abes
agence bibliographique de l'enseignement supérieur

SudocAD

Rapport final

6 décembre 2011

Ce document constitue le rapport final du projet SudocAD (octobre 2010 - octobre 2011).

L'objectif de SudocAD était de proposer un prototype qui sache « enrichir les métadonnées du métaportail Adonis du lien aux autorités Sudoc » (appelées aujourd'hui autorités IdRef).

Ce rapport final a été rédigé par l'ABES mais le projet SudocAD a été mené conjointement par l'ABES et l'équipe GraphIK du LIRMM. Il a été en partie financé par le TGE ADONIS, dans le cadre de son appel à projet 2009-2010.

Sommaire

Résumé

Mené par l'ABES et l'équipe de recherche GraphIK du LIRMM, co-financé par le TGE ADONIS dans le cadre de son appel à projets 2009-2010, le projet SudocAD vise à interconnecter entre eux différents corpus de métadonnées agrégés par la plateforme de recherche ISIDORE, en les reliant au référentiel IdRef. Ce qui est en jeu, ce n'est pas seulement l'efficacité de la recherche dans Isidore, mais l'intégration des données SHS françaises au web de données, auquel IdRef est déjà connecté.

L'objectif opérationnel du projet était d'enrichir automatiquement des notices d'articles du portail Persée, en identifiant (quand elle existe) l'autorité IdRef correspondant à chacun des auteurs de l'article. 13 444 notices ont ainsi été traitées et livrées à ADONIS et à l'équipe Persée.

Pour identifier la notice d'autorité IdRef qui correspond à l'auteur Persée, SudocAD ne se contente pas d'utiliser les informations contenues dans la notice d'autorité mais exploite les connaissances enfouies dans les notices bibliographiques Sudoc liées. Toutes ces connaissances sont exprimées en RDF, selon le vocabulaire FRBROO. Il devient possible alors de raisonner à propos de ces connaissances, grâce aux outils sémantiques conçus et développés par GraphIk.

Les principales étapes du traitement opéré par SudocAD sont les suivantes : le nom et le prénom de l'auteur Persée sont utilisés pour sélectionner une liste parfois longue d'autorités IdRef candidates ; le raisonneur du LIRMM charge un ensemble de données RDF composées de la notice Persée, des autorités candidates et des notices bibliographiques Sudoc liées à ces autorités ; enfin, après avoir analysé ces données au moyen de règles logiques, le raisonneur répartit les autorités candidates en sept catégories de liage, de *Strong* à *Impossible*.

SudocAD ne donne donc pas directement un verdict sur la bonne autorité à lier. Mais, à partir du rapport d'analyse en XML et des sept catégories, il est facile de définir un algorithme qui détermine automatiquement l'autorité à lier. Mais il existe plusieurs manières de construire un tel algorithme. Ce rapport distingue les algorithmes de liage automatique qui paraissent les plus pertinents.

A côté du liage automatique, le rapport d'analyse généré par SudocAD peut également être utilisé dans une perspective d'aide à la décision. Il s'agirait d'utiliser ce rapport pour présenter les autorités candidates d'une manière qui facilite et fiabilise le travail manuel du catalogueur qui cherche à lier une notice bibliographique à une autorité.

Afin d'évaluer l'approche de SudocAD, un protocole a été établi pour comparer les résultats d'un traitement automatique aux décisions de liage prises par un catalogueur. Sur un échantillon de 150 notices Persée, elle montre que SudocAD atteint un très bon taux de bonnes décisions (liage ou non liage), autour de 80%, et surtout un taux d'erreur (création de liens erronés) inférieur à 2%.

Au-delà du projet SudocAD, l'ABES et l'équipe GraphIK ont la volonté d'éprouver la validité de cette approche sur d'autres corpus de métadonnées et d'améliorer encore son efficacité en corrigeant les défauts actuels et surtout en élargissant le spectre des informations prises en compte, notamment en exploitant de manière sémantique les co-auteurs et le vocabulaire Rameau.

Contexte et objectifs

Contexte

Comme d'autre agrégateurs de métadonnées, la plateforme de recherche Isidore[1] est inévitablement confrontée à l'hétérogénéité des données qu'elle moissonne.

Au-delà de l'hétérogénéité des types de documents décrits, des formats de métadonnées, des vocabulaires de description et des niveaux de description, il existe ce qu'on pourrait appeler l'hétérogénéité sémantique : même si deux corpus de métadonnées à agréger contiennent chacun une notice faisant référence à la même entité, comment expliciter et exploiter le fait qu'il s'agit bien de la même entité ? Par exemple, comment savoir que deux notices décrivent le même document si chaque corpus de départ lui donne un identifiant unique différent ? Autre exemple, au cœur de la problématique de SudocAD : comment savoir que deux notices décrivent deux documents possédant le même auteur, que cet auteur soit seulement mentionné par une de ses appellations ou que chaque corpus l'associe à un identifiant unique… différent ? En effet, il faut rappeler qu'un identifiant unique est unique au sens où il est censé ne faire référence qu'à une et une seule entité, mais pas au sens où cette entité ne posséderait qu'un seul identifiant unique : une même entité peut posséder plusieurs identifiants uniques, même au niveau global du Web ou du Web de données.

Dans les catalogues de bibliothèques, cette hétérogénéité sémantique est traitée par le contrôle d'autorité. Afin d'expliciter le fait que, dans deux notices bibliographiques, derrière un même nom (ou derrière deux variantes du même nom voire deux noms différents) il s'agit de la même entité, on crée une troisième notice qui va fixer de l'information à propos de cette entité. Par exemple, afin de fixer une seule manière d'appeler cette entité, la *notice d'autorité* va fixer une forme linguistique privilégiée parmi des variantes possibles et avérées ou parmi des formes alternatives. C'est cette forme linguistique privilégiée, "forme retenue" selon le jargon catalographique, qui va être utilisée comme quasi-identifiant dans les notices bibliographiques afin de faire référence de manière normalisée et univoque à l'entité en question. SudocAD s'intéresse aux autorités de personne physique, mais toutes sortes d'entités peuvent être décrites dans une notice d'autorité : collectivités, lieux, concepts, familles, marques commerciales, navires…

Aussi précise soit-elle, comme " Lang, François (19..-20.. ; professeur de pharmacologie)"[2], une forme retenue offre une garantie d'unicité trop faible. C'est pourquoi certains catalogues de bibliothèques préfèrent identifier une autorité par un code. Dans le Sudoc, cet identifiant est

[1] Il faut préciser qu'au moment de l'appel à proposition, l'opportunité du méta-portail (qui devait devenir plus tard la plateforme de recherche Isidore) était à l'étude. A fortiori, l'ABES et le LIRMM ignoraient que l'appui sur des référentiels serait un des axes forts du cahier des charges. Il faut se féliciter de cette convergence de vues entre ADONIS, l'équipe GraphIK et l'ABES, reflet d'une tendance globale.

[2] Il s'agit de la forme linguistique retenue pour se référer à la personne décrite par une notice d'autorité du Sudoc, accessible à cette adresse : www.idref.fr/088802574.

composé de 9 chiffres. Et c'est cet identifiant unique qui sera injecté dans la notice bibliographique, établissant ainsi un lien logique et physique entre la notice bibliographique et la notice d'autorité, et par voie de conséquence entre un document et une entité bibliographique de type personne. Transformez les notices en triplets RDF, enchâssez l'identifiant dans une URL et vous retrouvez à l'œuvre les principes mêmes du web de données liées (*linked data*), mais confiné au sein d'un catalogue.

C'est précisément pour dépasser cette restriction et généraliser cette logique du contrôle d'autorité à travers des liens entres notices bibliographiques et notices d'autorité que l'ABES a, depuis 2006, engagé une stratégie de promotion des autorités Sudoc au-delà du périmètre du Sudoc. Pour ce faire, l'ABES a déployé divers moyens techniques qui permettent à d'autres catalogues de lier leurs propres notices bibliographiques aux notices d'autorité du Sudoc. Les autorités du Sudoc ne sont plus seulement celles du Sudoc, mais également celles de Calames, de theses.fr et bientôt d'autres bases. D'où leur nouvelle identité : IdRef, pour *Id*entifiants et *Réf*érentiels.

Objectifs

IdRef se positionne comme une solution générique de contrôle d'autorité pour les bases de métadonnées documentaires de l'enseignement supérieur et de la recherche en France. En d'autres termes, comme un référentiel partagé pour contribuer à construire un web de données de l'IST française, à connecter au web de données global. Or, le périmètre d'Isidore, restreint aujourd'hui aux SHS, est inclus dans le périmètre d'IdRef. C'est pourquoi il est apparu naturel de répondre à l'appel à projets ADONIS de fin 2008, en proposant de mettre IdRef au service d'une homogénéisation sémantique des données intégrées à Isidore, en particulier des références aux auteurs. L'objectif est qu'un agrégateur comme Isidore dispose de données interconnectées, et pas seulement juxtaposées.

