Date : 13/07/2006

VIAF (Virtual International Authority File) : un pont entre les fichiers d'autorité noms de personnes de Die Deutsche Bibliothek et de la Bibliothèque du Congrès

Rick Bennett
OCLC Online Computer Library Center
Dublin, Ohio, USA

Christina Hengel-Dittrich
Die Deutsche Bibliothek
Frankfurt am Main, Germany

Edward T. O'Neill
OCLC Online Computer Library Center
Dublin, Ohio, USA

Barbara B. Tillett
Library of Congress
Washington, D.C. USA

Meeting:	**123 Cataloguing**
Simultaneous Interpretation:	**Yes**

WORLD LIBRARY AND INFORMATION CONGRESS: 72ND IFLA GENERAL CONFERENCE AND COUNCIL
20-24 August 2006, Seoul, Korea
http://www.ifla.org/IV/ifla72/index.htm

Abstract

Die Deutsche Bibliothek, la Bibliothèque du Congrès et l'OCLC (Online Computer Library Center) sont en train de développer conjointement un fichier d'autorité international virtuel (VIAF, virtual international authority file) pour les noms de personnes. Ce fichier établira des liens entre les notices d'autorité produites par les agences bibliographiques nationales du monde entier, et sera disponible gratuitement sur le Web. Ce projet vise à démontrer la faisabilité de l'établissement automatique de liens entre notices d'autorité issues de différents fichiers d'autorité et à montrer les avantages que l'on en attend. Les fichiers bibliographiques et d'autorité de la Bibliothèque du Congrès et de Die Deutsche Bibliothek ont été utilisés pour créer le noyau initial du VIAF, qui contient plus de six millions de noms et plus d'un demi-million de liens. Un aspect primordial du projet a résidé dans le développement d'algorithmes d'appariement automatisé des noms, en utilisant à la fois des informations en provenance des notices d'autorité et des informations puisées dans les notices bibliographiques. Nous avons pu démontrer qu'il était réalisable de lier grâce à des algorithmes des noms de personnes

entre fichiers d'autorité nationaux ; soixante-dix pour cent des notices d'autorité noms de personnes communes aux deux fichiers ont été reliés automatiquement, avec un taux d'erreur inférieur à un pour cent. L'objectif à long terme du projet VIAF est de combiner les formes autorisées de noms en provenance d'un grand nombre de bibliothèques nationales ainsi que d'autres sources majeures d'information, pour les fondre en un service partagé d'autorités au niveau mondial.

Introduction

Plusieurs groupes de travail au sein de la Section de catalogage de la Fédération internationale des associations de bibliothécaires et de bibliothèques (FIAB/IFLA) ont reconnu le potentiel que présenterait un fichier d'autorité international virtuel (*virtual international authority file*, VIAF) [1], dans lequel les notices d'autorité produites par des agences bibliographiques nationales du monde entier et représentant la même entité seraient reliées entre elles et mises à la disposition de tous sur Internet. Ce VIAF constituerait une extension pratique du concept de contrôle bibliographique universel et s'appuierait sur le travail réalisé par chacune des agences bibliographiques nationales. Il permettrait la coexistence de variantes nationales ou régionales dans les formes autorisées, ce qui aiderait à répondre au besoin d'exprimer ces variantes dans la langue, l'écriture et la graphie que souhaite l'utilisateur.

Des scénarios actuels concernant l'avenir du Web décrivent le recours à des ontologies visant à rendre le Web plus intelligent dans un contexte de traitements automatisés. Le VIAF pourrait constituer l'une des briques fondamentales dans la construction d'un « Web sémantique » [2], une fois combiné à d'autres vocabulaires contrôlés et fichiers d'autorité provenant de sources telles que : services d'indexation et de production d'abstracts, archives, musées, éditeurs commerciaux, etc. Les bibliothèques tiennent là une occasion d'apporter leur pierre à cet édifice et devraient aider à concrétiser cette vision. Il est important pour la réalisation de cette vision commune que le VIAF soit rendu gratuitement accessible aux utilisateurs du monde entier.

D'autres projets se sont penchés sur les possibilités de réunir des noms de personnes dans des fichiers d'autorité. Le Projet LEAF [3] (*Linking and Exploring Authority Files)* se proposait de lier des notices d'autorités de sources nombreuses et variées (bibliothèques, archives, centres de documentation et de recherche). Ces notices sont dans des formats divers, et leur contenu varie énormément dans sa nature et son importance quantitative. Le projet LEAF proposait de lier automatiquement les notices au moment de les charger dans le système. En raison de la diversité des sources des notices d'autorité nom de personne, l'équipe du projet s'est aperçue que la seule information commune sur la base de laquelle il était possible d'établir les liens était le nom lui-même, en y incluant les formes de renvoi, et les dates associées. Comme il arrive fréquemment que les notices d'autorité nom de personne produites par les participants actuels au projet ne contiennent pas de dates, on s'attend à ce que le taux d'erreur atteigne des proportions inacceptables.

Le Projet InterParty [4] est un projet de démonstration subventionné par l'Union européenne visant à créer des fichiers d'autorité établissant des liens entre diverses

institutions, principalement dans le dessein d'aider à la gestion des droits en environnement numérique. Le système proposé par InterParty fournirait un seul point d'accès aux multiples bases de données impliquées dans le système, afin de fournir en premier lieu un service centralisé de recherche. Comme les liens sont identifiés manuellement dans toutes les bases de données, la personne qui établit une association entre deux noms peut saisir le lien correspondant. Ces liens peuvent ensuite être utilisés automatiquement. En fonction des institutions qui établissent ces liens, ces derniers peuvent être considérés comme plus ou moins fiables. La déclaration d'un lien par l'une des parties n'est pas soumise à accord des autres parties impliquées dans le système. Le projet est ouvert à l'appariement par algorithme, mais ne précise pas quelles sont les techniques ou les particularités des données qui sont requises pour permettre l'établissement automatisé de liens.

