

NNT 2017LYSEI072

THESE de DOCTORAT DE L'UNIVERSITE DE LYON

Préparée au sein de

l'INSA LYON

Ecole Doctorale ED512

Ecole doctorale d'informatique et mathématique de Lyon

Spécialité de doctorat : Informatique

Bilal BERJAWI

Integration of Heterogeneous Data from Multiple Location-Based Services Providers: a Use Case on Tourist Points of Interest

---

Devant le jury composé de :

Devogele, Thomas	Professeur	Université de Tours	Rapporteur
Favetta, Franck	MdC	Université Claude Bernard Lyon 1	Co-dir de thèse
Gaio, Mauro	Professeur	Université de Pau et des Pays de L'Adour	Examinateur
Lamarre, Philippe	Professeur	INSA de Lyon	Examinateur
Lbath, Ahmed	Professeur	Université Grenoble Alpes	Rapporteur
Libourel, Thérèse	Professeur	Université de Montpellier 2	Présidente
Maryvonne, Miquel	MdC (HDR)	INSA de Lyon	Dir de thèse
Zanin, Christine	MdC (HDR)	Université Paris Diderot	Examinatrice

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

Département FEDORA – INSA Lyon - Ecoles Doctorales – Quinquennal 2016-2020

SIGLE	ECOLE DOCTORALE	NOM ET COORDONNEES DU RESPONSABLE
CHIMIE	**CHIMIE DE LYON** **http://www.edchimie-lyon.fr** Sec : Renée EL MELHEM Bat Blaise Pascal 3e etage secretariat@edchimie-lyon.fr Insa : R. GOURDON	M. Stéphane DANIELE Institut de Recherches sur la Catalyse et l'Environnement de Lyon IRCELYON-UMR 5256 Équipe CDFA 2 avenue Albert Einstein 69626 Villeurbanne cedex directeur@edchimie-lyon.fr
E.E.A.	**ELECTRONIQUE, ELECTROTECHNIQUE, AUTOMATIQUE** **http://edeea.ec-lyon.fr** Sec : M.C. HAVGOUDOUKIAN Ecole-Doctorale.eea@ec-lyon.fr	M. Gérard SCORLETTI Ecole Centrale de Lyon 36 avenue Guy de Collongue 69134 ECULLY Tél : 04.72.18 60.97 Fax : 04 78 43 37 17 Gerard.scorletti@ec-lyon.fr
E2M2	**EVOLUTION, ECOSYSTEME, MICROBIOLOGIE, MODELISATION** **http://e2m2.universite-lyon.fr** Sec : Sylvie ROBERJOT Bât Atrium - UCB Lyon 1 04.72.44.83.62 Insa : H. CHARLES secretariat.e2m2@univ-lyon1.fr	M. Fabrice CORDEY CNRS UMR 5276 Lab. de géologie de Lyon Université Claude Bernard Lyon 1 Bât Géode 2 rue Raphaël Dubois 69622 VILLEURBANNE Cédex Tél : 06.07.53.89.13 cordey@univ-lyon1.fr
EDISS	**INTERDISCIPLINAIRE SCIENCES-SANTE** **http://www.ediss-lyon.fr** Sec : Sylvie ROBERJOT Bât Atrium - UCB Lyon 1 04.72.44.83.62 Insa : M. LAGARDE secretariat.ediss@univ-lyon1.fr	Mme Emmanuelle CANET-SOULAS INSERM U1060, CarMeN lab, Univ. Lyon 1 Bâtiment IMBL 11 avenue Jean Capelle INSA de Lyon 696621 Villeurbanne Tél : 04.72.68.49.09 Fax :04 72 68 49 16 Emmanuelle.canet@univ-lyon1.fr
INFOMATHS	**INFORMATIQUE ET MATHEMATIQUES** **http://infomaths.univ-lyon1.fr** Sec :Renée EL MELHEM Bat Blaise Pascal, 3e étage Tél : 04.72. 43. 80. 46 Fax : 04.72.43.16.87 infomaths@univ-lyon1.fr	M. Luca ZAMBONI Bâtiment Braconnier 43 Boulevard du 11 novembre 1918 69622 VILLEURBANNE Cedex Tél :04 26 23 45 52 zamboni@maths.univ-lyon1.fr
Matériaux	**MATERIAUX DE LYON** **http://ed34.universite-lyon.fr** Sec : Marion COMBE Tél:04-72-43-71-70 –Fax : 87.12 Bat. Direction ed.materiaux@insa-lyon.fr	M. Jean-Yves BUFFIERE INSA de Lyon MATEIS Bâtiment Saint Exupéry 7 avenue Jean Capelle 69621 VILLEURBANNE Cedex Tél : 04.72.43 71.70 Fax 04 72 43 85 28 Ed.materiaux@insa-lyon.fr
MEGA	**MECANIQUE, ENERGETIQUE, GENIE CIVIL, ACOUSTIQUE** **http://mega.universite-lyon.fr** Sec : Marion COMBE Tél:04-72-43-71-70 –Fax : 87.12 Bat. Direction mega@insa-lyon.fr	M. Philippe BOISSE INSA de Lyon Laboratoire LAMCOS Bâtiment Jacquard 25 bis avenue Jean Capelle 69621 VILLEURBANNE Cedex Tél : 04.72 .43.71.70 Fax : 04 72 43 72 37 Philippe.boisse@insa-lyon.fr
ScSo	**ScSo*** **http://recherche.univ-lyon2.fr/scso/** Sec : Viviane POLSINELLI Brigitte DUBOIS Insa : J.Y. TOUSSAINT Tél : 04 78 69 72 76 viviane.polsinelli@univ-lyon2.fr	M. Christian MONTES Université Lyon 2 86 rue Pasteur 69365 LYON Cedex 07 Christian.montes@univ-lyon2.fr

*ScSo : Histoire, Géographie, Aménagement, Urbanisme, Archéologie, Science politique, Sociologie, Anthropologie

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

À ma famille

Mon père Walid
Ma mère Aida
Mes frères Khaled et Mohamad
Mes soeurs Nagham et Mahassen

À ma belle épouse Moussina

Et à tous mes amis

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

Remerciements

Cette étude a été réalisée au Laboratoire d'Informatique en Images et Systèmes d' Information (LIRIS UMR 5205) au sein de l'Institut National des Sciences Appliquées de Lyon. Je tiens donc tout d'abord à adresser mes remerciements à Atilla BASKURT et Mohand-Said HACID pour m'avoir accueilli durant mes années de thèse. Celle-ci n'aurait pas été accomplie sans la bourse de projet de recherche UNIMAP et le support financier de Labex d'Intelligences des Mondes Urbains (IMU).

Ce travail n'aurait également pas pu voir le jour sans l'encadrement exemplaire, les inestimables conseils et les encouragements de Maryvonne MIQUEL et Franck FAVETTA, mes directeurs de thèse. Je les remercie pour leurs conseils avisés, leurs excellences scientifiques ainsi que pour leurs remarques intéressantes qui ont contribué à faire avancer et améliorer ce travail. Je remercie également Fabien DUCHATEAU pour ses précieux conseils et son soutien intellectuel qui m'ont permis de réaliser une très bonne expérience professionnelle. Je tiens aussi à remercier Robert LAURINI pour son aide précieuse.

Cette thèse est le fruit d'une collaboration entre les partenaires de projet UNIMAP : l'équipe Bases de Données (BD) du LIRIS, le laboratoire Environnement Ville Société (EVS UMR 5600) et les praticiens des offices de tourisme de Rhône-Alpes, Lyon et Saint-Etienne. Je tiens à remercier Thierry JOLIVEAU, Elisabeth CHESNEAU, Claire CUNTY, Karine FEIGE, Magali CAPELLE et Béatrice MASSON pour leurs apports afin de réussir ce projet.

Je remercie chaleureusement Thomas DEVOGELE et Ahmad LBATH pour l'honneur qu'ils m'ont fait en acceptant de rapporter cette thèse. Je voudrais également remercier Mauro GAIO, Thérèse LIBOUREL et Christine ZANIN pour l'intérêt qu'ils ont porté à ces travaux en acceptant de les examiner.

Mes remerciements s'adressent également à ma famille dont la présence a permis de poursuivre mes études jusqu'à ce jour. Ces remerciements seraient incomplets sans une pensée à mes meilleurs amis.

Mes derniers remerciements vont vers la personne avec laquelle j'ai choisi de partager ma vie, grand merci à mon épouse Moussina d'être toujours présente.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

Résumé

Les services de géolocalisation (LBS) sont impliqués dans de nombreuses applications pour fournir des informations géospatiales pertinentes basées sur une position ou une adresse géographique. La quantité de données géospatiales disponible augmente constamment et constitue des sources d'informations précieuses pour enrichir les applications LBS. Cependant, ces données géospatiales sont souvent incohérentes et contradictoires d'une source à l'autre. Aussi, pensons nous que l'intégration de données géospatiales à partir de plusieurs sources peut améliorer la qualité de l'information offerte aux utilisateurs.

Dans cette thèse, nous nous intéresserons plus spécifiquement aux données représentant les points d'intérêt (POIs) que les touristes peuvent obtenir grâce à des applications LBS. Techniquement, un POI est représenté par une entité géospatiale qui décrit ses informations terminologiques et spatiales. La récupération, l'alignement et la fusion de ces entités géospatiales mènent à plusieurs défis. Nous nous focalisons principalement sur trois principaux défis : (i) traiter les différents schémas et structures des entités, (ii) détecter et fusionner les entités correspondantes issues de multiples sources et (iii) tenir compte de l'incertitude liée aux entités intégrées et proposer leur représentation dans les applications LBS.

Tout d'abord, nous présentons un aperçu technique qui met en évidence les méthodes utilisées par les actuels fournisseurs LBS pour partager leurs POIs ainsi que leurs limites. Ensuite, nous définissons une taxonomie de différences et d'incohérences observées entre les entités qui reprśentent les POIs. Cette taxonomie permet de modéliser et de comprendre comment les données peuvent différer d'une source à l'autre, ce qui nous aide à étudier comment nous devrions les intégrer. En se basant sur cette taxonomie, nous présentons PABench, un benchmark pour l'alignement des entités géospatiales. PABench peut fournir une évaluation précise des différents aspects de la qualité des approches d'alignement d'entités géospatiales et également faciliter la compréhension de leurs capacités et faiblesses quant à l'intégration géospatiale.

En ce qui concerne l'intégration des données, nous nous concentrons sur deux étapes : l'alignement d'entités et la fusion d'entités. Nous proposons l'approche Global Similarity pour l'alignement des entités géospatiales qui utilise à la fois des informations spatiales et terminologiques pour détecter les entités correspondantes. Au préalable notre approche consiste à utiliser une méthode de blocage spatial pour réduire le nombre d'entités potentiellement correspondantes. Ensuite, les entités groupées sont comparées en utilisant des mesures de similarité afin de détecter les paires correspondantes. Pour les attributs spatiaux, nous utilisons une mesure que nous avons définie et comparée à d'autres mesures

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

existantes. Pour les attributs terminologiques, nous utilisons des mesures de similarité issues de la littérature que nous avons sélectionné selon le type de l'attribut. Une fois les entités correspondantes détectées, un algorithme de fusion de données est mis en œuvre pour fusionner les entités correspondantes et pour estimer l'incertitude des valeurs choisies. L'incertitude sera ensuite utilisée pour informer les utilisateurs de l'exactitude des informations qu'ils reçoivent.

Enfin, nous avons étudié la visualisation d'entités fusionnées et de l'incertitude dans des cartes interactives. Nous utilisons des tests cognitifs pour déterminer les variables visuelles à utiliser et les informations à représenter directement et les informations à représenter à la demande. Nous montrons la faisabilité et l'intérêt de notre étude en développant un prototype LBS multifournisseurs et en évaluant notre proposition pour les utilisateurs potentiels.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

Abstract

Location Based Services (LBS) had been involved to deliver relevant geospatial information based on a geographic position or address. The amount of geospatial data is constantly increasing, making it a valuable source of information for enriching LBS applications. However, these geospatial data are highly inconsistent and contradictory from one source to another. We assume that integrating geospatial data from several sources may improve the quality of information offered to users.

In this thesis, we specifically focus on data representing Points of Interest (POIs) that tourists can get through LBS. Technically, a POI is represented by a geospatial entity that describes the terminological and spatial information of the POI. Retrieving, matching and merging such geospatial entities lead to several challenges. We mainly focus on three main challenges including (i) dealing with different schemas and structures of entities, (ii) detecting and merging corresponding entities across multiple sources and (iii) considering the uncertainty of integrated entities and their representation in LBS applications.

First, we represent a technical overview to highlight the limitations and methods used by current LBS providers to share their POIs. Then, we define a taxonomy of observed differences and inconsistencies between the entities that represent the POIs. This taxonomy shows how data may differ from one source to another, which helps us understand how we should integrate them. Based on this taxonomy, we introduce PABench, a benchmark for geospatial entity matching. PABench can provide an accurate evaluation of the different quality aspects of geospatial entity matching approaches, and also facilitate an understanding of their weaknesses and abilities with respect to geospatial integration.

Concerning the data integration, we focus on two steps namely: entity matching and entity merging. We propose a geospatial entity matching approach namely Global Similarity that uses both spatial and terminological information to detect the corresponding entities. Our method uses a spatial blocking method to reduce the number of the potentially corresponding entities. Then, the grouped entities are compared using similarity measures in order to detect the corresponding pairs. We propose a spatial similarity measure and compare it to existing similar measures. We also compared a set of terminological similarity measures in order to select the appropriate measure to compare values of a given attribute. Once corresponding entities are detected, a data fusion algorithm is proposed to merge corresponding entities and to estimate the uncertainty of chosen values. The uncertainty is then used to inform users about the accuracy of the information they receive.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

Finally, we studied the visualization of merged entities in interactive maps. We use cognitive tests to find which visual variables to use and what information to be represented directly and what information to be represented on demand. We proved the feasibility and the benefits of our study by implementing a multi providers LBS prototype and by evaluating our proposal for potentially users.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

Contents

Contents viii

List of Figures xii

List of Tables xv

1 Introduction 1
- 1.1 Location-Based Services . . . 2
- 1.2 Motivation . . . 3
- 1.3 Issues . . . 5
- 1.4 Research Questions . . . 8
- 1.5 Methodology and Contributions . . . 8
- 1.6 Dissertation Outline . . . 11

2 Related Work 12
- 2.1 Characteristics of LBS . . . 13
- 2.2 Data Integration . . . 17
  - 2.2.1 Schema Matching . . . 19
  - 2.2.2 Entity Matching . . . 19
    - 2.2.2.1 Pre-matching . . . 21
    - 2.2.2.2 Matching . . . 22
    - 2.2.2.3 Post-matching . . . 24
  - 2.2.3 Geospatial Entity Matching . . . 24
    - 2.2.3.1 Approaches that use only spatial information . . . 25
    - 2.2.3.2 Approaches that combine terminological and spatial information without machine-learning . . . 30
    - 2.2.3.3 Approaches that combine terminological and spatial information using machine-learning . . . 33
  - 2.2.4 Data Fusion . . . 35
- 2.3 Data Matching Benchmark . . . 37
  - 2.3.1 Datasets . . . 37
  - 2.3.2 Metrics . . . 39
- 2.4 Visualization of Geospatial Integrated Data . . . 41
- 2.5 Conclusion and Positioning . . . 43

3 Taxonomy 46

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

3.1 Preliminary Definitions . . . 47
3.2 Entity Sets Characterization . . . 50
3.3 Taxonomy of Inconsistencies . . . 52
3.3.1 Schema Differences . . . 52
3.3.2 Terminological Inconsistencies . . . 54
3.3.3 Spatial Inconsistencies . . . 57
3.3.4 Entity's Availability . . . 59
3.4 Impact of Taxonomy . . . 60
3.4.1 Impact of Schema's Inconsistencies . . . 61
3.4.2 Impact of Spatial and Terminological Inconsistencies . . . 62
3.4.3 Impact of Availability's Inconsistencies . . . 62
3.5 Combination of Inconsistencies . . . 64
3.5.1 Simple Combinations . . . 64
3.5.2 Complex Combinations . . . 66
3.6 Conclusion . . . 68

4 An Evaluation Benchmark with Real World Data 70
4.1 GeoBench DB - Construction of Characterized Database . . . 71
4.1.1 Overview of GeoBench Aligner . . . 71
4.1.2 Schema Matching . . . 73
4.1.3 POIs' Types Matching . . . 74
4.1.4 Blocking Algorithm . . . 76
4.1.5 Basic Matching Algorithm . . . 76
4.1.6 GeoBench Aligner Prototype - User Interface . . . 78
4.1.7 Experimentation and Population of GeoBench DB . . . 79
4.2 PABench - An Evaluation Benchmark . . . 81
4.2.1 Model of Situations . . . 81
4.2.2 Test Cases and Metrics . . . 83
4.2.3 PABench Extractor . . . 84
4.3 Conclusion . . . 85

5 Matching Geospatial Data 86
5.1 Overview for Geospatial Entity Matching . . . 87
5.2 Normalized-Distance: A Spatial Similarity Measure . . . 88
5.3 Global Similarity: Combining Similarity Measures . . . 95
5.4 Decision Algorithm . . . 99
5.5 Experimental Evaluation Using Real-World Datasets . . . 101
5.5.1 Evaluation and Selection of Terminological Similarity Measures . . . 102
5.5.1.1 Test Cases Evaluation For Terminological Similarity Measures . . . 103
5.5.1.2 Full Datasets Evaluation For Terminological Similarity Measures . . . 106
5.5.2 Evaluation and Selection of Spatial Similarity Measures . . . 110
5.5.2.1 Test Cases Evaluation For Spatial Similarity Measures . . . 110
5.5.2.2 Full Datasets Evaluation For Spatial Similarity Measures 113
5.5.3 Evaluation of Hybrid Approaches Applied to Spatial and Terminological Attributes . . . 119

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

5.5.3.1 Test Cases Evaluation For Hybrid Approaches 119
5.5.3.2 Full Datasets Evaluation For Hybrid Approaches 121
5.6 Conclusion 125

6 Visualization of Geospatial Integration **127**
6.1 Data Fusion and Uncertainty 128
6.1.1 Merging Corresponding Entities 129
6.1.2 Uncertainty Level Computation 131
6.2 Uncertainty Visualization: Proposals and Assessment 133
6.2.1 Selection of Visual Variables to Portray Uncertainties 135
6.2.1.1 Experiment Set-up 136
6.2.1.2 Assessment and Result 137
6.2.2 Selection of Uncertainties Information to Portray on Map or on Demand 138
6.2.2.1 Proposal and Simulator 139
6.2.2.2 Assessment of Proposals 141
6.3 Impact of Visualizing Uncertainty in LBS: A Use Case for Tourists 143
6.3.1 Experimental Protocol and Simulator 143
6.3.2 Results and Analysis 147
6.4 Multi-providers LBS Prototype 150
6.4.1 Overview and Architecture 150
6.4.2 Interfaces and Navigation 152
6.5 Conclusion 153

7 Conclusions and Perspectives **155**
7.1 Contribution Summary 156
7.1.1 Taxonomy and Benchmarking 156
7.1.2 Spatial Similarity Measure and Geospatial Entity Matching 156
7.1.3 Visualization and SHS Evaluation 157
7.2 Short-term Perspectives 158
7.2.1 Enrichment of Taxonomy and Benchmark 158
7.2.2 Improving Geospatial Entity Matching 159
7.2.3 Different Estimations of Uncertainty 159
7.2.4 Experimenting New Visual Mode 160
7.2.5 Considering the Geographical Context 160
7.3 Long-term Perspectives 161
7.3.1 Enhancing Geospatial Entity matching 161
7.3.2 Combining Blocking Algorithm 161
7.3.3 Visualization and Navigation 162
7.4 Final Words 162

A Probability-based Combination of Individual similarities **163**

Bibliography **166**

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

Résumé long en français **180**
Services de Géolocalisation 180
Motivation 181
Les problématiques 184
Questions de recherche 187
Méthodologie et contributions 188
Conclusion 190

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

List of Figures

1.1 *Walser Hotel* in Courmayeur city, Italy, is located differently by three LBS providers. [Accessed: June 2016] . . . 3
1.2 *L'Ecluse restaurant* in France has different terminological data offered by two LBS providers. [Accessed: February 2016] . . . 4
1.3 Three different legends to represent an hotel on maps collected from Share Icon, Icon Archive and Google Maps, respectively. . . . 4
1.4 Benchmarking's process flow. . . . 9
1.5 Matching and merging process flow. . . . 10
1.6 Visualizing's process flow. . . . 10

2.1 Example of radius query with Nokia Here Places API. . . . 17
2.2 Schemas of semi-structured POI data offered by two LBS providers. . . . 20
2.3 Entity matching process's phases. . . . 21
2.4 Matching more than two datasets. . . . 24
2.5 Classification scheme for geospatial entity matching approaches. . . . 25
2.6 Example of One-Sided Nearest join: matching A to B. . . . 27
2.7 Example of Mutually-Nearest join. . . . 27
2.8 Example of NW: matching A and B with a 0.45 threshold. . . . 28
2.9 Set of visual variables to represent uncertainty. . . . 42

3.1 Different approximations of the area that contains spatial objects [BKS93]. MBR: Minimum Bounding Rectangle; RMBR: Rotated Minimum Bounding Rectangle; MBC: Minimum Bounding Circle; MBE: Minimum Bounding Ellipse; CH: Convex Hull; n-C: Minimum Bounding of n-corners (4-C, 5-C, 6-C). . . . 51
3.2 Schemas of data offered by two LBS providers. . . . 53
3.3 Example of the *Different Data* difference. Two corresponding entities referring to the *IUT Lyon 1 - Gratte-Ciel University* having SEMDD and SYNDD differences. . . . 55
3.4 Example of the *Missing Data* difference. Two corresponding entities refer to the *Colorado* restaurant having a MD. . . . 56
3.5 Example of the *Different Locations* difference. Two corresponding entities separated by 15 meters having DL difference. . . . 58
3.6 Example of the *Equipollent Positions* difference. Two corresponding entities refer to *IUT Lyon 1 - Gratte-Ciel University* having different locations but both are correct. . . . 59
3.7 Example of the *Duplicate Entities*. Two entities offered by one provider and referring to the same POI *Eat Sushi Lyon 2*. . . . 60

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

3.8 Two corresponding entities that have EP inconsistency and refer to *Wallace State Park* POI . . . 63
3.9 All possible combinations for differences. . . . 65
3.10 All possible combinations for resemblances. . . . 66
3.11 Example of complex combination of resemblances. $e_1^{'}$ SD e_1, $e_1^{'}$ SD e_2 and $e_1^{'}$ SUP e_3. . . . 67
3.12 Example of the complex combination of differences and resemblances. $e_1^{'}$ DL e_1, $e_1^{'}$ SD e_2 and $e_1^{'}$ SD e_3. . . . 68

4.1 Overview of the processes involved in GeoBench. . . . 72
4.2 Relational model of GeoBench DB. . . . 73
4.3 Mediator schema of GeoBench aligned to three LBS providers. . . . 74
4.4 Mediator hierarchy of GeoBench aligned to three LBS providers. . . . 75
4.5 Search Interface of GeoBench Aligner. . . . 79
4.6 Matching Interface of GeoBench Aligner. . . . 80

5.1 Process of geospatial entity matching approach. . . . 88
5.2 Curves of common mathematical functions that fit ND's requirements. . . . 91
5.3 Decay rate of several NDs functions. . . . 94
5.4 *F-measure*s of terminological similarity measures in the test cases experiment, applied to terminological attributes having SYNDD. . . . 104
5.5 *F-measure*s of terminological similarity measures in the full datasets experiment. . . . 107
5.6 *F-measures* of spatial similarity measures namely MN, NW and NDs. . . . 111
5.7 Effectiveness of NW and MN in the full datasets evaluation. . . . 113
5.8 Effectiveness of NDs in the full datasets evaluation with 1:n decision algorithm. . . . 115
5.9 Effectiveness of NW and MN in the full datasets evaluation. . . . 116
5.10 Effectiveness of NDs in the full datasets evaluation with 1:1 decision algorithm. . . . 117
5.11 Effectiveness of GSs in the full datasets experiment. . . . 122
5.12 Effectiveness of LRs in the full datasets experiment. . . . 124

6.1 Three different legends to represent an hotel on maps. . . . 133
6.2 Legend of Mapquest provider: uses color to categorize the POIs. . . . 134
6.3 Visual variables that may fit the uncertainties of integrated entities. . . . 137
6.4 Proposal #1: Spatial uncertainty is displayed on the map. . . . 139
6.5 Proposal #2: Terminological uncertainty is displayed on the map. . . . 139
6.6 Proposal #3: Spatial and terminological uncertainties are both portrayed. 140
6.7 Proposal #4: Global uncertainty is portrayed on the map. . . . 140
6.8 Proposal #5: All uncertainties are portrayed together on the map. . . . 140
6.9 Simulator: example of POIs location with proposal #1. . . . 141
6.10 Average score on the appreciation scale according to proposals. . . . 142
6.11 Interface of a first prototype for multi LBS providers. . . . 146
6.12 "Source mode" to compare the source entities on demand. . . . 146
6.13 Mean response times for the three missions. Black bars represent standard deviations. "*" indicates significant differences between groups, and "→*" a trend towards significance ($p = 5.6\%$). . . . 148

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

6.14 Sequence diagram for UNIMAP LBS framework. 151
6.15 Process flow of UNIMAP prototype. 151
6.16 Interface of UNIMAP prototype. 152
6.17 Source mode of an integrated entity. 153
6.18 A comparative table of an integrated entity. 154

A.1 *Hôtel Walser* dans la ville de Courmayeur, en Italie, est situé différemment par trois Fournisseurs LBS. [Accès: juin 2016] 182
A.2 Le restaurant L'*Ecluse* en France dispose de données terminologiques différentes, fournies par deux fournisseurs LBS. [Accès: Février 2016] . . . 183
A.3 Trois légendes différentes pour représenter un hôtel sur des cartes collectées respectivement par Share Icon, Icon Archive et Google Maps. . . . 183
A.4 Elaboration de benchmark. 188
A.5 Flux de processus de l'alignement et fusion. 189
A.6 Flux de processus de la visualisation. 190

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

List of Tables

2.1 Technical overview for several LBS providers . . . 16
2.2 State of the art of geospatial entity matching approaches. . . . 36
2.3 Contingency table of evaluation measures. . . . 40

3.1 Two entities x and y, offered by two different providers, that refer to *Eiffel Tower* POI, p_{ET}, with two different schemas. . . . 49
3.2 Example of the *Similar Data* resemblance. Two entities offered by one LBS provider, that have the same value for the terminological *name* attribute, and located in two different locations. . . . 54
3.3 Example of the *Superposition* resemblance. Two entities, offered by one LBS provider, that have the same location coordinates. . . . 57
3.4 Taxonomy's inconsistencies. . . . 60
3.5 Inconsistencies of the taxonomy and their impact. . . . 64
3.6 Example of complex combination of resemblances. $e_1^{'}$ SD e_1, $e_1^{'}$ SD e_2 and $e_1^{'}$ SUP e_3. . . . 67
3.7 Example of the complex combination of differences and resemblances. $e_1^{'}$ DL e_1, $e_1^{'}$ SD e_2 and $e_1^{'}$ SD e_3. . . . 68

4.1 Statistics on providers' datasets collected by GeoBench Aligner (June 2016). 81
4.2 Top ten test cases according to the number of correspondences (June 2016). 84

5.1 Set of common mathematical functions and curves that fit ND's requirements. . . . 93
5.2 Variations of ND's functions applied to the interval $[0; \beta]$ with $\beta = 100$. . . . 94
5.3 Test cases to evaluate string similarity measures applied to terminological attributes. . . . 103
5.4 Performance of terminological similarity measures in the test cases experiment, applied to terminological attributes having SYNDD. . . . 105
5.5 Statistic of terminological attributes in the full datasets evaluation with respect to differences. . . . 106
5.6 Performance of terminological similarity measures in the full datasets experiment. . . . 108
5.7 Selection of the most appropriate terminological similarity measure for each attribute at low, moderate and high thresholds. . . . 109
5.8 Characterized evaluation of spatial similarity measures. . . . 112
5.9 Efficiency of NDs, MN and NW. . . . 118
5.10 Effectiveness of GSs in the test cases experiment. . . . 120
5.11 Performance of GSs and LRs in the full datasets experiment. . . . 124

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

6.1 Results of Student t-test to evaluate the comprehension level of visual variables for spatial, terminological and global uncertainties, separately. . 138
6.2 Number of times the uncertainties are placed in the first position. 142
6.3 Number of times the proposals are chosen as the most relevant. 143
6.4 Mann-Whitney test results with 5% significance level: response time comparison of groups two by two. Significant results are indicated by bold, a trend towards significance are underlined. 148

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

Chapter 1

Introduction

Contents

1.1 Location-Based Services 2
1.2 Motivation 3
1.3 Issues 5
1.4 Research Questions 8
1.5 Methodology and Contributions 8
1.6 Dissertation Outline 11

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

1.1 Location-Based Services

A geographic information system (GIS) is a computer application designed to perform a wide range of operations on geographic information [SE90, LT92]. Geographic information can represent any location on the globe and has a variety of descriptions. Thus a GIS includes functions to capture, store, manipulate, analyze, visualize and export all kinds of geographic information. Recent years have seen an exciting evolution in information technology, which has led to a proliferation in new products such as Location-Based Services (LBS) [SV04]. The latter are information-oriented services issued from GIS and able to offer highly customized services. These services provide useful information based on a given address or geographical location via mobile, tablet or desktop PC. Examples of LBS include services to identify the location of a person or object, car navigation, vehicle tracking and personalized weather services. Also, they can include commerce when taking the form of coupons or advertising directed at customers based on their current location. LBS had faced several issues since its first launch such as the accuracy of GPS devices and the limitations of mobile devices.

This thesis considers LBS applied to the tourism field. It specifically focuses on LBS that offer information about Points of Interest (POIs), such as locating the nearest restaurant or discovering museums in a given city. LBS provide us with useful information about places anywhere, which makes these services of great interest for users and development communities. For instance, semantic routing systems use POIs to help people identifying their locations during their navigation [RLA+15, RLM+15]. In the tourism field, which has become a major economic resource for many countries, LBS providers (e.g., Google Maps[1], OpenStreetMap[2], Bing Maps[3], etc.) allow tourists to quickly and remotely search for POIs such as monuments, parks and hotels.

Technically, POIs are modeled as entities that are described by spatial information, such as location coordinates, and terminological information such as place name, phone and website. Usually, interactive map tools are proposed by LBS providers to facilitate the process of discovering POIs. These tools consist of a base map made of raster images or vector objects [Mac04]. Then, legends or icons are placed on the map to show POIs locations. A click on a legend displays a window containing the place terminological information. Traditionally, legends are designed to represent POIs types (e.g., park, lake and mall), so tourists can easily distinguish places on the map.

[1] Google Maps: http://maps.google.comm
[2] OpenStreetMap: http://openstreetmap.org
[3] Bing Maps: http://maps.bing.com

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

1.2 Motivation

The amount of data and the number of providers have been growing at a dramatic pace in recent years. The multiplication of geographic information describing the same reality casts doubts on the validity of the location displayed [DPS98]. Spatial entities referring to the same POI may include incomplete, inconsistent, inaccurate or even wrong data from one provider to another. In addition, some POIs may be included in the database of one provider but not by the others or they may be duplicated in the same database. This is due to different policies and strategies used by LBS providers to construct databases, update information and elaborate the results of queries. As a consequence, users obtain different and conflicted answers from one provider to another for the same query. For example, two LBS providers may give two different locations for the same POI. Figure A.1 shows the search results of *Walser Hotel* in Courmayeur city, Italy from three different providers, Figure A.1a is taken from GoogleMaps[4] where the hotel is located on the right side of the highway (the yellow street), Figure A.1b is taken from Nokia Here Maps[5] where the hotel is located on the left side of the highway (the red street) and Figure A.1c is taken from OpenStreetMap[6] where the *Walser hotel* is not represented at all. Also, inconsistencies may affect the terminological attributes. Figure A.2a and Figure A.2b show the terminological data of *L'Ecluse Restaurant* offered by GoogleMaps[7] and Nokia Here Maps[8] respectively. The two providers offer same address and phone number, but syntactic differences occur for place name ("La petite Ecluse des Grands-Augustins" vs. "L'Ecluse") and website ("lecluse-restaurant-paris.fr" vs. "leclusebaravin.com").

