**Projet de recherche**

# Raisonner avec la provenance sur les données du Web

**Concours CR2 06/03**

Antoine Amarilli

**Résumé.** Ce document présente mon projet de recherche sur la gestion d'annotations riches de provenance pour le raisonnement, appliquée aux sources de données structurées sur le Web. Il présente également l'intégration possible de ce projet dans l'équipe LINKS du laboratoire CRIStAL, l'équipe Automates et applications de l'IRIF, et l'équipe GraphIK du LIRMM.

De nombreuses sources de données structurées se sont récemment développées sur le Web. Des *bases de connaissances généralistes* se construisent notamment à partir du texte de Wikipédia, par exemple DBpedia [BLK+09], YAGO [SKW07], et, depuis trois ans, Wikidata [VK14]. Elles rejoignent de nombreuses sources plus spécifiques, comme OpenStreetMaps [HW08] pour la cartographie, et de nombreux jeux de données publiques actuellement mis en ligne suivant le principe des *données ouvertes*[1] : ceux de `data.gouv.fr` [Eta15], ceux du STIF [STI15], et bien d'autres. Parallèlement à cela, de nombreuses pages Web s'enrichissent d'*annotations sémantiques* : l'ontologie Schema.org [sch15] propose ainsi un vocabulaire utilisé par Google [Goo15] et Bing [Bin15] pour améliorer leurs résultats ; les annotations Open Graph [Fac14] lient les pages Web au graphe social de Facebook ; Google Scholar [Goo] lui-même s'appuie sur des annotations sémantiques. Le projet Web Data Commons a ainsi extrait et centralisé 20 milliards de faits tirés d'annotations sur le Web [MB12], qui complètent les milliards de faits regroupés dans les bases de connaissances.

Ces nouvelles données rendent possibles de nombreuses applications innovantes. Les moteurs de recherche peuvent par exemple les utiliser pour proposer des réponses riches ou structurées, au-delà des simples mots-clés. Ainsi, Google Search utilise déjà la base de connaissances Google Knowledge Graph pour répondre aux requêtes sur des entités avec des Answer Box structurées, ou pour proposer des réponses sémantiques à des questions simples comme « gdp of France ». Les données du Web permettraient d'étendre cette approche et de répondre à davantage de questions, par exemple « budget of cnrs ». La popularité des données ouvertes rend également possible la réutilisation des jeux de données publiques, pour les recombiner avec le socle commun des ontologies généralistes ou les intégrer à des données existantes, et pour créer de nouvelles visualisations pour ces données, ou de nouvelles applications.

Cependant, l'utilisation de ces sources de données pose des défis spécifiques. En premier lieu, ces données ne sont pas *fiables* : elles sont souvent saisies ou proposées par des internautes sans garanties de qualité, et sont parfois extraites automatiquement par des règles, un processus susceptible d'introduire de nouvelles erreurs. De plus, ces sources évoluent *rapidement*, avec des millions

1. En anglais, *open data*

d'éditions par mois sur Wikidata, par des milliers de contributeurs [Wik15b]. Enfin, ces sources peuvent contenir des données *biaisées*, *incomplètes*, ou *périmées*. En s'appuyant sur ces données, par exemple, Google répond directement à la requête « capital of Israel » par une Answer Box indiquant « Jerusalem », et à la requête « country crimea » par une Answer Box indiquant « Ukraine » ; des réponses qui dépendent de la source retenue, et du moment où l'on pose la question !

Il y a pourtant beaucoup de signaux à utiliser pour quantifier notre confiance en ces données. La fiabilité d'un fait sur une page Web dépend de la page, et celle d'un jeu de données dépend de qui le fournit : ces signaux ont déjà été utilisés pour la recherche de la vérité [2], par exemple par Google pour Knowledge Vault [DGM+15]. Plus spécifiquement, sur les sites collaboratifs comme OpenStreetMaps ou Wikidata, les faits sont fournis par des utilisateurs qui interagissent de différentes manières (discussions, révocations, etc.), et on peut ainsi apprendre le réseau signé [3] [LHK10] indiquant quels utilisateurs se font confiance. De plus, sur Wikidata, les utilisateurs peuvent indiquer d'où provient chaque fait individuel, conservant ainsi l'information de la *source originale* qui fait autorité : plus de 35 millions de faits ont ainsi une source, soit près de la moitié des faits [Wik15c]. Sur OpenStreetMaps, 40 millions de nœuds et 120 millions de voies indiquent une source, soit environ 40%. Enfin, Wikidata et OpenStreetMaps fournissent un *historique complet* des changements, que l'on pourrait utiliser pour savoir si une information risque d'être périmée, ou si à l'inverse elle a été ajoutée très récemment et n'a donc pas encore été suffisamment vérifiée.

Exploiter ces informations permettrait de répondre aux problèmes de fiabilité et de mise à jour posés par ces sources. Mon projet de recherche consiste donc à proposer une approche générale pour résoudre ces problèmes, qui s'appuie sur le développement récent des techniques de gestion de la *provenance*. La provenance est un outil abstrait pour annoter le résultat d'une requête avec des informations indiquant les sources et les dérivations utilisées pour produire chaque résultat. Elle a été développée pour les bases de données relationnelles, où des représentations générales à base de semi-anneaux ont été formulées [GKT07], et où elle a permis de généraliser des problèmes auparavant étudiés indépendamment : expliquer les résultats d'une requête, représenter les sources dont ils dépendent, mais aussi gérer des politiques de sécurité ou de coûts sur les données, maintenir des vues, etc. La provenance a depuis été appliquée avec succès à d'autres domaines plus généraux, comme la gestion des flux de travaux [4], ou des données scientifiques [DF08 ; FKS+08]. *Mon projet de recherche consiste donc à développer les fondements de la gestion d'une provenance expressive pour l'évaluation de requêtes et le raisonnement sur les sources de données du Web.*

La provenance permet en effet de proposer une réponse générale au problème de la confiance en les données, en calculant quelles sources, quels faits, et quels utilisateurs se cachent derrière les réponses que l'on obtient. On pourrait ainsi, en développant les bons outils, obtenir des notions bien fondées de confiance, de préférence, ou de récence des réponses, à partir des données initiales, et s'en servir pour *classer* les résultats par ordre d'intérêt pour l'utilisateur. La provenance d'une réponse peut également servir à *expliquer* la réponse à l'utilisateur, et à lui présenter les sources utilisées ; elle permet aussi d'*identifier les contradictions* dans les données. Enfin, elle permet de savoir comment les résultats d'une requête doivent être *mis à jour* à chaque changement des sources de données. La provenance capture et généralise ainsi des questions actuellement étudiées séparément, comme celle de la confiance [Sch08] ou de la temporalité [Mot12] sur les données du Web.

Cette nouvelle direction de recherche pose un défi scientifique majeur : *l'intégration de la provenance et du raisonnement*. En effet, les sources du Web sont diverses et doivent être intégrées et liées entre elles ; elles sont aussi incomplètes et doivent être étendues par des règles suivant l'hypothèse du monde ouvert. Ainsi, l'évaluation de requêtes sur ces sources se traduit généralement en du *raisonnement* suivant des contraintes logiques expressives dans des formalismes adaptés, comme les

2. En anglais, *truth finding*
3. En anglais, *signed network*
4. En anglais, *workflows*

*logiques de description* [Baa03], les *règles existentielles* [BLM+09], ou le langage *Datalog*$^{\pm}$ [CGL12]. Contrairement au contexte relationnel, on ne peut donc pas se contenter de définir la provenance pour un simple langage de requêtes ; il faut la définir pour des langages complexes de contraintes expressives, aux frontières de la décidabilité.