Le Projet VIAF

Au cours de la Conférence IFLA 2003 à Berlin, Die Deutsche Bibliothek (DDB), la Bibliothèque du Congrès (LC) et OCLC Online Computer Library Center (OCLC) se sont mis d'accord pour développer un Fichier d'Autorité International Virtuel (*Virtual International Authority File,* VIAF) pour les noms de personnes [5]. Les buts que poursuit le projet VIAF sont de démontrer qu'il est possible de lier automatiquement des notices d'autorité provenant de fichiers d'autorité nationaux différents, et de montrer l'intérêt qu'il y aurait à disposer d'un VIAF. Le projet VIAF va lier les fichiers d'autorité nom de personne de la Bibliothèque du Congrès et de Die Deutsche Bibliothek en un seul fichier d'autorité nom de personne virtuel. OCLC est en train de développer le logiciel d'appariement des notices d'autorité nom de personne entre les deux fichiers d'autorité. L'objectif à long terme du projet VIAF consiste à lier les formes retenues établies par de nombreuses bibliothèques nationales et autres sources faisant autorité, pour en faire un service d'autorité partagé au niveau mondial pour les noms de personnes, de collectivités, de manifestations temporaires, de lieux, etc.

Le projet VIAF se décompose en cinq phases :

1. Établir des notices d'autorité « élargies » à partir du fichier Personennormdatei (PND) et des notices d'autorité de la Bibliothèque du Congrès. Cet enrichissement consistera notamment à identifier les notices d'autorité qu'il convient d'intégrer et à déterminer tout besoin de traitement particulier concernant les fichiers sources.

2. Élaborer des algorithmes d'appariement, et apparier les notices d'autorité élargies afin de créer la version initiale du VIAF. Cela a constitué un processus itératif concomitant avec la Phase 1, dans la mesure où les résultats intermédiaires des appariements ont mis en lumière des informations supplémentaires qu'il était possible d'extraire et d'inclure dans les notices d'autorité élargies afin d'améliorer l'appariement.

3. Construire un serveur OAI (*Open Archive Initiative)* [6] pour donner accès au VIAF.

4. Pour assurer la maintenance de la base de données VIAF, il est nécessaire de disposer des ajouts et des modifications pratiqués par toutes les agences participantes tant dans leurs notices d'autorité que dans leurs notices bibliographiques. Ce système de mise à jour et de maintenance sera défini selon les protocoles établis par l'OAI pour rechercher les informations de mise à jour.

5. Afin d'accéder aux notices VIAF, une interface utilisateur sera rendue disponible sur le Web ouvert. Pour finir, la base de données et l'interface favoriseront Unicode, le multilinguisme et la multiplicité des écritures. Les requêtes directes sur la base de données (consistant par exemple à fournir un nom en usage à la Bibliothèque du Congrès et à demander le nom tiré du Personennormdatei qui lui a été apparié, sous la forme d'un simple lien HTML), peuvent servir dans le cadre des techniques du Web sémantique.

Il s'agit en priorité de démontrer la faisabilité du VIAF en liant les notices d'autorité nom de personne entre le Personennormdatei (PND) et le Fichier d'autorité Nom de la Bibliothèque du Congrès (*Library of Congress Name Authority File*, LCNAF). À la date du 31 décembre 2005, le fichier LCNAF contenait 4,2 millions de notices d'autorité nom de personne. À la même date, la Bibliothèque du Congrès avait créé et diffusé un total de 9,3 millions de notices bibliographiques.

À l'automne 2005, le fichier PND contenait 2,6 millions de notices d'autorité nom de personne. Le fichier PND est utilisé dans les notices bibliographiques de la DDB ainsi que dans les notices bibliographiques du Bibliotheksverbund Bayern (BVB). Ces deux fichiers totalisent 15 millions de notices bibliographiques associées aux notices d'autorité du PND.

Le problème de l'appariement des noms

Au départ, le VIAF fonctionnera comme un dictionnaire allemand-anglais et anglais-allemand des noms de personnes. Par exemple, pour un utilisateur américain qui chercherait **J. P. De Valk** (la forme du nom établie par la Bibliothèque du Congrès), ce nom pourrait être automatiquement « traduit » en **Johannes P. De Valk** (la forme établie par la DDB). Il est en effet courant que des agences internationales distinctes établissent des formes de noms différentes, ou bien, à l'inverse, qu'elles utilisent deux formes identiques pour représenter en fait deux auteurs distincts. (La forme **J. P. De Valk** aurait ainsi pu être établie par la DDB pour un auteur totalement différent.)

Les noms de personnes peuvent prendre des formes différentes pour la même personne, ou une même forme pour des personnes différentes, ce qui rend aléatoire l'appariement des noms en provenance de fichiers différents. Le domaine couvert par chacun des deux fichiers n'est pas tout à fait le même ; seule une petite fraction de noms de personnes figurent dans les deux fichiers. C'est pourquoi il faut faire appel à d'autres informations que le seul nom lui-même afin de permettre un appariement fiable. Dans les notices d'autorité nom de personne, les dates de naissance et/ou de mort d'une personne sont

souvent présentes. La combinaison des dates de naissance et de mort est souvent suffisante pour différencier des homonymes.

Afin de confirmer l'existence de ce problème d'appariement des notices d'autorité lorsqu'il n'est pas fait appel à des informations complémentaires, on a extrait des fichiers d'autorité LC et DDB un échantillon de noms figurant dans l'un comme dans l'autre. Ces paires de notices d'autorité ont ensuite été examinées manuellement afin de déterminer si elles correspondaient ou non à une seule et même personne. De cet examen il est ressorti qu'environ 10 % des paires de noms de personne correspondaient à deux personnes distinctes. Le taux d'erreur d'appariement, si l'on n'utilisait que la forme établie du nom, atteindrait donc un niveau inacceptable. Comme les formes de noms ne sont pas toujours identiques entre les deux fichiers d'autorité nationaux, le fait d'apparier des noms similaires mais pas tout à fait identiques conduirait à un taux d'erreur encore plus élevé. Cette approche simple ne permet pas non plus d'apparier les nombreux noms qui ont été établis sur des bases différentes.

