preuve de concept s'est déroulée de septembre à décembre 2019. La seconde preuve de concept a démarré en janvier 2019. Le travail a été fortement perturbé par l'irruption de la pandémie de Covid-19, faisant basculer l'équipe en télétravail et en activité partielle avec la fermeture des écoles, puis par la mort brutale d'un des co-responsables du projet, Stéphane Rey. Le projet a continué malgré ces circonstances et cette deuxième phase a été terminée en septembre 2020. La troisième et dernière preuve de concept a occupé la période d'octobre 2020 à février 2021. C'est donc le bilan de presque deux années de travail, à neuf puis huit personnes qu'il s'agit de tirer ici.

1/ Retour sur l'objectif

Conformément à la méthodologie Agile, l'équipe a défini au cours de la phase préliminaire une « vision du produit » : *mettre à disposition de l'Abes et des établissements de ses réseaux un système permettant comme le CBS l'ingestion, le traitement, le stockage et la fourniture de données bibliographiques mais en le faisant de manière unifiée dans un environnement technique ouvert, adaptable, évolutif donc maîtrisable par l'Abes et en s'appuyant sur l'exhaustivité des données utiles au métier, structurées selon un modèle de données unique compatible LRM (fondé sur les tâches utilisateurs).*

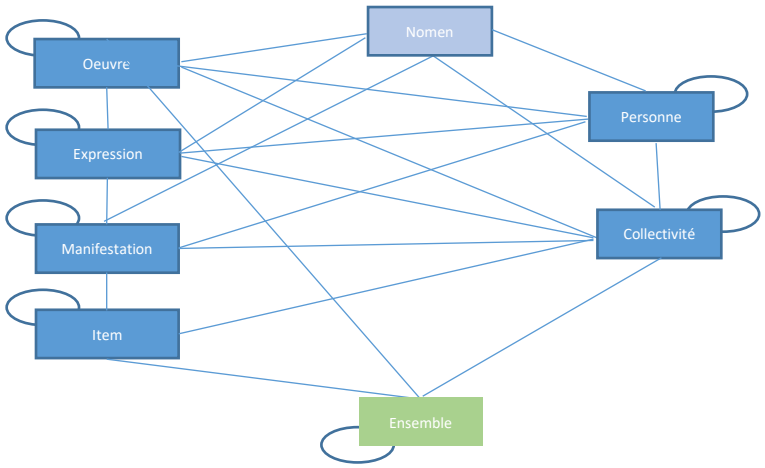
L'étude préalable d'avril 2018 avait produit un cahier des charges allégé listant les fonctionnalités et contraintes vu comme centrales en se focalisant sur celles nouvelles ou susceptibles de dissensus. La phase préliminaire est repartie de cette base pour construire les « *users stories* » ou historiettes et les organiser dans un tableau de suivi[1].

Outre ces cas d'usage à mettre en œuvre au sein des preuves de concept, le projet Sudoc21 a poursuivi deux objectifs associés : la connaissance et documentation de l'existant à l'Abes d'une part, et la prospection sur les outils et projets similaires en France comme à l'international d'autre part. Le lien entre ces trois dimensions a permis d'aboutir, si ce n'est à des réponses définitives, au moins à poser correctement les problèmes. Après ce projet, il apparait beaucoup plus clairement quelle sont les briques de l'architecture nécessaire et dans quel ordre peuvent s'articuler les actions pour le remplacement du SI.

2/ Les grands enseignements

Pour briser les silos, un modèle unifié

L'objectif du modèle unique était affiché dès la définition de la vision du projet et son importance s'est concrétisée à mesure de notre avancée. L'un des grands acquis de Sudoc21 est de pouvoir dire qu'il est possible et judicieux de réconcilier nos données en silos au moyen d'un modèle générique. Celui que nous avons développé est basé sur IFLA-LRM qui a été étendu selon nos besoins, en forgeant notamment l'entité « Ensemble » nécessaire à la description des bouquets (offres commerciales de documentation électronique), des fonds, des plans de conservation partagée, etc. Le schéma ci-dessous illustre le modèle simplifié des entités utilisées et des relations qu'elles entretiennent :



On peut préciser que LRM n'est pas un indispensable pour réconcilier les silos de données. Néanmoins, comme il constitue l'avenir des données aujourd'hui encodées en MARC, il était nécessaire de montrer qu'il n'y a pas d'obstacle à l'utiliser pour transformer les autres formats natifs. A ce titre nous rappelons que le format EAD utilisé pour encoder les données de Calames (archives et manuscrits) n'a pas pu être traité dans le cadre du projet. Toutefois,

[1] Via l'outil Kanboard et disponible à cette adresse : https://kanboard-prod.abes.fr/?controller=BoardViewController&action=show&project_id=110

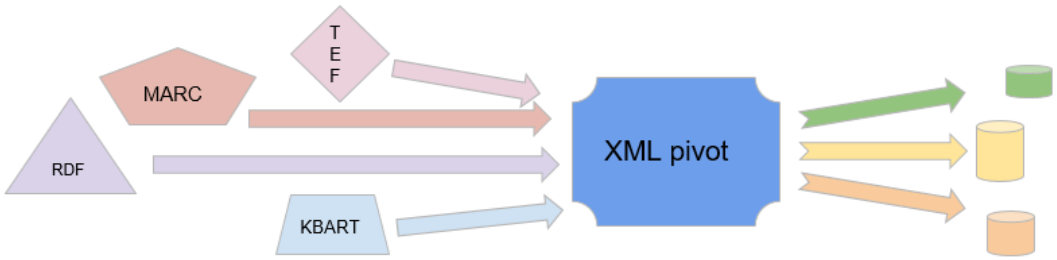
l'enchâssement qui le caractérise a été traité avec succès à partir du cas des ressources continues (bouquet, revue, fascicule, numéro, article) encodées dans les autres formats (MARC, KBART, RDF).