Mais selon quelle méthode travailler pour mettre en correspondance les noms des auteurs d'Isidore avec les identifiants des autorités IdRef ? En menant ce projet conjointement avec l'équipe GraphIK du LIRMM, l'ABES fait le pari d'une méthode fondée sur l'exploitation logique du Sudoc en tant que base de connaissance. En effet, l'équipe GraphIK[3] est spécialisée dans le domaine de la représentation de la connaissance, en particulier le raisonnement sur des bases de connaissance formalisées au moyen de graphes. L'approche de SudocAD revient donc à utiliser des technologies sémantiques qui sont celles du web de données (RDF, RDFS, OWL), mais avec un objectif de raisonnement, et non d'interopérabilité. En partant des notices d'entrée et en utilisant le Sudoc et IdRef comme base de connaissance enrichie d'une ontologie formelle et de règles logiques, trouver la bonne autorité doit être vu comme la conclusion d'un raisonnement logique.

[3] http://www2.lirmm.fr/~mugnier/graphik/index.html

Approche

Le programme actuel de liage automatique aux autorités dans le Sudoc : une approche classique perfectible

Depuis l'origine (2000), les procédures d'import du Sudoc incluent un programme de liage automatique des notices bibliographiques à importer aux autorités Sudoc. Ce programme s'appuie sur la comparaison stricte des chaînes de caractère présentes dans les notices à importer et des chaînes de caractère présentes dans les notices d'autorité, plus précisément des noms mais également des dates de naissance quand elles sont précisées.[4] Si le programme trouve plus d'une autorité[5] ayant le même nom que celui trouvé dans la notice d'import, alors il s'abstient de créer un lien. Cette méthode d'alignement purement linguistique repose sur l'hypothèse statistique que si un couple {nom, prénom} est fréquent, les autorités Sudoc (= IdRef) possède plusieurs occurrences de ce couple. En d'autres termes, si un auteur du Sudoc possède beaucoup d'homonymes *dans la réalité*, alors cet auteur possède au moins un de ces homonymes *dans le Sudoc*. Cette hypothèse s'avérant inexacte dans un nombre non négligeable de cas, ce programme crée fatalement un certain pourcentage de mauvais liens. Par ailleurs, puisqu'il s'appuie sur une comparaison stricte des chaînes de caractères, dans un grand nombre de cas, il échoue à lier une notice à importer à une autorité en raison des variations linguistiques mêmes mineures qu'on peut trouver dans la manière de transcrire le nom d'une personne.

La méthode de liage automatique décrite dans le paragraphe précédent présente deux types de défaillances :

- Elle repose sur une comparaison trop stricte des noms. De ce fait, cette méthode manque trop souvent de créer les bons liens.

- Elle ne repose que sur la comparaison des noms. De ce fait, cette méthode crée trop souvent de mauvais liens.

Une méthode alternative devrait reposer sur les principes suivants :

- **Effectuer une comparaison moins stricte entre les noms**. Il s'agit de tolérer une certaine dissimilarité entre les noms présents dans la notice d'entrée et les noms présents dans l'autorité. Par exemple, le "M. Caizergues" d'une notice d'entrée peut être la même personne que le "Marcel Caizergues" du Sudoc. Certes, appliqué tel quel à la méthode de liage automatique actuelle des imports Sudoc, ce relâchement des critères de comparaison risque d'augmenter encore le nombre de mauvais liens, mais c'est là une raison de plus de

[4] Même chose, *mutatis mutandis*, pour d'autres types d'autorités que celles des personnes physiques, par exemple les autorités Rameau.

[5] Ou aucune.

ne pas se contenter de limiter la comparaison au seul nom, comme y invite le principe suivant.

- **Exploiter plus d'informations que les seuls noms**. En effet, le Sudoc en sait beaucoup plus sur une personne décrite par une autorité que son seul nom. Non seulement la notice d'autorité possède souvent d'autres informations bibliographiques comme la langue d'expression, mais surtout la collection des notices bibliographiques liées à cette autorité en dit long sur cette personne. Elle dit qu'elle est l'auteur de tel ouvrage, publié à telle date, co-écrit avec telle autre personne, publié dans telle collection, portant sur tel sujet… Toutes ces informations sont autant de connaissances à exploiter, c'est-à-dire à comparer avec les connaissances analogues extraites de la notice d'entrée.

L'approche de SudocAD : beaucoup d'appelés et peu d'élus

Résumé de la démarche

Les grandes lignes de la démarche adoptée par SudocAD peuvent se résumer ainsi :

1. ***Beaucoup d'appelés***. En prenant pour critères de recherche le nom et le prénom d'une notice d'entrée, SudocAD lance une recherche dans le moteur de recherche d'IdRef afin d'en extraire une liste d'autorités candidates. Cette étape vise à maximiser le taux de rappel : ne pas manquer la bonne autorité, même si l'auteur en question est appelé de manière assez différente dans la notice d'entrée et dans IdRef.

2. ***Peu d'élus***. Pour chaque autorité IdRef candidate, SudocAD agrège toutes les connaissances pertinentes que le Sudoc possède à propos de la personne décrite par l'autorité, compare ces connaissances aux connaissances possédées à propos de l'auteur mentionné dans la notice d'entrée et, après traitement logique en fonction de règles métier, classe chaque candidat dans une des sept catégories, allant de "liage fort" à "liage impossible".

On peut exposer cette démarche sous un autre angle, en explicitant les entrées et les sorties du processus SudocAD :

- **Entrée** : une notice bibliographique, appartenant à un des corpus agrégés par Isidore, en l'occurrence, dans le cadre du prototype expérimental, une notice d'article Persée. Une notice peut mentionner zéro, un ou plusieurs auteurs.

- **Sortie** : pour chaque entrée, un rapport d'analyse en XML qui, pour chaque auteur, classe chacune des autorités candidates dans une catégorie de liage.

Il est important de noter que le processus SudocAD ne produit jamais *de lui-même* un lien entre un auteur de la notice d'entrée et une autorité IdRef. Ce qu'il produit, c'est un rapport d'analyse circonstancié et structuré, qui est exploitable dans un second temps :

- soit pour générer de manière automatique un lien,
- soit pour aider un agent humain à prendre la décision de lier ou non.

Cette déconnexion entre l'analyse par SudocAd et l'acte de liage est un choix délibéré de l'approche SudocAD. Certes, comme on le verra, il est techniquement trivial d'exploiter le rapport d'analyse pour générer un lien de manière automatique, mais nous considérons que, par défaut, c'est aux propriétaires des données d'entrée de décider dans quels cas ils souhaitent déclencher le liage automatique, en fonction de leur propre contexte (niveau d'aversion au risque d'erreur, intégration dans un workflow mixte homme/machine, prise en compte des spécificités des métadonnées du corpus concerné, etc.).

La démarche de SudocAD, étape par étape

On peut décomposer le processus SudocAD en dix étapes :

1. Les notices d'entrée qui ne satisfont pas à un minimum d'exigences sont ignorées (voir plus loin).
2. Chaque notice d'entrée est convertie en RDF/XML, en utilisant l'ontologie OntoSudocAD, extension de FRBROO (voir plus loin).
3. Pour chaque auteur de la notice d'entrée, le nom et le prénom sont passés comme paramètre au web service Find, qui lance une série de requêtes dans IdRef afin de ramener le plus d'autorités jugées pertinentes possibles. L'ensemble des autorités renvoyées constituent l'ensemble des *autorités candidates*.
4. Pour chaque autorité candidate, le web service Link1 est appelé et renvoie la liste des notices bibliographiques liées à cette autorité, en précisant le rôle.[6]
5. Chaque autorité candidate et chaque notice bibliographique liée sont converties en RDF/XML dans le vocabulaire OntoSudocAD.
6. Pour chaque notice d'entrée, l'ensemble de données RDF suivant est chargé par CoGui[7], l'application de raisonnement : la notice d'entrée + { pour chaque auteur : toutes les autorités candidates + toutes les notices bibliographiques liées }.

[6] On ne prend en compte que les notices bibliographiques dont le lien à la notice d'autorité est typé "auteur", directeur de thèse" ou "éditeur scientifique". Dans ce sous-ensemble, on ne prend en compte que les cent premières notices, pour des raisons de performance. On fait l'hypothèse que les cent premières notices bibliographiques rattachées à une autorité sont représentatives des domaines, dates et langues de l'ensemble de ses notices bibliographiques. Cette hypothèse, raisonnable, devra néanmoins être évaluée pour elle-même.

[7] http://www2.lirmm.fr/cogui/

7. CoGui enrichit chaque auteur de la notice Persée et chaque autorité IdRef en synthétisant au niveau de la description de la personne des informations qui portent sur les documents. Par exemple, on attribue à chaque auteur une période d'activité, calculée à partir des dates de publication des documents liés.

8. Pour chaque autorité candidate, CoGui compare un à un chaque attribut de l'autorité avec l'attribut correspondant de l'auteur Persée.

9. CoGui agrège ensuite ces comparaisons d'attributs au moyen de règles métier exprimées selon un formalisme logique, ce qui permet de comparer les descriptions des personnes elles-mêmes. Ceci revient à classer chaque autorité candidate dans une catégorie de liage : liage fort, moyen, faible, etc.

