Submitted on: 06.06.2017

Lier à IdRef. Un réseau pour la consultation et la production de données d'autorité, une collaboration entre les hommes et les machines

Aline Le Provost
Agence Bibliographique de l'Enseignement Supérieur, Montpellier, France.
le-provost@abes.fr

Yann Nicolas
Agence Bibliographique de l'Enseignement Supérieur, Montpellier, France.
nicolas@abes.fr

François Mistral
Agence Bibliographique de l'Enseignement Supérieur, Montpellier, France.
mistral@abes.fr

Copyright © 2017 by Aline Le Provost, Yann Nicolas and François Mistral. This work is made available under the terms of the Creative Commons Attribution 4.0 International License: http://creativecommons.org/licenses/by/4.0

Abstract:
Cet article présente IdRef (Identifiants et Référentiels), base d'autorités, mais aussi boîte à outils et services à destination de la communauté de l'enseignement supérieur français. Si les bibliothèques universitaires en sont les utilisatrices historiques (depuis la fondation du catalogue collectif Sudoc en 2000), des acteurs nouveaux sont venus s'y greffer, favorisant l'interopérabilité des systèmes associés aux données de la recherche. Pour accompagner cette ambition, l'ABES a cherché des moyens automatisés pour générer des liens vers les notices d'autorité, mais également diagnostiquer et soigner la qualité des liens existants. Bien que les données à relier soient massives, englobant les auteurs des articles, la qualité des liens est une priorité car l'interconnexion systémique et la réutilisation des données dans le LOD (Linked Open Data) sont un vecteur de contagion de ce qui est bon et mauvais. Seront détaillés une partie des résultats de Qualinca, projet de recherche sur la qualité des liens dans les catalogues documentaires. Outre des algorithmes et du code, Qualinca a produit une interface prototype conçue comme un lieu où les humains et les algorithmes collaborent. Pour un nom donné, les algorithmes calculent de nouveaux liens et font une évaluation des liens existants, que l'agent humain peut confirmer ou annuler. L'agent peut également donner un coup de pouce à l'algorithme lorsque celui-ci a en a besoin. Aujourd'hui, le catalogage unitaire manuel ne suffit plus. Il est nécessaire d'inventer de nouveaux environnements, à géométrie variable (enregistrement unitaire, ensemble d'enregistrements, bases de données) et à démographie mixte (humains plus algorithmes), dans lesquels la coopération et la qualité restent les piliers.

Keywords: Contrôle d'autorité, Qualité, Alignements de données, Identifiants, Automatisation

1 INTRODUCTION

L'ABES a été créée pour constituer et coordonner le Sudoc en tant que coopérative de catalogage des universités françaises en 2000. Progressivement, le périmètre initial s'est élargi

aux bibliothèques de nombreuses institutions de l'Enseignement Supérieur : à ce jour près de cent cinquante bibliothèques de taille et de collections variées.

Depuis cette initiative fondatrice, l'ABES s'est diversifiée en développant de nouveaux catalogues et bases de données à contenu spécifique : thèses électroniques[1] (2006), archives et manuscrits[2] (2007), archives d'éditeurs scientifiques[3] (2011). Afin de contrebalancer cette stratégie de diversification, contraire à la vision originale du Sudoc comme base de données exclusive et exhaustive, l'ABES a choisi de miser sur la puissance des données d'autorité. Plutôt que de regrouper toutes nos données dans une base intégrée à son propre fichier d'autorités local, il a été décidé d'accorder l'indépendance à ces données d'autorité, tant du point de vue technique que commercial, et de le rendre ouvert aux partenaires. Les autorités Sudoc sont mortes, vive "IdRef" (2010-).

2 IDREF – LIBERER LES AUTORITES EN MASSE

IdRef est une application générique et ouverte qui permet à n'importe qui d'accéder aux données d'autorité sur le web en lecture et à n'importe quel partenaire authentifié d'éditer, à partir de son propre environnement, les données.

IdRef en chiffres

- 3,2 millions de notices d'autorité, dont :
  - 2,6 millions de notices Personnes
  - 300 000 notices de Collectivités
  - 75 000 notices de Lieux géographiques
- 2 référentiels d'indexation matière : Rameau et FMeSH
- 180 000 notices d'autorité créées par an

Figure 1 – IdRef, un réseau de production dynamique

IdRef dispose de plusieurs dispositifs d'interopérabilité : base vivante et administrée de notices d'autorité identifiées de façon unique et pérenne ; site web pour rechercher, consulter et rebondir vers des catalogues documentaires ; outil professionnel full web pour produire des données d'autorité ; portail d'accès aux données ouvertes sous licence ouverte, réutilisables par web services, requêtables *via* un moteur Solr et moissonnables en OAI-PMH ; « interprète » des données en de multiples formats standards pour les humains et pour les machines (html, xml, rdf, json).

Enfin, depuis l'origine aussi, IdRef agit pour le web de données. Voilà maintenant 7 ans que toutes les données d'autorité sont exposées au format RDF, à la volée pour une parfaite fraîcheur. Ces données seront bientôt interrogeables dans un sparql endpoint dédié.

3 IDREF – PRENDRE EN CHARGE LA QUALITE DES DONNEES

Cette visibilité et cette accessibilité accrue des données d'autorité obligent à plus de responsabilité en matière de qualité. Dans un réseau décentralisé où l'activité de signalement

[1] https://www.theses.fr/
[2] http://www.calames.abes.fr/
[3] Au travers du projet ISTEX : http://www.istex.fr/istex-excellence-initiative-of-scientific-and-technical-information/

documentaire est continuelle et massive, c'est un véritable catalyseur. Parce que ces données sont ouvertes et réutilisées, les demandes de correction proviennent également de tiers. La question de la qualité des données d'autorité fait donc partie intégrante du projet IdRef, qualité entendue sous plusieurs aspects : repérage et correction des anomalies, augmentation du volume et de l'exactitude des informations dans les notices, expertise accrue de producteurs plus nombreux et diversifiés.