La mise sur pied du modèle a été un des premiers travaux théoriques, qui s'est concrétisé et développé tout au long du projet, avec des moyens divers car liés aux technologies de stockage testées. Lors de la troisième preuve de concept basée sur des bases RDF (triple stores) nous avons produit une ontologie, déduite des données. L'ontologie exprimée à l'aide des standards du RDF (OWL et RDFs) s'est révélée un outil précieux de manipulation et de mise en cohérence du modèle. Une ontologie RDF permet également de s'aligner avec un minimum d'efforts sur les vocabulaires existants.

Par conséquent, il nous apparait logique que le passage à l'échelle après les preuves de concept prenne comme point de départ cette ontologie afin de la compléter, de l'étendre pour les données non encore traitées, et de l'aligner avec les vocabulaires génériques souhaités.

Transformation des données

Une fois le modèle esquissé, une grande part du travail du projet Sudoc21 a consisté à « fabriquer » les données dont on voulait tester le stockage et la manipulation. C'est au cours de la première preuve de concept qu'a émergé l'idée d'un format XML pivot : un lieu agnostique où se trouveraient, décomposées depuis chacun des formats natifs, les informations relatives aux ressources. Ne se trouvent dans ce « pivot » que des entités et leurs relations, dans un vocabulaire ad hoc au projet. Sachant que nous aurions au moins trois types de bases à charger dans les preuves de concept, ce fonctionnement permet une économie, en séparant la transformation selon le modèle, de la pure production de données selon le format attendu.



Pour y parvenir, ce sont des XSLT[2] qui transforment les données natives, à raison d'au moins un par format initial. Ces programmes de transformation décomposent les documents initiaux à l'aide de raisonnements parfois complexes. La technologie XSLT a été choisie pour plusieurs raisons : la première, c'est qu'il était facile d'obtenir des documents XML pour tous les formats natifs. La seconde, c'est que ce mécanisme souple et plutôt lisible du point de vue des fonctionnels permet de nombreux allers/retours et un fonctionnement en mode essai/erreur plutôt que la rédaction de longues spécifications. La troisième, c'est qu'il s'agit d'un standard robuste qui peut ensuite être appelé indépendamment du langage. Enfin, il existait dans

[2] Programmes traitant des données structurées en XML :
https://fr.wikipedia.org/wiki/Extensible_Stylesheet_Language_Transformations

l'équipe deux personnes dotées d'une très bonne maîtrise de la technologie, qui a été abondamment utilisée à l'Abes depuis plusieurs années.

La conversion des informations issues des formats source a pu être soit totale (pour le KBART et le RDF), quasi-totale (pour le TEF) ou partielle (pour le MARC).
Concernant le MARC, nous avons choisi l'échantillon type mis à disposition pour les fournisseurs de SGB, représentatif des caractéristiques des données Sudoc : il s'agit de 100 000 notices bibliographiques décrivant tous types de documents, possédant des liens aux autorités, des liens entre notices bibliographiques, avec présence d'alphabets non latins. Nous y avons ajouté les données issues de l'expérimentation OCLC pour la FRBRisation (zones 579 en sus dans les notices, et notices d'autorités pour les regroupements).

Les notices ont été exportées en MARC XML depuis la base miroir XML, puis divisées en lots pour faciliter le traitement. L'XSLT a été appliqué via Oxygen au moyen du processeur Saxon *professional edition*. Les fichiers «pivots » ont été stockés sur serveur pour permettre ensuite l'application de l'XSLT choisi selon le format de destination (PropertyGraph, tables, RDF*).

Puisqu'il a fallu faire des choix, les zones et sous-zones traitées devaient nous permettre de tester des choix de modélisation (par exemple pour l'entité Nomen encore difficile à appréhender) et de concevoir des mécanismes que l'on pourrait ensuite étendre et généraliser (par exemple la zone 452, Autre édition sur un autre support, pour un lien réciproque entre notices). Nous avons pris soin de retenir au moins une zone pour chaque type de données (données codées : zones 101 et 102 pour les codes de langue et de pays, données textuelles : zone 200$a pour la transcription du titre, zone 320 de note bibliographique, données de lien entre notices : zone 410 titre de la collection, zone 7XX$3 pour les liens aux agents.)

Pour la transformation des données qui devra être faite, notre expérience nous amène à dire qu'il serait utile de scinder selon des critères les notices MARC en lots et de produire des XSLT plus spécifiques[3] qui n'auront pas besoin de calculer eux-mêmes des caractéristiques pour gagner en performance. On peut penser au travail mené par la BnF pour Noemi par exemple, qui a défini dans un moteur de règles tous les différents schémas d'éclatement OEMI des notices entrantes. A cette fin, il est important d'avoir des éléments statistiques sur les données – l'outil Catmandu[4] repéré et testé sur notre échantillon est prometteur.

Par ailleurs, en fonction du choix du format de destination, l'XSLT peut ne pas être la bonne option. Si l'on choisit de convertir en RDF toutes les données, il est coûteux et même inutile de passer par l'XML pour les données que l'Abes possède déjà en RDF[5]. Il faudrait alors creuser des pistes de conversion directe RDF à RDF.