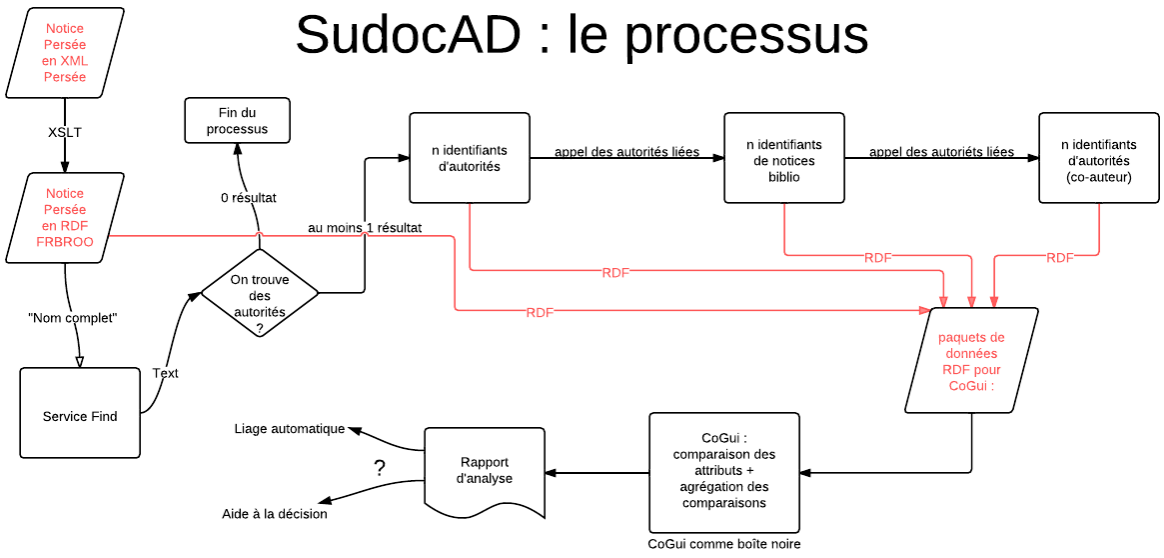
10. CoGui synthétise le résultat de ces traitements dans un rapport d'analyse en XML. Il existe un rapport par notice d'entrée.

L'étape suivante est extérieure au processus SudocAD proprement dit :

11. Un humain peut décider si ces rapports d'analyses sont exploités pour effectuer du liage automatique, pour faire de l'aide à la décision humaine ou encore pour structurer un workflow plus complexe qui combine liage automatique et décisions humaines.

Architecture

La figure suivante schématise le processus SudocAD.



Les étapes 1 à 6 sont entièrement implémentées en web service. Ces web services sont appelés par un programme Java. Ce dernier agrège toutes les données OntoSudocAD renvoyées par les web services et les envoie vers CoGui qui travaille sur les graphes conceptuels correspondant à ces données. En sortie, CoGui fournit au programme Java de quoi générer un rapport d'analyse en XML.

Les notices Sudoc et les notices IdRef sont stockées dans une base Oracle, sous le format UNIMARC/XML. C'est le processeur XSLT d'Oracle qui génère les données OntoSudocAD.

Le service Find est un web service qui agrège un ensemble de requêtes gérées par le moteur de recherche Solr.

Le corpus de test : les métadonnées des articles Persée

Les données livrées par l'équipe Persée

En accord avec ADONIS et l'équipe Persée, ce sont les données de Persée qui ont été choisies pour évaluer les résultats de l'approche de SudocAD. Le lot de données à traiter correspond aux articles des revues du domaine Economie. L'équipe Persée a fourni à l'équipe SudocAD une série de fichiers XML, chaque fichier contenant une année de chacune des neuf revues du périmètre. Ces fichiers ont été découpés et convertis de façon à obtenir un fichier XML par article, c'est-à-dire 13 444 fichiers.[8]

Les notices d'articles ont été fournies dans un format XML propre à Persée, contenant un certains nombre d'éléments descriptifs. En voici la liste :

- Identifiant interne court
- Titre
- Identifiant interne long
- Identifiant DOI
- URL publique
- **Langue**
- **Date de publication**
- ISSN
- Auteur(s)
    - Rang si plusieurs Auteurs
    - Rôle (toujours aut = auteur)
    - Nom complet (pour affichage)
    - Identifiant local de l'auteur (= identifiant unique à l'échelle de la revue)
    - Autorité Persée (éventuelle)
        - Identifiant Persée (unique à l'échelle de tout Persée)

[8] Voir Annexe 1

- **Nom**
- **Prénom**

Les éléments en gras correspondent aux informations qui jouent un rôle effectif dans la recherche de la bonne autorité IdRef, en l'état actuel de SudocAD. Les prochains développements de SudocAD devraient exploiter plus d'informations fournies en entrée. C'est le cas du titre de l'article, par exemple, source d'information importante pour le catalogueur mais ignorée par SudocAD pour l'instant.

Contraintes liées aux données Persée et solutions adoptées

Les données Persée présentaient les défauts suivants :

- Absence de métadonnées sur le contenu (ni classification, ni indexation matière, ni mots-clés)
- Incohérence, parfois, entre le nom complet d'une part et le nom de famille et le prénom d'autre part
- Absence d'auteur

Les notices Persée ne contenaient aucun élément d'information portant sur le contenu intellectuel de l'article décrit : ni indice de classification thématique ou disciplinaire, ni indexation matière qui s'appuierait sur un vocabulaire contrôlé, ni mots-clés libres. Or, dès le début du projet, il paraissait évident que l'analyse du contenu serait un élément décisif pour le liage.

En commun accord avec l'équipe Persée, il a été décidé d'injecter au niveau des notices la classification disciplinaire qui organise les revues Persée dans le portail et dans son serveur OAI-PMH. Cette classification est une liste de disciplines propre à Persée. Chaque revue peut appartenir à une ou plusieurs disciplines.

Ces métadonnées *Discipline* ne peuvent servir au liage que si on peut les comparer à des métadonnées analogues du côté des notices bibliographiques Sudoc. Or, même si le Sudoc encourage l'indexation par la classification Dewey ou permet l'indexation par la classification de la Bibliothèque du Congrès, toutes les notices ne contiennent pas d'indice de classification. Il a donc fallu imaginer une solution *ad hoc* afin de rattacher *indirectement* le maximum de notices bibliographiques Sudoc à une classification comparable à celle de Persée.

Pour ce faire, nous nous sommes appuyés sur les constats suivants :

1. De nombreuses notices Sudoc sont indexées avec RAMEAU (environ 2/3).

2. RAMEAU possède un cadre de classement : théoriquement, chaque terme RAMEAU est associé à un ou plusieurs domaines, codé par un indice relevant d'un cadre de classement issu de Dewey.[9]

Nous avons donc décidé de faire hériter le domaine d'un terme RAMEAU vers chaque notice possédant ce terme en élément d'entrée d'une vedette matière. Par exemple, la notice bibliographique 007344368 contient une vedette Rameau dont la tête est la notice IdRef 027237893[10] ("Monnaie") : via la zone 686, cette dernière est associée à la classe Dewey 330 ("Economie politique"). Ainsi, indirectement, sauf exception, chaque notice bibliographique indexée en RAMEAU est indexée en Dewey.

La dernière étape consiste à permettre la comparaison des disciplines Persée aux Indices Dewey. Pour ce faire, nous avons établi une correspondance entre les disciplines Persée et les indices Dewey. Au moment de la conversion des notices Persée en OntoSudocAD, nous avons fait hériter la ou les discipline(s) d'une revue au niveau de chacun des articles de la revue, en y injectant l'indice Dewey correspondant.

Nous avons donc désormais un indice Dewey (au moins) dans chaque notice Persée et un indice Dewey (au moins) dans un grand nombre de notices bibliographiques Sudoc. Les deux types de notices deviennent donc comparables sous cet aspect.

Désormais, nous ferons allusion à cet indice Dewey sous l'appellation "Domaine".

Code de revue Persée	Titre de la revue	Domaine(s) exprimé(s) en Dewey
ecop	Economie et prévision	330
ecoru	Economie rurale	330
estat	Economie et Statistique	330
hes	Histoire, économie & société	330, 900
ofce	Revue de l'OFCE	330
reco	Revue économique	330
rei	Revue d'économie industrielle	330
rfeco	Revue française d'économie	330
tiers	Revue Tiers monde	300, 320, 330, 390, 915

[9] http://rameau.bnf.fr/utilisation/rameau4_2.htm

[10] Cette autorité est visible en UNIMARC/XML à cette adresse : http://www.idref.fr/027237893.xml

Ontologie

Afin d'établir des liens entre les auteurs Persée et les autorités Sudoc en utilisant les capacités de raisonnement de CoGui, il était nécessaire d'exprimer les données Persée et les données Sudoc (notices bibliographiques et notices d'autorité) dans un vocabulaire commun.

C'est l'ontologie FRBROO[11] qui a été retenue. FRBROO est une réinterprétation du modèle FRBR de l'IFLA[12], modèle conceptuel du domaine de l'information bibliographique. FRBROO réinterprète les concepts des FRBR dans une perspective dite "orientée objet" (OO). Comme le CRM[13], modèle de l'information muséographique dont elle est une extension, l'ontologie FRBROO est exprimée sous la forme d'un schéma RDFS (et le sera aussi, probablement, comme CRM, sous la forme d'une ontologie OWL).