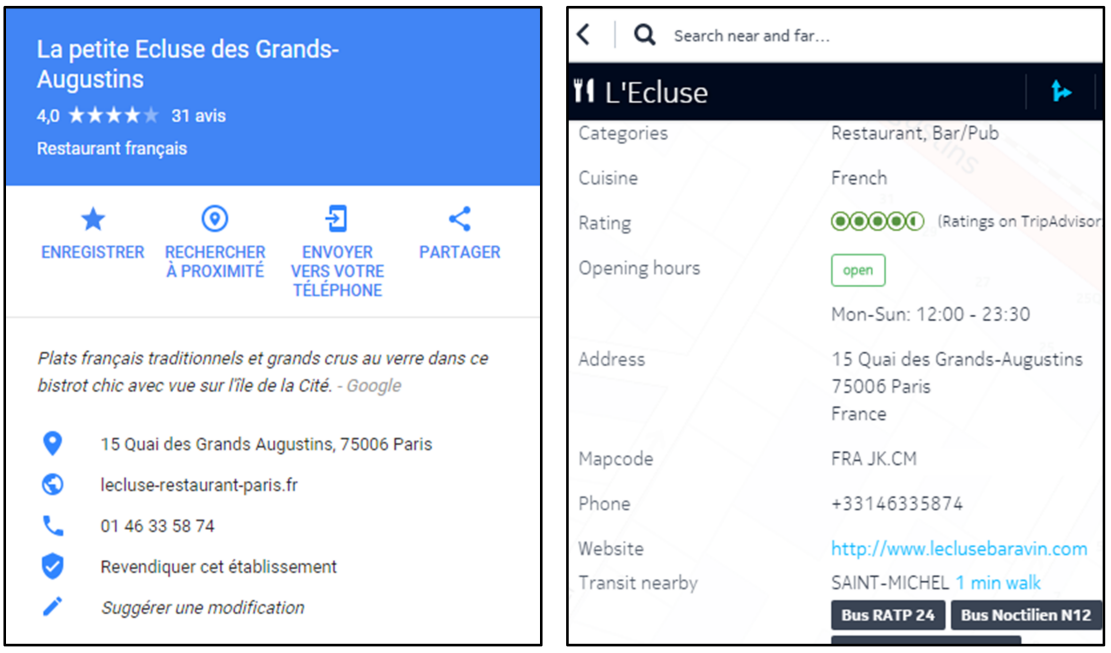


(a) GoogleMaps (b) HereMaps (c) Openstreetmaps

FIGURE 1.1: *Walser Hotel* in Courmayeur city, Italy, is located differently by three LBS providers. [Accessed: June 2016]

[4] *Walser Hotel* by GoogleMaps. [Accessed: June 2016]
[5] *Walser Hotel* by Nokia Here Maps. [Accessed: June 2016]
[6] *Walser Hotel* by OpenStreetMap. [Accessed: June 2016]
[7] *L'Ecluse Restaurant* by GoogleMaps. [Accessed: June 2016]
[8] *L'Ecluse Restaurant* by Nokia Here Maps. [Accessed: June 2016]

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés



(a) *L'Ecluse restaurant* by GoogleMaps (b) *L'Ecluse restaurant* by HereMaps

FIGURE 1.2: *L'Ecluse restaurant* in France has different terminological data offered by two LBS providers. [Accessed: February 2016]

The above examples show the heterogeneity of geospatial data issued from several LBS providers. The heterogeneity may also affect the representation of POIs on maps. Each LBS provider uses its own set of legends to represent POIs. For instance, Figure A.3 shows three different legends to represent hotels collected from Share Icon[9], Icon Archive[10] and Icons DB[11], respectively. Furthermore, with respect to the real area of places, POIs could be represented in different ways such as point (0D), line (1D), polyline (2D) or volume (3D) from one provider to another.



FIGURE 1.3: Three different legends to represent an hotel on maps collected from Share Icon, Icon Archive and Google Maps, respectively.

In database research domain, data integration has been widely proposed to solve the heterogeneity of data issued from several sources in order to improve the data quality [HRO06]. In our context, this thesis aims to integrate the existing LBS sources of POIs to create a better service with more complete and accurate information with respect to tourist field. Geospatial integration has been widely studied under the term "map conflation" where vector and raster maps are integrated [RAUB11]. This thesis focuses

[9] Share Icon hotel legend: https://www.shareicon.net/hotel-accomodation
[10] Icon Archive hotel legend: http://www.iconarchive.com/show/ios7-icons-by-icons8/Hotel.html
[11] Icons DB hotel legend: http://www.iconsdb.com/royal-azure-blue-icons/hotel-icon.html

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

on the integration of geospatial punctual objects; it does not consider the integration of base maps or the integration of complex geographic objects such as streets or buildings. The integration of punctual objects requires the matching of heterogeneous data structures and the reconciliation of inconsistent geospatial data. Besides, the integration of data describing the same reality is prone to uncertainty due to different kinds of heterogeneity [CD99]. Ignoring that uncertainty can, at best, lead to slightly incorrect predictions and at worst can be completely fatal to the use of integrated data. This thesis also considers the uncertainties of integrated data in order to offer accurate and meaningful representation of integrated data.

1.3 Issues

The scope of this research is to improve the results' quality of users' queries so that tourists can search and find POI with accurate and complete information. Data of several LBS providers should be integrated to ensure merging of spatial and terminological information. In other words, entities that refer to the same place should be integrated to obtain more complete information and to avoid duplication. The main issues facing LBS integration are detailed below.

1. Data access: LBS integration requires an access to providers' data to be integrated. Most LBS providers, in particular commercial providers, give a free and limited access to their databases. They offer an on-demand access using web services Application Programming Interface (API). Providers' APIs are surrounded by limitations concerning the number of queries per day, the number of returned POIs per query, etc. Accessing all POIs from all LBS providers is subject to technical constraints and randomness, such as advertisement and POI rating. These affect the availability and numbers of POIs, and therefore their integration. Besides, data of providers are subject for usage policies. These policies differ from one provider to another. For example, free data may be used for personnel or research projects, but not for commercial projects.

2. Source data quality: LBS providers have distinct strategies to construct their databases. For example, Apidae-Tourisme[12] provider, also known as SITRA, employs geographic experts to visit the physical locations of places in order to check and validate POIs. On the other hand, OpenStreetMap provider allows users from anywhere to add, edit and delete POIs. Various strategies, which are detailed further in the state-of-the-art, lead to different quality of data sources. The integration of low and high quality data may not end with better information.

[12] Apidae-Tourisme provider, also known as Sitra: http://apidae-tourisme.com

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

3. Schema heterogeneity: The schema of a database describes its structure. In 2010, a W3C working group[13] was created to develop standard specifications for the representation of POI information on the web. In 2013, this working group was moved to Open Geospatial Consortium (OGC) under the Points Of Interest Standards Working Group (POISWG) [POI13]. The intent of this group was to produce an encoding standard of POIs data that includes an abstract data model and JSON and XML schema implementations of that data model. The creation of the standard model had been initiated[14], but abandoned in 2014. Nowadays, there is no standard schema that can be used to construct a POIs' database; each LBS provider uses its own model. This means the data of distinct providers have different labels and hierarchies.

4. Attributes' values heterogeneity: Databases created by different people and for different purposes may have variations at the instance level. This reflects on data integration even when they refer to the same object. Variations in values of POIs attributes arise for multiple reasons. Concerning spatial values, real-world objects may be represented differently. They may be different due to the accuracy of GPS devices, human mistake when locating a place or distinct strategies to locate a POI to be represented on a map. For instance, a park may be represented as a polygon that covers the park's area or a point located at the park's gate. Terminological values may be (1) syntactically different such as use of abbreviations and mis-spelling, or (2) semantically different such as synonyms and hyponyms. Values variations over time may also cause data heterogeneity if data are not updated regularly. Other frequent issue concerning data heterogeneity is the language used to describe the POIs. For instance, the place names of POIs offered by OpenStreetMap are described in different languages. On the other hand, Bing Maps uses the language of the country where the POIs are located. In other words, data of several providers may be described in different languages, which causes additional heterogeneity. These distinct heterogeneities create obstacles for entity matching and LBS integration.

5. Evaluation of geospatial entity matching: LBS integration requires matching the entities that refer to the same real-world POI. In some contexts, there are keys identifier that may be used for entity matching. For instance, the Social Security Number (SSN) may be used as a key to identify the citizens in several databases, because each person has a unique SSN. Unfortunately, LBS providers use specific internal identifier in their databases. This means that POIs lack a global unique identifier. In this case, entity matching can be done by comparing the values

[13] W3C working group for POI standard model: https://w3.org/2010/POI
[14] POI standard model project: https://github.com/opengeospatial/poi

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

of attributes describing POIs details. Several approaches have been proposed to solve the matching of geospatial entities. This can be done by measuring the values' similarities between entities in order to identify whether they do correspond. These approaches have been evaluated in different contexts using different metrics and datasets. These datasets are often small, not well characterized, chosen randomly and not made available for researchers, which makes the results of the approaches' experiments incomparable. Benchmarks have been proposed for entity matching in different context such as publications and commerce [KTR10]. But, all of these benchmarks do not consider the spatial aspect of entities. To the best of our knowledge, there is no benchmark for evaluating approaches that match geospatial entities.

6. Evaluation of data fusion: Once the corresponding entities of several sources are identified, a crafted algorithm is needed to fuse them in order to obtain new integrated entities that are better than the sources entities. Some attributes may have different possible values from one provider to another. Data fusion is a decision process that helps selecting the value that seems most correct. For example, one basic solution is to choose the value that comes from a source with a higher quality, which introduces new issues about the quality's classification of LBS providers. To evaluate the values selected by the fusion algorithm, we need to compare them with correct values. Unfortunately, there is no data source with 100% accuracies to be used as a ground-truth. The only way to achieve this evaluation is to manually check values from reality, which is infeasible due to the increasingly voluminous of POIs. Instead, it is possible to estimate the certainty of merged entities according to the sources information, but not to reality. In other words, when an attribute has only one possible value, then this attribute represents certain information. Conversely, if an attribute has several contradictory values, the chosen value cannot be viewed as reliable. A fusion algorithm should be able to evaluate the uncertainty of integrated entities according to the contradiction of sources entities. This uncertainty may arise for spatial information, terminological information or both kinds of information. The uncertainty of data causes additional information that must be delivered to users through LBS applications.

7. Data visualization: When considering merged entities, additional information about data uncertainty and data sources needs to be provided for users so they can estimate the quality of data and compare values of different sources in order to make confident decisions when choosing POIs. Practical solutions are needed to provide all additional information without overloading maps, especially for mobile users. Such solutions need to be evaluated to make sure that users can understand

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

additional information without any contradiction with basic information. One additional issue concerns the visualization of integrated entities. As was mentioned earlier, legends are used to represent POIs on maps. However, each LBS provider has its own set of legends. A solution is needed to solve the heterogeneity at this level.

1.4 Research Questions

The main goal of this dissertation is to create unified LBS by integrating the data of several providers. After listing the issues of LBS integration, this section represents the questions that this thesis is addressing.

1. How LBS providers differ one from another? In other words, what kinds of heterogeneities and inconsistencies exist between LBS providers? And, how LBS data can be integrated? (Issues: Schema heterogeneity and Attributes' values heterogeneity)
2. Are techniques for geospatial entity matching effective enough? What are the necessary specifications that allow a fair evaluation and comparison of geospatial entity matching approaches? (Issue: Evaluation of geospatial entity matching)
3. How entities referring to the same real-world object should be merged together? And how to estimate the uncertainty resulting from such merging? What is the most appropriate way to deliver integrated POIs and to represent their uncertainty in tourist context? How tourists interpret this uncertainty? Does it reflect on tourists' choices when they search for POI? (Issues: Evaluation of data fusion and Data visualization)

Note other issues, namely data access and source data quality are not considered in this thesis. Although the APIs for data access are surrounded by several technical limitations, they are still sufficient for our research project. In addition, several existing work discusses the quality of geospatial data [TKA07, PZLL11]. But in this thesis, we use LBS data sources as they are, without considering their qualities.

1.5 Methodology and Contributions

After developing the issues and research questions, and following the state of the art, we will expose in depth our research method.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

Concerning the first contribution, we already highlighted the necessity of creating a benchmark to ensure a fair evaluation and comparison of geospatial entity matching approaches. Figure A.4 shows the process flow to create such a benchmark using real-world data collected from several LBS providers. To do so, we start by creating a taxonomy that describes and formalizes all kinds of inconsistencies and heterogeneities between LBS providers. Then, a tool called GeoBench Aligner is implemented; it consists of a semi-automatic process that collects data from existing LBS providers. A user's validation is required to indicate the inconsistencies between entities and decide whether two entities correspond. The output of this tool is a database, namely GeoBench DB, characterized according to the taxonomy and it contains real entities for which we know the relevance of correspondence. Following this, a second tool namely PABench Extractor takes GeoBench DB as an input to generate datasets describing a given situation. These characterized database will be used later to evaluate geospatial entity matching approaches. Our benchmark, namely PABench (POI Alignment Benchmark), consists of those datasets and a list of metrics to assess the performance and quality of matching approaches. Thus, PABench allows the evaluation of geospatial entity matching approaches in different situations, which helps discovering their weak and strong points.

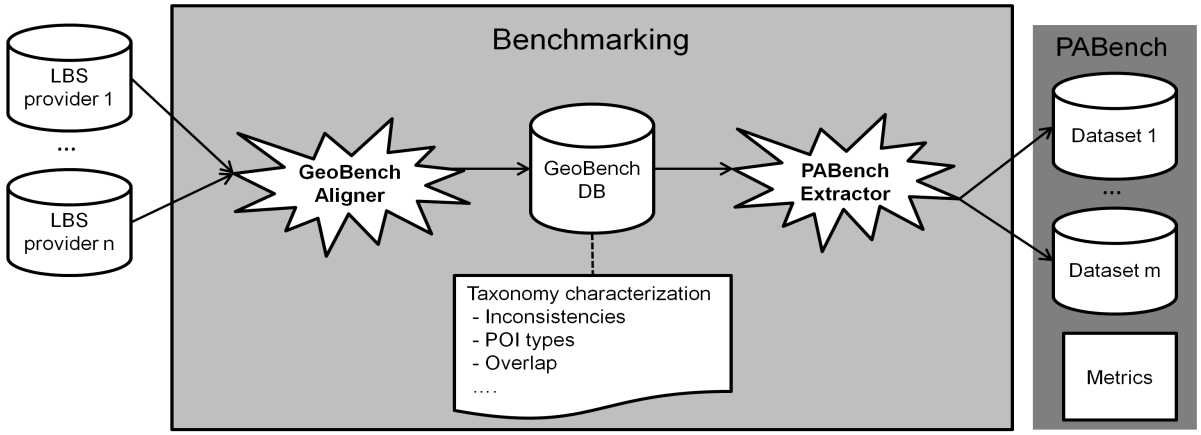


FIGURE 1.4: Benchmarking's process flow.

Then, we focus on the matching process to detect the entities of two datasets that refer to the same place in the real-world. Hence, the second contribution consists of elaborating a matching and merging process to integrate entities of several providers. Figure A.5 shows the process flow of this contribution. We start by proposing a generalization for spatial similarity measure, namely Normalized-Distance (ND). This measure is calculated using the Euclidean distance between two compared entities. Now, spatial and terminological information of entities are compared using distinct similarity measures. Then, we propose a method, namely Global Similarity (GS), to numerically combine several similarities in order to obtain an overview between compared entities. Using

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

this combined similarity, we decide whether two entities correspond. Concerning the merging phase, a basic algorithm will be used based on the state of the art. The final output is a dataset containing merged entities where each entity is accompanied with a degree representing its uncertainty. This latter is estimated using the combined and initial similarity measures. Note that our matching contribution will be compared to existing approaches using PABench.

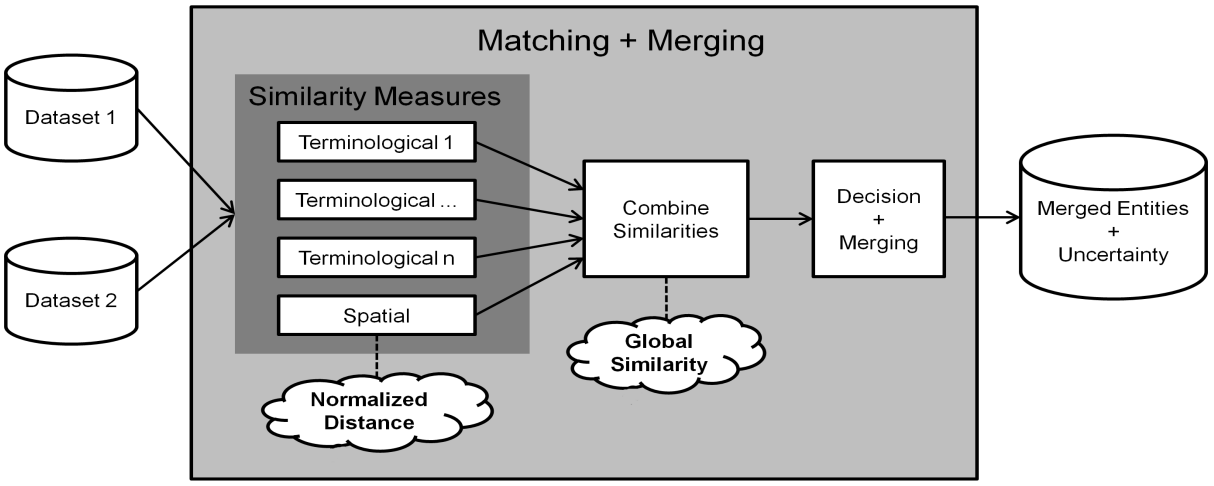


FIGURE 1.5: Matching and merging process flow.

The final contribution concerns the delivery of integrated entities and their uncertainties to end-users. Figure A.6 shows the process flow of this contribution. After investigating existing solutions for representing integrated POIs and uncertainty information, we select and adapt what can be used in our context. On these bases, several proposals are suggested and a prototype of these proposals is implemented. Then, a first psycho-cognitive test is conducted to find the most appropriate proposal to represent integrated POIs and their uncertainty at three levels namely spatial uncertainty, terminological uncertainty and global uncertainty. Also, a second psycho-cognitive test is conducted to analyze how uncertainty reflects on users' decisions when searching for POIs.

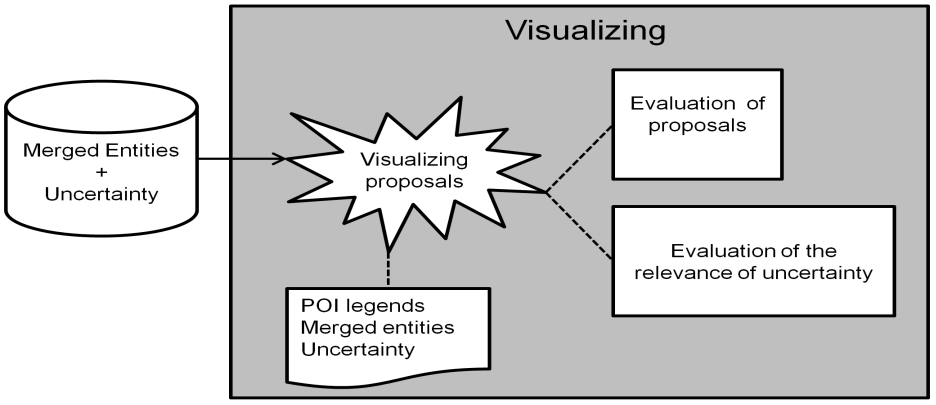


FIGURE 1.6: Visualizing's process flow.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

1.6 Dissertation Outline

This chapter introduced the problems and challenges of LBS integration and outlined the goals and contributions of this research.

Chapter 2 presents the state-of-the-art of this study in details and develops the theoretical and technical background to overcome the different challenges towards LBS integration.

Chapter 3 introduces taxonomy for LBS providers. This taxonomy describes a model for LBS that offer POI and highlights various inconsistencies that exist between LBS providers.

Chapter 4 details the necessary specification to create a benchmark in order to evaluate and compare geospatial entity matching approaches.

Chapter 5 defines a new approach for matching geospatial entities and provides the experimental validation that evaluates and compares the matching approaches based on our benchmark.

Chapter 6 adapts existing solutions of uncertainty visualization to fit our context. These solutions are then analyzed using psycho-cognitive tests in order to find the most suitable visualization and to understand the effect of uncertainty on tourists' choices. Also, we present a prototype that includes our final results concerning data integration and uncertainty visualization.

Chapter 7 concludes this dissertation and identifies the perspectives in this research area.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

Chapter 2

Related Work

Contents

2.1 Characteristics of LBS 13

2.2 Data Integration 17

2.2.1 Schema Matching 19

2.2.2 Entity Matching 19

2.2.3 Geospatial Entity Matching 24

2.2.4 Data Fusion 35

2.3 Data Matching Benchmark 37

2.3.1 Datasets 37

2.3.2 Metrics 39

2.4 Visualization of Geospatial Integrated Data 41

2.5 Conclusion and Positioning 43

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

This chapter presents related work which covers the main topics of this thesis. First, a technical state of the art concerning the LBS providers is given. Second, we present data integration with special focus on geospatial entity matching. Then, we present existing benchmarks in data integration. Next, the visualization of spatial data is discussed with respect to data uncertainty. Finally, a summary is given and the positioning of this thesis with respect to the state-of-the-art is discussed.

2.1 Characteristics of LBS

Ponce-Medellin et al. define LBS as a new technology field, issued from GIS that focuses on providing useful information via mobile or desktop PC, based on a given spatial position [PSAR09]. Also, they represent a list of frameworks involved in the LBS and GIS market. A spatial position may be a location coordinates that is automatically detected by as a GPS receiver or a simple address denoted by the user. In this thesis, we specifically focus on LBS that provide tourist POIs. The development of several domain areas such as WEB 2.0, Database Management System (e.g. NOSQL), Internet network and wireless connection, allowed LBS providers to offer efficient services that have arisen a great interest in users and development communities, such as find POIs and Geocoding [XFC12, MRANIG14]. Now, end-users can search for POIs and locate them on map using their personal devices (e.g., mobile, tablet, laptops). When users request location-based queries (e.g. find the nearest restaurants, find hotels in Paris, etc.), these queries are submitted to a provider's database server over the net, along with a simple address or a location detected by GPS device. Finally, the provider's server treats the user query and returns the results. The POIs results may be presented in a simple list or be located on a dynamic or static map. Some intelligent LBS take into account the user's profile and context (e.g., weather) in order to offer more suitable results for users' requests [KFKL10]. For example, if a user specifies in his profile that he likes *Lebanese food*, when he searches for nearby restaurant, the Lebanese restaurants will be ranked up in the results. Another example about context, when a user requests POIs for outdoor activities in winter, the beaches are eliminated from results or ranked at the bottom.

The different strategies used by LBS providers to construct and update their POIs databases lead to the emergence of data heterogeneity between distinct providers. In the following list, we classify these methods:

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

1. **Practitioners method:** it consists of experts who visit places on the ground in order to check and validate POIs positioning and attributes such as the Apidae-Tourisme[1] provider. Generally, this method should produce accurate data, and it is very efficient for a limited region. Otherwise, two drawbacks are distinguished: (i) covering a large geographical area requires a lot of practitioners and time, and (ii) guaranteeing an efficient data updating process is a very hard task.

2. **Collaborative method:** the main source of POIs is users contributions. People from everywhere may add, edit and delete POIs using many technologies such as web user interface or Application Programming Interface (API). The main benefit of this method is the unlimited contributions. Some providers, such as OpenStreetMap[2], use this method and do not verify users' contributions, which casts doubts on the quality of information since people may contribute wrong information. In contrast, some other providers, such as Google Maps[3], require a verification for users' contributions.

3. **Reuse method:** it consists of LBS providers that integrate external sources (e.g., yellow pages, POIs of another LBS providers, old databases, etc.) to enhance their databases. This method may cause data duplication. For instance, BingMaps[4] uses the POIs of Nokia Here Maps[5].

4. **Knowledge method:** it consists in analyzing unstructured data (e.g., text, images) in order to extract information about POIs. For instance, the smart cars of Google take images for *Street View*. These images can be analyzed to extract locations and places' names of POIs.

Some LBS providers may use multiple methods at the same time. For instance, Google Maps uses the collaborative method; it allows people to add the places they own through Google Business tool[6]. Users' contributions go online only after a verification by Google employees. Also, Google Maps uses the knowledge method by extracting information about POIs by analyzing the images detected by Google smart cars. In addition, Google Maps uses the reuse method; it collects all popular and well known touristic POIs (e.g., Eiffel tower and Pisa tower) and integrates them in the base maps.

Since its appearance, LBS has faced a number of core constraints and issues, some have been resolved while others have not been adequately addressed [PSAR09, Kar11].

[1] Apidae-Tourisme provider: http://apidae-tourisme.com
[2] OpenStreetMap provider: http://www.openstreetmap.com
[3] Google Maps provider: http://maps.google.com
[4] Bing Maps provider: https://bing.com/maps
[5] Nokia Here Maps provider: http://www.here.com
[6] Google Business: https://www.google.com/business

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

1. **LBS for mobile:** in the past, mobile devices presented limitations that prevent LBS from reaching their potential. These limitations concerns several aspects: (1) restrictions of mobile computing such as: energy capacity, limited computing power, amount of memory and storage space, (2) limited bandwidth and high costs of wireless internet access, (3) visual representation due to low screen resolution and (4) diversity of mobile devices and operating systems; it was necessary to develop particular devices applications for each. Most of these constraints have been remedied after the release of smart phones and unified operating systems such as Android and IOS.

2. **Positioning techniques:** providers may use different devices and technologies to measure the positions of POIs. These different positioning techniques have different levels of precision and accuracy. On the other hand, all POIs, even those which refer to large area (e.g., parks, mountains), are represented as points. This issue remains pertinent, because there is no standard strategy to locate a point for representing a POI. For instance, some providers locate the point at the center of gravity of POI [BDK+05], while others locate it at the entrance gate of POI[7].

3. **Access to geographical data:** At the beginning, LBS were dedicated to end-users only. Most of geographical databases created by laboratories, governments, organizations and companies have not been made fully available because of confidentiality and commercial issues. In 2013, the Open Geospatial Consortium (OGC) POI Working Group[8] had intentions to create an open POI database[9] where data would be collected from public POIs' lists. A prototype[10] should allow the access to this database through an interactive map and a published web service API. Unfortunately, there is no available information if this database has been created and the prototype is not in service yet. Nevertheless, this issue has been resolved thanks to the protocols developed by OGC such as Web Feature Service (WFS), Web Mapping Service (WMS), Web Catalog Service (WCS), Web Integrator Service (WIS) and Web Processing Service (WPS). For instance, WMS serves to produce maps of spatially referenced data with dynamic geographic information generated by a map server using data from a GIS database and WFS provides an interface allowing requests for geospatial features across the WEB using platform-independent calls (e.g. formatted URL). Based on these protocols and after the emergence of WEB 2.0 and web services, a variety of LBS providers (e.g., Google Maps, MapQuest, Nokia Here Maps, etc.) decided to sell or share their data with developers, researchers, companies, etc.

[7] Example of Nokia Here Maps locates a park at its entrance gate. [Accessed: June 2016]
[8] OGC POI Working Group: http://www.opengeospatial.org/pressroom/pressreleases/1940
[9] Open POI database: http://openpois.net
[10] Prototype for Open POI database: http://openpois.net/map.html

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

Table 2.1 gives a summary of technologies and limits for accessing the POIs data of the most popular LBS providers namely: Google Maps, Bing Maps, OpenStreetMap, MapQuest[11], Apidae-Tourisme and Nokia Here Maps.

	Google Maps	**Bing Maps**	**OpenStreetMap**	**MapQuest**	**Apidae-Tourisme**	**Here**
Owner	Google	Microsoft	OpenStreetMap Community	AOL	Rhône-Alpes Tourisme	Nokia
Launched	2005	2010	2004	1996	2004	2007
Package	Free, commercial	Free, commercial	Free	Free, commercial	Free	Free, commercial
Technologies / API	REST, Javascript	REST, Javascript, SOAP	REST, Download (xml, csv)	REST, Javascript	REST, Export (xls)	REST, Javascript
# of Queries (free package)	1000 / 24h	125 000 / year	1 000 000 / 24h[12]	15000 / month	Unlimited	2500 / 24h
# of Queries (commercial)	100 000 / 24h	Unlimited	N/A	Unlimited	N/A	10 000 / 24h
# POIs / query	200	250	Unlimited	Unlimited	200	100
Constructing databases	Collaborative, Reuse, Knowledge	Reuse	Collaborative	Collaborative, Reuse	Practitioners	Collaborative, Reuse

TABLE 2.1: Technical overview for several LBS providers

According to Table 2.1, most of providers use the REST API technology to allow people accessing their databases. The concept of the REST API is to call a provider's server with a specific HTTP query, and then the server retrieves the POIs and returns them in a semi-structured format such as XML and JSON. People can use these POIs to offer new services, construct new POIs database, etc. There are several types of queries that can be requested over a LBS provider's server:

- **By identifier:** users can request one single POI if they know the internal ID of this POI at the provider's database.
- **By keywords:** users can send a free-text query such as *Restaurants in Paris*. Providers deal with such queries in several ways which produces different results between distinct providers.
- **By spatial criteria and filters:** users can request POIs in a given geographical area. In addition, they can add filters to queries such as requesting POIs of specific types (e.g., restaurant, park) and obtaining results in a specific language or format. There are several queries to specify a geographical area:
  - **Radius:** users can request the nearby POIs within a given distance of a location or address.
  - **Rectangle:** this type of search returns POIs that are within a specific bounding box (rectangle). The rectangle is formed by defining upper left and lower right points.

[11] MapQuest provider: http://www.mapquest.com
[12] One million query per day for all users.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

– **Polygon:** this type of search returns POIs that are within a custom polygon shape. The polygon is defined by a series of location coordinates that end with the same point it has started with.

The *Identifier* and *Radius* queries are common between all providers. Figure 2.1 shows an example of a radius query using Nokia Here Places API[13].

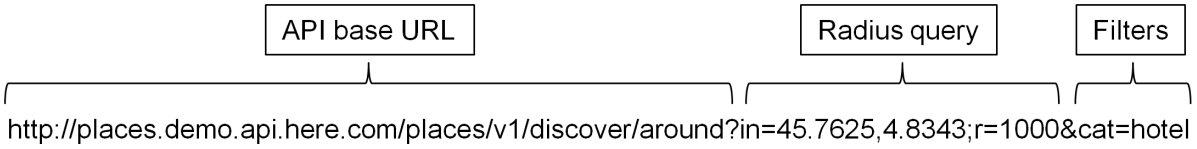


FIGURE 2.1: Example of radius query with Nokia Here Places API.

The base URL concerns the API's server, the radius query concerns the location and distance parameters and the filters concerns the parameters that refine the results according to users' needs. The above query will return all POIs of type *hotel* near the location (45.7625,4.8343) within a distance of 1000 meters.

2.2 Data Integration

Geospatial data are growing continuously over the internet. Some of these data describe the same reality which causes a multiplicity of data. This multiplication may include incomplete, inconsistent, inaccurate or even wrong data from one source against another. Quite a large number of papers have investigated the integration of several geospatial data sources in order to obtain more accurate, complete and up-to-date data. Geospatial integration has been widely studied under the term "*map conflation*" where base maps are integrated. Integration of maps consists of identifying the corresponding entities and fusing them [Cas06]. Ruiz et al. present a wide description of the art with respect to map conflation [RAUB11]. The authors describe existing works in map conflation regarding their formats (raster and vector) and their criteria such as spatial data, terminological data and topological relationships between entities. Some works have been proposed in map conflation using punctual entities [Saa85, CTKS03, Vol06], linear entities [Saa88, Doy00, ZSM05] and polygonal entities [ACH+91, GZK03, Mas06]. In our study, we specifically focus on (semi-) structured spatial databases that describe touristic places. Several studies have been proposed for spatial database integration and spatial entity matching [SGV06, KFKL10, SKS+10]. Database integration, also referred to as data interoperability, is the process of unifying several autonomous data sources. Data integration is a critical task to resolve the problem of heterogeneity between data

[13] Nokia Here Places API: https://developer.here.com/api-explorer

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

sources that describe the same reality. In addition, most new databases are built by integrating existing data sources such as yellow pages or older databases. Parent et al. define three levels for data integration [PS00]:

1. Low level: allowing one database management system (DBMS) to request data from another DBMS.
2. Intermediate level: giving users the possibility to simultaneously manipulate data from several sources in some uniform way.
3. Higher level: building a mediator system on top of existing sources to provide the desired level of integration of the data sources.