[3] On peut imaginer les critères suivants : arbre OEM simple *vs* arbre complexe, puis différents arbres complexes : traductions / notice multi-manifestation / agrégat type collection d'expression ; mais aussi monographie *vs* publication en série ; types de documents ; année de création de la notice (pour traiter au mieux les strates issues d'évolution des consignes de catalogage), etc.

[4] Catmandu : outil en ligne de commandes pour analyser et transformer des lots de données de bibliothèques en format natif, dont le MARC et le PICA+, format de stockage de CBS – documenté ici : https://librecat.org/Catmandu/

[5] Nous avons ainsi renoncé à charger un gros graphe RDF issu d'ISTEX parce que le chemin aurait été très fastidieux : l'export depuis Virtuoso en RDFXML scinde le graphe en plusieurs fichiers. Malheureusement l'XSLT pour produire les entités du tronc OEMI suit une logique qui implique d'avoir dans le même document les triplets

Choix relatifs au format et à la base de stockage

L'étude SCAN de 2018 recensait les outils susceptibles d'être candidats pour les preuves de concept. Au cours des preuves de concept, trois types de stockages ont été abordés : le graphe de propriété, les tables relationnelles, et le RDF - le dernier comprenant un test de la syntaxe RDF* qui « traduit » de façon compatible avec RDF le graphe de propriété et est aujourd'hui en discussion au sein du W3C. Il n'est pas possible au stade des preuves de concept de déterminer le choix du futur système à cet endroit. Néanmoins l'élimination de certaines pistes est déjà un gain indubitable.

Lors de la première preuve de concept, il est apparu que tous les cas d'usages répertoriés des bases de type *property graph*, qu'il s'agisse de Neo4j ou d'Oracle PG, se portaient vers de l'analyse de vastes ensembles de données non structurées : ces bases ne semblent à ce jour pas utilisées pour du stockage natif. Le graphe de propriété nous paraissait une piste sérieuse pour gérer plus facilement un modèle avec des liens qualifiés, sur lesquels le format UNIMARC et la modélisation RDF/Abes ont buté ; pour encoder une information de type « tel chercheur a publié tel article en tant qu'affilié à telle université » ou bien « telle œuvre a été publiée par tel auteur sous cette variante de nom ». S'il peut effectivement gérer ces cas avec facilité et finesse, le graphe de propriété est malheureusement plus démuni pour traiter la majorité des cas d'usages des données bibliographiques : ce sont des données très normées, sur lesquelles nous avons un niveau de contrainte fort pour garantir leur homogénéité. C'est très important dans un système où le catalogue collectif est issu d'un catalogage partagé[6]. L'inconvénient majeur du graphe de propriété pour notre cas est l'absence de mécanisme coercitif des couches basses : pour bâtir dessus un système de gestion de métadonnées bibliographiques hautement normé, il est nécessaire de rajouter par-dessus la base elle-même tous les mécanismes de contraintes pour garantir la qualité de la base.

La deuxième preuve de concept a cherché à établir comment combiner la souplesse d'un modèle de type entités/relations avec la robustesse et la solidité d'une base de données relationnelle. Nous avons fait le choix de définir une table par type d'entité ainsi que des tables de relations en fonction des entités liées. La grosse difficulté à laquelle nous nous sommes heurtés a été le chargement des données, pour lequel nous avons dû opérer des contorsions afin de satisfaire les exigences structurelles d'une BDDR sur clés primaires et étrangères. Pour faire simple, il est délicat de faire entrer dans des tables des données pensées selon des modalités de graphe. La souplesse que l'on souhaite avoir dans la gestion du modèle en se libérant du carcan de l'UNIMARC et de l'extrême lourdeur que représentent aujourd'hui les implémentations de nouvelles zones, semble difficile à concilier avec la rigidité d'une base où les tables dépendent du modèle. Par ailleurs, les cas de multi-valeurs nous ont amené à dé-normaliser. Par exemple, pour stocker plusieurs résumés du même document dans différentes langues, nous avons organisé les couples résumé/langue au sein d'un tableau JSON à l'intérieur d'un champ de la table. Cette dé-normalisation est potentiellement gênante pour le fonctionnement des index, et pour les performances en général. Une solution alternative aurait pu être de créer des tables très génériques sans suivre le modèle, mais l'on se retrouve

relatifs à l'arbre entier. Et concaténer les fichiers de sortie en un seul énorme fichier pose des problèmes de performance de processeur pour lancer l'XSLT.

[6] L'important travail de prospection mené dans le projet nous a montré que les cas où un catalogue collectif est issu de catalogage partagé sont rares. Les échanges avec les collègues de Suède, pays dans cette configuration, ont été précieux en ce sens.

alors dans une situation proche de celle du stockage du CBS aujourd'hui, où les index sont difficiles à mettre en œuvre.

La troisième preuve de concept s'est appuyée sur le format RDF et des bases de données graphe natives ou *triple stores*. Le chargement en a été grandement facilité et c'est une des raisons de ce choix : étant donné le temps court d'une preuve de concept, il nous apparaissait nécessaire de choisir un outil permettant d'explorer d'autres aspects. Les tests relatifs à la syntaxe RDF* (et à l'extension SPARQL* idoine pour l'interroger) ont été plutôt concluants, même s'ils ne portent que sur des points assez précis de la modélisation. RDF* est une solution de réification concise et qui correspond à nos cas d'usage ; il n'en reste pas moins que toute réification induit un alourdissement et une complexification du modèle. Les apports ne sont pas substantiels par rapport aux inconvénients actuels : comme cette syntaxe encore en discussion n'est pas une recommandation officielle du W3C, elle pose des problèmes de compatibilité avec des structures et des outils plus génériques[7]. Le choix du RDF signifie aussi disposer des standards assortis pour le langage de requête (SPARQL) et les langages permettant la description, qui peut être performative, de la représentation des connaissances et des règles métier pour structurer des données (RDFS, OWL et SHACL). Cette panoplie présente le gros avantage d'être accessible à la fois à des informaticiens et à des fonctionnels, ce qui en fait un possible support de dialogue et de coopération. Nous avons ainsi expérimenté la mise sur pied d'une ontologie, extraite de nos données, et le langage de contraintes SHACL[8] qui permet d'écrire directement en RDF des règles de validation, de manière similaire à Schematron pour XML.