FRBROO a été retenu pour les raisons suivantes :

- C'est un vocabulaire très riche car, couplé avec CRM, il ambitionne de couvrir l'ensemble du champ de la documentation culturelle. Il promet donc de pouvoir exprimer l'essentiel de nos besoins actuels et à venir.
- S'appuyant sur le travail effectué sur le modèle/ontologie CRM CIDOC et le modèle FRBR de l'IFLA, il possède une réelle maturité conceptuelle.
- Il existe déjà une version RDFS de ce vocabulaire.

Pour exprimer en RDF certaines informations présentes dans les notices Persée ou les notices Sudoc et jugées exploitables dans le cadre de SudocAD, il a fallu étendre le vocabulaire FRBROO, en ajoutant dans un nouvel espace de nom les éléments d'information suivants, déclarés comme sous-propriétés ou sous-classes du vocabulaire de base :

- Propriétés correspondants aux codes de fonction UNIMARC
- Classes permettant d'exprimer avec précision tel ou tel type d'identifiant (DOI, identifiant BnF, etc.)
- Classes et propriétés exprimant certaines notions propres aux notices d'autorité : forme retenue, élément d'entrée… Ces extensions sont provisoires car FRBROO intègrera bientôt les concepts de FRAD[14], conceptualisation par l'IFLA des données d'autorité.

[11] http://www.cidoc-crm.org/frbr_inro.html

[12] http://www.ifla.org/publications/functional-requirements-for-bibliographic-records

[13] http://www.cidoc-crm.org/index.html

[14] http://www.ifla.org/publications/ifla-series-on-bibliographic-control-34

- Propriétés permettant d'associer un domaine à une ressource
- Propriétés permettant d'exprimer en RDF des zones de note UNIMARC précises
- Classe "Record" pour décrire la notice

Rétrospectivement, on peut relever que toutes ces extensions n'ont pas été exploitées dans le cadre du prototype, mais elles pourraient l'être dans une phase postérieure, au-delà du projet expérimental.

Comme la stratégie de liage de SudocAD repose sur le raisonnement, tout ce qui contribue au raisonnement et tout ce qui est exprimé en conclusion du raisonnement doivent être exprimés en RDF, comme les notices elles-mêmes, sur lesquelles portent le raisonnement. C'est pourquoi, *Liage*, une ontologie complémentaire, a été créée pour exprimer :

- les propriétés des auteurs résultant de l'enrichissement des autorités (phase 7). Par exemple, cette ontologie possède une propriété exprimant la période d'activité d'un auteur, calculée à partir des dates de publication des documents liés.
- les relations entre auteurs, résultant de la comparaison des auteurs un à un et propriété par propriété (phase8). Par exemple, cette ontologie possède une propriété disant que tel auteur Persée et tel auteur Sudoc ont des domaines correspondant fortement.
- les relations entre auteurs, résultant de l'agrégation des comparaisons de la phase 8 (phase 9). Par exemple, cette ontologie possède une propriété disant que tel auteur Persée et tel auteur IdRef ont une relation de "liage fort'.

L'ontologie utilisée dans le cadre de SudocAD, baptisée OntoSudocAD, peut être résumée ainsi :

OntoSudocAD = CRM + extensions FRBROO + extensions Sudoc + extensions Liage

OntoSudocAD étend des ontologies existantes, mais il s'est avéré impossible d'utiliser tels quels les documents RDFS et OWL publiés par les groupes de travail à l'origine du CRM ou du FRBROO. Sur quelques points mineurs ou formels, il a fallu les amender ou les compléter afin de les rendre opérationnels dans le contexte du raisonneur CoGui. Par ailleurs, une fois OntoSudocAD chargée dans CoGui, quelques propriétés logiques supplémentaires ont été ajoutées, quand le langage OWL ne pouvait les exprimer.

Les principales étapes de SudocAD

Enrichissement des autorités

Pour comparer des autorités, SudocAD compare leurs propriétés. La première étape consiste donc à identifier les propriétés des personnes qui paraissent utiles.

Côté Sudoc

Côté Sudoc, ne s'appuyer que sur le contenu des notices d'autorité pour obtenir les propriétés des personnes correspondantes, ce serait ignorer toute la connaissance implicite contenue dans l'ensemble des notices bibliographiques liées à cette autorité. Au contraire, SudocAD enrichit l'autorité de la personne en faisant "remonter" à ce niveau certaines informations calculées à partir des propriétés des documents liés :

- *Domaine*. A partir de l'indexation Rameau des notices bibliographiques, une ou plusieurs classes Dewey sont attribuées à la personne elle-même. Chaque classe Dewey est pondérée, en raison du nombre de documents qui relèvent de cette classe. Voici un exemple de liste pondérée de domaines rattachés à une autorité, telle qu'on la trouve formulée en XML dans le rapport d'analyse :

```
<domaines>
<domaine_publi domain="650" occurs="7.888888888888889"/>
<domaine_publi domain="330" occurs="0.444444444444444"/>
<domaine_publi domain="510" occurs="0.2222222222222222"/>
<domaine_publi domain="350" occurs="0.2222222222222222"/>
<domaine_publi domain="150" occurs="0.2222222222222222"/>
</domaines>
```

  Ces domaines Dewey sont obtenus de manière indirecte, faute d'une classification Dewey des documents Sudoc : pour chaque vedette Rameau d'une notice, on analyse le domaine Dewey associé à la tête de vedette. Cette information est donnée en zone 686 des notices Rameau – pas de toutes, hélas, ce qui a un impact négatif sur les résultats de SudocAD. Par exemple, la notice bibliographique 007344368 contient une vedette Rameau dont la tête est la notice IdRef 027237893 ("Monnaie") : cette dernière est associée à la classe Dewey 330 ("Economie politique"), ce qui n'est pas encore le cas de la notice 027285332 ("Capitalisme").

- *Période*. La période d'activité d'une personne est calculée en construisant l'intervalle entre la date de la première publication liée à cette personne et la date de la dernière.

- *Langue*. A partir des langues de publications liées à une personne, on détermine son profil linguistique : une liste pondérée de langues.

Ces nouvelles propriétés attribuées à la personne décrite par l'autorité sont ajoutées à la principale propriété que contient celle-ci : la ou les appellations de la personne, qu'il s'agisse de la forme retenue ou des formes rejetées, les deux types de formes étant traités de manière équivalente par SudocAD.

Faute de temps, l'expérimentation n'a pas travaillé sur les propriétés suivantes, qui devraient être exploitées dans un second temps pour améliorer l'efficacité de la méthode de liage :

- Les *sujets* des notices bibliographiques n'ont pas été exploités en eux-mêmes car il n'y en avait pas côté Persée. Côté Sudoc, comme on l'a vu plus haut, l'indexation matière en Rameau a été exploitée, mais de manière très partielle et indirecte : pour chaque vedette Rameau d'une notice, on s'est intéressé au domaine Dewey associé à la tête de vedette. Ce pis-aller est doublement restrictif :
    - On n'utilise pas toute la puissance du réseau sémantique de Rameau : synonymes, termes généraux, termes spécifiques, etc. D'ailleurs, ces relations sémantiques entre concepts ne sont pas encore exprimables en FRBROO, mais le seront d'ici quelques mois. Dès lors, il sera plus facile de comparer l'indexation (libre ou contrôlée) des notices d'entrée avec les notices bibliographiques Sudoc. On aura là un vecteur de liage entre auteurs et autorités très prometteur.
    - Tous les termes Rameau ne sont pas associés à un domaine Dewey. Ces lacunes expliquent un certain nombre des échecs de SudocAD à proposer un liage fort entre un auteur et une autorité : si un auteur n'a pas de domaine, il ne peut être dans une relation de liage fort. Or, si un auteur Sudoc est lié à des notices bibliographiques dont aucune vedette Rameau n'est liée à un domaine Dewey, l'auteur n'est lié à aucune domaine, tout comme s'il n'était lié à aucun document ou qu'aucun de ses documents n'était indexé en Rameau.
- La relation de *co-auteur* n'a pas encore été exploitée. Mais, sans aucun doute, dans un certain nombre de cas, elle aurait permis d'établir une relation de liage fort entre un auteur Persée et une autorité Sudoc.
- Le *rôle* d'une personne par rapport aux documents auxquels il est lié n'a été que partiellement exploité : seuls les documents liés à une personne par la fonction d'auteur, de directeur de thèse ou d'éditeur scientifique ont été retenus.

Côté Persée

Suivant la même logique que pour les notices Sudoc, les notices bibliographiques Persée ont été interprétées comme une source pour attribuer des propriétés à une personne, l'auteur. Ici,

contrairement au Sudoc, un seul document est lié à l'auteur. Les propriétés de l'auteur exploitables par SudocAD sont donc assez pauvres :

- Appellation
- Date
- Domaine
- Langue

Il faut également mentionner les propriétés suivantes, qui ne sont pas exploitées par SudocAd aujourd'hui :

- Titre de l'article dont il est l'auteur
- Nom de la revue qui publie cet article
- ISSN de cette revue
- etc.

