Candidat *Applicant*
Nom *Last Name* AMARILLI
Prénom *First Name* ANTOINE

DOSSIER DE CANDIDATURE
AU CONCOURS EXTERNE
DE CHARGÉS DE RECHERCHE DE DEUXIÈME CLASSE
POUR L'ANNÉE 2016

APPLICATION TO
A YOUNG GRADUATE SCIENTIST POSITION
FOR YEAR 2016

DÉPÔT DES CANDIDATURES
SUBMITTING APPLICATIONS

Le dossier de candidature doit comprendre :

- Les rapports de thèse ou de doctorat (si disponibles)
- Une copie des derniers titres et diplômes
- Une photographie récente de la candidate / du candidat (facultative)

The application should include:

- *Ph D. dissertation reports (when available)*
- *A copy of most recent titles and diplomas*
- *A recent photography of the applicant (optional)*

La date limite de dépôt des dossiers de candidature est fixée au **15 février 2016**.
Les candidates / candidats doivent remettre leur **dossier en 1 exemplaire** revêtu de la signature **à l'une ou plusieurs des adresses énumérées ci-dessous selon le(s) souhait(s) d'affectation** :

- soit en déposant ce dossier à l'une ou plusieurs de ces adresses avant le **15 février 2016**, 16 heures ;
- soit en l'envoyant par la poste à l'une ou plusieurs de ces adresses avant le **15 février 2016** minuit, le cachet de la poste faisant foi.

The deadline to file an application is **February 15th, 2016**.

Applicants must supply **1 copy of their application** (with the original signature), to **one or several of the following addresses** according to the research centre(s) the applicant wishes to be assigned to:

- either by depositing this application in person at one or several of these addresses before 4:00 PM, **February 15th, 2016**;
- or by sending this application by mail, postmarked by midnight **February 15th, 2016**, to one or several of these addresses.

Adresses/Addresses :

- Service des ressources humaines du centre de recherche Inria Bordeaux – Sud-Ouest,
  200, avenue de la Vieille Tour
  33405 TALENCE Cedex
- Service des ressources humaines du centre de recherche Inria Grenoble - Rhône-Alpes,
  Inovallée - 655, avenue de l'Europe, Montbonnot
  38334 SAINT-ISMIER Cedex
- Service des ressources humaines du centre de recherche Inria Lille - Nord Europe,
  Parc Scientifique de la Haute Borne
  40, avenue Halley - Bât. A, Park Plaza
  59650 VILLENEUVE D'ASCQ
- Service des ressources humaines du centre de recherche Inria Nancy - Grand Est,
  Technopôle de Nancy Brabois, 615, rue du Jardin Botanique, BP 101,
  54602 VILLERS-LES-NANCY
- Service des ressources humaines du centre de recherche Inria de Paris,
  2 rue Simone Iff,
  75012 Paris
- Service des ressources humaines du centre de recherche Inria Rennes - Bretagne Atlantique,
  Campus universitaire de Beaulieu,
  35042 RENNES Cedex
- Service des ressources humaines du centre de recherche Inria Saclay - Île-de-France,
  1 rue Honoré d'Estienne d'Orves, Bâtiment Alan Turing, Campus de l'École Polytechnique
  91120 PALAISEAU
- Service des ressources humaines du centre de recherche Inria Sophia Antipolis - Méditerranée,
  2004, route des Lucioles, BP 93,
  06902 SOPHIA ANTIPOLIS

Attention/*Warning*:

Dans l'état actuel de la réglementation française, **seul le dossier original signé constitue le document officiel de candidature**[1].
According to present French regulations, ***the original application with the applicant's signature is considered as the sole official application document***[1].

Transmission du dossier de candidature par courrier numérique/
Transmitting the application packet via e-mail

Il est demandé à la candidate / au candidat d'**envoyer** le dossier de candidature[2] **par courrier numérique** (formulaires 1 à 7 dans l'ordre), **en un seul fichier**. Ce fichier au format PDF sera enregistré sous le nom de la candidate / du candidat (`nom-prenom.pdf` ; exemple : `dupond-jean.pdf`).

Applicants are asked to ***send a digital version***[2] *of the application packet, (with forms 1 to 7 in this order),* ***in a single file****. This file in PDF format is sent under the name of the applicant (*`lastname-firstname.pdf`*; for example* `smith-john.pdf`*).*

Ce document doit être envoyé à l'une ou plusieurs des adresses énumérées ci-dessous selon le(s) souhait(s) d'affectation / *This document should be sent to one or several of the following addresses according to the research centre(s) the applicant wishes to be assigned to:*

`cr2-2016-bordeaux@inria.fr`	pour une affectation au centre de recherche Inria Bordeaux – Sud-Ouest *for an assignment in the Inria Bordeaux – Sud-Ouest research centre*
`cr2-2016-grenoble@inria.fr`	pour une affectation au centre de recherche Inria Grenoble – Rhône-Alpes *for an assignment in the Inria Grenoble – Rhône-Alpes research centre*
`cr2-2016-lille@inria.fr`	pour une affectation au centre de recherche Inria Lille – Nord Europe *for an assignment in the Inria Lille – Nord Europe research centre*
`cr2-2016-nancy@inria.fr`	pour une affectation au centre de recherche Inria Nancy – Grand Est *for an assignment in the Inria Nancy – Grand Est research centre*
`cr2-2016-paris@inria.fr`	pour une affectation au centre de recherche Inria de Paris *for an assignment in the Inria Paris research centre*
`cr2-2016-rennes@inria.fr`	pour une affectation au centre de recherche Inria Rennes – Bretagne Atlantique *for an assignment in the Inria Rennes – Bretagne Atlantique research centre*
`cr2-2016-saclay@inria.fr`	pour une affectation au centre de recherche Inria Saclay – Île-de-France *for an assignment in the Inria Saclay – Île-de-France research centre*
`cr2-2016-sophia@inria.fr`	pour une affectation au centre de recherche Inria Sophia-Antipolis – Méditerranée *for an assignment in the Inria Sophia-Antipolis – Méditerranée research centre*

Un accusé de réception sera envoyé par mail à la candidate/au candidat dans un délai maximum de 7 jours ouvrés après réception du dossier de candidature.
An acknowledgment of receipt will be sent by e-mail to the applicant within a maximum of 7 working days, after reception of the application file.

[1] Les informations fournies par la candidate / le candidat feront l'objet d'un traitement informatisé, et les listes nominatives des candidates/candidats admis à concourir, présélectionnés, admissibles et admis au concours seront accessibles sur le serveur web d'Inria. Le droit d'accès prévu par l'article 34 de la loi n°78-17 du 6 janvier 1978 modifiée relative à l'informatique, aux fichiers et aux libertés (communication et rectification des données concernant les candidates / candidats) s'exerce auprès de la Direction des ressources humaines d'Inria.

[1] *The data provided in your application will be data processed. The name lists of the selected applicants will be posted on Inria web site. The access right as stated in art. 34 of the law N°78.17, January 6th 1978, modified, related to data processing, files and liberty (communication and correction of the data related to your application) is filed to Inria's Human Resources Department.*

[2] Ce document transmis par courrier numérique sera utilisé pour faciliter le travail des jurys du concours.