Il y a une forme d'évidence à considérer que le RDF est le cadre qui nous fournit à la fois une extensibilité, une excellente souplesse et des mécanismes inhérents de contrôle et de validation des données. En outre, des données RDF peuvent être stockées de différentes manières : au sein d'un triple store de manière native, au sein de bases de données de type document au format JSON-LD[9], au sein de bases de données relationnelles, voire en combinant ces différentes options[10].

3/ Les points qui resteront à traiter

Le projet Sudoc21 avait des ambitions importantes, qui sont devenues d'autant plus difficile à satisfaire que les conditions de travail ont radicalement changé au milieu et que l'équipe a été amputée de son responsable informatique. Les points sur lesquels nous n'avons pas ou peu d'éléments significatifs sont :

- Les transformations en sortie, pour exporter les données communes dans des formats connus

On sait que les partenaires de l'Abes, au premier rang desquels se trouvent les bibliothèques du réseau Sudoc, fonctionnent et continueront encore pendant nombre d'années à fonctionner avec des outils traitant uniquement des notices MARC. En dépit des progrès réalisés en central sur le modèle, une voie d'export classique devra perdurer. Les XLST écrits

[7] Par exemple, la librairie Apache Jena, important *framework* Java qui permet notamment de lancer des validations de données à partir de contraintes SHACL, ne supporte pas le RDF* à ce jour.
[8] SHACL est une recommandation W3C depuis 2017. Elle est disponible ici : https://www.w3.org/TR/shacl/
[9] C'est le choix fait par le projet suédois LIBRIS XL.
[10] C'est le cas de l'infrastructure derrière Wikibase, qui combine une base documents MariaDB, un moteur ElasticSearch et duplique les données dans le triplestore BlazeGraph.

pour transformer les données ne sont pas réversibles, et il n'était pas envisageable pour l'équipe de passer du temps à écrire des XSLT de rétroconversion appauvrissants à seule fin de montrer qu'il est possible de reconstruire des notices MARC. Cette tâche devra tout de même être accomplie.

- Les workflows d'entrée et de modification de données

A ce stade de l'expérimentation, nous n'avions aucun circuit standard ou répétable. Les modifications dans les XSLT nous ont certes conduits à recharger les données à de multiples reprises mais les interventions sont restées très manuelles. Ce n'est pas étonnant si l'on songe que les circuits WORME[11] mis en place à l'Abes ne se sont sédimentés que graduellement sur plusieurs années au cours du projet Hub de métadonnées : il faut d'abord définir les étapes, les rôder une à une ; et ce n'est qu'en constatant que le processus sera répété qu'il devient intéressant de l'automatiser.

- La gestion des gros volumes.

Le temps nécessaire pour produire le modèle et transformer les données n'a permis de traiter que des volumes modestes et sans commune mesure avec l'intégralité des données aujourd'hui conservées à l'Abes. L'état le plus riche de la base en fin de PoC3 était en entrée de 100 000 notices MARC avec leurs exemplaires, 14 thèses en TEF, 2 millions de triplets RDF issus des corpus ISTEX et une poignée de fichiers KBART. Cela représente, en fin de PoC3 dans la base, 40 millions de triplets, parmi lesquels le MARC est surreprésenté (près de 90%).
A titre de comparaison, l'ABES gère aujourd'hui un gros millier de fichiers KBART, 114000 thèses en TEF, 18 millions de notices MARC, et près de 4 milliards de triplets dans ses bases RDF. Il est discutable de considérer qu'on se situe ou non dans le domaine du « big data » ; en effet il n'existe pas dans les données bibliographiques gérées par l'Abes d'accroissement continu et massif comme le sont par exemple, les données industrielles issus de multiples capteurs qui pointent des valeurs à des intervalles d'une seconde.

- La montée en charge pour soutenir la concurrence multi-utilisateurs (accès multiples en écriture)

Une des caractéristiques importantes de la base maîtresse sur laquelle se pratique le catalogage (aujourd'hui CBS) est de pouvoir tenir la charge d'un millier d'utilisateurs humains simultanés opérant des recherches, modifications et suppressions, de manière concomitante à des traitements de masse opérés par des machines (alignements, corrections, etc.) Un des apports du modèle qui décompose les notices en entités, et qui sépare notamment les niveaux bibliographiques et exemplaires aujourd'hui organiquement liés, est de réduire la probabilité des cas de concurrence multi-utilisateurs. Toutefois, deux ambitions du nouveau système risquent de conduire à des flux de modifications intensifiés : celle de mieux refléter les données d'exemplaires de tous les membres des réseaux, et celles de faciliter les traitements de masse automatisés effectués par l'Abes. Au stade des preuves de concept, nous n'avons pas pu fournir d'éléments sur la robustesse et la capacité à monter en charge des bases testées.