Par commodité, on appellera ces ensembles de propriétés rattachés aux auteurs Persée les "autorités Persée". Pourtant, il ne s'agit pas là de véritables autorités, qui décriraient des personnes indépendamment d'*un* document particulier. Au contraire, pour SudocAD, une même personne a autant d'"autorités Persée" qu'elle a d'articles. L'autorité Persée est liée à un document particulier. C'est pourquoi, selon le document (et donc selon le domaine auquel appartient la revue qui le publie), une même personne peut théoriquement être liée à une autorité IdRef ou non, voire à telle autorité IdRef ou telle autre.

Il faut noter que, par ailleurs, l'équipe Persée a commencé à créer de véritables autorités Persée, qui rassemblent autour d'un identifiant unique toutes les publications d'un auteur Persée, soit à l'échelle d'une revue, soit à l'échelle de tout le portail. Pour l'instant, l'équipe SudocAD n'a pas exploité ces véritables autorités SudocAD, pour des raisons méthodologiques et pratiques. Mais il serait intéressant, à tout le moins, de comparer les résultats de ce travail manuel par l'équipe Persée et les résultats de SudocAD : y a-t-il ou non une bijection entre les véritables autorités Persée et les identifiants IdRef associés à des auteurs Persée ?

Comparaison des attributs

Une fois les autorités Sudoc et les "autorités Persée" constituées et enrichies, il s'agit de les comparer. Dans un premier temps, il s'agit de comparer leurs propriétés seulement. Cette opération consiste à isoler une des propriétés et de mesurer si, *sous ce rapport*, la similarité est forte, faible, distante, nulle, etc.

Comparaison des appellations

Même si, selon l'approche logique de SudocAD, la mesure de similarité entre le nom dans Persée et le nom dans IdRef n'est pas un critère suffisant, elle joue un rôle important. Cette mesure utilise différents algorithmes du marché (comme la distance Levenshtein) mais s'attache également à prendre en compte le fait qu'il s'agit de noms propres et non de chaînes de caractère indifférenciées (ex : gestion des prénoms-initiales). Certaines erreurs ou certains silences de SudocAD, d'ailleurs, tiennent à une mesure défaillante de la similarité entre noms quand le nom propre est composé – ce type de défaillance sera corrigé à l'avenir.

Cette mesure va permettre d'établir de nouvelles relations entre les autorités Persée et les autorités Sudoc :

- *dissimilar denomination*
- *distant denomination*
    - *close denomination*
        - *same denomination*

Comparaison des domaines

Le fait qu'une autorité Persée et une autorité IdRef aient des domaines similaires est un facteur important, mais qui peut s'interpréter de différentes manières.

Dans le cas le plus simple, les deux autorités relèvent toutes les deux du même domaine, 330 par exemple. On établira entre elles la relation *domain strong correspondence*.

Mais si l'une relève de 330 (économie politique) et l'autre de 320, va-t-on établir la relation *domain without correspondence* comme on le ferait si l'une des deux relevait de 615 (phamacie) ? ou va-t-on établir la relation *domain weak correspondence*, comme on le ferait si l'une des deux relevait de 900 (histoire) ? C'est plutôt une relation comme *domain intermediate correspondence* qui semble ici pertinente.

On comprend donc que la comparaison des domaines nécessite quatre relations :

- *domain without correspondence*
- *domain weak correspondence*
    - *domain intermediate correspondence*
        - *domain strong correspondence*

Pour comparer les domaines, SudocAD a établi un tableau qui mesure la distance de tous les domaines jugés pertinents pour Persée et ce deux à deux. En voici un extrait :

	330	320	300
330	1	0,9	0,6
320	0,9	1	0,6
300	0,6	0,6	1
900	0,2	0,2	0,1

Ce tableau des distances entre domaines est une pièce essentielle dans le dispositif SudocAD. Cependant, les points suivants doivent être notés :

- Tous les domaines Dewey n'ont pas été traités.
- La mesure de la distance entre domaines comporte nécessairement une part d'arbitraire. Nous n'avons pas pu exploiter de travaux antérieurs qui auraient cherché à fonder une telle mesure sur des bases théoriques ou empiriques plus solides.
- La classification des revues dans Persée se limite au périmètre des SHS. Or, l'évaluation montre que l'absence de domaine 630 (agriculture) pour la revue *Economie rurale* explique une part significative des ratés de SudocAD.

Jusqu'à présent, nous n'avons envisagé que le cas où chaque autorité à comparer ne possède qu'un domaine. Or, côté Persée, une revue (donc ses articles) peut relever de plusieurs domaines. C'est le cas de la revue, *Tiers Monde*, par exemple, qui émarge à quatre domaines. Côté Sudoc, c'est encore plus fréquent : par l'intermédiaire des publications qui lui sont liées, une autorité peut relever de nombreux domaines. Il s'agit donc de comparer entre eux l'ensemble des domaines de l'autorité Persée et l'ensemble des domaines de l'autorité IdRef. Ultime complexité enfin : chacun de ses ensembles est une liste *pondérée* de domaines. En effet, un auteur qui a publié 20 fois en 330 et 1 fois en 900 ne présente pas le même profil qu'un autre auteur rattaché aux mêmes domaines mais en proportions inverses (20 fois 900 et 1 fois 330). Ce sont donc bien des listes pondérées de domaines que SudocAD doit comparer.

Comparaison des périodes

Là encore, la comparaison des périodes d'activité aboutit à poser quatre relations possibles entre autorités :

- *date without correspondence*
- *date weak correspondence*
    - *date intermediate correspondence*
        - *date strong correspondence*

Cette comparaison des périodes s'appuie à la fois sur les dates de vie mentionnées dans certaines autorités IdRef et les dates de publication des documents liés aux autorités.

En l'état, cette mesure ne prend pas en compte le fait que certaines publications sont des rééditions, ce qui trouble l'interprétation des dates. Cette prise en compte des rééditions est envisagée pour la suite, d'autant que le vocabulaire FRBROO s'y prête bien.

Comparaison des langues

La comparaison des langues ne prévoit que deux relations possibles :

- *language without correspondence*
- *language strong correspondence*

En l'état, cette mesure ne prend pas en compte le fait que certaines publications sont des traductions, qui ne reflètent donc pas systématiquement la langue de l'auteur. Là encore, le choix du vocabulaire FRBROO devrait permettre de prendre en compte ce type de situations.

Agrégation des résultats de la comparaison des attributs

C'est à travers la déclaration de règles logiques censées refléter les connaissances des bibliothécaires que la comparaison des propriétés des autorités va permettre la comparaison des autorités elles-mêmes, ou plutôt l'établissement de relations de liage entre ces autorités :

- LiageAutoriteStrong
- LiageAutoriteMedium
- LiageAutoriteWeak
- LiageAutoritePoor
- LiageAutoriteNeutral
- LiageAutoriteUnrelated
- LiageAutoriteImpossible

Par exemple, la règle liageStrong2 peut s'écrire :

si

SameDenomination(id_Persee,id_sudoc) et

DateIntermediateCorrespondence(id_Persee,id_sudoc) et

DomainStrongCorrespondence(id_Persee,id_sudoc) et

LanguageStrongCorrespondence(id_Persee,id_sudoc)

alors

StrongLinkage(id_Persee,id_sudoc)

Cette règle signifie que, pour le système, l'autorité Persée identifiée par id_Persee et l'autorité IdRef identifiee par id_sudoc, ne peuvent être distinguées.

A ce stade, vingt-et-une règles ont été déclarées. L'architecture de SudocAD rend aisé l'ajout de nouvelles règles logiques, reflet des connaissances et des raisonnements des catalogueurs.

Les rapports d'analyse et leurs diverses exploitations possibles

L'enchaînement automatique de toutes les étapes précédentes constitue une chaîne de traitement dont l'entrée est une notice bibliographique en OntoSudocAD (une notice Persée, en l'occurrence) et la sortie un rapport d'analyse en XML qui, pour chaque auteur de la notice, liste les candidats possibles et les répartit entre les sept catégories de liage. Ce rapport mentionne également, pour chaque candidat, les résultats de chaque étape intermédiaire : résultats des requêtes du service Find ; indicateur de similarité des domaines, des dates, des langues et des noms ; liste pondérée des domaines…

L'annexe 4 renvoie vers un exemple de rapport d'analyse, correspondant à la notice Persée reco_0035-2764_1959_num_10_6_407388.

Grâce aux informations précisées et structurées qu'il contient, ce rapport d'analyse XML peut être exploité de manière automatique

- soit pour alimenter un processus de liage automatique,
- soit pour configurer une fonctionnalité d'aide à la décision de liage dans une interface homme-machine.

Exploiter le rapport d'analyse pour faire du liage automatique

Pour enclencher un processus de liage automatique à partir du rapport d'analyse, il faut utiliser un algorithme spécifique qui détermine quelle autorité IdRef unique est sélectionnée parmi toutes les autorités candidates réparties entre les sept catégories de liage. En voici deux exemples possibles :

- On sélectionne l'autorité IdRef qui est la seule autorité dans la catégorie LiageAutoriteStrong
- On sélectionne l'autorité IdRef qui est la seule dans la catégorie la plus forte parmi les catégories suivantes : LiageAutoriteStrong, LiageAutoriteMedium, LiageAutoriteWeak.

