Depuis la création du Sudoc, l'Abes et ses réseaux ont compris, comme d'autres, les bienfaits des liens bibliographiques. Un aperçu de la démarche Qualinca pour résoudre les problèmes inhérents au liage dans les grandes bases de données documentaires : une véritable complémentarité entre l'homme et la machine.

« Faire le lien », *un besoin vital*

Traditionnellement, les liens permettent le contrôle d'autorité et la navigation dynamique de notices en notices. Avec le modèle FRBR et l'ouverture des catalogues sur le web de données, leur importance ne cesse de croître. Pour la FRBRisation, l'exploitation de certains liens facilite la mise au jour des entités œuvre et expression[1]. En matière d'ouverture, IdRef avait déjà émancipé les ex-« autorités Sudoc » en leur permettant d'être utilisées et alimentées par d'autres applications. Quant aux alignements à grande échelle, ils se nourrissent des liens bibliographiques sur un mode ricochet : entre deux entités sur le web de données, plusieurs entités peuvent se donner la main. La qualité des liens internes au catalogue est primordiale car elle se répercute bien au-delà d'une seule base de données et d'un seul usage.

DÉFINIR ET MESURER LA QUALITÉ DU LIAGE

Appuyés par des théories et méthodes en intelligence artificielle, les chercheurs du projet Qualinca, parmi lesquels des membres de l'équipe GraphIK (*voir encadré*), s'aventurent sur le terrain des grandes bases de données documentaires comme le Sudoc pour résoudre les problèmes de liage. La réconciliation d'entités dans les bases de données a fait couler beaucoup d'encre depuis la fin des années 1950[2]. Il s'agit de confier à une machine le soin de dire si des descriptions d'entités (une personne, une collectivité, une œuvre, etc.) décrivent ou non la même entité « réelle ». Il existe plusieurs approches, dont celle à base de connaissances dans laquelle s'intègre le projet. L'objectif est de disposer de connaissances formalisées sur lesquelles appliquer des opérations logiques censées simuler les opérations humaines. *In fine*, on aimerait pouvoir repérer et caractériser les problèmes de qualité des liens et disposer de schémas de réparation.

DES ERREURS DE LIAGE, MAIS POURQUOI ?

Les raisons pour lesquelles les catalogueurs peuvent faire des erreurs de liage sont multiples : doublons d'autorité, ambiguïté des autorités (informations associés insuffisantes), erreurs de liage préexistantes, absence d'autorité. D'autres erreurs ne sont pas humaines, car pendant plusieurs années un algorithme de liage a fonctionné sur la base Sudoc. Cet outil comparait la chaîne de caractères d'une appellation (nom, prénom) dans une notice bibliographique avec les formes retenues des appellations dans les notices d'autorité. Problème : la machine n'intégrait pas qu'une forme retenue pouvait être constituée d'une appellation accompagnée de qualificatifs (dates de vie, fonction). Fatalement, dans les cas d'homonymie, les autorités dont la forme retenue n'avait pas de qualificatif ont été les seules à être liées par cet algorithme. Autre inconvénient : si cet outil ne faisait pas de lien lorsqu'il était confronté à plusieurs possibilités de liage (ce qui est bien), il ne capitalisait pas cette hésitation (ce qui est dommage). Avec un système à base de connaissances et de règles, la collaboration humain/machine, alors absente, sera fortement mise en valeur.

NOURRIR LA MACHINE : TRANSFORMER UN CATALOGUE EN BASE DE FAITS

Le choix des données est le premier pas vers la construction du système. Le travail de l'analyste consiste à identifier dans les notices ce qui est utile à un catalogueur dans une décision de liage. L'environnement normatif (ISBD, Unimarc, etc.) et

LE PROJET GRAPHIK

Créé en 2010, GraphIK (*Graphs for Inferences on Knowledge*) est une équipe commune de l'Inria (Sophia Antipolis), de l'université de Montpellier 2, du CNRS et de l'Inra. Leurs travaux reposent sur la représentation des connaissances et les moyens de raisonner à partir de ces dernières *via* des approches logiques.
Cette équipe de recherche a développé une bonne connaissance des données et des problématiques de l'Abes, avec qui elle a déjà travaillé (projet SudocAD, 2011).
https://team.inria.fr/graphik

[1] Dont on peut voir les résultats dans des applications comme data.bnf.fr.

[2] Depuis l'article fondateur : H. Newcombe, et al, « Automatic Linkage of Vital Records », *Science*, 1959, vol. 130, p. 954-959.

l'historique de construction et d'alimentation du Sudoc imposent une connaissance experte du catalogue, afin de ne pas surinterpréter ou mal interpréter le sens des données, c'est-à-dire l'information qu'elles véhiculent. En résumé, qu'une valeur existe dans le catalogue ne signifie pas qu'elle soit exploitable pour n'importe quel usage.

Il faut ensuite se départir du format source (Unimarc) pour transformer le catalogue en base de connaissances. Pour cela, on fait appel à une ontologie, FRBRoo[3] pour ce qui concerne le Sudoc[4]. Là encore, c'est l'humain qui, interprétant la grammaire de l'ontologie, modélise les données sélectionnées. L'attribution d'un identifiant à chaque entité nommée mentionnée dans une notice bibliographique achève ce travail de formalisation.