3.1 L'expertise accrue des producteurs historiques

La valorisation de l'activité de catalogage des autorités est au cœur des ambitions d'IdRef. Au sein du réseau de catalogage partagé, chaque établissement a son expert mandaté : le "correspondant autorité". A l'aune des enjeux scientifiques et politiques liés à la visibilité de la recherche[4], les fonctions de ce professionnel, sensibilisé et compétent sur les questions d'identification des auteurs et d'attribution fidèle de leurs publications, tendent à se déplacer vers les services à la recherche et aux chercheurs : archives institutionnelles, pages de chercheurs, applications bibliométriques. IdRef est là pour accompagner ces évolutions.

3.2 L'extension du réseau des utilisateurs d'IdRef

Par ailleurs, le dynamisme des réseaux traditionnels de l'ABES (Sudoc, Thèses, Archives et manuscrits) attire vers IdRef de nouveaux partenaires aux profils hétérogènes et aux finalités variées. Ces nouveaux utilisateurs-producteurs viennent renforcer à leur tour la vitalité d'IdRef par leurs usages renouvelés, leurs exigences différentes mais néanmoins stimulantes et contribuent à accroître la qualité des données (détection des erreurs préexistantes, apport d'information complémentaire).

Parmi ces nouveaux acteurs, on trouve : Persée, portail d'accès libre et gratuit à des publications scientifiques numérisées ; ADUM, réseau de gestion des écoles doctorales ; les chercheurs du laboratoire de recherche historique Rhône-Alpes (LARHRA), spécialisé en histoire moderne et contemporaine et particulièrement investi dans le champ des Humanités Numériques ; le projet Prelib de recherche en littérature de langue bretonne de la Maison des Sciences de l'Homme en Bretagne (USR 3549).

Au-delà, IdRef entend devenir la principale source d'identification des chercheurs français. A ce titre, IdRef apporte une contribution tout à fait particulière au projet Conditor de base de signalement exhaustive de la production scientifique française.

4 IDREF AU SEIN DES INITIATIVES INTERNATIONALES

Au niveau national, la coopération se manifeste par la génération d'alignements et par l'existence d'un réseau actif de production en commun. Au niveau international, la coopération passe essentiellement par des alignements à grande échelle. IdRef est pleinement intégré à ces entreprises internationales.

Depuis 2012, sous l'étiquette historique mais caduque de « Sudoc », IdRef fournit à Viaf ses notices d'autorité et ses notices bibliographiques. Chaque mois, ce dump est récupéré et chargé à la fois dans une base relationnelle et dans une base RDF internes. L'ABES exploite ce réservoir de liens pour générer de nouveaux alignements et détecter des anomalies dans ses

[4] À ce sujet, voir par exemple (Pontille, et al., 2016)

propres bases. L'ABES est aussi membre d'ISNI, dont les identifiants sont intégrés autant que possible dans les notices IdRef. A ce jour, 1 900 000 autorités IdRef se sont vu assigner un ISNI. Entérinée en mai 2016 par la signature d'un *Memorandum of Understanding*[5], la coopération avec Orcid vise à plus d'interopérabilité entre les deux applications. Dans le même temps, des opérations d'alignement IdRef-Orcid ont été programmées pour l'identification des personnes présentes dans chacun des référentiels.

5 LES REALISATIONS EN MATIERES D'ALIGNEMENT A L'ABES

Disséminer des identifiants fiables et pérennes dans l'environnement ouvert du web, rendre interopérable les données : des missions importantes qui nous poussent, comme d'autres, à renforcer nos moyens techniques en matière d'alignement. En premier lieu, cela passe par des processus automatisés que nous voulons fiables et autonomes.

5.1 Dériver

Lorsque les ressources sont des ebooks, il est très probable que d'autres éditions de la même œuvre soient cataloguées dans le Sudoc ou dans Worldcat. Si les identifiants auteurs sont présents, on peut alors les dériver en profitant de la souplesse de RDF[6]. Pour trouver ces ressources dans le Sudoc, on s'appuie en priorité sur l'ISBN de la version papier ou à défaut sur le titre et le nom d'auteur. Ensuite, dans le cas où le document a plusieurs auteurs, il faut s'assurer que l'on attribue bien le bon identifiant au bon auteur. Pour cela, on utilise une fonction de comparaison des noms. Cette stratégie a néanmoins ses limites : elle ne s'applique pas aux chapitres et aux articles, qui ne sont pas catalogués dans le Sudoc ou Worldcat.

5.2 Clustériser

Les métadonnées provenant des éditeurs contiennent des informations qui dépasse le champ bibliographique traditionnel mais qui peuvent être fort utiles comme données de gestion : affiliations, biographies ou encore adresses électroniques. Ces dernières sont très précieuses pour regrouper des mentions d'auteurs dans des documents différents qui réfèrent à la même personne. En effet, puisqu'une adresse électronique ne change en général pas de propriétaire, si deux mentions d'auteur homonymes ont la même, il doit s'agir de la même personne. Grâce à cette information, nous rassemblons des mentions d'auteur autour d'un même identifiant provisoire. C'est ce dernier que l'on tentera ensuite d'aligner vers IdRef.

