Acronyme / Acronym	**Qualinca**		
Titre du projet	Qualité et interopérabilité de grands catalogues documentaires		
Proposal title	Quality and Interoperability of Large Catalogues of Documents		
Axe(s) thématique(s) / theme(s)	☐ 1 ☑ 2 ☐ 3 ☐ 4 ☐ 5 Enrichissement sémantique et raisonnement Agrégation de contenu et de connaissances		
Type de recherche / Type of research	☑ Recherche Fondamentale / Basic Research ☐ Recherche Industrielle / Industrial Research ☐ Développement Expérimental : Experimental Development		
Types de projets spécifiques	☐ Plate-forme (cocher si ce projet est une mise en place/construction de plate-forme au sens de l'appel à projets) /Platform		
Coopération internationale / International cooperation	☐ Le projet propose une coopération internationale / International cooperation		
Aide totale demandée / Grant requested		**Durée du projet / Projet duration**	42 mois

1. RESUME DE LA PROPOSITION DE PROJET

Les grands catalogues documentaires sont en train de passer de l'ère de la gestion de bases de métadonnées dans des formats spécifiques issus de la communauté des Sciences de l'Information et des Bibliothèques (SIB) à l'ère du Web dans les langages standards du web sémantique (RDF/S, OWL). Cette évolution, qui présente de nombreux avantages (meilleure exposition des fonds documentaire, augmentation des possibilités d'échange de données, création de nouveaux services de recherche/d'exploitation des fonds), pose des problèmes importants concernant la qualité des bases documentaires.

Ce projet se propose d'élaborer des mécanismes permettant de :

- qualifier le niveau de qualité d'une base documentaire existante ;
- maintenir un niveau de qualité donné en contrôlant les opérations de mise à jour de ces bases ;
- améliorer le niveau de qualité d'une base ;
- disposer de méthodes génériques d'exploitation de ces bases dépendants de leur niveau de qualité (par exemple pour la recherche de documents ou l'interconnexion de bases).

Grâce à la représentation des données dans les langages du web sémantique une approche « représentation des connaissances » de ces problèmes est possible. Cette approche permettra, d'une part, de donner une sémantique logique à la notion de qualité et, d'autre part, d'utiliser des mécanismes de raisonnement pour traiter les divers problèmes. Cette approche repose sur la formalisation des connaissances présentes dans les catalogues documentaires, l'élaboration d'un modèle de qualité pour la problématique de l'identification des entités individuelles (ou entités nommées) dans une base de connaissances, la définition d'un modèle original de confiance adapté à la réconciliation et à la fusion d'informations provenant de différentes sources, et la découverte de caractéristiques d'identification d'entités et leur exploitation selon différentes approches (logique, numérique, probabiliste, …).

Une large part du projet est dévolue à l'évaluation de l'approche proposée par des expérimentations menées sur des corpus de tests et par le développement de démonstrateurs adaptés au contexte métier de deux gestionnaires de bases documentaires.

Le consortium regroupe cinq partenaires complémentaires : deux acteurs nationaux majeurs des systèmes documentaires et trois équipes de chercheurs en informatique. L'Agence Bibliographique de l'Enseignement Supérieur (ABES) et l'Institut National de l'Audiovisuel (Ina) sont détenteurs et gèrent de très grandes bases documentaires, ils sont très fortement investis, au plan national et au plan international, dans l'exposition, la standardisation, l'interconnexion et la valorisation de leurs métadonnées. Les équipes du LIG, du LIRMM et du LRI impliquées dans ce projet, possèdent de leur côté une expertise reconnue en bases de données, représentation des connaissances et web sémantique. De plus, de multiples et anciens liens existent entre les partenaires de ce projet. Les compétences des partenaires et les liens scientifiques tissés entre eux dans le cadre de projets communs sont très importants pour le succès de ce projet pluridisciplinaire qui concerne aussi bien les Sciences de l'Information et des Bibliothèques que l'Informatique, et qui devrait avoir des retombées, non seulement dans le domaine des bases documentaires mais aussi dans le Web des données (« Linked Data »).

2. CONTEXTE, POSITIONNEMENT ET OBJECTIFS DE LA PROPOSITION

Les acteurs de la gestion des grands catalogues documentaires sont à l'heure actuelle sur le point de passer de l'ère de la gestion de bases de métadonnées dans des formats spécifiques issus de la communauté des Sciences de l'Information et des Bibliothèques (SIB) à celle du Web Sémantique (cf. les projets VIAF[1], Isidore[2],…). Cette évolution pour l'essentiel due aux perspectives que le web sémantique, renforcé par l'initiative « Linking Open Data », ont ouvertes en termes d'exposition de leur fonds documentaire, d'augmentation des possibilités d'échanges de données, ou de création de nouveaux services de recherche/d'exploitation de leur fonds, pose des problèmes similaires à ceux du « linked data », ceci dans un contexte spécifique.

En effet, la communauté des SIB dispose d'une longue tradition normative et a depuis longtemps intégré, à côté de ses catalogues de notices documentaires, des catalogues de notices d'autorités, qui recensent par catégorie les différentes entités utiles au catalogage des documents (les personnes, les collectivités, les lieux, les matières, …). Ces catalogues d'entités permettent, d'une part, de normaliser les termes utilisés pour référencer ces entités dans les différentes notices documentaires et, d'autre part, de rassembler dans une seule notice (celle de l'autorité) des informations concernant l'entité plutôt que de les dupliquer dans chaque notice. Ainsi, une base documentaire est à l'heure actuelle constituée d'un (ou de plusieurs) catalogue de notices documentaires possédant des liens qualifiés vers des notices de catalogues d'autorités (auteur, éditeur, sujet…). Ces liens sont issus d'un immense travail collaboratif fourni par l'ensemble des indexeurs de cette base documentaire. L'intégration automatique de catalogues issus d'autres bases documentaires, ainsi que la

[1] Virtual International Authority File http://viaf.org/.
[2] Plateforme unifiée d'accès aux données SHS http://www.rechercheisidore.fr.

nature coopérative de la création des notices ont conduit à des catalogues de qualité diverse. Si le contenu de chaque notice prise séparément est généralement considéré comme sûr, la confiance dans les liens entre notices est l'objet de toutes les attentions. En effet, l'accès principal aux autorités se fait au moyen d'un (ou plusieurs) terme choisi pour désigner cette autorité (nom et prénom pour une personne, nom commun pour une matière…). Au fil du temps, l'explosion du nombre d'autorités, en particulier pour les autorités « personne » et « collectivité », que l'on ne peut pas borner (de nouveaux auteurs apparaissant chaque jour), s'est heurtée au problème de l'homonymie des termes utilisés (il y a par exemple 18 *Alexandre Dumas* recensés dans VIAF). Pour pallier ce problème, les indexeurs ont complété ces termes avec des informations de diverses natures (dates de naissance, nationalité, fonction… Par exemple : « *Alexandre Dumas, père* », « *Alexandre Dumas, fils* » ou « *Alexandre Dumas, 18..-1916* ») qui si elles résolvent le problème de l'homonymie ne permettent pas toujours d'identifier l'entité du monde réel référencée par l'autorité (en particulier quand il ne s'agit pas de personnes célèbres). Face à cette situation, deux phénomènes se sont développés : l'accroissement d'erreurs de liage attribuant des documents à un auteur inadéquat (certes homonyme) et celle de doublons d'autorités dus à l'absence d'informations suffisantes dans les notices d'autorités permettant à un indexeur de reconnaître l'entité réelle référencée par une autorité et le conduisant, par principe de précaution, à créer une nouvelle autorité homonyme qu'à son tour il différenciera des autres par une information non forcément pertinente.

Dans le cadre d'une base documentaire isolée, ces liens entre notices documentaires et notices d'autorité sont utilisés pour permettre, par exemple, la recherche de documents par autorité. Les erreurs dans les liens augmentent le bruit et le silence du système de recherche. De telles erreurs deviennent cruciales dans les cas où les bases sont sémantisées (c'est-à-dire formalisées dans un langage de représentation de connaissances permettant de valider les inférences réalisées) et les liens exploités comme des connaissances pour permettre, par exemple, la fusion de plusieurs bases ou la liaison d'une base à d'autres bases (par exemple, dans le cadre de la recherche des ayants-droits d'une œuvre pour le versement de royalties, ou l'attribution de documents à un auteur dans le cas d'une étude bibliométrique).

L'objectif du projet Qualinca est de s'appuyer sur une sémantisation des bases documentaires, via la possibilité qu'elle offre de mettre en œuvre des techniques d'intelligence artificielle, pour contrôler et améliorer la qualité des liens dans les bases documentaires. Plus précisément, quatre objectifs principaux d'amélioration de la qualité seront poursuivis :

1. détection et réparation des erreurs de liage dans les bases documentaires ;
2. détection et fusion des doublons dans les catalogues d'autorités ;
3. enrichissement des autorités pour permettre l'identification par un humain des entités qu'elles représentent ;
4. représentation explicite du degré de confiance en la qualité des liens d'une base documentaire.

Pour atteindre ces objectifs, il est nécessaire de conduire des recherches fondamentales afin d'une part de disposer d'un cadre théorique fondé sur les principes de la représentation des

connaissances (bases de connaissances, ontologies, règles, contraintes) permettant de formaliser ces problèmes de qualité, et d'autre part de mettre au point des méthodes de raisonnement permettant de résoudre ces problèmes, notamment par la prise en compte de l'incertitude sur la qualité des données et sur la pertinence des règles permettant d'établir/valider des liens.

De plus, et bien que ce projet n'ait pas pour ambition de développer des outils directement destinés à un usage « en production », l'objectif « qualité » rend nécessaire le développement de prototypes permettant l'évaluation des méthodes proposées sur des données réelles. Les grandes bases documentaires de l'ABES et de l'Ina permettront cette évaluation.

2.1. CONTEXTE ET ENJEUX ECONOMIQUES ET SOCIETAUX

La qualité des grandes bases documentaires est essentielle pour répondre aux usages et à la nature spécifique de ces bases :

- la majorité des utilisateurs finaux est composée de professionnels (journalistes, enseignants, chercheurs, …) ou d'étudiants, c'est un public averti plus exigeant que le grand public ;
- les documentalistes qui ont en charge ces bases s'appuient sur ces bases elles-mêmes pour les maintenir (par exemple, pour l'insertion de nouvelles notices dans une base, l'enrichissement de notices existantes ou la fusion de bases) ; ils ont donc besoin que ces bases soient de qualité ;
- ces bases documentaires sont le résultat d'un travail considérable de professionnels qualifiés, elles ont un caractère patrimonial et leur pérennité nécessite qu'elles ne se dégradent pas au cours du temps, voire que leur qualité soit améliorée ;
- ces bases peuvent être utilisées dans des outils de pilotage, en particulier lorsque des procédures d'évaluation utilisent la bibliométrie, ou contenir des informations critiques, comme les ayants-droits d'œuvres.

Nous validerons nos méthodes sur les bases de l'ABES (Sudoc, 12 millions de notices documentaires) et de l'Ina (6 millions de notices documentaires) pour lesquelles l'enjeu principal est de *valoriser les investissements importants et de longue durée consentis par l'Etat, les universités et l'Ina pour décrire ces millions de documents*. Cette valorisation sera stimulée par :

- l'amélioration de la qualité des bases, ce qui permettra, notamment, d'améliorer les outils de recherche de documents proposés à l'utilisateur final ;
- la mise en conformité des données détenues avec les standards émergeants ;
- l'intégration des notices (documentaires et d'autorités) provenant d'autres organismes (comme la BNF pour l'ABES, FR3 ou RFI pour l'Ina) ;
- le développement de la politique d'ouverture des données sur le web des données, en décrivant ces données dans les langages du Web sémantique (RDFS-OWL) et en interconnectant des bases (par exemple, dans le cadre d'Europeana[3]).

D'autres enjeux, plus spécifiques mais également importants, seront pris en compte :

[3] Catalogue de ressources numériques des musées, bibliothèques, archives et collections audiovisuelles européennes http://www.europeana.eu.

- pour l'ABES, contribuer à la construction de l'identité numérique des chercheurs qui vise à consolider sous la forme de données structurées et fiables l'ensemble des activités scientifiques publiques des chercheurs et veiller à ce que cette consolidation des données soit maîtrisée par les scientifiques et leurs institutions ;
- pour l'Ina, améliorer la qualité de service et faire baisser les coûts des processus de libération des droits d'auteur et de reversement des royalties aux ayants-droits.

L'importance, aussi bien en termes de taille que d'usages, des bases de l'ABES et de l'Ina, est telle que nos méthodes devraient pouvoir s'appliquer aux autres grands catalogues documentaires français (par exemple ceux de la BNF) ou internationaux (par exemple, WorldCat).

La généricité des méthodes devrait aussi permettre, en facilitant le signalement et donc l'accès aux documents, de valoriser des archives numériques (par exemple celles issues des licences nationales).

2.2. POSITIONNEMENT DU PROJET

Positionnement par rapport au contexte des SIB

En tant que gestionnaire de catalogues (Sudoc, Calames, theses.fr), l'ABES fournit des services à l'utilisateur final mais également à d'autres gestionnaires de catalogues (OCLC pour Worldcat[4] par exemple) ou agrégateurs (Isidore du TGE ADONIS[5] du CNRS). L'enjeu de l'interconnexion des grands catalogues est désormais un enjeu global, qu'il s'agisse d'un enjeu pour le secteur des bibliothèques ou bien qu'il soit intégré dans la question plus générale du Web de données. OCLC[6], en particulier, opérateur global des bibliothèques, à la fois fournisseur et partenaire de l'ABES, poursuit une activité de recherche & développement importante, dont le projet VIAF.