[2] *This document sent by e-mail will be used by the committees involved in the competitive selection process.*

Formulaire 1 / *Form 1*

DÉCLARATION DE CANDIDATURE
STATEMENT OF INTENT TO APPLY

Je soussigné(e)[1] / *I undersigned*[1] AMARILLI ANTOINE déclare présenter ma candidature au(x) concour(s) de recrutement de Chargés de recherche de deuxième classe d'Inria pour l'année 2016 / *hereby declare that I apply for the 2016 competitive selection(s) for Inria young graduate scientist (Chargés de recherche de deuxième classe) positions.*
(Cocher au moins une case / *Check at least one box)*

☐ Concours n° 1 :
Affectation : Centre de recherche Inria Bordeaux - Sud-Ouest
☐ Concours n° 2 :
Affectation : Centre de recherche Inria Grenoble - Rhône-Alpes
☒ Concours n° 3 :
Affectation : Centre de recherche Inria Lille - Nord-Europe
☐ Concours n° 4 :
Affectation : Centre de recherche Inria Nancy - Grand-Est
☐ Concours n° 5 :
Affectation : Centre de recherche Inria de Paris
☐ Concours n° 6 :
Affectation : Centre de recherche Inria Rennes - Bretagne-Atlantique
☒ Concours n° 7 :
Affectation : Centre de recherche Inria Saclay - Île-de-France
☒ Concours n° 8 :
Affectation : Centre de recherche Inria Sophia-Antipolis - Mediterranee

Mon programme de recherche s'intitule / *Title of my research program* :

Raisonner avec la provenance sur les données du Web

Sujet de recherche ciblé (optionnel) / *Research topic I apply for (optional):*

. . . . . .

En cas de réussite au(x) concours je demande à être affecté(e) au sein des équipes ou équipes-projet(s) suivantes[2] / *If I am recruited by Inria I wish to be assigned to the following teams or project-teams*[2]:

- INRIA Lille – Nord-Europe, équipe LINKS;
- INRIA Saclay – Île-de-France, équipe DAHU;
- INRIA Sophia-Antipolis – Méditerranée, équipe GraphIK.

[1] Ecrire en lettres capitales

[1] *Please print.*

[2] Les Chargés de recherche de deuxième classe d'Inria sont recrutés au sein de l'une des équipes-projets Inria existantes (ou en cours de création au moment du concours). C'est pourquoi il est demandé aux candidates/candidats d'indiquer la ou les équipes-projets auxquelles ils souhaitent être rattachés en cas de recrutement. Pour chaque équipe-projet mentionnée, indiquer le centre de recherche considéré ; si la candidate/le candidat postule au sein d'une équipe-projet localisée dans deux centres de recherche, il doit mentionner le ou les centres de recherche choisi(s). Voir la liste des équipes-projets d'Inria sur `http://www.inria.fr/recherche/equipes/listes/index.fr.html`. Dans le cadre des souhaits émis par la candidate/ le candidat, la direction d'Inria se réserve le droit de déterminer le centre de recherche d'accueil.

[2] *Inria young graduate scientist(Chargés de recherche de deuxième classe) are recruited within one of the existing research project-teams (or in one of the research project-teams being currently under creation). The applicant is asked to indicate the research project-team(s) he or she wishes to be assigned to. For each research project-team mentioned, indicate the research centre. If the applicant is applying to a research project-team based in two research centres, the chosen research centre(s) must be mentioned. See the list of Inria research projects-teams on* `http://www.inria.fr/recherche/equipes/listes/index.en.html`. *Inside the wishes expressed by the applicant, Inria's management reserves the right to determine the research centre assigned.*

Avez-vous une reconnaissance administrative de travailleur handicapé ? *Do you hold an administrative certificate as a worker with a disability ?*	□
Si oui, souhaitez-vous déposer une demande d'aménagement des épreuves ? *If so, do you need any adjustments for the tests?*	□

Avant de rédiger leur programme de recherche (formulaire 5), les candidates / candidats sont fortement invitées / invités à prendre contact avec les responsables des équipes ou équipes-projet(s) dans lesquelles elles / ils postulent. / *Before writing their research program (form 5), the applicants are strongly encouraged to contact the team or project-team leaders concerned by their applications.*

J'ai pris connaissance des conditions requises pour concourir[3], et je certifie sur l'honneur l'exactitude des renseignements fournis dans ce dossier / *I am aware of the conditions required*[3] *for the consideration of my application and I certify that the information I have supplied is true and correct.*

À/*City* Paris, le/*Date* December 22, 2016
Signature

[3] Voir la brochure d'information / *See the information booklet:*
http://www.inria.fr/medias/recrutement-metiers/pdf/informations-generales-concours-chercheurs-2016.

Formulaire 2 / *Form 2*

CURRICULUM VITÆ DÉTAILLÉ
DETAILED CURRICULUM VITÆ

Nom / *Last Name*: AMARILLI Prénom / *First Name*: ANTOINE
Date et lieu de naissance / *Date and place of birth*: 07/02/1990, Colmar
Nationalité / *Citizenship*: Sexe / *Sex*: ☐F ☒M Âge / *Age*: 26years
Adresse postale / *Mailing address*: 42, Street Name
12345 City
N° de téléphone / *Telephone*: (+33) 01 23 45 67 89
Adresse électronique / *E-mail*: a3nm@a3nm.net
Page Web personnelle / *Web page*: https://a3nm.net/
Centre de recherche Inria (si applicable) / *Inria Research Centre (if applicable)*:
Équipe-projet de recherche (si applicable) / *Project-team (if applicable)*:

Ce document suivra obligatoirement le plan indiqué ci-dessous (conserver la numération des sections même si certaines d'entre elles restent vides). / *This document should follow the guidelines given below (adhere to the order of the sections below, even if some of them are non-applicable).*

1) Diplômes / *Diplomas*

Doctorat(s) / *Ph.D.(s)*
Intitulé / *Title* : Tirer parti de la structure des données incertaines
Date de soutenance / *Date of the defense of the Ph.D.* : Soutenance prévue le 14 mars 2016 (attestation jointe)
Établissement ayant délivré la thèse / *Granting institution* : Télécom ParisTech
Entité d'accueil (laboratoire, équipe, etc.) pour la préparation de la thèse[1] / *Host entity (laboratory, team, etc.) for the preparation of the Ph.D.*[1] : CNRS LTCI, équipe DBWeb

Master ou équivalent / *Master's or equivalent*
Intitulé / *Title* : Master Parisien de Recherche en Informatique
Date / *Date* : 2011–2012
Établissement ayant délivré le diplôme / *Granting institution* : École normale supérieure (Paris)
Organisme où s'est déroulé le stage / *Institution where the training course took place* : Télécom ParisTech

Autres diplômes / *Other diplomas*:

- .....
- .....