Comme le processus d'évaluation le montre (voir plus loin), selon l'algorithme de liage automatique choisi, le nombre de bons liens créés, le nombre de bons liens manqués ou encore le nombre de mauvais liens créés peuvent varier. Ainsi, selon qu'un fournisseur de données à lier à IdRef souhaite maximiser le nombre de liens créés (au risque de créer de mauvais liens) ou, au contraire, minimiser le nombre de mauvais liens créés (au risque du silence), il choisira tel ou tel algorithme. Il est du ressort du fournisseur de données de prendre cette décision, en fonction de ses préférences (notamment de son niveau aversion au risque d'erreurs) et peut-être également en fonction de la nature ou de la qualité des données d'entrée.

Dans une prochaine version, cette décision sera probablement intégrée dans le processus global de SudocAD. En entrée, le fournisseur fournirait non seulement la notice bibliographique mais également un indicateur correspondant à un algorithme de liage automatique. En sortie, le système pourrait ajouter au rapport d'analyse la mention de *la* notice d'autorité IdRef sélectionnée pour le liage automatique *en fonction de l'algorithme donné en entrée*. C'est précisément ce fait le fichier de synthèse livré après traitement des 13 444 notices Persée (Voir l'annexe 6).

Exploiter le rapport d'analyse pour faire de l'aide à la décision

Le rapport d'analyse peut aussi être utilisé pour aider un catalogueur à lier une notice d'entrée à IdRef. Cette situation d'aide à la décision peut être vue soit comme une alternative au liage automatique (le fournisseur de données considérant que l'acte de lier doit toujours être assumé par un humain), soit comme une étape complémentaire (un humain étant censé prendre une décision pour les seuls cas que le liage automatique n'a pas réglés, cas plus ou moins nombreux selon l'algorithme de liage automatique choisi).

Comme pour le liage automatique, il existe différentes possibilités d'exploiter le rapport d'analyse pour configurer une interface d'aide à la décision. En voici quelques exemples :

- L'interface propose la liste de tous les candidats, mais triés dans l'ordre des catégories de liage, de LiageAutoriteStrong à LiageAutoriteImpossible.
    - Variante : l'interface ne propose pas les candidats des catégories LiageAutoriteUnrelated et LiageAutoriteImpossible.
- L'interface propose différents paliers de candidats, chaque palier correspondant à une catégorie de liage. Le catalogueur progresse de palier en palier, tant qu'il ne lie pas ou

n'abandonne pas (soit pour laisser l'auteur de la notice sans lien à IdRef, soit pour créer une nouvelle autorité IdRef) .

- Variante : l'interface ne propose pas de palier pour les candidats des catégories LiageAutoriteUnrelated et LiageAutoriteImpossible.

- L'interface se contente de demander à l'humain s'il accepte ou non le candidat sélectionné par l'algorithme de liage automatique.

Bien que cette question soit abordée plus loin, le projet SudocAD n'avait pas pour objectif d'étudier en détail les conditions d'exploitation du rapport d'analyse pour de l'aide à la décision. Mais il est acquis qu'au-delà du présent projet, il s'agit d'un axe de recherche utile et prometteur pour les équipes ABES et GraphIK.

Evaluation

Le protocole d'évaluation

Evaluer le traitement effectué par SudocAD consiste à se demander si le programme a eu raison ou pas de créer tel lien ou de ne pas créer de lien. Or, comme on vient de le voir, le programme ne crée pas de lien, mais répartit les candidats entre sept catégories. Comment donc comparer les résultats du programme avec les choix d'un catalogueur, considéré comme point de référence ?

On pourrait demander à un catalogueur d'examiner les auteurs des notices Persée et de répartir toutes les autorités IdRef candidates entre les sept catégories de liage. Il serait alors facile de comparer les choix humains et les résultats automatiques. Nous avons rejeté cette méthode d'évaluation car il nous semblait artificiel d'exiger du catalogueur une telle capacité de discrimination, dans une situation où son objectif pratique est strictement binaire : faut-il lier ou non ? Il aurait été difficile d'interpréter les choix humains et le catalogueur, lui-même, aurait été bien incapable d'expliquer pourquoi telle autorité était classée en LiageAutoriteWeak plutôt qu'en LiageAutoritePoor.

Nous avons préféré mettre le catalogueur dans une situation plus familière. Nous avons présenté à un catalogueur 150 notices Persée. Ces notices contenaient également le titre de l'article ou de la revue, à savoir des informations que le programme n'a pas exploitées. Nous avons demandé au catalogueur de chercher dans IdRef l'autorité correspondant à chacun des auteurs de la notice Persée, de noter la décision prise (liage ou non, avec doute ou pas, en faveur d'autres notices ou pas) et de commenter sa décision. Les décisions qu'il pouvait prendre étaient les suivantes :

- Lier sans aucun doute
- Lier avec doute
    - Avec mention ou pas d'autres autorités qui auraient pu convenir
- S'abstenir de lier, sans aucun doute
- S'abstenir de lier, avec un doute
    - Avec mention ou pas d'autorités qui auraient pu convenir

Voici un extrait du tableau Excel que le catalogueur avait à remplir :

Auteur Nom	Auteur localId	Vous liez ou non ? (O/N)	Si oui, vers quel PPN ?	Des doutes sur votre liaison ? (O/N)	Autres PPNs qui auraient pu convenir ?	Pourquoi avez-vous fait ce choix (de liaison ou de non liaison) ?
Louis Cassagnes	auteur_reco_9157	oui	081379064	non		Nom + prénom + discipline + période chronologique correspondent. Pas d'homonyme

Dans un premier temps, le catalogueur a opéré dans des conditions proches des conditions ordinaires, à savoir : pas de recours à des informations extérieures, temps limité (maximum cinq minutes par auteur), consigne de veiller à ne surtout pas créer de mauvais lien. En comparant les résultats du programme à ces résultats humains, certaines différences observées nous ont conduits à vérifier certaines décisions humaines en ayant recours à des ressources extérieures (sites Web d'auteurs, bibliographies) et sans s'imposer de temps limité. Ce faisant, certains choix humains sous contraintes sont apparus comme erronés et ont été corrigés.

Grâce à ces investigations et ces corrections, nous disposons donc d'un nouveau corpus de liages très fiable, qui nous sert de corpus de référence. C'est ce corpus de liages qui a été comparé aux résultats du programme de SudocAD. Il pourrait également être comparé à d'autres méthodes de liage automatiques ou semi-automatiques, dans le cadre d'un benchmark. Mais il peut également être comparé aux premiers liages humains, effectués sous certaines contraintes de temps et de sources d'information. On observerait alors que le taux d'erreur (création de mauvais liens) des catalogueurs travaillant sous ces contraintes est faible mais non négligeable, et tout à fait comparable au taux d'erreur de SudocAD.

Voici la synthèse des liages effectués par un catalogueur sans contrainte de temps ni de sources d'information sur les 150 notices de notre échantillon Persée :

	Pas de doute	Doute			Total
		Autre candidat en lice	Pas d'autre candidat en lice		
Liage	**146**	3	19	**22**	*168*
Pas de liage	**37**	7	0	**7**	*44*
Total	**183**			*29*	*212*

On constate que le catalogueur a décidé de lier dans 79% des cas (168 liages sur 212 auteurs Persée). Il doute de ces liens dans 13 % des cas (22 doutes sur les 168 liages).

Enfin, précisons que certaines autorités Persée, mentionnées dans les 150 notices, n'ont pas été évaluées pour les raisons suivantes :

- Six auteurs ont été écartés parce qu'il y avait une incohérence entre le nom complet utilisé par le catalogueur et le couple {nom, prénom} utilisé par le programme. Par exemple, une notice Persée nommait le même auteur avec "Gérard Winter" et le couple {Winter, Georges). Il n'aurait pas été judicieux d'inclure ces cas incohérents dans l'évaluation.

- Pour des raisons pratiques, cinq autres autorités ont été écartées quand la requête Find correspondante renvoyait plus de 100 résultats. C'est le cas, par exemple de {Louis, Jacques}, surtout sous la forme {Louis, J*} comme le reformule une des requêtes de Find. Il est d'autant plus regrettable d'avoir écarté ces autorités prolixes qu'il est probable que, précisément dans ce type de situations, l'approche SudocAD aurait été bien supérieure à d'autres approches, y compris le lien manuel par le catalogueur. Cette mise à l'écart n'est que provisoire.

Indicateurs de liage automatique

Comme nous l'avons écrit plus haut, il n'est pas possible de comparer directement les décisions de liage ou de non-liages humaines au traitement SudocAD, qui distribue les autorités candidates parmi sept catégories. Certes, nous pourrions comparer les décisions humaines à la seule catégorie LiageAutoriteStrong, mais pourquoi ne pas plutôt comparer cette catégorie de liage SudocAD la plus forte avec la décision humaine la plus ferme, c'est-à-dire quand le doute est absent ?