VIAF est une base qui agrège des fichiers d'autorité fournis par les bibliothèques nationales ou de grands catalogues collectifs comme celui du Sudoc (depuis cet automne 2011). VIAF s'intéresse aux notices de personnes physiques, de collectivités et d'œuvres. VIAF contribuera aux initiatives ISNI[7] et ORCID[8], qui visent à associer un identifiant unique et global aux auteurs. L'interconnexion entre ces différents fichiers s'appuie sur des algorithmes propriétaires, qui comparent le contenu textuel des notices d'autorité elles-mêmes et les notices bibliographiques associées. En employant d'autres méthodes de rapprochement, Qualinca contribuera non seulement à la qualité du Sudoc et des bases nationales associées, mais également à l'interopérabilité au niveau global.

[4] Service d'accès aux catalogues documentaires de bibliothèques à travers le monde http://www.worldcat.org/.
[5] Très grand équipement pour l'accès unifié aux données SHS http://www.tge-adonis.fr/.
[6] Online Computer Library Center http://www.oclc.org.
[7] International Standard Name Identifier, norme ISO en cours de standardisation, http://www.isni.org/.
[8] Open Researcher & Contributor ID, Initiative communautaire de résolution du problème de l'ambiguïté des noms de contributeurs, http://orcid.org/.

L'Ina est partenaire du projet européen ICP/PSP ASSETS[9] destiné à fournir de nouveaux services à Europeana, le portail culturel européen. Dans le cadre de ce projet, l'Ina développe un outil d'annotation manuel destiné à la correction/complétion des notices documentaires ingérées dans le système. Cet outil offre à l'annotateur des fonctionnalités d'enrichissement s'appuyant sur la recherche d'entités nommées (personnes physiques et morales, lieux, artefacts, événements) dans trois grandes bases du web de données : DBPedia, FreeBase et Geonames. La réalisation d'alignements simples entre les différentes ontologies utilisées par ces bases ainsi que l'utilisation d'une métrique de similarité et de règles de fusion de données permet de présenter à l'utilisateur une vision unifiée « cross-bases » des entités disponibles et donc de faciliter leur sélection et in fine leur référencement dans les notices documentaires.

Europeana est actuellement en cours d'adoption de son nouveau modèle de données EDM (Europeana Data Model) en remplacement d'ESE (Europeana Semantic Elements). Avec ce nouveau modèle, Europeana tourne le dos à une représentation purement documentaire de ses données pour se tourner vers un modèle davantage proche de celui du Web Sémantique qui assurera une plus grande interopérabilité de ces données et permettra l'émergence de nouveaux services de recherche et de consultation. Ce nouveau format prendra notamment en charge le référencement d'entités nommées, fonctionnalité qui n'était pas incluse dans ESE.

Positionnement par rapport à la problématique du web de données

Le web des données trouve son origine dans une série d'initiatives éparses destinées à rendre accessible sur le Web non plus des documents mais des données structurées dans le but de rendre interprétables ces informations par la machine et notamment de créer de nouveaux services complexes mettant en jeu de véritables agents logiciels. DBPedia ou Freebase, par exemple, s'affirment comme d'immenses bases encyclopédiques trouvant leur origine dans l'exploitation de Wikipedia, la grande encyclopédie collaborative symbole du « crowd sourcing ». Geonames ou MusicBrainz, par exemple, sont, elles, spécialisées respectivement dans la géographie et la musique. Outre la mise à disposition de grandes quantités d'informations structurées permettant de lier entre eux des objets, personnes et événements, ces initiatives ont également pour but de proposer des identifiants uniques permettant d'améliorer l'interopérabilité des bases et *in fine* d'améliorer la transmission d'informations ou leur fusion.

En proposant des méthodes permettant à la fois de détecter / corriger des erreurs de liages dans des bases de métadonnées, de qualifier le niveau de description d'une entité nommée, de réconcilier des entités provenant de bases différentes, de fusionner des informations réparties dans ces bases, le projet Qualinca se place au cœur de la problématique du web des données. En effet, les méthodes de Qualinca peuvent s'appliquer en amont d'une phase de publication pour en évaluer et augmenter la qualité d'une base, pendant la publication pour identifier des liens d'interconnexion avec d'autres bases ou après publication pour fournir de véritables services d'identification d'entités individuelles.

[9] Advanced Service Search and Enhancing Technological Solutions for the European Digital Library http://www.assets4europeana.eu/.

Qualinca est donc complémentaire des projets DataLift (ANR Contint, http://datalift.org/) et LOD2 (FP7-ICT-2009-5, http://lod2.eu/) qui visent à développer des plateformes pour la publication, l'interconnexion et l'exploitation de données publiques sur le web des données.

Positionnement par rapport aux axes thématiques de l'appel

Les problématiques de l'axe 2 –*des contenus aux connaissances*– sont au cœur du projet Qualinca. En effet, tous les objectifs de Qualinca peuvent être décrits en utilisant exclusivement les différents domaines de cet axe. Nous nous intéressons à de grandes masses d'informations complexes et hétérogènes (des grandes bases documentaires concernant des documents textuels ou audiovisuels et composées de divers types de catalogues), desquelles nous extrairons automatiquement des informations (e.g., des clés) –*archivage, indexation, fouille de contenus*– sur lesquelles nous ferons des raisonnements à partir de la modélisation des connaissances présentes dans ces bases documentaires, en utilisant des standards et des ontologies représentant des référentiels et en mobilisant les technologies du web sémantique et de l'intelligence artificielle –*enrichissement, sémantique et raisonnement*– en vue d'améliorer la qualité des bases, de permettre leur interopérabilité –*agrégation de contenus et de connaissances*– et d'être utile dans l'amélioration de services existants ou la création de nouveaux services (e.g., bibliométrie, versement de royalties) –*nouveaux services et personnalisation*.

Positionnement par rapport aux projets précédents des partenaires

Certains des problèmes que nous aborderons dans Qualinca ont été rencontrés par des membres du présent consortium dans le cadre de divers projets. Citons par exemple :

- dans le cadre de Saphir (projet ANR-RIAM 05-09) et Logos (projet européen 05-08) : la nécessité de regrouper dans un référentiel les entités nommées importantes d'une application (avec suffisamment d'information pour pouvoir les identifier) ;
- dans le cadre d'Opales (projet ANR-PRIAMM 00-03) et de Mogador (projet Ministère de la Culture 00-01) : la nécessité d'avoir des procédures (semi) automatiques de liage entre termes d'une métadonnée et termes d'un référentiel pour diminuer le temps nécessaire pour faire manuellement des annotations sémantiques ;
- dans le cadre de SudocAd (projet CNRS-TGE ADONIS 09) : la nécessité d'avoir une base la plus correcte possible en ce qui concerne les liages pour mettre en œuvre des procédures automatiques d'enrichissement d'une base ;
- dans le cadre de Dataring (projet ANR-VERSO) : la nécessité de modéliser la confiance dans les systèmes pair-à-pair de gestion des données pour pouvoir distinguer la qualité des réponses à une requête fournies selon les sources de données dont elles proviennent ;
- dans le cadre de GeOnto (projet ANR-Masse de données et de connaissances) : la nécessité de réconcilier des instances pour améliorer l'alignement d'ontologies géographiques et vice-versa.

2.3. ÉTAT DE L'ART

Représentation de connaissances et raisonnements

Les systèmes à base de connaissances exploitent une représentation formelle des connaissances pour résoudre des problèmes. Le formalisme de base pour représenter et raisonner avec des connaissances est celui de la logique classique pour laquelle la déduction est un problème semi-décidable. Dès les années 80, de nombreux travaux ont cherché à identifier des langages qui d'une part disposent de syntaxes adaptées à la spécification des connaissances de domaine, les ontologies (Gruber, 1993), et d'autre part disposent d'une expressivité suffisante tout en conservant de bonnes propriétés calculatoires (cf. les réseaux sémantiques. (Lehman, 1992)). Une partie de ces travaux s'est concentrée sur les raisonnements par classification et a donné naissance aux logiques de description (Baader, et al., 2003). Une autre partie s'est intéressée à l'interrogation de bases de faits et a notamment conduit aux représentations à bases de graphes (cf. par exemple les graphes conceptuels (Sowa, 1984), (Chein, et al., 2009)).

L'avènement du web sémantique a conduit à promouvoir au rang de standard de représentation (RDF, RDFS, OWL) certains de ces travaux. Si l'apport de ces standards en termes d'interopérabilité sémantique est indéniable, la complexité des raisonnements résultant de l'utilisation des langages ontologiques dans un contexte d'interrogation de bases de connaissances ne permet pas le passage à l'échelle. Depuis, une dizaine d'années, les travaux se sont donc concentrés sur l'identification de sous-classes de ces standards prenant en compte les besoins en raisonnement et les caractéristiques des bases de connaissances actuelles : grandes bases de faits souvent distribuées, besoin en expressivité limité pour les ontologies, nécessité de traiter le problème de la déduction que ce soit pour l'interrogation, l'application de règles ou la vérification de contraintes. Cette communauté tend à se fédérer autour d'une famille de langages à bases de règles, baptisée *Datalog+/-* (Calì, et al., 2010), (Baget, et al., 2011), (Ortiz, et al., 2011). Le projet Qualinca exploitera cette famille de langages qui permet à la fois de représenter le contenu des catalogues documentaires, les connaissances ontologiques contenues dans les normes de catalogage utilisées par les SIB (CRM-CIDOC, FRBROO, FRAD, FRSAR, RDA…), mais aussi des règles de réconciliation, d'enrichissement ou de création d'autorité. L'un des objectifs de ce projet est d'étendre cette famille à la prise en compte de différentes catégories de « termes logiques », en s'appuyant sur les travaux de Levesque et Lakemeyer (Levesque, et al., 2000), afin de proposer une problématique de qualité des liens.

Réconciliation et fusion de données

Depuis l'article fondateur de Newcombe, Kennedy, Axford et James publié en 1959 dans la revue Science (Newcombe, et al., 1959), de nombreux travaux sur la réconciliation concernent les bases de données. Les problèmes d'identification d'entités dans les bases de données ont reçu des noms divers – « record linkage », « entity resolution », « reference resolution », « de-duplication », « object identification », etc. – et, depuis 1959, la plupart des méthodes de résolution proposées utilisent des techniques de classification (cf. les références compilées par Winkler (Winkler, 2006)). Comme le disent Smalheiser et Torvik (Smalheiser, et al., 2009), et même en se restreignant aux bases bibliographiques : « There is no single paradigmatic author name disambiguation task – each bibliographic database, each digital library, and each collection of publications has its own unique set of problems and issues. »

Ces dernières années, des approches à base de connaissances se sont développées. Ces approches exploitent des connaissances exprimées sous la forme d'un ensemble de règles de réconciliation (Doan, et al., 2003), (Dong, et al., 2005), (Bhattacharya, et al., 2007) ou déclarées dans l'ontologie (schéma) (Saïs, et al., 2009). L'approche LN2R développée par (Saïs, et al., 2009) est une approche qui combine deux méthodes de réconciliation de références, une méthode logique nommée L2R et une méthode numérique nommée N2R. Dans LN2R les connaissances déclarées dans l'ontologie (e.g., contraintes de clés ou disjonctions entre classes) sont traduites sous la forme d'un ensemble de règles d'inférence logiques de (non-) réconciliation et/ou exploitées dans un calcul de similarité numérique pour et de non-réconciliation entre références. Un raisonnement logique est appliqué pour inférer toutes les (non-) réconciliations sûres. La méthode N2R calcule des scores de similarité entre références en exploitant les contraintes de clés pour modéliser les influences entre similarités.

Les contraintes de clés, éventuellement incertaines, sont difficiles à définir par un expert alors qu'elles sont d'une importance cruciale. L'approche développée par (Sismanis, et al., 2006) permet de découvrir des clés composées dans une relation de base de données. Cependant, ce type d'approche n'est pas adapté aux données RDF incomplètes, multi-valuées et conformes à une ontologie. L'approche de découverte de contraintes de clés proposée par (Nikolov, et al., 2010) dans le contexte du Web de données permet d'apprendre des propriétés (inverses) fonctionnelles à partir de données réconciliées. Cependant, ces dernières ne sont pas souvent disponibles.

Une fois les références réconciliées se présente le problème de fusionner les références pour les représenter par une référence unique. Fusionner les références réconciliées a plusieurs intérêts : d'une part, réduire le nombre total de références, ce qui permet un stockage plus facile et une interrogation plus rapide de ces dernières ; d'autre part, la réponse renvoyée suite à une requête d'utilisateur est plus lisible. La plupart des approches de réconciliation s'intéressent exclusivement au problème de la détection de redondances et délèguent entièrement la tâche de fusion à l'utilisateur final. Dans (Papakonstantinou, et al., 1996), un langage à base de règles est utilisé par l'administrateur du système d'intégration pour définir différentes fonctions de fusion de références (e.g. privilégier les sources les plus fiables en cas conflits). La fusion peut ainsi être réalisée sans prendre en compte les valeurs provenant des autres sources, masquant de ce fait les conflits entre valeurs. Dans (Subrahmanian, et al., 1995), la fusion est également réalisée à l'aide de règles de fusion spécifiées par l'administrateur du système d'intégration. Dans (Bleiholder, et al., 2005), les auteurs proposent un nouvel opérateur *FUSE BY* utilisé dans des requêtes SQL qui prend comme arguments un ensemble de fonctions prédéfinies (e.g., vote, max, min) associées à des attributs impliqués dans la requête. Enfin, des travaux plus récents (Saïs, et al., 2008), (Saïs, et al., 2010) ont intégré d'autres préoccupations : prise en compte de plusieurs critères complémentaires (fiabilité des sources, fréquence d'apparition, homogénéité, etc.) et représentation de la variabilité des valeurs prises à l'aide d'un modèle d'incertitude.