La solution pour permettre l'appariement des noms

Il est donc, de toute évidence, nécessaire de faire appel à des informations complémentaires pour confirmer ou infirmer un appariement potentiel de noms de personnes. Par exemple, prenons les informations d'autorité suivantes, enregistrées par la Bibliothèque du Congrès pour Diane Glynn :

```
100 10 $a Glynn, Diane, $d 1946-
400 10 $a O'Connor, Diane, $d 1946- $w nna
670    $a Country western dancing, 1994:  $b CIP t.p. (Diane
        Glynn) pub. info. (an avid country w. dancer & co-author
        of How to make your man more sensitive)
```

Les seules données directement utilisables sont les noms et la date de naissance. Deux titres sont mentionnés dans le champ 670 (Source des données) et pourraient être extraits de manière automatisée. En pratique, il n'est possible d'extraire de ces champs de manière fiable que certains de ces titres.

Les notices bibliographiques constituent une source évidente d'informations complémentaires au sujet d'une personne. Ces notices bibliographiques peuvent être orpaillées pour y trouver des attributs suplémentaires de l'œuvre d'une personne, qui peuvent permettre de distinguer cette personne de ses homonymes. On trouve par exemple, sur l'une des notices bibliographiques :

```
100 1  $a Glynn, Diane, $d 1946- -
245 10 $a How to make your man more sensitive / $c by Diane and
        Dick O'Connor.
700 1  $a O'Connor, Dick, $d 1938- $e joint author –
```

Les notices bibliographiques comportent deux types d'informations complémentaires. Les notices bibliographiques comprennent généralement des informations spécifiques à l'œuvre, telles que le titre, et des informations spécifiques à la manifestation, telles que

l'ISBN. Un appariement des titres permet d'établir de façon quasi certaine un appariement de noms de personnes.

La notice bibliographique contient en outre des informations complémentaires qui peuvent s'appliquer à plusieurs œuvres d'une même personne. Ces informations peuvent aider le processus d'appariement des auteurs quand il n'est pas possible d'apparier des titres spécifiques. Le co-auteur Dick O'Connor fournit un exemple de ce type d'information. Dick O'Connor peut être co-auteur de plus d'un ouvrage de Diane Glynn, ce qui corrobore un appariement des noms entre fichiers d'autorité. Même si le même ouvrage figure dans les deux bases de données nationales, mais seulement à l'état de traduction dans l'une des deux, il peut s'avérer difficile de procéder à un appariement automatique des titres. Dans ce cas, le nom du co-auteur présentera encore plus probablement des similitudes entre les deux bases de données, ce qui corrobore l'appariement.

Toutes les notices bibliographiques disponibles où le nom figure comme accès principal, accès secondaire, ou sujet, sont transformées pour créer une notice intermédiaire appelée « autorité dérivée ». Ces notices d'autorité dérivées sont ensuite combinées avec la notice d'autorité d'origine afin de créer une notice d'autorité élargie. Comme les notices d'autorité élargies comportent des informations complémentaires associées au nom tiré des notices bibliographiques, elles viennent appuyer un processus d'appariement plus rigoureux que les notices d'autorité elles-mêmes.

Confirmation de l'appariement de noms

Une simple comparaison de noms entre deux fichiers d'autorité nationaux constitue une façon raisonnable de retrouver un même individu. On peut s'attendre à des variations dans la forme du nom, ce qui diminue les chances d'avoir affaire à un seul et même individu. Pour confirmer de manière automatisée un appariement de noms, nous partons du principe que (1) les noms doivent être compatibles, et (2) il faut qu'il y ait suffisamment d'informations complémentaires qui confirment l'appariement.

La notion de compatibilité signifie que l'on ne relève pas de différences qui interdiraient que deux noms représentent une même personne. Les noms peuvent différer en complétude : John A. Smith / John Allen Smith, par exemple. Ces noms sont compatibles parce que le « A. » peut être l'initiale de Allen. En revanche, John A. Smith et John B. Smith ne sont pas compatibles parce qu'ils n'ont pas la même initiale médiane. Les tests de compatibilité prennent en considération tant les variantes de forme que la forme autorisée du nom.

Une fois que l'on a déterminé la compatibilité des noms, on utilise les informations complémentaires collectées sur ces noms afin de corroborer l'appariement. Les fichiers bibliographiques peuvent comporter beaucoup de titres similaires mais non identiques, ainsi que beaucoup de noms de personnes similaires mais non identiques. Toutefois, si une paire nom de personne/titre est similaire dans les deux fichiers, il est vraisemblable

que les noms de personne désignent une même personne. Cette stratégie fondamentale est étendue à d'autres types d'informations tirées des notices bibliographiques.

Les dates sont considérées séparément comme une corrélation positive. Quand les dates divergent de plus d'un an, les noms sont réputés non compatibles, et l'appariement est rejeté. On autorise des divergences d'un an dans les dates. Au cours du développement du VIAF, il s'est avéré relativement courant de relever de petites différences dans quelques dates, et les informations complémentaires d'appariement ont suffi pour corroborer l'appariement même avec de légères divergences dans les dates.