2) Parcours Professionnel / *Professionnal history*

2.1) Situation professionnelle actuelle / *Current professionnal status*

Statut et fonction[2] / *Position and statute*[2]: Doctorant
Etablissement (ville - pays) / *Institution (city -country)*: Télécom ParisTech
Date d'entrée en fonction / *Start*: 01/09/2013
[ ] Sans emploi / *Without employment*

[1] Dans le cas où la thèse s'est déroulée au sein d'une équipe-projet Inria, veuillez indiquer le centre de recherche.
[1] *If the thesis took place within an Inria project-team, do please indicate the research centre.*

2.2) Expériences professionnelles antérieures / *Previous professionnal experiences*

Dates début / *Start*	Dates fin / *End*	Établissements / *Institutions*	Fonctions et statuts[2] / *Positions and status*[2]
01/09/2013	31/08/2016	Télécom ParisTech	Doctorant (allocation ENS)
01/09/2009	31/08/2013	École normale supérieure	Normalien (fonctionnaire-stagiaire)

Nombre d'années d'exercice des métiers de la recherche après la thèse / *Number of years of professionnal research experience after thesis*: 0

3) Prix et distinctions / *Prizes and awards*

- Première place au concours de programmation Google Hash Code en 2015.
- Médaille d'argent au concours de programmation ACM ICPC SWERC en 2010 et 2011.
- Première place au concours national d'informatique Prologin en 2008, seconde place en 2010.

4) Encadrement d'activités de recherche / *Supervision of research activities*

5) Responsabilités collectives / *Responsibilities*

J'ai relu ou co-relu des articles pour les revues et conférences suivantes :

- *VLDB Journal*, 2016
- Principles of Database Systems (PODS), 2016
- *Journal of Logic and Computation*, 2015
- *Review of Symbolic Logic*, 2015
- International Conference on Data Engineering (ICDE), 2015
- *Distributed and Parallel Databases*, 2014
- Latin American Theoretical Informatics Symposium (LATIN), 2014
- ACM Special Interest Group on Management of Data (SIGMOD), 2013

Je ferai partie du comité de programme de la conférence Scalable Uncertainty Management (SUM) 2016.

6) Collaborations, mobilité / *Collaborations, visits*

Séjours à l'étranger:

Tel Aviv University (2012–2013, 3 mois). Premier stage à l'étranger effectué pendant ma quatrième année à l'École normale supérieure. J'ai travaillé avec Tova Milo et Yael Amsterdamer sur l'interrogation de la foule (*crowdsourcing*). Notre travail a été publié à **ICDT'14** [5].

University of Oxford (2013, 5 mois). Second stage à l'étranger effectué pendant ma quatrième année à l'école normale supérieure. J'ai travaillé avec Michael Benedikt sur la réponse aux requêtes en monde ouvert. Notre travail a été publié à **IJCAI'15** [2] et **LICS'15** [4].

[2]Indiquer avec précision chaque situation statutaire. Par exemple : pour une situation d'agent titulaire de la fonction publique, préciser le corps et le grade de rattachement, pour une situation de salarié du secteur privé ou d'agent non titulaire d'un établissement public, préciser la nature du contrat salarial, etc.

[2]*For each position, indicate grade or rank. For example, for a tenured civil service position, indicate the branch and rank, for a private sector position or non-tenured position in a public institution, indicate the nature of the work contract, etc.*

Google New York (2011, 4 mois). J'ai effectué un stage de M1 à Google New York sur l'implémentation de techniques de reconnaissance vocale avec des transducteurs à états finis pondérés, pour des méthodes de décodage par minimisation du risque bayésien (*Minimum Bayes Risk decoding*). Ce travail a été valorisé sous la forme d'un **brevet** [7].

Collaborations en cours:

Pierre Bourhis (équipe LINKS INRIA Lille/CRIStAL). Nous travaillons sur la gestion de données relationnelles probabilistes, en imposant des *hypothèses de structure sur les instances* (formellement, des bornes sur la largeur d'arbre[2]). Nos travaux ont été publiés à **ICALP'15** [3] et dans deux soumissions actuellement en relecture : [20] ainsi qu'une soumission en double aveugle (avec Michael Benedikt et Michael Vanden Boom, voir point suivant).

Michael Benedikt (University of Oxford). Nous étudions le *raisonnement en monde ouvert* sous des contraintes logiques, dans les contextes des logiques de description et des règles existentielles, et dans le contexte des bases de données sous l'hypothèse de finitude, pour démontrer la décidabilité de nouveaux langages de règles. En plus de nos articles publiés à **IJCAI'15** [2] et à **LICS'15** [4], nous avons un article actuellement en soumission (en double aveugle) avec Pierre Bourhis et Michael Vanden Boom (University of Oxford).

Tova Milo et Yael Amsterdamer (Tel Aviv University). Nous avons travaillé sur la *fouille de données à partir de la foule*[3], notamment sur l'identification des ensembles d'objets fréquents[4] et sur l'extrapolation de valeurs numériques manquantes suivant un ordre partiel. Après notre article publié à **ICDT'14** [5], nous avons publié un article de vision au workshop **UnCrowd'14** [10] et avons un autre article actuellement en soumission [22].

Daniel Deutch (Tel Aviv University) et M. Lamine Ba (Qatar Computing Research Institute). Nous développons de nouvelles représentations pour la gestion de bases de données relationnelles avec un *ordre sur les faits*, notamment pour représenter l'incertitude sur cet ordre, et pour étudier le problème des réponses possibles et des réponses certaines. Notre travail est actuellement en soumission [21].

7) Enseignement / *Teaching*

J'ai effectué une mission d'enseignement complète de 64 heures par an pendant mes trois années de doctorat. J'ai enseigné les cours suivants à Télécom ParisTech, dans le cadre de son offre de cours ou des masters associés :

Uncertain Data Management (2015–2016) Cours du **M2 Data and Knowledge**, Université Paris–Saclay. Conception et enseignement du cours avec Silviu Maniu : cours magistraux, TP (12 heures).

Projet de programmation : problèmes pratiques et concours (2012–2016). Cours de **Télécom ParisTech, de niveau master**. Réalisation de supports, responsable d'un groupe (24 heures/an, trois ans).

Théorie des langages (2012–2016). Cours de l'enseignement de tronc commun de 1e année à **Télécom ParisTech**. Responsable d'un groupe : cours magistraux, TDs, TPs (16.5 heures/an, trois ans).

Technologies du Web (2012–2015). Cours du **master COMASIC** (Polytechnique, Télécom ParisTech, ENSTA ParisTech). Conception du cours et enseignement : cours magistraux et TP (12 heures/an, deux ans).

8) Diffusion de l'information scientifique / *Dissemination of scientific knowledge*

9) Eléments divers / *Other relevant information*

[2] En anglais, *bounded treewidth*
[3] En anglais, *crowd data sourcing*
[4] En anglais, *frequent itemsets*

Formulaire 3 / *Form 3*

DESCRIPTION SYNTHÉTIQUE DE L'ACTIVITÉ PASSÉE
SUMMARY OF YOUR PAST ACTIVITY

Nom / *Last name*: AMARILLI Prénom/*First name*: ANTOINE

Ma recherche s'est principalement intéressée aux fondements de la gestion de données *incertaines*, sous différentes formes. La recherche sur les données incertaines vise à étendre les techniques classiques de gestion de données développées pour les bases de données relationnelles, afin de leur permettre de gérer des données plus hétérogènes et moins fiables, notamment issues des nombreuses sources de données du Web, ou obtenues par des techniques comme l'extraction automatique d'information ou l'apprentissage. Je me suis notamment intéressé aux données *probabilistes* et aux représentations de la *provenance* sur les données; au raisonnement sur des données *incomplètes*, sous l'hypothèse du monde ouvert; à la représentation de l'incertitude sur des données issues de la *foule*[1]; et aux données munies d'un ordre incertain ou d'autres *nouvelles structures d'incertitude*.