Modèles de confiance

La plupart des travaux sur la confiance portent sur les systèmes multi-agents et les systèmes pair-à-pair ainsi que plus récemment sur les réseaux sociaux. Dans de tels

systèmes, fondés sur l'interaction entre entités (logiciels ou humains), la confiance et la réputation sont des critères essentiels à prendre en compte pour garantir la sécurité des échanges et des transactions. On peut distinguer les approches numériques où la confiance est modélisée comme une quantité (parfois une probabilité) qui évolue en fonction de retours d'expériences (feedback) positifs ou négatifs, des approches symboliques où la confiance est modélisée par des règles de bons comportement que les agents s'engagent à respecter pour être considérés dignes de confiance (c'est-à-dire fiables). On peut se référer à (Ramchurn, et al., 2004) pour une revue plus complète de ces différentes approches. Dans le contexte de ce projet, les approches dont on pourra s'inspirer sont celles qui se fondent sur la réputation de certaines sources documentaires connues pour être fiables (au moins pour certains types d'information), qu'il s'agira de combiner avec des approches à base de retour d'expérience permettant de faire évoluer la confiance en fonction de la qualité des résultats obtenus et de leur provenance. Comme point de départ, on pourra s'appuyer sur des travaux récents d'un des partenaires du projet (le LIG) sur la modélisation probabiliste de la confiance dans des systèmes pair-à-pair de partage de ressources (Nguyen, et al., 2008), (Atencia, et al., 2011). Il s'agira de les étendre à la propagation d'indices de confiance par règles (de réconciliation ou d'enrichissement). Ce problème est une instance du problème difficile de la combinaison d'approches logiques et probabilistes, qui a aussi été abordé par le LIG dans le cadre de l'alignement d'ontologies (Tournaire, et al., 2011).

2.4. OBJECTIFS ET CARACTERE AMBITIEUX/NOVATEUR DU PROJET

Les objectifs scientifiques généraux du projet peuvent se formuler de la manière suivante :

- Construire des bases de connaissances représentant les connaissances issues de bases documentaires ;
- Attaquer les problèmes d'identification/validation/réparation de liens entre entités individuelles dans ces bases de connaissances documentaires ;
- Attaquer les problèmes d'identification d'entités individuelles issues de différentes bases dans un objectif d'alignement/fusion de bases et/ou d'enrichissement d'une base avec des connaissances issues d'une autre base.

Il s'agit, plus précisément de :

- développer un cadre logique pour qualifier la qualité d'une base de connaissances documentaires vis-à-vis de l'identification des entités individuelles et des liens entre entités individuelles ;
- proposer des principes/méthodes/outils pour passer d'un certain niveau de qualité à un meilleur niveau. Ces méthodes s'appuieront sur les approches logiques et numériques utilisées en réconciliation et fusion de références qui, d'une part, seront adaptées aux nouvelles problématiques engendrées par ces bases de connaissances documentaires :
  - identification d'une autorité à lier à une notice de documents,
  - détection d'erreurs de liage,
  - création automatique d'autorités,
  - enrichissement d'autorités en particulier par exploitation du web de données,

et, d'autre part, seront étendues pour prendre en compte des connaissances incertaines :

- découvertes de clés d'autorités dans ces bases de connaissances,
- représentation de la confiance dans les connaissances,
- adaptation des méthodes numériques à ces nouvelles données,
- intégration des raisonnements probabilistes dans les méthodes logiques,
- proposition de différentes intégrations des approches numériques et logiques,
- étudier les conditions de passage à l'échelle (filtrage des données à traiter, algorithmes incrémentaux, etc).

Les verrous scientifiques à attaquer

➢ **La définition d'un modèle de qualité**

Nous aborderons les différents problèmes en utilisant un langage de représentation de connaissances permettant d'exprimer des connaissances « fines », sous la forme de contraintes et de règles (par exemple, pour éliminer les doublons ou pour réaliser des liages) et de faire des raisonnements correspondant à ceux des documentalistes.

Contrairement au cas des bases de données traditionnelles, ces problèmes sont à traiter dans un cadre de monde « ouvert » dans le sens où il faut, en plus de l'incomplétude intrinsèque de ce type de données, des règles permettant de créer de nouveaux individus (par exemple, tout document induit l'existence d'une œuvre à son origine).

De plus, les cadres classiques ayant des modèles en logique du premier ordre doivent être étendus pour prendre en compte les spécificités de ces problèmes en particulier la notion de "*surrogate*", certaines constantes du langage formel ne devant avoir qu'une interprétation, ainsi que l'incertitude et l'incomplétude intrinsèques à ce type de données. Nous nous appuierons en particulier sur les « standard names » introduits par Levesque et Lakemeyer (Levesque, et al., 2000).

➢ **La problématique cœur du liage**

Au cœur des problèmes de qualité on trouve la problématique du liage : étant donnée une entité individuelle importante au sein d'une notice (ou d'un ensemble de notices), il s'agit de retrouver dans un référentiel (appartenant ou pas à la base d'où est issue la notice) l'autorité qui lui correspond.

On peut distinguer 3 niveaux dans les mécanismes de liage :

- niveau 1 : comparaisons de données élémentaires (noms, dates, …) ;
- niveau 2 : utilisation de raisonnements pour se prononcer sur le liage (règles logiques, préférences, propagation d'indice de confiance,…) ;
- niveau 3 : la propagation de résultats concernant le liage de données d'une certaine complexité vers le liage de données plus complexes.

Si cette problématique n'est pas sans rappeler celle de la réconciliation de références, elle s'en distingue par le fait qu'elle ne s'applique pas sur des données structurées mais sur des connaissances dont le vocabulaire conceptuel (classes et relations) est fixé par une (voire plusieurs) ontologie. Ainsi les comparaisons des données élémentaires ne se font pas attribut

par attribut de même nature mais nécessite la mobilisation de connaissances codées dans les ontologies pour identifier les données élémentaires comparables. Par exemple, les vocabulaires prévus pour décrire une personne prévoient que l'on peut lui associer une appellation, un nom, un prénom, des initiales… Pour comparer deux personnes, on peut être amené à inférer certaines valeurs de propriétés à partir d'autres (e.g. une appellation à partir d'initiales et d'un nom).

De plus, les liens que l'on cherche à établir vont des notices documentaires qui contiennent des connaissances contextuelles sur les autorités - par exemple, la description d'un document, sur un sujet particulier, publié à une certaine époque par l'autorité personne que l'on cherche à lier - vers les notices d'autorités qui contiennent des connaissances (censées être) indépendantes de tout contexte sur les autorités. Du coup, peu d'informations présentes à la fois dans les notices de document et dans les notices d'autorité sont comparables. Face à cette spécificité des bases documentaires, nous envisageons de bâtir le liage sur la comparaison des connaissances contextuelles présentes dans la notice du document contenant une entité que l'on cherche à lier aux connaissances contextuelles de l'ensemble des notices de documents liées aux autorités candidates à ce liage.

Enfin, nous souhaitons exploiter des connaissances externes à la base documentaire sur laquelle la problématique de liage se pose. Ceci afin de permettre de compléter les connaissances hors-contexte sur une autorité et/ou de disposer d'un plus grand nombre de « d'instances contextuelles » de cette autorité, voire de disposer d'une notice d'autorité quand elle n'existe pas dans la base d'origine. Nous envisageons deux types de connaissances externes :

- l'accès à d'autres bases documentaires : le liage nécessite alors de résoudre le problème de l'interconnexion/alignement de ces bases que ce soit au niveau des autorités ou des documents ;
- l'accès au Web de Données : le liage se pose alors dans le contexte du « linked data » qui pose des problèmes de découverte de liens possibles, de validation de rapprochement et de sélection/importation de connaissances.

➢ **La représentation explicite et le calcul de la confiance en la qualité des liages**

A cause de l'incomplétude et/ou de l'incertitude des données, de nombreux liages, tout en étant probables, ne sont pas certains. Associer des indices de confiance aux liages détectés est essentiel pour aider les experts à valider certaines décisions de réconciliations, mais aussi pour en tirer des mesures de qualité sur les bases de connaissances ainsi enrichies et sur les réponses obtenues quand on les interroge. Le problème de la modélisation de la confiance est un problème difficile qui se pose de manière critique dans de nombreux domaines de l'Informatique. Le verrou principal sera de définir une sémantique rigoureuse aux indices de confiance et à leur propagation, qui soit interprétable par les humains et calculable facilement par la machine. Jusqu'à présent, deux grands types d'approches ont été explorées de façon indépendante pour modéliser et calculer la confiance (dans les réseaux pair-à-pair ou dans les systèmes multi-agents) : des approches logiques fondées sur des logiques modales appropriées (du type BDI, Belief Desire Intention) et des approches numériques fondées sur les probabilités. Le défi sera de combiner la logique et les probabilités pour

calculer et inférer des indices de confiance en intégrant, dans un cadre uniforme et fondé mathématiquement, des informations sur la provenance des données, des avis d'experts, des calculs de similarités sur des valeurs d'attributs, des contraintes sur le domaine et des règles de réconciliation et d'enrichissement.

Caractère ambitieux/novateur

Ce projet est ambitieux dans la mesure où il concerne des problèmes fondamentaux et difficiles actuels en SIB, avec des projets comme WorldCat ou VIAF, ainsi qu'en Informatique (« Semantic Data Integration »), avec des projets comme le « Linked Data » (le noyau du « Linked Data » consistant à établir des liens entre des données, i.e., des entités nommées par des URIs).

Les aspects novateurs de notre projet résident dans la définition d'un cadre formel permettant d'identifier les problèmes de qualité des liens, et dans la combinaison d'approches logiques et numériques pour résoudre ces problèmes, et ceci dans le contexte de grandes bases documentaires existantes. Gageons que la réussite de ce projet permettra donc tout à la fois d'offrir des techniques et outils de contrôle de qualité de catalogues pour la communauté des sciences de l'information et des bibliothèques et fournira un cadre théorique et des méthodes associées permettant de résoudre la problématique fondamentale de la génération de liens posées par le « linked data ».

Critères de réussite et d'évaluation

Les critères de réussite pour le projet Qualinca sont de plusieurs ordres.

En SIB, les résultats devraient permettre de :

- fournir des méthodes opérationnelles permettant d'apprécier la qualité de grands catalogues aussi bien en ce qui concerne le contenu des notices d'autorités, qu'en ce qui concerne les liens présents entre ressources du catalogue. Ces méthodes devront permettre la mise en place de plans d'amélioration de la qualité ;
- offrir des outils automatiques ou semi-automatiques d'amélioration de la qualité des catalogues : correction d'anomalies de catalogage, enrichissement de notices, alignement avec des standards. Les outils semi-automatiques seront également jugés en fonction de leur pourcentage de fausses alarmes, critère pouvant être rédhibitoire pour leur adoption ;
- offrir des outils facilement intégrables aux systèmes documentaires de l'entreprise et capables de traiter des catalogues importants (passage à l'échelle).

L'évaluation de la réussite du projet sur ces points se fera par des expérimentations sur des échantillons représentatifs de bases documentaires. Différents prototypes seront développés pour les contextes applicatifs de l'ABES et de l'Ina et des campagnes d'évaluations sur des données réelles seront menées par les professionnels du catalogage de ces organismes.

En Informatique, les résultats devraient permettre de disposer d'un cadre formel de diagnostic de qualité des « liens » dans les bases de connaissances et de méthodes d'amélioration de cette qualité. La réussite sur ce point se mesurera par la publication dans des résultats dans des conférences et journaux de premier plan.

3. PROGRAMME SCIENTIFIQUE ET TECHNIQUE, ORGANISATION DU PROJET

3.1. PROGRAMME SCIENTIFIQUE ET STRUCTURATION DU PROJET

Les objectifs de Qualinca nécessitent de se doter d'un cadre formel d'identification de la qualité des bases documentaires et de proposer des méthodes d'amélioration de cette qualité.