5.3 Calculer

Ce qui n'est pas dérivable peut être examiné par un algorithme dédié au calcul d'alignement vers les autorités IdRef, les identités ISNI ou encore ORCID. À partir du nom et du prénom d'une mention d'auteur à lier, l'algorithme procède dans un premier temps à une recherche floue dans le référentiel cible afin d'isoler un ensemble de candidats au liage pertinents. Ensuite, il considère toutes les informations associées à la mention d'auteur pour les comparer, *via* des critères précis, aux informations associées à chaque autorité candidate.

[5] https://orcid.org/blog/2016/07/11/abes-and-orcid-memorandum-understanding

[6] Depuis 2012 sous l'étiquette « hub de métadonnées », l'ABES a adopté la technologie RDF pour mettre au point des outils d'import et d'enrichissement des données fournies par les éditeurs.

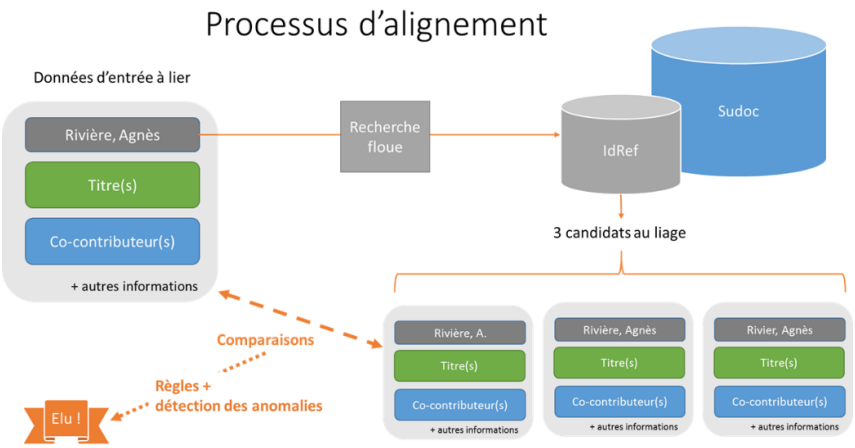

Figure 2 – Schéma du processus d'alignement par calculs

Si une mention d'auteur prise dans son contexte de publication est riche d'informations (nom, prénom, titre du document, date de publication, indexation matière, co-contributeurs, etc.), on ne sait bien souvent que peu de choses sur une personne décrite dans une notice d'autorité (nom, prénom, dates de vie). Pour ces dernières, des informations riches peuvent néanmoins être récupérées en suivant les liens d'autorité vers les notices bibliographiques : titres, dates, sujets, co-contributeurs, etc. viennent enrichir l'autorité pour constituer ce que nous appelons une « superautorité ». Ce principe ne s'applique pas seulement aux autorités *stricto sensu*, il est aussi valable pour les mentions d'auteurs clustérisées (voir supra).

Finalement, les résultats de comparaison sont analysés en fonction de règles simples d'alignement vers un candidat unique (voir figure 3). En cas de doute, l'algorithme s'abstient de lier et fournit parfois une alerte qui sera prise en compte par des experts, qui pourront procéder à des corrections dans la ou les bases de données concernées. Bien souvent les anomalies ainsi détectées correspondent à des doublons d'autorité ou à des erreurs de liage préexistantes.

3 règles simples d'alignement
Règle 1 : l'appellation n'est pas différente et les titres ou les co-contributeurs sont proches.
Règle 2 : l'appellation est un peu différente et les titres ou les co-contributeurs sont quasi identiques.
Règle 3 : il existe un seul candidat au liage (malgré une recherche élargie) et le nom est identique et aucun critère ne tend à l'éloignement.

Figure 3 – Des règles simples pour déduire des alignements fiables

6 DIAGNOSTIQUER, CONTROLER, ASSISTER

Le projet de recherche Qualinca (2012-2016) a permis de questionner en profondeur les problèmes de qualité du liage. Mesurant l'ampleur des erreurs et des lacunes dans notre propre catalogue, nous avons collaboré étroitement avec des chercheurs de l'équipe GraphIK[7] de Montpellier pour mettre au point SudoQual, un programme dédié au diagnostic qualité. Après avoir présenté son fonctionnement, nous montrerons comment un utilisateur humain peut interagir avec ce programme grâce à une interface web dédiée.

[7] GraphIK (Graphs for Inferences on Knowledge) est une équipe commune de l'Inria (Sophia Antipolis), de l'Université de Montpellier, du CNRS et de l'INRA https://team.inria.fr/graphik/. Travaille sur la représentation des connaissances et les moyens de raisonner à partir de ces dernières via des approches logiques.

Dans un projet précédent, (Chein, Leclère, & Nicolas, 2014) ont montré que la stratégie d'alignement *via* la génération de superautorités (voir supra) présente le risque de fédérer des informations issues de notices bibliographiques indûment liées à l'autorité, ce qui peut conduire le système d'alignement à se tromper à son tour. SudoQual gère ce risque grâce à un processus incrémental : tout nouvel enrichissement de la superautorité se fait après validation du lien d'autorité. A l'initialisation, c'est le lien dit « source »[8] qui est considéré comme valide. La plus-value de cette méthode : neutraliser les anomalies en amont pour ne pas les subir en aval.

6.1 Un modèle d'expertise de liage

Mentions d'auteur, identifiants d'autorité : des références aux entités du monde réel
Pour (Chein, Gutierrez, & leclère, 2015) l'identifiant est une référence unique à une entité du monde réel, ici des personnes. Au sein d'une notice bibliographique (e.g. mention d'auteur), c'est une référence contextuelle. Si l'entité fait l'objet d'une notice d'autorité, c'est une référence d'autorité. Le lien d'autorité est une relation *sameAs* établie entre une référence contextuelle et une référence d'autorité. Absent des catalogues de bibliothèque, le « non-lien d'autorité » est une relation *differentFrom*.