La difficulté majeure pour calculer de la provenance sur les bases de connaissance est donc de faire ce calcul *à travers le raisonnement* : comprendre comment les résultats d'une requête dépendent, non seulement des données, mais aussi des *règles logiques* utilisées pour compléter les données. En effet, les règles avec lesquelles on raisonne ne sont généralement pas fiables : les bases de connaissances sont alignées entre elles [ES07] de façon automatique et sujette à erreur, ou bien avec des appariements de schémas [5] incertains [DHY07]. Les bases sont également alimentées par des règles qui peuvent produire des résultats faux : YAGO conserve par exemple [BKS13] une forme de provenance indiquant quel extracteur a généré quel fait, afin d'identifier les extracteurs faillibles. Bientôt, ces bases pourront même être étendues par des conséquences plausibles obtenues en généralisant à partir de règles statistiques, que l'on commence déjà à extraire automatiquement [GTH+13]. La provenance sur les données du Web nous fournirait de nouvelles fondations pour raisonner sur ces données en tolérant les contradictions [LS08] et les erreurs possibles des règles. En conservant le lien entre les données, les règles, et les conséquences, la provenance nous permettrait d'identifier les dérivations fiables et de réviser nos estimations, et ainsi de relever le défi du *raisonnement avec des données et règles incertaines.*

Je présente dans la suite de ce document le détail des axes de mon projet. Le premier défi sera de formuler les *fondements d'une provenance pour le raisonnement* (Axe 1). Une fois définie cette notion générale, j'entends étudier son application aux problèmes de confiance, de récence, et de préférence des données. Je compte en particulier l'utiliser *qualitativement* pour aboutir à un classement sur les résultats d'une requête en propageant un ordre de préférence ou de temporalité sur les données et règles initiales (Axe 2). Je compte aussi m'intéresser aux *représentations quantitatives* de la confiance et de l'incertitude à travers la provenance, notamment les représentations probabilistes, et explorer de nouvelles directions pour assurer la tractabilité de ces problèmes (Axe 3). À plus long terme, j'envisage aussi d'étudier la question de la *révision* des estimations de confiance, en remontant aux données initiales avec la provenance (Axe 4). Je conclus ce document par une mise en perspective de mon projet de recherche au regard de mes travaux antérieurs et de mes collaborations en cours, et par une étude de sa faisabilité et des étapes concrètes que j'espère atteindre. Je présente enfin mon projet d'intégration dans trois équipes possibles.

## Axe 1 – Provenance pour le raisonnement en monde ouvert

La réponse aux requêtes sur des données en monde ouvert sur le Web s'effectue à travers des techniques de raisonnement. Le problème fondamental est celui de la *réponse aux requêtes en monde ouvert* : étant donné des faits initiaux, des contraintes logiques et une requête, on souhaite déterminer les réponses à la requête qui sont *certaines*, c'est-à-dire, qui sont vraies dans toutes les complétions des faits initiaux qui satisfont les contraintes. Ce problème a été étudié en détail par les communautés des logiques de description [Baa03] et des règles existentielles [BLM+09], et a été exploré depuis plus longtemps encore pour les bases de données relationnelles [JK84 ; CLR03a].

Pourtant, à ma connaissance, peu de méthodes ont été proposées pour effectuer ce raisonnement en maintenant une information riche de *provenance*, afin d'indiquer comment les réponses dépendent des faits initiaux. La provenance a été définie dans le cadre classique des bases de données relationnelles, avec des langages comme Datalog [GKT07], mais Datalog ne fait pas vraiment

5. En anglais, *schema mappings*

l'hypothèse du monde ouvert, car il ne permet pas d'affirmer l'existence de nouveaux objets. Le langage Datalog$^{\pm}$ corrige ce problème, mais seules des approches très récentes [LVP+14] esquissent des notions de provenance pour ce langage. Au-delà du contexte des bases de données, la provenance a certes été définie pour les données du Web [Mor10], avec des représentations standardisées [DMF06 ; MCF+11], ainsi que pour les langages de requêtes du Web sémantique, comme SPARQL [GKC+13], mais ces approches étudient seulement l'évaluation de requêtes et ne s'intéressent pas au raisonnement. Lorsque l'on souhaite raisonner sur les données, les approches actuelles sont spécifiques à des tâches particulières [FFP+09 ; BHP+11 ; SDS11 ; DHS12], en particulier l'explication ou la justification de résultats [KPS+05 ; BCR08], ou limitées à des langages peu expressifs comme RDF Schéma [ZLP+12] ; elles ne fournissent pas de définition générale de la provenance pour des langages complexes. Des notions générales de *graphe de provenance* ont par ailleurs été proposées pour les flux de travaux, mais ces approches traitent les tâches effectuées comme des boîtes noires, sans s'intéresser à leur structure.

Dans le cas du raisonnement, on s'attend pourtant à ce que la provenance reflète précisément les hypothèses et les dérivations utilisées, pour obtenir des annotations symboliques abstraites sur chaque résultat. Ces annotations ne devraient pas être spécifiques au problème de la gestion de la confiance, mais devraient capturer et généraliser de nombreux problèmes, comme dans le contexte relationnel [GKT07 ; KG12]. En d'autres termes, l'évaluation de requêtes et le raisonnement devraient calculer des annotations *abstraites* de provenance pour chaque résultat, en fonction de variables représentant les annotations encore inconnues sur les données, par exemple, la confiance sur les sources. Les propriétés de *spécialisation* et de *commutation avec les homomorphismes* permettraient ensuite de calculer les annotations concrètes de provenance, une fois fixées la tâche précise et les confiances exactes. On pourrait par exemple imaginer que l'utilisateur puisse évaluer une requête, et ensuite choisir différents critères de tri, ou différents profils, correspondant à différentes manières d'attribuer une confiance aux sources, et menant à différentes évaluations des mêmes annotations abstraites de provenance. La provenance doit ainsi permettre de *découpler* l'évaluation de requêtes par le raisonnement avec propagation d'annotations expressives de provenance (présentée dans cet axe), du problème de spécialiser et d'utiliser ces annotations (présenté dans les deux prochains axes).

Pour définir ainsi une provenance générale, propagée à travers le raisonnement, qui indique la dépendance des résultats en les données initiales, les difficultés à surmonter dépendent de l'*expressivité* attendue. Par exemple, si l'on souhaite simplement généraliser la Why-Provenance [BKT01], il suffirait d'indiquer quel sous-ensembles de faits initiaux permet de déduire la requête avec les règles. Plus ambitieusement, on peut vouloir définir une $\mathbb{N}[X]$-provenance [GKT07], qui dans le contexte des bases de données est *universelle* et capture toutes les autres formes de provenance à base de semi-anneaux. Il faut dans ce cas garder trace du *nombre de fois* qu'un fait initial a été utilisé dans la dérivation, et du *nombre de manières* de dériver chaque réponse avec les mêmes faits.