Données probabilistes et provenance (Fiche 1). J'ai étudié le problème d'évaluer des requêtes sur des bases de données relationnelles probabilistes, où chaque fait est annoté par une probabilité d'être vrai. Ce problème est la généralisation naturelle de l'évaluation classique de requêtes par des systèmes de gestion de données, mais appliqué à des données dont on ne sait pas si elles sont correctes : il est alors intractable dans de nombreuses situations. Nous avons développé une nouvelle approche pour ce problème, qui s'appuie sur des automates d'arbres, pour des requêtes expressives sur des instances quasi-arborescentes[2] [3]. Nous procédons par l'approche *intensionnelle* consistant à calculer d'abord une représentation de la *provenance* de la requête, et ensuite à en déterminer la probabilité : nous généralisons ces résultats à des représentations expressives de provenance à base de semianneaux. Sous certaines hypothèses, on peut montrer que ces résultats sont optimaux, et que l'évaluation probabiliste de certaines requêtes du premier ordre sont toujours infaisables si on ne borne pas la largeur d'arbre des instances d'entrée, quelles que soient les autres restrictions que l'on impose [20].

Données incomplètes (Fiche 2). J'ai étudié la gestion de données incomplètes sous l'angle du problème des *réponses certaines en monde ouvert* : calculer les réponses à une requête qui sont impliquées par les données connues et par des règles logiques imposées sur les données manquantes. Ce problème est étudié par plusieurs communautés (gestion de données, intelligence artificielle, et représentation des connaissances), qui ont proposé plusieurs formalismes (données relationnelles ou graphe) et plusieurs langages de règles (dépendances, règles existentielles, logiques de description). J'ai travaillé à étendre et à combiner ces approches pour concevoir de nouveaux langages logiques expressifs en assurant la décidabilité du problème des réponses certaines. J'ai notamment proposé des langages décidables hybrides intégrant logiques de description et règles existentielles [2], étudié de nouvelles contraintes relationnelles sous l'hypothèse de finitude [4], et étudié des assertions de transitivité ou d'ordre que les formalismes existants ne peuvent pas exprimer (actuellement en soumission en double aveugle).

Données issues de la foule (Fiche 3). J'ai travaillé sur la question de la fouille de données à partir de la foule, pour extraire des données ou des motifs d'intérêt en posant des questions simples à un grand nombre d'utilisateurs. Mon principal travail dans cette direction [5] étudie la recherche d'ensembles d'éléments fréquents[3] par des questions posées à la foule en présence d'une taxonomie sur les éléments. Nous étudions la complexité d'identifier les éléments fréquents, ce qui revient à apprendre une fonction booléenne monotone sur un treillis distributif, en bornant le nombre de questions à poser et la complexité computationnelle d'identifier la prochaine bonne question à poser. Nous avons esquissé ensuite des approches plus générales [10] pour nous intéresser plus récemment au problème de compléter des valeurs numériques manquantes obtenues à partir de la foule [22].

Nouvelles structures pour l'incertitude. Je travaille également à la gestion de données munies d'un *ordre* incertain, par exemple sur des séquences d'événements ou des préférences utilisateur. Nos travaux, actuellement en soumission [21], étudient la recherche de réponses certaines pour l'algèbre relationnelle avec agrégation sur des relations munies d'un ordre partiel, et étudient la complexité de ce problème selon les opérateurs autorisés. J'ai également travaillé sur les documents XML probabilistes, avec une étude de la complexité du problème consistant à déterminer si un document est un monde possible d'une représentation probabiliste [11, 18].

[1] En anglais, *crowdsourcing*
[2] En anglais, *treelike*
[3] En anglais, *frequent itemsets*

Formulaire 4 / *Form 4*

CONTRIBUTIONS MAJEURES
MAJOR CONTRIBUTIONS

Nom / *Last name*: AMARILLI Prénom/*First name*: ANTOINE

Fiche 1 : Restrictions structurelles pour le calcul efficace de probabilités et de provenance

1. Description de la contribution / *Description of the contribution*

La théorie des bases de données relationnelles a récemment étudié diverses généralisations de l'évaluation de requêtes. Pour la gestion de données *probabilistes*, on souhaite calculer la *probabilité* des résultats, en fonction d'une distribution de probabilité sur la base de données d'entrée. Pour la gestion de la *provenance*, on souhaite munir les résultats d'une *annotation symbolique* (par ex., dans un semianneau) indiquant comment ils dépendent des faits d'entrée. J'ai proposé une nouvelle méthode à base d'automates d'arbre pour résoudre ces problèmes sous des hypothèses structurelles sur les données, et montré leur faisabilité en données si l'on limite la largeur d'arbre des bases de données d'entrée, en utilisant les méthodes introduites par Courcelle pour récrire de telles instances à des arbres. J'ai réciproquement montré (sous certaines conditions techniques) que le problème de calcul de probabilités est infaisable sur toute famille de données d'entrée si l'on n'impose pas cette limite.

2. Contribution personnelle de la candidate / du candidat / *Personal contribution of the applicant*

Cette contribution est l'une des directions majeures de mon travail de thèse, et représente une collaboration avec mon directeur de thèse Pierre Senellart et avec Pierre Bourhis (équipe LINKS INRIA Lille/CRIStAL). J'ai travaillé au développement de ces techniques avec eux et je suis l'auteur principal de nos travaux sur ces questions.

3. Originalité et difficulté / *Originality and difficulty*

Le calcul de probabilités ou de provenance sur les instances a été étudié jusqu'à présent par des méthodes inspirées de l'algèbre relationnelle, indépendamment des méthodes logiques pour évaluer des requêtes par compilation vers des automates. Pourtant, ces méthodes logiques s'appliquent à des langages de requêtes plus expressifs et permettent de tirer parti de la structure des données. Nous les avons ainsi généralisées au calcul de la provenance et de la probabilité pour obtenir de nouvelles garanties de faisabilité pour ces tâches.
Pour concevoir notre méthode, il a ainsi fallu développer les fondements d'une nouvelle notion de provenance pour les automates d'arbre, l'appliquer à des représentations nouvelles pour la provenance expressive (à base de circuits arithmétiques), et rattacher ces définitions aux notions classiques de provenance et de calcul probabiliste.
Notre résultat d'infaisabilité s'appuie sur des outils techniques totalement différents : l'extraction de mineurs topologiques dans des graphes de grande largeur d'arbre. Là encore, ces techniques n'avaient pas été appliquées jusqu'à présent hors du cadre traditionnel de l'évaluation de requêtes sans probabilités.

4. Validation et impact / *Validation and impact*

Ma contribution généralise des résultats antérieurs sur l'évaluation efficace de requêtes sur des documents XML probabilistes en les généralisant au cadre théorique des métathéorèmes algorithmiques et des automates d'arbres. Nos travaux sont les premiers à lier ce type de résultats à l'étude de la provenance ou de l'évaluation probabiliste de requêtes, qui s'appliquait surtout jusqu'à présent au contexte relationnel classique. Ces travaux ont déjà donné lieu à un stage de master par Mikaël Monet pour développer un prototype logiciel et utiliser ces techniques en pratique ; nous comptons continuer à le développer en collaboration avec Pierre Senellart et Silviu Maniu (LRI).