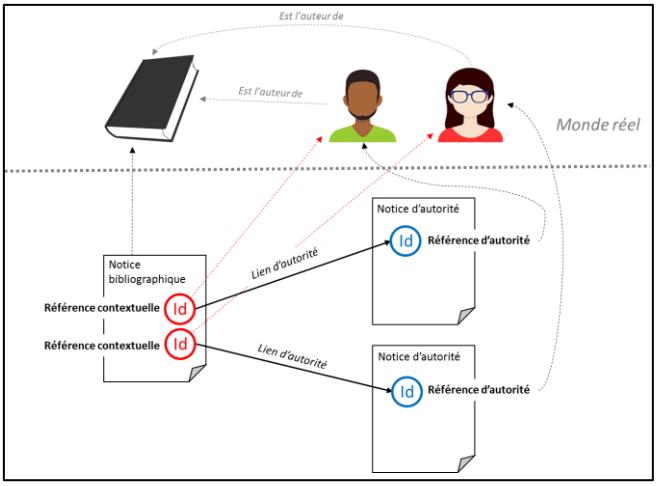

Figure 4 – Les liens dans un catalogue documentaire sous contrôle d'autorités

Des connaissances tacites aux connaissances explicites
Dans l'esprit d'un catalogueur confronté à un problème de liage s'opèrent des raisonnements successifs fondés sur des connaissances souvent tacites. La décision de liage, de non liage ou de correction de lien aboutit certes lorsqu'un faisceau de présomptions existe, mais les indices qui l'alimentent restent implicites. La première partie de notre travail a consisté à transformer ces connaissances tacites en connaissances explicites.

Des experts autorités de l'ABES ont été invités à formuler des affirmations au sujet du lien entre des publications et leurs contributeurs (voir figure 5).

[8] Dans un notices d'autorité, une zone (810 en UNIMARC) permet de mentionner la source d'information consultée. Lorsque cette source est un document décrit dans le catalogue, on déduit un lien source entre ce document et l'autorité en question.

« Une personne a tendance à publier plusieurs fois au sein d'une même organisation (e.g. laboratoire) »
« Les publications d'une personne portent souvent sur les mêmes thématiques »
« Une personne a tendance à republier chez le même éditeur »
« Deux manifestations qui ont le même titre ont en général les mêmes contributeurs »
« Une personne a tendance à publier plusieurs fois avec les mêmes co-contributeurs »
« Une personne ne peut pas publier alors qu'elle n'est pas encore née »

Figure 5 - Des affirmations sur le liage

Mais comment exprimer le degré de certitude de chaque affirmation ? Comment, de plus, expliciter les éléments de connaissance qui les justifient ? Nous avons élaboré un modèle d'expertise de liage qui permet la formalisation du savoir-faire des catalogueurs par des règles logiques. Ce modèle comporte plusieurs niveaux : le niveau attribut, le niveau critère de comparaison, le niveau règle de coréférence ou de différence.

L'attribut est un **élément de connaissance** associé à une référence. Un attribut est extrait de la notice bibliographique ou d'autorité de la référence à laquelle il est associé.

Forme du nom : « Paul Otlet »
Date de naissance : « 1868 »
Titre : « Traité de Documentation : le livre sur le livre »
Indice Dewey : « 020.9 »
Mot-clé : « Bibliothéconomie »
Collectivité : « Mundaneum (Mons, Belgique) »

Figure 6 – Exemple d'attributs d'une référence contextuelle

Le critère est une **fonction de comparaison** sur les valeurs de certains attributs permettant d'obtenir un indice de proximité ou d'éloignement de deux références.

Titre : *identique, très proche, assez proche*
Mot-clé : *très proche, assez proche*
Dewey : *très proches, assez proche, différent*
Forme du nom : *homonyme, proche, distant, différent*

Figure 7 – Des critères, avec leurs indices de proximité ou d'éloignement possible

La règle est une **affirmation sur la relation entre deux références** conclue à partir d'un ensemble de conditions incarnées par les critères. Une règle peut conclure au rapprochement (coréférence) ou à l'éloignement (différence), avec plus ou moins d'intensité. Pour cela, des valeurs qualitatives de confiance ont été définies et ordonnées (voir figure 8).

Règles de coréférence : 5 valeurs de confiance

→← Certain → →← Très probable → →← Probable → →← Plausible → →← Concevable

Exemple (X = référence contextuelle, Y = référence d'autorité)

Si la *forme du nom* de X est homonyme à celle de Y
et si la *collectivité* associée à X est identique à une de celles associées à Y
et si le document associé à X a des *mots-clés* très proches de ceux associés à Y,
ALORS il est **Très probable** que X soit la même personne que Y.

Règles de différence : 4 valeurs de confiance

←→ Certain → ←→ Très probable → ←→ Probable → ←→ Plausible

Exemple (X = référence contextuelle, Y = référence d'autorité)

Si la *forme du nom* de X est seulement proche (non homonyme) de celle de Y
Si le document associé à X a des indices Dewey différents de ceux associés à Y,
ALORS il est **Très probable** que X ne soit pas la même personne que Y.

Figure 8 – Les valeurs de confiance des règles de coréférence/différence

Partitionnement initial et calcul des relations

Nous avons conçu SudoQual, un programme qui s'appuie d'une part sur ce modèle d'expertise, d'autre part sur le processus incrémental, pour diagnostiquer les liens d'autorité.

A partir d'une appellation, i.e. d'un nom et d'un prénom, une recherche floue est effectuée dans le Sudoc et dans IdRef afin d'isoler un sous-ensemble composée de références d'autorité et de références contextuelles, ces dernières pouvant être pourvues ou dépourvues de lien d'autorité. Le but est de construire le sous-ensemble (« partition ») le plus pertinent possible par rapport aux paramètres d'entrée. Pour cette raison, les appellations des résultats obtenus peuvent être assez différents des paramètres d'entrée.