- La gestion de l'historicisation (suivi fin des modifications, retour en arrière)

Ce point apparaissait comme très important dans l'étude préalable, en raison d'un manque constaté sur les notices MARC hébergées dans CBS – les données permettant de connaître le

[11] WORkflow de Modification de d'Edition des données Abes mis en place lors du projet Hub de métadonnées, qui reposent sur des briques agencées de manière semi-autonome par les agents pour lancer et suivre les processus d'imports massifs des données. Une base Oracle joue le rôle d'orchestrateur.

dernier modificateur, mais non les modifications ; ce travers étant en partie corrigé au niveau de la base miroir en XML. Néanmoins il est apparu pendant le projet, et c'est là que l'on voit le caractère précieux de l'étude de l'existant, que le CBS possède depuis quelques années une fonctionnalité de log, non connue donc non exploitée, mais qui recense précisément les modifications au niveau de la sous-zone et permet de les annuler. Cette découverte est cruciale, car elle offre un corpus réel pour mieux définir les usages, dans un futur système, de l'historicisation des modifications. Que voulons-nous suivre et pourquoi ?

Par ailleurs, une approche théorique a tout de même été tentée en fin de PoC 2, après l'essai infructueux de l'outil Liquibase (destiné à suivre les modifications des schémas de BDD et non des données elles-mêmes). On peut imaginer que la base de production est dupliquée. Dans la base dupliquée, la fonction d'audit transforme un journal de transaction propriétaire (WAL – Write Ahead Log) en des tables SQL exploitables et interrogeables. Le fait d'avoir ces tables dans une copie de la base permet de ne pas ralentir la base de production lorsqu'on exploite ces données de différentiel. Ce fonctionnement est illustré par un schéma[12].

4/ Les enjeux à la sortie de ce projet

Comme nous l'avons rappelé en introduction, le projet Sudoc21 fait suite à une longue gestation. Il a pris à bras-le-corps l'ambition de concevoir un nouveau système de gestion des métadonnées (NSGM) défini comme l'objectif numéro 1, priorité 1, du projet d'établissement en cours. Il nous apparait donc logique et légitime de capitaliser sur le travail issu des preuves de concept pour démarrer une réalisation concrète.

En interne

La menée du projet selon la méthode Agile nous a conduit à tenir informées l'ensemble des équipes de l'Abes au moyen des revues d'itérations, et des « revues de Poc » revenant sur chaque preuve de concept de manière globale. Ces revues ont rassemblé en moyenne 30 à 40 personnes. Il s'agissait d'un effort de pédagogie pour expliquer le travail en cours et recueillir les avis et suggestions des collègues. Le dialogue a été fructueux, mais le passage forcé au télétravail massif a compliqué la transmission. Il nous parait important d'insister sur deux points : faire en sorte qu'il n'y ait pas de déperdition à la dissolution de l'équipe (car tout travail comporte une part d'incorporé) et faire « infuser » dans l'établissement les acquis du projet. La construction de fait d'un nouveau système ne peut pas se penser comme un projet séparé des activités courantes de l'Abes.

Au sein des réseaux Abes

Depuis les Journées Abes 2019, aucune communication n'a été faite aux réseaux de l'Abes au sujet du projet Sudoc21. Conçu comme un projet interne, très technique, et dépourvu de conséquences immédiates sur les usagers, il n'avait pas vocation à être dépeint et chroniqué en détail. Les JABES 2021 qui auront lieu cet automne seront probablement l'occasion de revenir sur le travail effectué à la condition de faire le lien avec un ou des chantiers concrets

[12] Le schéma est visible et commenté sur cette page – billet de blog : https://medium.com/@ramesh.esl/change-data-capture-cdc-in-postgresql-7dee2d467d1b

à venir. Ceux-ci, qui concerneront directement les usagers professionnels, seront nécessairement conçus en partenariat avec eux, selon des modalités à définir.

A l'extérieur

Le projet Sudoc21 s'est abondamment nourri des expériences extérieures, également documentées et mises à disposition sur un wiki interne. Il ne serait que justice de partager à notre tour l'expérience acquise à l'occasion d'un article (ex : revue code4lib) ou de participations à des conférences du domaine (ex : ELAG, SWIB). Il appartient également à l'Abes de saisir l'occasion pour mieux s'insérer dans les réseaux internationaux pour échanger avec des pairs (projets Share-VDE[13], LD4P2[14], etc.). Par ailleurs, conformément à la politique informatique, le code produit (les XLST) a été déposé sur Github[15], documenté pour les humains, et rendu librement accessible et réutilisable.

5/ Un nouveau départ : les pistes proposées

Les pistes proposées ci-dessous sont issues de la réflexion collective de l'équipe Sudoc21. Elles sont présentées pour guider la réflexion et pourront faire l'objet d'une présentation aux différentes instances selon l'avis du comité de pilotage.

Piste 1 : avancer dans la LRM-isation des données du Sudoc

Du point de vue Abes, il faut rebondir après l'expérimentation FRBR qui a porté sur une grande partie du Sudoc (exclusion des agrégats, dont les périodiques). Il s'agit de passer d'une logique de grappe à une logique de redistribution des informations entre les notices d'entités OEMI. La décomposition des notices en entités pourra, pour une partie importante des notices, se faire de manière simple : lorsqu'à une notice correspond un arbre OEMI sans embranchements –ce qu'on peut appeler « les cyprès ». Le gros du travail va consister à organiser correctement les entités issues des notices en arbres complexes (qu'on pourrait appeler « la mangrove »). C'est un double jeu de rapprochement (identifier la même œuvre décrite par x notices) et d'éclatement (sur quels critères au sein d'une notice MARC on crée tant d'œuvres, tant d'expressions, tant de manifestations.) L'expérience de la BnF pour Noemi et des collègues suédois pour Libris XL nous enseigne que l'éclatement sans rapprochements préalables crée de grosses quantités de doublons à traiter.