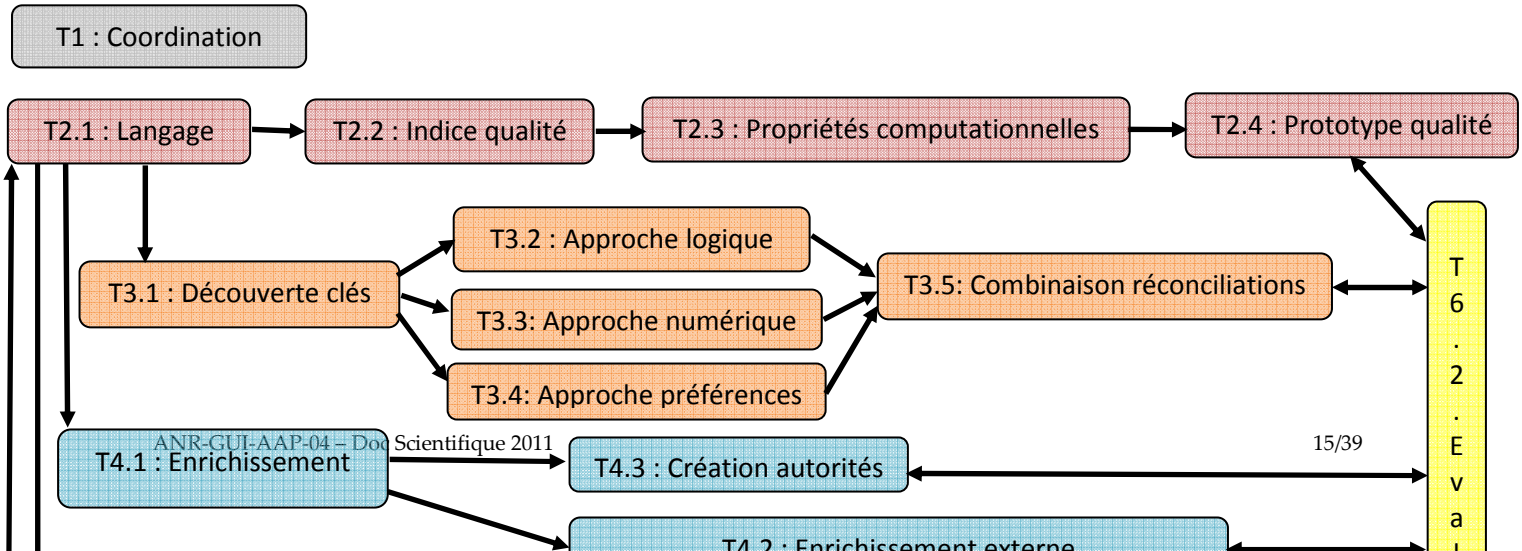
Nous définirons ce cadre formel en nous appuyant sur une sémantisation de ces bases dans des langages de représentation des connaissances. Il s'agit en particulier d'identifier les caractéristiques de ces bases (normes de description utilisées, notices documentaires, notices d'autorités, notion de « vedette », confiance en les données…), de proposer une représentation de leur contenu dans des primitives adaptées, et de formaliser des indicateurs de qualité de ces bases en terme de qualité des liens des notices documentaires vers les notices d'autorités (sont-ils corrects, complets ?), et de qualité des contenus des fiches d'autorités (sont-elles en « bijection » avec les entités qu'elles réfèrent ? Sont-elles suffisamment renseignées pour permettre l'identification de l'entité qu'elle réfère, ou permettre le respect d'une norme de contenu ?). La tâche 2 est dédiée à la définition de ce cadre.

Nous proposerons alors des méthodes d'amélioration de leur qualité. Deux axes seront étudiés : l'un sur l'identification des entités individuelles, pour traiter les problèmes de qualité des liens et de détection de doublons d'autorité, qui s'appuiera sur les travaux sur la réconciliation de références et les règles existentielles ; l'autre sur la description de ces entités individuelles, pour traiter les problèmes de désambiguïsation, et d'enrichissement du contenu des autorités, qui s'appuiera sur les travaux sur les systèmes à règles et sur les techniques de fusion de données. Les tâches 3 et 4 sont dédiées à ces deux axes d'amélioration.

Enfin, pour permettre l'interconnexion de bases documentaires de qualité diverse et exploiter les connaissances du web de données, nous proposerons un modèle de calcul de la confiance en une source de données et étendrons les méthodes d'amélioration proposées à ce modèle. La tâche 5 est dédiée à cette extension.

L'ensemble des techniques d'identification et d'amélioration de la qualité sera implanté et évalué sur les bases documentaires de l'ABES et l'Ina. C'est l'objectif de la tâche 6.

Le schéma ci-après présente les liens entre les différentes tâches et sous-tâches.



3.2. MANAGEMENT DU PROJET

Le LIRMM, porteur du projet, assurera la coordination globale du projet. Pour faciliter cette coordination, un coordinateur local est identifié pour chaque partenaire :

1. LIRMM : Michel Leclère
2. LRI : Fatiha Saïs
3. LIG : Marie-Christine Rousset
4. ABES : Yann Nicolas
5. INA : Patrick Courounet

Chaque coordinateur local a la responsabilité scientifique, technique et administrative de son groupe que ce soit au niveau de la qualité du travail scientifique, de l'organisation des réunions au sein de son entité, de la gestion des ressources locales ou des activités liées à la dissémination des résultats.

Le projet comprend 6 tâches qui seront conduites sous la responsabilité d'un des partenaires qui assurera son animation, la préparation biannuelle du rapport d'avancement de la tâche (activités menées, réunions, résultats, publications, réalisations logicielles) et la production du rapport final de la tâche. Pour chaque tâche, au moins deux réunions par an seront tenues. Dans la mesure du possible, ces réunions seront regroupées pour faciliter la communication entre les tâches. Afin d'assurer le suivi général du projet, une tâche de coordination est identifiée.

3.3. DESCRIPTION DES TRAVAUX PAR TACHE

3.3.1 TACHE 1 : COORDINATION

Objectifs

L'objectif de cette tâche est d'assurer le suivi général du projet, la communication entre les partenaires, la communication avec l'ANR et la dissémination des travaux.

Partenaires impliqués

LIRMM (responsable de la tâche), LIG, LRI, ABES, INA

Programme

En ce qui concerne l'organisation générale du projet, le coordinateur, Michel Leclère (LIRMM), s'assurera que l'avancement des travaux des tâches, en particulier en terme de production des livrables attendus et d'atteinte des objectifs annoncés, permet le respect de l'ordonnancement prévu des différentes tâches. Par ailleurs, il se chargera de coordonner la rédaction de rapports biannuels décrivant les avancées réalisées dans chaque tâche, les résultats obtenus, les livrables produits, les publications et prototypes réalisés et rendra compte des réunions tenues et de l'état d'avancement du projet vis-à-vis des objectifs initiaux. Il sera également responsable de la production du rapport final. Enfin, il s'occupera de la communication avec l'ANR.

Afin d'assurer une bonne compréhension de l'ensemble des attentes, travaux et résultats par l'ensemble des partenaires au projet, une première réunion plénière sera tenue au lancement du projet et au moins une réunion plénière par an sera organisée tout au long du projet. On peut d'ailleurs noter qu'une réunion préparatoire à l'appel à projet s'est déjà tenue en juillet 2011.

Enfin un site web permettant d'une part de présenter le projet et d'afficher ses résultats, et d'autre part d'assurer par l'intermédiaire d'un espace intranet les échanges de documents entre partenaires tout au long du projet sera mis en ligne. Cette dissémination sera renforcée par l'organisation d'ateliers dans différentes manifestations d'audience nationale et internationale (à titre d'exemple : journées de l'ABES, conférence d'extraction et gestion des connaissances (EGC), conférence Extended Semantic Web Conference (ESWC)…)

Livrables

L.1.1 Site web de dissémination des résultats du projet et d'échange de documents internes au projet.

L.1.2 Rapport biannuel d'avancement du projet

L.1.3 Rapport final du projet

Contributions

Le LIRMM pour le site web, la planification des réunions plénières et la communication avec l'ANR. Tous les partenaires pour la rédaction des rapports, la dissémination des résultats (organisation de workshop), et l'organisation locale des réunions.

3.3.2 TACHE 2 : FORMALISATION DE LA QUALITE D'UNE BASE DOCUMENTAIRE

Objectifs

L'objectif principal de cette tâche est de définir un modèle théorique permettant de formaliser les problèmes de qualité des bases documentaires. Il s'agit donc de définir un langage logique spécifique permettant de représenter :

- les différentes techniques de référencement des entités individuelles qu'elles soient directes par des identifiants standardisés (ex. ISBN d'une manifestation), par des identifiants non-sûrs (ex. lien vers une fiche d'autorité), par l'intermédiaire de données cibles explicites ou implicites de propriétés de l'entité (ex. nom et prénom d'un personne, sigle d'une institution), ou indirectes via les liens qu'elle entretient avec une autre entité (ex. l'auteur de « la princesse de Clèves ») ;
- les différentes caractéristiques prises par les connaissances sur les entités individuelles : connaissances pour l'identification / connaissances pour information, connaissances générales / connaissances contextuelles, sûre / incertaine, complète / incomplète, donnée / inférée, …

Dans un deuxième temps ce langage devra permettre d'exprimer le niveau de qualité d'une base par la caractérisation formelle des liens manquants, erronés ou douteux, des autorités potentiellement co-référentes, des incohérences au sein des notices, des incomplétudes des notices d'autorité vis-à-vis de leur rôle d'identification de l'entité qu'elles réfèrent.

Partenaires impliqués

LIRMM (responsable de la tâche), LIG, LRI, ABES, INA

Programme

Sous-tâche 2.1 : Dans un premier temps, nous définirons un langage L susceptible de formaliser les propriétés essentielles des identifiants (« surrogates ») d'entités importantes dans la représentation de connaissances issues de bases documentaires. L sera une variante de la logique du premier ordre, s'appuyant d'une part sur les « Standard Names » proposés par Levesque et Lakemeyer (Levesque, et al., 2000), et d'autre part sur des règles (Datalog+/-) telles que celles étudiées au LIRMM par GraphIK (Baget, et al., 2011). Il devra permettre, entre autres, l'expression de règles ayant en conclusion une égalité entre deux termes (l'hypothèse d'une telle règle étant alors une condition suffisante d'égalité de ces termes) et, toujours en conclusion, la présence de variables quantifiées existentiellement (une telle règle signifiant que certaines conditions --l'hypothèse de la règle— implique l'existence d'une entité).

Sous-tâche 2.2 : Nous nous intéresserons alors à la représentation dans ce langage des propriétés fondamentales concernant l'identification des autorités et les relations entre autorités dans les bases documentaires. Nous envisageons en particulier d'exploiter les travaux (Croitoru, et al., 2007) sur la génération d'expressions référentielles connus comme « GRE problem » pour caractériser l'existence de suffisamment de connaissances dans une notice d'autorité pour désambiguïser l'entité qu'elle représente parmi l'ensemble des entités de la base. Le GRE est une tâche cruciale en TAL qui modélise la façon dont les humains désignent verbalement des entités (par exemple, « Alexandre Dumas, l'auteur de la dame aux Camélias »). Appliquée aux bases de connaissances, cela revient étant donné une entité de la base à identifier un ensemble de caractéristiques permettant d'identifier cette entité et aucune autre.

Sous-tâche 2.3 : Nous étudierons les propriétés computationnelles de ce langage.

Sous-tâche 2.4 : Nous développerons un prototype de diagnostic qualité adapté au langage développé.

Livrables

L2.1 Etat de l'art sur les représentations logiques des notions d'identification et de création d'entités en mettant l'accent sur leurs propriétés computationnelles

L2.2 Définition du langage L et de la formalisation de la qualité d'une base documentaire

L2.3 Etude des propriétés computationnelles de ce langage

L2.4 Un prototype de diagnostic qualité

Contributions

Cette tâche est un travail d'équipe de l'ensemble des partenaires impliqués. Elle tirera profit des compétences en réconciliation de référence du LRI et du LIG, et de la longue expérience du LIRMM en représentation des connaissances. Les partenaires Ina et ABES seront mis à contribution pour leur expertise sur le contenu des catalogues documentaires, les normes utilisées dans les SIB, mais aussi sur les bonnes pratiques de catalogage. Les collaborations passées entre le LIRMM et l'ABES d'une part et le LIRMM et l'Ina d'autre part faciliteront l'appropriation de cette expertise par l'ensemble des partenaires.

Risques

Que certains aspects importants pour les démonstrateurs soient difficiles à modéliser dans le cadre logique proposé ou que la complexité algorithmique soit trop importante pour traiter les grandes bases cibles de ce projet. Dans ce cas nous nous attacherons à proposer une sémantique opérationnelle précise des algorithmes utilisés.

3.3.3 TACHE 3 : APPROCHES AUTOMATIQUES DE RECONCILIATION DE DONNEES DOCUMENTAIRES

Objectifs

L'objectif de cette tâche est de définir des méthodes automatiques de réconciliation des entités décrites dans les catalogues pour aborder le problème du dédoublonnage des notices d'autorité et celui de la génération de liens entre les notices documentaires et les notices d'autorité. Ces décisions de réconciliation seront utilisées dans un processus d'aide à la décision lors de la rédaction de nouvelles notices d'objets (liage avec les notices d'autorité), lors de l'intégration de listes d'autorité, ou lors du liage des entités des catalogues avec des entités décrites dans une ressource externe (par exemple issue du web de données). Ces méthodes exploiteront des contraintes sûres et incertaines du domaine mais également toute la richesse des descriptions des entités pour inférer des réconciliations sûres ou possibles. Elles devront également permettre la propagation de similarités de descriptions entre entités. Cette propagation permettra d'inférer de nouveaux liens mais aussi de détecter des liens erronés en cas d'incohérence. Compte tenu du volume important des bases documentaires,

ces méthodes devront être définies en vue de leur passage à l'échelle. Deux approches seront développées.

L'une s'appuiera sur des contraintes de clés composées pour modéliser les influences entre similarités de descriptions en distinguant les contraintes sûres des contraintes incertaines. Dans le cadre de données décrites dans des catalogues, certaines contraintes sont sûres et bien connues (par exemple, un même ISBN ne peut pas être associé à deux livres différents). En revanche, il est difficile pour un expert de définir toutes les contraintes et de leur associer un degré de confiance. Ainsi, développer une telle approche nécessite de découvrir automatiquement des contraintes de clés à partir des données documentaires. Cette phase de découverte devra tenir compte des caractéristiques des données, en particulier de la présence ou de l'absence de l'hypothèse du nom unique, de leur incomplétude et du niveau d'hétérogénéité syntaxique des descriptions. Ces clés seront alors exploitées par des approches logiques (pour les clés sûres) et numériques (pour les clés sûres et incertaines) de réconciliation. Les clés incertaines pourront également être exploitées par le modèle logique de prise en compte de la confiance développée dans la tâche 5.