Pour cette partition, on va constituer une base de faits où à chaque référence sont associés des attributs, issus de l'extraction des données du catalogue et du fichier d'autorités. Le programme dispose alors des informations nécessaires pour comparer chaque référence contextuelle à chaque référence d'autorité, indépendamment du lien initial ; le processus de calcul itératif peut commencer. Dès qu'une règle logique s'applique entre deux références, une relation (positive ou négative) est immédiatement déduite. Une fois cette étape achevée, le programme a constitué son propre graphe de relations entre les références du sous-ensemble (voir figure 9).

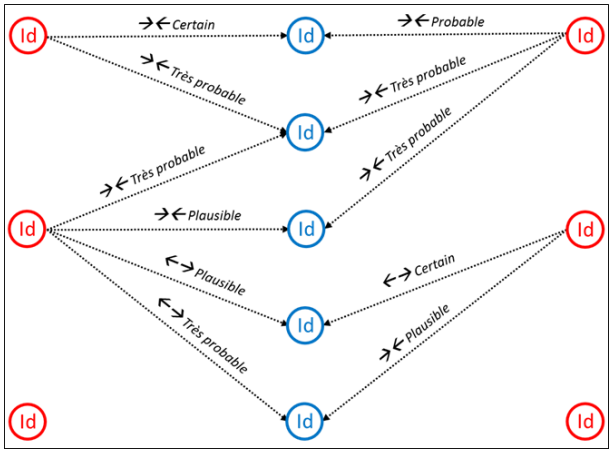

Figure 9 – Schéma des relations établies suite à l'application des règles logiques

Ensuite, s'appuyant sur la nature (→← coréférence ou ←→ différence) et le degré de confiance des relations, une heuristique établit différents types de liens : liens sûrs, liens suggérés, liens impossibles. (voir figure 10)

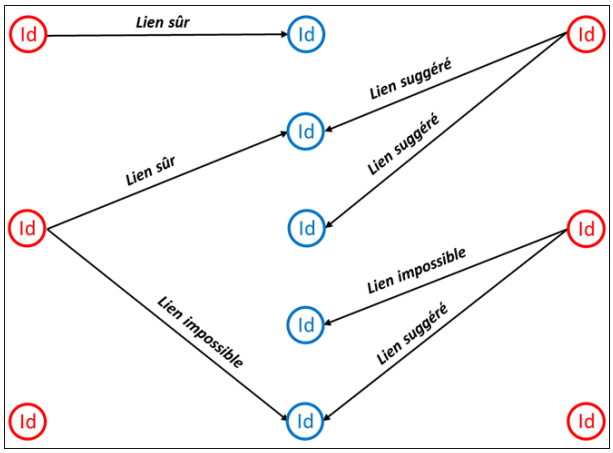

Figure 10 – Liens établis par l'heuristique de calcul des liens

Calcul itératif en n étapes
Dans la figure 10, on voit qu'au bout de cette première étape d'établissement des liens, certaines références contextuelles restent isolées, sans relation avec une référence d'autorité. Pour d'autres, la relation n'a pas permis de calculer un lien sûr. L'enrichissement des références d'autorité avec les attributs des références contextuelles dont le lien a été validé ouvre la possibilité de découvrir des relations. Le programme relance donc une nouvelle étape de calcul. Tant que de nouveaux liens sûrs seront calculés, SudoQual procède à une nouvelle itération. Comme nous l'expliquions au début de cette section, ce fonctionnement garantit la qualité de la superautorité de manière progressive.

Le diagnostic
Après la dernière itération, le programme confronte les liens dit calculés avec les liens initialement établis dans la base : il propose un diagnostic qualité des liens initiaux. Ce diagnostic prend la forme d'un statut (*status*) jugeant la qualité du lien que la référence contextuelle possède dans la base (voir figure 11).

validatedLink
Le lien initial est validé par SudoQual.
erroneousLink
Le lien initial est jugé invalide par SudoQual et des propositions de correction sont faites.
almostValidatedLink
Le lien initial est jugé fort probable par SudoQual sans pour autant qu'il soit complètement validé.
doubtfulLink
Le lien initial est jugé comme douteux par SudoQual et des propositions de correction peuvent être proposées.
missingLink
Aucun lien initial n'existe et SudoQual peut faire des propositions de correction.

Figure 11 - Les 5 statuts de diagnostic d'une référence contextuelle

6.2 Evaluation

Nous avons mené une première évaluation de SudoQual sur un échantillon de 57 notices d'autorité et 359 notices bibliographiques. Un expert de l'ABES a d'abord vérifié et corrigé si nécessaire tous les liens d'autorité à l'aide de l'outil de catalogage traditionnel. Ce diagnostic humain a ensuite été comparé au diagnostic de SudoQual. Dans plus de 70% des cas SudoQual a fait les mêmes choix que l'expert –dont plus de la moitié de choix certains (liens sûrs), alors qu'il ne s'est trompé que dans 0,3% des cas. Pour le reste des cas, SudoQual n'a pas été capable de faire un choix, mais il a pu faire des suggestions.

6.3 Entre automatisation et aide à la décision

Ces résultats intermédiaires montrent le potentiel de notre programme en matière de diagnostic automatique, mais aussi d'aide à la décision. En effet, pour tous les cas où SudoQual reste indécis, le catalogueur doit pouvoir prendre le relais facilement. D'autant plus qu'il peut bénéficier de suggestions qui lui faciliteront la tâche. Mieux, il doit pouvoir donner un coup de pouce à SudoQual pour que ce dernier relance de nouveaux calculs. Dans la section suivante, nous présentons cet outil d'aide à la décision.