To integrate several datasets, three tasks are required: schema matching, entity matching and data fusion.
Schema matching can be defined as the discovery of correspondences between schema elements as well as mapping functions to transform source instances into target instances [BBR11]. Data sources may be in different formats, unstructured data (e.g., free text) requires sophisticated mechanisms for extraction of semantics. While structured (e.g., relational database) and semi-structured (e.g., XML and JSON) data may be easily processed.
Entity matching, also referred to as entity resolution, duplicate identification, record linkage or reference reconciliation, is the task of identifying corresponding entities that refer to the same real-world object [KTR09]. Traditionally, similarity measures are used to compare elements in schema and entity matching. These measures quantifies the concept of proximity between two objects [ZCB87]. Gruyer et al. define a similarity measure as a function $S : \mathbb{O} \times \mathbb{O} \rightarrow [0, 1]$ that verifies the following properties [GRB03]:

$$\begin{aligned} &\forall\, o_1, o_2 \in \mathbb{O} \qquad S(o_1, o_2) = S(o_2, o_1) \\ &\forall\, o_1 \in \mathbb{O} \qquad\quad S(o_1, o_1) = 1 \end{aligned} \tag{2.1}$$

where $\mathbb{O}$ is a set of objects. A similarity measure equals 1 means that the two objects are totally similar. Conversely, a similarity equals 0 means that the two objects are dissimilar. The string similarity measures have been widely used to compare the labels and values of attributes [SGV06, SSL12, RM08, DORB14]. Cohen et al. compare the performance of several string similarity measures on the task of matching entity names [CRF03]. For instance, Levenshtein distance [Lev66], also referred to as Edit-distance, computes the similarity between two strings based on the minimum number of required edit operations (insertion, deletion, substitution) to transform the first string into the other.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

The last step concerns data fusion [DGH+14], once the entity matching process is done, the corresponding entities of several sources may be merged together in order to obtain a more complete integrated database. The task of data fusion is to identify the true values of entities among multiple observed values drawn from different sources of varying reliability.

2.2.1 Schema Matching

Database schema is the structure of data described in a formal language; it refers to the organization of data as a blueprint of how a database is constructed. Schemas that are independently developed in different domains or by different people often have different structures. Schema matching is a critical task in database integration; it allows us to discover the relations between the elements or attributes of two or more data structures. The schema matching process produces a set of corresponding attributes with the mapping functions between them. The mapping functions allow the normalization of corresponding attributes in order to compare them. For instance, Figure 3.2 shows the schemas of two LBS providers, the *lat* and *lng* attributes in Figure 3.2a correspond to the *position* attribute in Figure 3.2b, the *lat* and *lng* attributes need to be concatenated to be mapped and compared to the *position* attribute. Based on the schema matching results, data of corresponding attributes are later compared in the entity matching process in order to find corresponding entities that describe the same objects in the real-world.

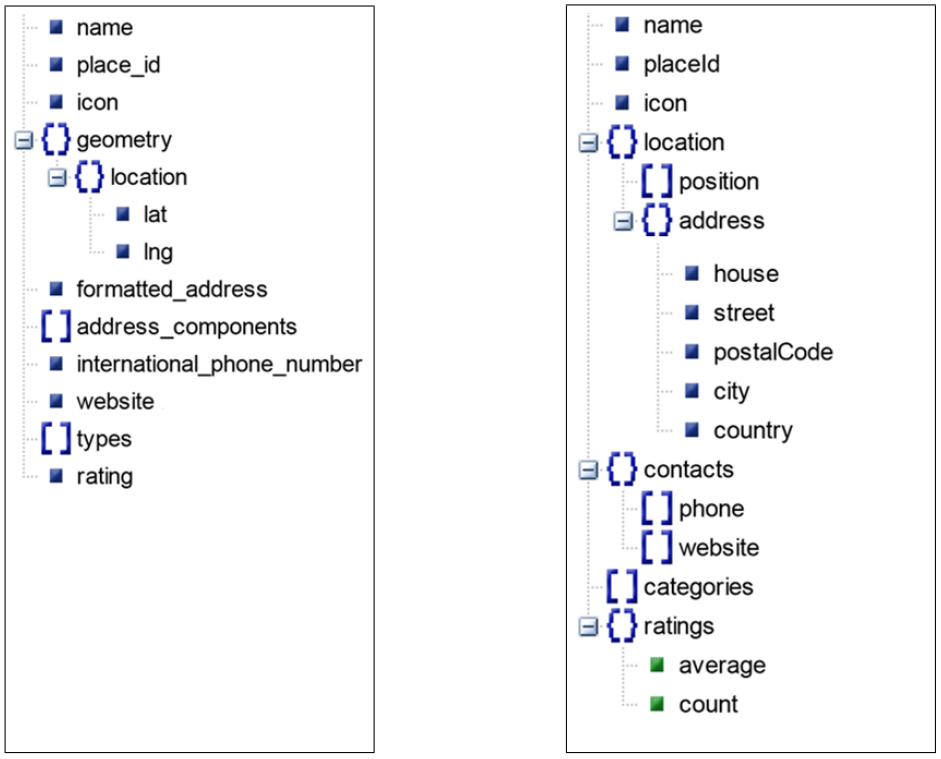
Schema matching may be done manually by an expert in case of simple and small schemas. Otherwise, various approaches have been proposed for semi-automatic and fully-automatic schema matching. Bernstein et al. present a survey that compares different approaches to schema matching [RB01, BMR11]. Most of these approaches combine several techniques, this can be done in two ways: (i) hybrid approach that integrates multiple matcher criteria [MBR01, DR02, ADMR05] and (ii) composite approach that combines the results of independently executed matchers [DDL00, DDH01].

No further details are given about schema matching since this thesis mainly focuses on entity matching. It is important to note that the requirements and specifications of schema matching are similar to those of entity matching (see Section 2.2.2).

2.2.2 Entity Matching

Entity matching is a crucial task for both data integration and de-duplication. It is a challenging task for entities that are highly heterogeneous or of limited data quality

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés



(a) Google POI schema. (b) Nokia Here POI schema.

FIGURE 2.2: Schemas of semi-structured POI data offered by two LBS providers.

where a unique identifier is not available [KTR09, KTR10, TR07]. Generally, entity matching approaches should meet four requirements:

- **Effectiveness:** high-quality matching result where only real corresponding entities should be included in the result.
- **Efficiency:** execution time of a matching approach is a critical factor. An approach should perform as fast as possible even for large datasets.
- **Generic:** an approach should be applicable to match data from various domains (e.g., enterprise data, life science data) and for different data models (e.g., relational, XML, JSON).
- **Self-tuning:** the manual effort to employ and parameterize an approach should be as low as possible.

Two kinds of entity matching approaches are distinguished: Offline and Online. The former deals with local and large datasets. For instance, entity matching during the Extract, Transform and Load (ETL) process of data warehouses is a sample case for offline matching [EEV02, KR08]. The latter arises from interactive data integration steps such as mediated queries or data mash-ups based on specific user input [BBS05, BG11].

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

Köpcke et al. present a survey for entity matching and they compare 11 frameworks [KR10]. For instance, FEBRL (Freely Extensible Biomedical Record Linkage) is one of the rare frameworks that are freely available on the web under an open source software license [Chr08]. It was originally developed for entity matching in the biomedical domain. FEBRL offers a variety of blocking methods and a large selection of 26 different string similarity measures for attributes values matching. Additionally, FEBRL allows the combination of different similarity measures and it supports a training-based numerical combination, which uses machine learning approaches such as support vector machine. Figure 5.1 shows the three main phases for entity matching namely, pre-matching, matching and post-matching.

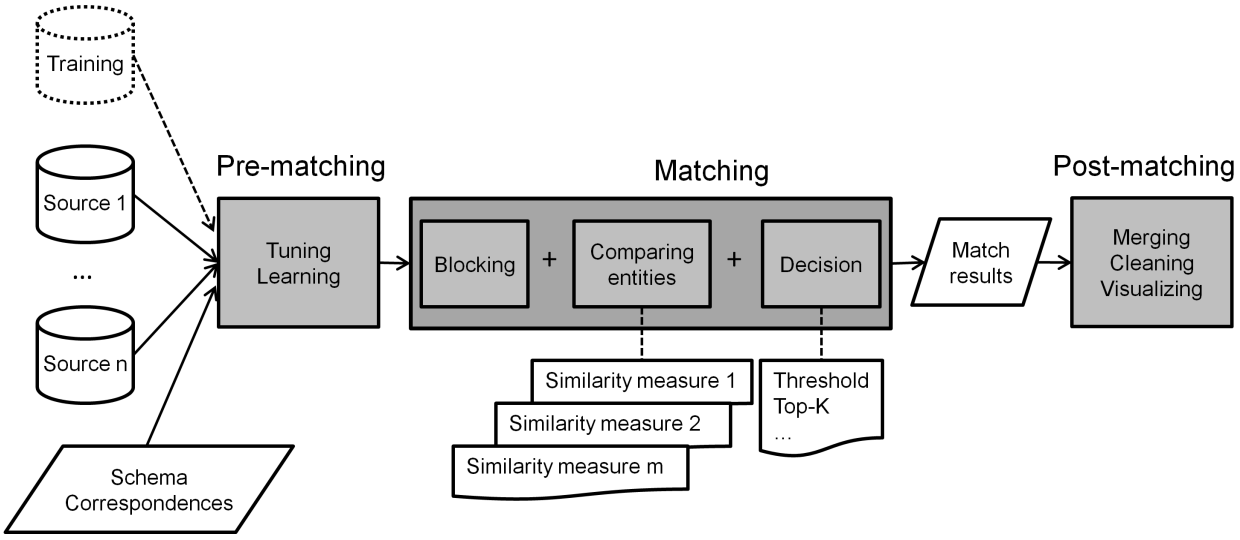


FIGURE 2.3: Entity matching process's phases.

2.2.2.1 Pre-matching

Entity matching approaches begin by considering two or more source sets of data, which need to be integrated, and then also the alignment list between their schemas. For de-duplicating purpose, only one dataset is required. In this phase, a user might intervene to provide the necessary elements for the matching process. Users may specify whether they want to use a blocking algorithm or not. External sources, such as thesauri or dictionaries, can be also specified during this phase. A matching approach may need to convert data formats into a specific format. The configuration of a matching approach must be specified by tuning its parameters, such as weights or thresholds. This configuration may be set manually or by using a machine learning technique [TKM02, BM03]. This latter requires a training dataset and a learning-based approach to find the parameterization that will lead to the best results. Once all the necessary elements are ready, the second phase can be processed.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

2.2.2.2 Matching

The matching phase benefits of three steps namely Blocking, Compare entities and Decision algorithm. Concerning the first step, when matching two large datasets, it is generally not feasible to exhaustively match the Cartesian product of all input entities. Usually, a blocking task is needed for large inputs in order to reduce the search space for entity matching from the Cartesian product to a small subset of the most likely matching entity pairs. Matching of an entity can then be restricted to the other source's entities in the same block. Typically, a key is used to partition the entities to be matched into blocks. A standard blocking method clusters entities into blocks where they share an identical blocking key. An example of this key is the values of a given attribute or multiple attributes. It is preferable to use the least error-prone attributes available. One possible example is matching two datasets that contain entities about people in France where entities that share the same postal code (blocking key) are clustered and compared together. The key may be determined manually or automatically based on training data [BKM06]. Additionally, more advanced blocking methods have been proposed such as processing iteratively blocks in order to use the results of one block in the processing step of another block [WMK+09]. Peter Christen proposes a survey that describes and evaluates several blocking methods [Chr12] namely: (1) Standard Blocking, (2) Sorted Neighborhood, (3) Q-gram Indexing, (4) Suffix-Array Based, (5) Canopy Clustering and (6) String-Map Based.

The second step consists of comparing entities of the same block. For each couple of entities, a similarity is produced. This step is very specific to each matching approach. Generally, entity matching approaches employ approximate similarity measures on attributes' values to compare entities. The similarity measures have two strategies to compute the similarity between a couple of entities.

1. Element-based: Similarity measures that compare entities in isolation [SGV06]. This means that, each couple of entities are compared in isolation to estimate their similarity.

2. Context-based or Neighborhood-based: Similarity measures consider the neighbors of compared entities. They utilizes semantic relationships between different entities to improve the similarity approximation [TR07, DHM05].

In addition, some approaches apply only one similarity measure on a specific attribute [SKS+10]. Some others apply several similarity measures on several attributes and then combine the similarities in order to obtain a general view of the compared entities [TKM02]. The combination of several measures includes three main strategies:

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

1. Numeric-based: numerically combining the values of several similarity measures [SGV06, RM08]. This combination may be done in a variety of ways such as counting the weighted average or using probabilistic considerations. A final decision may then be made using the combined value.

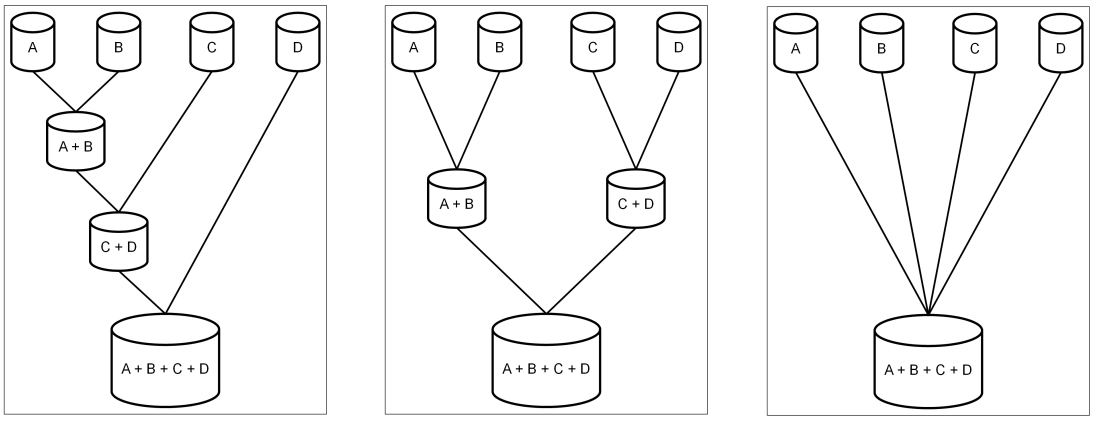
2. Rule-based: logically combining several similarity measures [ARS09, SSL12]. For example, two spatial entities may be considered matching if the similarity value of their *names* is above a given threshold and the distance between them is less than a given value.

3. Composite-based: combining the results of independently executed matchers [SKSD06]. For example, the final result is the union or the intersection of the results obtained by several similarity measures independently executed.

The final step in the matching phase consists in selecting the corresponding entities using their similarities. To do so, a decision algorithm is needed. A basic method is that we select the couple with the highest similarity, but this latter may not always be sufficient enough to consider two entities as corresponding. The most used decision algorithm in entity matching is the threshold-based [KTR10]. This latter serves as a baseline to indicate if two entities are corresponding or not [SGV06, SKS+10, SSL12]. If the similarity exceeds a given threshold, then entities are selected as corresponding. The similarity threshold should be provided as a parameter in the pre-matching phase. The choice of the threshold is not obvious, the result of the matching approach depends strongly on that choice. At the end of the matching phase, a list of corresponding entities and a list of singleton entities are produced.

Matching several datasets has been discussed [SKS+10]. A 2-join algorithm consists in taking more than two datasets and matching them two-by-two in order to produce one virtual dataset. There are three approaches to match several datasets namely serial join, hierarchical join and holistic join. The first starts by joining two datasets. Then, in each step, a dataset that has not yet been used is joined with the result of the previous step. For example, in Figure 2.4a, we consider four datasets A, B, C, and D; a possible serial join is to match A with B, then match the result with C, and finally, match D to the virtual dataset generated from matching A, B, and C. The second approach partitions the input datasets into pairs. Then, each pair is replaced with the matching results of its datasets. This process is repeated for the new produced datasets until one dataset remains. For example, in Figure 2.4b, we consider four datasets A, B, C, and D. A possible hierarchical join of these datasets is to match in the first step A with B, and C with D. In the second step, we join the two datasets that were produced in the first step. Note that the results' quality varies depending on the order in which the input datasets

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés



(a) Serial join. (b) Hierarchical join. (c) Holistic join.

FIGURE 2.4: Matching more than two datasets.

have been matched. The third approach addresses the input datasets all together at the same time in order to produce a final dataset by one step (see Figure 2.4c).

2.2.2.3 Post-matching

Finally, the post-matching phase addresses the corresponding entities detected in the previous phase. These entities may be processed for different purposes. In case of de-duplication, the duplicated entities should be cleaned. For integration purpose, corresponding entities should be merged together in order to obtain new entities. To merge entities, a crafted algorithm is needed to decide how the values must be merged. Corresponding entities may also be used to assess the quality of spatial datasets [TKA07]. For instance, when matching one spatial dataset of known quality with a second dataset of unknown quality, the comparison between their corresponding entities allows estimating the quality of the second dataset.

2.2.3 Geospatial Entity Matching

The geospatial entity matching field is similar to entity matching but enhanced by a spatial aspect, which already has a long history of research [Dev97, DPS98]. The geospatial entity matching has similar process, specifications and requirements as entity matching. In addition, the spatial information may be used as a key for the blocking phase. Three spatial blocking methods are distinguished:

1. **Blocking distance:** an entity from a first source is compared to an entity from a second source only if the radial distance between them does not exceed a given blocking distance [SKS+10].

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

2. **Blocking bounding:** an entity from a first source is compared to an entity from a second source only if the second entity lies within the borders of a bounding box. This latter is determined around the first entity[SSL12].

3. **Blocking tiles:** it consists of dividing the entire world into equal sized rectangular tiles. Entities of different sources that are located in the same tile are considered in the same block and can be addressed together [PD96, DORB14].

Figure 2.5 shows our classification scheme together with some approaches. Approaches that use only terminological information can be used in our context, but we believe that spatial information can be a benefit to improve the matching results. Köpcke et al. present a survey that describes and compares several approaches that use only terminological information[KR10]. In this section, we will focus on approaches that use either only spatial information or combine terminological and spatial information. On this bases, the geospatial entity matching approaches are classified into three classes namely (i) approaches that utilize only spatial information without machine-learning, (ii) approaches that combine terminological and spatial information without machine-learning and (iii) approaches that combine terminological and spatial information using machine-learning. In the follow, we detail existing approaches of these classes.

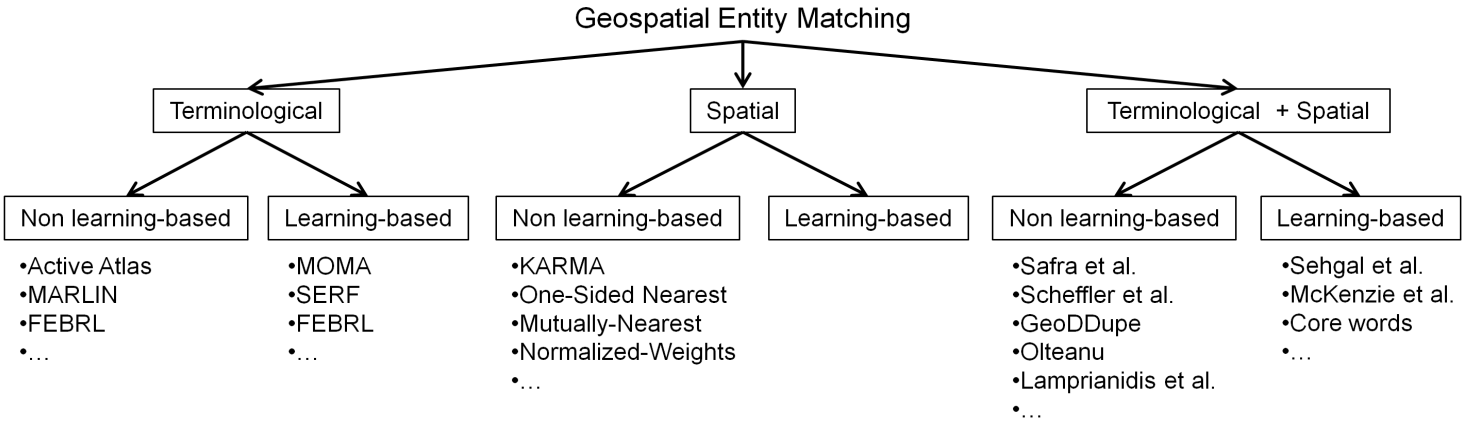


FIGURE 2.5: Classification scheme for geospatial entity matching approaches.

2.2.3.1 Approaches that use only spatial information

1. **Karma:**

   Zhang et al. propose a tool, named KARMA [KSA+12], to integrate heterogeneous geospatial data [ZCSK13]. Authors collect data from two providers namely Wikimedia and OpenStreetMap. The data concerns buildings in a given area, a building may be presented as a polygon or a point geographic object. KARMA

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

uses machine-learning techniques, such as conditional random field (CRF), to allow a semi-automatic alignment of providers' schemas to a pre-defined ontology. This alignment structures the data in the same format, which facilitates the integration process. Concerning the matching of entities, only spatial information without machine-learning is used to detected correspondences. Authors suppose that if two entities are polygons, and one of them is included in the other, then they are similar. Otherwise, if they are overlapping and the spheroid distance between them is smaller than a specified value, then they are the same. In the case that at least one of the entities is a point, the similarity between them is given by the following formula:

$$\text{similarity} = \left| 1 - \frac{distance}{\text{threshold}} \right|$$

In their experiments, authors use a threshold of 25 meters. If the similarity of two entities exceeds 0.97, then they are considered corresponding. Authors do not explain what happens when the distance exceeds the *threshold* defined in the formula. We assume that the similarity is set to 0 as the distance gets greater than the given threshold. Otherwise, this formula may produce a value between 0 and infinity. For example, consider two entities separated by 75 meters; their similarity is equal to 2. Concerning the merging of corresponding entities, authors propose the union of values of all sources. They also give users the ability to modify and reformat the values of merged entities.

2. **One-Sided Nearest join:**
   The One-sided Nearest join is commonly used in commercial geographic information systems [MSW00]. Given an entity $a_i \in A$, we consider that an entity $b_j \in B$ is the corresponding entity of a_i if b_j is the closest entity to a_i amongst all the entities of B. For example, Figure 2.6 shows the entities of two datasets $A = \{a_1, a_2\}$ and $B = \{b_1, b_2, b_3\}$. If A is matched to B, then both a_1 and a_2 will choose b_1 as corresponding entity (b_1 is the closest entity to a_1 and a_2), while b_2 and b_3 will be considered as singleton. Otherwise, if B is matched to A, then b_1, b_2 and b_3 will choose a_2 as corresponding entity (a_2 is the closest entity to b_1, b_2 and b_3), while a_1 will be considered as singleton. This approach produces n:1 correspondences, and it is asymmetry (i.e., the results of matching A with B are different from matching B with A). The performance of this method is evaluated with the two next approaches namely Mutually-Nearest join and Normalized-Weights method [BKSS04, BDK+05, SKS+10]. The evaluation is given with Normalized-Weights method (see below).

3. **Mutually-Nearest join (MN):**
   Authors of [BKSS04, BDK+05, SKS+10] propose the Mutually-Nearest join (MN).

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

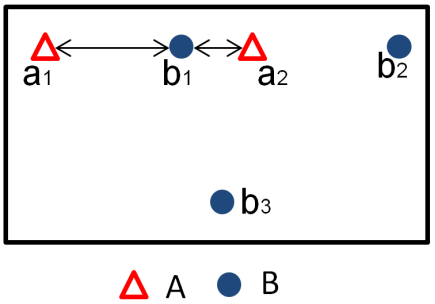


FIGURE 2.6: Example of One-Sided Nearest join: matching A to B.

This method is element-based and uses the Euclidean distance between entities. Given two entities $a_i \in A$ and $b_j \in B$, they are considered as corresponding entities if a_i and b_j are mutually nearest to each other and the distance between them is less than a given blocking distance. This means that, a_i is the closest entity to b_j amongst all the entities of A and b_j is the closest entity to a_i amongst all the entities of B. If b_j is the nearest entity to a_i but a_i is not the nearest one to b_j then a_i is considered as singleton (i.e., does not have a corresponding entity in B) and vice versa. For example, Figure 2.7 shows the entities of two datasets $A = \{a_1, a_2\}$ and $B = \{b_1, b_2, b_3\}$. Among all entities of B, b_1 is closest entity to a_1, but a_1 is not the closest entity to b_1, so a_1 and b_1 cannot be considered as corresponding entities. Similarly, among all entities of A, a_2 is the closest entity to b_2, but b_2 is not the closest entity to a_2, so a_2 and b_2 cannot be considered as corresponding entities. Otherwise, the two entities a_2 and b_1 will be considered as corresponding because they are mutually nearest to each other, while a_1, b_2 and b_3 will be considered as singletons. The evaluation of this approach is given with Normalized-Weights method (see below).

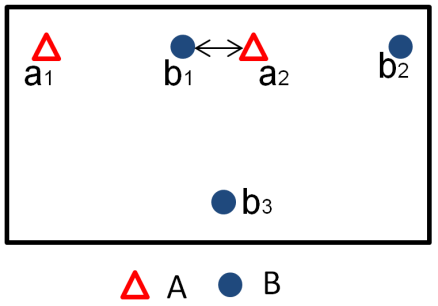


FIGURE 2.7: Example of Mutually-Nearest join.

4. **Normalized-Weights method (NW):**
   The Normalized-Weights method (NW) is proposed by the same authors of MN [BKSS04, BDK+05, SKS+10]. This method is a context-based approach that consists in computing the probability that two entities $a_i \in A$ and $b_j \in B$ choose

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

each other as correspondence. This probability is based on the Euclidean distance between entities. The shorter the distance between entities means the closer to 1 the probability value for coordinates will be. If the distance between a_i and b_j is greater than a blocking distance β, then the probability that a_i chooses b_j is set to 0. Otherwise, the probability that a_i chooses b_j, denoted $P_{a_i}(b_j)$, is equal to the Euclidean distance between a_i and b_j, denoted $d(a_i, b_j)$, above the sum of β and the distances between a_i and all the remaining entities of B.

$$P_{a_i}(b_j) = \frac{d(a_i, b_j)^{-\alpha}}{\sum_{k=1}^{m} d(a_i, b_k)^{-\alpha} + \beta^{-\alpha}}$$

where m is the number of remaining entities of B and α is a *decay factor* that allows the increasing of the probability when the distance decreases. The probability that b_j chooses a_i is calculated similarly. In addition, they calculate the probability that a given entity does not choose any entity from the other set. The probability that the entity a_i does not choose any entity from B, denoted $P_{a_i}(\perp_B)$, is given by the following formula:

$$P_{a_i}(\perp_B) = \frac{\beta^{-\alpha}}{\sum_{k=1}^{m} d(a_i, b_k)^{-\alpha} + \beta^{-\alpha}}$$

The probability that an entity b_j does not choose any entity from A is calculated similarly. Based on this concept, a matrix is created (entities of A in rows and entities of B in columns), the score for each pair (a_i, b_j) is $P_{a_i}(b_j) \times P_{b_j}(a_i)$. In the last column, the score is equal to $P_{a_i}(\perp_B) \times \prod_{k=1}(1 - P_{b_k}(a_i))$. Similarly, in the last row the score is equal to $P_{b_j}(\perp_A).\prod_{k=1}(1 - P_{a_k}(b_j))$. Then, the rows and the columns of the matching matrix are normalized to one (i.e., the sum of each row and column is equal to 1), except for the last column and row. Finally, we obtain a matching matrix that contains a similarity score for each pair (a_i, b_j). Pairs whose similarity does not exceed a given threshold are filtered out. To better understand this method, consider the following example. Let $A = \{a_1, a_2\}$ and $B = \{b_1, b_2, b_3\}$ be two datasets and the blocking distance $\beta = 15$. The positions of the entities are shown in Figure 2.8.

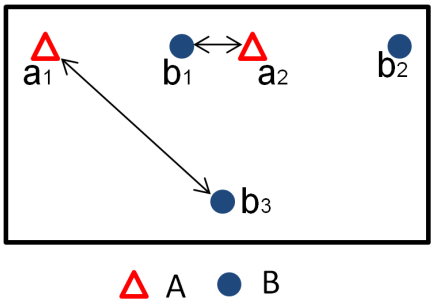


FIGURE 2.8: Example of NW: matching A and B with a 0.45 threshold.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

The distances between entities are given in the following table:

Distances			
	b_1	b_2	b_3
a_1	7	17	12.8
a_2	3	10	10.2

We use the above distances to compute the probabilities according to the above formula with $\alpha = 2$. The choice probabilities of A are presented in the next table:

Choice probabilities for A				
	b_1	b_2	b_3	$\perp_B$
P_{a_1}	0.66	0	0.20	0.14
P_{a_2}	0.82	0.07	0.07	0.03

The following table shows the choice probabilities of B:

Choice probabilities for B			
	P_{b_1}	P_{b_2}	P_{b_3}
a_1	0.15	0	0.30
a_2	0.82	0.69	0.48
$\perp_A$	0.03	0.31	0.22

The initial weights in the matching matrix are as follows:

Initial Weights				
	b_1	b_2	b_3	$\perp_B$
a_1	0.1	0	0.06	0.09
a_2	0.67	0.05	0.03	0
$\perp_A$	0	0.28	0.16	0

Finally, we start normalizing the matrix, in each iteration, we normalize the rows then the columns or vice-versa. In this example, the normalization algorithm terminates after 9 iterations and returns the following matrix:

After Normalization					
	b_1	b_2	b_3	$\perp_B$	Sum
a_1	0.27	0	0.47	0.26	1
a_2	0.72	0.18	0.1	0	1
$\perp_A$	0.01	0.82	0.43	0	
Sum	1	1	1		

For a 0.45 threshold, the algorithm returns the join sets $\{a_1, b_3\}, \{a_2, b_1\}$ as correspondences and $\{b_2\}$ as singleton. If a 0.6 threshold is used, then only the sets

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

$\{a_2, b_1\}$ and $\{b_2\}$ are returned. A drawback appears in this case because a_1 and b_3 have not been chosen neither as correspondence nor as singleton.

Experiments in [BKSS04] show that NW and the MN perform better than the One-Sided Nearest Join. These three approaches have been evaluated by matching automatically generated spatial datasets and by matching two real datasets. The real datasets contain 86×74 entities which represent hotels in one city. Two small and non-characterized datasets are not sufficient enough to create a conclusion about the performance of matching approaches. On the other hand, results of matching generated datasets depend on the generator algorithm which does not necessarily reflect realistic conditions. For instance, as a first step, authors generate 1000 entities. For the second step, the user specifies an error interval and the number of entities for two datasets which are namely source and target. A source or target entity is randomly associated with one of the 1000 entities. The error interval controls how far a source or target entity is from its associated entity. Three pairs of datasets are randomly generated of sizes 100, 500 and 1000 entities in each dataset, respectively. Authors suppose that the pair that contains 100 entities has a small overlap with a very high probability, the pair that contains 500 entities has a medium overlap and the pair that contains 1000 entities has a complete overlap. The conclusions made with these datasets are not necessarily true, since the pair that contains 100 entities may have 100, 50 or 0 correspondences. Hence, the overlap may be complete, medium or null regardless of the number of entities (similarly for the pair that contains 500 entities). Nevertheless, NW and MN have been extended in [BDK+05, SKS+10] to holistically match three or more datasets at the same time.

2.2.3.2 Approaches that combine terminological and spatial information without machine-learning

1. **Safra et al.:**

   Safra et al. propose three algorithms to combine spatial and terminological similarity measures [SKSD06]. First and second algorithms are based on a composite-based combination (see Section 2.2.2.2). The first consists of detecting a first set of corresponding entities by applying a terminological similarity measure to one attribute. Then, it applies a spatial similarity measure to the remaining entities, i.e., entities that are not detected in the first step. The final result is the union of the results of the two steps. The second examines the intersection of the results of independently executed spatial and terminological similarity measures. The third algorithm starts by applying a terminological similarity measure. Then, for each

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

pair of entities that are not considered as corresponding, their Euclidean distance is multiplied by a given factor α. Note that increasing the distance between two entities lowers their spatial similarities. The second step consists in applying a spatial similarity measure with the new distances. Concerning the spatial information, authors use the similarity measures proposed in [BKSS04, BDK+05], while a simple equality function is used for the terminological information. A first experiment is done by matching two small datasets representing POIs of type *Hotel*. 28 entities have been collected from Google Maps and 39 entities have been collected from Yahoo Maps, in which 21 entities are corresponding. Result shows that the first algorithm gives the best performance. A second experiment is done by matching generated data. Authors conclude that if the values of terminological attributes are either accurate or missing, then combining the three algorithms gives the best performance. Otherwise, if terminological attributes always have values, but some of those values are inaccurate, then combining the first and the third algorithms gives the best performance. Unfortunately, the combination between algorithms has not been described.