L'autre sera fondée sur la modélisation de l'importance relative des différentes caractéristiques d'une entité, lors d'une décision de liage, par un modèle de préférences ordinales. Les critères de préférence seront exprimés sous la forme d'un ordre partiel sur l'importance relative des caractéristiques (par exemple, les nom et prénom d'une personne sont un critère plus important que son type d'activité). Considérant que chaque caractéristique permet de partitionner un ensemble de références de manière plus ou moins fine selon le degré de précision de la comparaison utilisée, nous étudierons comment exploiter ces préférences pour sélectionner une partition préférée (chaque classe de cette partition préférée représentant alors les réconciliations de références). Des travaux dans le domaine des réseaux sociaux, tel que « l'agrégation par choix social » pourront être exploités.

Partenaires impliqués

LRI (responsable), LIG, LIRMM, ABES, INA

Programme

Le plan de travail de cette tâche sera le suivant :

- **Sous-tâche 3.1 :** définition d'une approche efficace de découverte automatique de contraintes de clés composées sûres et incertaines à partir de données documentaires et de (non) réconciliations existantes en tenant compte des caractéristiques des données (UNA, incomplétude et hétérogénéité). La découverte pourra être effectuée au niveau d'une seule ou plusieurs sources tout en garantissant le passage à l'échelle.
- **Sous-tâche 3.2 :** définition d'une approche logique de réconciliation intégrant les contraintes de clés composées sûres (validées par un expert) et découvertes en sous-tâche 3.1.
- **Sous-tâche 3.3 :** définition d'une approche numérique de réconciliation en intégrant les contraintes de clés composées sûres (validées par un expert) et incertaines découvertes en sous-tâche 3.1.

- **Sous-tâche 3.4 :** définition d'une approche à base de préférences pour la réconciliation.
- **Sous-tâche 3.5 :** adaptation/combinaison des approches précédentes pour :
  - la réconciliation d'autorités redondantes ;
  - la génération de liens manquants des notices documentaires vers les autorités ;
  - détection et correction des liens erronés déjà existants ou inférés.

Livrables

L.3.1 Méthodes de découverte de clés composées sûres et incertaines.

L.3.2 Définition de l'approche numérique pour la réconciliation exploitant les contraintes de clés sûres et incertaines.

L.3.3 Définition de l'approche à base de préférences pour la réconciliation de références.

L.3.4 Adaptation / combinaison des approches précédentes pour la détection de doublons d'autorité et la génération / réparation des liens vers les notices d'autorité.

L3.5 Réalisation de prototypes correspondants pour la tâche 5.

Contributions

Le LRI contribuera à la découverte automatique des clés. Le LRI et le LIG à la définition d'une approche logique et numérique de réconciliation intégrant ces clés. Le LIRMM contribuera au développement d'une approche à base de préférences. Le LIRMM et le LRI contribueront à la combinaison et à l'adaptation des méthodes précédentes au problème de la génération et de la correction des liens entre notices documentaires et notices d'autorité. Ina et ABES participeront aux expertises nécessaires.

Risques

Le risque est minimisé par l'expérience des partenaires en réconciliation de données et par les travaux préliminaires sur la découverte de contraintes de clés dans le cadre de données relationnelles représentées en RDF. La validation des outils développés est garantie par la mise à disposition de données documentaires par deux partenaires différents.

3.3.4 TACHE 4 : ENRICHISSEMENT

Objectifs

L'objectif de cette tâche est d'étudier différentes voies d'enrichissement de notices d'autorités à partir de sources de natures très diverses et de définir des mécanismes d'enrichissement adaptés aux différentes caractéristiques de ces mêmes sources. Outre l'enrichissement de listes d'autorités existantes, cette tâche abordera également la problématique de création de listes d'autorités à partir de connaissances contextuelles internes et d'informations provenant de bases externes.

Cet enrichissement vise à terme à améliorer le processus d'identification des entités nommées et la mise en place de services innovants de recherche et de navigation basés sur une sémantisation accrue des ressources du catalogue.

Nous nous intéresserons à trois types de sources de données différentes : les informations structurées et non structurées présentes dans les catalogues eux-mêmes, les bases de données spécialisées dont la qualité est reconnue et les bases généralistes du web de données dont les connaissances sont issues d'un processus de « crowd-sourcing ». Chacune de ces sources possède des caractéristiques très différentes en termes de niveau de structuration des données, de qualité des informations stockées, d'interopérabilité des référentiels utilisés, de la nature contextuelle ou non contextuelle de l'information qualifiant les entités nommées. Il est donc nécessaire de mettre en place suivant les cas des méthodes d'extraction d'information, d'alignement de référentiels, de gestion de l'incertitude et de fusion de données prenant en compte des données contextuelles et/ou contradictoires.

Partenaires impliqués

INA (responsable), LIRMM, ABES, LRI, LIG

Programme

Cette tâche sera divisée en trois sous-tâches, la première et les deux dernières étant indépendantes. La deuxième sous-tâche visant notamment l'exploitation de connaissances du web de données pourra utiliser et adapter les mécanismes définis dans la tâche 5.

Sous-tâche 4.1 : Enrichissement par exploitation des connaissances internes

L'objectif principal de cette sous-tâche est d'exploiter les connaissances non structurées présentes dans les différents types de notices ainsi que les connaissances contextuelles du catalogue afin d'enrichir les notices d'autorités. Cet enrichissement doit notamment permettre d'informer les méthodes de découvertes de propriétés caractéristiques et les méthodes de réconciliation développées en tâche 3.

- Analyse et extraction des données non-structurées

Les notices d'autorité ainsi que les notices documentaires sont souvent porteuses d'informations non structurées qui concourent à l'identification de l'entité nommée. Qualité, type, nationalité, dates, périodes sont autant de types d'information utilisés dans ce processus d'identification. L'objectif de cette activité est de mettre au point des mécanismes automatiques d'extraction permettant de décorer à la fois notices d'autorités et notices documentaires avec respectivement des ensembles de connaissances contextuelles et non-contextuelles potentiellement pertinents pour une phase d'enrichissement. La réalisation de ces mécanismes sera basée à la fois sur l'analyse des données textuelles et sur l'exploitation des méthodologies et pratiques documentaires utilisées pour la production de ces données. Nous utiliserons donc des techniques de TAL que nous adapterons au contexte de l'enrichissement.

- Agrégation des connaissances contextuelles de l'autorité

Les connaissances contextuelles d'une autorité présentes dans les notices documentaires qui la réfèrent contribuent à l'identification de l'entité que l'autorité représente. L'objectif de cette activité est d'agréger certaines de ces connaissances contextuelles au niveau de l'autorité. Cette tâche d'agrégation conduira à la mise en œuvre d'une part des mécanismes de sélection automatique des connaissances à agréger et d'autre part des techniques

d'agrégation/fusion s'appuyant sur une logique possibiliste. Ces dernières permettront de proposer, pour chacune des valeurs intervenant dans un conflit, un degré de confiance obtenu par la combinaison de plusieurs critères complémentaires (syntaxiques, caractéristiques des sources de données, …). Cette approche pourra être exploitée de façon flexible lors de la présentation des données à l'utilisateur. Cette activité visera notamment à agréger les connaissances issues de la phase précédemment décrite d'extraction d'informations contextuelles.

- Fusion d'autorités

Le processus de dédoublonnage de la liste d'autorité doit s'accompagner d'une phase de fusion des informations provenant des doublons, phase qui doit conduire à une unique notice résultat. Nous élaborerons des mécanismes à base de règles de fusions qui permettront d'obtenir des notices d'autorités bien formées.

Sous-tâche 4.2 : Enrichissement par exploitation de connaissances externes

L'objectif est ici de permettre l'exploitation de connaissances issues d'autres bases documentaires et notamment du Web de données. Il est donc nécessaire de se doter de mécanismes d'enrichissement capables de prendre en compte les spécificités du Web des données : données incomplètes, contradictoires, alignement de référentiels. Dans ce contexte, se fonder sur des indices de confiance (cf. Sous-Tâche 5.3) est une aide précieuse pour prendre les bonnes décisions d'enrichissement. Cette activité couvre plusieurs points :

- Alignement de référentiels

Nous utiliserons des méthodes d'alignement d'ontologies à l'état de l'art afin de permettre des processus de réconciliation (tâche 3 et 5) entre autorités issues de bases différentes et pour faciliter la fusion de données pour l'enrichissement

- Règles d'enrichissement

Des patrons de règles de fusion seront élaborés afin de collecter et intégrer les connaissances potentiellement contradictoires provenant d'autorités externes (après réconciliation) dans les notices d'autorités du catalogue à enrichir. Ces règles prendront notamment en charge la collecte d'identifiants uniques garantissant l'interopérabilité future des données du catalogue.

Sous-tâche 4.3 : Création de listes d'autorités

L'objectif est suite à l'identification de l'existence implicite d'un nouveau type d'entité non référencée dans les catalogues (par exemple des œuvres musicales), de se doter de mécanismes de construction de sa notice d'autorité associée à la fois à partir des informations contextuelles disponibles sur cette entité mais aussi à partir d'informations provenant de bases externes spécialisées. Nous nous intéresserons également à l'utilisation de ces bases spécialisées pour améliorer le processus d'extraction d'autorités candidates en sus des connaissances documentaires.

Nous nous intéresserons à identifier pour certains types courants d'autorité (notamment œuvres et événements), les ressources externes disponibles et les contextes d'occurrences

dans les notices documentaires, et les technologies à utiliser en terme d'extraction et d'agrégation/fusion.

Les mécanismes mis en œuvre devront permettre de lier ces autorités avec d'autres autorités de même type ou de types différents afin de faciliter leur identification et de mettre en place des scénarios plus complexes de recherche d'information dans le catalogue ainsi enrichi.

Livrables

L.4.1 : Etat de l'art sur les techniques d'extraction utilisables et leur adaptation ai contexte des notices documentaires et notices d'autorités.

L.4.2 : Méthodologie d'agrégation/fusion et de gestion de doublons.

L.4.3 : Méthodologie de réalisation de patrons de règles d'enrichissement à partir de bases externes.

L.4.4 : Méthodologie de création de listes d'autorités.

L4.5 : Prototype de création/enrichissement/fusion de notices d'autorités.

L4.6 : Prototype étendu au contexte du web de données.

Contributions

L'Ina élaborera des mécanismes d'extraction d'informations à partir de données contextuelles et de sources externes (T4.1 et T4.3). L'ABES et l'INA définiront divers modèles de notices d'autorités en relation avec les types d'autorité ciblés, puis identifieront des bases spécialisées et généralistes du web de données susceptibles de satisfaire aux besoins d'enrichissements. L'Ina, le LIRMM et le LRI proposeront des mécanismes d'agrégation de connaissances contextuelles et de fusion de connaissances. Le LIRMM, l'ABES et l'Ina proposeront un mécanisme de création automatique d'autorité basé sur la combinaison de règles de réconciliation et de règles d'inférences de nouveaux individus.

Risques

Chacune des activités de cette tâche est potentiellement génératrice de types de risque différents. L'activité d'extraction d'information est déterminante pour initier les activités de réconciliation et d'enrichissement dans le cas de listes d'autorités très peu structurées. Le risque est minimisé par la maturité des technologies disponibles en TAL et concomitamment par l'existence de pratiques documentaires de catalogage utilisables pour le paramétrage de ces technologies. Les risques liés à l'exploitation de bases externes reposent en partie sur le problème d'alignement de référentiel, ce risque sera minimisé par l'utilisation de technologies disponibles, notamment au LIG et par l'utilisation d'alignements manuels lorsqu'il s'agit des modèles de descriptions des autorités (un certain nombre d'alignements a été réalisés par l'Ina dans le cadre du projet ASSETS pour DBPedia,Freebase et Geonames).

Les risques liés à l'agrégation de connaissances et à la fusion d'autorités est minimisé par l'existence de d'un nombre suffisant de notices d'autorités redondantes et par l'exploitation de données provenant de deux partenaires et du LOD.

3.3.5 TACHE 5 : MODELE ET CALCUL DE LA CONFIANCE EN LA QUALITE DES LIAGES ET DES ENRICHISSEMENTS

Objectifs

L'objectif de cette tâche est de formaliser et de combiner dans un cadre rigoureux et uniforme différentes connaissances permettant d'inférer des décisions de réconciliations ou non réconciliations avec des indices de confiance fondés sur la provenance des données. Il est en effet essentiel de pouvoir prendre en compte que certains catalogues contiennent des informations de référence et d'autres des informations qui tout en étant riches et potentiellement utiles ne sont pas totalement fiables. Il s'agira d'étendre et d'adapter les méthodes logiques sûres mais incomplètes, fondées sur l'exploitation de contraintes fortes du domaine permettant de déduire des décisions sûres de réconciliation ou de non réconciliation entre références. Par exemple, deux descriptions de livres faisant tous deux référence au même numéro ISBN correspondent de façon certaine au même livre, car l'ISBN est un identifiant unique (une clef) dans la description d'un livre dans tous les catalogues du monde. Mais comment savoir si deux descriptions de personnes font référence au même individu ? Avoir le même nom n'est pas suffisant pour être sûr qu'il s'agit de la même personne mais il peut être utile d'inférer qu'on est presque sûr qu'il s'agit de la même personne, en associant à cette réconciliation probable mais non certaine un indice de confiance. Cet indice de confiance peut être défini à différents niveaux de granularité. Il peut-être calculé, inféré ou fourni par un expert, et peut dépendre de la provenance des données sur lesquelles le calcul de la confiance se fait. Pour pouvoir être comparés et propagés de façon pertinente dans d'autres décisions de réconciliation ou non réconciliation, il est essentiel de pouvoir donner une sémantique rigoureuse à ces indices de confiance et à leur propagation via des règles logiques. Nous explorerons pour cela un cadre combinant la logique et les probabilités. Interpréter les indices de confiance comme des probabilités que deux références identifient la même entité est un plus pour l'aide à la décision mais contraint la façon dont ils peuvent être combinés par des connecteurs logiques et propagés par des règles logiques.