6.4 Une interface pour faire collaborer l'humain et la machine

La figure 12 montre le principe général de l'outil : matérialiser le delta entre les liens initiaux (y compris les liens manquants, représentés en bleu) dans la base et les liens calculés par SudoQual. Par un système de *drag and drop*, l'utilisateur peut créer ou modifier des liens

d'autorité. En sélectionnant une référence contextuelle, il peut valider ou invalider son lien d'autorité.

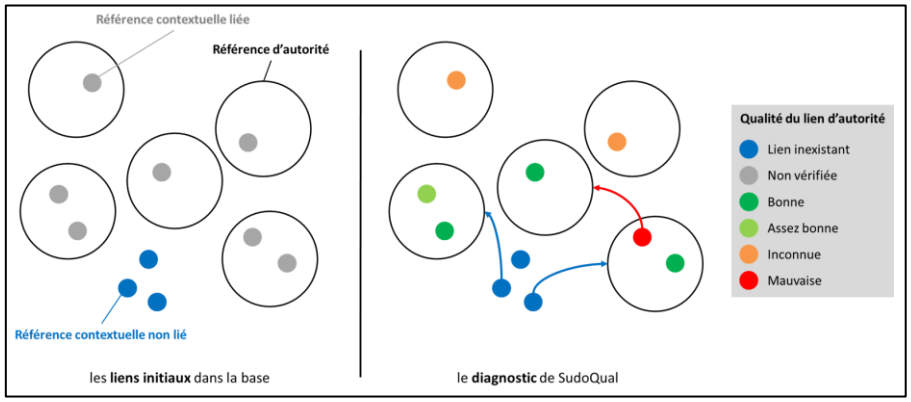

Figure 12 - Vue schématique de l'interface de visualisation

Se projetant sur le long terme, l'ABES a décidé de pousser la réflexion assez loin pour produire un outil innovant en matière de représentation des autorités enrichies de leurs liens bibliographiques. Cette interface est pour l'instant adossée au catalogue Sudoc et à IdRef. Elle offre à l'utilisateur des moyens de validation, de correction et de création de liens d'autorité. La représentation imagée des éléments (références contextuelles, références d'autorité et liens) facilite la compréhension rapide du problème et des actions possibles. Les icônes, les symboles, les interactions simples et fluides (*drag and drop*, clic, survol…) servent à rendre l'interface moins rébarbative, voire ludique. De manière complémentaire, des messages clairs permettent à l'utilisateur de poursuivre son travail étape par étape, de maîtriser la série d'actions qui s'achèvera sur une modification réelle dans le catalogue Sudoc.

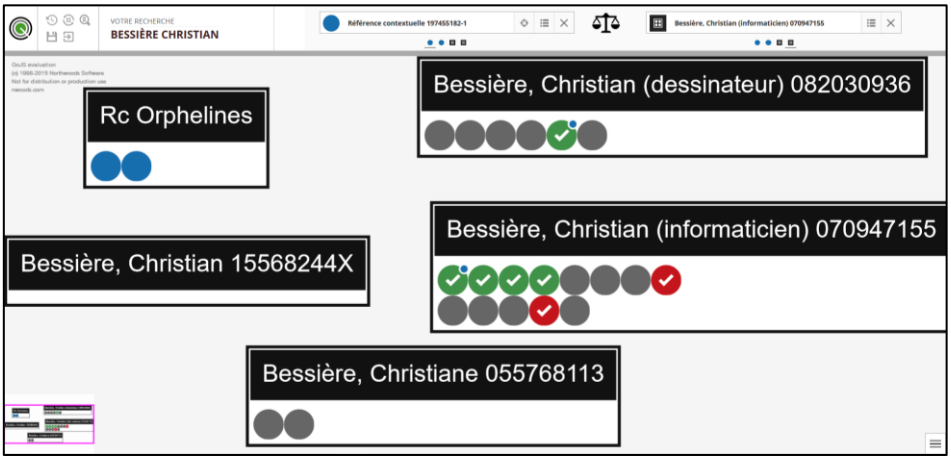

Figure 13 – Vue de la partition « Bessière, Christian » : Les couleurs et icônes indiquent les liens validés, invalidés ou restant à diagnostiquer

L'utilisateur reste maître de la situation. Selon ses besoins, il peut agir en autonomie ou bien s'appuyer sur les diagnostics de SudoQual. Le cas échéant, il pourra valider les choix de ce dernier et profiter de ses « conseils » (liens suggérés et liens interdits). Il n'y est pas pour autant soumis : en cas de nécessité, il dispose d'un droit de contestation. En définitive, ce seront toujours ses décisions qui primeront sur celles du programme informatique.

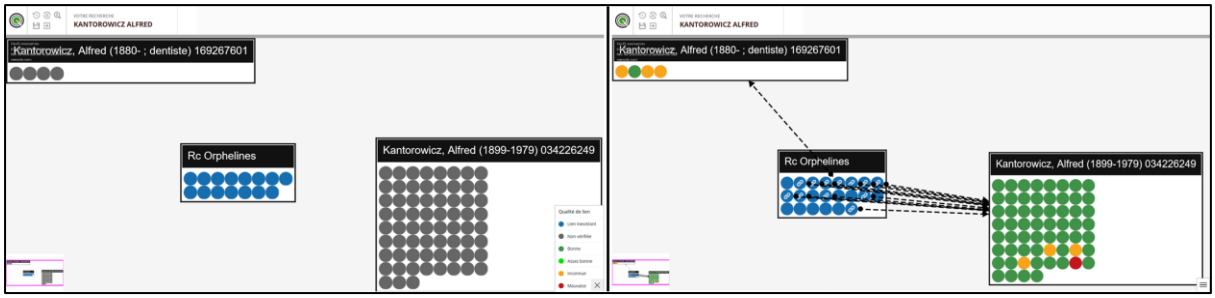

Figure 14 – à gauche : vue de la partition initiale ; à droite : vue de la partition diagnostiquée par SudoQual

Le contrôle qualité des liens bibliographiques requiert l'accès à des informations dont les niveaux de détail et de complexité sont plus ou moins élevés. C'est pourquoi l'utilisateur doit pouvoir accéder à des vues synthétiques (zoom out) comme à des vue analytiques (zoom in), tout en bénéficiant d'une navigation fluide entre les différents modes d'affichage.