FABRIQUER LE SYSTÈME D'INTERPRÉTATION

À ce stade, la machine peut, en effectuant des requêtes, isoler des « grains de connaissance » et les attribuer à des références[5] *(voir graphique ci-contre)*, entre lesquelles des liens potentiels existent. Ces attributs constituent le matériau du raisonnement humain : le catalogueur les combine, les questionne, les compare et, en associant cela à ses propres connaissances, décide de lier, de corriger un lien ou de s'abstenir de lier. Décortiquer ces raisonnements pour les reproduire en machine est un défi majeur du projet. Distinguer des attributs permet d'alimenter la grille d'interprétation du système. Celle-ci comprend des critères, qui comparent des attributs ou combinaisons d'attributs, et des règles, qui combinent les résultats des critères en un résultat d'identification (rapprochement de références) ou de différenciation (éloignement de références). Ce modèle permet de disposer d'un système capable de prendre des décisions de liage/non-liage justifiées et contextualisées. Par ailleurs, ce système n'est pas figé : il bénéficie d'améliorations itératives *via* des ajustements de paramétrage.

De l'intuition à la règle

- *Intuition métier*

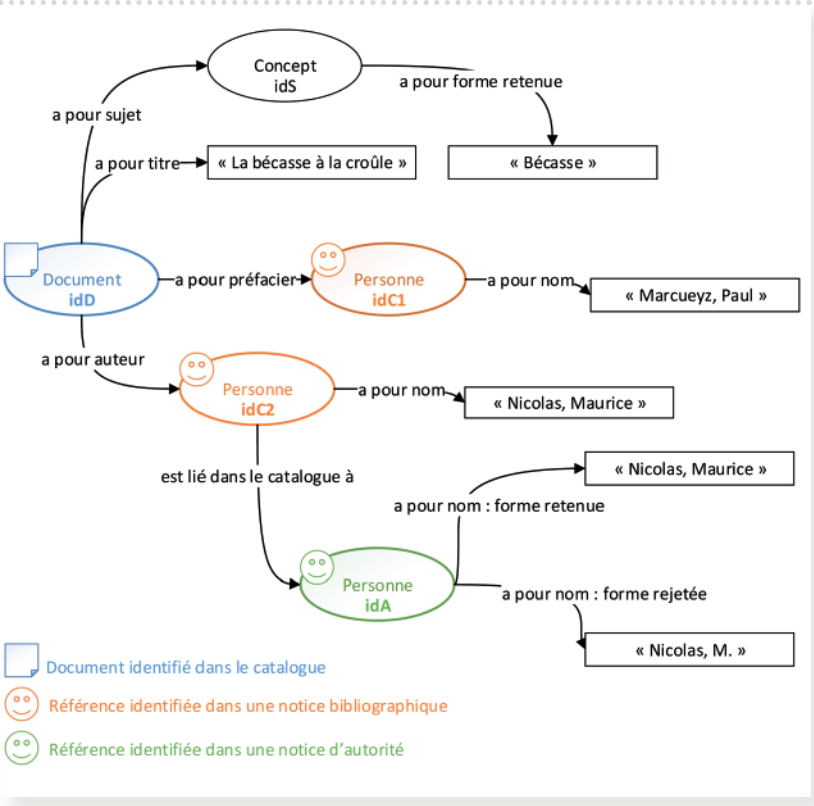
> « Un contributeur a tendance à republier avec les mêmes co-contributeurs ».

- *Traduction règle formelle*

> Si X a publié avec Y, et si X a publié un autre livre avec Z qui a le même nom que Y, alors il est quasiment certain que Z = Y.

- *Explication*

On compare la référence X avec la référence Y. Les attributs impliqués sont l'appellation et les co-contributeurs. Deux résultats de critères sont combinés dans la règle de liage : « identique » pour l'appellation et « identique » pour co-contributeur. Le système décide, selon cette règle, que le liage entre X et Y est quasiment certain.



Du catalogue au graphe : identifier les références et formaliser les relations qu'elles entretiennent avec les entités (documents, personnes, concepts…) et les valeurs (noms, dates, titres…) du catalogue.

FAIRE CONFIANCE À LA MACHINE : OUI, MAIS PAS AVEUGLÉMENT !

Dans la démarche Qualinca, le diagnostic qualité sollicite fortement la collaboration de l'humain et du système automatique. Ce dernier sera capable de proposer des rapprochements ou des éloignements de référence plus ou moins forts. L'établissement de seuils de confiance est la clé de la réussite. L'analyse humaine intervient encore une fois en ajoutant une heuristique, c'est-à-dire un système de règles contrôlées qui valide ou non la décision brute. Réparer un catalogue de l'intérieur serait ensuite possible, en se fondant sur des combinaisons de décisions fiables de liage/non-liage. Ces scénarios modéliseront des types de problèmes détectables par le système et dont les modalités de corrections auront été décrites. Plus proche du travail quotidien, l'exploitation de décisions de liage pour l'aide à la décision dans une application comme IdRef est également envisagée. À n'en pas douter, la philosophie de la zone 309[6], qui permet aux mains expertes du réseau de prendre le relais lorsque les traitements de masse se sont épuisés, a de beaux jours devant elle.

Aline Le Provost
Abes
le-provost@abes.fr

[3] FRBRoo est une ontologie formelle destinée à capturer et représenter la sémantique sous-jacente de l'information bibliographique et de faciliter l'intégration, la médiation et l'échange de l'information bibliographique et muséographique : **www.cidoc-crm.org/frbr_inro.html**

[4] L'Institut national de l'audiovisuel, second partenaire du projet en gestion documentaire, a choisi de développer une ontologie maison pour ses propres données.

[5] Une référence est une entité nommée dans une notice (bibliographique ou d'autorité).

[6] À ce sujet, lire : **http://punktokomo.abes.fr/2014/05/02/une-zone-309-pour-coordonner-le-travail-collectif-sur-la-qualite-des-donnees-sudoc**