Selon l'expressivité que l'on vise, le premier défi consiste donc à trouver une bonne *définition* de ces provenances expressives pour des tâches de raisonnement qui vont au-delà de l'évaluation de requêtes. Ces définitions devraient en particulier coïncider avec la définition classique dans le cas particulier de l'évaluation sans règles, et respecter les propriétés de spécialisation et de commutation avec les homomorphismes, pour pouvoir être utilisées pour des tâches plus spécifiques. Le second défi consiste à *calculer* efficacement cette provenance : on souhaiterait en particulier le faire en temps polynomial en les données, à requête fixée, comme on peut le faire pour l'algèbre relationnelle positive [GKT07]. On voudrait enfin *représenter* cette provenance de façon concise en développant de nouveaux formalismes, comme la notion récente de *circuits de provenance* [DMR+14].

Bien sûr, la représentation de la provenance peut également dépendre de la *méthode de raisonnement* qu'on utilise, parmi celles couramment employées en gestion de données. Une première

approche courante est le *chaînage avant*, en particulier l'algorithme de *poursuite*[6] qui construit une complétion universelle des données sous les contraintes, sur laquelle on peut lire les réponses certaines. Il faudrait ainsi étudier comment définir et représenter la provenance des faits au cours de la poursuite, une direction récemment explorée pour optimiser la reformulation de requêtes [ICD+14]. Ces questions sont particulièrement délicates dans les cas où le résultat de la poursuite est infini, et doit être manipulé de façon implicite, par exemple comme un arbre infini régulier. Une seconde approche pour la réponse aux requêtes en monde ouvert est le *chaînage arrière*, ou la réécriture [CLR03b], où l'on réécrit la requête en incorporant les contraintes pour l'évaluer directement sur les données initiales. Il faudrait ainsi développer des techniques de réécriture qui maintiennent les informations nécessaires au calcul de provenance.

Si l'on parvient à définir une notion de provenance en fonction des *faits* pour le raisonnement, une tâche plus ambitieuse serait ensuite de définir une provenance qui reflète aussi les *règles de raisonnement* que l'on utilise, afin de rendre possible un raisonnement robuste sous des règles incertaines à l'aide d'annotations symboliques. Là encore, suivant l'expressivité que l'on désire atteindre, on peut simplement parler des sous-ensembles de faits *et de règles* qui permettent d'obtenir la réponse, ou on peut représenter davantage d'informations. Par exemple, si l'on sait que nos règles de raisonnement ne sont pas parfaites et sont parfois fausses, il est crucial de mesurer le *nombre de fois* qu'elles doivent être appliquées pour parvenir à un résultat (car chaque application rend le résultat moins fiable) ; mais il faut aussi mesurer le *nombre de manières* de les appliquer pour dériver le résultat par différents biais (car cela augmente notre confiance, à condition que les dérivations soient indépendantes, en un sens à définir). Dans tous ces cas, la question de l'efficacité du calcul et de la représentation se pose à nouveau.

D'autres questions nouvelles sur la provenance sont posées par les sources de données en monde ouvert, par exemple la gestion de l'*exhaustivité* : risque-t-on de rater certains résultats à cause de l'incomplétude des sources ? Sur des données en monde ouvert, ce risque existe toujours, mais dans le contexte récemment étudié d'un monde *mixte*, ouvert suivant certaines dimensions et fermé suivant d'autres, il est plus délicat d'estimer ce risque [RKN+15]. OpenStreetMap est ainsi complet pour le tracé des communes françaises [Fra13], mais pas pour les rues ; Wikidata indique parfois que l'absence d'une valeur est délibérée, par exemple qu'un personnage historique n'a *pas* d'enfant [Wik15a] ; ces informations de complétude peuvent parfois être déduites automatiquement [GTH+13]. Cependant, comme ces informations elles-mêmes ne sont pas fiables, une question naturelle se pose : comment raisonner avec de telles informations de complétude pour décider si notre résultat est exhaustif, tout en représentant la *provenance* de cette information d'exhaustivité ?

On pourrait aussi s'intéresser, sur les sources de données du Web, à la provenance sur les *valeurs manquantes*. Wikidata permet par exemple d'indiquer [Wik15a] qu'une valeur existe mais qu'elle est inconnue ; de telles informations peuvent aussi être générées lors du raisonnement, notamment lorsque des règles d'intégration de données impliquent l'existence d'un fait mais sans déterminer tous ses champs. Les données incomplètes ont été étudiées depuis longtemps dans le contexte relationnel [AHV95, Chapitre 19], mais la provenance reste à définir sur de telles données, par exemple pour représenter comment la requête dépend des choix effectués pour compléter les données manquantes.

## Axe 2 – Propagation d'ordres à travers la provenance

Une fois définie une notion générale de provenance, qui représente comment les résultats d'une requête dépendent des données initiales et des règles, mon objectif est d'utiliser cette information pour l'application que j'ai présentée : propager des jugements sur les sources et sur les règles jusqu'aux résultats, pour la récence, la fiabilité ou la pertinence. Une première approche, que je

6. En anglais, *chase*

développe dans cet axe, est *qualitative* : si l'utilisateur indique qu'une source est *plus fiable*, ou *plus récente*, qu'une autre source, ou que certaines règles sont *plus probables* que d'autres, on peut s'en servir pour déduire que certains résultats sont *meilleurs* que d'autres, et en tirer un *classement* sur les résultats qui respecte les préférences de l'utilisateur.

Considérons d'abord une vision *implicite* de la provenance, où l'on propage des annotations à travers le raisonnement sans que les règles logiques les mentionnent. On pourrait alors utiliser les relations d'ordre définies par l'utilisateur sur les sources et les règles, et les propager structurellement aux annotations symboliques complexes qui annotent les résultats du raisonnement. Cette tâche devient ardue si les règles de raisonnement sont suffisamment expressives : comment déterminer alors quels ordres sur les résultats sont réellement conformes aux préférences initiales ? Pire encore, si l'on doit imposer un ordre *total* sur les résultats, par exemple parce qu'on veut restreindre aux meilleures réponses et afficher une liste triée, il va falloir faire des choix pour compléter l'ordre sur les annotations en un ordre total, en extrapolant les préférences indiquées par l'utilisateur. Idéalement, on voudrait pouvoir *garder trace* de ces choix, en étendant l'annotation de provenance pour les couvrir, notamment pour recalculer facilement l'ordre des résultats suivant les filtres et critères choisis. Par exemple, si l'on décide pour l'affichage de faire davantage confiance à Wikidata qu'à YAGO, cette information devrait être reflétée par la provenance, pour qu'il soit possible d'expliquer ce choix et de le remettre facilement en question.

Une vision plus ambitieuse est celle d'une provenance *explicite* sur laquelle l'utilisateur pourrait définir des relations complexes de préférence, dans le même langage logique que celui utilisé pour le raisonnement. L'utilisateur pourrait ainsi indiquer des contraintes comme « un fait sur Wikidata ne peut pas être plus récent que la plus récente de ses sources », ou « si deux faits sont incompatibles, préférer la révision la plus ancienne, sauf si j'ai indiqué que je faisais davantage confiance à une des sources ». Pour définir l'*ordre* de pertinence sur les résultats d'une requête, il faudrait alors raisonner sur les annotations de provenance elles-mêmes, suivant ces contraintes logiques d'ordre sur le temps, la confiance et la préférence.