Lorsque nous comparons deux notices d'autorité élargies, chaque élément qui correspond est considéré comme un « point d'appariement ». Les points d'appariement sont ventilés en trois catégories : fort, moyen, faible. Pour des noms compatibles, un point d'appariement fort est réputé suffisant pour confirmer que deux individus sont une seule et même personne. Les points d'appariement forts sont : les titres, les ISBN, les dates de naissance et de mort, les co-auteurs. La date de naissance toute seule n'était pas suffisante pour différencier des noms, et nous la considérons donc plutôt comme un point d'appariement moyen. Les points d'appariement moyens sont des indicateurs de l'environnement de travail des personnes : éditeurs commerciaux, domaine thématique, fonction de la personne par rapport aux documents (illustrateur, compositeur…). Un grand éditeur publie les œuvres de nombreux auteurs, dont au moins quelques-uns peuvent avoir des noms semblables. Des correspondances sur plusieurs points d'appariement moyens suffisent pour confirmer l'appariement. Les points d'appariement faibles ne sont réputés suffisants que pour différencier des appariements par ailleurs ambigus. Ces points d'appariement faibles comportent la langue, le domaine thématique, et le pays de publication.

Pour combiner des points d'appariement, des valeurs numériques sont attribuées à chaque point d'appariement. Pour un numéro tel que l'ISBN, la correspondance doit être absolue, ou ce n'est pas une correspondance. L'ISBN se voit donc affecter une valeur de 1 s'il y a correspondance ou de 0 s'il n'y a pas correspondance. Pour une donnée textuelle telle que le titre, on attribue une valeur entre 0 et 1, en fonction du degré de similitude entre les données textuelles. On utilise une technique de valuation fondée sur des trigrammes pour noter le degré de similitude entre deux textes. Les valeurs individuelles sont pondérées selon la force du point d'appariement (fort, moyen, faible), et additionnées. Lorque le résultat dépasse un seuil déterminé par le processus de test, l'appariement est corroboré. Dans l'algorithme d'appariement réel, l'examen d'un grand nombre de notices a permis d'ajuster les valuations au sein de chaque catégorie, et nous serons sans doute amenés à affiner encore ces ajustements au fur et à mesure que des fichiers d'autorité seront intégrés au système et que nous acquerrons plus d'expérience.

Élaboration des notices d'autorité élargies

Les techniques décrites ci-dessus ont été utilisées pour créer des notices d'autorité élargies pour les notices d'autorité du PND et de la Bibliothèque du Congrès. Les fichiers bibliographiques de la Bibliothèque du Congrès ont été traités pour que des notices

d'autorité dérivées élargissent le fichier d'autorité de la Bibliothèque du Congrès, et les fichiers bibliographiques de la DDB et du BVB ont été traités pour élargir les notices d'autorité du PND. La figure 1 montre un schéma simplifié du flux d'information qui a permis ce processus d'élargissement des notices d'autorité.

Pour le fichier d'autorité élargi de la Bibliothèque du Congrès, 3,8 sur les 4,2 millions (soit 90 %) des notices d'autorité ont pu être élargies. Seuls 2,6 millions (soit 60 %) ont été élargies avec des informations provenant des notices bibliographiques, au total 7,4 millions de titres. D'autres élargissements ont été réalisés, en faisant appel à 4,1 millions de titres tirés du champ 670 (Source des données) des notices d'autorité. Les titres représentent l'élément le plus important d'élargissement pour procéder à des appariements de noms, comme on pourra le voir dans la section consacrée aux résultats.

Pour le fichier d'autorité élargi du PND, 2,4 sur 2,6 millions (soit 90 %) des notices d'autorité ont reçu un élargissement, mais seuls 2 millions (soit 20 %) ont été élargis grâce aux notices bibliographiques. Les 400 000 notices restantes ont été élargies avec des titres tirés des notices d'autorité du PND elles-mêmes.

Techniques de test des appariements

Les participants au projet VIAF ont aidé à l'élaboration du processus d'appariement en en vérifiant l'exactitude et en en commentant les résultats. Par exemple, les titres de collection ont été utilisés au départ, mais il s'est avéré qu'ils débouchaient fréquemment sur des appariements qui n'avaient pas lieu d'être. Chaque examen a débouché sur des modifications soit dans le sens de la diminution du nombre d'appariements erronés soit dans le sens d'une hausse du nombre d'appariements valides. Au cours de cette période, on a procédé à une estimation de la valeur de seuil d'exactitude (dans des proportions raisonnables) et au développement de l'algorithme de valuation. On ne décrira ici que les tests finaux de confirmation.

Afin de confirmer l'exactitude et l'efficacité du processus d'appariement, des échantillons de noms appariés ont été examinés par des catalogueurs chevronnés de la DDB et de la LC. Le premier échantillon visait deux objectifs : déterminer le taux de recouvrement des noms entre les deux fichiers d'autorité, et trouver quelle proportion sur ces paires de noms pouvait être identifiée par le processus d'appariement. Le deuxième échantillon a servi à rechercher toutes les erreurs ou déficiences systématiques qui pouvaient être rectifiées, et à estimer le taux d'erreur général.

Le premier échantillon comportait 391 notices d'autorité extraites de manière aléatoire du PND. On a lancé des requêtes, à la fois manuellement et selon une procédure automatisée, sur le fichier d'autorité de la LC pour récupérer les notices qui leur correspondaient. La requête automatisée portait la correspondance entre noms de famille, et a ramené 74 000 paires de notices. L'algorithme d'appariement a été appliqué à l'ensemble de ces 74 000 paires et a débouché sur 79 paires de notices d'autorité.

L’examen manuel de l’ensemble des 391 notices d’autorité du PND a permis de trouver en outre 35 noms qui avaient un pendant dans le fichier d’autorité de la LC, mais soit l’appariement ne portait pas sur le nom de famille, soit l’algorithme d’appariement ne permettait pas de confirmer l’appariement. Cet examen manuel a permis de confirmer les 79 appariements qui avaient été réalisés de manière automatique. Sur la base de l’échantillon du PND, on estime qu’environ 30 % des noms du PND figurent également dans les notices d’autorité de la LC, et que l’algorithme peut apparier environ 70 % de ces noms qui figurent dans les deux bases. Ce qui revient à dire que l’on estime à 800 000 le nombre de noms qui existent dans les deux bases, sur lesquels le processus d’appariement automatisé devrait permettre d’identifier 550 000 véritables paires.