L'idée d'une collaboration avec la société Progilone qui a développé (notamment lors de la thèse CIFRE de Joffrey Decourselle, lue et documentée lors du projet Sudoc21[16]) des algorithmes de FRBRisation de notices est aujourd'hui sur la table. Elle n'a encore fait l'objet d'aucune formalisation.

Il nous parait intéressant de continuer à **creuser l'approche de l'algOCLC2**[17] qui constitue des grappes de notices. Ce rapprochement de niveau œuvre permettrait de faciliter la constitution

[13] https://www.share-vde.org/sharevde/clusters?l=en
[14] https://wiki.lyrasis.org/display/LD4P2/
[15] https://github.com/abes-esr/abes-format-pivot
[16] Lien vers la thèse : http://www.theses.fr/2018LYSE1183/document
[17] L'algOCLC2 comme son nom l'indique est la V2 de l'algorithme proposé par OCLC, fournisseur de CBS, pour identifier dans le Sudoc les regroupements de notices décrivant la même œuvre. Le rapport final est un

de petits ensembles à convertir en limitant les doublons d'entités Oeuvre. Par ailleurs, elle permettrait à la transition bibliographique de prendre corps, comme une étape transitoire.

Plusieurs pistes techniques sont possibles. **L'algOCLC2** a le défaut d'être un algorithme propriétaire sur lequel nous n'avons qu'une maîtrise partielle. Il a pourtant l'avantage d'avoir été éprouvé, son fonctionnement considérablement documenté. La dépense a par ailleurs déjà été réalisée. Il serait possible de consolider en laissant tourner l'algorithme en place, mais en infléchissant son fonctionnement pour qu'une intervention de l'humain soit prépondérante et ne puisse être annulée.

Par ailleurs, **le framework Sudoqual**[18], aujourd'hui utilisé à travers l'application Paprika pour des opérations de liage et de diagnostic entre personnes et de personnes à références bibliographiques, a été conçu dès le départ pour s'appliquer à d'autres cas, notamment pour construire un scénario de liage au niveau de l'entité Œuvre. Son caractère modulable et le support qu'il offre pour un dialogue fécond entre bibliothécaires et informaticiens sont des atouts face à des algorithmes proposés par des sociétés extérieures.

Quel que soit l'outil choisi, **une démarche générique** pourrait être mise en œuvre avec les éléments suivants :

- Construire un outil pour visualiser et corriger les grappes
- Préfigurer des « notices d'œuvres » en se focalisant sur l'indexation matière et les liens aux autorités personnes
- Répercuter le travail sur les grappes dans les outils de visualisation existants (PSI, IdRef)

L'outil de visualisation et de modification des grappes de notices bibliographiques et des notices elles-mêmes pourrait être un exemple à la fois de curation des données pour préparer la migration au plan technique, et de familiarisation des bibliothécaires, de manière concrète et en partant de données réelles, avec la notion d'œuvre. Il pourrait s'agir d'une interface de type Paprika, qui donnerait à voir des ensembles de notices bibliographiques et permettrait de faire et défaire les liens vers des pré-notices d'œuvres. Ces pré-notices d'œuvres (avec pour base les actuelles notices Tr) pourraient récupérer depuis les notices bibliographiques liées des informations de manière intelligente (par ex proposer la date la plus ancienne comme date de l'œuvre, l'indexation sujet commune..) et le bibliothécaire pourrait ajuster, corriger ces remontées d'informations. Le travail sur ces pré-notices d'œuvres, conformes aux évolutions récentes du format UNIMARC, pourrait être l'occasion d'étendre l'indexation des notices bibliographiques, qu'il s'agisse de l'indexation matière ou de l'indexation auteur : si les notices appartiennent à la même grappe, alors les liens aux autorités sujet et auteur (lorsque le rôle est de niveau Œuvre) sont répercutés dans toutes les notices.

---

document interne. Pour plus d'informations, voir les billets de blog OUBIPO sur l'expérimentation FRBR :
https://oubipo.abes.fr/experimentation-sudoc-frbr-ii-portrait-robot-dun-algo/
https://oubipo.abes.fr/experimentation-sudoc-frbr-ii-levaluation-algo-vs-humain-1-3/
https://oubipo.abes.fr/experimentation-sudoc-frbr-ii-levaluation-algo-vs-humain-2-3/
https://oubipo.abes.fr/experimentation-sudoc-frbr-ii-levaluation-algo-vs-humain-3-3/

[18] Le framework Sudoqual a été développé dans le cadre du projet ANR Qualinca en collaboration avec le LIRMM. Le code écrit par le prestataire a été internalisé en 2019 avec formation des informaticiens Abes.