Ce problème est délicat, car les langages de raisonnement actuellement utilisés sont peu adaptés pour manipuler des relations d'ordre. Ces relations sont notamment *transitives* (un utilisateur qui préfère Wikidata à YAGO et YAGO à MusicBrainz préférera Wikidata à MusicBrainz), mais les langages logiques utilisés pour la réponse aux requêtes ne peuvent généralement pas imposer la transitivité en conservant la décidabilité du raisonnement : les règles existentielles gardées [ANB98] et frontière-gardées [BLM10] ne l'autorisent pas et deviennent indécidables si on les étend en ce sens [GPT13] ; les logiques de description qui peuvent l'exprimer [EOS+12] restreignent souvent son interaction avec les autres contraintes (notamment la cardinalité). Les relations d'ordre sont également *antisymétriques* (il est impossible que A soit plus récent que B, que B soit plus récent que C, et C plus récent que A) ; cette condition aussi n'est généralement pas exprimable. Enfin, on peut souhaiter imposer que l'ordre soit *total*, notamment pour l'ordre final sur les résultats : cette contrainte n'est généralement pas exprimable non plus dans les formalismes gardés et les logiques de description.

Ainsi, on ne dispose pas aujourd'hui des bonnes techniques de gestion de la provenance pour exploiter des indications de préférence, et on dispose encore moins des langages de règles qui nous permettraient de raisonner sur de telles annotations explicites avec des relations d'ordre. Ce sont ces défis que je me propose de relever, pour utiliser les annotations qualitatives de provenance sur les sources et règles.

## Axe 3 – Confiance quantitative et tractabilité

Une autre manière d'utiliser la provenance, au lieu de faire des choix *qualitatifs* sur la pertinence des faits et des résultats, est de procéder de manière *quantitative*, c'est-à-dire, calculer des scores de confiance numériques à partir d'une estimation de la qualité des sources, et en déduire un score pour les résultats. C'est en particulier l'objet des applications *probabilistes* de la provenance, souvent appelée lignage [7] dans ce contexte : la provenance nous permet de calculer la *probabilité* d'une réponse, à partir de la probabilité que les faits d'entrée soient corrects, que les utilisateurs soient dignes de confiance, etc. Fonder l'étude du raisonnement probabiliste sur la provenance offre ainsi une base formelle pour le raisonnement avec des faits et des règles incertaines, et permet de découpler le calcul symbolique de la provenance et le calcul numérique des probabilités. En étendant ces techniques aux données du Web, on devrait ainsi pouvoir capturer les approches probabilistes déjà étudiées séparément, par exemple pour les logiques de description [dFL08 ; LS08].

Le passage d'une confiance qualitative à une confiance quantitative pose cependant de nouvelles difficultés. Certaines sont *définitionnelles* : peut-on définir un modèle probabiliste raisonnable sur les résultats d'une requête, si les probabilités initiales portent sur les faits, mais également sur la probabilité d'application des règles ? Comment faire si l'univers peut être étendu, dans le cadre du raisonnement, par l'introduction d'un nombre arbitraire de nouveaux éléments ? D'autres difficultés sont *computationnelles* : même sans raisonnement, dans le cadre traditionnel de l'évaluation d'une requête fixée, le passage aux représentations probabilistes introduit souvent de l'intractabilité [DS07]. Par ailleurs, les langages existants pour le raisonnement probabiliste [RKT07 ; BCK+] imposent généralement de renoncer aux garanties que l'on voudrait exiger sur la précision des probabilités calculées ou sur le temps de calcul.

J'ai déjà travaillé sur les questions de tractabilité pour l'évaluation probabiliste [**ABS15** ; **ABS16**], et je compte ainsi m'intéresser dans un premier temps à ces questions sur les données du Web pour l'évaluation de requêtes au sens classique, et entreprendre l'étude de nouvelles approches que j'ai récemment esquissées avec mes collaborateurs. Une première direction serait de combiner l'évaluation probabiliste basée sur les requêtes avec nos méthodes basées sur les instances [**ABS15**], par exemple pour choisir des plans d'évaluation pour les requêtes qui privilégient les parties *simples* de l'instance. La deuxième serait d'étudier des méthodes tractables d'évaluation approchée, avec des bornes d'erreur contrôlées, par exemple en réécrivant l'instance pour la rendre *plus simple* pour la requête : cette direction fait écho à de récents travaux dans le même sens [GS15], qui procèdent du point de vue de la requête, en l'approximant par des requêtes plus simples. Enfin, la troisième direction serait d'utiliser des méthodes *hybrides* pour le calcul probabiliste [MCS14], qui effectuent du calcul exact lorsque c'est possible, et du raisonnement approché par échantillonnage [8] sur les parties des données où les approches exactes deviennent infaisables.

Je compte ensuite étudier les nouveaux défis posés par l'extension du calcul probabiliste au contexte du raisonnement, notamment à celui de l'incertitude sur les règles. En particulier, en monde ouvert, l'ensemble de faits potentiellement utiles pour répondre à la requête devient infini, donc il devient plus complexe de définir une distribution de probabilités sur les mondes possibles. Par ailleurs, lorsqu'une règle affirme l'*existence* d'un objet, la distribution sur cet objet n'est pas évidente à formaliser : quelle est la probabilité qu'il s'agisse d'un objet déjà connu dans les données dont on dispose, ou bien qu'il s'agisse d'un nouvel objet, peut-être déjà utilisé pour d'autres applications de cette règle ? En présence de contraintes d'ordre, il est également délicat de définir un modèle probabiliste bien fondé sur les manières possibles de compléter l'ordre. Enfin, lorsqu'une réponse peut être dérivée par différentes règles à partir d'hypothèses plus ou moins certaines, on ne sait pas

7. En anglais, *lineage*
8. En anglais, *sampling*

encore comment combiner la probabilité des hypothèses et des règles, pour savoir quelle probabilité donner à la réponse.

Une direction intéressante dans ce contexte serait d'étendre les algorithmes de chaînage avant, et de généraliser la poursuite[6] pour travailler avec des règles probabilistes. On pourrait ainsi construire un *modèle universel probabiliste*, et, pour les fragments gardés qui garantissent que ce modèle est arborescent, on pourrait peut-être évaluer des requêtes même si le modèle est infini, en utilisant les techniques habituelles d'automates d'arbres [CDG+07] ou de Datalog monadique [GPW10], par exemple en calculant les probabilités dans le formalisme des chaînes de Markov récursives [EY09].

## Axe 4 – Retours utilisateur et révision de la confiance

Après avoir exploré les questions de représentation de la provenance en monde ouvert, et de propagation d'ordres qualitatifs ou de valeurs quantitatives à travers le raisonnement, une direction à plus long terme de mon programme de recherche serait de pouvoir *réviser* des informations de confiance ou de préférence sur les sources ou sur les règles, notamment en intégrant des retours fournis par les utilisateurs, ou en identifiant des contradictions.

En effet, dans la vision d'un système pour raisonner en monde ouvert sur des données du Web, on s'attend à ce qu'un utilisateur soit en mesure de réagir aux résultats fournis par le système. L'utilisateur pourrait valider ou invalider les résultats, soit directement, soit à travers des indices indirects comme le temps passé sur chaque résultat ou les clics effectués. Comme l'ordre sur ces résultats serait choisi à partir de leurs annotations de provenance, ces retours fourniraient des informations précieuses pour réviser la confiance que l'on peut accorder aux données initiales et aux règles, à condition de parvenir à remonter jusqu'à elles. Bien sûr, dans des contextes comme l'intégration de données, des travaux ont déjà étudié [TJM+08] comment utiliser les retours utilisateurs et la provenance pour apprendre les bons résultats. Il s'agirait, là encore, d'étendre de telles approches au problème du raisonnement en monde ouvert.