5. Diffusion / *Dissemination*

Nos résultats ont été publiés dans un article à **ICALP'15** [3], et dans une nouvelle publication actuellement en soumission [20]. J'ai donné huit exposés depuis 2015 pour présenter ces travaux.

Fiche 2 : Réponse aux requêtes en monde ouvert sous contraintes expressives

1. Description de la contribution / *Description of the contribution*

Le problème de *réponse aux requêtes en monde ouvert* consiste à déterminer les *réponses certaines* à une requête à partir d'une base de données incomplètes et de contraintes logiques que l'on impose sur les données manquantes. Formellement, une réponse est *certaine* si elle est vraie sur toutes les manières de compléter les données qui satisfont les contraintes logiques. Une difficulté importante dans ce contexte consiste à trouver des langages logiques qui sont suffisamment expressifs pour représenter des contraintes utiles en pratique, mais où le problème de réponse aux requêtes est néanmoins décidable.
J'ai étudié ce problème et montré sa décidabilité pour de nouveaux langages expressifs de contraintes, qui combinent plusieurs aspects habituellement étudiés indépendamment : signatures de haute arité (plutôt que des graphes), opérateurs expressifs (comme la disjonction), restrictions de cardinalité (par exemple, exprimer qu'un objet a un unique voisin), et hypothèses de finitude.

2. Contribution personnelle de la candidate / du candidat / *Personal contribution of the applicant*

Cette contribution représente une longue collaboration avec Michael Benedikt (University of Oxford) que j'ai entamée avant ma thèse pendant un séjour de 5 mois à Oxford en 2013, et poursuivie depuis lors. J'ai mené la recherche sur ces thèmes avec lui et suis le principal auteur de nos deux premières publications dans cette direction. Cette collaboration s'est récemment élargie à Pierre Bourhis (équipe LINKS INRIA Lille/CRIStAL) et à Michael Vanden Boom (University of Oxford), pour étendre nos résultats à des contraintes de transitivité et d'ordre, en utilisant de nouvelles méthodes à base d'automates pour un fragment logique gardé expressif avec points fixes à paramètres.

3. Originalité et difficulté / *Originality and difficulty*

Le problème de réponse aux requêtes en monde ouvert est étudié actuellement par diverses communautés pour différents types de langages logiques : les logiques de description (qui travaillent en arité 2), les règles existentielles (qui n'autorisent pas les contraintes de cardinalité ou la disjonction), ou les dépendances en théorie des bases de données (dont la forme est elle aussi limitée). Les langages que nous étudions ont l'originalité de combiner certaines de ces caractéristiques, et permettent de tisser des liens entre les approches étudiées par ces communautés. Nous étudions aussi l'hypothèse de finitude, spécifique aux bases de données, mais pour des langages de dépendances plus expressifs qu'auparavant, qui incluent notamment des contraintes de cardinalité comme les dépendances fonctionnelles.
Nos résultats nécessitent ainsi d'appliquer des techniques logiques récentes (comme certaines formes de dépliement[1] ou la construction de modèles finis universels) en sachant les adapter aux contextes et aux langages étudiés par ces différentes communautés. Ainsi, certains de nos résultats sont très techniques : le résultat principal de [4] découle ainsi de plusieurs dizaines de pages de preuves, dans la version étendue que je prépare actuellement.

4. Validation et impact / *Validation and impact*

Nos nouvelles méthodes permettent de répondre aux requêtes en monde ouvert pour des langages de contraintes expressifs pour lesquels la décidabilité de ce problème n'avait pas été établie. Nos approches hybrides [2] ont ainsi montré que les langages logiques expressifs en arité 2 pouvaient se combiner à certains types de règles existentielles. Dans le contexte des bases de données, nos travaux [4] sous l'hypothèse de finitude sont les premiers à montrer la décidabilité de ce problème pour un langage de contraintes en haute arité qui incluent des restrictions de cardinalité sensibles à l'hypothèse de finitude (i.e., qui ne sont pas finiment contrôlables[2]).

5. Diffusion / *Dissemination*

Nos travaux sur les langages hybrides avec logiques de description et règles existentielles ont été publiés à **IJCAI'15** [2] et nos travaux sur les langages de dépendances dans les bases de données sous hypothèse de finitude ont été publiés à **LICS'15** [4]. Un nouveau travail sur les contraintes de transitivité et d'ordre est actuellement en soumission en double aveugle.

[1] En anglais, *unraveling*
[2] En anglais, *finitely controllable*

Fiche 3 : Incertitude sur les données de la foule[3]

1. Description de la contribution / *Description of the contribution*

L'extraction d'informations à partir de la foule[4] permet d'obtenir ou de compléter des informations structurées qu'il n'est pas possible de recueillir autrement parce qu'elles sont inaccessibles aux machines : notamment, le résultat de tâches qui ne peuvent pas encore être parfaitement traitées par des approches automatiques (classification de produits, traitement de la langue naturelle...) ou des connaissances qui ne sont pas centralisées dans des bases de données (par exemple, des combinaisons d'activités populaires lorsqu'on visite une ville donnée).
Cette contribution décrit mon travail pour développer des fondements théoriques s'appliquant aux techniques de fouille de données[5] dans ce contexte nouveau où on les applique aux utilisateurs de la foule.

2. Contribution personnelle de la candidate / du candidat / *Personal contribution of the applicant*

Au cours d'un séjour de 3 mois à Tel Aviv, je me suis intégré au projet ERC MoDaS ("Mob Data Sourcing") porté par Tova Milo (Tel Aviv University), et j'ai également collaboré avec Yael Amsterdamer. J'ai étudié la question de la fouille de données sur la foule sous une taxonomie d'objets en collaboration avec eux, et j'ai co-écrit nos travaux de recherche sur cette question.
Cette contribution s'est ensuite poursuivie au cours de ma thèse, notamment par une visite de recherche de deux semaines à Tel Aviv.

3. Originalité et difficulté / *Originality and difficulty*

L'application de techniques de fouille de données à la foule pose de nouvelles questions de recherche liées aux méthodes d'accès inhabituelles et à la nature des coûts à optimiser. La plupart des techniques actuelles de fouille de données mesurent en effet leur performance suivant le nombre de lectures de la base de données d'entrée, mesure qui n'a pas de sens dans le contexte de la foule. Lorsque l'on pose des questions à la foule, il faut plutôt rechercher des compromis entre la complexité en termes du nombre de requêtes élémentaires à poser, et la complexité computationnelle pour choisir les bonnes questions à poser.
Les réponses obtenues à partir de la foule sont par ailleurs hautement incertaines (les utilisateurs de la foule ne sont pas fiables) et elles ne représentent qu'une vision incomplète des données à recueillir. Ceci impose de compléter les informations manquantes, et de choisir les questions dont on estime qu'elles peuvent nous donner le plus d'informations utiles dans la suite du processus d'acquisition.