Partenaires impliqués

LIG (responsable), LIRMM, LRI, INA, ABES

Programme

Le plan de travail de cette tâche sera le suivant.

- **Sous-tâche 5.1** : Calcul et propagation d'indices de confiance par des règles logiques extraites du schéma de la base documentaire. Certains indices de confiance sur la réconciliation entre valeurs de propriétés peuvent être calculés, par exemple par des mesures de similarité entre chaînes de caractères. Ils peuvent très facilement être interprétés comme la probabilité que les deux chaînes de caractères identifient la même valeur pour une certaine propriété (par exemple, le nom d'auteur). Le problème qui sera étudié dans cette sous-tâche est celui de l'inférence de nouveaux indices de confiance par propagation et agrégation d'indices existants (donnés, calculés ou inférés) à partir de

règles logiques de réconciliation ou non réconciliations de références. Il s'agit d'un cas particulier d'inférence probabiliste où les faits sont incertains mais les règles sont certaines.

- **Sous-tâche 5.2** : Etude et mise en œuvre d'inférences probabilistes pour propager des indices de confiance par des règles logiques non sûres. Pour cela, le premier problème sera de modéliser les règles ou les contraintes non sûres sur les réconciliations ou les non réconciliations par des règles associées à des probabilités. Le second problème consistera à étendre le cadre d'inférence probabiliste mis en œuvre dans la sous-tâche 5.1 au cas où les règles sont également incertaines.
- **Sous-tâche 5.3** : Extension de cette modélisation en logique probabiliste aux règles d'enrichissement. Cette sous-tâche est complémentaire de la sous-tâche 4.2 visant l'enrichissement par des données externes, en particulier provenant du Web. Associer des indices de confiance à l'ajout de certaines connaissances permettra de guider l'aide à la décision de fusion d'informations.

Livrables

L.5.1 : Prototype intégrant le calcul initial d'indices de confiance, leur propagation et leur agrégation en fonction de règles logiques de réconciliation et de non réconciliation données en entrée.

L.5.2 : Modèle probabiliste de règles incertaines de réconciliation et de non réconciliation et extension du prototype précédent à l'inférence probabiliste sur des règles incertaines.

L.5.3 : Mise en œuvre et expérimentation de règles d'enrichissement à base d'indices de confiance.

Contributions

Le LIG contribuera sur la modélisation de la confiance. Le LRI et le LIRMM fourniront les règles d'enrichissement et de réconciliation. Les 3 partenaires définiront ensemble du modèle d'inférence probabiliste à mettre en œuvre pour la propagation et l'agrégation des indices de confiance et leur utilisation pour la fusion d'informations provenant de différentes sources. L'Ina et l'ABES contribueront à la validation du modèle dans les applications documentaires du projet.

Risques

Le risque est balisé par les travaux préliminaires du LIG sur le calcul d'indices de confiance pour l'alignement d'ontologies et l'interrogation de sources de données dans des systèmes distribués de gestion de données.

3.3.6 TACHE 6 DEMONSTRATEURS ET EVALUATION

Objectifs

L'objectif de cet ensemble de tâches est double :

- tester et évaluer les méthodes de diagnostic qualité, de réconciliation et d'enrichissement ;

- intégrer ces méthodes dans un contexte métier, au moyen de démonstrateurs, et évaluer cette intégration.

Tester les méthodes de diagnostic qualité, de réconciliation et d'enrichissement signifie les appliquer à des corpus de métadonnées ABES et Ina et obtenir en sortie des rapports d'analyse structurés qui permettent de connaître, de visualiser et d'exploiter une série d'indicateurs quantitatifs sur le résultat des diagnostics, des réconciliations et des enrichissements.

Il s'agit ensuite d'évaluer ces résultats par différents moyens, notamment en les confrontant à des pratiques manuelles réalisées en parallèle pour les besoins de l'évaluation. Cette évaluation doit permettre d'améliorer les méthodes de manière itérative (cycle test/évaluation/test/…), de mesurer leurs limitations intrinsèques (qui ne pourraient être levées que par d'autres approches théoriques), de catégoriser les profils de données auxquelles elles peuvent s'appliquer de manière efficace et enfin de spécifier les conditions dans lesquelles ces méthodes peuvent être intégrées dans un contexte métier opérationnel.

Cette intégration dans un contexte métier fait l'objet du second volet de test. Si les rapports d'analyse contiennent des indicateurs sur des opérations visant à l'amélioration de la qualité, à la réconciliation ou à l'enrichissement de données, l'acte même de correction effective, de réconciliation effective ou d'enrichissement effectif, ayant pour effet la modification du corpus de données, relève d'un autre niveau et exige d'être l'objet de tests et d'évaluation dédiées. Ces tests seront menés à travers des démonstrateurs, inscrits dans les contextes ABES et Ina.

Il s'agira de développer un (ou plusieurs) prototype(s) simulant, en situation proche des conditions de travail réelles, des opérations effectives de correction, de réconciliation et d'enrichissement. Ces opérations effectives peuvent être supervisées ou entièrement automatiques, ce qui suppose de déterminer en amont comment on extrait du rapport d'analyse des critères de choix automatique. Chaque prototype fera l'objet d'une évaluation.

Les sous-tâches de tests et de prototypage seront menées en parallèle par l'ABES et l'Ina. Cependant, ces sous-tâches seront précédées d'une sous-tâche transversale, dont l'objectif sera méthodologique. Il s'agira d'identifier des critères de sélection des corpus de test, de déterminer la structure et le contenu des rapports d'analyse et de choisir les protocoles d'évaluation.

Partenaires impliqués

ABES (responsable), INA (responsable d'une sous-tâche dédiée), LIRMM, LRI, LIG

Programme

Sous-tâche 6.1. Méthodologie

- Identifier les critères de sélection des corpus de métadonnées de test.
- Sélectionner les corpus.
- Déterminer la structure et le contenu des rapports d'analyse.
- Choisir les protocoles d'évaluation.

Sous-tâche 6.2. Test et évaluation des méthodes de diagnostic qualité, de réconciliation et d'enrichissement des données sur les corpus ABES et Ina

- Appliquer les méthodes issues des tâches 2, 3, 4 et 5 sur les données sélectionnées ;
- Evaluer la pertinence et la qualité des méthodes à travers ces tests.

Sous-tâche 6.3. Intégration de ces méthodes dans un contexte métier ABES, au moyen de démonstrateurs, et évaluation de cette intégration

Il s'agira de développer des prototypes utilisant les méthodes précédentes comme outils d'aide à la décision et d'évaluer l'utilisation de ces prototypes.

- Démonstrateur « Validation des liens internes au catalogue Sudoc » permettant de :
  - parcourir les données actuelles du Sudoc pour identifier des erreurs de liaison entre notices bibliographiques et notices d'autorité ;
  - corriger les erreurs ou suggérer des corrections à un opérateur.
- Démonstrateur « Identifier les notices d'œuvres implicites » (FRBR) permettant de :
  - analyser les données actuelles du Sudoc et éventuellement d'autres catalogues ou bases afin d'identifier les œuvres correspondant aux éditions (« manifestations » selon le modèle FRBR) ;
  - générer de nouvelles notices d'œuvres liées aux notices d'éditions, ou suggérer à un opérateur de les créer.
- Démonstrateur "Interconnecter le Sudoc aux Licences Nationales" permettant de :
  - utiliser le corpus de métadonnées "Licences Nationales" (LN), ressources électroniques achetées par l'ABES pour la communauté nationale ;
  - analyser ce corpus LN pour établir des liens aux notices d'autorité existantes (personnes ou œuvres) ou création de notices d'œuvres ;
  - générer automatiquement ou suggérer de nouveaux liens ou autorités.

Sous-tâche 6.4. Intégration de ces méthodes dans un contexte métier Ina, au moyen de démonstrateurs, et évaluation de cette intégration

Il s'agira de développer des prototypes utilisant les méthodes précédentes comme outils d'aide à la décision.

- Démonstrateur « Interface pour un opérateur de gestion des listes d'autorités ». Cette Interface technique simple permet une visualisation / validation des données basée sur l'analyse des alertes posées dans les rapports des démonstrateurs précédents.
- Intégration au sein d'un prototype d'annotation manuelle existant de fonctionnalités permettant à un documentaliste d'être piloté dans le choix des entités nommées à utiliser.

Livrables

L.6.1 Rapport sur les critères de sélection des corpus.

L.6.2 « Dump » des corpus de test choisis.

L.6.3 Dossier de présentation des protocoles d'évaluation : contenu et format de sortie des prototypes de diagnostic, modalités d'évaluation.

L.6.4 Résultats d'application des méthodes sur les corpus (L6.4.1 ABES et L6.4.2 Ina).

L.6.5 Rapport d'évaluations des méthodes sur les corpus (L6.5.1 ABES et L6.5.2 Ina).

L.6.6 Cahier des charges des démonstrateurs : fonctionnalités attendues, modalités de prise en compte des prototypes de diagnostic, mode opératoire d'utilisation et d'évaluation du démonstrateur (L.6.6.1 ABES et L.6.6.2 Ina).

L.6.7 Démonstrateurs (L.6.7.1 ABES et L.6.7.2 Ina).

L.6.8 Résultats de l'évaluation des démonstrateurs.

Contributions

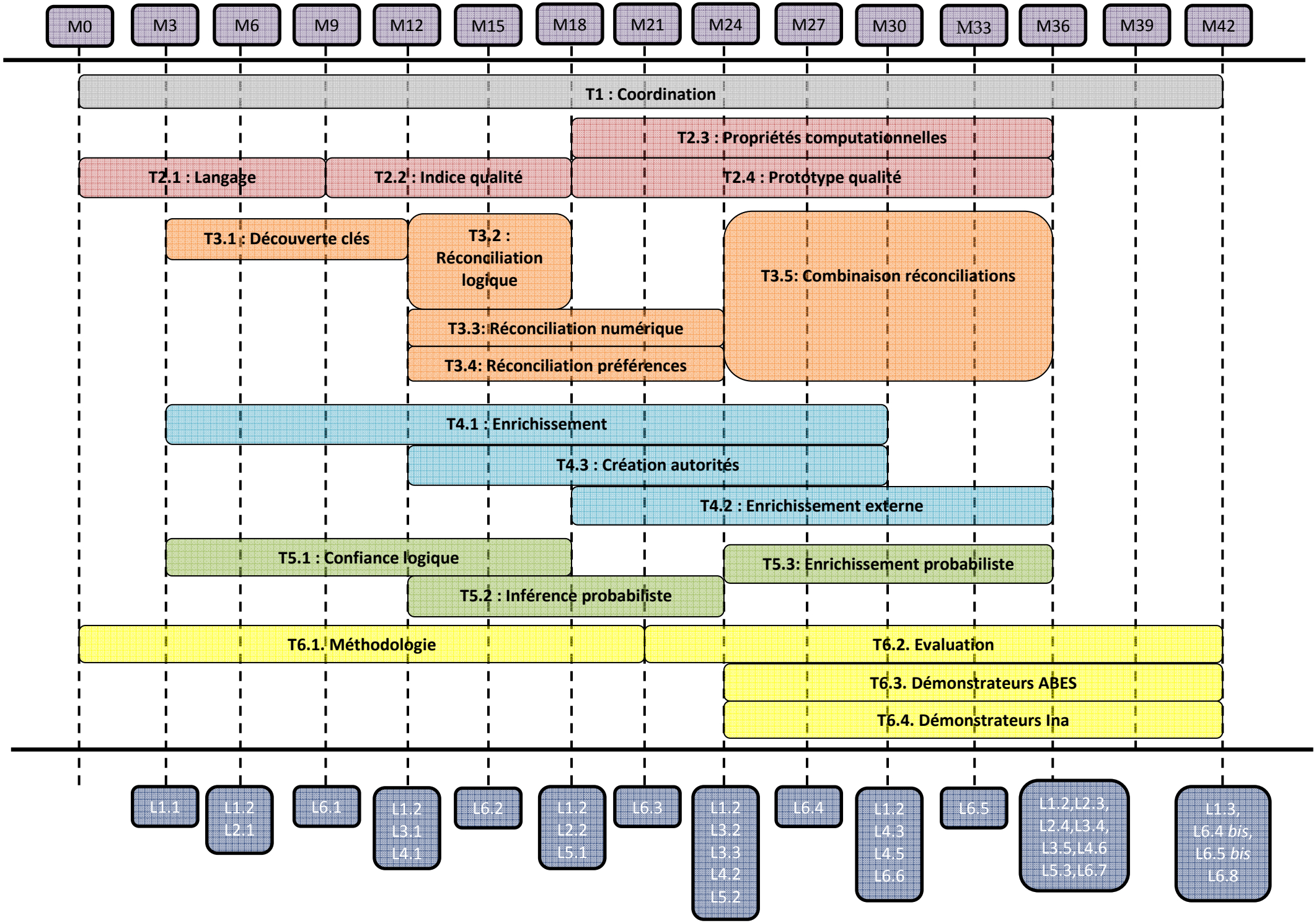
ABES, Ina, LIRMM et LIG sur la sous-tâche 6.1 et la sous-tâche 6.2. ABES et Ina s'occupent séparément de l'évaluation des méthodes sur leurs données respectives. ABES sur la sous-tâche 6.3. Ina sur la sous-tâche 6.4

Risques

Les risques concernent essentiellement la sélection des corpus qui peuvent, rétrospectivement, comportés des biais. De plus, il faudra organiser différentes étapes d'évaluations afin de permettre d'adapter, améliorer les méthodes et de les réévaluer.