Même en l'absence de retours utilisateur, la question de la réévaluation de la confiance se pose tout de même, dans le contexte de la détection de contradictions et du raisonnement sous ces contradictions. En effet, si l'on veut raisonner avec des règles et des données incertaines, il est souvent utile d'avoir des règles *négatives*, comme des *dépendances fonctionnelles*, pour restreindre la création de nouveaux faits, et limiter la *dérive*[9] : par exemple, il ne faut généralement pas inférer plus de deux parents pour une même personne. Ces règles négatives peuvent bien sûr être elles-mêmes incertaines : la plupart des personnes ont une seule nationalité, mais certaines en ont deux. Lorsque l'application de règles nous conduit à prédire des faits peu probables, il faut pouvoir raisonner malgré les contradictions apparentes, mais il faut aussi *réviser* notre confiance en les règles et en les faits initiaux, pour détecter les contradictions ou les règles incohérentes, et éventuellement proposer des réparations possibles à l'utilisateur.

Dans le contexte général du raisonnement en monde ouvert, cependant, il est particulièrement complexe d'utiliser judicieusement de telles informations sur les résultats. Si une règle négative nous suggère que deux résultats ne peuvent probablement pas être vrais en même temps, si l'utilisateur nous indique qu'une réponse est fausse ou moins intéressante qu'une autre, comment *généraliser* à partir de cette information ? Comment en tirer les bonnes conclusions sur les confiances initiales à attribuer aux faits et aux règles ? Pour comprendre cela, il faut *remonter* la provenance, et ramener ces informations sur les annotations complexes à leur meilleure traduction possible sur les faits de base. À ma connaissance, ce problème n'a pas été étudié sur la provenance de manière générale, même s'il se rapproche du *conditionnement* des bases de données probabilistes [KO08], qui vise à modifier une représentation probabiliste (sans provenance) pour intégrer une nouvelle information.

9. En anglais, *drift*

Une autre direction intéressante qui concerne les retours utilisateur est l'étude de contextes où le système peut choisir de *susciter* des retours sur des données ou des réponses de son choix. Par exemple, le système pourrait attirer l'attention de l'utilisateur sur un résultat bien choisi, pour savoir s'il est pertinent ; le système pourrait valider certains résultats en posant des questions à des experts, ou à la *foule*[10]. Dans ces situations, il faut déterminer *quels retours* nous donneraient le plus d'information, pour choisir comment sonder les utilisateurs, sans poser trop de questions. Lorsqu'il s'agit d'apprendre ainsi à identifier les résultats pertinents pour l'utilisateur, le problème ressemble à de l'apprentissage de requêtes, étudié auparavant dans le cadre relationnel [Zlo75 ; AAP+13]. On peut aussi penser à des liens plus généraux avec l'*apprentissage actif*[11] sur des données structurées, ou le problème d'*apprendre à classer*[12] dans le cas des relations d'ordre.

L'étude de cette dernière direction complèterait ainsi cette vision d'un système capable d'utiliser la provenance pour raisonner sur de grands volumes de données, en maintenant des informations de confiance et de pertinence, et en interagissant avec l'utilisateur : le système saurait non seulement exploiter des retours sur les réponses qu'il calcule, mais saurait aussi poser les bonnes questions.

## Originalité et faisabilité du projet

Cette section décrit le contraste entre mon projet de recherche et mes travaux de recherche antérieurs, et présente certaines des collaborations que j'ai déjà entreprises autour des ces questions. J'étudie ensuite la faisabilité à court et à long terme des différentes directions de mon projet.

Ma recherche a principalement porté jusqu'à présent sur la gestion des données incertaines et probabilistes, et sur la réponse aux requêtes en monde ouvert. J'ai ainsi pu m'intéresser aux semi-anneaux de provenance, et à leur extension au-delà du cadre relationnel pur, en proposant notamment une notion de provenance pour les automates d'arbre [**ABS15**], et en appliquant ces techniques à la gestion probabiliste de l'incertitude [**ABS16**]. Pour la réponse aux requêtes en monde ouvert, j'ai adopté une approche transversale visant à unifier les travaux entrepris par les communautés des logiques de description et des règles existentielles [**AB15a**], tout en m'intéressant aussi au contexte habituel des bases de données [**AB15b**].

Ce positionnement de ma recherche m'amène naturellement à la nouvelle problématique que pose mon projet : l'extension des techniques de gestion de la provenance vers le nouveau contexte du raisonnement en monde ouvert, afin de gérer l'incertitude et la confiance sur les données. Mon projet combine ainsi les deux principales directions de ma recherche, tout en posant des questions résolument nouvelles : développement de nouvelles définitions de la provenance, interaction entre règles de raisonnement et provenance, nouvelles applications de la provenance pour la confiance sur les sources du Web, et gestion des retours fournis par les utilisateurs via la provenance.

Suivant cette perspective, plusieurs collaborations que j'entretiens actuellement ont commencé à aborder certaines directions à court terme du projet de recherche présenté ici :

— Je travaille avec Daniel Deutch (University of Tel Aviv), M. Lamine Ba (Qatar Computing Research Institute) et Pierre Senellart sur la propagation d'informations d'ordre partiel sur les faits [**ABD+16**] à travers l'évaluation de requêtes de l'algèbre relationnelle ; ces travaux peuvent être vus comme un cas particulier de certaines questions de l'Axe 2.

— Ma collaboration avec Michael Benedikt (Université d'Oxford) se poursuit avec Pierre Bourhis (CNRS CRIStAL) et Michael Vanden Boom (Université d'Oxford) avec l'étude de contraintes de transitivité et d'ordre pour les langages gardés, de sorte à préserver la décidabilité du

10. En anglais, *crowdsourcing*
11. En anglais, *active learning*
12. En anglais, *learning to rank*

raisonnement en monde ouvert, et sa tractabilité en fonction des données. Cette question rejoint aussi l'Axe 2, mais n'étudie pas de lien avec la provenance.

— Les prochaines étapes de mon travail avec Pierre Bourhis (CNRS CRIStAL) et Pierre Senellart (Télécom ParisTech) consistent à étudier de nouvelles techniques pour assurer la tractabilité de l'évaluation probabiliste dans le contexte classique des données relationnelles, ce qui constitue les premières questions de l'Axe 3.

— La gestion des retours utilisateurs (Axe 4) s'inspire de ma collaboration avec Tova Milo et Yael Amsterdamer (Tel Aviv University) portant sur la foule[13] [**AAM14a**; **AAM14b**; **AAM+16**], qui constitue un champ d'application naturel pour ces questions.

— Enfin, j'ai commencé à travailler avec Fabian Suchanek et son doctorant Luis Galárraga (Télécom ParisTech) sur des questions liées à leur ontologie YAGO [SKW07] ; nous travaillons notamment sur le raisonnement sous des règles probabilistes et sur la gestion de la complétude.

Je compte m'appuyer sur ce réseau de collaborations, et sur les questions immédiates que nous étudions, pour m'attaquer au projet général que j'ai présenté. S'il est particulièrement ambitieux dans son ensemble, je suis néanmoins persuadé que certaines de ses directions peuvent être réalisées à court ou à moyen terme :

**Axe 1.** La totalité de mon projet impose de parvenir à définir une forme de provenance générale qui puisse être appliquée pour le raisonnement. Ceci me semble faisable à court terme, au moins pour des semi-anneaux de provenance faiblement expressifs, et permettrait déjà d'abstraire certaines des tâches spécifiques étudiées hors du contexte général de la provenance, ainsi que de poursuivre l'étude des axes suivants.