Les résultats ont également fait l’objet d’un examen afin d’améliorer le processus d’appariement des noms. En n’utilisant que les noms de famille, près de 1 000 paires de noms devraient être soumises au processus complet d’appariement. Le test d’appariement manuel a permis d’établir qu’une stratégie fondée sur le nom de famille, le prénom et quelques informations de date pouvait être utilisée comme estimation grossière de la compatibilité des noms. Ce simple indice s’est avéré utile pour 95 % des appariements, avec seulement 4 paires de noms à comparer à chaque fois. Ce simple indice est donc pleinement efficace, et des ajustements minimes devraient permettre de l’améliorer encore.

L’objectif du deuxième échantillon était d’estimer le taux d’erreur dans les appariements. Au cours du processus, l’échantillon a permis de tester la pertinence de la valeur-seuil, et de l’ajuster si nécessaire. Lorsque l’on utilise une valeur-seuil, le taux d’erreur pour des appariements dotés d’une valeur proche du seuil devrait être supérieur à celui d’appariements dont la valeur est franchement supérieure au seuil. La plupart des notices d’autorité nom de personne qui se correspondaient avaient des valeurs largement supérieures au seuil. Afin d’estimer au plus juste le taux d’erreur en faisant appel le moins possible à un examen manuel, l’échantillon a été scindé en quatre sous-échantillons en fonction de la valeur attribuée. Des examens manuels ont permis d’identifier toutes les erreurs d’appariement, et le taux d’erreur a été établi pour chaque sous-échantillon. Ces résultats partiels ont été pondérés et additionnés pour déterminer le taux d’erreur global de la technique d’appariement. Le nombre d’appariements qui n’avaient pas lieu d’être s’est révélé inférieur à 1 %.

L’un des sous-échantillons semblait se situer juste en dessous du seuil. Si le seuil était abaissé, on aurait un faux appariement supplémentaire pour trois appariements valides. Il n’est donc, de toute évidence, pas opportun d’abaisser le seuil. Juste au dessus du seuil, il n’y avait qu’un faux appariement sur 25. Comme il y a relativement peu de faux appariements dans cette zone de valeurs, l’impact global sur le taux d’erreur est faible, et on garde un grand nombre d’appariements valides. Le seuil a donc été maintenu tel qu’il avait été initialement défini.

Développement du noyau initial du VIAF

Les fichiers d'autorité élargis issus des deux fichiers sources ont été passés à l'algorithme d'appariement, et les notices résultantes, qu'elles soient appariées ou non, ont été converties en notices VIAF. Ce processus est représenté à la figure 2. Il y a 6,3 millions de notices dans le fichier VIAF résultant, avec 500 000 notices liées, 3,7 millions de notices non appariées provenant du fichier d'autorité de la LC, et 2,1 millions de notices non appariées provenant du fichier d'autorité PND. Ces résultats sont très proches de l'estimation fournie par les tests manuels. On estime qu'il y a en outre 250 000 paires de notices d'autorité représentant une même personne qui n'ont pas pu être appariées automatiquement faute d'informations utilisables. Le système final permettra de lier à la main des notices qui sont dans ce cas et de réaliser d'autres appariements identifiés intellectuellement. Les notices d'autorité comporteront un numéro de notice VIAF attribué séquentiellement.

La figure 3 donne un exemple de notice VIAF au format MARC21. L'objectif principal du VIAF étant de créer des liens entre fichiers , la notice VIAF contient une entrée pour chaque forme de nom en champ 700 (liaison des vedettes établies), avec indication de sa source. Comme il n'y a pas de forme autorisée unique, le champ 100 (vedette nom de personne) n'est pas utilisé. Quand un appariement est déterminé par l'algorithme, deux vedettes liées figurent dans la notice. Quand un nom n'est pas apparié, un seul champ 700 y figure.

Les informations complémentaires figurent également dans les notices d'autorité élargies, dans les champs de données locales (9XX). Les champs de données locales utilisés dans les notices d'autorité élargies sont sommairement décrits sur la figure 4. Pour simplifier l'appariement, l'ensemble du texte est normalisé grâce à une version modifiée des règles de normalisation du NACO (*Name Authority Cooperative Program of the Program for Cooperative Cataloging*) [7]. Le nombre d'occurrences d'un terme donné est stocké en sous-champ $9. Comme cette information n'est destinée *a priori* qu'à un traitement par machine, elle n'est pas forcément visible à l'affichage des notices pour les utilisateurs. Au fur et à mesure que d'autres fichiers d'autorité nationaux viendront s'intégrer au VIAF, ils seront d'abord comparés avec les notices VIAF élargies existantes, et de nouveaux appariements seront intégrés au fur et à mesure qu'ils seront établis.

Dans un nombre non négligeable de cas, un nom autorisé dans un fichier correspond à plusieurs noms autorisés dans l'autre fichier. L'un des buts du VIAF étant d'établir des liens de cardinalité 1-1, les appariements ne sont pas confirmés dans de tels cas. 70 000 appariements obtenus par algorithme ont ainsi été rejetés. Deux raisons au moins ont été identifiées pour expliquer ce cas de figure.

Premièrement, il existe un certain nombre de noms non différenciés dans le PND, qui correspondent chacun à deux noms désambiguïsés ou davantage dans le LCNAF. S'appuyant sur les règles de catalogage allemandes RAK-WB, la pratique catalographique allemande ne différenciait pas les noms de personnes. Quand la DDB a commencé à cataloguer avec un fichier d'autorité, cette pratique a été abandonnée et la DDB ne crée plus de notices d'autorité nom de personne pour des noms non différenciés. Toutefois, le PND contient encore de nombreux noms non différenciés. La DDB

désambiguïsera ces noms en leur affectant plusieurs appariements, et ce de manière automatisée autant que faire se peut, sur la base d'appariements entre les titres mentionnés dans les notices d'autorité élargies de la LC et de la DDB, puis de manière intellectuelle. Les corrections seront versées dans le VIAF dans le cadre des fréquentes mises à jour, ce qui générera des liens non ambigus entre les notices appariées.