Plus encore, un tel outil pourrait être utilisé comme support pédagogique lors des formations aux catalogueurs pour rendre concret cette « transition bibliographique » qui tarde à l'être. Il pourrait aussi être un pont vers le web de données, en proposant des rebonds au-delà du catalogue Sudoc pour enrichir les pré-notices d'œuvres, par exemple vers Wikidata, Wikipedia, ou dans le cas des œuvres étrangères, vers des interfaces LOD telles datos.bne.
Enfin, des aménagements modestes dans les outils existants (PSI, IdRef) pourraient permettre de donner à voir ces progrès. Dans PSI, on pourrait envisager de faire apparaître à l'affichage d'une notice bibliographique le lien vers la pré-notice d'œuvre (aujourd'hui zone B579) et afficher la notice d'autorité correspondante ou bien permettre de rebondir sur toutes les autres notices liées de la grappe. Dans IdRef, on pourrait afficher les pré-notices d'œuvres au même titre que les autres notices d'autorité et ainsi présenter les notices bibliographiques liées dans le Sudoc.

Pour synthétiser, voici les **grandes étapes** :

- Exploitation de l'algOCLC2, tout en donnant la priorité aux ajustements manuels des catalogueurs
- Poursuivre la LRM-isation de cas plus complexes, avec une autre solution
- Passer d'une logique de « grappe » à une logique de redistribution et d'héritage des informations entre les notices d'entités OEMI.
- Mettre entre les mains des catalogueurs les liens OEMI et les informations redistribuées, pour apprendre en faisant.

Piste 2 : la première brique du NSGM, un environnement unifié pour les périodiques.

La construction d'un nouveau système de gestion des métadonnées est un des objectifs phares de l'actuel projet d'établissement. Il s'agit d'un énorme chantier qui a deux visées : d'une part unifier les données, en sortant du carcan des formats actuels pour aller vers un modèle conforme à LRM, et d'autre part rationaliser le SI en s'appuyant sur des briques génériques et des fonctionnements standardisés, pour faciliter la maintenance et le développement. L'unification sur ces deux plans, modèle de données et stockage/accès technique, facilitent les modifications de données unitaires ou en masse.

Cette proposition consiste à circonscrire un sous-ensemble de notre système actuel pour bâtir le nouveau système, en partant d'un type de documents : **les périodiques.**

Plusieurs raisons nous poussent à faire des périodiques le fer de lance du NSGM :

- L'Abes a une responsabilité particulière sur ce type de documents, en tant que coordinatrice du réseau Sudoc-PS qui brasse au-delà de l'enseignement supérieur.
- L'attente des établissements est grande au sujet des périodiques électroniques, depuis longtemps, et nous y répondons imparfaitement.
- La O-E-M-I-sation des périodiques, conformément à LRM, est une opération relativement simple, puisqu'à chaque périodique correspond un arbre simple, un « cyprès ». La transformation des données pourra donc être plus rapide.
- Nonobstant, les données descriptives de périodiques contiennent déjà beaucoup de liens (zones 4XX du MARC) que nous pourrons exploiter pleinement en passant à un modèle entité/relation.

- La gestion de ce type de document par l'Abes passe aujourd'hui par une mosaïque d'applications : du côté du système informatique, comme pour les fonctionnels, il y a beaucoup à gagner d'une unification.
- Les périodiques bénéficient d'un identifiant (l'ISSN) stable et reconnu dans tout l'écosystème de la production scientifique et éditoriale : la logique des données liées ouvertes peut s'y développer avec rapidité et profit.

Ce projet permettrait de fédérer les forces humaines et les initiatives déjà entamées en plusieurs points. Les initiatives actuelles autour des périodiques sont légion : on peut citer le plan de développement de BACON, la refonte de Périscope, la collaboration avec Mir@bel, ou encore le projet de synchronisation Sudoc/Alma pour les exemplaires. Leur coordination peut reposer sur une ambition nouvelle : la mise sur pied d'un réservoir de données unifiées.

Le travail mené dans Sudoc21 a montré qu'il était possible de rassembler les données aujourd'hui stockées en silos dans des formats distincts, au sein d'un modèle unifié. La logique de « pot commun » s'illustre parfaitement pour les périodiques, puisque ce type de documents est décrit à la fois dans les notices MARC, dans les fichiers KBART et dans les données RDF des bases internes (désormais pour certaines publiées avec https://scienceplus.abes.fr/).

La brique « périodiques » du NSGM vise à **remplacer les outils existants à moyen terme.** En rassemblant dans une seule base, avec un seul modèle, toutes les informations descriptives qui les concernent, on peut imaginer des services qu'il est aujourd'hui très difficile à mettre en place. En termes de volume, cette base rassemblerait toutes les informations issues des 2 millions de notices MARC du Sudoc décrivant des périodiques, ainsi que les 1013 fichiers KART stockés aujourd'hui dans BACON, et tous les graphes RDF de revues de la base de travail du Hub. La mise en commun des données concerne donc les périodiques imprimés et électroniques, et tous les niveaux de description existants : le bouquet, la revue, le fascicule, le numéro, l'article. On pourrait parler d'un « IdRef des périodiques » : autour du point focal de la revue, rassembler toutes les informations pertinentes.

Sur cette base unique peuvent être construites **une interface professionnelle** et **une interface publique** (un OPAC LRM).