On peut ensuite étendre la provenance suivant plusieurs directions indépendantes : définitions plus expressives ; dépendance en les règles ; représentations efficaces et tractabilité ; valeurs manquantes et exhaustivité. L'objectif à long terme est bien sûr d'intégrer ces approches, mais on peut d'abord les étudier séparément.

**Axe 2.** La propagation d'un ordre qualitatif sur la provenance implicite me semble accessible à court terme, mais pose aussi à moyen terme des questions de tractabilité et de représentation.

Le raisonnement sur la provenance explicite est une direction plus ambitieuse. Elle peut d'abord prendre la forme d'un prolongement de mes travaux actuels sur le raisonnement avec relations d'ordre, pour les étendre spécifiquement ensuite dans la direction de la provenance.

**Axe 3.** L'évaluation probabiliste tractable est délicate, mais il me semble utile d'étudier de nouvelles approches à base d'échantillonnage, d'approximation, et de gestion jointe des données et des requêtes. On peut d'abord étudier ces questions indépendamment du raisonnement.

Les questions autour des règles probabilistes sont un défi à plus long terme. Je pense toutefois qu'on peut déjà définir et étudier des modèles dans ce contexte, comme la poursuite probabiliste, mais qu'ils ne seront pas bien fondés ou tractables dans toutes les situations.

**Axe 4.** La problématique générale de l'intégration des retours et du choix des questions, présentée dans cet axe, est plus prospective. Certains problèmes semblent toutefois suffisamment bien définis pour être étudiés dès maintenant, par exemple en commençant par le cadre mieux compris de la provenance relationnelle, pour généraliser ensuite à nos propres représentations.

## Projet d'intégration

Je détaille ici comment mon projet de recherche pourrait s'intégrer au sein d'équipes d'accueil dans trois laboratoires CNRS, présentés par ordre alphabétique :

13. En anglais, *crowdsourcing*

**CRIStAL (Lille).** L'équipe LINKS du laboratoire CRIStAL s'intéresse spécifiquement au thème de mon projet de recherche, à savoir, la gestion du Web des données[14] et des sources hétérogènes, en utilisant des approches logiques, et en gérant l'incertitude et l'incomplétude sur les données. Mon projet se rattache notamment à leur étude de l'intégration de données appliquée aux sources hétérogènes du Web, problème que l'on peut voir comme une forme de raisonnement. Plus spécifiquement, je pourrais tirer parti de la forte expertise de cette équipe en matière de raisonnement sous contraintes expressives, notamment avec Pierre Bourhis, avec qui j'ai déjà eu l'occasion de collaborer [**ABS15** ; **ABS16**] ; et également pour certains outils techniques importants comme les automates d'arbres, avec notamment Sophie Tison ; je pourrais pour ma part apporter à l'équipe mes compétences en matière de gestion de données probabilistes et de gestion de la provenance.

Le dernier axe de mon projet, où je propose un système capable d'apprendre les préférences de l'utilisateur sur les résultats et sur les sources, se rapproche également des questions d'apprentissage étudiées par LINKS et par l'équipe voisine MAGNET (notamment Marc Tommasi) : apprentissage de requêtes, apprentissage symbolique, et apprentissage statistique sur des données structurées.

**IRIF (Paris).** L'équipe Automates et applications de l'IRIF s'est récemment enrichie d'une thématique portant sur les bases de données, avec l'arrivée de Cristina Sirangelo et d'Amélie Gheerbrant. Leur travail porte notamment sur l'évaluation de requêtes sur les données incomplètes, un problème directement lié à mon projet de recherche. Plus généralement, mon intégration à l'IRIF renforcerait ce groupe de recherche sur les bases de données, qui est appelé à croître, et lui apporterait de nouvelles perspectives, comme la provenance ou la gestion de données probabilistes.

L'équipe a également une forte composante portant sur les automates d'arbres, un outil théorique important pour le raisonnement en monde ouvert et l'évaluation de requêtes, que mes propres travaux [**ABS15**] ont par ailleurs connecté à la notion de provenance. On peut notamment mentionner Thomas Colcombet, qui les a appliqués pour les logiques gardées, et Olivier Serre. Ma recherche pourrait également trouver des liens avec l'étude des graphes à grande échelle actuellement menée par l'équipe projet INRIA GANG, par exemple pour la gestion de grandes bases de connaissances.

**LIRMM (Montpellier).** L'équipe GraphIK du LIRMM s'intéresse à la représentation des connaissances et au raisonnement sous contraintes expressives. Cette équipe a la spécificité de réunir en France des spécialistes de différents formalismes expressifs de raisonnement : les logiques de description, avec notamment Meghyn Bienvenu, et les règles existentielles, avec par exemple Jean-François Baget et Marie-Laure Mugnier. Ma venue apporterait une compétence complémentaire en théorie des bases de données, qui est actuellement recherchée par l'équipe.

L'équipe GraphIK travaille en particulier sur l'interrogation de bases de connaissances, notamment sur l'accès aux données basé sur une ontologie[15], que l'on peut voir comme de l'interrogation de données en monde ouvert sous contraintes expressives. Les membres de GraphIK ont par exemple exploré la frontière de décidabilité pour l'interrogation sous contraintes transitives [BBM+15], ce qui s'apparente au raisonnement sous des relations d'ordre (Axe 2). Certains membres, notamment Meghyn Bienvenu et Madalina Croitoru, étudient également la gestion de données incohérentes, avec notamment des travaux sur l'explication de résultats [BBG], en lien direct avec la provenance, ou sur l'exploitation de retours utilisateurs. GraphIK s'intéresse aussi aux données incertaines et probabilistes, notamment dans le cadre de ses collaborations avec le centre INRA à Montpellier.

14. En anglais, *linked data*
15. En anglais, *ontology-based data access*

## Auto-références


[AAM+16] Antoine Amarilli, Yael Amsterdamer, Tova Milo et Pierre Senellart. "Top-$k$ Queries on Unknown Values under Order Constraints". Preprint : https://a3nm.net/publications/amarilli2016top.pdf. 2016.

[AAM14a] Antoine Amarilli, Yael Amsterdamer et Tova Milo. "On the Complexity of Mining Itemsets from the Crowd Using Taxonomies". In : *Proc. ICDT*. 2014, p. 15–25. URL : https://arxiv.org/abs/1312.3248.

[AAM14b] Antoine Amarilli, Yael Amsterdamer et Tova Milo. "Uncertainty in Crowd Data Sourcing Under Structural Constraints". In : *Proc. UnCrowd*. T. 8505. LNCS. Springer Berlin Heidelberg, 2014, p. 351–359. URL : https://arxiv.org/abs/1403.0783.

[AB15a] Antoine Amarilli et Michael Benedikt. "Combining Existential Rules and Description Logics". In : *Proc. IJCAI*. AAAI Press, 2015, p. 2691–2697. URL : https://arxiv.org/abs/1505.00326.

[AB15b] Antoine Amarilli et Michael Benedikt. "Finite Open-World Query Answering with Number Restrictions". In : *Proc. LICS*. 2015, p. 305–316. URL : https://arxiv.org/abs/1505.04216.

[ABD+16] Antoine Amarilli, Lamine M. Ba, Daniel Deutch et Pierre Senellart. "Possible and Certain Answers for Queries over Order-Incomplete Data". Preprint : https://a3nm.net/publications/amarilli2016possible.pdf. 2016.