4. Validation et impact / *Validation and impact*

J'ai tout d'abord proposé de nouveaux algorithmes pour la recherche d'ensemble d'éléments fréquents[6], en supposant l'existence d'une taxonomie connue sur les objets, dans ce cadre de la fouille de données appliquée à la foule. J'ai notamment étudié comment utiliser les relations de spécialisation dans la taxonomie et d'inclusion entre les ensembles d'éléments, pour déduire qu'un ensemble d'éléments est nécessairement fréquent (ou peu fréquent) si on sait qu'un autre l'est aussi, et ainsi réduire le nombre de questions à poser. En plus d'algorithmes nouveaux, j'ai également montré des résultats d'infaisabilité en proposant des bornes inférieures pour la complexité de certains de ces problèmes.
Nous avons par la suite étudié des questions similaires pour le problème de compléter des informations numériques incomplètes obtenues à partir de la foule, sous des contraintes d'ordre partiel. Nous avons proposé un algorithme permettant de reconstituer les valeurs manquantes à partir de celles qui sont déjà connues, et montré sa faisabilité pour certains ordres partiels arborescents, et son infaisabilité en général. Ma contribution s'intègre ainsi au volet théorique du projet ERC MoDaS portant sur les données de la foule.

5. Diffusion / *Dissemination*

Notre travail a été d'abord publié à la conférence **ICDT 2014** [5], puis un papier vision présentant des extensions possibles a été présenté à un workshop [10]. La suite de cette collaboration est actuellement en soumission [22].

[4] En anglais, *crowdsourcing*
[5] En anglais, *data mining*
[6] En anglais, *frequent itemsets*

Formulaire 5 / *Form 5*

PROGRAMME DE RECHERCHE
RESEARCH PROGRAM

Nom / *Last name*: AMARILLI Prénom/*First name*: ANTOINE

Équipes-Projets d'affectation souhaitées / *Project-teams assignation wishes*:

- INRIA Lille – Nord-Europe, équipe LINKS;
- INRIA Saclay – Île-de-France, équipe DAHU;
- INRIA Sophia-Antipolis – Méditerranée, équipe GraphIK.

Intitulé du programme de recherche : **Raisonner avec la provenance sur les données du Web**

De nombreuses sources de données structurées se développent aujourd'hui sur le Web : des bases de connaissances comme Wikidata, DBPedia, et YAGO ; des sources spécifiques comme OpenStreetMaps ou les données ouvertes de `data.gouv.fr` ; des annotations sémantiques sur des pages Web écrites dans les langages de Schema.org, Open Graph, etc. L'utilisation de ces données permettrait de nombreuses applications nouvelles : par exemple, apporter des réponses sémantiques à des questions structurées, comme le fait déjà Google avec ses Answer Box, et plus généralement offrir de nouveaux moyens de visualiser et d'intégrer ces sources de données.
Malheureusement, contrairement aux situations classiques en gestion de données, ces sources ne sont pas *fiables* : elles sont souvent créées ou extraites automatiquement par des règles faillibles, ou contribuées directement par les internautes. Les données peuvent ainsi être biaisées, incomplètes, voire périmées. On pourrait pourtant compléter les techniques actuelles de gestion pour ces données en utilisant des signaux liés à leur origine : les sites collaboratifs comme OpenStreetMaps ou Wikidata indiquent par exemple les différentes révisions de leurs données et les utilisateurs qui les ont créées, ainsi que leur *source* : Wikidata contient 35 millions de faits sourcés et OpenStreetMap 120 millions de voies sourcées.
Mon projet de recherche consiste à proposer une approche générale pour raisonner sur les données du Web en intégrant ces informations, en s'appuyant sur les techniques de gestion de la *provenance*. La provenance est une technique permettant de propager des annotations symboliques sur les résultats d'une requête, afin d'indiquer les sources et faits dont ils proviennent et la manière dont ils ont été calculés. La provenance a d'abord été définie pour les bases de données, où elle a fourni une solution générale à de nombreux problèmes auparavant étudiés indépendamment (gestion de politiques de sécurité, maintenance de vues, coûts d'accès, etc.) ; elle a été plus récemment étendue à d'autres domaines comme la gestion de données scientifiques ou les *flux de travaux*[1]. 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.*

Provenance et raisonnement. La principale difficulté pour définir et utiliser la provenance sur les données du Web provient du fait que la réponse aux requêtes sur le Web doit s'effectuer en *monde ouvert* : les données peuvent être incomplètes, et on souhaite évaluer des requêtes en déterminant leurs réponses certaines, en appliquant des techniques de *raisonnement* sous contraintes : logiques de description, règles existentielles, appariements de schémas[2], etc. La provenance, en revanche, a surtout été définie et étudiée pour l'instant dans le contexte de l'évaluation de requêtes relationnelles au sens classique ; c'est ainsi un défi majeur que de la généraliser au raisonnement logique sous des contraintes expressives. Mon projet de recherche consiste en premier lieu à développer une telle définition. Il vise également à étendre les approches actuelles pour le raisonnement, afin de calculer efficacement cette nouvelle notion de provenance ; et étudiera comment représenter cette provenance de façon concise, par exemple dans le formalisme récent des circuits de provenance, ou sous d'autres formes à développer.
Pour définir cette notion de provenance sur les données du Web, il faudrait étudier comment représenter les faits utilisés comme *hypothèses* pour le raisonnement, voire même les *règles logiques* utilisées pour aboutir à une conclusion. On pourrait ainsi *comprendre* et *expliquer* les résultats du raisonnement en monde ouvert : l'utilisateur pourrait remonter aux sources et aux contraintes qui ont été utilisées pour déduire un résultat. Ceci nécessite une notion de provenance qui soit *abstraite* et indépendante de l'application visée, pour pouvoir la *spécialiser* ensuite,

[1] En anglais, *workflows*
[2] En anglais, *schema mappings*

comme pour la provenance relationnelle à base de semianneaux. Ce projet est ambitieux mais toutefois réaliste : sa difficulté peut être modulée selon l'expressivité et le degré de généralité que l'on désire obtenir pour la provenance.

Utilisation qualitative de la provenance. Une fois définie une notion de provenance pour annoter les résultats d'une requête en monde ouvert, il faut pouvoir l'utiliser pour en déduire des jugements sur les résultats. Un premier type de jugements est *qualitatif* : déterminer quel résultat est nécessairement meilleur que les autres résultats, si on sait que certaines sources sont meilleures. Notamment, si l'utilisateur a indiqué qu'une source était plus fiable qu'une autre, qu'on dispose d'un ordre temporel sur les données, ou d'un ordre de fiabilité sur les règles, il faut en déduire comment classer les résultats par ordre de pertinence pour l'utilisateur. On peut chercher à propager de tels jugements de pertinence de manière implicite sur les annotations de provenance, ou même envisager de raisonner sous des contraintes logiques qui peuvent utiliser cette relation d'ordre : "si deux faits sont incompatibles, préférer le plus ancien, sauf si l'utilisateur fait davantage confiance à une des sources qu'à l'autre".
Ce problème s'apparente au raisonnement en monde ouvert sous *relations d'ordre*, qui est un problème théorique délicat car la plupart des langages de contraintes décidables pour le raisonnement en monde ouvert ne permettent pas d'exprimer la transitivité de l'ordre. J'ai récemment commencé à étudier ces questions avec Michael Benedikt, Michael Vanden Boom (University of Oxford) et Pierre Bourhis (équipe LINKS INRIA Lille/CRIStAL), et envisage de poursuivre cette étude et d'en appliquer les résultats à la gestion de la provenance pour le raisonnement sur les données du Web.