Deuxièmement, un certain nombre de notices d'autorité de la LC reflètent la pratique des AACR2 qui consiste à avoir des notices d'autorité séparées pour chaque identité bibliographique utilisée par une personne, par exemple dans le cas des pseudonymes. C'est le cas inverse des notices non différenciées du PND. Dans ce cas de figure, plusieurs notices d'autorité sont créées pour une seule personne réelle. Le PND, qui suit les règles RAK-WB, n'a qu'une seule notice d'autorité pour les noms de toutes les identités bibliographiques d'une personne. À l'instar des noms non différenciés, ces notices d'autorité « sur-différenciées » posent des problèmes pour lesquels nous n'avons pas encore trouvé de solution pleinement satisfaisante.

Les noms liés peuvent être utilisés directement pour la traduction automatique des autorités LC vers les autorités PND ou vice versa. Cela peut alimenter les techniques du Web sémantique ou les systèmes de recherche fédérée. Le maintien des formes de renvoi peut fournir des informations supplémentaires à l'observateur humain.

Les identifiants des autorités dans les fichiers des institutions participantes ou les identifiants VIAF eux-mêmes peuvent aussi constituer la base d'URI. On pourrait y voir les bases d'un service résolveur pour des URI d'autorité. À partir de n'importe quelle mention d'URI dans un document, une notice ou sur un site Web, l'utilisateur pourrait être conduit vers tous les documents, toutes les notices, toutes les ressources, etc., auxquels sont reliées les autorités représentées dans les URI, ainsi qu'aux notices d'autorité elles-mêmes.

Système actuel

Les fichiers nationaux d'autorité noms et les bases de données bibliographiques changent continuellement. Pour une base de donnée liée qui est construite sur deux ou plus de deux fichiers évolutifs, il est indispensable de revérifier fréquemment les liens et de les mettre à jour. La logique et le logiciel du système VIAF initial sont en cours de modification pour permettre la mise à jour permanente des notices. Lorsque de nouvelles notices bibliographiques ou d'autorité sont reçues, les notices d'autorité élargies existantes sont modifiées, et l'appariement entre bases de données est rafraîchi. De nouveaux appariements ne cessent d'être établis, et les appariements qui n'ont plus lieu d'être en raison de changements dans les notices sources sont rompus. Lorsque des appariements sont rompus, chacune des deux notices de l'ancien couple comporte un historique de l'appariement, à des fins de documentation.

Dans l'avenir, le système VIAF sera alimenté via le protocole OAI à partir des bases de données sources, lorsque le protocole OAI aura été implémenté dans ce contexte. D'ici là,

des procédés plus traditionnels d'accès aux fichiers, tels que FTP, seront mis en œuvre pour tester le projet.

Vu la quantité de données disponibles à un seul endroit, on peut envisager de nombreuses méthodes différentes pour l'accès aux données et leur utilisation. Les liens peuvent servir à traduire un nom de personne dans le format voulu par l'utilisateur final, dans le cadre du Web sémantique. On pourrait construire des outils qui permettraient la recherche automatique dans d'autres bases de données [que celle où la requête est initialement lancée] en fournissant la forme de nom appropriée pour ces autres bases de données. On pourrait construire des outils de catalogage et de contrôle des autorités de manière analogue, en identifiant la forme appropriée d'un nom figurant dans une notice. Bien sûr, il sera également possible de lancer des requêtes directement sur la base de données VIAF.

Conclusions

Le fichier PND a déjà tiré des profits substantiels du projet. Les tests d'auto-appariement dans les deux fichiers ont permis d'améliorer sensiblement le PND, et la DDB espère trouver une aide considérable dans les paires de notices élargies pour différencier les noms de personnes grâce aux titres appariés. Les processus et algorithmes d'appariement développés pour le projet peuvent être adaptés à nombre d'autres applications. On cherche à mettre sur pied des services qui feront appel aux données d'appariement des noms de personnes pour améliorer l'accès à l'information bibliographique et pour apporter un soutien aux activités de catalogage des institutions participantes.

Le projet a montré qu'il est faisable de lier automatiquement les noms de personnes de deux fichiers d'autorité nationaux. Soixante-dix pour cent des notices d'autorité noms de personnes figurant dans les deux fichiers ont pu être reliés avec un taux d'erreur inférieur à un pour cent. La stratégie consistant à apporter aux notices d'autorité originales un complément d'information à partir des notices bibliographiques a grandement amélioré le taux d'appariement, tout en réduisant le nombre d'appariements erronés. Il suffirait de changements mineurs à apporter aux notices d'autorité pour améliorer les appariements. Nombre d'appariements ratés viennent de ce qu'il n'a pas été possible de parser correctement le champ 670 (Source des données). On pourrait tirer un parti non négligeable de structures complémentaires, en évitant l'utilisation de noms ou de titres abrégés, ou de liens explicites à la notice bibliographique source. L'identification explicite de la fonction ou du domaine d'activité [des personnes] (compositeur, illustrateur, mathématicien, etc.) améliorerait encore les possibilités d'appariement, tant à la main que de manière automatisée, de même que l'enregistrement des formes complètes de noms, au moins sous la forme de renvois.