[ABS15] Antoine Amarilli, Pierre Bourhis et Pierre Senellart. "Provenance Circuits for Trees and Treelike Instances". In : *Proc. ICALP*. T. 9135. LNCS. Springer Berlin Heidelberg, 2015, p. 56–68. URL : https://arxiv.org/abs/1511.08723.

[ABS16] Antoine Amarilli, Pierre Bourhis et Pierre Senellart. "Tractable Lineages on Treelike Instances : Limits and Extensions". In : *Proc. PODS*. To appear. 2016. URL : https://a3nm.net/publications/amarilli2016tractable.pdf.


## Références


[AAP+13] Azza Abouzied, Dana Angluin, Christos Papadimitriou, Joseph M. Hellerstein et Avi Silberschatz. "Learning and Verifying Quantified Boolean Queries by Example". In : *Proc. PODS*. 2013.

[AHV95] Serge Abiteboul, Richard Hull et Victor Vianu. *Foundations of Databases*. Addison-Wesley, 1995.

[ANB98] Hajnal Andréka, István Németi et Johan van Benthem. "Modal Languages and Bounded Fragments of Predicate Logic". In : *J. Philosophical Logic* 27.3 (1998).

[Baa03] Franz Baader. *The Description Logic Handbook : Theory, Implementation, and Applications*. Cambridge University Press, 2003.

[BBG] Meghyn Bienvenu, Camille Bourgaux et François Goasdoué. "Explaining Inconsistency-Tolerant Query Answering over Description Logic Knowledge Bases". Accepté à AAAI 2016. URL : http://www.lirmm.fr/~meghyn/papers/BieBouGoa-AAAI16.pdf.

[BBM+15] Jean-François Baget, Meghyn Bienvenu, Marie-Laure Mugnier et Swan Rocher. "Combining Existential Rules and Transitivity : Next Steps". In : *IJCAI*. 2015.

[BCK+] Vince Barany, Balder ten Cate, Benny Kimelfeld, Dan Olteanu et Zografoula Vagena. "Declarative Statistical Modeling with Datalog". Accepté à ICDT 2016. URL : http://arxiv.org/abs/1412.2221.

[BCR08] Alexander Borgida, Diego Calvanese et Mariano Rodriguez-Muro. "Explanation in the DL-Lite Family of Description Logics". In : *Proc. OTM*. 2008.

[BHP+11] Piero A. Bonatti, Aidan Hogan, Axel Polleres et Luigi Sauro. "Robust and Scalable Linked Data Reasoning Incorporating Provenance and Trust Annotations". In : *Web Semantics : Science, Services and Agents on the World Wide Web* 9.2 (2011).

[Bin15] Bing. *Marking Up Your Site with Structured Data*. https://www.bing.com/webmaster/help/marking-up-your-site-with-structured-data-3a93e731. 2015.

[BKS13] Joanna Biega, Erdal Kuzey et Fabian M Suchanek. "Inside YAGO2s : A Transparent Information Extraction Architecture". In : *Proc. WWW*. 2013.

[BKT01] Peter Buneman, Sanjeev Khanna et Wang-Chiew Tan. "Why and Where : A Characterization of Data Provenance". In : *Proc. ICDT*. 2001.

[BLK+09] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak et Sebastian Hellmann. "DBpedia - A Crystallization Point for the Web of Data". In : *J. Web Semantics* 7.3 (2009).

[BLM+09] Jean-François Baget, Michel Leclère, Marie-Laure Mugnier et Eric Salvat. "Extending Decidable Cases for Rules with Existential Variables". In : *Proc. IJCAI*. 2009.

[BLM10] Jean-François Baget, Michel Leclère et Marie-Laure Mugnier. "Walking the Decidability Line for Rules with Existential Variables". In : *Proc. KR*. 2010.

[CDG+07] H. Comon, M. Dauchet, R. Gilleron, C. Löding, F. Jacquemard, D. Lugiez, S. Tison et M. Tommasi. *Tree Automata : Techniques and Applications*. `http://www.grappa.univ-lille3.fr/tata`. 2007.

[CGL12] Andrea Calì, Georg Gottlob et Thomas Lukasiewicz. "A General Datalog-Based Framework for Tractable Query Answering over Ontologies". In : *Web Semantics : Science, Services and Agents on the World Wide Web* 14 (2012).

[CLR03a] Andrea Calì, Domenico Lembo et Riccardo Rosati. "On the Decidability and Complexity of Query Answering over Inconsistent and Incomplete Databases". In : *Proc. PODS*. 2003.

[CLR03b] Andrea Calì, Domenico Lembo et Riccardo Rosati. "Query Rewriting and Answering under Constraints in Data Integration Systems". In : *Proc. IJCAI*. 2003.

[DF08] Susan B. Davidson et Juliana Freire. "Provenance and Scientific Workflows : Challenges and Opportunities". In : *Proc. SIGMOD*. 2008.

[dFL08] Claudia d'Amato, Nicola Fanizzi et Thomas Lukasiewicz. "Tractable Reasoning with Bayesian Description Logics". In : *Proc. SUM*. 2008.

[DGM+15] Xin Luna Dong, Evgeniy Gabrilovich, Kevin Murphy, Van Dang, Wilko Horn, Camillo Lugaresi, Shaohua Sun et Wei Zhang. "Knowledge-Based Trust : Estimating the Trustworthiness of Web Sources". In : *PVLDB* 8.9 (2015).

[DHS12] Mariangiola Dezani-Ciancaglini, Ross Horne et Vladimiro Sassone. "Tracing Where and Who Provenance in Linked Data : a Calculus". In : *TCS* 464 (2012).

[DHY07] Xin Dong, Alon Y. Halevy et Cong Yu. "Data Integration with Uncertainty". In : *Proc. VLDB*. 2007.

[DMF06] Paulo Pinheiro Da Silva, Deborah L. McGuinness et Richard Fikes. "A Proof Markup Language for Semantic Web Services". In : *Information Systems* 31.4 (2006).

[DMR+14] Daniel Deutch, Tova Milo, Sudeepa Roy et Val Tannen. "Circuits for Datalog Provenance." In : *Proc. ICDT*. 2014.

[DS07] Nilesh Dalvi et Dan Suciu. "Efficient Query Evaluation on Probabilistic Databases". In : *VLDBJ* 16.4 (2007).

[EOS+12] Thomas Eiter, Magdalena Ortiz, Mantas Simkus, Trung-Kien Tran et Guohui Xiao. "Query Rewriting for Horn-SHIQ Plus Rules". In : *Proc. AAAI*. 2012.

[ES07] Jérôme Euzenat et Pavel Shvaiko. *Ontology Matching*. Springer, 2007.

[Eta15] Etalab. *Plateforme ouverte des données publiques françaises*. `https://www.data.gouv.fr/`. 2015.

[EY09] Kousha Etessami et Mihalis Yannakakis. "Recursive Markov Chains, Stochastic Grammars, and Monotone Systems of Nonlinear Equations". In : *J. ACM* 56.1 (2009).

[Fac14] Facebook. *The Open Graph Protocol*. `http://ogp.me/`. 2014.

[FFP+09] Giorgos Flouris, Irini Fundulaki, Panagiotis Pediaditis, Yannis Theoharis et Vassilis Christophides. "Coloring RDF Triples to Capture Provenance". In : *Proc. ISWC*. 2009.