3.4. CALENDRIER DES TACHES, LIVRABLES ET JALONS

Le diagramme de Gantt ci-après situe dans la durée du projet les différentes tâches et sous-tâches et les dates de rendu des livrables (MO, mois 0 correspond à la date de démarrage du projet). Les contenus de chaque livrable sont décrits dans les tâches afférentes. Concernant la mise au point des différentes méthodes automatiques de diagnostic qualité, de réconciliation et d'enrichissement, il est prévu plusieurs étapes durant l'évaluation (du mois 24 au mois 30, du mois 30 au mois 36, et du mois 36 au mois 42) afin de permettre d'exploiter les résultats d'évaluation pour améliorer ces méthodes.


M0
M3
M6
M9
M12
M15
M18
M21
M24
M27
M30
M33
M36
M39
M42
T1 : Coordination
T2.3 : Propriétés computationnelles
T2.1 : Langage
T2.2 : Indice qualité
T2.4 : Prototype qualité
T3.1 : Découverte clés
T3.2 : Réconciliation logique
T3.5: Combinaison réconciliations
T3.3: Réconciliation numérique
T3.4: Réconciliation préférences
T4.1 : Enrichissement
T4.3 : Création autorités
T4.2 : Enrichissement externe
T5.1 : Confiance logique
T5.3: Enrichissement probabiliste
T5.2 : Inférence probabiliste
T6.1. Méthodologie
T6.2. Evaluation
T6.3. Démonstrateurs ABES
T6.4. Démonstrateurs Ina
L1.1
L1.2 L2.1
L6.1
L1.2 L3.1 L4.1
L6.2
L1.2 L2.2 L5.1
L6.3
L1.2 L3.2 L3.3 L4.2 L5.2
L6.4
L1.2 L4.3 L4.5 L6.6
L6.5
L1.2,L2.3, L2.4,L3.4, L3.5,L4.6 L5.3,L6.7
L1.3, L6.4 bis, L6.5 bis L6.8

4. STRATEGIE DE VALORISATION, DE PROTECTION ET D'EXPLOITATION DES RESULTATS

Valorisation des résultats de Qualinca par l'ABES et l'Ina

Les problématiques et les débouchés de ce projet sont au cœur de la stratégie et des priorités du prochain projet d'établissement de l'ABES, à savoir : ouverture et interconnexion des données à l'échelle du Web de données, construction d'un hub des données et des métadonnées de l'Information Scientifique et Technique au service des universités et des institutions de recherche, restructuration des catalogues bibliographiques selon de nouveaux modèles (FRBR, RDA, …). L'investissement de l'ABES dans ce projet est donc en cohérence avec ses priorités. Les résultats du projet seront intégrés au sein des applications professionnelles de l'ABES, qui en a la maîtrise (développements en interne).

Les gains espérés par l'ABES sont triples :

- améliorer la productivité du catalogage, activité coûteuse en ressources humaines mobilisées dans les bibliothèques universitaires et les centres de documentation ;
- améliorer l'efficacité et la fiabilité des traitements automatiques sur les données, au sein d'un seul catalogue ou dans le contexte de l'interconnexion entre plusieurs catalogues ;
- fournir des données plus riches et plus fiables permettant des réutilisations plus variées et plus puissantes, avec pour objectif ultime la valorisation des ressources documentaires produites ou acquises par l'enseignement supérieur français.

L'Ina a initié depuis près de dix ans une série de chantiers de modernisation de l'ensemble de son système technique afin de s'adapter à l'accroissement constants des données qu'elle a à traiter et afin de diversifier les différentes voies de valorisation à destination de ses publics.

Les thématiques les plus importantes pour l'institut concernent le développement de solutions permettant l'accélération de la documentation et notamment la gestion des imports, l'évolution contrôlée de ses ressources d'indexation, l'amélioration des différents moteurs de recherche de contenus mis à la disposition de ses clients sur Internet et intranet, la présentation de son contenu en liaison avec d'autres ressources du Web et l'accélération de ses capacités de libération des droits. Dans ce cadre, assurer la qualité et l'interopérabilité de ses métadonnées devient un enjeu majeur pour l'institut. Le projet permettra à l'Ina d'évaluer des solutions automatiques et semi-automatiques d'enrichissement et d'amélioration de la qualité de ses catalogues.

Les gains attendus sont multiples :

- améliorer les processus coûteux d'imports de métadonnées provenant de différentes chaines ;
- résoudre les problèmes dus à l'absence de gestion des homonymies dans le système actuel ;

- rendre interopérable les systèmes de gestion documentaire et de reversement des droits de l'entreprise et assurer la prise en compte des standards d'identification d'œuvres et de personnes, ceci dans le but d'accélérer les processus de libération des droits.

Etant donné l'importance de ce projet pour ces deux partenaires, un accord de consortium sera élaboré, comme demandé, dans un délai de un an.

Plus généralement, les outils développés dans le cadre de Qualinca devraient pouvoir être utilisés dans d'autres domaines pour lesquels disposer d'une base d'autorités commune à un ensemble de bases d'informations est important, comme par exemple en économie (entreprises/employés/actionnaires …). En effet, les modèles et outils de ce projet seront suffisamment génériques pour permettre d'améliorer le processus de création de nouvelles métadonnées ainsi que l'exploitation de métadonnées existantes dans le cadre de l'interopérabilité, dirigée par des référentiels, de grands catalogues. Nous espérons ainsi que les retombées scientifiques et techniques de Qualinca seront significatives dans le domaine du web sémantique, et plus particulièrement dans le domaine du « Linked Data ».

Publications, Dissémination

Les résultats scientifiques de Qualinca pourront être présentés dans de nombreuses revues et conférences nationales et internationales dans la mesure où nos travaux concernent aussi bien la science de l'information et des bibliothèques que l'informatique (en particulier le web sémantique et l'intelligence artificielle).

Ils seront également présentés dans des ateliers spécifiques, par exemple dans le cadre des Journées de l'ABES ou du congrès EGC (Extraction et Gestion de Connaissances).

5. DESCRIPTION DU PARTENARIAT

5.1. DESCRIPTION, ADEQUATION ET COMPLEMENTARITE DES PARTENAIRES

5.1.1 LIRMM / INRIA

Le **LIRMM** (Laboratoire d'Informatique, Robotique et Microélectronique de Montpellier http://www.lirmm/fr), est une unité mixte de recherche de l'Université Montpellier 2 et du CNRS (UMR 5506) qui regroupe tous les chercheurs en Informatique, Robotique et Microélectronique de la région Languedoc-Roussillon. Lors de la dernière évaluation AERES, le LIRMM a été noté A+. Les 360 chercheurs (doctorants compris) du LIRMM sont répartis dans 17 équipes-projets.

GraphIK (Graphs for Inferences on Knowledge, http://www.lirmm.fr/graphik) est une équipe projet commune avec l'INRIA et l'INRA dont les travaux se situent dans le domaine de la Représentation des Connaissances et Raisonnements. Elle possède une forte expertise sur les langages à la fois graphiques et logiques comme : RDFS le langage standard du web sémantique, les règles et contraintes (TGD) des bases de données, les graphes conceptuels, certains fragments de la logique du premier ordre et des logiques de description.

L'un des problèmes centraux auxquels l'équipe s'intéresse est l'interrogation de grandes bases de connaissances en présence d'une ontologie. GraphIK a développé une approche reconnue internationalement basée sur l'utilisation de règles existentielles (cf. (Baget, et al., 2009), (Baget, et al., 2011), (Baget, et al., 2011))

A côté de ses travaux théoriques, GraphIK poursuit une importante action de développement d'une suite logicielle dédiée à l'acquisition, l'exploitation et la maintenance de grandes bases de connaissances (cf. Cogui : http://www.lirmm.fr/cogui et Cogitant : http://cogitant.sourceforge.net/).

Ses travaux théoriques et ses outils logiciels ont été développés et mis en œuvre dans de nombreux projets concernant, en particulier, des systèmes basés sur des annotations sémantiques construites à partir d'une ontologie (en collaboration avec l'Ina : Opales, Saphir, et Logos, et aussi le projet Eiffel), ou des données bibliographiques (en collaboration avec l'ABES : Mogador et SudocAd), financés par l'ANR, l'union européenne ou le CNRS. Les notices documentaires ou d'autorités peuvent être considérées comme de telles annotations sémantiques.

5.1.2 LRI / INRIA

Le **LRI** (Laboratoire de Recherche en Informatique) est une unité mixte de recherche (UMR8623) de l'Université Paris-Sud et du CNRS. Le laboratoire accueille plus de 260 personnes dont environ 115 permanents et 110 doctorants, organisés en 12 équipes de recherche. Lors de la dernière évaluation AERES, le LRI a été noté A. Huit des 12 équipes de recherche sont communes avec l'INRIA Saclay - Ile-de-France qui est ainsi le partenaire privilégié du laboratoire.

L'équipe **IASI/Leo** commune au LRI et à l'INRIA Saclay Île-de-France regroupe des compétences reconnues au niveau national et international en intelligence artificielle et en bases de données. Elle travaille sur la gestion d'informations sous toutes ses formes (données RDF, documents XML, services) pouvant être sémantiquement hétérogènes et distribuées sur le Web. Certains travaux de l'équipe portent en particulier sur l'intégration sémantique de données hétérogènes, la réconciliation de données et l'alignement d'ontologies, dans le cadre du Web sémantique. Les travaux sur la réconciliation de données et sur l'alignement d'ontologies ont été validés par la publication de plusieurs articles dont des conférences et revues internationales reconnues dans le domaine du Web sémantique ou de l'intelligence artificielle (AAAI 2007, JoDS 2009, EKAW 2010, DEXA 2011).

Les approches développées ont été confrontées à des données réelles dans le cadre de campagnes d'évaluation internationales (OAEI 2007 à 2010) et lors de la participation à des projets académiques et industriels dont les plus récents sont : PICSEL (avec France Telecom R&D) sur l'intégration sémantique de données, ANR GeOnto sur l'interopérabilité de données relatives à l'information géographique, WebContent (projet RNTL) sur la construction d'une plate-forme générique et flexible de gestion de contenus intégrant les technologies du Web Sémantique et l'activité Data Bridges menée dans le cadre du EIT-KIC ICT Labs (thème Digital Cities) sur l'intégration sémantique de données pour les villes numériques. Dans ce projet l'objectif de ce partenaire est de partager son expérience en

réconciliation de données pour développer des outils adaptés au problème de liage de données documentaires dans les catalogues.

5.1.3 LIG

Le **LIG** (Laboratoire d'Informatique de Grenoble, http://www.liglab.fr/) est une unité de mixte recherche entre le CNRS, les universités Joseph Fourier, Grenoble INP et Pierre-Mendès-France de Grenoble, et regroupant plus de 400 chercheurs et enseignants chercheurs, doctorants et post-doctorants, ITA/IATOS. Lors de la dernière évaluation AERES, le LIG a été noté A+.

L'équipe **HADAS** (Heterogeneous Autonomous Distributed Database Systems) est une des 23 équipes du LIG. Sa thématique de recherche est la gestion de données distribuées et hétérogènes, avec un focus sur les données du Web et leur intégration dans le cadre du Web sémantique. Elle a participé ou participe à plusieurs projets sur la composition et la continuité de services pour le traitement de données (projet ANR CONTINUUM), sur les systèmes de gestion de données pair-à-pair (projet ANR Dataring), ainsi que sur des algorithmes de fouille de données capables d'exploiter le parallélisme des machines multi-cœurs (projet SocTrace du pôle de compétitivité Minalogig).

Le personnel permanent impliqué dans le projet sera Marie-Christine Rousset qui a une expertise reconnue en représentation des connaissances pour l'intégration d'informations, les systèmes pair-à-pair de gestion de données et le web sémantique. Un doctorant et un post-doctorant seront impliqués dans le projet sur les aspects liés à la modélisation de la confiance : Mustaf Al Bakri, allocataire MENRT, et Manual Atencia, postdoc UJF.

5.1.4 ABES

L'**ABES** (Agence Bibliographique de l'Enseignement Supérieur http://www.abes.fr) est un établissement public administratif (EPA), placé sous la tutelle du Ministère de l'Enseignement Supérieur. Elle a été créée en 1994 pour construire le Sudoc, catalogue collectif du Supérieur et animer un réseau de catalogage partagé et de prêt entre bibliothèques. Créé en 2001, le Sudoc est aujourd'hui un outil incontournable, de grande envergure : 2000 bibliothèques travaillent au sein de son réseau ; 10 millions de notices bibliographiques décrivent 32 millions de documents, en s'appuyant sur 2 millions de notices d'autorité.

Dès 2002, l'ABES s'est diversifiée en proposant d'autres services comme :

- STAR : application de dépôt national des thèses électroniques (2006) ;
- Calames : catalogue collectif des archives et des manuscrits (2007) ;
- IdRef : application Web dédiée aux autorités Sudoc, ouverte à toute autre application documentaire Web, en lecture et en écriture (création de notices) (2010) ;
- theses.fr, portail des thèses françaises (2011).

Au cours de ces évolutions, l'ABES n'a cessé d'augmenter le nombre de bases de métadonnées nationales placées sous sa responsabilité, en veillant à respecter deux exigences cruciales :

- leur mise en cohérence, notamment au moyen des autorités Sudoc, pivot commun à ces différentes bases ;
- leur ouverture sur le Web et, en particulier, le Web de données : en 2011, toutes les données de l'ABES sont exposées selon les principes des *Linked data*.