2. **Scheffler et al.:**
   Scheffler et al. use a rule-based approach to combine several similarity measures [SSL12]. Authors use spatial information only for the blocking aspect. A preprocessing step is applied to normalize the names of POIs (e.g., lower casing, filtering stop words). Then, the Levenshtein distance is applied to the name attribute. If the similarity exceeds 90%, then entities are considered corresponding. Otherwise, names are converted to documents and the TF-IDF (Term Frequency times Inverse Document Frequency) cosine similarity is applied. If the similarity exceeds 50%, then entities are considered corresponding. Authors do not specify how the thresholds have been chosen. This approach is compared to two basic methods. The former, called Nearest Point of Interest (NP), use only the spatial information; it selects the nearest POI. While the latter, called Longest Common Substring (LCS), use only the name attributes; it selects the target whose title shares the longest common substring with the source entity. Two experiments are done by matching 50 POIs from Facebook places[14] and 50 POIs from Qype[15] located in Berlin with OSM POIs separately. Results show that the rule-based approach outperforms LCS, and LCS outperforms NP.

3. **GeoDDupe:**
   Kang et al. propose a semi-automatic tool, called GeoDDupe, to detect the corresponding geospatial entities [KSG07], which takes two sources of entities as input.

[14] Facebook Places: www.facebook.com/places
[15] Qype: www.qype.com

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

The user has to tune the tool by choosing a blocking algorithm, a similarity measure and a weight for each attribute; the final similarity score is the weighted average. Authors propose to use the Euclidean distance as a spatial similarity measure, but without giving any precision on how the distance is quantified as a similarity. Concerning the terminological information, several measures are proposed such as Levenshtein, Jaccard, Jaro, JaroWinkler, MongeElkan, etc. The potentially corresponding entity pairs are automatically detected if their similarities exceed a predefined threshold. Also, these correspondences are ranked according to their similarities. Then, the user has to analyze each pair of the top-K correspondences to make the final decision for whether it should be considered corresponding or not. In addition, Geoddupe gives a visual representation for each pair and their neighbors in order to facilitate the decision making. For instance, the potentially corresponding entities who share the same neighbors may be considered corresponding. This tool is available on demand[16].

4. **Olteanu:**

   The "Dempster-Shafer Theory" [Sha76] is also called the "Evidence Theory" or "Belief Theory." This mathematical theory allows the combination of evidence from different sources and arrives at a degree of belief that takes into account all the available pieces of evidences. Olteanu uses the "Dempster-Shafer Theory" to numerically combine several similarity measures in order to match two spatial datasets [OR07]. Let A and B be two spatial datasets, where each entity $b_j \in B$ is a possible candidate to correspond to a given entity $a_i \in A$. For each pair (a_i, b_j), two independent weights are calculated based on the similarities of *location* and *name* attributes. Then, the two weights are combined using Dempster's rule in order to compute a belief mass (i.e., global similarity). Authors use the Levenshtein distance to measure the similarity of the *name* attribute. Concerning the spatial attribute, they suppose that the closer the entities to each other, the more likely there would be a high similarity. Yet, authors do not give any additional information about how to convert this similarity into a weight between 0 and 1. For their experiments, authors used two real datasets about geographic reliefs to show some use cases of applying the "Belief Theory." However, they did not give the performance results of the whole matching.

5. **Lamprianidiset al.:**

   Lamprianidiset al. propose a rule-based geospatial entity matching approach in order to identify emerging regions of interest, i.e., geographical areas with high density of POIs of certain types [LSPP14]. In a first stage, authors create a reference POI types list and manually mapped the POI types of several LBS providers

[16] Geoddupe: http://linqs.umiacs.umd.edu/projects/geoddupe/ [Accessed: June 2016]

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

to their reference list. Then, they match entities of the same type by comparing their location coordinates and names. A blocking tiles method is used to cluster entities, then two entities being compared are considered as corresponding only if the distance between their locations is lesser than 0.002° (approximate ≈ 200 meters) and the Levenshtein similarity of their names is higher than 0.7. For their experiments, authors collect and match thousands entities of several POI types from several providers. For each pair of providers, results show the percentage of matching between the collected sets for each POI type. But, the evaluation of the matching process is not given.

2.2.3.3 Approaches that combine terminological and spatial information using machine-learning

1. **Sehgal et al.:**
   Sehgal et al. use machine-learning techniques to learn how to numerically combine several similarity measures [SGV06]. String similarity measures (e.g., Jaccard, Levenshtein) are used to compute the similarity of terminological attributes. Concerning the spatial measure, the similarity between two entities is equal to the inverse of the Euclidean distance between them. This spatial similarity makes it difficult to estimate the similarity between two entities. For example, for a distance equal to 1, the similarity is 1 and for a distance of 2 meters the similarity is 0.5. Concerning these authors' experiments, they use two real datasets that represent POIs such as cemeteries and airfields. The former contains 202210 entities distributed in Afghanistan, while the latter contains only 2096 distributed in *Helmand Province, Afghanistan*. The first experiment consists of comparing the results of matching using a single similarity measure. The Levenshtein applied to the name attribute outperforms the other measures. The second experiment considers a variety of learning-based methods including logistic regression, support vector machines and voted perceptron to discover out how to combine and tune several similarity measures. This experiment has been repeated by changing the ratio between the number of corresponding and singletons entities. Results show that a learned weighted average, using logistic regression, outperforms all the other methods, whatever the ratio is. The datasets used in experiments are not challenging because such POIs with large geographic area, do not express an interesting heterogeneity. For instance, two different positions for an airfield can easily be detected because it is impossible to find another large POI inside the airfield that may confuse the choice. On the other hand, if we request hotels in some quarter in Paris from two different LBS providers, the results may be hundreds of entities in a small area, and the matching between them would be a hard task.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

2. **McKenzie et al.:**

   McKenzie et al. also propose a learning-based method for geospatial entity matching [MJA13]. Concerning the terminological information, similarity measures like Levenshtein distance, Metaphone and Latent Dirichlet Allocation (LDA). On the other hand, authors use the great circle distance to measure the spatial similarity, with the smallest distance representing the best estimated match, but without giving any precision on how the similarity is quantified. A blocking distance is used for this method, this distance has been specified according to the farthest distance between two corresponding entities from their datasets, which is equal to 1000 meters. As a first step, authors evaluate the performance of matching using one single measure. Concerning the POIs' names, authors measure their similarity using Levenshtein or Metaphone, while LDA is used for user-contributed reviews about POIs. Results show that the Levenshtein and Metaphone measures separately applied to the name attribute outperforms the other methods. In the second step, they propose to combine several measures using three algorithms: (1) unweighted combination that considers the average of several measures, (2) ordinal weight that specifies a weight for each measure according to its performance; the higher the performance, the higher the weight and (3) a learned weight using the logistic regression method. Experiments are done by matching POIs collected from location-based social networks (LBSN) such as Yelp and Foursquare. 200 entities have been collected from Yelp and manually matched to entities from Foursquare where 140 correspondences have been found. Results show that the three algorithms are equivalent and give more accuracy than applying a single measure. They achieve 87% of accuracy when combining three measures namely Levenshtein, Metaphone and spatial similarity. After adding a fourth measure namely LDA, the accuracy increases to 95%, 94% and 97% for the first, second and third algorithms respectively. The datasets used in this paper were chosen randomly. Also, the context of LBS is very dynamique; providers databases contain millions of entities that have different types and located using different strategies. Hence, use small datasets for learning purposes makes it hard to conclude an hypothesis.

3. **Core Words:**

   Dalvi et al. propose an approach to address the challenging problem of deduplicating places in Facebook's database using the name attribute and the spatial context [DORB14]. This approach presents a language model that tokenizes the name of POIs by finding the core and background terms. For example, consider the POI *Sam restaurant Time Square* that is located near the Time Square, New York. The tokens *restaurant Time Square* are considered as a background terms because they indicate the properties of the POI (i.e., address and type), while the

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

token *Sam* consists of the core term because it indicates the POI name. Authors use machine-learning techniques and spatial context (i.e., location and address) to find the core and background terms. Then, in order to estimate the similarity between two entities, they calculate the probability that two POI names have the same cores. A subset of places database at Facebook is used for experiments. The subset is restricted to POIs in USA; it contains 7k candidate pairs, of which around 2K are duplicates. Evaluation's results show that this approach outperforms baseline approaches such Levenshtein distance and TF-IDF cosine similarity applied to place names.

Table 2.2 summarizes the above geospatial entity matching approaches. Unfortunately, most of these approaches' implementations has not been made available for researchers.

2.2.4 Data Fusion

After matching the corresponding entities from the different sources, a final step requires merging their values to obtain one single entity. Ideally, the merged entity should include more complete and accurate information. However, corresponding entities may have inconsistent values, such as two different phone numbers. Data fusion aims to solve the conflicted issue by identifying the true value amongst multiple observed values drawn from different sources [DGH+14]. Basic approaches have been proposed for data fusion such as considering value from the most recent up-to-date source or taking the average, maximum or minimum for numerical values [BN08, DN09]. However, the baseline strategy for data fusion is the voting. Among conflicting values, each value has one vote from each data source. The value with the highest vote count is considered as the correct value. A classic approach is to consider the most frequent value. In other words, the correct value is the one provided by the largest number of sources. More advanced approaches have been proposed for voting, they can be classified into three classes:

1. Relation-based: methods that consider the relationships between sources [DBS09, DBHS10]. For instance, a source that copies data from another source is not allowed voting for its copied values.

2. Quality-based: methods that consider the quality of sources [PR10, DSS12]. These methods compute a higher vote count for a high-quality source. The quality of the sources may be set manually or automatically. For instance, if data are extracted from web pages, the quality of the sources is then estimated using the rank of a web search engine [BP12].

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

Class	**Approach**	**Authors / Ref.**	**Similarity measures**		**Combination**	**Learning**	**Blocking**
			Spatial	**Terminological**			
Spatial without machine-learning	KARMA	Zhang et al. [ZCSK13]	similarity = $\left\|1 - \frac{d}{\text{threshold}}\right\|$				
	One-Sided Nearest	Minami et al. [MSW00]	Distance-based Nearest entities				
	Mutually Nearest (MN)	Beeri et al. [BKSS04]	Distance-based Mutually Nearest entities				Blocking distance
	Normalized Weights (NW)	Beeri et al. [BKSS04]	$P_{a_i}(b_j) = \frac{d(a_i,b_j)^{-\alpha}}{\sum_{k=1}^{m} d(a_i,b_k)^{-\alpha+\beta-\alpha}}$				Blocking distance
Spatial & terminological without machine-learning	Safra et al.	Safra et al. [SKSD06]	One-Sided Nearest Mutually-Nearest Normalized-Weights	Simple equality	Composite		Blocking distance
	Scheffler et al.	Scheffler et al. [SSL12]	(Spatial data are used for blocking)	Levenshtein TF-IDF	Rule-based		Bounding box
	GeoDDupe (semi-automatic)	Kang et al. [KSG07]	Distance-based (no precision)	Levenshtein Jaccard ...	Numeric-based		Several
	Olteanu	Olteanu [OR07]	Distance-based (no precision)	Levenshtein	Numeric-based (Belief Theory)		Blocking distance
	Lamprianidis et al.	Lamprianidis et al. [LSPP14]	Distance-based	Levenshtein	Rule-based		Tiles
Spatial & terminological with machine-learning	Sehgal et al.	Sehgal et al. [SGV06]	similarity = $\frac{1}{d}$	Levenshtein Jaccard JaroWinkler	Numeric-based	Logistic regression	
	McKenzie et al.	McKenzie et al. [MJA13]	Distance-based (no precision)	Levenshtein Metaphone LDA	Numeric-based	Logistic regression	Blocking distance
	Core Words	Dalvi et al. [DORB14]	(Spatial data are used for analyzing terminological data)	Probability-based			Tiles

TABLE 2.2: State of the art of geospatial entity matching approaches.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

3. Context-based: this class, also known as knowledge fusion [DGH+14], is an extension of quality-based data fusion. In addition to the quality of sources, knowledge fusion considers the context from where data are extracted and methods used for data extraction. For instance, requesting information from a well known API is not the same as extracting information from random web pages using text-wrapper.

Xian Li et al. propose a survey that compares advanced data fusion methods [LDL+12]. The authors focus on Deep Web Data, where each website is a source of information. In their experiments, 13 data fusion methods are applied to data concerning *Stock* and *Flight*. Authors conclude that data fusion methods are unstable with respect to data, there is no method that definitely dominates others on all datasets. In other words, the result of a fusion method strongly depends on the data; each situation has its advantages and disadvantages. Such techniques have also been extended to handle the volume of data. Authors of [LDOS11] propose a framework named SOLARIS for online data fusion. The approach is based on the sources' qualities. Instead of waiting for data fusion to complete and returning answers in a batch, SOLARIS starts by querying the highly ranked sources, then refreshes the answer as it queries more. It is not necessary to query all sources, once a value receives a high probability, SOLARIS returns it as the correct value, and then it starts analyzing new values.

2.3 Data Matching Benchmark

Benchmarks are used to compare approaches among each other by evaluating their effectiveness and efficiency. A benchmark defines a set of test cases, which provides the basis for determining the quality and performance of an approach. These test cases are examined using characterized datasets and a list of metrics that measure the degree of success of an approach to handle the matching issue. A single test case is designed to evaluate a particular aspect of matching approaches. While the overall quality of the matching approaches can be determined by observing the quality through all test cases. The goal of this kind of evaluation is to find out the weak and strong points of an approach and to improve it.

2.3.1 Datasets

In **ontology matching**, the objective is to discover semantic correspondences between concepts and properties of different ontologies [ES07]. Ontology alignment researchers

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

have designed Ontology Alignment Evaluation Initiative (OAEI[17]) to compare ontology alignment tools. It includes ontologies describing the domain of bibliography; the datasets fulfill various criteria. For instance, the *benchmark* dataset gathers many ontologies in which a specific type of information has been altered (modifications, deletions, etc.). Each of the altered ontologies has to be matched against the initial ontology. This facilitates the detection of weaknesses of a tool according to available information. Other datasets might be very specific like the *Food and Agriculture Organization* ontologies, which usually require external resources such as dictionaries to obtain acceptable results. In 2013, the initial datasets have been extended with synthetic ones [ERT13]. ONTOBI benchmark has been developed for evaluating instance-based ontology matching systems [Zai10, ZCV10]. A large amount of instances have been collected from external sources (e.g., Wikipedia data). Simple modifications (e.g., spelling mistakes, changed format) and complex modifications (e.g., synonyms, expanded structure) are applied to source instances to build heterogeneous datasets. The different combinations of modification produces 16 distinct test cases.

In **schema matching and mapping**, as defined before, it concerns the discovery of correspondences between schema elements as well as the mapping functions to transform source instances into target instances [BBR11]. The community has designed benchmarks for evaluating these two tasks. XBenchMatch enables the assessment of schema matching tools [DB14]; it includes a classification of task-oriented datasets.STBenchmark aims at evaluating the quality of the mapping functions [ATV08]. Datasets are gathered by compiling a source schema as input and applying several transformations (e.g., copying, flattening). In addition, they can be enriched using instance generators, which can be tuned with configuration parameters (e.g., kinds of join, nesting levels).

The **entity matching** task, which is directly related to spatial entity matching, consists of discovering correspondences between equivalent objects. It also benefits from two benchmarks. The former proposes a set of four datasets about e-commerce and scientific publications [KTR10]. These static datasets were used to compare entity matching tools. On the other hand, EMBench is based on importing existing entities (e.g., from Wikipedia[18], IMDB[19] and Amazon[20]) and applying modifiers to their features (e.g., abbreviation, synonyms) [IRV13]. These changes generate a set of modified entities, which form an entity matching dataset when grouped with the original entities. Although these two entity matching benchmarks are useful in most contexts, they are insufficient when dealing with spatial matching.

[17] OAEI: http://oaei.ontologymatching.org [Accessed: June 2016]
[18] Wikipedia: http://www.wikipedia.org
[19] IMDB: http://www.imdb.com
[20] Amazon: http://www.amazon.com

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

To the best of our knowledge, there is no benchmark for evaluating geospatial entity matching approaches. Kang et al. propose a tool to detect the corresponding spatial entities [KSG07]. This tool may be interesting to create a training dataset but it is not characterized enough for a benchmark. Beeri et al. implemented a random dataset generator to evaluate their matching approach [BDK+05]. They generate two datasets of spatial entities in which some entities are corresponding. Unfortunately, generated entities are described only by spatial information (longitude and latitude) because their proposed matching approach exploits only the spatial information to detect the correspondences. Nomao labs[21] have published a dataset containing the results of POIs comparison from multiple sources without sharing the POIs values. Two POIs being compared are represented by their IDs, a label to indicate if they correspond and the similarity scores for attributes' values. Unfortunately, location coordinates (i.e., latitude and longitude) are compared using terminological similarity measures. This dataset allows an individual to discover the best similarity measure for each attribute to detect the correspondences. Otherwise, the datasets used in geospatial entity matching papers are not made fully available for various reasons including confidentiality issues. A few attempts are available, such as datasets about restaurants[22]; yet, they cannot be exploited either. Some of them are not challenging (e.g., a simple equality metric applied to the phone numbers in the restaurant dataset discovers all the correct corresponding entities). Also, a specific dataset may be required, for instance to include all POI types (e.g., restaurants, museums, mountains) or all entities from a given area.

2.3.2 Metrics

The performance of matching approaches can be evaluated by measuring users effort in the pre-matching and post-matching phases [DB14] or by measuring the amount of memory allocated by an approach and its executed time [ATV08]. Also, there are several metrics to measure the result's quality of these approaches. The most common metrics used in matching approaches are the standard performance measures that come from the information retrieval domain, *Precision*, *Recall* and *F-measure* [Rij79]. These measures evaluate the quality of a matching approach by comparing its result to the actual result. For instance, consider two datasets, namely source and target, for which a list of actual correspondences, called ground-truth, is known in advance. The correspondences returned by a matching approach are compared to the ground-truth correspondences in order to measure how successful the matching approach detects the expected answer.

[21] Nomao Database: http://archive.ics.uci.edu/ml/datasets/Nomao [Accessed: June 2016]
[22] Restaurants datasets: http://www.cs.utexas.edu/users/ml/riddle/data.html [Accessed: June 2016]

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

These measures are formally defined as follows. The set of derived correspondences consists of True Positives (TP) and False Positives (FP), the former refers to the true correspondences that are correctly identified by the matching approach. The latter refers to the wrong correspondences detected by the matching approach. In the other hand, the False Negatives (FN) are true correspondences according to the ground-truth, that have not been identified by the matching approach. In contrast, True Negatives (TN) are false correspondences according to the ground-truth, which have been correctly discarded by the matching approach. Table 2.3 classifies the contingency of evaluation measures' base.

Matching approach \ Ground truth	Correspondences	Non correspondences
Correspondences	True Positive (TP)	False Positive (FP)
Non correspondences	False Negative(FN)	True Negative (TN)

TABLE 2.3: Contingency table of evaluation measures.

Precision calculates the proportion of correct correspondences detected by the matching approach among all detected correspondences. Using the notations of Table 2.3, the *Precision* is given by formula (2.2). A 100% precision means that all correspondences detected by the matching approach are true.

$$Precision = \frac{TP}{TP + FP} \tag{2.2}$$

Recall computes the proportion of correct correspondences detected by the matching approach among all correct correspondences. The *Recall* is given by formula (2.3). A 100% recall means that all correct correspondences have been found by the matching approach.

$$Recall = \frac{TP}{TP + FN} \tag{2.3}$$

F-measure is a trade off between *Precision* and *Recall* and it is calculated with the formula (2.4). The β parameter of formula (2.4) regulates the respective influence of *Precision* and *Recall* ($\beta \in \mathbb{R}^+$). It is often set to 1 to give the same weight to these two evaluation measures.

$$F - measure\,(\beta) = \frac{\left(\beta^2 + 1\right) \times precision \times recall}{(\beta^2 \times precision) + recall} \tag{2.4}$$

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

2.4 Visualization of Geospatial Integrated Data

Cartography can be defined as a science dealing with representation, communication and exploration of spatial knowledge; it mainly concerns design and use of the map [MT94]. Map design's goal is to address practical requirements, e.g., the need in information systems for efficient models of geographic objects and their spatial relationships for various urban planning requirements [Lau01]. Cartography has heavily changed since Information Technology appeared; paper has been replaced by screen. Besides computers enable quick processing of big data, the Internet widely broadcast maps through an easy and fast access [vE00], as illustrated by cartographic services. Web have made maps dynamic and real time modifiable to fit context and user requirements. For instance, LBS providers use interactive map to represent POIs and deliver their related information to end-users and tourists.

Our context considers inconsistent geospatial data collected from several LBS providers and having distinct quality. Integration and merging of such data may improve the quality but uncertainty sill remains on process's output. This uncertainty refers to the level of reliability of integrated data, it is a complex notion composed of a large scale of doubt and incompatibility [GE00]. Thus, a map is often seen as absolute truth for the general public. Information about reliability of data on a map is essential for an objective analysis. Evans specifies that "we get responsibility towards users to provide reliability information on cartographic data and their representation, so map-based decisions may be done knowing map limits" [Eva97].

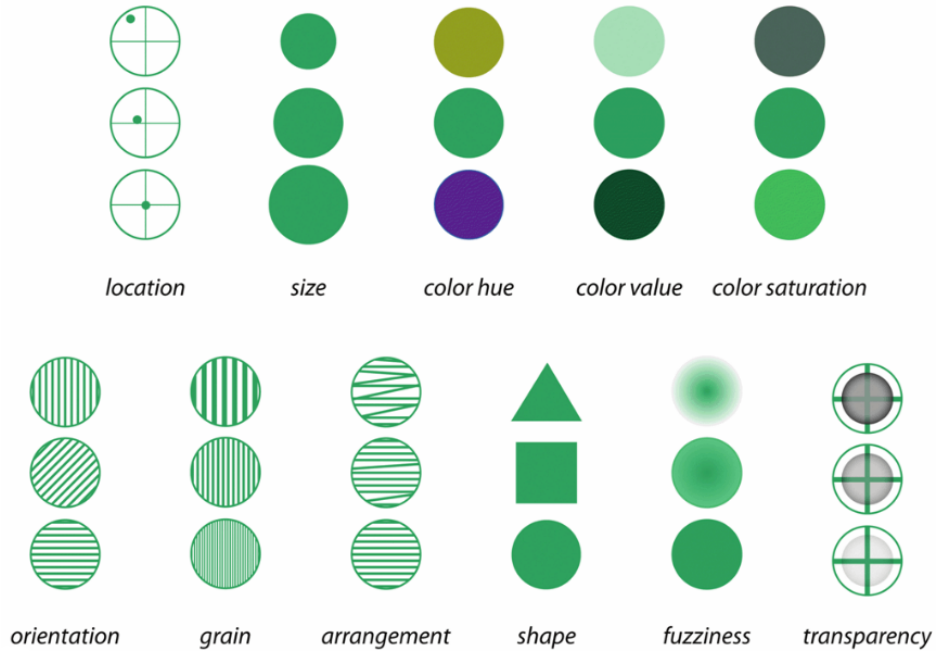
According to Thomson et al., the uncertainty of geospatial data concerns three components namely spatial attributes, time and terminological attributes (i.e., non spatial attributes) [THM+05]. Authors pair these components with nine categories:

1. **Accuracy/Error:** difference between observation and reality.
2. **Precision:** exactness of measurement.
3. **Completeness:** extent to which info is comprehensive.
4. **Consistency:** extent to which info components agree.
5. **Lineage:** conduit through which info passed.
6. **Currency/Timing:** temporal gaps between occurrence, info collection and use.
7. **Credibility:** reliability of info source.
8. **Subjectivity:** amount of interpretation or judgment included.

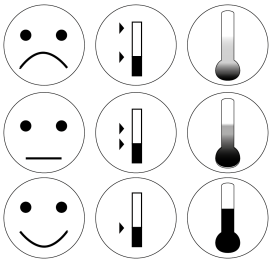
Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

9. **Interrelatedness/Trustworthiness:** source independence from other information.

On this basis, MacEachren et al. make an empirical study to characterize the kind of visual signification that is appropriate for representing those different categories of uncertainty [MRH+05]. The authors use a set of abstract visual variables collected from [Ber83, Mac95, Mor74] and shown in Figure 6.3: *Location, Size, Color Hue, Color Value, Color Saturation, Orientation, Grain, Arrangement, Shape, Fuzziness and Transparency.* Their symbol sets are points and for each visual variable, three degrees are specified coming from high to low uncertainty. In addition, they added iconic symbols such as Smiley, Filled bar with Slider, and Thermometer (Figure 2.9b) to compare their efficiency with respect to abstract and geometric symbols.



(a) Set of visual variables defined by [MRO+12].



(b) Smiley, Filled bar with Slider, and Thermometer icons proposed by [MRO+12].

FIGURE 2.9: Set of visual variables to represent uncertainty.

Two empirical perception tests were realized to judge the suitability of different symbol sets for representing variation in uncertainty by manipulating one single visual variable

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

(i) for each category of uncertainty regardless of the components of geographic information and (ii) for each component of the geographic information with respect to three specific categories namely accuracy, precision, and trustworthiness. Authors conclude that abstract visual variables lead to quicker judgments than iconic visual variables. Also, these latter only work well if users understand both the aspect of uncertainty being signified and the metaphor upon which the icon are based. Anyhow, *Location*, *Size* and *Fuzziness* variables are the most appropriate to portray spatial uncertainty. Whilst, *Smiley*, *Filled bar associated with Slider* and *Thermometer* are interesting to portray terminological uncertainty. Finally, *Fuzziness*, *Location* and *Color value* are well suited to portray global uncertainty.

2.5 Conclusion and Positioning

The main concern of this thesis is to improve the quality of LBS that provide POIs for tourists. To do so, existing LBS must be integrated in order to obtain more complete and correct information. The subject areas that hold priority of this thesis are outlined in this section.

First, a technical overview was given to highlight how the current LBS providers share their data with end-users and developers. A common method is used by all LBS providers to share POI; it consists of on-demand access through web service API. Although the providers' APIs are surrounded by several technical limitations, they are still sufficient for our research project.

Concerning the data of LBS providers, assessing the quality of original data sources has already been studied [TKA07]. But, this thesis uses the LBS data sources without considering their qualities. Concerning the multilingual issue, as mentioned before, OpenStreetMaps already offers POIs in different languages. Recently, some providers such as Google Maps, started to add filters to their API to allow users to request POIs in a specific language; this filter is not in production yet. In addition, all providers represent POIs as a point geographical object (0D), even those who refer to lines or polygons. On this basis, we focus on integrating POIs that are represented as punctual objects and described by the same language.

Secondly, we investigated the data integration where three tasks are required namely schema matching, entity matching and data fusion.

Schema matching allows us to discover the relations between the structures of data sources to be integrated. Quite a large number of papers have investigated the schema matching. As was mentioned earlier, the LBS providers have different schemas that need

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

to be matched. Currently, this matching can be done manually due to the simplicity of the LBS providers' schemas. This means that there is no need for semi-automatic or fully-automatic approaches to handle the schema matching task.

Then, we focus on the geospatial entity matching field. Possible inconsistencies of terminological data have been classified, such as synonym, hyponym, etc. Similarly, the inconsistencies in spatial context also need to be classified, which has not yet been accomplished. On the other hand, several approaches based on different concepts have been proposed for geospatial entity matching. These works may exploit terminological information, spatial information or combine both in order to find the corresponding entities. Also, some works have proposed to use machine learning techniques to produce better results. Unfortunately, most of these approaches have not been implemented into prototypes or frameworks. Also, several papers propose using spatial similarity measures in their approaches. Those measures are often not well detailed or face drawbacks.

The last step in data integration consists in merging the corresponding entities of several sources to create new entities with more complete information. Corresponding entities may have contradictory values, such as different phone numbers or different websites. The data fusion task helps to decide which value to pick for the new entity. According to the state of the art, all advanced data fusion approaches are based on the quality of data sources. But, as we mentioned, this thesis does not consider the quality of LBS providers.

On these bases, we intend to adapt existing approaches of data integration in order to elaborate a process for matching and merging geospatial entities from LBS providers.

Thirdly, geospatial entity matching approaches have been evaluated in different contexts using different datasets. These datasets are often times small, not enough characterized, chosen randomly and not made available for researchers. This makes the results of the approaches' experiments incomparable and difficult to understand the degree these approaches can handle the matching issue. In addition, we highlighted the absence of a benchmark to evaluate the geospatial entity matching approaches. This lack of a benchmark does not facilitate a fair and accurate comparison of different geospatial matching approaches. We also argue that the properties of a dataset are useful, both for understanding whether a geospatial entity matching approach is effective or not, and for using appropriate characterized data when needed. On this basis, this thesis intends to propose a benchmark and describe the necessary specifications that allow a correct evaluation for results'quality of geospatial entity matching.

Finally, LBS providers use interactive maps to allow users searching for POIs. The merging of contradictory entities from several sources may not be 100% reliable, and

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

users need to be informed about the certainty of the data for which they are searching. We highlighted some studies related to the visual presentation of data uncertainty that can be used for LBS purposes. Empirical studies have been done to characterize the kind of symbols that are appropriate for representing different categories of uncertainty. In this thesis, we consider existing solutions and adapt them to suit our context. Then, we validate our choices using psycho-cognitive tests. In addition, we intend to investigate the interest of such uncertainty to tourists and how it affects their decisions to find the desirable POIs.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

Chapter 3

Taxonomy

Contents

3.1 Preliminary Definitions 47
3.2 Entity Sets Characterization 50
3.3 Taxonomy of Inconsistencies 52
3.3.1 Schema Differences 52
3.3.2 Terminological Inconsistencies 54
3.3.3 Spatial Inconsistencies 57
3.3.4 Entity's Availability 59
3.4 Impact of Taxonomy 60
3.4.1 Impact of Schema's Inconsistencies 61
3.4.2 Impact of Spatial and Terminological Inconsistencies 62
3.4.3 Impact of Availability's Inconsistencies 62
3.5 Combination of Inconsistencies 64
3.5.1 Simple Combinations 64
3.5.2 Complex Combinations 66
3.6 Conclusion 68

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

In this chapter, we formalize a taxonomy that categorizes all kinds of inconsistencies concerning the LBS providers. Identifying the inconsistencies helps in both investigating statistics about the LBS providers' entity sets and evaluating integration's performance in different situations. The inconsistencies can appear between entities that refer to the same POI (i.e., corresponding entities), these would be named differences. For example, two corresponding entities can have different locations. Also, inconsistencies can appear between entities that do not refer to the same POI but share similar properties, these would be named resemblances. For example, two distinct entities can have identical location. In addition, these inconsistencies may occur within the entity set of one single provider, denoted as Intra-Inconsistency class, and between the entity sets of several providers, denoted as Inter-Inconsistency class. The former class helps evaluate the quality of the entity set of one provider such as completeness and redundancies. The latter helps compare the entity sets of several providers.

We start by introducing preliminary definitions that describe a model of LBS providers and identify the factors that characterize their entity sets. Then, the taxonomy of inconsistencies is defined using this model. Finally, in order to understand how inconsistencies impact the results' quality of integration, we intend to analyze their impact on the *Precision* and *Recall*.

3.1 Preliminary Definitions

It is necessary to represent the context of the LBS in order to construct a process to integrate them. In this section, we illustrate a model that describes the LBS context of multi-providers.

Definition 3.1. Point of Interest (POI)
A POI is a geographical object described by a set of properties. Among these properties, there is a name, a type (e.g., restaurant, castle), a location (positioning coordinates) and a geometric shape such as a point, line or polygon. It is defined by the tuple:

$$POI = (name, type, coordinates, shape)$$

For example, the tuple p_{ET} below represents the *Eiffel Tower* POI:

$$p_{ET} = \text{(Eiffel Tower, Tourist, (48.858439, 2.294474), Point)}$$

Let us consider the following set $\mathbb{P} = \{p_1, ..., p_q\}$ that contains all of the POIs of the real world where q is the number of POIs. Each LBS provider offers a set of entities that refer

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

to a subset of existing POIs. Currently, the APIs of LBS providers represent the entities with a geometrical point. Regarding the entities that refer to POIs with large areas, they are approximated by points such as computing their center of gravity [BDK+05], locating the entrance gate of POI or arbitrary choosing any point in the POI area. The entities offered by a provider are derived from a specific schema of that provider.