On le voit, il n'y a pas une seule manière de comparer les résultats humains aux résultats du programme SudocAD. Sans même parler de l'aide à la décision, on peut construire plusieurs *modèles de liage automatique* possibles, construits à partir des sept catégories de liage. Ce sont ces modèles de liage automatique que nous allons comparer aux décisions humaines.

Voici les modèles de liage automatique retenus :

- 14G1 : liage automatique si un seul candidat dans la catégorie LiageAutoriteStrong

- 14G2 : liage automatique si un seul candidat dans la catégorie de liage la meilleure parmi LiageAutoriteStrong et LiageAutoriteMedium

- 14G3 : liage automatique si un seul candidat dans la catégorie de liage la meilleure parmi LiageAutoriteStrong, LiageAutoriteMedium et LiageAutoriteWeak

- 14G4 : liage automatique si un seul candidat dans la catégorie de liage la meilleure parmi LiageAutoriteStrong, LiageAutoriteMedium, LiageAutoriteWeak et LiageAutoritePoor

Il s'agit maintenant de comparer chacun de ces modèles aux décisions humaines. Chaque décision automatique de liage ou de non liage est comparée à chaque décision humaine et cette comparaison peut prendre une des valeurs suivantes :

- Bonne décision
    - Liage humain et liage automatique correct
    - Pas de liage humain et pas de liage automatique
- Décision acceptable
    - Liage humain et liage automatique vers un candidat en doute
    - Pas de liage humain et liage automatique vers un candidat en doute
- Mauvaise décision (un mauvais lien est créé)
    - Liage humain et liage automatique incorrect
    - Pas de liage humain et liage automatique incorrect
- Décision de prudence
    - Liage humain (avec ou sans doute) et pas de liage automatique

Voici le tableau qui résume cette comparaison, pour chacun des modèles de liage automatique :

		Bonne décision	Décision acceptable	Mauvaise décision	Prudence
14G1	1 candidat dans S	54,7%	0%	1,89%	43,4%
14G2	1 candidat dans + fort de {SM}	77,36%	0,47%	1,89%	20,28%
14G3	1 candidat dans + fort de {SMW}	80,19%	0,47%	3,77%	15,57%
14G4	1 candidat dans + fort de {SMWP}	86,79%	0,94%	6,6%	5,66%

Nous avons également mesuré l'efficacité d'un autre modèle de liage automatique qui associe les exigences logiques de 14G2 avec le fait statistique de la rareté des homonymes. Ce modèle, 14G2H peut se résumer ainsi :

- 14G2H : liage automatique si
    - même situation que 14G2

      ou

    - si un seul candidat dans l'union des catégories LiageAutoriteStrong, LiageAutoriteMedium, LiageAutoriteWeak et LiageAutoritePoor et si ce candidat est le seul à avoir la même *denomination* que l'autorité Persée (la valeur de l'attribut *denomination* de l'autorité IdRef est "same")

Le modèle 14G2H dit en substance : dans le cas où 14G2 n'est pas avéré, si une autorité IdRef est la seule à n'être ni *impossible*, ni *unrelated* ni *neutral* et que c'est la seule à posséder le *même* nom que l'autorité Persée, alors on peut considérer que c'est la bonne. De fait, les résultats de ce modèle sont très bons, comme le montre le tableau qui suit : le taux de bonne décision monte à plus de 88% et le taux d'erreur n'est pas supérieur à celui de 14G2.

		Bonne décision	Décision acceptable	Mauvaise décision	Prudence
14G2H	1 candidat dans S	88,21%	0,94%	1,89%	8,96%

Le fait que ce modèle repose en partie sur une dimension statistique le rend fragile, malgré ses résultats excellents appliqué à cet échantillon Persée. Il faudra mesurer son efficacité sur d'autres corpus. Néanmoins, le tableau de synthèse du traitement des 13 444 notices Persée possède une colonne qui contient les identifiants des autorités IdRef sélectionnées par 14G2H, à côté d'une autre colonne correspondant aux décisions de 14G2, dont les résultats sont un peu moins bons mais qui semblent plus solides d'un point de vue théorique.

Indicateurs d'aide à la décision

Afin de commencer à mesurer le potentiel d'une exploitation du rapport d'analyse de SudocAD au sein d'un dispositif d'aide à la décision, trois indicateurs ont été calculés :

- L'indicateur de **rappel** cherche à mesurer si les meilleures catégories de liage contiennent bien le candidat choisi par le catalogueur (ou un candidat mis en doute). Il s'agit d'éviter d'imposer du **silence** au catalogueur en situation de liage. Un taux de rappel de 100% signifie qu'aucun bon candidat ne manque.

- L'indicateur de **précision** cherche à mesurer combien de candidats refusés par le catalogueur demeurent dans les meilleures catégories. Il s'agit d'éviter d'imposer du **bruit** au catalogueur en situation de liage. Un taux de précision de 100% signifie qu'on ne propose que le bon candidat au catalogueur (ou un candidat acceptable, car "en doute").

- L'indicateur de **pertinence** cherche à mesurer combien de candidats le catalogueur va devoir passer en revue pour trouver la bonne autorité, s'il commence par le candidat classé dans la catégorie la plus forte. Il s'agit d'éviter au catalogueur d'avoir à examiner trop de candidats avant de trouver le bon. Un taux de pertinence de 100% signifie ici que le premier candidat à examiner est le bon. Un taux de pertinence de 50 % signifie que le catalogueur doit examiner deux candidats (dont le bon) pour trouver le bon.

Ces indicateurs sont mesurés dans deux configurations :

1. On ne prend pas en considération les autorités candidates classées dans LiageAutoriteImpossible.

2. On ne prend pas en considération les autorités candidates classées dans LiageAutoriteImpossible et LiageAutoriteUnrelated .

	On ignore LiageAutoriteImpossible			On ignore LiageAutoriteUnrelated et LiageAutoriteImpossible		
	Rappel	Précision	Pertinence	Rappel	Précision	Pertinence
Pas de liage sans doute	-	-	-	-	-	-
Pas de liage avec doute	100,00%	45,41%	60,71%	92,86%	43,94%	56,12%
Liage sans doute	100,00%	77,57%	94,32%	100,00%	78,76%	94,32%
Liage avec doute	100,00%	68,01%	95,24%	98,48%	68,25%	91,71%

On constate que le taux de rappel est toujours très élevé, avec un léger fléchissement pour la catégorie peu fournie par ailleurs *Pas de liage avec doute*.

Le taux de précision est toujours supérieur à 50%.

Dans le cas où le catalogueur a choisi de lier, le taux de pertinence est proche de 100%. Cela signifie que, à de rares exceptions près, si le catalogueur parcourt les autorités candidates en partant de la catégorie de liage la plus forte, la première est la bonne – et ce même dans les cas où sa catégorie de liage n'est pas très forte, comme LiageAutoriteWeak, et qu'on n'envisagerait pas un liage automatique .

Conclusions et perspectives

Enseignements généraux

Les résultats de l'évaluation montrent la validité de l'approche de SudocAD pour un objectif de liage *automatique*. Certes, il faudrait entreprendre un benchmarking pour confirmer le fait qu'un taux de bonne décision compris entre 77 et 88% est un excellent taux, mais on peut d'ores et déjà établir qu'un taux d'erreur inférieur à 2%, s'il est confirmé sur d'autres corpus, est très bon et probablement guère différent du taux d'erreur des humains en situation de catalogage ordinaire, soumis aux contraintes de temps. En effet, d'après nos analyses, le modèle 14G2 a créé 4 mauvais liens mais le catalogueur sous contrainte de temps en a créé au moins 3.

De même, comme le montre l'excellent taux de pertinence, SudocAD semble prometteur pour de nombreux scénarios d'aide à la décision. Le scénario le plus convaincant semble celui qui ferait coexister liage automatique avec le modèle 14G2 et aide à la décision pour traiter manuellement mais efficacement le reliquat. Dans un tel scénario, le gain de temps procuré par SudocAD est double :

- Travail évité grâce au liage automatique
- Travail facilité pour le catalogueur grâce au filtrage et au tri proposés par SudocAD

Par ailleurs, l'analyse des ratés de l'expérimentation confirme également le potentiel de l'approche de SudocAD. En effet, 3 des 4 mauvais liens créés par SudocAD s'expliquent

- soit par les limites des fonctions actuelles de comparaison des noms en présence d'un nom composé,
- soit par des erreurs de liage dans le Sudoc,
- soit par le fait que la revue "économie rurale" ne soit classée qu'en 330 (économie) et non en 630 (agriculture), ce qui s'explique par la dimension SHS de Persée mais peut être considéré comme une anomalie dans les données d'entrée.

Ces raisons expliquent aussi une grande partie des décisions trop prudentes de certains modèles de liage automatique de SudocAD qui, dans le doute, s'abstiennent. Par ailleurs, ces décisions s'expliquent souvent par le peu de notices bibliographiques rattachées à certaines autorités, ou par l'absence d'indexation Rameau ou encore l'absence de domaine Dewey associé à une notice Rameau.