Ce triple processus (diversification des données, interconnexion des données, ouverture des données) va continuer, avec la montée en charge d'une nouvelle mission : l'acquisition des ressources électroniques au nom des établissements (groupements de commande) ou de l'ensemble de la communauté (licences nationales). Cette mission se concrétisera par l'abondance de nouvelles métadonnées, de plus en plus variées en qualité et quantité, que l'ABES entend intégrer dans un hub de données au service de l'enseignement supérieur.

C'est dans cette perspective stratégique centrale et en continuité avec son investissement dans les technologies sémantiques (ouverture au web de données, projet SudocAd avec le LIRMM et ADONIS) que l'ABES s'engage dans le projet Qualinca.

Ce projet étant au cœur des enjeux actuels et à venir pour l'ABES, ses experts en information bibliographique, en modélisation et en traitement des données seront directement impliqués. Par ailleurs, l'ABES ayant aujourd'hui des ressources en développement qui lui permettent de construire de façon *ad hoc* les outils spécifiques dont elle a besoin, elle a la capacité d'encadrer le développement des démonstrateurs et de planifier leur intégration éventuelle dans des outils en production. Enfin, une partie du réseau de catalogueurs du Sudoc pourra être mobilisée au moment de l'évaluation.

5.1.5 INA

L'**Ina** est un Établissement Public de l'État à caractère Industriel et Commercial, avec un effectif moyen de 980 personnes et un budget annuel de 100 M euros. Créé par la loi sur l'audiovisuel de 1975, l'Ina assure, entre autres missions la conservation et l'exploitation des archives de la radio et de la télévision, ainsi que le Dépôt Légal Audiovisuel qui a débuté au 1er janvier 1995 et qui permet une exploitation non commerciale à destination des chercheurs.

Dans ce projet l'Ina est en charge de modéliser les problèmes de qualité, d'enrichissement et de fusion des données de catalogage ainsi qu'un corpus de test prétraité (à la fois en terme de format et de connaissances a priori) afin de permettre l'évaluation des solutions retenues. L'Ina s'attachera tout particulièrement a élaborer des solutions d'enrichissement des données par utilisation de ressources externes disponibles soit sur le web des données soit sur des catalogues spécifiques de partenaires proches (sociétés d'auteurs par exemple) et l'extraction d'information pertinentes à partir de son propre corpus pour élargir son catalogue (création de notices d'autorité « œuvre » par exemple) afin d'améliorer sa communication avec ces mêmes partenaires.

L'équipe concernée bénéficiera pour ce faire de son expérience acquise sur les problèmes d'indexation sémantique au cours de ses cinq derniers projets de recherche. Ces projets lui ont procuré à la fois de multiples collaborations étroites avec le LIRMM (projet ANR OPALES et SAPHIR, projet européen LOGOS), une expérience sur la manipulation du

format RDF acquise sur les projets européens MeSH et ASSETS. Ce dernier, dédié à la bibliothèque numérique Europeana, a permis de développer un outil d'annotation de contenus multimédias connecté avec plusieurs bases du Web de données.

Pour mener les modélisations puis évaluations internes, cette équipe bénéficiera, au sein de l'Ina, du support de l'un des trois juristes investis dans le PCJ (Plan de Capitalisation Juridique) qui vise à terme à simplifier les procédures de libération des droits selon un nouveau paradigme (Blasi, 2011) ainsi que la responsable de la maintenance du thésaurus, ce qui permettra de recenser et tester les cas concrets d'import d'information (notices et référentiels) qui se présentent chaque année.

5.2. COMPLEMENTARITE DU CONSORTIUM

Le consortium est constitué d'un ensemble complémentaire et non redondant de partenaires. Les partenaires sont complémentaires car ce projet regroupe, d'une part, deux acteurs nationaux fondamentaux des systèmes d'informations documentaires détenteurs de très grandes bases et largement investis, au plan national mais aussi international, dans les problématiques d'exposition, de standardisation, d'interconnexion et de valorisation de leurs métadonnées, en particulier dans le contexte du web sémantique, et, d'autre part, trois équipes de chercheurs en informatique possédant une expertise reconnue en représentation des connaissances, bases de données et web sémantique. Les partenaires sont non redondants car les activités de l'ABES sont centrées sur les données et métadonnées bibliographiques, celles de l'Ina sur les données et métadonnées concernant les documents audiovisuels, les activités du LIG sur la gestion de données distribuées et hétérogènes, avec un focus sur les données du Web et leur intégration dans le cadre du Web sémantique, celles du LIRMM sur l'interrogation de grandes bases de connaissances en présence d'une ontologie, et celles du LRI sur la réconciliation de données et sur l'alignement d'ontologies.

Par ailleurs, de nombreux liens tissés dans la cadre de projets communs existent entre les cinq partenaires (par exemple l'ABES et le LIRMM dans les projets Mogador et SudocAd, l'Ina et le LIRMM dans les projets Opales, Saphir et Logos, le LIG et le LRI dans de nombreux projets puisque Marie-Christine Rousset dirigeait l'équipe IASI du LRI avant d'aller à Grenoble, le LRI et le LIRMM puisque Fatiha Saïs a fait un post-doc au LIRMM, etc.). Ces liens multiples et anciens assurent la garantie d'une compréhension mutuelle des attentes et apports de chacun.

6. ANNEXES

6.1. REFERENCES BIBLIOGRAPHIQUES

Abdallah, N., Goasdoué, F. et M.C., Rousset. 2009. DL-liteR in the Light of Propositional Logic for Decentralized Data Management. *IJCAI.* 2009.

Abiteboul, S., et al. 2010. Distributed Datalog revisited. *International Workshop on Datalog 2.0.* 2010.

Atencia, M., et al. 2011. Alignment-based trust for resource finding in semantic P2P networks. *10th International Semantic Web Conference (ISWC).* LNCS, Springer, 2011.

Baader, F., et al. 2003. *The Description Logic Handbook: Theory, Implementation, and Applications.* s.l. : Cambridge University Press, 2003.

Baget, J. F., et al. 2009. Extending Decidable Cases for Rules with Existential Variables. *IJCAI.* 2009, pp. 677-682.

—. 2011. On Rules with Existential Variables: Walking the Decidability Line. *Artificial Intelligence Journal.* 2011, Vol. 175, 9-10, pp. 1620-1654.

Baget, J. F., et al. 2011. Walking the Complexity Lines for Generalized Guarded Existential Rules. *IJCAI.* 2011, pp. 712-717.

Baget, J.F. 2005. RDF Entailment as a Graph Homomorphism. *International Semantic Web Conference.* 2005, pp. 82-96.

Baget, J.F., et al. 2010. Translations between RDF(S) and Conceptual Graphs. *ICCS.* 2010, pp. 28-41.

Baget, J.F., Leclère, M. et Mugnier, M.L. 2009. Walking the Decidability Line for Rules with Existential Variables. *International Conference of Knowledge Representation and Reasoning.* AAAI Press, 2009, pp. 466-476.

Bhattacharya, I. et Getoor, L. 2007. Collective Entity Resolution in Relational Data. *ACM Transactions on Knowledge Discovery from Data (ACM-TKDD).* 2007, pp. 1-36.

Blasi. 2011. Commercialisation d'un fond d'archives audiovisuelles: quelles données juridiques produire. *GFII.* [En ligne] 2011. http://gfii.fr/uploads/docs/commercialisation-d-un-fonds-d-archives-audiovisuelles-quelles-donnees-juridiques-produire-prix-de-la-meilleure-contribution-retours-d-experiences-i-expo-2011.pdf?symfony=lm0qorh0u6l8j9k9jghrh8rbl6.

Bleiholder, J. et Naumann, F. 2005. Declarative data fusion – Syntax, semantics, and implementation. *Proc. of the 9th East European Conference on Advances in Databases and Information Systems.* 2005, pp. 58-73.

Bouquet, P., et al. 2008. Entity Name System: The Backbone of an Open and Scalable Web of Data. 2008, pp. 554-561.

Buche, P., et al. 2011. An Ontology-Based Method for Duplicate Detection in Web Data Tables. *DEXA.* 2011, pp. 511-525.

Calì, A., et al. 2010. Datalog+/-: A Family of Logical Knowledge Representation and Query Languages for New Applications. 2010, pp. 228-242.

Chein, M. et Mugnier, M.L. 2009. *Graph-based Knowledge Representation and Reasoning—Computational Foundations of Conceptual Graphs.* s.l. : Springer, 2009.

Croitoru, M. et Van Deemter, K. 2007. A Conceptual Graph Approach for the Generation of Referring Expressions. *IJCAI.* 2007, pp. 2456-2461.

Doan, A., et al. 2003. Profile-Based Object Matching for Information Integration. *IEEE Intelligent Systems.* 2003, Vol. 18, 5, pp. 54-59.

Dong, X., Halevy, A. et Madhavan, J. 2005. Reference Reconciliation in Complex Information Spaces. *ACM SIGMOD International Conference on Management of Data.* ACM Press, 2005, pp. 85-96.

Genest, D. et Chein, M. 2005. A Content-search Information Retrieval Process Based on Conceptual Graphs. *KAIS.* 2005, pp. 292-309.

Genest, D. 2000. *Extension du modèle des graphes conceptuels pour la recherche d'informations.* Thèse d'Informatique : Université Montpellier 2, 2000.
Gruber, T. R. 1993. A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition.* 1993, Vol. 5, 2, pp. 199-220.
Hamdi, F., Reynaud, C. et Safar, B. 2010. Pattern-Based Mapping Refinement. *EKAW.* 2010, pp. 1-15.
Huhtala, Y., et al. 1999. TANE : An efficient algorithm for discovering functional and approximate dependencies. *Comput. Journal.* 1999, Vol. 42, 2, pp. 100-111.
Lehman, F. 1992. *Semantic Networks in Artificial Intelligence.* New York : Elsevier, 1992.
Levesque, H. J. et Lakemeyer, G. 2000. *The Logic of Knowledge Bases.* s.l. : MIT Press, 2000.
Moreau, N., et al. 2007. Formal and graphical annotations for digital objects. *SADPI.* 2007.
Moreau, N., Leclère, M. et Croitoru, M. 2009. Distinguishing Answers in Conceptual Graph Knowledge Bases. *ICCS.* 2009, pp. 233-246.
Newcombe, H., et al. 1959. Automatic Linkage of Vital Records. *Science.* 1959, Vol. 130, pp. 954-959.
Nguyen, G.H., Chatalic, P. et M.C., Rousset. 2008. A Probabilistic Trust Model for Semantic Peer to Peer Systems. *Prooceedings of international workshop on Data management in peer-to-peer systems (EDBT).* ACM International Conference Proceedings Series, 2008, Vol. 261, pp. 59-65.
Nikolov, A., Uren, V. et Motta, E. 2010. Data linking: Capturing and utilising implicit schema-level relations. *Workshop: Linked Data on the Web, WWW 2010.* 2010.
Ortiz, M., Rudolph, S. et Simkus, M. 2011. Query Answering in the Horn Fragments of the Description Logics SHOIQ and SROIQ. *IJCAI.* 2011, pp. 1039-1044.
Papakonstantinou, Y., Abiteboul, S. et Garcia-Molina, H. 1996. Object fusion in mediator systems. *VLDB.* 1996, pp. 413-424.
Ramchurn, S. D., Huynh, D. et N., Jenning. 2004. Trust in multi-agent systems. *he Knowledge Engineering Review.* 2004, Vol. 19, 1, pp. 1-25.
Saïs, F. et Thomopoulos, R. 2008. Reference Fusion and Flexible Querying. *Proceedings of OTM Conferences.* LNCS 5332, Springer, 2008, pp. 1541-1549.
Saïs, F., Pernelle, N. et Rousset, M.-C. 2009. Combining a Logical and a Numerical Method for Data Reconciliation. *Journal of Data Semantics.* LNCS Springer, 2009, Vol. 12, pp. 66-94.
Saïs, F., Pernelle, N. et Rousset, M.C. 2007. L2R: A Logical Method for Reference Reconciliation. *AAAI.* 2007, pp. 329-334.
Saïs, F., Thomopoulos, R. et Destercke, S. 2010. Ontology-Driven Possibilistic Reference Fusion. *Proceedings of OTM Conferences.* LNCS 6427, Springer, 2010, pp. 1079-1096.
Sismanis, Y., et al. 2006. Gordian: efficient and scalable discovery of composite keys. *Proceedings of the 32nd International conference on Very Large Data Bases (VLDB).* 2006, pp. 691-702.
Smalheiser, N. R. et Torvik, V. I. 2009. Author Name Disambiguation. *Annual Review of Information Science and Technology.* Information Today, Inc., 2009, Vol. 43, pp. 287-313.
Sowa, J. F. 1984. *Conceptual Structures : Information Processing in Mind and Machine.* s.l. : Addison-Wesley, 1984.
Subrahmanian, V., et al. 1995. *Hermes: A heterogeneous reasoning and mediator system.* s.l. : Technical Report, Univ. of Maryland, 1995.

Tournaire, R., et al. 2011. Discovery of Probabilistic Mappings between Taxonomies: Principles and Experiments. *Journal of Data Semantics.* 2011, Vol. XV.
Wang, D.Z., et al. 2009. Functional dependency generation and applications in pay-as-you-go data integration systems. *12th International Workshop on the Web and Databases.* 2009.
Winkler, W. E. 2006. *Overview of Record Linkage and Current Research Directions.* Washington : Bureau of the Census, 2006.