Definition 3.2. Schema of a provider
The schema $\mathbb{S}_k$ describes the structure of entities offered by the provider k. It is defined by:

$$\mathbb{S}_k = \mathbb{I}_k \cup \mathbb{L}_k \cup \mathbb{G}_k \cup \mathbb{A}_k \cup \mathbb{B}_k$$

where

- $\mathbb{I}_k = \{id_k\}$ is an internal identifier attribute that represents a given entity for the provider k.
- $\mathbb{L}_k = \{\text{LONGITUDE}_k.label, \text{LATITUDE}_k.label\}$ is a pair of spatial attributes that are mandatory and standing for the spatial coordinates.
- $\mathbb{G}_k = \{att\mathbb{G}_k^1.label, ..., att\mathbb{G}_k^u.label\}$ is another set of spatial attributes that are optionally provided where $u = |\mathbb{G}_k|$. We call them secondary spatial attributes because they may be either missing from some schemas or have null values.
- $\mathbb{A}_k = \{\text{NAME}_k.label, \text{TYPE}_k.label\}$ is a pair of terminological attributes that are mandatory and standing for the POI name and type. We call them primary attributes because they exist in the schemas of all providers and always have values.
- $\mathbb{B}_k = \{att\mathbb{B}_k^1.label, ..., att\mathbb{B}_k^r.label\}$ is another set of terminological attributes that are optionally provided where $r = |\mathbb{B}_k|$. We call them secondary terminological attributes because they may be either missing from some schemas or have null values.

Hypothetically, a schema of any provider k includes at least all attributes in $\mathbb{I}_k \cup \mathbb{L}_k \cup \mathbb{A}_k$. We note att_k^i any attribute of the schema $\mathbb{S}_k$. The attribute's label, denoted as $att_k^i.label$, refers to the name of the attribute. For instance, the attribute $\text{NAME}_k \in \mathbb{A}_k$ refers to the POI name and may have several labels such as *POI_name* or *place_name*. We denote $att_k^i.type$ as the abstract data type of att_k^i including the most frequent data types: string, number, array or associative array. Note that a schema may be static or dynamic. A static schema has fixed labels and structures. Conversely, labels and structures of dynamic schema can be modified. As an example, the provider OpenStreetMap has

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

a dynamic schema in which the user can add new attributes for entities. In contrast, Google Maps has a static schema, so that the number and the labels of the attributes are identical for all entities.

Definition 3.3. Entity of POI
An entity of a POI of a provider k, denoted by e, is an instance of the schema $\mathbb{S}_k$ and refers to one real-world POI $p \in \mathbb{P}$.

$$\begin{aligned} e \quad = \quad & \{(id_k.label, id_k.val), (\text{LATITUDE}_k.label, \text{LATITUDE}_k.val), \\ & (\text{LONGITUDE}_k.label, \text{LONGITUDE}_k.val), \\ & \left(att\mathbb{G}_k^1.label, att\mathbb{G}_k^1.val\right), ..., (att\mathbb{G}_k^u.label, att\mathbb{G}_k^u.val), \\ & (\text{NAME}_k.label, \text{NAME}_k.val), (\text{TYPE}_k.label, \text{TYPE}_k.val), \\ & \left(att\mathbb{B}_k^1.label, att\mathbb{B}_k^1.val\right), ..., (att\mathbb{B}_k^r.label, att\mathbb{B}_k^r.val)\} \end{aligned}$$

where u and r are the number of spatial and terminological secondary attributes, respectively. The entity set of a provider k is denoted by $\mathbb{E}_k = \{e_1, ..., e_n\}$ where n is the number of entities. We denote $\mathbb{E} = \bigcup_{k=1}^{z} \mathbb{E}_k$, the union set of z providers' entities sets.

Table 3.1 shows an example of two entities x and y that refer to the *Eiffel Tower* POI, p_{ET}. These two entities are offered by two different providers with two different schemas for which the model $\mathbb{I} \cup \mathbb{L} \cup \mathbb{G} \cup \mathbb{A} \cup \mathbb{B}$ is distinguish.

Model	**Entity x (offered by provider 1)**	**Entity y (offered by provider 2)**
$\mathbb{I}$	placeId : *250u09tu-4561...*	place_id : *ChIJLU7jZClu...*
$\mathbb{L}$	location : {position: [*48.858606*, *2.293971*]}	geometry : { location : { lat : *48.85837*, lng : *2.294481*}}
$\mathbb{G}$	shape : *point* altitude : 324	geometric_shape : *point*
$\mathbb{A}$	name : *Tour Eiffel* categories : [*7999*]	name : *Eiffel Tower* types : [*point_of_interest*]
$\mathbb{B}$	contacts : { phone : [*0892701239*], website : [ ]} location : {address : { street : *Champ De Mars, Avenue Anatole*, postalCode : *75007*, ...	formatted_phone_number : *+33892701239* website : *http://www.tour-eiffel.fr* formatted_address : *Champ de Mars,* *5 Avenue Anatole France, 75007 Paris,* *France* ...

TABLE 3.1: Two entities x and y, offered by two different providers, that refer to *Eiffel Tower* POI, p_{ET}, with two different schemas.

The above example will be used to illustrate the following definitions.

Definition 3.4. Corresponding attributes
Two attributes $att_1^i \in \mathbb{S}_1$ and $att_2^j \in \mathbb{S}_2$ are two corresponding attributes, denoted $att_1^i \equiv att_2^j$, iff they refer to the same concept.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

In Table 3.1, the attribute *categories* of the entity x and the attribute *types* of the entity y are two corresponding attributes ($categories \equiv types$). In the literature of schema matching, the correspondences between attributes are represented by a relationship [BBR11] such as equivalence, overlap, disjointness or exclusion. But in the context of LBS providers, we only consider the equivalence relationship because providers' schemas are small and simple as shown in Table 3.1.

Definition 3.5. Association function between entity and POI
We define f as a function that associates a given entity, $e \in \mathbb{E}$, to a POI, $p \in \mathbb{P}$, such that e refers to p.

$$\begin{aligned} f \quad : \quad & \mathbb{E} \rightarrow \mathbb{P} \\ & e \rightarrow f(e) = p \end{aligned}$$

For example, the two entities x and y of Table 3.1 refer to the *Eiffel Tower* POI and $f(x) = f(y) = p_{ET}$.

Definition 3.6. Corresponding entities
Two entities $e_i \in \mathbb{E}_1$ and $e_j \in \mathbb{E}_2$ are corresponding entities, denoted $e_i \equiv e_j$, iff

$$\exists\, p \in \mathbb{P} \setminus f\left(e_i\right) = f\left(e_j\right) = p$$

For example, the two entities x and y of Table 3.1 are corresponding entities ($x \equiv y$) since they refer to the same POI p_{ET}.

The above definitions will be used to formalize the inconsistencies between the entity sets of LBS providers.

3.2 Entity Sets Characterization

This section presents the characteristics of entity sets of LBS providers.

We denote $\mathcal{A}(box)$ as the area of a given geographical *box*.

Definition 3.7. Density
The density of a dataset $\mathbb{E}_1$ represents the fraction of entities present in the area of a given geographical *box* that contains all the entities of $\mathbb{E}_1$.

$$\mathcal{D}(\mathbb{E}_1, box) = \frac{|\mathbb{E}_1|}{\mathcal{A}(box)}$$

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

To create a fair comparison between two entity sets, $\mathbb{E}_1$ and $\mathbb{E}_2$, their densities are measured with respect to a fixed *box*. This latter is defined as the smallest geographical area that contains all entities of $\mathbb{E}_1 \cup \mathbb{E}_2$. Brinkhoff et al. propose several approximations for the smallest area that covers a set of spatial objects as shown in Figure 3.1 [BKS93]. Any of these propositions may be used to find the smallest area of $\mathbb{E}_1 \cup \mathbb{E}_2$ since the same box is used to calculate both densities. In our context, we consider the standard Minimum Bounding Rectangle (MBR) that is equal to the smallest bounding rectangle containing the entities of both datasets. The MBR, also referred to as the *envelope* or Minimum Bounding Box, is identified by two coordinates: the minimum *longitude*- and *latitude*-coordinates $(longitude_{min}, latitude_{min})$, at the lower left of the coordinate space, and the maximum *longitude*- and *latitude*-coordinate $(longitude_{max}, latitude_{max})$, at the upper right.

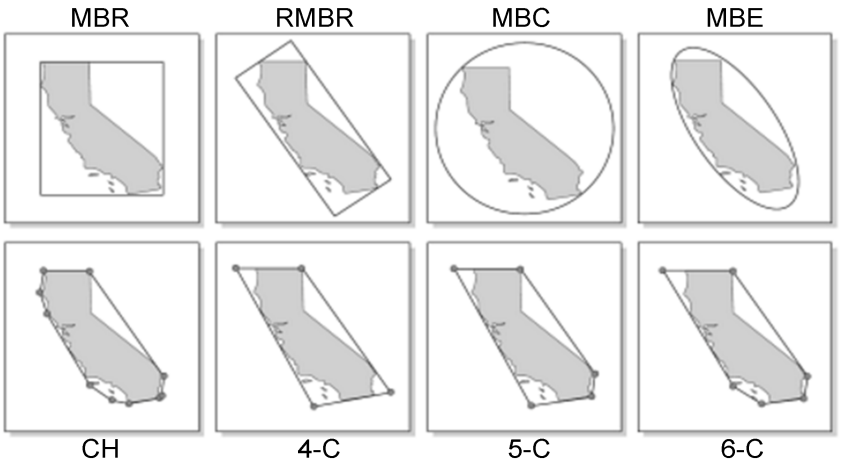


FIGURE 3.1: Different approximations of the area that contains spatial objects [BKS93]. MBR: Minimum Bounding Rectangle; RMBR: Rotated Minimum Bounding Rectangle; MBC: Minimum Bounding Circle; MBE: Minimum Bounding Ellipse; CH: Convex Hull; n-C: Minimum Bounding of n-corners (4-C, 5-C, 6-C).

Definition 3.8. Overlap

The overlap $\theta(\mathbb{E}_1, \mathbb{E}_2)$ is a measure of the fraction of corresponding entities between two datasets $\mathbb{E}_1$ and $\mathbb{E}_2$. The overlap is defined by [SKS+10] as follows:

$$\theta(\mathbb{E}_1, \mathbb{E}_2) = \sqrt{\frac{c}{|\mathbb{E}_1|} \cdot \frac{c}{|\mathbb{E}_2|}}$$

where c is the number of corresponding entities between $\mathbb{E}_1$ and $\mathbb{E}_2$.

We denote $d(e_i, e_j)$ as the distance between two entities $e_i \in \mathbb{E}_1$ and $e_j \in \mathbb{E}_2$. This distance may be measured using different methods including Euclidean distance [SX08a], Manhattan Distance [SX08b] and Haversine distance[1].

[1] Haversine distance: http://mathworld.wolfram.com/GreatCircle.html [Accessed: June 2016]

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

Definition 3.9. Farthest distance
The Farthest distance $\delta(\mathbb{E}_1, \mathbb{E}_2)$ is defined as the maximum distance between two corresponding entities among all the pairs of corresponding entities in two given entity sets $\mathbb{E}_1$ and $\mathbb{E}_2$.

$$\begin{aligned}
&\delta(\mathbb{E}_1, \mathbb{E}_2) = d(e_x, e_y) \text{ where} \\
&e_x \equiv e_y \wedge \\
&\forall\, e_i \in \mathbb{E}_1, \forall\, e_j \in \mathbb{E}_2 \\
&e_i \equiv e_j \wedge d(e_x, e_y) \geq d(e_i, e_j)
\end{aligned}$$

3.3 Taxonomy of Inconsistencies

In this section, the taxonomy is introduced by formalizing the various differences and resemblances that may arise between the entities of two providers. To illustrate this comparison, let us consider $\mathbb{E}_1 = \{e_1, e_2, \ldots\}$ and $\mathbb{E}_2 = \{e_1^{'}, e_2^{'}, \ldots\}$ as entity sets of two LBS providers. Let $\mathbb{S}_1 = \mathbb{I}_1 \cup \mathbb{L}_1 \cup \mathbb{G}_1 \cup \mathbb{A}_1 \cup \mathbb{B}_1$ and $\mathbb{S}_2 = \mathbb{I}_2 \cup \mathbb{L}_2 \cup \mathbb{G}_2 \cup \mathbb{A}_2 \cup \mathbb{B}_2$ be the schemas of $\mathbb{E}_1$ and $\mathbb{E}_2$ respectively. We will analyze the inconsistencies between two corresponding entities $e_a \in \mathbb{E}_1$ and $e_b^{'} \in \mathbb{E}_2$ that refer to the same POI $p_{ab} \in \mathbb{P}$ (i.e., $e_a \equiv e_b^{'}$), and between two non-corresponding entities $e_c \in \mathbb{E}_1$ and $e_d^{'} \in \mathbb{E}_2$ that refer to two distinct POIs, $p_c \in \mathbb{P}$ and $p_d \in \mathbb{P}$, respectively (i.e., $e_c \not\equiv e_d^{'}$).

The entities of several sets will be compared according to four levels: 1) schema, 2) terminology, 3) spatial and 4) entities' availability.

3.3.1 Schema Differences

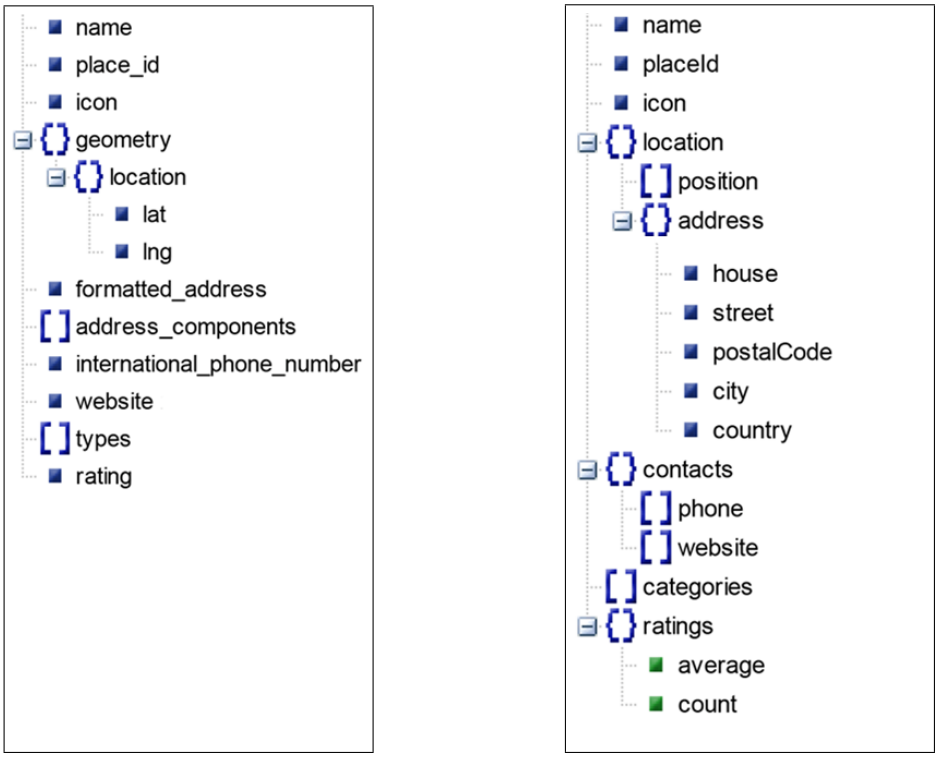
This level shows the differences between distinct schemas where two inconsistencies are distinguished. Generally, inconsistencies between schemas involve two providers, i.e., Inter-Inconsistency. In the case of a provider with a dynamic schema, inconsistencies may be classified as Intra-Inconsistency.

Attribute Heterogeneity (AH)
The *Attribute Heterogeneity* consists of two corresponding attributes belonging to two distinct schemas and have different labels or different abstract data types.
The two schemas, $\mathbb{S}_1$ and $\mathbb{S}_2$, have *Attribute Heterogeneity*, denoted as $\mathbb{S}_1$ AH $\mathbb{S}_2$, iff

$$\begin{aligned}
&\exists\, att_i \in \mathbb{S}_1, \exists\, att_j' \in \mathbb{S}_2 \backslash \\
&\big(att_i \equiv att_j'\big) \wedge \big(att_i.label \neq att_j'.label \vee att_i.type \neq att_j'.type\big)
\end{aligned}$$

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

Figure 3.2 represents two schemas of two LBS providers. The attribute *types* in Figure 3.2a and the attribute *categories* in Figure 3.2b are two corresponding attributes that have *Attribute Heterogeneity* inconsistency, where both refers to the type of the POI but with different labels.



(a) Schema of provider 1. (b) Schema of provider 2.

FIGURE 3.2: Schemas of data offered by two LBS providers.

Complex Correspondences (CC)

Schemas may have various structures. One attribute of one schema may correspond to two or more attributes of another schema. Therefore, a concept is described by one attribute of the schema $\mathbb{S}_1$ and by two or more attributes of the schema $\mathbb{S}_2$, or vice versa.

The two schemas, $\mathbb{S}_1$ and $\mathbb{S}_2$, have *Complex Correspondences*, denoted as $\mathbb{S}_1$ CC $\mathbb{S}_2$, iff

$$att_i \equiv (att'_1, att'_2, \ldots) \vee (att_1, att_2, \ldots) \equiv att'_j$$

Returning to Figure 3.2, the attribute *formatted_address* in Figure 3.2a represents the full address while this latter is represented by several attributes in Figure 3.2b (*house, street, postalCode, ...*). Note that there are more complex correspondences between the structures of the schemas [RB01]. For instance, more than one attribute of a schema may correspond to more than one attribute of another (i.e., [n:m] correspondences). These cases are not considered due to the simplicity of schemas in LBS context.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

3.3.2 Terminological Inconsistencies

This level focuses on the inconsistencies of values for both primary and secondary terminological attributes of two corresponding entities.

Similar Data (SD)

The *Similar Data* resemblance consists of two non-corresponding entities that have similar values for terminological attributes (denoted $att_i.val \cong att'_j.val$). This resemblance is classified as Intra-Inconsistency and Inter-Inconsistency.

The two distinct entities, e_c and e'_d, have *Similar Data* resemblance, denoted as e_c SD e'_d, iff

$$
\begin{aligned}
& e_c \not\equiv e'_d \ \wedge \\
& \exists\, att_i \in \mathbb{A}_1 \cup \mathbb{B}_1, \exists\, att'_j \in \mathbb{A}_2 \cup \mathbb{B}_2 \setminus \\
& \left(e_c.att_i \equiv e'_d.att'_j\right) \wedge \left(e_c.att_i.val \cong e'_d.att'_j.val\right)
\end{aligned}
$$

For example, *Similar Data* resemblance may appear for chains stores or agencies of the same companies. Table 3.2 shows an example[2] of SD, where two entities offered by Google Maps, referring to two distinct POIs located in different areas that have the same type *post office*, similar place name *La poste* and not corresponding to each other.

Entity 1	**Entity 2**
geometry : { lat : 45.758730, lng : 4.853554}	geometry : { lat : 45.758018, lng : 4.862524}
name : La poste	name : La poste
types : post office	types : post office
address : 6 Rue du Lac, 69003 Lyon	address : 72 Rue Maurice Flandin, 69003 Lyon

TABLE 3.2: Example of the *Similar Data* resemblance.
Two entities offered by one LBS provider, that have the same value for the terminological *name* attribute, and located in two different locations.

Different Data (DD)

This case consists of two corresponding entities that have different values for their corresponding terminological attributes (primary or secondary). It is classified as Inter-Inconsistency.

The two corresponding entities, e_a and e'_b, have *Different Data* difference, denoted as e_a DD e'_b, iff

$$
\begin{aligned}
& e_a \equiv e'_b \ \wedge \\
& \exists\, att_i \in \mathbb{A}_1 \cup \mathbb{B}_1, \exists\, att'_j \in \mathbb{A}_2 \cup \mathbb{B}_2 \setminus \\
& \left(e_a.att_i \equiv e'_b.att'_j\right) \wedge \left(e_a.att_i.val \neq e'_b.att'_j.val\right)
\end{aligned}
$$

[2] Example of SD: two entities offered by Google Maps having the same name *La poste*. [Accessed: June 2016]

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

Note that the degree of difference between the values varies. This variation distinguishes two classifications of *Different Data*:

1. **Semantic Different Data (SEMDD)** denoted as e_a SEMDD e'_b, it consists of two corresponding attributes, where their values are built on the same concept and meaning but with different words, such as synonyms.

2. **Syntactic Different Data (SYNDD)** denoted as e_a SYNDD e'_b, it is about the syntax of corresponding attributes' values. There are many different ways that a value can be expressed in real life, without any alteration of its meaning, or a result of human errors (i.e., misspellings, word permutations, aliases, different standards, acronyms, abbreviations and multilingualism).

Figure 3.3 shows two corresponding entities from two LBS providers that have SEMDD and SYNDD differences. The green marker (top left) represents the entity[3] offered by Google Maps and the blue marker (bottom right) represents the entity[4] offered by Nokia Here Maps. These two entities refer to the same POI *IUT Lyon 1 - Gratte-Ciel University, France*. The names given by these entities are syntactically different, "IUT Lyon 1 Site de Gratte-Ciel" vs. "Univ. Lyon 1-I.U.T. (B)". In addition, the types given by these entities are semantically different, "university" vs. "education".

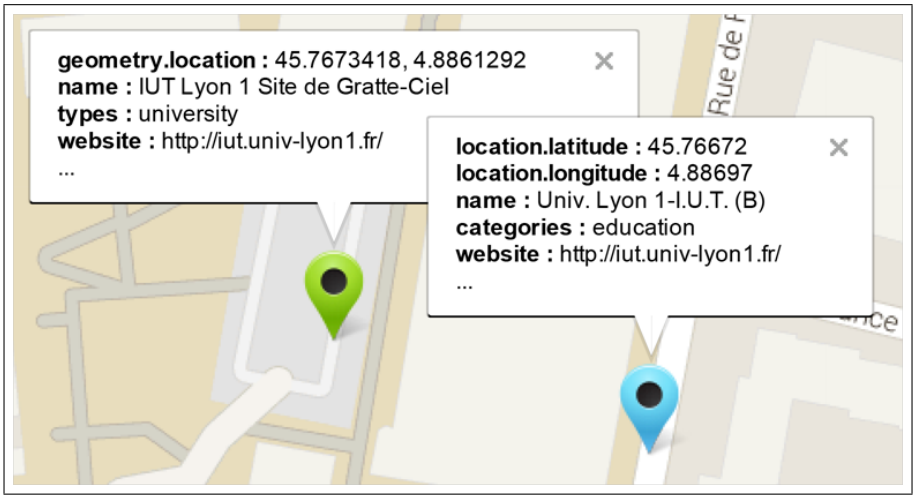


FIGURE 3.3: Example of the *Different Data* difference.
Two corresponding entities referring to the *IUT Lyon 1 - Gratte-Ciel University* having SEMDD and SYNDD differences.

Missing Data (MD)

The *Missing Data* difference appears when information is only given by one of two corresponding entities. This difference is classified as Inter-Inconsistency.

[3] Example of DD: *IUT Lyon 1 - Gratte-Ciel University* offered by Google Maps. [Accessed: June 2016]
[4] Example of DD: *IUT Lyon 1 - Gratte-Ciel University* offered by Nokia Here. [Accessed: June 2016]

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

The two corresponding entities, e_a and e'_b, have *Missing Data* difference, denoted as e_a MD e'_b, iff

$$
\begin{aligned}
& e_a \equiv e'_b \ \wedge \\
& \Big(\exists\, att_i \in \mathbb{B}_1, \exists\, att'_j \in \mathbb{B}_2 \backslash \\
& \Big(att_i \equiv att'_j \ \wedge \big(e_a.att_i.val = NULL \vee e'_b.att'_j.val = NULL \big) \Big) \Big) \vee \\
& \Big(\exists\, att_i \in \mathbb{B}_1, \forall\, att'_j \in \mathbb{B}_2 \backslash \big(att_i \not\equiv att'_j \big) \Big)
\end{aligned}
$$

This difference is due to two cases. The first consists of two corresponding entities that have two corresponding attributes, where one of them has a null value. The second consists of two corresponding entities, where one of them has an attribute that does not have any correspondence with the attributes of the other entity. Figure 3.4 shows two corresponding entities from two LBS providers. The green marker (bottom left) represents the entity[5] offered by Google Maps and the blue marker (top right) represents the entity[6] offered by Nokia Here Maps. These two entities refer to the same POI (*Colorado* restaurant, France). As shown, the website is given by one entity, while it has a NULL value by the other.



FIGURE 3.4: Example of the *Missing Data* difference.
Two corresponding entities refer to the *Colorado* restaurant having a MD.

[5] Example of MD: *Colorado restaurant* offered by Google Maps. [Accessed: June 2016]
[6] Example of MD: *Colorado restaurant* offered by Nokia Here Maps. [Accessed: June 2016]

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

3.3.3 Spatial Inconsistencies

At this level, we investigate the problem of positioning between the corresponding entities. One resemblance and two differences can be distinguished.

Superposition (SUP)

This resemblance consists of two entities having the same location but refer to two distinct POIs, generally it is classified as Intra-Inconsistency and Inter-Inconsistency. The two distinct entities, e_c and $e_d^{'}$, have *Superposition* resemblance, denoted as e_c SUP $e_d^{'}$, iff

$$e_c \not\equiv e_d^{'} \wedge$$
$$\Big(e_c.\text{LATITUDE}.val \cong e_d^{'}.\text{LATITUDE}.val\Big) \wedge \Big(e_c.\text{LONGITUDE}.val \cong e_d^{'}.\text{LONGITUDE}.val\Big)$$

For example, this case may appear in shopping centers where two POIs of the same type are located one above the other on two different floors. For instance, Table 3.3 shows two distinct entities, the first refers to the *Chez Leon*[7] restaurant POI and the second refers to the *Maison Rousseau*[8] restaurant POI. These entities are offered by the Nokia Here Maps and having exactly the same location coordinates.

Entity 1	**Entity 2**
geometry : { lat : 45.763527 , lng : 4.85023}	geometry : { lat : 45.763527 , lng : 4.85023}
name : Chez Leon	name : Maison Rousseau
phone_number : +33478623028	phone_number : +33478623765
address : 102 cours lafayette, Lyon, France	address : 102 cours lafayette, Lyon, France

TABLE 3.3: Example of the *Superposition* resemblance.
Two entities, offered by one LBS provider, that have the same location coordinates.

Different Locations (DL)

This difference consists of two corresponding entities that have different values for their corresponding spatial attributes. This difference is classified as Inter-Inconsistency.
The two corresponding entities, e_a and e_b', have *Different Locations*, denoted as e_a DL e_b', iff

$$e_a \equiv e_b' \wedge$$
$$\Big(e_a.\text{LATITUDE}.val \neq e_b'.\text{LATITUDE}.val \vee e_a.\text{LONGITUDE}.val \neq e_b'.\text{LONGITUDE}.val\Big)$$

Figure 3.5 shows two corresponding entities from two LBS providers. These two entities have a DL difference since they have different longitude and latitude values and the distance between them is approximately equal to 15 meters.

[7] Example of SUP: *Chez Leon*POI offered by Nokia Here Maps. [Accessed: June 2016]
[8] Example of SUP: *Maison Rousseau* POI offered by Nokia Here Maps. [Accessed: June 2016]

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés



FIGURE 3.5: Example of the *Different Locations* difference.
Two corresponding entities separated by 15 meters having DL difference.

Equipollent Positions (EP)

This difference appears when the corresponding entities have different locations, but these locations are correct with respect to the location of the POI. This shows that the corresponding entities' positions are equivalent in terms of concept but not in terms of values. This difference is classified as Inter-Inconsistency.

The two corresponding entities, e_a and e'_b that refer to p_{ab}, have *Equipollent Positions* difference, denoted as e_a EP e'_b, iff

$$
\begin{aligned}
& e_a \equiv e'_b \ \wedge \\
& (e_a.\text{LONGITUDE}, e_a.\text{LATITUDE}) \subset p_{ab}.coordinates \ \wedge \\
& (e'_b.\text{LONGITUDE}, e'_b.\text{LATITUDE}) \subset p_{ab}.coordinates \ \wedge \\
& \Big(e_a.\text{LONGITUDE}.val \neq e'_b.\text{LONGITUDE}.val \ \vee \ e_a.\text{LATITUDE}.val \neq e'_b.\text{LATITUDE}.val\Big)
\end{aligned}
$$

Figure 3.6 shows two corresponding entities that refer to *IUT Lyon 1 - Gratte-Ciel University*. The green marker (top left) represents the entity[9] offered by Google Maps and the blue marker (bottom right) represents the entity[10] offered by Nokia Here Maps. These two entities have different locations (center of gravity vs. entrance gate) but both are correct.

[9] Example of EP: *IUT Lyon 1 - Gratte-Ciel University* offered by Google Maps. [Accessed: June 2016]
[10] Example of EP: *IUT Lyon 1 - Gratte-Ciel University* offered by Nokia Here. [Accessed: June 2016]

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

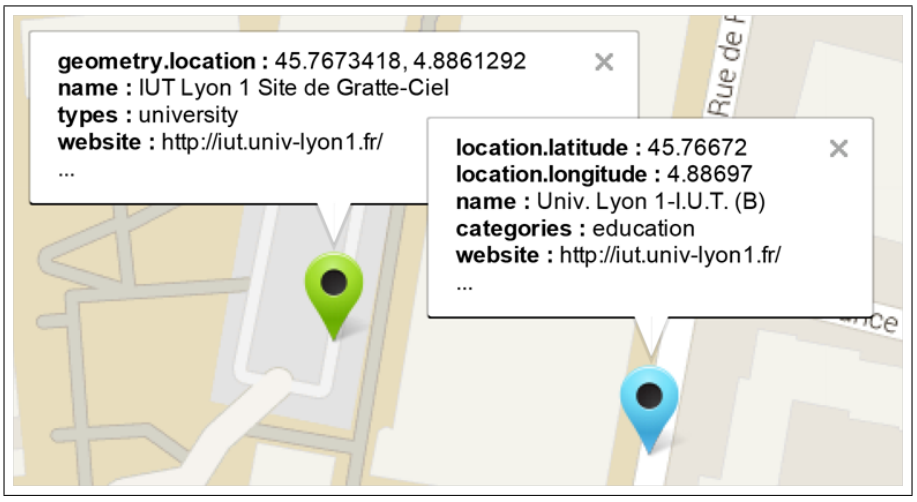


FIGURE 3.6: Example of the *Equipollent Positions* difference.
Two corresponding entities refer to *IUT Lyon 1 - Gratte-Ciel University* having different locations but both are correct.

3.3.4 Entity's Availability

The entity's availability category takes into account the inconsistencies that can be found in the entity set of a provider. Two cases can be distinguished at this level.

Not Found Entity (NFE)

This case is classified as Inter-Inconsistency, which consists of a POI that is given by one provider, but not by the other.

The entity $e_c \in \mathbb{E}_1$ is a *Not found Entity* in $\mathbb{E}_2$, denoted as e_c NFE $\mathbb{E}_2$, iff

$$\forall e_j^{'} \in \mathbb{E}_2, f(e_c) = p_c \wedge f(e_j^{'}) \neq p_c$$

For example, suppose that a new restaurant opens and there is only one LBS provider who added this restaurant to its entity set. This new POI remains as NFE until it is included in the entity sets of other LBS providers.

Duplicate Entities (DE)

This case is classified as Intra-Inconsistency, which corresponds to two entities of the same provider referring to the same POI.

Two entities, $e_i \in \mathbb{E}_1$ and $e_j \in \mathbb{E}_1$, are *Duplicate Entities*, denoted as e_i DE e_j, iff

$$\exists p_{ij} \in \mathbb{P} \backslash (f(e_i) = f(e_j) = p_{ij})$$

Figure 3.7 shows two duplicate entities, offered by Google Maps, that refer to the same POI *Eat Sushi Lyon 2*[11] located at 27 Quai Jean Moulin, 69002 Lyon. Note that it is

[11] Example of DE: two duplicate entities offered by Google. [Accessed: June 2016]

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés



FIGURE 3.7: Example of the *Duplicate Entities*.
Two entities offered by one provider and referring to the same POI *Eat Sushi Lyon 2*.

not necessary that two duplicated entities have the same values for all attributes.

Table 3.4 summarizes the taxonomy's inconsistencies that are grouped into four categories including schema, terminological, spatial and availability.