Facteurs d'amélioration de SudocAD

Une analyse fine des ratés (erreurs ou prudence) permet de lister les points sur lesquels les prochaines versions de SudocAD devront apporter des améliorations :

- Amélioration des fonctions de comparaison des propriétés des autorités. Exemple : comparaison des noms, comparaison des dates, comparaison des domaines (exemple : meilleure pondération des domaines en fonction du nombre des publications).
- Prise en compte de nouvelles propriétés :
    - Titre des notices bibliographiques. Utilisation de techniques de traitement automatisé des langues naturelles pour comparer les titres des notices d'entrée et les titres des notices rattachées à une autorité IdRef.
    - Indexation matière. Comparaison des mots-clés ou descripteurs normalisés des notices d'entrée avec les vedettes Rameau. Exploitation des relations sémantiques internes à Rameau.
    - Indices de classification (CDU, Dewey, etc.)
    - Co-auteurs
- Prise en compte de davantage de rôles. Prise en compte du rôle "traducteur", par exemple, et de son impact sur la propriété "langue".
- Evaluation et correction préalables des liens internes au Sudoc. En effet, SudocAD se trompe parfois parce que le Sudoc se trompe lui-même, en raison d'un mauvais lien établi par un catalogueur Sudoc ou par un programme de liage automatique rustique. Les bons liens internes au Sudoc sont la condition de possibilité des bons liages externes.
- Affinage et enrichissement des règles logiques.

Conditions de passage à l'échelle et de généralisation

A ce stade, l'approche SudocAD a été testée sur un échantillon de données Persée. Dans les prochaines étapes, il s'agira d'appliquer la même approche à des données diverses et plus massives. En d'autres termes, les défis à venir sont la généricité de l'approche et le passage à l'échelle.

Généricité

Dès sa conception, SudocAD a été conçu de manière générique, de façon à ce que les programmes puissent s'appliquer à des données d'entrée qui ne soient pas Persée et au moyen

d'une base de connaissance qui ne soit pas le Sudoc. Les points qui plaident en faveur de la généricité de SudocAD sont les suivants :

- Fonctionnement par web services
- Vocabulaire neutre : FRBROO
- Représentation formelle de la connaissance sous forme de règles logiques extensibles

Pour autant, en l'état, le prototype de SudocAD n'est pas aussi générique qu'il pourrait l'être, pour les raisons suivantes :

- L'absence d'indexation matière dans Persée a rendu moins urgent l'intégration des mots-clés et des vocabulaires contrôlés comme Rameau dans le dispositif de comparaison et de raisonnement.
- La comparaison des domaines est configurée pour les domaines de Persée. Il faut donc poursuivre le travail engagé en complétant le tableau de distances entre domaines Dewey.
- Toute notice d'entrée doit être exprimée en FRBROO, qui n'est pas un vocabulaire très utilisé. Mais, précisément, FRBROO se positionne comme une ontologie pivot, avec laquelle la plupart des autres vocabulaires peuvent être alignés. On pourrait même imaginer qu'un service de liage automatique basé sur SudocAD proposerait en entrée un service de conversion en FRBROO, ce qui dispenserait les fournisseurs de données d'effectuer eux-mêmes cette conversion.

Passage à l'échelle

En entrée, le processus SudocAD part d'une seule notice. En sortie, un seul rapport d'analyse est produit. Mais, dans l'intervalle, une grande quantité de notices peut être appelée et analysée. Chaque article peut contenir plusieurs auteurs. Chaque auteur peut appeler plusieurs dizaines d'autorités IdRef candidates. Chaque autorité candidate peut appeler des dizaines de notices bibliographiques. Chaque notice bibliographique peut contenir plusieurs vedettes Rameau qu'il faut aller consulter pour en connaître le domaine. De requêtes en rebonds, de notices en notices, une grande quantité de données sont extraites d'une base de données MARCXML, converties en FRBROO, transférées par HTTP puis chargées dans Cogui et enfin intégrées dans le processus logique de raisonnement. En moyenne, ce système traite 500 notices à l'heure, ce qui revient à traiter une notice d'entrée toutes les sept secondes.

Pour un traitement automatique de quelques dizaines de milliers de notices d'entrée, ces performances sont viables. Au-delà de cet ordre de grandeur, la solution actuelle n'est guère praticable. Par exemple, il ne serait pas envisageable d'utiliser cette chaîne telle quelle pour traiter le Sudoc lui-même afin de vérifier la validité des liens internes actuels entre notices bibliographiques et notices d'autorité.

Pour une aide à la décision, c'est-à-dire dans un contexte d'interaction hommes-machines, une attente de sept secondes ne semble pas intolérable, surtout quand on rapporte cette attente au

gain de temps procuré par ailleurs. Mais cette moyenne de sept secondes cache de grandes disparités. Certaines notices exigent un temps de traitement de plusieurs dizaines de secondes.

Toutefois, il faut garder à l'esprit que ces performances sont dues en partie au souci de conserver une architecture qui soit la plus générique possible. Ainsi, la base de données qui gère les notices IdRef et Sudoc n'a pas été configurée pour répondre de manière spécifique aux attentes de SudocAD : aucune des propriétés exploitées par SudocAD n'a été pré-calculée. De même, l'utilisation de web services n'optimise pas l'accès aux données. On peut donc prévoir que, dans une situation où la question du temps de traitement deviendrait critique, des solutions *ad hoc* pourraient être trouvées. Ainsi, la mise en cache des notices bibliographiques et des notices IdRef a déjà permis de gagner 25% de temps de traitement.

Perspectives

Au-delà de l'échantillon de 150 notices, le processus SudocAD a été appliqué à plus de 13 000 notices Persée. Les résultats seront livrés sous la forme d'un fichier Excel, avec une ligne par auteur Persée (voir l'annexe 6). Ce fichier contient les colonnes suivantes : nom de l'auteur, identifiant de l'auteur, identifiant de l'article, colonnes de liage automatique 14G2 et 14G2H, une colonne par catégorie de liage.

A court terme, les priorités de l'ABES et de GraphIK sont les suivantes :

- Améliorer encore les résultats en travaillant sur certaines marges de progression identifiées et mentionnées plus haut.
- Appliquer le processus SudocAD à d'autres corpus de métadonnées en SHS.

A moyen terme, les objectifs sont les suivants :

- Appliquer le processus SudocAD à d'autres corpus de métadonnées hors SHS.
- Intégrer l'indexation Rameau et l'analyse des co-auteurs dans le processus.
- Elaborer un prototype qui proposerait une analyse SudocAD comme aide à la décision au sein de l'application en production IdRef.

L'ABES et l'équipe GraphIK ont la volonté de prolonger ces premiers travaux. C'est pourquoi, en septembre 2011, elles ont déposé un dossier de projet ANR à l'occasion d'un appel d'offres du programme "Contenu et Interactions".

Annexes

Annexe 1. Notice Persée telle que fournie par l'équipe Persée

<notice id="reco_407388">

<titre>Les variations de la circulation fiduciaire de 1954 à 1958</titre>

<id type="Persee">reco_0035-2764_1959_num_10_6_407388</id>

<id type="DOI">10.2307/3498605</id>

<id type="URL">http://www.persee.fr/web/revues/home/prescript/article/reco_0035-2764_1959_num_10_6_407388</id>

<langue>fre</langue>

<datePub>1959</datePub>

<issn>0035-2764</issn>

<auteur rang="1" marcRole="aut">

<nom type="affichage">Pierre Berger</nom>

<localId>auteur_reco_5463</localId>

<autorite id="persee_24348">

<nom>Berger</nom>

<prenom>Pierre</prenom>

</autorite>

</auteur>

<auteur rang="2" marcRole="aut">

<nom type="affichage">Louis Cassagnes</nom>

<localId>auteur_reco_9157</localId>

<autorite id="persee_31094">

<nom>Cassagnes</nom>

<prenom>Louis</prenom>

</autorite>

</auteur>

</notice>

Annexe 2. Notice Persée convertie en FRBROO étendu (OntoSudocAD)

Voir le fichier reco_0035-2764_1959_num_10_6_407388_frbroo.rdf

Annexe 3. Tableau listant les métadonnées fournies par Persée et leur utilisation dans le processus SudocAD

	Utilisation dans le processus SudocAd		
	Converti en FRBROO ?	Utilisé par le service FIND ?	Attribut comparé ?
Id interne court	■		
Titre	■		
ID interne long	■		
DOI	■		
URL			
Langue	■		■
Date	■		■
ISSN			
Auteur-Rang			
Auteur-Rôle	■		
Auteur-Nom	■		■
Auteur-IDlocal	■		
Auteur-Autorité-IDPersée	■		
Auteur-Autorité-NomFamille	■	■	■
Auteur-Autorité-	■	■	■

Prénom			
Co-Auteur	■		
Domaine	■		■

Annexe 4. Exemple de rapport d'analyse, correspondant à la notice Persée reco_0035-2764_1959_num_10_6_407388

Voir le fichier reco_0035-2764_1959_num_10_6_407388.rapport.xml

Annexe 5. Tableau d'analyse de l'échantillon de test (150 notices Persée)

Voir le fichier EVAL14G2.xls

Annexe 6. Tableau de synthèse du traitement SudocAd des 13 444 notices Persée

Voir le fichier PerseeSynthese.xls