Utilisation quantitative de la provenance. En plus des utilisations qualitatives visant à déterminer quels résultats sont plus pertinents que d'autres, la provenance sur un résultat peut également être utilisée de manière *quantitative*. L'utilisation la plus fréquente est d'adopter des modèles *probabilistes*, où les faits que l'on connaît sont annotés par des probabilités (estimées suivant leur source, ou par des techniques d'apprentissage ou de recherche de la vérité[3]). On cherche alors à calculer, pour chaque réponse à la requête, la probabilité qu'elle soit correcte.
Une telle utilisation probabiliste de la provenance pose cependant de nombreux défis de recherche. La définition d'une sémantique précise est déjà délicate, surtout si l'on veut aussi pouvoir indiquer que les *règles de raisonnement* sont également incertaines. Cependant, le principal problème est celui de l'efficacité : l'évaluation probabiliste est généralement hautement infaisable (#P-difficile). J'ai déjà étudié des méthodes pour rétablir la faisabilité [3, 20] dans ce contexte, et je compte m'intéresser à de nouvelles manières de faire cela dans le cas des données du Web, notamment en ayant recours à des techniques d'approximation et d'échantillonnage.

Retours utilisateur et révision. Lorsqu'on dispose d'annotations de provenance sur des résultats de raisonnement sur les données du Web, et que l'on peut les utiliser quantitativement et qualitativement pour estimer la qualité et la pertinence des résultats à partir des données et contraintes initiales, la dernière étape consiste à intégrer les retours que fournissent les utilisateurs. Une application permettrait en effet à un utilisateur d'indiquer qu'un résultat est correct ou incorrect, soit directement, soit par des indices indirects comme le temps passé sur chaque résultat ou les clics effectués. On peut également vouloir utiliser des règles négatives comme des dépendances fonctionnelles pour savoir, par exemple, que deux résultats sont incompatibles et ne peuvent pas être vrais simultanément.
Ainsi, les perspectives à plus long terme de mon projet seraient d'utiliser de telles informations sur les *résultats* pour remonter, à travers la provenance, à des jugements sur les *données initiales*, notamment pour réviser la confiance qu'on leur accorde. Si on prend ainsi en compte la possibilité de solliciter des retours de la part de l'utilisateur, on peut même chercher à étudier quelles questions il sera le plus informatif de poser, ce qui se rapproche du problème d'interrogation de la foule[4] sur lequel j'ai déjà travaillé.

Intégration. Mon projet s'intègre à l'équipe LINKS, qui s'intéresse spécifiquement au thème de la gestion du Web des données[5] et des sources hétérogènes, en utilisant des approches logiques, et en gérant l'incertitude et l'incomplétude sur les données. Les techniques de provenance sur ces données, qui sont l'objet de mon projet, peuvent notamment s'appliquer au problème de l'intégration de données sur le Web, une question centrale étudiée par l'équipe et que l'on peut formuler comme une tâche de raisonnement. Je pourrais ainsi apporter à l'équipe de nouvelles compétences en matière de gestion de données probabilistes et de provenance.
Mon intégration dans LINKS est naturelle, comme le montre notamment ma collaboration actuelle avec Pierre Bourhis, membre de l'équipe : nous avons déjà publié ensemble un travail à **ICALP'15** [3] et soumis deux articles ([20] et un article en double aveugle). De plus, certaines problématiques que j'étudie (notamment celle des retours utilisateur) permettraient des interactions avec d'autres équipes lilloises, en particulier avec MAGNET (notamment Marc Tommasi) pour les questions liées à l'apprentissage.

[3] En anglais, *truth finding*
[4] En anglais, *crowdsourcing*
[5] En anglais, *linked data*

Formulaire 6 / *Form 6*

LETTRES DE RECOMMANDATION
(Coordonnées des personnalités scientifiques sollicitées)[1]
RECOMMENDATION LETTERS
(Names and addresses of professional references)[1]

5 NOMS MAXIMUM / *MAXIMUM 5 NAMES*

Nom de la candidate / du candidat / *Applicant's Last Name*: AMARILLI Prénom/*First name*: ANTOINE

1. **Prof. Firstname Lastname**
   Address details
   Phone: 01 23 45 67 89
   Email: name@example.com
2. Same format as above
3. Same format as above
4. Same format as above
5. Same format as above

[1]Les candidates/candidats solliciteront directement les lettres de recommandations à leurs recommandants (5 au maximum) en leur demandant d'adresser leur courrier directement au Président Directeur Général de l'Institut via l'alias suivant : concours-cr2-2016@inria.fr. Les candidates/candidats transmettront aux recommandants les demandes de lettres de recommandation au moyen du modèle de courrier ci-après. La direction d'Inria demandera aussi un avis au(x) responsable(s) scientifiques(s) des équipes ou des équipes-projets et au(x) directeur(s) de centre(s) de recherche concerné(s) par la candidature.

[1]*Applicants must contact directly the referees (maximum 5 names) to request recommendations letter. These referees have to send in confidence their letters to the President of Inria via the following email: concours-cr2-2016@inria.fr. Applicants must use the model of letter below to send their requests. Inria will also seek advice from the project-team leader(s) and the director(s) of the research centre(s) where the applicant wishes to apply.*

Modèle de lettre :

Suite au dépôt de mon dossier d'inscription au concours de recrutement de Chargés de recherche de deuxième classe au sein d'Inria, je me permets de vous solliciter pour emettre un avis sur ma candidature. À cet effet, je vous saurais gré de bien vouloir adresser une lettre de recommandation avant le **15 février 2016** à l'adresse concours-cr2-2016@inria.fr (à l'attention du Président-Directeur Général d'Inria) sous le format suivant : *Nom-Candidat_recomm_NomReferent.pdf*.

Je vous remercie de bien vouloir porter avis sur mes travaux, mes contributions scientifiques et mes réalisations en matière de transfert technologique mais aussi sur tous les plans qu'il est intéressant de considérer pour la carrière d'un chercheur : son potentiel, son leadership, son ouverture et son dynamisme.

En vous remerciant vivement pour votre collaboration, je vous adresse mes sentiments très cordiaux.

Model of letter :

Following my application to a young tenured scientist position at Inria, I would like to ask you to express your opinion on my candidacy. Would you please send a letter of recommendation before **February 15th, 2016** *to: concours-cr2-2016@inria.fr (to the attention of Inria's CEO) under the format: NameApplicant_recomm_Name-Referent.pdf.*

Please express yourself in a free manner on my work, my scientific results and my technological achievements, and also on any matter that you considered as relevant for a researcher, for example: potential, leadership, open-mindedness and dynamism.

I thank you very much for your kind assistance.