Category	**Inconsistency Type**	**Inconsistency**	**Notation**	**Intra-Incon.**	**Inter-Incon.**
Schema		Attribute Heterogeneity	AH	X(dynamic schema)	X
		Complex Correspondences	CC	X(dynamic schema)	X
Terminology	Differences	Semantic Different Data	SEMDD		X
		Syntactic Different Data	SYNDD		X
		Missing Data	MD		X
	Resemblances	Similar Data	SD	X	X
Spatial	Differences	Different Locations	DL		X
		Equipollent Positions	EP		X
	Resemblances	Superposition	SUP	X	X
Availability		Not Found Entity	NFE		X
		Duplicate Entities	DE	X	

TABLE 3.4: Taxonomy's inconsistencies.

The above inconsistencies may influence the results of the geospatial entity matching approaches. The next section gives more details about how they impact the performance of spatial entity matching approaches.

3.4 Impact of Taxonomy

This section analyzes how the taxonomy impacts the quality of integrating multiple LBS. The impact of taxonomy varies depending on the inconsistencies. As mentioned in Section 2.3.2, the performance of a matching approach is measured in terms of *Precision* and *Recall*. These metrics depend on three measures including True Positive (TP), False

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

Negative (FN) and False Positive (FP). The inconsistencies may impact the effectiveness of matching approaches, which is related *Precision* and *Recall*. On the other hand, they may impact the efficiency of matching approaches, which is related to performance, such as execution time and amount of memory allocated by the approach. In addition, some inconsistencies may impact the usability of LBS.

3.4.1 Impact of Schema's Inconsistencies

Concerning the schema level, two inconsistencies are distinguished namely *Attribute Heterogeneity* and *Complex Correspondences*. Usually, an entity matching approach uses the corresponding attributes of datasets' schemas in order to compare their values and detect the corresponding entities. Hence, any error made in the schema matching, due to *Attribute Heterogeneity* or *Complex Correspondences*, must necessarily impact the effectiveness of entity matching approaches. To understand the impact of these inconsistencies, let us consider $\mathbb{E}_1 = \{e_1, e_2, \ldots\}$ and $\mathbb{E}_2 = \{e_1^{'}, e_2^{'}, \ldots\}$ as entity sets of two LBS providers. Let $\mathbb{S}_1 = \{\text{poi_N, poi_types}, \ldots\}$ and $\mathbb{S}_2 = \{\text{poi_name, poi_T}, \ldots\}$ be the schemas of $\mathbb{E}_1$ and $\mathbb{E}_2$, respectively. The attributes poi_N and poi_name are two corresponding attributes (poi_N $\equiv$ poi_name) that have *Attribute Heterogeneity* and refer to the POI name. Similarly, the attributes poi_types and poi_T are two corresponding attributes (poi_types $\equiv$ poi_T) that have *Attribute Heterogeneity* and refer to the POI type. In addition, suppose that poi_N, poi_name and poi_T have a string abstract type and poi_types has an integer abstract data type. A schema matching approach may make a mistake and decide that poi_N corresponds to poi_T because they have similar labels and same abstract data type. In this case, the effectiveness will be impacted. For instance, consider two corresponding entities $e_1 = \{\text{poi_N} : \text{Berger, poi_types} : 5882\}$ and $e_1^{'} = \{\text{poi_name} : \text{Le Berger, poi_T} : \text{hotel}\}$. Comparing e_1.poi_N.val with $e_1^{'}$.poi_T.val (Berger vs. hotel) will give a very low similarity for e_1 and $e_1^{'}$. Hence, these two corresponding entities will not be detected, one False Negative (FN) is produced and one True Positive (TP) is missed, which decreases the *Recall*. Now, consider the entity $e_2 = \{\text{poi_N} : \text{Hotel Ly, poi_types} : 5882\}$ that does not correspond to $e_1^{'}$. Comparing e_2.poi_N.val with $e_1^{'}$.poi_T.val (Hotel Ly vs. hotel) will give a high similarity to e_2 and $e_1^{'}$. Hence, these two entities will be detected as corresponding and one False Positive (FP) is produced, which decreases the *Precision*. Similar impact may arise due to *Complex Correspondences* inconsistency. By contrast, if the matching between the schemas is not provided during the entity matching process, there is an additional rise in impact. In this case, entity matching approaches will be forced to compare the values of all attributes of both schemas, i.e., Cartesian product comparison, which reduces the performance by increasing the execution time and amount of memory.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

The datasets' schemas of LBS providers are small and simple. In this work, we consider that the schema matching can be easily performed manually by an expert to guarantee its correctness and avoid any problems at the entity matching process. In other words, manual schema matching allows overcoming the inconsistencies of schema level. On this basic, schema's inconsistencies are not considered in the following of this thesis.

3.4.2 Impact of Spatial and Terminological Inconsistencies

In this section, we analyze how the spatial and terminological inconsistencies impact the effectiveness of entity matching approaches. The differences may prevent a matching approach from detecting two correct corresponding entities, which impacts the *Recall*. The resemblances may force a matching approach to detect two non-corresponding entities, which impacts the *Precision*.

Concerning the differences, two corresponding entities may have one of the following differences: SEMDD, SYNDD, MD, DL or EP. If a matching approach fails to detect the correspondence due to one of these differences, then one FN is produced and one TP is missed, which decreases the *Recall*. For instance, consider two corresponding entities that have a SYNDD for their names, such as "Kentucky Fried Chicken" vs. "KFC" or "Les 3 Collèges" vs. "Les Trois Collèges". If an entity matching approach compares their names and fails to detect that they are corresponding, then this correspondence would be missing from the matching results. This means that one TP is missed, which decreases the *Recall*. Similarly, Figure 3.8 shows two corresponding entities that refer to *Wallace State Park* POI and have EP inconsistency. These two entities are separated by 550 meters, if a matching approach cannot detect them as corresponding, then the *Recall* decreases.

Concerning the resemblances, consider two non-corresponding entities that have one of the following resemblances: SD or SUP. If an entity matching approach detects them as corresponding due to one of these resemblances, then one FP is produced, which decreases the *Precision*. For example, consider two entities that have a SD resemblance (see Table 3.2 in Section 3.3.2). If an entity matching approach compares their names and finds them as corresponding, then one incorrect correspondence is produced (i.e., 1 FP) and the *Precision* decreases.

3.4.3 Impact of Availability's Inconsistencies

The availability level contains two inconsistencies namely *Not Found Entity* and *Duplicate Entities*.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

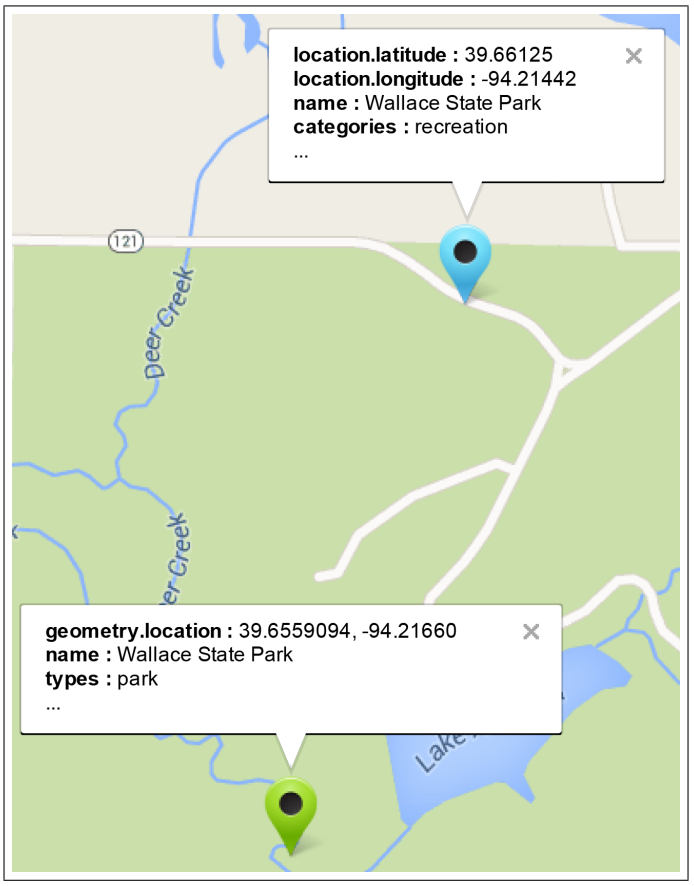


FIGURE 3.8: Two corresponding entities that have EP inconsistency and refer to *Wallace State Park* POI .

The *Not Found Entity* has similar influence to the resemblances, i.e., it impacts the *Precision*. Consider an entity with NFE, the *Precision* decreases if the matching approach decides that this NFE entity has a correspondence.

Concerning the *Duplicate Entities*, a low quality dataset may contain duplicate entities that refer to the same POI. On one hand, the DE inconsistency impacts the usability of LBS. For instance, users would be confused when they requests a POI and the results show two entities located one next to other, with the same names and maybe some contradictory terminological information, such as different phone numbers. On the other hand, matching such kind of datasets impacts the effectiveness by increasing the difficulty of evaluating the quality of results. For example, consider two entities in the same dataset, $e_j \in \mathbb{E}_2$ and $e_k \in \mathbb{E}_2$, that have DE inconsistency, and their corresponding entity $e_i \in \mathbb{E}_1$. If a matching approach produces 1:1 correspondences, then only one correspondence can be assigned to a given entity. Now, this matching approach may decide that either none of the duplicate entities correspond to e_i, which produces two FN, or only one of them corresponds to e_i, which produces one FN. However, both cases decrease the *Recall*. Otherwise, if matching approaches, that produce 1:n or n:n

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

correspondences, detect both duplicate correspondences, then the DE causes no impact on the effectiveness. Otherwise, one or two FN will be produced, which decreases the *Recall*. Furthermore, DE impacts the efficiency of matching approaches because the duplicate entities increase the number of comparisons between the entities of two entity sets, which reduces the performance by increasing the execution time and amount of memory.

Table 3.5 summarizes the impact of taxonomy's inconsistencies.

Category	**Inconsistency Type**	**Inconsistency**	**Notation**	**Impact**
Schema		Attribute Heterogeneity	AH	Effectiveness, Efficiency
		Complex Correspondences	CC	Effectiveness, Efficiency
Terminology	Differences	Semantic Different Data	SEMDD	*Recall*
		Syntactic Different Data	SYNDD	*Recall*
		Missing Data	MD	*Recall*
	Resemblances	Similar Data	SD	*Precision*
Spatial	Differences	Different Locations	DL	*Recall*
		Equipollent Positions	EP	*Recall*
	Resemblances	Superposition	SUP	*Precision*
Availability		Not Found Entity	NFE	*Precision*
		Duplicate Entities	DE	*Recall*, Effectiveness, Efficiency

TABLE 3.5: Inconsistencies of the taxonomy and their impact.

3.5 Combination of Inconsistencies

When working with LBS providers' data (real-world data), it is not rare to find entities with several inconsistencies. To better characterize the datasets, we studied the different possibilities of combining the inconsistencies defined above. In this section, we investigate the combinations of inconsistencies that may appear between entities. Two kinds of inconsistencies' combinations are distinguished (i) simple combinations that appear between two entities exclusively and (ii) complex combinations that appear between more than two entities.

3.5.1 Simple Combinations

Two entities being compared may have several inconsistencies, such as EP for spatial attributes and SYNDD for terminological attributes. Some of these inconsistencies may be combined with any other inconsistency. Some others are contradictory between one another and cannot be combined. For instance, the differences (i.e., the inconsistencies between two corresponding entities) cannot be combined with the inconsistencies that appear between non-corresponding entities, such as resemblances.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

To simplify the analysis, the two inconsistencies SEMDD and SYNDD are grouped together. We denote DD as the combination of SEMDD and SYNDD. Recall that schema's inconsistencies are not considered anymore.

Concerning the differences (i.e., when the compared entities are corresponding), in terminological level, there are several attributes (e.g., name, type or phone number), that may have any of the two terminological differences, namely DD and MD. This means that DD and MD can appear together. In contrast, there is only one attribute (location) in spatial level, that may have one of the two spatial differences, namely DL and EP. This means that DL and EP cannot appear together. Finally, the differences of terminological and spatial levels can be combined together since they are independent from each other. On this basis, 11 combinations are obtained by combining the following five inconsistencies: DL, DD, MD, and EP. Figure 3.9 shows all possible combinations for differences.

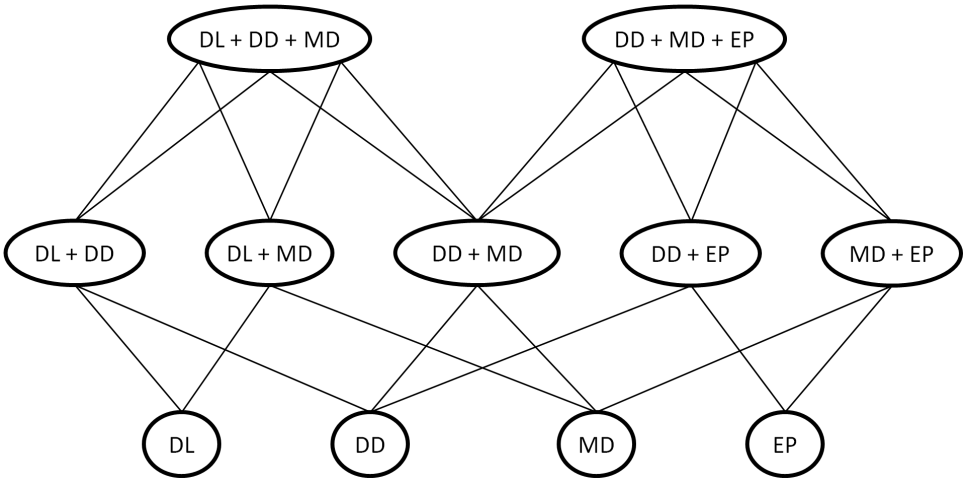


FIGURE 3.9: All possible combinations for differences.

Concerning the resemblances (i.e., when compared entities are not corresponding), two inconsistencies are distinguished, SUP for spatial level and SD for terminological level. These two resemblances may appear together since they are independent. For example, two non-corresponding entities may have similar data and location. Concerning the availability level, two inconsistencies are also distinguished, *Not Found Entity* and *Duplicate Entities*. The former refers to one entity from one provider that does not have any correspondence in the entity set of a second provider. This means that NFE arises between one entity and one entity set. Hence, this inconsistency cannot be combined since we focus only on the inconsistencies that appear between two entities exclusively. Concerning the latter, two duplicate entities may have similar data or location, which means that DE can be combined with SD and SUP. On this basis, 7 additional combinations are obtained for the following fourth inconsistencies: SD, SUP and DE (see Figure 3.10).

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

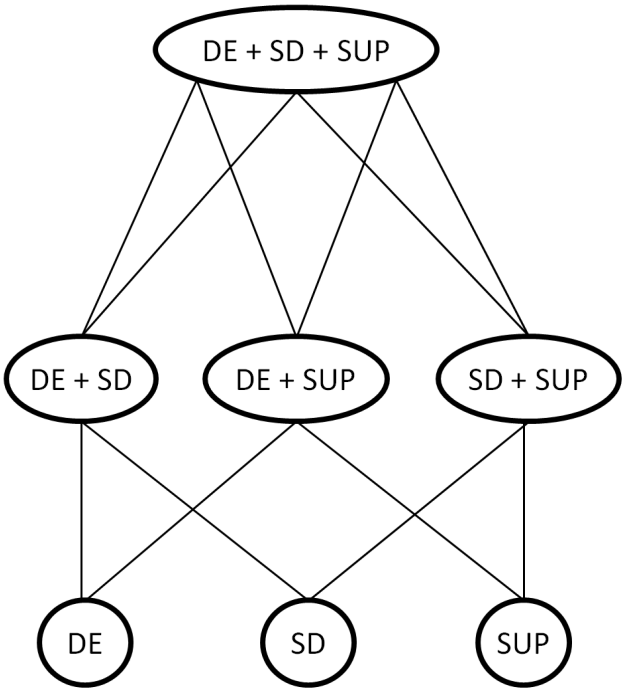


FIGURE 3.10: All possible combinations for resemblances.

The overall sum of all possible combinations is equal to 18. These combinations do not cause any additional impact to the results' quality. For example, if a matching approach detects two non corresponding entities that have SUP resemblance, then one FP is produced. Similarly, if they have a SD resemblance and a matching approach detects them as corresponding, then one FP is produced. Now, if they have a combination of SUP and SD and a matching approach detects them as corresponding, the impact remains as one FP.

3.5.2 Complex Combinations

The previous section studied the inconsistencies between two entities. However, during the matching of two entity sets, complex combinations arise when an entity from one dataset has inconsistencies with two or more entities from the other set. Some of these combinations may change the impact of inconsistencies. An entity may have resemblances with more than one non-corresponding entities. In contrast, an entity may have differences with only its corresponding entity, since each entity has only one correspondence; except the case where one entity corresponds to two duplicate entities and consequently, it may have differences with two entities. We do not consider the duplicate case in this section because it has already been analyzed in Section 3.4.3. On this basis, two remaining cases will be analyzed. To illustrate these cases, let $\mathbb{E}_1 = \{e_1, ..., e_n\}$ and $\mathbb{E}_2 = \{e_1^{'}, ..., e_m^{'}\}$ be two entity sets of two LBS providers.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

1. The first case concerns an entity from one dataset that has resemblances with two or more entities from the other set. Table 3.6 and Figure 3.11 show an example of the entity $e_1^{'} \in \mathbb{E}_2$ that has *Superposition* with $e_3 \in \mathbb{E}_1$ and *Similar Data* with $e_1 \in \mathbb{E}_1$ and $e_2 \in \mathbb{E}_1$, i.e., $e_1^{'}$ SD e_1, $e_1^{'}$ SD e_2 and $e_1^{'}$ SUP e_3. Usually, if two non-corresponding entities share a resemblance and are detected as corresponding, then one FP is produced. If $\mathbb{E}_1$ and $\mathbb{E}_2$ are matched using an approach that produces 1:1 correspondence, then the worst case scenario would be if $e_1^{'}$ is detected with one of the following three entities: e_1, e_2 and e_3, which produces only one FP. On the other hand, suppose that $\mathbb{E}_1$ and $\mathbb{E}_2$ are matched with an approach that produces n:n or 1:n correspondences. Then, at worst, $e_1^{'}$ may be detected with e_1, e_2 and e_3, which produces three FPs and the impact remains the same as if each pair is analyzed separately.

$e_1^{'}$	e_1	e_2	e_3
position : [45.75803, 4.86256]	lat : 45.76701, lng : 4.86333	lat : 45.76792, lng : 4.86430	lat : 45.75803, lng : 4.86256
place_name : Hotel Leyla	name : Hotel Leya	name : Hôtel Leyal	name : Movenpick

TABLE 3.6: Example of complex combination of resemblances. $e_1^{'}$ SD e_1, $e_1^{'}$ SD e_2 and $e_1^{'}$ SUP e_3.

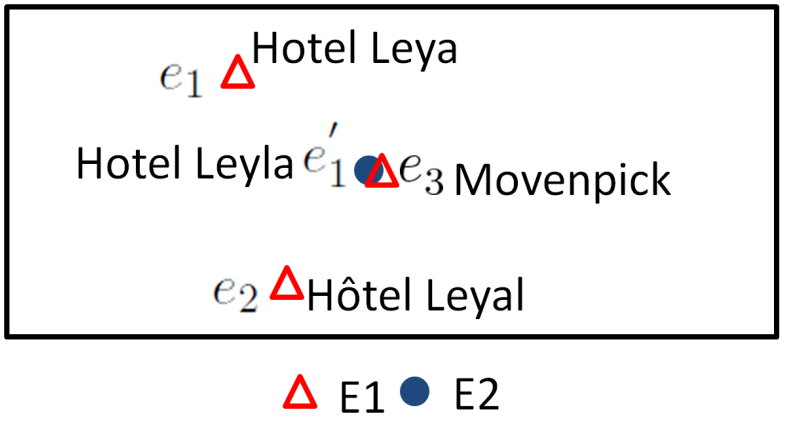


FIGURE 3.11: Example of complex combination of resemblances. $e_1^{'}$ SD e_1, $e_1^{'}$ SD e_2 and $e_1^{'}$ SUP e_3.

2. The second case concerns an entity from one dataset that has a difference with its corresponding entity and one or more resemblances with entities from the other set. Table 3.7 and Figure 3.12 show an example of the entity $e_1^{'} \in \mathbb{E}_2$ that has *Different Locations* with $e_1 \in \mathbb{E}_1$ and *Similar Data* with $e_2 \in \mathbb{E}_1$ and $e_3 \in \mathbb{E}_1$, i.e., $e_1^{'}$ DL e_1, $e_1^{'}$ SD e_2 and $e_1^{'}$ SD e_3. Usually, if two entities have a resemblance are detected as corresponding, then one FP is produced. Otherwise, if two entities having a difference are not detected, then one FN is produced. If $\mathbb{E}_1$ and $\mathbb{E}_2$ are matched using an approach that produces 1:1 correspondence, then the worst case arises if $e_1^{'}$ is detected with e_2 or e_3, which produces one FP, and one FN since $e_1^{'}$ has not

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

been assigned with its corresponding entity. On the other hand, suppose that $\mathbb{E}_1$ and $\mathbb{E}_2$ are matched with an approach that produces n:n or 1:n correspondences. Then, at worst, $e_1^{'}$ may not be detected as corresponding to e_1, which produces one FN. Instead, it may be detected with e_2 and e_3, which produces two FPs and the impact remains the same as if each pair is analyzed separately.

$e_1^{'}$	e_1	e_2	e_3
position : [45.76701, 4.86333]	lat : 45.76709, lng : 4.86344	lat : 45.75803, lng : 4.85256	lat : 45.73303, lng : 4.9547
place_name : Hotel Leya	name : Hotel Leya	name : Hôtel Leyal	name : Hotel Leyla

TABLE 3.7: Example of the complex combination of differences and resemblances. $e_1^{'}$ DL e_1, $e_1^{'}$ SD e_2 and $e_1^{'}$ SD e_3.

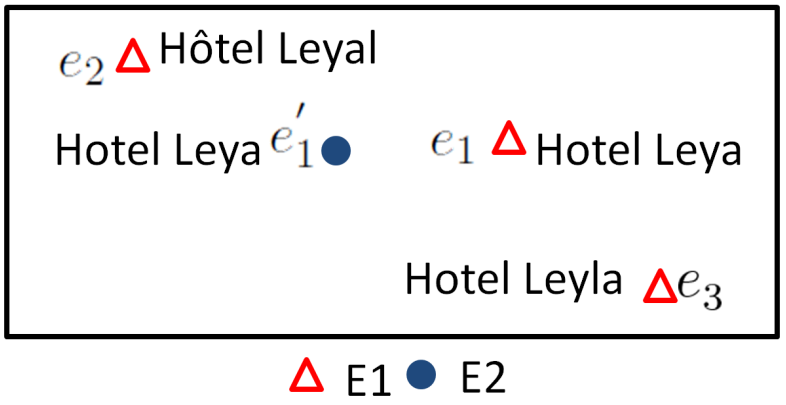


FIGURE 3.12: Example of the complex combination of differences and resemblances. $e_1^{'}$ DL e_1, $e_1^{'}$ SD e_2 and $e_1^{'}$ SD e_3.

According to these cases, entity matching approaches that produces 1:1 correspondence may reduce the impact of resemblances that appear in a complex combination. As shown, when an entity has resemblances with n entities from the other set, the impact of $(n-1)$ resemblances may be avoided thanks to 1:1 correspondence.

3.6 Conclusion

This chapter describes the preliminary definitions that constitute the necessary elements to present a model of multiple LBS providers. More specifically, these definitions describe the composition of entity sets of LBS providers and highlight their characterizations. However, as mentioned in Section 2.1, most LBS providers share their data using REST API, with limitations concerning the number of queries and the number of entities returned per query. This means that in reality, users are unable to collect the whole entity set of a given provider. Instead, a dataset that refers to a subset of the whole entity set can be requested.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

The definitions of the LBS model allow the construction of a taxonomy that formalizes all kinds of inconsistencies, such as differences and resemblances. These inconsistencies influence the integration of multiple LBS by impacting the effectiveness, the efficiency and the usability. The *Duplicate Entities* may be resolved using de-duplication techniques to verify the quality of a dataset before starting the entity matching process [MAK10]. However, the remaining inconsistencies should be addressed during the entity matching process. The more the matching approach resolves inconsistencies, the better the quality of results are. Although the inconsistencies described in this taxonomy are elementary, two compared entities may have a combination of distinct inconsistencies. This makes the detection of the corresponding entities a hard task. To ensure the effectiveness of matching, several approaches have been proposed [SGV06, OR07, SKSD06]. These approaches measure the similarities of several attributes in order to obtain a general overview about the similarity of compared entities. Moreover, when an entity shares inconsistencies with two or more entities, the impact of the effectiveness decreases when using entity matching approaches that produce 1:1 correspondence. In the next chapter, we use the taxonomy to build a benchmark that helps evaluate and compare the entity matching approaches in a spatial context.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

Chapter 4

An Evaluation Benchmark with Real World Data

Contents

4.1 GeoBench DB - Construction of Characterized Database . 71
4.1.1 Overview of GeoBench Aligner 71
4.1.2 Schema Matching 73
4.1.3 POIs' Types Matching 74
4.1.4 Blocking Algorithm 76
4.1.5 Basic Matching Algorithm 76
4.1.6 GeoBench Aligner Prototype - User Interface 78
4.1.7 Experimentation and Population of GeoBench DB 79
4.2 PABench - An Evaluation Benchmark 81
4.2.1 Model of Situations 81
4.2.2 Test Cases and Metrics 83
4.2.3 PABench Extractor 84
4.3 Conclusion 85

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

In the previous chapter, we built a taxonomy that formalizes all kind of differences between corresponding entities (see Section 3.3 in Chapter 3). In this chapter, we discuss the evaluation issue of geospatial entity matching approaches. The evaluation of a given approach consists of matching characterized data, which allows us to analyze strengths and weaknesses of this approach. On this base, we intend to create a database characterized according to the taxonomy's differences. A tool called GeoBench Aligner[1] is implemented to build a characterized database namely GeoBench DB. This tool consists of a semi-automatic process that collects data from existing LBS providers and a user's validation is required to indicate the differences between entities and to decide whether two entities correspond. A second tool called PABench Extractor takes GeoBench DB as an input to generate datasets describing a given situation of differences. Thus, our benchmark namely PABench[2] (POI Alignment Benchmark), consists of these datasets and a list of metrics that assess the performance and quality of matching approaches. In the following, the construction of GeoBench DB is given and the necessary specifications of PABench are discussed.

4.1 GeoBench DB - Construction of Characterized Database

The goal of this section is to build a database characterized according to the taxonomy's differences. GeoBench Aligner is a web-based application addressed to experts; it serves to build such characterized database in a semi-automatic matching process through the sets of several LBS providers. We start by representing an overview of GeoBench Aligner and describing its main algorithms. Then, a prototype of GeoBench Aligner is represented. Finally, experimentation and population of GeoBench DB are given.

4.1.1 Overview of GeoBench Aligner

GeoBench Aligner aims to help users find and characterize the corresponding entities between several LBS providers. To do so, among the available providers, we consider one of them as a source provider, while remaining providers are considered as target providers. The concept behind this tool is to consider a source entity from the source provider and to finds the potentially corresponding target entities at each target provider. Then, an expert has to choose the correct corresponding target entity (for each target provider) and identify the differences between the source and target entities.

[1] GeoBench Aligner: http://liris-unimap01.insa-lyon.fr/GeoBench_Aligner [Accessed: June 2016]
[2] PABench: http://liris-unimap01.insa-lyon.fr/benchmark/test_cases [Accessed: June 2016]

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

As mentioned earlier, to match the entities of two sets we first need to match their schemas. The matching between the schemas of LBS providers involved in GeoBench Aligner is done manually (see Section 4.1.2). Also, each provider has its own hierarchy of labels for representing POIs' types. An alignment has been manually produced between the hierarchies of the LBS providers (see Section 4.1.3); this alignment serves for blocking purposes. Thus, Figure 4.1 illustrates the processes of GeoBench Aligner. The tool takes two inputs, one of them being a parameter κ that represents the maximal number of potentially corresponding target entities proposed by the tool for each target provider. The other would be a POI specified by the user or randomly chosen by the tool. Then, GeoBench Aligner searches for this POI at the source provider and returns a list of entities that may refer to the specified POI. The user selects the desired source entity and then GeoBench Aligner queries the target providers using a blocking algorithm (see Section 4.1.4). For each target provider, the tool obtains a set of potentially corresponding target entities. A basic matching algorithm is in charge of ranking the target entities (see Section 4.1.5). The top κ correct target entities are proposed to the user who compares them to the source entity. The user makes the final decision to choose the correct corresponding target entity and to select the differences that exist between the source and target entities at each level (spatial, primary and secondary terminological). This process can be repeated by choosing a new POI in order to create GeoBench DB.

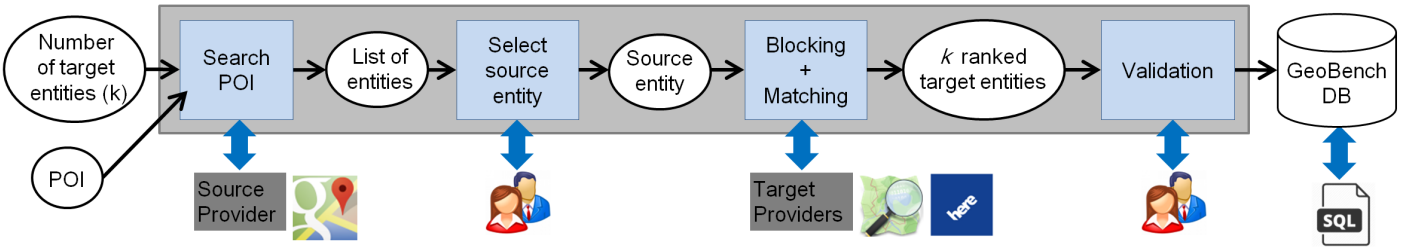


FIGURE 4.1: Overview of the processes involved in GeoBench.

Thus, GeoBench DB contains real-world entities collected from several LBS providers. For each pair of entities, we know their relevance of correspondence and their situation of differences. Figure 4.2 shows the relational model of GeoBench DB; it consists of three tables. The "entities" table contains all entities processed in GeoBench Aligner. Each tuple represents an entity that refers to a POI. The attribute *POI_id* represents an internal identifier, the two attributes *provider* and *provider_POI_id* represent the LBS provider that offers the entity and the identifier of this entity in the set of its provider, respectively. Remaining attributes describe the spatial, primary and secondary terminological information of the POI. The "correspondences" table contains the pair of corresponding entities. The attribute *id* represents an internal identifier of a correspondence, the two attributes *id_Entity_Source* and *id_Entity_Target* refer to the identifiers

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

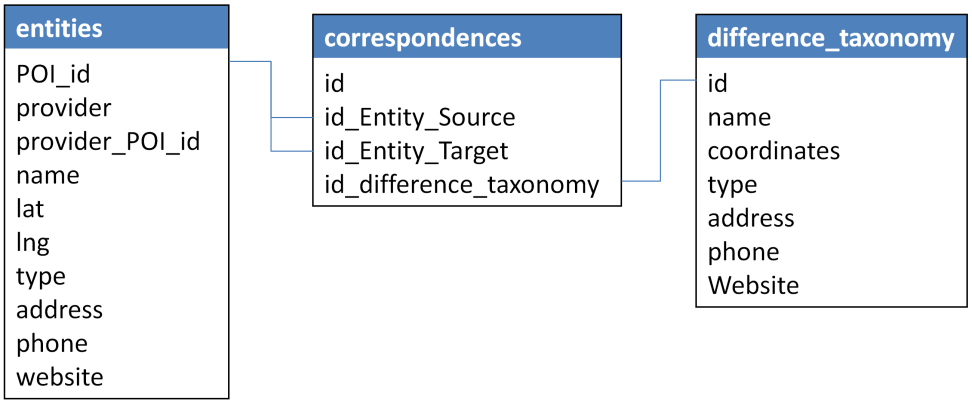


FIGURE 4.2: Relational model of GeoBench DB.

of two corresponding entities. The attribute *id_Difference_Taxonomy* points to the situation of differences represented in "difference_taxonomy" table. This latter represents the situations of differences for each pair of corresponding entities. Each attribute represents the difference (e.g. Different Locations, Missing Data, etc.) for each specific information of the POI.