Figure 15 – Visualisation des attributs d'une référence contextuelle (à gauche) et d'une référence d'autorité (à droite)

Enfin, l'interface donne également à l'utilisateur la possibilité de donner un coup de pouce au programme. Dans l'exemple ci-dessous, on voit que la simple validation manuelle d'un lien permet à SudoQual de lancer une nouvelle étape de calcul aboutissant à la validation d'un grand nombre supplémentaire de liens, ainsi qu'à la création de nouveaux liens. Ces allers-retours peuvent connaître plusieurs itérations.

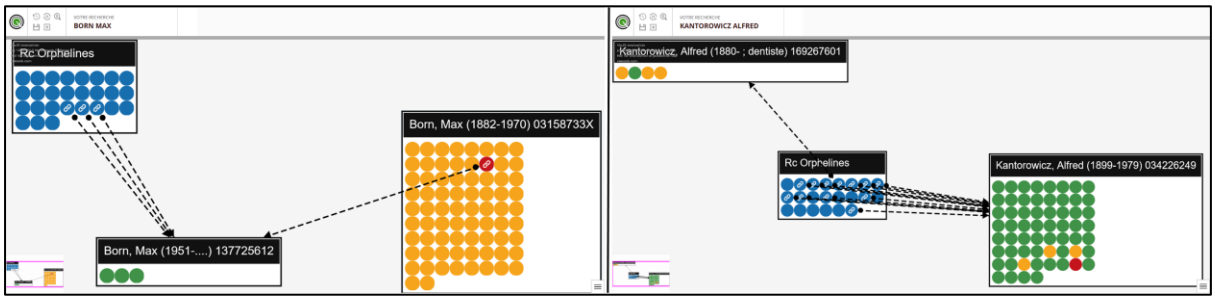

Figure 16 – à gauche : vue de la partition diagnostiquée une fois par SudoQual ; à droite : vue de la partition diagnostiquée par SudoQual après validation manuelle d'un lien

6.5 Perspectives

SudoQual comme l'application de visualisation sont encore des prototypes issus d'un travail de recherche. Pour envisager le traitement de gros volumes de données ou pour fournir des outils de travail en temps réel aux membres de nos réseaux, un important travail d'optimisation des performances est nécessaire. Une étude est en cours afin de définir et d'évaluer les coûts techniques et humains qu'il nous faudra investir pour répondre à la question cruciale de l'amélioration des temps de calcul.

Nous avons également du chemin à faire pour maîtriser parfaitement le paramétrage de SudoQual afin, d'une part d'interpréter plus facilement ses conclusions, d'autre part de les optimiser. Interpréter signifie être capable de remonter facilement à la source : telle(s) règle(s), tel(s) critère(s), tel(s) attributs(s) et de mesurer les impacts des modifications –même infimes du paramétrage. Optimiser signifie régler au plus près des besoins : cela concerne tant le mode de constitution des sous-ensembles de travail, que la pertinence d'une règle logique appliquée à un corpus donné.

Si nous disposons aujourd'hui d'une solide preuve de concept, nous ne pouvons pas nous passer d'une évaluation à grande échelle. La nature des données traitées, leur richesse, mais aussi leur niveau de qualité seront autant de critères à envisager pour évaluer en profondeur et peut-être amender notre modèle d'expertise.

7 CONTROLE QUALITE ET CATALOGAGE QUOTIDIEN

Nous venons de présenter en détail un outil dédié au contrôle de la qualité des liens entre notices d'autorité et notices bibliographiques. Un tel outil ne doit pas être pensé en vase clos, indépendamment du fonctionnement global et quotidien du catalogage coopératif. Les opérations de contrôle qualité sont coûteuses. Elles appellent donc des incitations en amont et un impact durable et visible en aval.

Côté amont, les opérations de contrôle qualité ne peuvent reposer sur le seul bon vouloir individuel du catalogueur. L'ABES a déjà mis en place différents dispositifs qui incitent les bibliothèques du réseau Sudoc à initier ou participer à des chantiers de correction, dont les objectifs et le périmètre sont fixés en commun. Ainsi, l'investissement dans la qualité, par certains, au service de tous, est reconnu et valorisé. Et cette valorisation nourrit l'esprit de réseau.

Mais, cet investissement n'est possible que s'il promet un retour sur investissement durable. Cela passe notamment par la capitalisation de ce travail : les liens validés par les catalogueurs, dans l'environnement Qualinca, ne peuvent être considérés comme des liens ordinaires, que le quotidien du catalogage fait et défait. Ces liens ont une valeur et une validité supérieures. Ils doivent être qualifiés comme tels dans la base de données du Sudoc. Cette qualification servira à la fois aux humains et aux machines. Elle rendra leurs futures conclusions plus sûres et plus rapides, en évitant de mener à nouveau certaines analyses ou certains calculs.