Notre recherche plaide de manière convaincante en faveur du contrôle des autorités, de l'utilisation des notices d'autorité, du travail en réseau et de la création de liens inter-bases, et de l'élaboration d'un Web sémantique pour les bibliothèques. Pour les bibliothèques et réseaux de bibliothèques en Allemagne qui ont des notices bibliographiques comportant des points d'accès issus du LCNAF, le VIAF pourrait servir

de pivot entre fichiers d'autorité, soit pour transcrire les points d'accès issus du LCNAF dans les notices bibliographiques sous la forme de points d'accès de type PND, soit pour permettre la recherche dans le VIAF avec des vedettes issues du PND. Implémenté dans des portails multinationaux ou multilingues tels que le portail de The European Library, le VIAF pourrait combiner automatiquement des requêtes dans le LCNAF et le PND, ce qui conduirait l'utilisateur vers des notices bibliographiques provenant des deux sources.

Avec les techniques d'appriement qui ont été mises en place, on projette un système qu'il sera possible de mettre à jour et qui collectera les données actuelles d'autorité nom de personne et bibliographiques produites par les institutions participantes, au moyen du protocole OAI. Ce système devra être modulable et de nouveaux participants désireux de partager leurs notices bibliographiques et d'autorité seraient les bienvenus. Nous ne pourrons pas établir les limites de la modularité du VIAF tant qu'un plus grand nombre d'institutions ne se sera pas associé au projet.

Le projet VIAF s'est concentré sur le problème de l'appariement des notices d'autorité. Il nous faudra un service à long terme et une stratégie de gouvernance pour maintenir, étendre et implémenter le VIAF. Des décisions doivent être prises concernant l'extension du projet aux noms de collectivités et l'accroissement du nombre d'institutions participantes. Nous avons des projets pour étendre les possibilités du système en faisant appel aux polices de caractères Unicode. Unicode permettra d'intégrer les écritures non latines mais l'extension de l'algorithme d'appariement constituera un vrai défi, notamment pour des écritures telles que le hangul, les caractères idéographiques chinois, ou les différentes écritures du japonais.

Références bibliographiques

1. IFLA Core Activity: IFLA-CDNL Alliance for Bibliographic Standards (ICABS) http://www.ifla.org.sg/VI/7/icabs.htm [Mai 2006]

2. Berners-Lee, Tim, James Hendler, et Ora Lassila. "The semantic Web: a new form of Web content that is meaningful to computers will unleash a revolution of new possibilities." *Scientific American*, May 17, 2001. http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21 [Mai 2006]

3. LEAF Project, http://www.leaf-eu.org [Mai 2006]

4. Project InterParty: From Library Authority Files to E-Commerce, Andrew MacEwan, http://www.haworthpress.com/store/E-Text/View_EText.asp?a=3&fn=J104v39n01_11&i=1%2F2&s=J104&v=39 [Mai 2006]

5. VIAF: The Virtual International Authority File, http://www.oclc.org/research/projects/viaf [Mai 2006]

6. Open Archives Initiative - Protocol for Metadata Harvesting, http://www.openarchives.org/OAI/openarchivesprotocol.html [Mai 2006]

7. Hickey, Thomas B., Jenny Toves, et Edward T. O'Neill. "NACO Normalization: A detailed Examination of the Authority File Comparison Rules", *Library Resources & Technical Services*, Vol. 50, No. 3, p. 18-24. [à paraître]