[FKS+08] Juliana Freire, David Koop, Emanuele Santos et Cláudio T Silva. "Provenance for Computational Tasks : A Survey". In : *Computing in Science & Engineering* 10.3 (2008).

[Fra13] OpenStreetMap France. *Achèvement du tracé collaboratif des limites communales françaises dans OpenStreetMap*. `http://www.openstreetmap.fr/36680-communes`. 2013.

[GKC+13] Floris Geerts, Grigoris Karvounarakis, Vassilis Christophides et Irini Fundulaki. "Algebraic Structures for Capturing the Provenance of SPARQL Queries". In : *Proc. ICDT*. 2013.

[GKT07] Todd J. Green, Grigoris Karvounarakis et Val Tannen. "Provenance Semirings". In : *Proc. PODS*. 2007.

[Goo] Google. *Google Scholar Help : Inclusion Guidelines for Webmasters*. https://scholar.google.fr/intl/en/scholar/inclusion.html#indexing.

[Goo15] Google. *Promote Your Content with Structured Data Markup*. https://developers.google.com/structured-data/. 2015.

[GPT13] Georg Gottlob, Andreas Pieris et Lidia Tendera. "Querying the Guarded Fragment with Transitivity". In : *Proc. ICALP*. 2013.

[GPW10] Georg Gottlob, Reinhard Pichler et Fang Wei. "Monadic Datalog over Finite Structures of Bounded Treewidth". In : *TOCL* 12.1 (2010).

[GS15] Wolfgang Gatterbauer et Dan Suciu. "Approximate Lifted Inference with Probabilistic Databases". In : *PVLDB* 8.5 (2015).

[GTH+13] Luis Antonio Galárraga, Christina Teflioudi, Katja Hose et Fabian Suchanek. "AMIE : Association Rule Mining under Incomplete Evidence in Ontological Knowledge Bases". In : *Proc. WWW*. 2013.

[HW08] Mordechai Haklay et Patrick Weber. "OpenStreetMap : User-Generated Street Maps". In : *Pervasive Computing* 7.4 (2008).

[ICD+14] Ioana Ileana, Bogdan Cautis, Alin Deutsch et Yannis Katsis. "Complete yet Practical Search for Minimal Query Reformulations under Constraints". In : *Proc. SIGMOD*. 2014.

[JK84] David S. Johnson et Anthony C. Klug. "Testing Containment of Conjunctive Queries under Functional and Inclusion Dependencies". In : *JCSS* 28.1 (1984).

[KG12] Grigoris Karvounarakis et Todd J Green. "Semiring-Annotated Data : Queries and Provenance". In : *ACM SIGMOD Record* 41.3 (2012).

[KO08] Christoph Koch et Dan Olteanu. "Conditioning Probabilistic Databases". In : *PVLDB* 1.1 (2008).

[KPS+05] Aditya Kalyanpur, Bijan Parsia, Evren Sirin et James Hendler. "Debugging Unsatisfiable Classes in OWL Ontologies". In : *Web Semantics : Science, Services and Agents on the World Wide Web* 3.4 (2005).

[LHK10] Jure Leskovec, Daniel Huttenlocher et Jon Kleinberg. "Predicting Positive and Negative Links in Online Social Networks". In : *Proc. WWW*. 2010.

[LS08] Thomas Lukasiewicz et Umberto Straccia. "Managing Uncertainty and Vagueness in Description Logics for the Semantic Web". In : *Web Semantics : Science, Services and Agents on the World Wide Web* 6.4 (2008).

[LVP+14] Thomas Lukasiewicz, Maria Vanina Martinez, Livia Predoiu et Gerardo I. Simari. "Information Integration with Provenance on the Semantic Web via Probabilistic Datalog+/–". In : *Uncertainty Reasoning for the Semantic Web III*. Springer, 2014.

[MB12] Hannes Mühleisen et Christian Bizer. "Web Data Commons-Extracting Structured Data from Two Large Web Corpora". In : *Proc. LDOW* 937 (2012).

[MCF+11] Luc Moreau, Ben Clifford, Juliana Freire, Joe Futrelle, Yolanda Gil, Paul Groth, Natalia Kwasnikowska, Simon Miles, Paolo Missier, Jim Myers, Beth Plale, Yogesh Simmhan, Eric Stephan et Jan Van den Bussche. "The Open Provenance Model Core Specification (v1.1)". In : *Future Generation Computer Systems* 27.6 (2011).

[MCS14] Silviu Maniu, Reynold Cheng et Pierre Senellart. "ProbTree : A Query-Efficient Representation of Probabilistic Graphs". In : *Proc. BUDA*. 2014.

[Mor10] Luc Moreau. "The Foundations for Provenance on the Web". In : *Foundations and Trends in Web Science* 2.2–3 (2010).

[Mot12] Boris Motik. "Representing and Querying Validity Time in RDF and OWL : A Logic-Based Approach". In : *Web Semantics : Science, Services and Agents on the World Wide Web* 12 (2012).

[RKN+15] Simon Razniewski, Flip Korn, Werner Nutt et Divesh Srivastava. "Identifying the Extent of Completeness of Query Answers over Partially Complete Databases". In : *Proc. SIGMOD*. 2015.

[RKT07] Luc De Raedt, Angelika Kimmig et Hannu Toivonen. "ProbLog : A Probabilistic Prolog and Its Application in Link Discovery". In : *Proc. IJCAI*. 2007.

[Sch08] Simon Schenk. "On the Semantics of Trust and Caching in the Semantic Web". In : *Proc. ISWC*. 2008.

[sch15] schema.org. *Home – schema.org*. https://schema.org/. 2015.

[SDS11] Simon Schenk, Renata Queiroz Dividino et Steffen Staab. “Using Provenance to Debug Changing Ontologies”. In : *J. Web Sem.* 9.3 (2011).

[SKW07] Fabian M. Suchanek, Gjergji Kasneci et Gerhard Weikum. “Yago : a Core of Semantic Knowledge”. In : *Proc. WWW*. 2007.

[STI15] STIF. *Portail Open data STIF*. `https://opendata.stif.info/`. 2015.

[TJM+08] Partha Pratim Talukdar, Marie Jacob, Muhammad Salman Mehmood, Koby Crammer, Zachary G. Ives, Fernando Pereira et Sudipto Guha. “Learning to Create Data-Integrating Queries”. In : *PVLDB* 1.1 (2008).

[VK14] Denny Vrandečić et Markus Krötzsch. “Wikidata : a Free Collaborative Knowledgebase”. In : *CACM* 57.10 (2014).

[Wik15a] Wikidata. *Unknown or no values*. `https://www.wikidata.org/wiki/Help:Statements#Unknown_or_no_values`. 2015.

[Wik15b] Wikimedia Foundation. *Other Project Statistics : Wikidata : Monthly counts and Quarterly rankings*. `https://stats.wikimedia.org/wikispecial/EN/TablesWikipediaWIKIDATA.htm`. 2015.

[Wik15c] Wikimedia Foundation. *Wikidata Stats : Overview*. `https://tools.wmflabs.org/wikidata-todo/stats.php`. 2015.

[Zlo75] Moshé M. Zloof. “Query by Example”. In : *AFIPS NCC*. 1975.

[ZLP+12] Antoine Zimmermann, Nuno Lopes, Axel Polleres et Umberto Straccia. “A General Framework for Representing, Reasoning and Querying with Annotated Semantic Web Data”. In : *Web Semantics : Science, Services and Agents on the World Wide Web* 11 (2012).