4.1.2 Schema Matching

Although the study of schema matching and ontology alignment has generated many approaches [GLM04], the schemas of LBS providers are sufficiently small and static enough to be manually matched. To do so, we use the concept of the holistic approach [BMR11]; it consists of constructing a single mediator schema and aligning the elements of a large corpus of schemas. Devogele uses this approach to create a unified schema for geographic information describing road data [Dev97]. The benefit of this approach is that each time we need to add a new provider to GeoBench, it is sufficient to align its schema only with the mediator schema. In addition, eliminating one provider from the tool will not affect the schema matching of the other providers.

Figure 4.3 shows the schema matching of three LBS providers with the mediator schema of GeoBench Aligner. As mentioned earlier in the taxonomy, spatial and primary terminological attributes exist in the schemas of all providers (see Section 3.1 in Chapter 3). This means that our mediator schema requires at least three attributes including POI name, type and location coordinates (e.g. latitude and longitude). Concerning the secondary terminological attributes, they differ from one provider to another. In GeoBench Aligner, only the most common secondary information are considered namely phone number, website and address. Note that for the *Complex Correspondences* between attributes (see Section 3.3.1 in Chapter 3), basic mapping functions are used to normalize

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

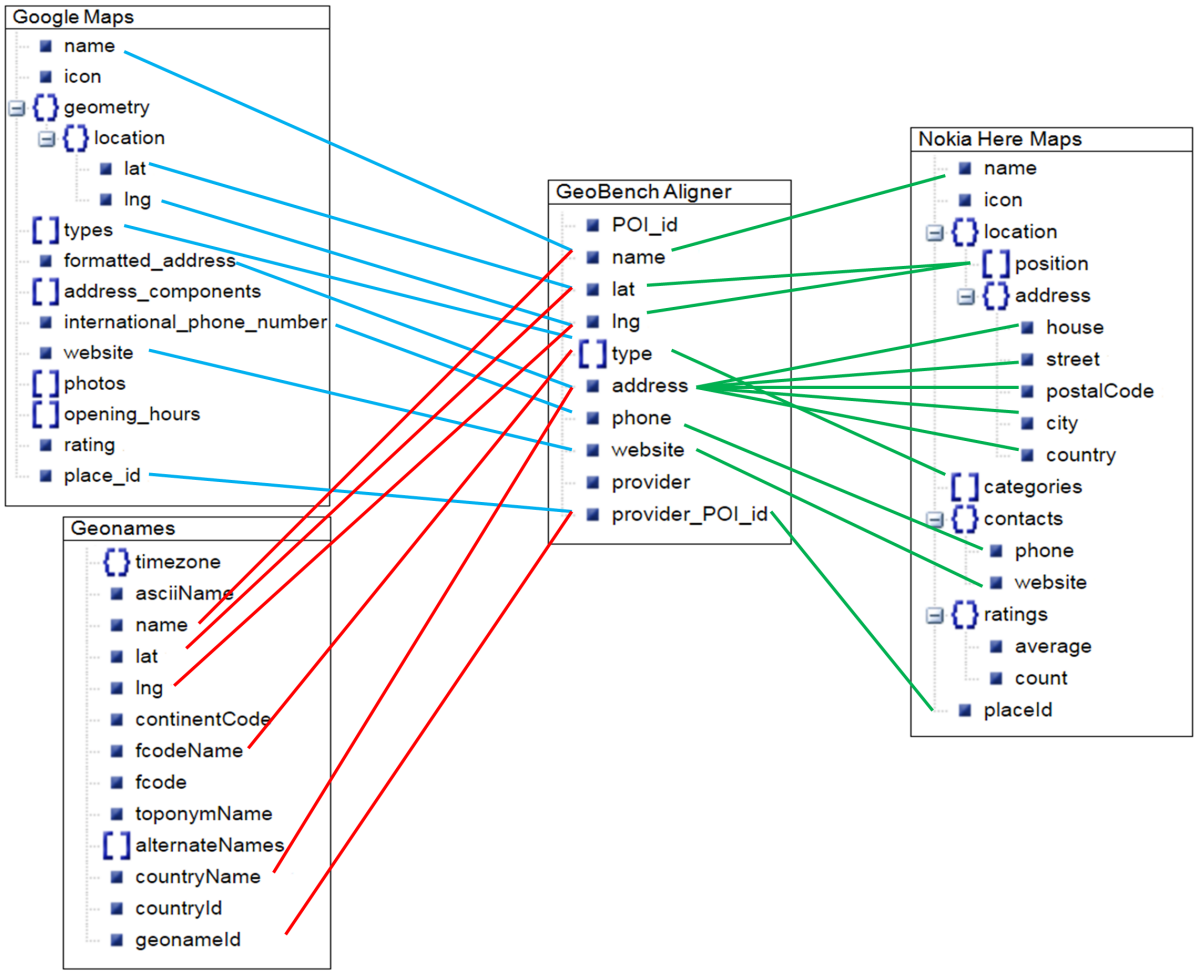


FIGURE 4.3: Mediator schema of GeoBench aligned to three LBS providers.

them, such as split and concatenate. For instance, concerning Nokia Here Maps schema, the *location.position* attribute needs to be split to extract *latitude* and *longitude*.

4.1.3 POIs' Types Matching

As mentioned previously, the APIs of LBS providers allow users to request POIs and to refine the results by adding filters such as find the nearest POI of a given type (see Section 2.1 in Chapter 2). The matching between the POIs' types allows us to perform a blocking for compared entities, which may improve the quality and the performance of entity matching. Each LBS provider has its own hierarchy of labels for representing POIs' types. For instance, as for May 2016, Google Maps[3] supports 97 types, Nokia Here Maps[4] supports 82 types and Geonames[5] supports approximately 670 distinct types. Similar to the schema matching, a mediator hierarchy is created for the GeoBench

[3]POI types of Google Maps: https://developers.google.com/places/supported_types
[4]POI types of Nokia Here Maps: http://places.demo.api.here.com/places/v1/categories/places
[5]POI types of Geonames: http://www.geonames.org/export/codes.html

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

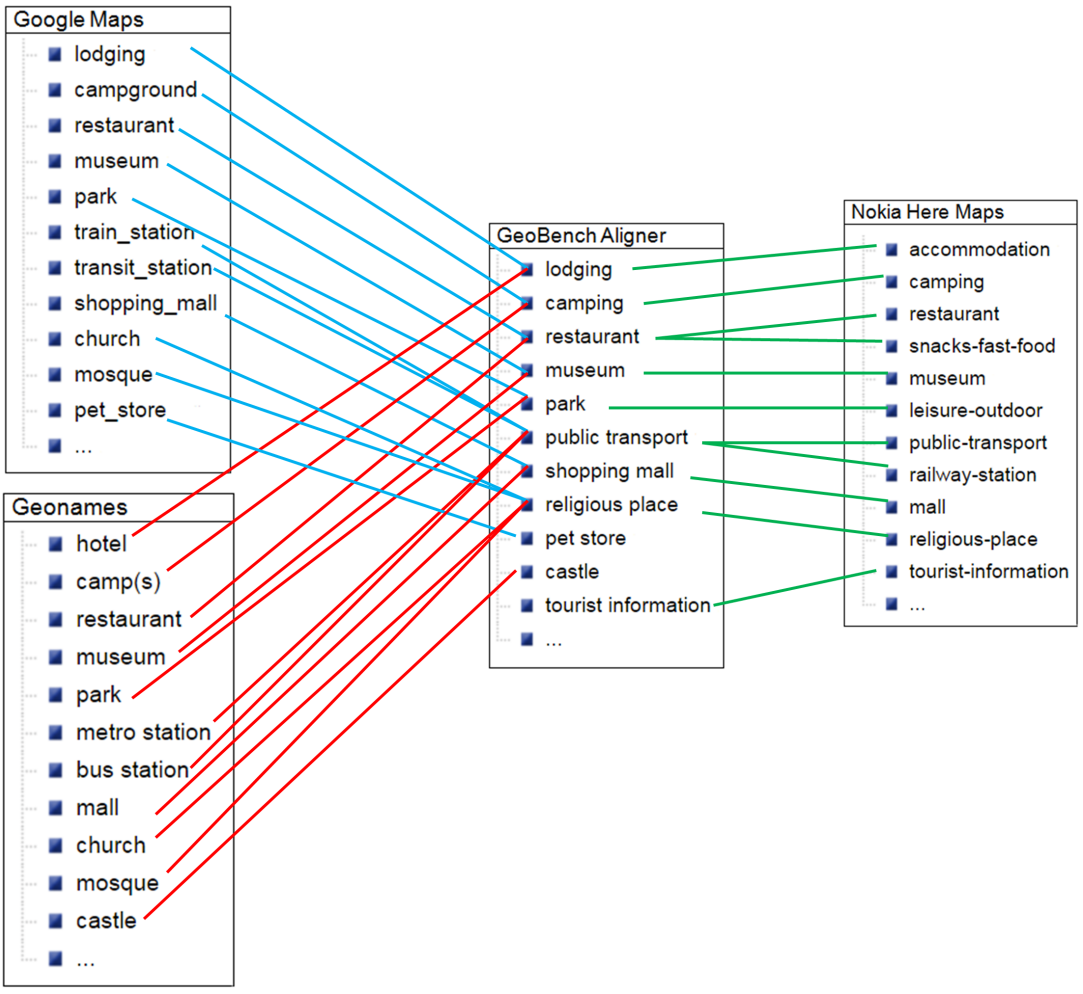


FIGURE 4.4: Mediator hierarchy of GeoBench aligned to three LBS providers.

Aligner that is manually aligned to the hierarchies of LBS providers. The mediator hierarchy considers a subset of the most commonly used types in tourist field like lodging, restoration, museum, etc. Figure 4.4 shows a part of the hierarchy matching of three LBS providers with the mediator hierarchy of GeoBench Aligner.

The categorization of POI types differ from one provider to another. For instance, Nokia Here Maps distinguishes between two types such as “restaurant” and “snacks-fast-food”, while Google Maps and Geonames have only one type for food, namely “restaurant”. Besides, some types are not supported by all providers like “pet_store” from Google Maps, “castle” from Geonames and “landmark” from Nokia Here Maps. For such cases, a blocking using the POI type cannot be performed. However, the next section presents an algorithm that uses the POI types and other information to perform an efficient blocking.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

4.1.4 Blocking Algorithm

Each provider owns millions of geospatial entities. A blocking algorithm aims at quickly constraining a relevant subset of entities among all those available. In our context, the blocking algorithm needs to select a few target entities which likely represent the source entity. In other words, for each source entity, we create a block that contains the target entities that may correspond to the source entity. According to the state-of-the-art, most LBS providers allow a *blocking distance* through a *radius query* that requests the nearby POIs within a given distance of a location (see Section 2.1 and Section 2.2.3 in Chapter 2). Intuitively, we could perform the blocking based on the coordinates of the source entity and include all entities within a radius. Yet, this is not sufficient for two reasons. First, even a small area (e.g., city centers) may contain thousands of entities, thus limiting the benefit of the blocking. Second, larger POIs, such as mountain ranges or parks, may have their coordinates either in the center of the POI or at one of the entrance. In order to perform an efficient blocking, terminological information such as POI name and type, are used in addition to spatial coordinates. We start by defining a blocking area using the source entity coordinates and a radius. The radius value is adjusted according to the POI's type. For example, for restaurants or hotels, the radius value is set to 100 meters while for a park it is equal to 1000 meters. Then, the result of the blocking consists of the union set of the two following queries executed separately. The first query returns all target entities of the blocking area which have the same type as the source entity. The second query returns all entities of the blocking area whose name shares a token with the source entity's name. Concerning the first query, we use the alignment between the POIs types that has been done in Section 4.1.3. Concerning the second query, it can be done using a *keywords query* offered by providers' APIs (see Section 2.2.3 in Chapter 2). Thus, the limited result of the blocking is a set of potentially corresponding entities that can now be compared with the source entity using state-of-the-art matching techniques.

4.1.5 Basic Matching Algorithm

The blocking algorithm has constrained the number of target entities to be matched, and the matching process aims at computing a confidence score between each of those selected entities and the initial source entity. The challenge of the matching algorithm is to produce relevant confidence scores for ranking entities resulting from the blocking process. A confidence score close to 0 means that a target entity is completely dissimilar to the source entity. A confidence score equal to 1 indicates that both entities are equivalent, according to the matching algorithm. Contrary to the blocking algorithm, which quickly identifies potential corresponding entities using three attributes,

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

the matching algorithm is based on sophisticated but costly similarity measures applied to all attributes.

To obtain the confidence score, we compute similarity values between the attributes of a blocked entity and those of the source entity. Let us describe the different attributes and how we compare them. The coordinates of two entities are compared according to the Euclidean distance. The shorter the distance between both entities is, the closer to 1 the similarity value for coordinates. Inspired by Zhang et al., the following formula $\text{Similarity} = 1 - \frac{\text{distance}}{\text{radius}}$ is used to calculate the coordinates similarity [ZCSK13]. Concerning the terminological information, there are several string similarity measures such as Levenshtein, Jaccard, JaroWinkler and MongeElkan [CRF03]. Each similarity measure has its advantages; for example, Levenshtein is able to capture misspellings, while JaroWinkler performs well for matching person names and abbreviations. The choice of similarity measure function depends upon the application. Sehgal et al. show that the best performance for matching geospatial names is obtained by the Levenshtein [SGV06]. Hence, in our context, the Levenshtein measure is applied between the names of compared entities. The attributes corresponding to the concept phone number and website are also matched using the Levenshtein measure. The address attribute requires a pre-processing step to normalize the different formats. Comparing each of the individual elements of an address (e.g., postcode, city, etc.) involves a new problem for smartly combining the different similarity values. Hence, the individual elements of an address are merged into one normalized element during the schema matching phase, so that the Levenshtein measure can be applied. The normalization of the address is as follows: street number, street name, postcode, city and country. The main advantage of such normalization is that a difference in the postcode value or in the street number value (which are common mistakes) does not strongly affect the similarity computed between two addresses. When all the individual scores have been computed, we need to compute a confidence score. A weighted average is traditionally used for combining the individual similarity values. GeoBench Aligner also combines them with this technique and it provides more weight to the most important attributes. Indeed, the secondary attributes such as phone or address may be missing for a provider. Thus, we tune their weight to one-third, while the primary attributes have a weight equal to two-thirds.

A decision step is finally required to select the potential correspondences. Various methods such as a threshold or the top-K enable this automatic selection [BBR11]. In semi-automatic approaches, proposing the top-K correspondences to the user is the most appropriate choice because the user has to manually verify these suggested correspondences [KSG07]. Additionally, it is easier for an end-user to tune a parameter related to a number of propositions to be validated (κ in our case) rather than tuning a mysterious threshold value. At the end of the matching process, GeoBench Aligner outputs for

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

each target provider an ordered list of at most κ entities which are ranked according to their confidence score, and the user validates those entities which may correspond to the source entity.

4.1.6 GeoBench Aligner Prototype - User Interface

LBS providers share their POIs data using web services REST APIs (see Section 2.1 in Chapter 2). Also, they offer additional web services (e.g. Javascript libraries) that help to develop LBSs' tools. To benefit from these techniques, a prototype of GeoBench Aligner has been implemented as a web-based application. A simple configuration file allows the user to set up the necessary parameters such as the APIs' URLs of providers, selecting the source provider and specifying the radius for each POI type. The current version deals with three LBS providers namely Google Maps, Nokia Here Maps and Geonames. But, it is possible to add more providers through the configuration file, the user must manually add (i) the APIs URLs and (ii) the schemas and POIs' types alignment between new providers and GeoBench Aligner.

The prototype of GeoBench Aligner benefits from two interfaces. The first one namely "Search Interface", it concerns the phase in which the user searches for a POI and selects a source entity. The second one namely "Matching Interface", concerns the matching phase between the selected source entity and the target entities proposed by the tool. Figure 4.5 shows a use case of the "Search Interface", which is composed of four panels. Panel #1 allows the user to search for a random or specific POI. The parameter κ is set to 5 by default, but users can change it by clicking on "Advanced Search". In this use case, the user searches for *Universite toulouse* and the type of POI is set to *Univeristy*. Panel #2 displays a list of source entities that match the user's query. These source entities are retrieved from the source provider, which is Google Maps in this use case. When a user selects an entity from panel #2, its information will be shown in panel #3 and it will be located on the map in panel #4. In Figure 4.5, *Université Toulouse - Jean Jaurès* has been selected and located. To start the matching phase for the selected source entity, the user has to click on the "Start matching" button in panel #3.

Figure 4.6 shows the matching phase. Panel #5 contains a ranked list of the top κ potentially corresponding entities retrieved from the target providers, which are Nokia Here Maps and Geonames in this use case. The user can toggle between the list of target providers. For each target provider, the user looks for the correct corresponding target entity from panel #5. Once found and selected, it will be located in panel #6 and its information will be shown in panel #7. Then, the user compares the locations of selected source and target entities (panel #4 vs. panel #6) and their attributes' values (panel #3

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

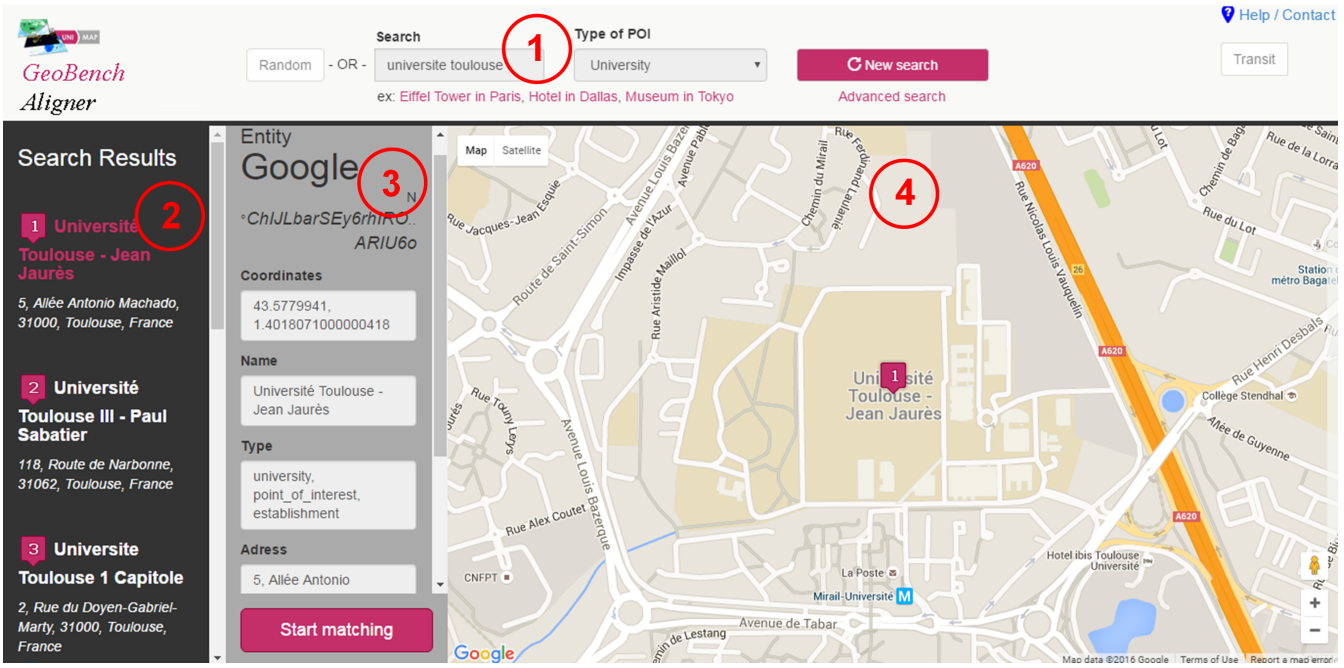


FIGURE 4.5: Search Interface of GeoBench Aligner.

vs. panel #7). For each attribute in panel #7, the user selects the difference between source and target entities using the radio buttons. For instance, source and target entities have different values for coordinates attributes, but both entities are correctly located on the maps. This means that the coordinates attributes have an Equipollent Difference and the "EP" radio button is selected. Concerning the terminological information, the name attributes' of source and target entities have the same value (*Université Toulouse - Jean jaurès*), so no difference is selected. Otherwise, the type attributes have semantically distinct values (*university* vs. *Educational Facility*). This means that the type attributes have a Semantic Different Data difference and the "SEMDD" radio button is selected. Remaining terminological attributes are compared similarly. Once the user finishes the comparison, the "$\surd$" green button at the bottom of panel #7 is used to save the result into GeoBench DB. Panel #8 contains a progress bar that shows the number of target entities processed for each target provider. Finally, suppose that the user has matched a source entity e_1 with two target entities e_2 and e_3 from the two target providers, respectively. Consequently, e_2 and e_3 are also two corresponding entities, the "Transit" button at the left of panel #1 allows the user to compare and match e_2 with e_3. This transitivity increases the number of corresponding entities in GeoBench DB.

4.1.7 Experimentation and Population of GeoBench DB

The current version of GeoBench Aligner includes three LBS providers namely Google Maps, Nokia Here Maps and Geonames. During the experimentation, we alternate the

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

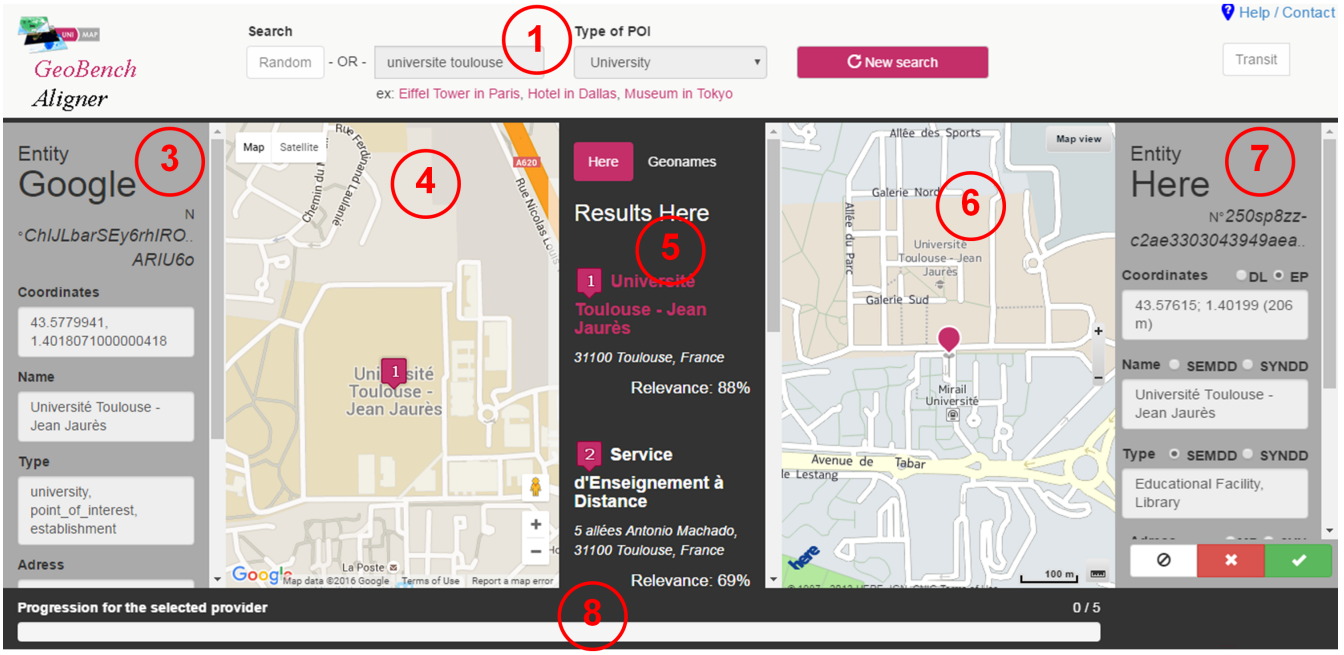


FIGURE 4.6: Matching Interface of GeoBench Aligner.

use of three providers as a source provider. We matched entities of several POI types such as hotel, restaurant, park, museum, hospital and university. These entities are distributed in several geographical zone and countries. Note that for the differences of spatial level, we assume that if the distance between two entities is less than 10 meters, then no spatial difference is considered. Thus, using GeoBench Aligner, we created GeoBench DB that contains entities describing real-world POIs and, for each pair of entities, we know the relevance of correspondence and the situation of differences. This database does not contain any redundancies or duplicated entities. GeoBench DB is available online for researchers in SQL Dump format[6]. It can be easily parsed in ways which researchers and evaluators want. In addition, the datasets of providers are available separately, as is the ground-truth between each pair of them.

Currently, as of June 2016, GeoBench DB contains approximately 3150 entities including 1700 correspondences. Let $\mathbb{E}_{GM}$, $\mathbb{E}_{NH}$ and $\mathbb{E}_{GN}$ be the datasets of Google Maps, Nokia Here Maps and Geonames respectively. Table 4.1 provides some statistics on these datasets such as the number of entities of each provider and the number of correspondences between them.

Note that the *Farthest distance* between two corresponding entities (see Section 3.2 in Chapter 3) may reach tens of kilometers, this is due to many reasons such as when a POI changes its address and only one provider updates its dataset or when two corresponding entities refer to a POI with a very large geographic zone (e.g., lake, park). The last

[6]Download GeoBench DB: liris-unimap01.insa-lyon.fr/benchmark/test_cases [Accessed: June 2016]

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

Pairs of data sets	# of entities	# of correspondences	Farthest distance (km)	AVG / SD (km)
$\mathbb{E}_{GM}$ vs. $\mathbb{E}_{NH}$	1487 × 1131	989	13.74	0.14 / 0.72
$\mathbb{E}_{GM}$ vs. $\mathbb{E}_{GN}$	1487 × 526	418	23.96	0.22 / 1.44
$\mathbb{E}_{NH}$ vs.$\mathbb{E}_{GN}$	1131 × 526	300	22.79	0.29 / 1.48

TABLE 4.1: Statistics on providers' datasets collected by GeoBench Aligner (June 2016).

column AVG/SD represents the average and the standard deviation of the distances between corresponding entities. For the three pairs of datasets, the standard deviation has a relatively high value compared to the average. This means that the majority of corresponding entities are separated with distances less than the average. In other words, the average is not an accurate representative for the variation of distances. For example, between $\mathbb{E}_{GM}$ and $\mathbb{E}_{NH}$, there are 847 correspondences out of 989 having distances less than 0.14 km and between $\mathbb{E}_{GM}$ and $\mathbb{E}_{GN}$ there are 384 correspondences out of 418 having distances less than 0.22 km.

4.2 PABench - An Evaluation Benchmark

PABench has been constructed based on the differences defined in our taxonomy (see Section 3.3 in Chapter 3), i.e., inconsistencies that exist between corresponding entities. Deciding whether two geospatial entities correspond is a challenging task due to the differences that occur between them. As previously mentioned, two corresponding entities being compared may have a combination of differences where each combination is a distinct situation of differences. To understand the weaknesses and strengths of a geospatial entity matching approach, the evaluation must be characterized according to the situations of differences that may occur between entities. In other words, it is required to evaluate an approach based on each situation of differences.

4.2.1 Model of Situations

The possible situations of differences are computed based on the taxonomy with respect to the entity matching task. Since the entity matching goal is to detect the corresponding entities, only the differences concerning corresponding entities are considered, namely *Different Locations* (DL) and *Equipollent Positions* (EP) from the spatial category and *Missing Data* (MD), *Semantic Different Data* (SEMDD) and *Syntactic Different Data* (SYNDD) from the terminological category. The inconsistencies of the schema category namely *Attribute Heterogeneity* and *Different Structures* are excluded from PABench because they will be handled during a manual schema matching phase.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

Spatial information is only expressed by an entity's location, it may have zero (i.e., no difference) or only one difference in the spatial category differences. The set of spatial differences S_dif is given by:

$$S_dif = \{\emptyset, \text{DL}, \text{EP}\}$$

Primary terminological information is expressed by an entity's name and type, it may have zero, one (i.e., at least one attribute has the difference) or two differences of the terminological category differences. MD cannot be considered because the primary terminological attributes are always provided and have values (see Section 3.1 in Chapter 3). The set of primary terminological differences PT_dif is given by:

$$PT_dif = \{\emptyset, \text{SEMDD, SYNDD, (SEMDD, SYNDD)}\}$$

Secondary terminological information varies from one provider to another, it may have zero, one (i.e., at least one attribute has the difference), two (i.e., each difference appears at least once) or three differences of the terminological category differences. The set of secondary terminological differences ST_dif is given by:

$$\begin{aligned} ST_dif \quad = \quad & \{\emptyset, \text{SEMDD, SYNDD, MD, (SEMDD, SYNDD)}, \\ & \text{(SEMDD, MD), (SYNDD, MD), (SEMDD, SYNDD, MD)}\} \end{aligned}$$

Let $Situations_dif$ be the set of all possible combinations of differences that may occur between two corresponding entities at all levels (spatial, primary terminological and secondary terminological):

$$\begin{aligned} Situations_dif = \{(a, b, c) \setminus \\ a \in S_dif, b \in PT_dif, c \in ST_dif\} \end{aligned}$$

where $|Situations_dif| = 3 \times 4 \times 8 = 96$

Note that each situation $s \in Situations_dif$ must be unique and exclusive, in the sense that the situations do not share any relation between them such as intersection or inclusion. For instance, consider two corresponding entities that have the following combination of differences. For spatial information they have DL, for primary terminological information they have SYNDD for the POI name and for secondary terminological information they have SYNDD for the phone number and MD for the website. These two

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

entities have only one situation $s \in Situations_dif$ that is given by:

$$s = (\text{DL, SYNDD, (SYNDD, MD)})$$

Hence, a geospatial entity matching approach must be evaluated with each of these situations in order to discover it weaknesses and strengths. The next section describes the evaluation's process of a given situation.

4.2.2 Test Cases and Metrics

After describing the situations of differences that may appear between two corresponding entities, we intend to create test cases to evaluate the geospatial entity matching approaches against these situations. To do so, for each situation $s \in Situations_dif$, we define a test case denoted TestCase(s). This latter consists of a source dataset $\mathbb{E}_S \subset$ GeoBench DB, a target dataset $\mathbb{E}_T \subset$ GeoBench DB and a ground-truth between source and target datasets.

$$\text{TestCase(s)} = (\mathbb{E}_S, \mathbb{E}_T, \text{ ground-truth})$$

All corresponding entities between $\mathbb{E}_S$ and $\mathbb{E}_T$ should have the situation of differences s. On this basis, to evaluate the performance of a matching approach according to a given situation, we first generate the test case of this situation, then we match the source and target datasets of the test case using the matching approach. Thus, if the matching approach returns the expected answer according the ground-truth, then this approach is able to deal with the given situation. Table 4.2 provides the top ten test cases according to the number of correspondences. The entire table is available online along with PABench[7].

Note that in the practice of the LBS context, some situations of differences rarely occur. For instance, although GeoBench DB contains 1700 correspondence, there are 48 situations that do not appear to any correspondences. This is due to the nature of information represented by some attributes. For example, two phone numbers can never be semantically different. However, it is still possible to develop an entity generator tool that takes a subset of entities to modify the values of their attributes in order to create a target dataset that expresses these rare situations.

In order to differentiate a matching approach from other similar approaches, we need some metrics to measure the performance of matching source and target datasets of a test case. According to the state-of-the-art, there are several common metrics for

[7] PABench Extractor: liris-unimap01.insa-lyon.fr/benchmark/test_cases [Accessed: June 2016]

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

Test Case #	Situation of Differences	Number of Correspondences
84	(EP, SYNDD, MD)	177
17	(∅, SYNDD, ∅)	152
20	(∅, SYNDD, MD)	115
52	(DL, SYNDD, MD)	112
19	(∅, SYNDD, SYN)	92
60	(DL, (SEMDD, SYNDD), MD)	71
92	(EP, (SEMDD, SYNDD), MD)	69
68	(EP, ∅, MD)	68
76	(EP, SEM, MD)	64
36	(DL, ∅, MD)	52

TABLE 4.2: Top ten test cases according to the number of correspondences (June 2016).

evaluation (see Section 2.3.2 in Chapter 2). In PABench, we mainly focus on measuring the efficiency and effectiveness of matching approaches. The former consists of measuring the time and amount of memory allocated by an approach to detect the correspondences. While the latter ensures the result's quality using the standard performance measures that come from the information retrieval domain:

- *Precision*: computes the proportion of correct correspondences detected by the matching approach among all detected correspondences.
- *Recall*: computes the proportion of correct correspondences detected by the matching approach among all correct correspondences given the ground-truth.
- *F-measure*: computes trade off between *Precision* and *Recall*.

These metrics help us understand to what degree a matching approach is able to handle a given situation. Notet that evaluators are free to use other metrics which may be more convenient for their contexts. The next section describes a tool that generate the test cases using the real-world data of GeoBench DB.