L'interface professionnelle serait un lieu unique pour accomplir les tâches relatives aux périodiques effectués dans les établissements : catalogage, demande de numérotation, renseignement des états de collection, récupération des informations de bouquets, gestion des PCP, politique documentaire. Le stockage unifié de toutes les informations relatives aux périodiques pourraient rendre les services suivants :

- Permettre une recherche sur tous les éléments (qu'ils soient de niveau «bibliographique» ou de niveau « exemplaire ») – par ex, dans le cadre d'un projet Collex, retrouver tous les exemplaires des revues dans une discipline (indexation Rameau=info biblio) qui proviennent d'un ancien fonds disséminé (note sur la provenance=info d'exemplaire)
- Visualiser les données, sous forme de frises pour comparer des états de collection selon des critères libres

- Visualiser l'historique d'un titre (« métarevue ») en rassemblant les différents états d'une revue pour lisser les changements
- Exporter les données de manière autonome, selon des critères libres et non prédéfinis
- Apporter des modifications sur une sélection de données – par ex, le statut PEB ou la cote de tous les exemplaires de ma bibliothèque pour les revues sélectionnées
- Signaler le détail des états de collection de manière pratique et facile – par ex, à partir d'une frise ou d'un damier pour « décocher » facilement les lacunes
- Harmoniser la présentation des informations de niveau Item, qu'il s'agisse d'états de collection (laps, continus ou non) ou de numéros isolés (points)[19]
- Signaler ses états de collection de manière automatisée par une remontée des informations depuis les systèmes locaux, à la fois pour l'accès électronique (choix de bouquets) et pour le papier (remontée des informations de bulletinage)
- Donner des outils de politique documentaire pour les revues électroniques : pouvoir comparer facilement des offres commerciales (bouquet) pour un titre ou bien une liste de titres – par ex, si je veux m'abonner à telle liste de périodiques, quelle est la meilleure combinaison ? ou bien pour tel titre, quelles sont toutes les options d'accès possible ?
- Faciliter la vie des services de PEB avec une interrogation « scihub-like » : à partir d'un DOI d'article, via l'API Crossref[20], trouver facilement une copie appropriée au sein du réseau.

L'interface publique permettrait aux usagers de consulter de manière fluide la disponibilité d'un titre quel que soit son format (papier ou électronique). Elle serait un exemple abouti des gains présentés par le modèle LRM pour les utilisateurs. Elle pourrait, à l'exemple des interfaces des SGB des établissements ou de celle de Mir@bel, mettre en avant les informations que nous jugeons utiles et pertinentes : les liens aux personnes, les possibilités d'accès ouvert…

Un apport important de ce projet sur les périodiques pourra être de produire des résultats concrets et visibles dans un périmètre circonscrit, pour se donner confiance et peaufiner les idées pour les autres types de documents. Les logiques centrales seront déjà à l'œuvre : liens entre entités (exploitation des actuels liens entre notices biblio et de notices biblio vers notices d'autorité), emboîtements (bouquet/revue/numéro/article, qu'on retrouvera dans monographie/chapitre, ou à une échelle plus grande dans les données d'archives en EAD) et alignement sur des données externes dans une logique *Linked Open Data*. Un tel projet est aussi l'occasion parfaite de donner corps à une des ambitions majeures du projet d'établissement qu'est la co-construction, pour réussir au mieux l'interconnexion entre le nouveau système et les systèmes locaux des établissements.

[19] Aujourd'hui les établissements peuvent signaler dans le Sudoc - donc en MARC, leur état de collection sous la notice de périodique et créer des notices de type monographie pour les numéros de périodiques isolés. La notice du numéro renvoie à la notice du périodique (zone 461) mais pas l'inverse.

[20] L'API Crossref permet de décoder, à partir d'un DOI, d'où vient un article (dans quelle revue/numéro il a été publié). On pourrait ensuite faire le lien, si jamais la revue électronique n'est pas dispo, avec la version papier.

Pour synthétiser, voici les **grandes étapes** :

- Définir le modèle de données unifié, appuyé sur des standards existants, pour réconcilier toutes les informations relatives aux périodiques issues des formats MARC, KBART et RDF
- Transformer les données
- Définir le socle technique et charger les données modélisées dans la base
- Construire sur cette base les services aux utilisateurs (pro et public) pour la visualisation, la modification, l'export des métadonnées
- Fermer à mesure des services développés les anciennes applications
- Exploiter cette nouvelle architecture pour concrétiser les partenariats avec ceux qui disposent, créent ou diffusent des informations concernant les périodiques (Mir@bel, Couperin, éditeurs, ISSN, etc)

Conclusion

Ce rapport auprès du Conseil Scientifique tente de présenter synthétiquement tous les éléments utiles issus du projet Sudoc21 pour en apprécier les tenants et aboutissants. Les jalons qu'il apporte sur la conception d'un futur système sont précieux : intérêt du modèle unifié, organisation de la transformation des données, exclusion de pistes pour le stockage. Cette succession de preuves de concept reste pourtant insuffisante sur de nombreux points, et a vocation à être complétée par des études techniques spécifiques, qui prendront d'autant plus sens qu'elles s'inscriront dans un projet de services délimités permettant la co-construction avec les établissements.

Il est dédié à la mémoire de Stéphane Rey, pilier de l'Abes de longue date, fin connaisseur de CBS, précurseur du projet, rapporteur de la mission Sudoc2 en 2015, co-responsable du projet Sudoc21 avec le rôle de *scrum master*, notre collègue et ami. Nous espérons avoir fait honneur à ses attentes.

Composition de l'équipe projet : Laetitia Bothorel, Théo Chambon, Michaël Jeulin, Carole Melzac, Thomas Michaux, Stéphane Rey, Emilie Romand-Monnier, Olivier Rousseaux, Mickaël Seror

abes

Rapport de fin de projet à l'attention du Conseil Scientifique de l'Abes, mars 2021
Auteur : Carole Melzac, responsable de projet Sudoc 21

Agence bibliographique de l'enseignement supérieur
avenue du Professeur Jean-Louis Viala - CS 84308 - 34193 Montpellier Cedex 5 - Tél. 04 67 54 84 10

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR, DE LA RECHERCHE ET DE L'INNOVATION