Mais, à l'opposé du contrôle qualité, le catalogage quotidien est dicté par la file d'attente des documents à traiter. Dans ce contexte, la qualité systématique du catalogue, et au-delà, du web de données, n'est qu'un horizon ou une valeur, pas un objectif. Que peut promettre Qualinca au catalogueur soumis à la pression des livres en attente ? Il s'agit à la fois de lui faciliter la vie et de valoriser ses interventions.

Lui faciliter la vie consiste à le dispenser de tâches inutiles et de rendre plus aisées les tâches irréductibles. Nous avons vu que l'opération de liage n'est qu'un cas particulier des opérations de diagnostic et réparations des liens : le diagnostic établit l'absence d'un lien et la réparation propose le lien à créer. Dans un certain nombre de cas, les propositions de lien générées par machine sont suffisamment fiables pour être suivies à la lettre et enregistrées dans le catalogue. Autant de liens que le catalogueur n'aura pas à établir. Dans les autres cas, la machine fournit non pas la solution toute crue, mais une aide à la décision. Cette assistance doit être intégrée au sein même de l'environnement de catalogage, quel qu'il soit (Sudoc ou autre) : le catalogueur ne doit pas être obligé de basculer dans une interface de contrôle qualité dont les fonctions sont bien plus riches que le besoin immédiat à satisfaire. D'un point de vue ergonomique, cette intégration doit être assez intuitive pour faciliter une décision rapide et assez informative pour permettre une *bonne* décision ! Parfois, la bonne décision consiste à ne pas lier : la machine doit non seulement suggérer au catalogueur de ne pas lier, mais également pré-fabriquer la nouvelle notice d'autorité à créer, ou bien aller la chercher dans un référentiel d'appui, ou encore préparer l'opération de fusion entre deux autorités identifiées comme doublons.

Si le travail du catalogueur est facilité par la machine, celui-ci peut se concentrer sur les tâches à plus grande valeur ajoutée : faire ce que la machine ne sait pas faire, ou pas assez bien ; faire ce qui permettra à la machine de faire davantage et mieux. Certains liages sont trop difficiles pour être confiés à la machine. Si elle est bien configurée, elle déclarera forfait d'elle-même, passant le relais à l'humain. Pour établir le lien à la main, ce dernier aura besoin de prendre un peu de temps pour analyser les autorités en présence. Qu'il en profite pour faire du vrai contrôle qualité, systématique, aux effets durables. Cela implique d'interrompre le flux du catalogage pour *investir* quelques instants au contrôle qualité. Le catalogueur peut ainsi décider de vérifier systématiquement les liens relatifs à cette appellation, avec l'outil décrit ci-dessus, mais sa contribution à la qualité globale et durable peut être plus modeste, quoique précieuse : en vérifiant ou complétant les liens sources, par exemple, le catalogueur effectue une opération simple et à fort rendement, qui fournit à la machine un socle solide pour de futurs calculs itératifs.

8 CONCLUSION

Le point de mire du projet IdRef est un objectif largement partagé : doter la communauté de l'IST de référentiels solides pour la diffusion et la valorisation la plus ouverte possible des connaissances scientifiques. Les bibliothèques, expertes historiques en la matière, sont un chaînon important ; les liens étroits qu'elles entretiennent avec nombre d'acteurs (éditeurs, universités, institutions culturelles) les placent en mesure de fédérer les initiatives éparses.

Nous avons montré comment, par la mise à disposition de moyens techniques, par l'écoute des besoins et des attentes exprimées par les membres d'un réseau devenu multiculturel, une agence bibliographique d'envergure nationale peut piloter la décentralisation des efforts de production et de contrôle des autorités, devenues jeux de données référentielles ouvertes et liées.

Dans l'océan mondial de la connaissance, la fiabilité de ces données est primordiale. Des systèmes fédérateurs tels VIAF ou ISNI, s'abreuvant à de multiples sources, construisent peu à peu des canaux d'irrigation indispensables. Mais pour que l'océan atteigne le plein et qu'il ne soit pas pollué à cause de canaux défectueux, les sources de données doivent être entretenues

dès l'amont. C'est ce à quoi appelaient déjà (Angjeli, Mac Ewan, & Boulet, 2014) et c'est ce que nous nous efforçons de faire… contre vents et marées.

Acknowledgments
Les auteurs remercient chaleureusement Michel Chein, Alain Gutierrez, Michel Leclère, Clément Sipieter et tous les membres de l'équipe GraphIK, ainsi que Brett Choquet pour la collaboration fructueuse et durable qu'ils ont entretenue ces dernières années. Ils remercient également Cornelia Stratulat et Christophe Bonnefond, leurs collègues informaticiens pour leur ténacité et leur implication dans ce projet de longue haleine.

References

Angjeli, A., Mac Ewan, A., & Boulet, V. (2014). ISNI and VIAF – Transforming ways of trustfully consolidating identities. *IFLA WLIC 2014 - Lyon - Libraries, Citizens, Societies: Confluence for Knowledge.*

Chein, M., Gutierrez, A., & leclère, M. (2015). *Un problème d'identification d'entités nommées dans des bases de donnés documentaires.* Rapport de recherche, LIRMM.

Chein, M., Leclère, M., & Nicolas, Y. (2014). Sudocad: a knowledge-based system for the author linkage problem. *Knowledge and Systems Engineering*, pp. 65-83.

Pontille, D., Séné, A., Prêtre, V., Pothier, V., Deniau, N., Durand-Barthez, M., & Girard, F. (2016). Éloge de la complexité: la signature des chercheurs et le millefeuille de l'affiliation institutionnelle dans les processus d'évaluation de la recherche. *Ethics, Medicine and Public Health, 2*(3), pp. 456-465.