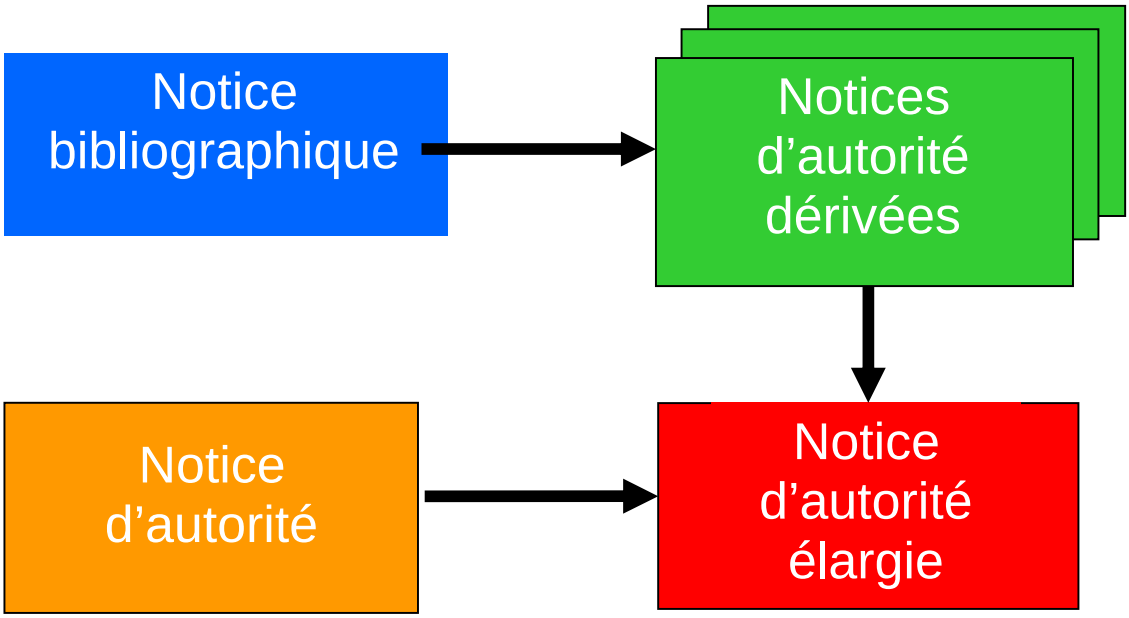

Figure 1. Création de la notice d’autorité élargie

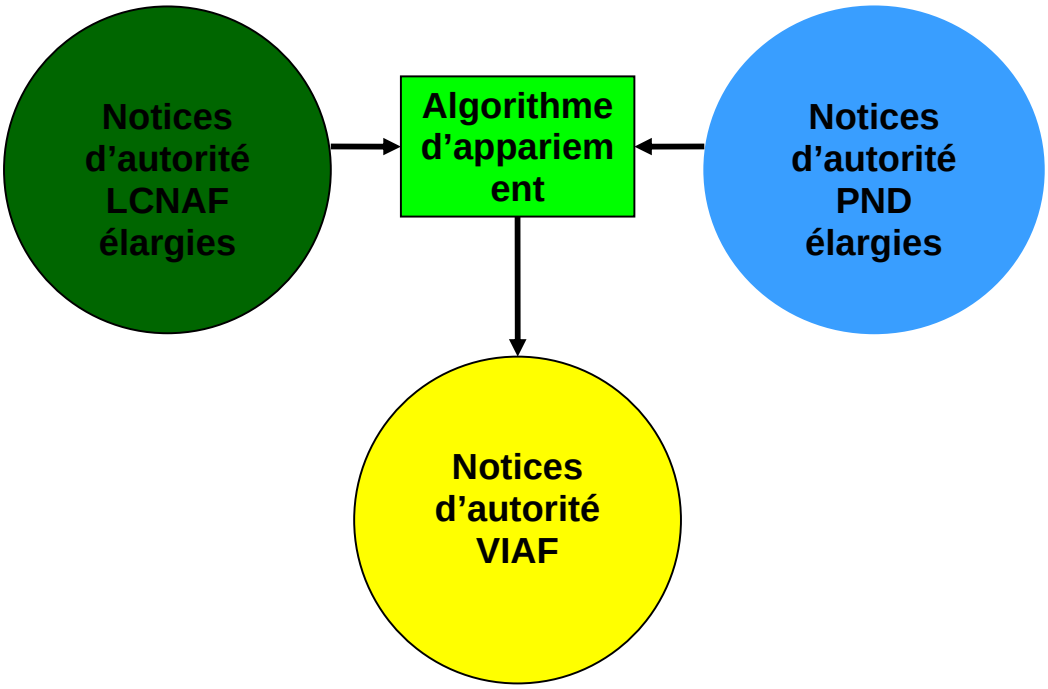

Figure 2. Création des notices d'autorité VIAF

```
000    nz n
001    viaf 30543
005    20050826163535.0
008    050826n||anannabbn |a aaa
040    VIAF   $c VIAF
400 10    $w nnaO'Connor, Diane,   $d 1946-
700 17 Glynn, Diane,   $d 1946-   $2 DLC   $0 n 94057411
700 17 O'Connor, Diane   $2 DDB   $0 108982424
901    052512920   $9 1
901    349917275   $9 1
901    350215532   $9 1
903    75014386   $9 1
910 11 how to make your man more sensitive   $9 3
910 11 macht eure manner zartlicher   $b liebevolle ratschlage
fur e neues rollenverhalten   $9 1
910 11 macht eure manner zartlicher   $b wie e frau ihrem mann
helfen kann e verstandnisvoll   $9 1
919    country western dancing,   $9 1
920    0-525   $9 1
920    3-499   $9 1
920    3-502   $9 1
921    dutton   $9 1
921    rowohlt   $9 1
921    scherz   $9 1
922    gw   $9 2
922    nyu   $9 1
940    eng   $9 1
940    ger   $9 2
942    18   $9 1
943    197x   $9 3
944    am   $9 3
950 11 oconnor, dick   $9 2
950 11 oconnor, dick   $d 1938   $9 1
999    1   $b 75014386 //r94   $2 DLC
999    1   $b n 94057411   $2 LoCNA
999    2   $b 780147766   $b 790425319   $2 DDB
```

Figure 3. Notice VIAF

Figure 4
Format des notices élargies

90x	**Numéros de contrôle**	
901	ISBN	$a Portion numérique de l'ISBN (sans caractère de contrôle ni barres obliques)
902	ISSN	$a Portion numérique de l'ISSN (sans caractère de contrôle ni barres obliques)
903	LCCN	$a Portion numérique du LCCN (sans caractère de contrôle ni barres obliques)
91x	**Champs titre**	
910	Titre du 245	Sous-champs a & b
911	Titre abrégé du 210	Sous-champs a & b
913	Titre uniforme du 130 ou du 240	Sous-champs a & b
914	Titre traduit du 242	Sous-champs a & b
915	Titre uniforme collectif du 243	Tous les sous-champs
916	Variante de titre du 246	Sous-champs a & b
917	Titre uniforme notice d'autorité	Tiré des notices d'autorité auteur/titre, champ 100 $t
919	Titre tiré d'un autre texte	Diverses notes ou champs similaires
92x	**Champs éditeur**	
920	Identifiant de l'éditeur	$a Identifiant de l'éditeur tiré de l'ISBN
921	Nom de l'éditeur	$a Nom de l'éditeur tiré du 260 b ou du 533 c
922	Lieu d'édition	$a Code du pays d'édition tiré du 008
93x	**Utilisation**	
930	Utilisation du nom	$a Forme du nom trouvée dans la mention de responsabilité, 245 sous-champ c
94x	**Attributs**	
940	Langue	$a Code de langue tiré du 008 ou du 041 sous-champ a
941	Fonction de l'auteur	$a Code de fonction tiré du 700, sous-champs e et/ou 4
942	Sujet NATC	$a Identifiant NATC
943	Décennie d'édition	$a Décennie d'édition
944	Format	$a Type et niveau bibliographique (008/06-07)
945	Sujet Conspectus	Voir discussions sur le PND
95x	**Co-auteurs**	
950	Auteurs personnes physiques	Sous-champs $a, $b, $c, $d, et $q des champs 100 et 700
951	Auteurs collectivités	Sous-champ $a des champs 110 et 710
96x	**Sujets noms**	
960	Nom en sujet	Sous-champs $a, $b, $c, $d, et $q du champ 600
969	Utilisation matière	Texte "Subject" signalant que la vedette était utilisée en matière, et a été tirée d'un champ 600
99x	**Champs spécifiques**	
999	Notices bibliographiques associées	$a Nombre total de notices $b Numéro de contrôle de la notice $2 Source de la notice