Yours sincerely

Formulaire 7 / *Form 7*

LISTE COMPLÈTE DES CONTRIBUTIONS[1]
COMPLETE LIST OF CONTRIBUTIONS [1]

Nom / *Last name*: AMARILLI Prénom/*First name*: ANTOINE

1) Publications

1.1) Revues internationales / *International journals*

[1] Ruiming Tang, Antoine Amarilli, Pierre Senellart, and Stéphane Bressan. "A Framework for Sampling-Based XML Data Pricing". In: *Transactions on Large-Scale Data and Knowledge-Centered Systems* 24 (2016), pp. 116–138. URL: https://a3nm.net/publications/tang2014framework.pdf.

1.2) Conférences internationales avec comité de lecture / *Reviewed international conferences*

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

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

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

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

[6] Ruiming Tang, Antoine Amarilli, Pierre Senellart, and Stéphane Bressan. "Get a Sample for a Discount". In: *Proc. DEXA*. Vol. 8644. LNCS. Springer International Publishing, 2014, pp. 20–34. URL: https://a3nm.net/publications/tang2014get.pdf.

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

1.3) Livres et chapitres de livres / *Books and book chapters*

1.4) Autres publications (posters, short papers...) / *Other publications (posters, short papers...)*

[7] Antoine Amarilli, Cyril Allauzen, and Mehryar Mohri. *Minimum Bayesian Risk Methods for Automatic Speech Recognition*. United States Patent 9123333. 2015. URL: https://a3nm.net/publications/amarilli2014minimum.pdf.

[8] Aliaksandr Talaika, Joanna Biega, Antoine Amarilli, and Fabian M. Suchanek. "IBEX: Harvesting Entities from the Web Using Unique Identifiers". In: *Proc. WebDB*. ACM, 2015, pp. 13–19. URL: https://arxiv.org/abs/1505.00841.

[9] Antoine Amarilli. "Structurally Tractable Uncertain Data". In: *Proc. PhD Symposium of SIGMOD/PODS*. ACM, 2015, pp. 39–44. URL: https://arxiv.org/abs/1507.04955.

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

[1] Les contributions les plus significatives (publications, logiciels) devront être consultables sur la page web de la candidate / du candidat.
[1] *Most relevant contributions (publications, software) have to be available for consultation via the web page of the applicant.*

[11] Antoine Amarilli. “The Possibility Problem for Probabilistic XML”. In: *Proc. AMW*. 2014. URL: http://ceur-ws.org/Vol-1189/paper_2.pdf.

[12] Marilena Oita, Antoine Amarilli, and Pierre Senellart. “Cross-Fertilizing Deep Web Analysis and Ontology Enrichment”. In: *Proc. VLDS*. 2012, pp. 5–8. URL: http://pierre.senellart.com/publications/oita2012crossfertilizing.pdf.

***Articles invités** (sans comité de lecture) :*

[13] Antoine Amarilli, Silviu Maniu, and Pierre Senellart. “Intensional Data on the Web”. In: *SIGWEB Newsletter* Summer 2015 (2015), 4:1–4:12. URL: https://a3nm.net/publications/amarilli2015intensional.pdf.

[14] Antoine Amarilli, Luis Galárraga, Nicoleta Preda, and Fabian M. Suchanek. “Recent Topics of Research around the YAGO Knowledge Base”. In: *Proc. APWEB*. Vol. 8709. LNCS. Springer International Publishing, 2014, pp. 1–12. URL: https://zenodo.org/record/34912.

[15] Antoine Amarilli, Fabrice Ben Hamouda, Florian Bourse, Robin Morisset, David Naccache, and Pablo Rauzy. “From Rational Number Reconstruction to Set Reconciliation and File Synchronization”. In: *Proc. TGC*. Vol. 8191. LNCS. Springer Berlin Heidelberg, 2013, pp. 1–18. URL: https://zenodo.org/record/33991.

[16] Antoine Amarilli, Sascha Müller, David Naccache, Daniel Page, Pablo Rauzy, and Michael Tunstall. “Can Code Polymorphism Limit Information Leakage?” In: *Proc. WISTP*. Springer-Verlag, 2011, pp. 1–21. URL: https://eprint.iacr.org/2011/099.

[17] Antoine Amarilli, David Naccache, Pablo Rauzy, and Emil Simion. “Can a Program Reverse-Engineer Itself?” In: *Proc. IMACC*. Ed. by Liqun Chen. Vol. 7089. LNCS. Springer Berlin Heidelberg, 2011, pp. 1–9. URL: https://eprint.iacr.org/2011/497.

1.5) Revues nationales / *National journals*

[18] Antoine Amarilli. “Possibility for Probabilistic XML”. In: *Ingénierie des Systèmes d’Information* 20.5 (2015), pp. 53–75. URL: https://arxiv.org/abs/1404.3131.

1.6) Conférences nationales avec comité de lecture / *Reviewed national conferences*

[19] Antoine Amarilli and Pierre Senellart. “On the Connections between Relational and XML Probabilistic Data Models”. In: *Proc. BNCOD*. 2013, pp. 121–134. URL: http://pierre.senellart.com/publications/amarilli2013connections.pdf.

1.7) Rapports de recherche et articles soumis / *Research reports and publications under review*

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

[22] Antoine Amarilli, Yael Amsterdamer, Tova Milo, and Pierre Senellart. “Top-k Queries on Unknown Values under Order Constraints”. Preprint: https://a3nm.net/publications/amarilli2016top.pdf. 2016.

2) Développements technologiques : logiciel ou autre réalisation / *Technology development : software or other realization*

PARIS A-2; SO-4; SM-1; EM-1; SDL-4; OC: DA-4; CD-4; MS-3; TPM-2

Le logiciel d’alignement d’ontologies PARIS, disponible à l’adresse http://webdam.inria.fr/paris/, est un prototype développé pour un travail de recherche antérieur entre Serge Abiteboul, Pierre Senellart, et Fabian Suchanek. PARIS permet de déterminer automatiquement des liens entre de grandes bases de connaissances structurées afin de les intégrer ; il utilise une approche générique qui a été testée sur DBPedia, IMDB et YAGO.

J’ai effectué mon stage de M2 sur PARIS, dont j’ai largement retravaillé le code, en termes d’ingénierie (exécution parallèle, accélération en travaillant en mémoire principale) et en termes d’ajout de nouvelles fonctionnalités. Une de ces nouvelles directions a été présentée comme un papier vision [12]. Nous avons par ailleurs été contactés par plusieurs chercheurs qui ont souhaité utiliser l’implémentation pour comparer leurs résultats à ceux de PARIS.

Décodage MBR pour la reconnaissance vocale A-1; SO-3; SM-1; EM-1; SDL-2; OC: DA-4; CD-4; MS-1; TPM-1
Au cours d'un stage de 4 mois à Google New York, j'ai travaillé à l'expérimentation de nouvelles techniques de décodage pour la reconnaissance vocale sur la plateforme de Google, en utilisant des techniques de minimisation du risque bayésien pour réduire l'erreur moyenne sur les transcriptions. Le prototype utilisait des techniques de transducteurs finis pondérés avec la librairie OpenFST développée entre autres par Google. Il n'a pas été distribué hors de Google pour des raisons de confidentialité, mais a donné lieu à la publication d'un **brevet** [7].

3) Impact socio-économique et transfert / *Socio-economic impact and transfer*