4.2.3 PABench Extractor

To facilitate the generation of test cases, a web-based tool, called PABench Extractor[8], has been implemented. It uses GeoBench DB to generate the source and target datasets of each test case. This tool allows configuring the characteristics of a test case through a set of parameters in order to control aspects such the number of correspondences and data formats. Once the characteristics of a test case are configured, source and target

[8] PABench Extractor: liris-unimap01.insa-lyon.fr/benchmark/test_cases [Accessed: June 2016]

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

datasets are retrieved with a ground-truth file, so the evaluators can assess the results of their matching approaches. The tool searches for all pairs of entities that match the requested situation of differences, then the entities of each pair are distributed between the source and the target datasets. Afterwards, the internal identifiers of corresponding entities are listed together in the ground-truth file. Current version of PABench Extractor allows users to add Not Found Entities (NFE), i.e., entities that do not have any correspondences in the other dataset. These NFE entities are chosen randomly. We do not consider the other inconsistencies namely Superposition (SUP), Similar Data (SD) and Duplicate Entities (DE), because they are hard to be detected with GeoBench Aligner. But, it is possible to develop an extension tool that takes a subset of source entities and modify the values of their attributes (spatial, primary and secondary terminological) in order to create target entities havin resemblances, i.e., entities that are not corresponding but have similar data.

4.3 Conclusion

This chapter deals with the evaluation of geospatial entity matching approaches. In a first stage, we implemented the semi-automatic tool, GeoBench Aligner, that helps in collecting and comparing POIs from several LBS providers. Using GeoBench Aligner, we created GeoBench DB, a POIs database that contains entities having distinct types and distributed in several countries. These entities are characterized according to the taxonomy of LBS; for each pair of entities, GeoBench DB indicates the relevance of correspondence and the differences of each attribute. As of June 2016, GeoBench DB contains approximately 3150 entities including 1700 correspondences. In a second stage, we presented the necessary specifications to create an evaluation benchmark called PABench. We defined the situation of differences list that covers 96 possible combinations of differences which may appear between two entities. Then, for each situation, we defined a test case that allows researchers to evaluate a matching approach against the given situation. The PABench Extractor tool is in charge to generate the test cases through the data of GeoBench DB. Hence, PABench allows researchers to find out if a geospatial entity matching approach is able to detect the corresponding entities that have a given situations of differences, which allows the discovery of weaknesses and strengths of matching approaches. PABench is therefore used to evaluate and compare the geospatial entities matching approaches proposed in the next chapter.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

Chapter 5

Matching Geospatial Data

Contents

5.1 Overview for Geospatial Entity Matching 87

5.2 Normalized-Distance: A Spatial Similarity Measure 88

5.3 Global Similarity: Combining Similarity Measures 95

5.4 Decision Algorithm 99

5.5 Experimental Evaluation Using Real-World Datasets 101

5.5.1 Evaluation and Selection of Terminological Similarity Measures 102

5.5.2 Evaluation and Selection of Spatial Similarity Measures 110

5.5.3 Evaluation of Hybrid Approaches Applied to Spatial and Terminological Attributes 119

5.6 Conclusion 125

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

Matching two datasets consists of two main phases namely schema matching and entity matching. Current LBS providers have simple schemas. Hence, similar to GeoBench Aligner (see Section 4.1.2 in Chapter 4), the schema matching task is done manually and it will no longer be discussed in this chapter. Concerning the entity matching, it consists in measuring the similarity between attributes' values in order to detect whether two entities correspond (i.e., refer to the same POI). In the literature, there are several measures such as Levenshtein [Lev66] and Trigram [Ull77] that measure the similarity between terminological values. This chapter compares and evaluates some of these measures that have been used in geospatial entity matching approaches [SGV06, MJA13, OR07, SSL12]. On the other hand, the state-of-the-art of geospatial entity matching represents several approaches that measure the similarity of spatial attribute only [SKS+10, ZCSK13]. This chapter investigates whether existing spatial similarity measures are compatible with our context and proposes a generalization for a spatial similarity measure, namely Normalized-Distance (ND). In addition, several approaches have proved that measuring the similarities of several attributes, spatial and terminological, and combining them improve the result of matching [SGV06, MJA13]. This chapter discusses these approaches and propose a new method namely Global Similarity (GS), that numerically combines several similarities. Moreover, our taxonomy (see Section 3.5.2 in Chapter 3) proved that the resemblances impose a negative impact on the result of matching. In this chapter, a new decision algorithm is proposed in order to reduce the impact of resemblances and improve the result of matching.

Thus, this chapter represents an overview of our approach for geospatial entity matching and describes its related contributions namely Normalized-Distance (ND), Global Similarity (GS) and Decision algorithm. In addition, experimental evaluations using the real-world and characterized data of PABench are given in order to evaluate and compare our propositions to some existing works. First, these evaluations consist in evaluating distinct similarity measures separately in order to select the most appropriate one for each attribute. Then, we evaluate the matching process by combining the similarities of selected measures.

5.1 Overview for Geospatial Entity Matching

Figure 5.1 shows the process of our approach for geospatial entity matching. Firstly, given two datasets from two LBS providers, a blocking phase is performed on these datasets to restrict the matching search area and avoid a costly Cartesian product comparison. There are several blocking methods described in Section 2.2.2.2 of Chapter 2 that can be applied using spatial attributes (i.e., location coordinates) such as blocking

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

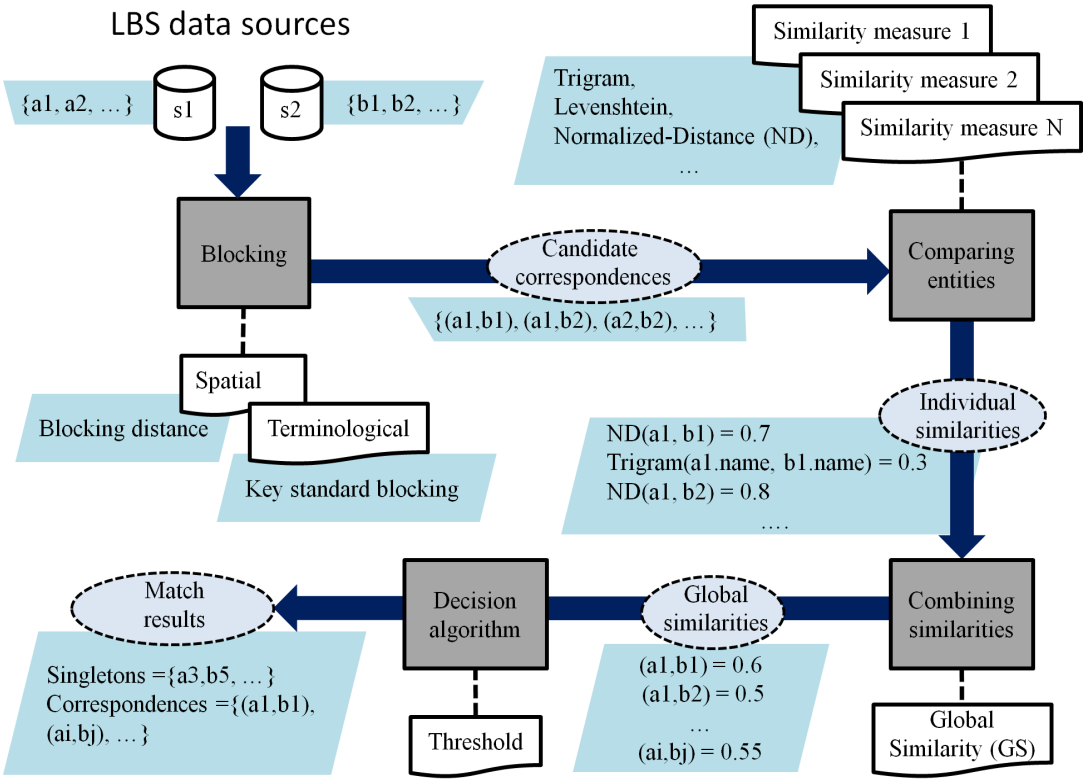


FIGURE 5.1: Process of geospatial entity matching approach.

distance, blocking bounding and tiles, or terminological attributes such as Key standard blocking and Canopy Clustering. In our context, blocking distance using location coordinates and Key standard blocking using POI types mapping (see 4.1.3 Section in Chapter 4) are used to perform such blocking. This latter produces a set of pair wise candidates that may be potentially corresponding. The second phase consists of comparing entities by measuring the similarities of corresponding attributes for each candidate; one similarity per attribute. The similarity measure of each attribute has been selected beforehand by a performance assessment of distinct spatial and terminological similarity measures. The comparison of entities produces several individual similarities for each candidate. The next phase consists of combining these individual similarities in order to obtain one global similarity per candidate. Finally, the last phase proposes a threshold-based decision algorithm that automatically selects singletons and corresponding entities using the global similarities.

5.2 Normalized-Distance: A Spatial Similarity Measure

In the context of geospatial entity matching, similarity measures or similarity functions are used to quantify the similarity between two objects. A spatial similarity measure

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

quantifies the similarity between two geospatial entities using their spatial information such as location coordinates. According to the state-of-the-art (see Section 2.2.3 in Chapter 2), several spatial similarity measures for punctual objects have been proposed in distinct approaches. Intuitively, the spatial similarity is related to the distance between compared entities. Usually, researchers calculate the Euclidean or Haversine distance between the locations coordinates of two entities, and then estimate their similarity using the calculated distance. Beeri et al. propose two approaches that use only spatial information namely Mutually-Nearest Join (MN) and Normalized-Weights (NW) [BKSS04, BDK+05, SKS+10]. The former considers that two entities are similar only if they are mutually nearest to each other; it produces a Boolean similarity and does not quantify the similarity. The latter uses probability; the similarity between two compared entities is equal to the distance between them over the sum of distances between them and remaining entities. This probability-based method quantifies the similarity and produces a value between 0 and 1. A similarity equals 1 means that compared entities are completely similar. Conversely, a similarity equals 0 means that compared entities are completely dissimilar. These two approaches are context-based; the similarity between two entities strongly depends on their neighbor entities. In other words, two compared entities may have different similarities when their neighbor entities change. For instance, consider two entities mutually nearest to each other, their similarity is 1 and they are considered as corresponding according to MN. But, if a new entity is placed between them, then the similarity of the two initial entities is set to 0 and they will no longer be considered as corresponding although the distance between them remains unchanged. Remaining spatial similarity measures are non context-based and use one concept; the closer the entities, the higher the spatial similarity is. Some approaches do not explain how the similarity is calculated [OR07, KSG07]. Other approaches calculate the similarity by defining mathematical functions. Sehgal et al. use the inverse of the distance, $Similarity = \frac{1}{\text{distance}}$ [SGV06]. Statistics of GeoBench DB (see Section 4.1.7 in Chapter 4) shows that the distances' average of corresponding entities exceeds 100 meters. Now consider two corresponding entities having Different Locations difference and separated by 50 meters, they have a 0.02 similarity. Also consider two corresponding entities having Equipollent Positions difference and separated by 200 meters, they have a 0.005 similarity. This function has a sharp decreasing rate; the similarities calculated with this function are too small and do not express a good quantification of reality. Zhang et al. propose the following function for spatial similarity measure, $Similarity = \left|1 - \frac{\text{distance}}{\text{threshold}}\right|$ [ZCSK13]. Authors do not explain what happens when the distance exceeds the *threshold* defined in the formula. We assume that the similarity is set to 0 as the distance gets greater than the given threshold. This measure can produce a similarity between 0 and 1 with a linear decreasing rate and without any dependency to the neighbor entities. This spatial similarity measure seems to express a reasonable quantification for distances in

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

our context. But, we can also imagine other mathematical functions such as quadratic and exponential, that can act like this measure with different decreasing rates.

On these bases, we propose a generalization of a spatial similarity measure to quantify the distances between geospatial entities. A spatial similarity measure should fit the following requirements:

1. The similarity depends only on the compared entities, without any considerations to other entities.
2. The similarity value should belong to an interval with defined minimum and maximum values [min, max], which are usually set to 0 and 1, respectively.
3. The maximum similarity value is obtained when the compared entities share the same locations, i.e., the distance between them is equal to 0.
4. The minimum similarity value is obtained when the compared entities have a distance greater than a given value β called *blocking distance* (see Section 2.2.3 in Chapter 2).
5. The spatial similarity measure should be a decreasing function over the interval of distance, i.e., the similarity decreases as the distance increases. Typically, the Euclidean distance belongs to the interval $\mathbb{R}^+$.

Definition 5.1. Let Normalized-Distance (ND) be a spatial similarity measure:

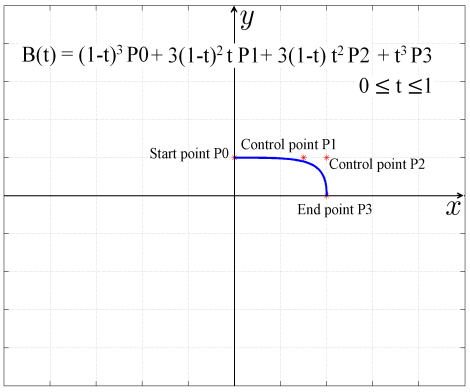
$$\begin{aligned} \mathrm{ND}\colon \mathbb{E} \times \mathbb{E} &\rightarrow [0,1] \\ (e, e') &\rightarrow \mathrm{ND}(d(e, e')) \end{aligned}$$

where $d(e, e')$ is the Euclidean distance between (e, e') and $\mathrm{ND}(d(e, e'))$ is their similarity value.

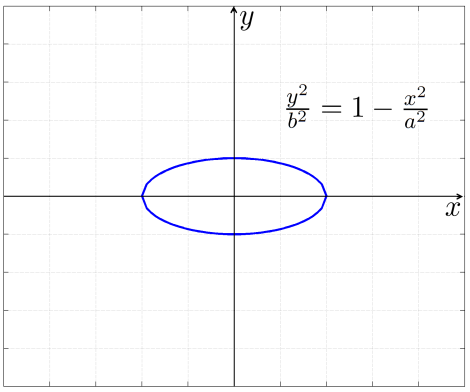
Hypothetically, there are an unlimited number of continuous and discontinuous mathematical functions that can fit the requirements of ND. Figure 5.2 shows a set of common curves and mathematical functions namely Cubic Bezier, Ellipse, Quadratic, Gaussian, Linear, Hyperbolic and Exponential. Table 5.1 summarizes these functions by adapting their parameters to fit ND's requirements. The last column represents the curves of these functions where x-axis refers to the distance. These curves plot the decay of NDs as the distance increases.

Figure 5.3 compares the decay rate of the above NDs. For more precision, let's consider a 100 meters *blocking distance* (i.e., $\beta = 100$). Table 5.2 shows the variations of similarity values when the distance increases from 0 to β with steps of 10 meters. $\mathrm{ND}_{\mathrm{BezierMax}}$

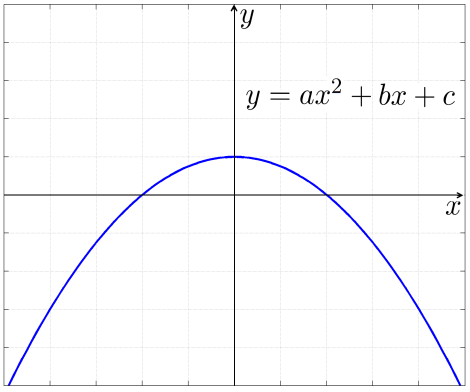
Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés



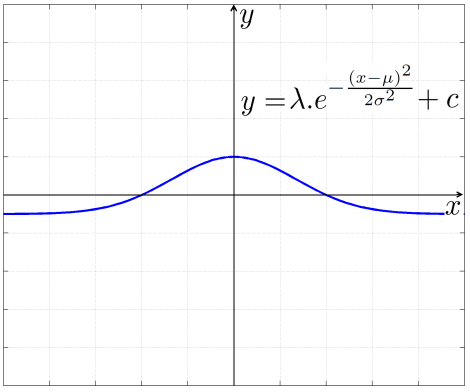
(a) Cubic Bezier with logarithmic decay



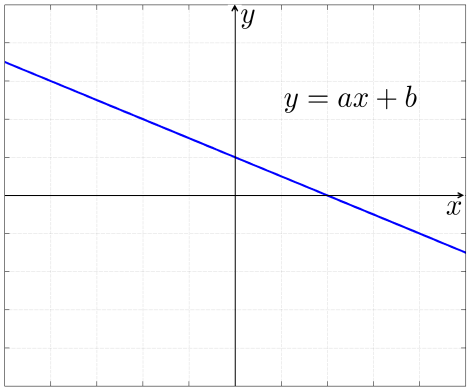
(b) Ellipse



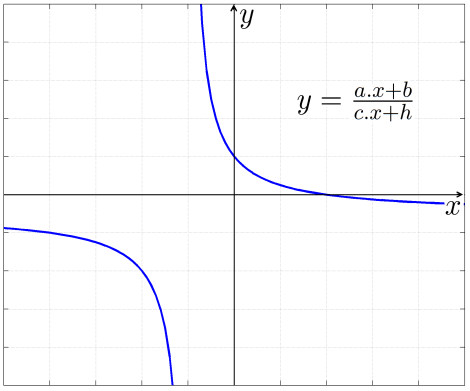
(c) Quadratic



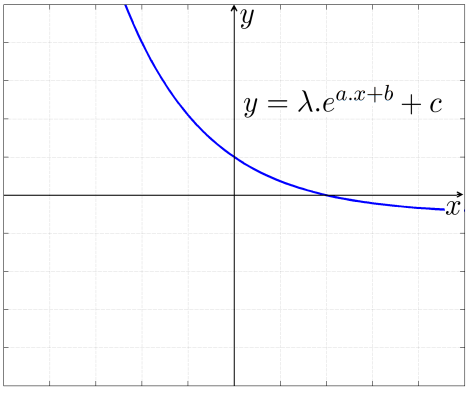
(d) Gaussian



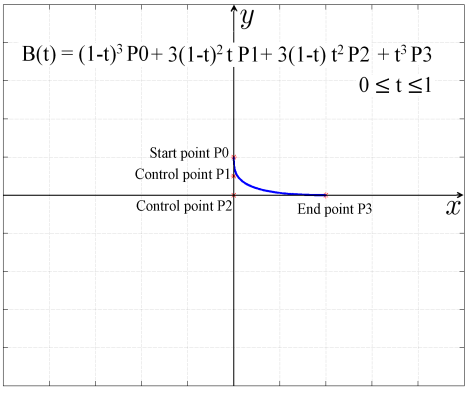
(e) Linear



(f) Hyperbolic



(g) Exponential



(h) Cubic Bezier with exponential decay

FIGURE 5.2: Curves of common mathematical functions that fit ND's requirements.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

Function	General equation	Similarity equation	Similarity curve
BezierMax	$B(t)=(1-t)^3P0$ $+3t(1-t)^2P1$ $+3t^2(1-t)P2$ $+t^3P3$ where $0 \leq t \leq 1$	$\mathrm{ND}_{\mathrm{BezierMax}}(d) = \begin{cases} y_P \quad P0=(0,1), P1=P2=(\beta,1), \\ \quad\quad P3=(\beta,0) \\ \quad\quad \text{for } d \leq \beta, P \in B(t), x_P = d, \\ 0 \quad \text{for } d > \beta \end{cases}$	
Ellipse	$\frac{y^2}{b^2} = 1 - \frac{x^2}{a^2}$	$\mathrm{ND}_{\mathrm{Ellipse}}(d) = \begin{cases} \sqrt{1-\frac{d^2}{\beta^2}} & \text{for } d \leq \beta \\ 0 & \text{for } d > \beta \end{cases}$	
Quadratic	$y = ax^2 + bx + c$	$\mathrm{ND}_{\mathrm{Quadratic}}(d) = \begin{cases} 1-\frac{d^2}{\beta^2} & \text{for } d \leq \beta \\ 0 & \text{for } d > \beta \end{cases}$	
Gaussian	$y = \lambda.e^{-\frac{(x-\mu)^2}{2\sigma^2}} + c$	$\mathrm{ND}_{\mathrm{Gaussian}}(d) = \begin{cases} (1-c)e^{d^2.\frac{ln(-c)-ln(1-c)}{\beta^2}} + c & \text{for } d \leq \beta, c < 0 \\ 0 & \text{for } d > \beta \end{cases}$	

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

Linear	$y = ax + b$	$\mathrm{ND}_{\mathrm{Linear}}(d) = \begin{cases} 1 - \frac{d}{\beta} & \text{for } d \leq \beta \\ 0 & \text{for } d > \beta \end{cases}$	
Hyperbolic	$y = \frac{a.x+b}{c.x+h}$	$\mathrm{ND}_{\mathrm{Hyperbolic}}(d) = \begin{cases} \frac{d-\beta}{c.d-\beta} & \text{for } d \leq \beta, c < 0 \\ 0 & \text{for } d > \beta \end{cases}$	
Exponential	$y = \lambda.e^{a.x+b} + c$	$\mathrm{ND}_{\mathrm{Exponential}}(d) = \begin{cases} (1-c)e^{d.\frac{ln(-c)-ln(1-c)}{\beta}} + c & \text{for } d \leq \beta, c < 0 \\ 0 & \text{for } x > \beta \end{cases}$	
BezierMin	$B(t)=(1-t)^3P0$ $+3t(1-t)^2P1$ $+3t^2(1-t)P2$ $+t^3P3$ where $0 \leq t \leq 1$	$\mathrm{ND}_{\mathrm{BezierMin}}(d) = \begin{cases} & \text{for } d \leq \beta, P \in B(t), x_P = d, \\ y_P & P0 = (0,1), P1 = P2 = (0,0), \\ & P3 = (\beta, 0) \\ 0 & \text{for } d > \beta \end{cases}$	

TABLE 5.1: Set of common mathematical functions and curves that fit ND's requirements.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

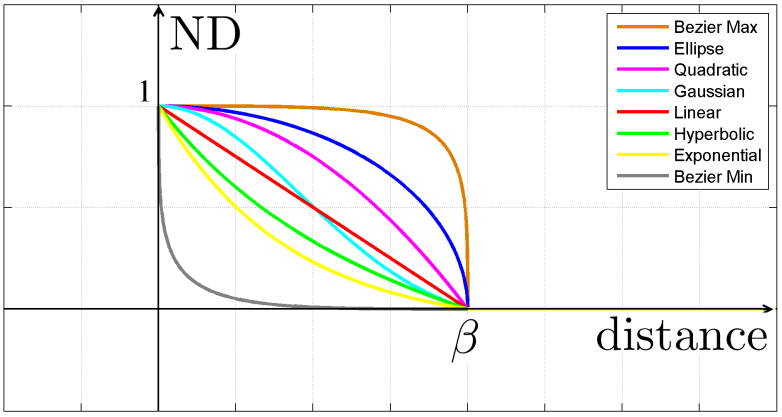


FIGURE 5.3: Decay rate of several NDs functions.

		Distance d										
		0	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**	**100**
ND(d)	**BezierMax**	1	1	1	1	1	0.99	0.98	0.96	0.93	0.84	0
	Ellipse	1	0.99	0.98	0.95	0.92	0.87	0.8	0.71	0.6	0.44	0
	Quadratic	1	0.99	0.96	0.91	0.84	0.75	0.64	0.51	0.36	0.19	0
	Gaussian	1	0.98	0.94	0.86	0.76	0.64	0.51	0.38	0.24	0.12	0
	Linear	1	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0
	Hyperbolic	1	0.86	0.73	0.61	0.5	0.4	0.31	0.22	0.14	0.07	0
	Exponential	1	0.84	0.7	0.58	0.47	0.37	0.28	0.2	0.12	0.06	0
	BezierMin	1	0.16	0.07	0.04	0.02	0.01	0	0	0	0	0

TABLE 5.2: Variations of ND's functions applied to the interval $[0; \beta]$ with $\beta = 100$.

refers to the cubic Bezier curve with the highest logarithmic decay with respect to ND requirements; it has the slowest decreasing rate and produces very high similarities. For instance, the similarity equals 1 for a distance smaller than 40 meters and it achieves 0.84 for a distance of 90 meters. In contrast, $\text{ND}_{\text{BezierMin}}$ refers to the cubic Bezier curve with the lowest exponential decay with respect to ND requirements; it has the fastest decreasing rate and produces very low similarities. For instance, the similarity equals 0 for a distance greater than 50 meters.

These different ND functions allow us measuring the spatial similarity between two entities. Later in this chapter, the performances of these ND functions will be evaluated and compared to existing spatial similarity measures by matching real-world datasets.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

Such evaluation allows selecting a spatial similarity measure to be combined with terminological similarity measures.

5.3 Global Similarity: Combining Similarity Measures

According to the state-of-the-art, some geospatial entity matching approaches use one information, spatial or terminological, to detect the corresponding entities. Some others use both kinds of information, which should give more accuracy when selecting corresponding entities. Several strategies such as numeric-based, rule-based and composite-based are used for combining similarities (see Section 2.2.2.2 in Chapter 2). Note that we distinguish between the hybrid similarity measures that combine several similarity measures applied to one information and the hybrid matching approaches that combines distinct similarity measures applied to several independent information. The former intends to improve the quantification of the similarity of a given information, while the latter intends to find an agreement between the similarities of several information. Scheffler et al. use spatial information for blocking purpose, then apply a hybrid similarity measure that combines two string similarity measures applied to the *name* attribute using rule-based technique [SSL12]. Two experiments were done by matching 50 POIs from Facebook places and 50 POIs from Qype with OpenStreetMap POIs separately, which produces 64% and 76% accuracy, respectively. Safra et al. propose a composite-based combination, such as the union or intersection of the results of similarity measures applied separately [SKSD06]. Experiments are done by matching 28 entities from Google Maps with 39 entities from Yahoo; these entities represent POIs of type hotels. The highest *F-measure* equals 93% obtained with a union combination. Olteanu propose a numeric-based combination using the "Belief Theory" that is based on probability consideration [OR07]. For their experiments, authors used two real datasets about geographic reliefs to show some use cases of applying the "Belief Theory." However, they did not give the performance results of the whole matching. Sehgal et al. propose an hybrid approach that numerically combine several similarities using machine-learning techniques [SGV06]. Authors learn the weights of each similarity measure for a weighted average combination; three learning algorithms were compared namely logistic regression, voted perceptron and support vector machines. These algorithms have been evaluated by matching two big datasets that represent POIs such as cemeteries and airfields. The datasets used in experiments are not challenging because such POIs with large geographic area, do not express an interesting heterogeneity. For instance, two different positions for an airfield can easily be detected because it is impossible to find another large POI inside the airfield that may confuse the choice. However, the combination of the name, location and type similarities using logistic regression achieves

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

94.1% *F-measure* and outperforms the two other algorithms, as well as it outperforms the results of using one single similarity measure for matching. McKenzie et al. also propose a learning-based hybrid approach [MJA13]. Experiments are done by matching 140 correspondences that have been manually extracted and matched from Yelp and Foursquare. Combining the name, location and users reviews similarities using logistic regression learned weights achieves 97% accuracy, while combining the same similarities using the standard unweighted average achieves 95% accuracy.

The context of LBS is very dynamic; providers databases contain millions of entities that are located using different strategies, refer to POIs of different types and have different number of independent attributes. For instance, if we request hotels in some quarter in Paris from two different LBS providers, the results may be hundreds of entities in a small area, and the matching between them would be a hard task. Learning weights and evaluations of existing approaches have been done using small, random and non characterized datasets. This does not guarantee neither enough precision nor a fair comparison of their results.

On these bases, we intend to propose an hybrid geospatial entity matching approach that produces one global similarity for two compared entities. This approach must respect the following specifications:

- Non context-based: the hybrid approach combines the similarities of independent attributes of two given entities. This means that an individual similarity between two corresponding attributes (e.g., "POI name" vs. "place_name") depends only on the values of these attributes regardless of the values of other attributes, the number of entities offered by providers, the overlap or the density of providers datasets. Hence, the global similarity of the hybrid approach is not affected by the variation and dynamicity of LBS context.
- Numeric-based: the hybrid approach numerically combines the independent similarities of two entities and produces one global similarity that belong to an interval with defined minimum and maximum values [min, max], which are usually set to 0 and 1, respectively. Such combination facilitates the estimation of data's certainty of corresponding entities in order to inform the end users and tourists about the quality of information.

Consider two entities e and e' offered by two LBS providers that are in the same block. For each pair of corresponding attributes between e and e', that does not have a Missing Data difference, we calculate an individual similarity s_i using a simple or hybrid similarity measure that returns a numeric value between 0 and 1.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

Definition 5.2. Let Global Similarity (GS) be an hybrid geospatial entity matching approach that combines n independent similarities of two entities:

$$\begin{aligned} \mathrm{GS}\colon \mathbb{E} \times \mathbb{E} &\to [0,1] \\ (e, e') &\to \mathrm{GS}(s_1, s_2, ..., s_n) \end{aligned}$$

where $s_i \in [0,1]$ is an individual similarity and n is the number of pairs of corresponding attributes.

There are several methods in which GS can combine the individual similarities. With respect to the state-of-the-art, we intend to cover all possible numeric combinations:

1. A pessimistic combination returns the lowest similarity among the available individual similarities.

   Definition 5.3. Let $\mathrm{GS}_{\mathrm{Min}}$ be the pessimistic combination of n similarities between two entities:

   $$\begin{aligned} \mathrm{GS}_{\mathrm{Min}}\colon \mathbb{E} \times \mathbb{E} &\to [0,1] \\ (e, e') &\to \mathrm{GS}_{\mathrm{Min}}(s_1, ..., s_n) = \min(s_1, ..., s_n) \end{aligned}$$

   To achieve the highest performance of $\mathrm{GS}_{\mathrm{Min}}$, for each pair of corresponding attributes, we use a similarity measure that can achieve a satisfying matching result with low similarities.

2. An optimist combination returns the highest similarity among the available individual similarities.

   Definition 5.4. Let $\mathrm{GS}_{\mathrm{Max}}$ be the optimist combination of n similarities between two entities:

   $$\begin{aligned} \mathrm{GS}_{\mathrm{Max}}\colon \mathbb{E} \times \mathbb{E} &\to [0,1] \\ (e, e') &\to \mathrm{GS}_{\mathrm{Max}}(s_1, ..., s_n) = \max(s_1, ..., s_n) \end{aligned}$$

   Conversely to $\mathrm{GS}_{\mathrm{Min}}$, the highest performance of $\mathrm{GS}_{\mathrm{Max}}$ is achieved by combining similarity measures that produce a satisfying matching result with the high similarities.

3. An extreme average combination returns the average between lowest and highest similarities among the available individual similarities.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

Definition 5.5. Let $\mathrm{GS}_{\mathrm{ExAvg}}$ be the extreme average combination of n similarities between two entities:

$$\mathrm{GS}_{\mathrm{ExAvg}} : \mathbb{E} \times \mathbb{E} \rightarrow [0,1]$$
$$(e, e') \rightarrow \mathrm{GS}_{\mathrm{ExAvg}}(s_1, ..., s_n) = \frac{\min(s_1, ..., s_n) + \max(s_1, ..., s_n)}{2}$$

To achieve the highest performance of $\mathrm{GS}_{\mathrm{ExAvg}}$, for each pair of corresponding attributes, we select the similarity measure that can achieve a satisfying matching result with moderate similarities.

4. An un-weighted average combination returns the average of all available individual similarities.

   Definition 5.6. Let $\mathrm{GS}_{\mathrm{Avg}}$ be the un-weighted average combination of n similarities between two entities:

$$\mathrm{GS}_{\mathrm{Avg}} : \mathbb{E} \times \mathbb{E} \rightarrow [0,1]$$
$$(e, e') \rightarrow \mathrm{GS}_{\mathrm{Avg}}(s_1, ..., s_n) = \mathrm{avg}(s_1, ..., s_n)$$

   Similar to $\mathrm{GS}_{\mathrm{ExAvg}}$,the highest performance of $\mathrm{GS}_{\mathrm{Avg}}$ is achieved by combining the similarity measures that produce a satisfying matching result with moderate similarities.

5. A probability-based combination returns a trade-off between n individual similarities.

   Definition 5.7. Let $\mathrm{GS}_{\mathrm{Pr}}$ be the probability combination of n independent similarities between two entities:

$$\mathrm{GS}_{\mathrm{Pr}} : \mathbb{E} \times \mathbb{E} \rightarrow [0,1]$$
$$(e, e') \rightarrow \mathrm{GS}_{\mathrm{Pr}}(s_1, ..., s_n) = \begin{cases} p & \text{if } \exists s_i, s'_i \backslash s_i = 0 \wedge s'_i = 1 \\ \frac{p. \prod_{i=1}^{n} s_i}{p. \prod_{i=1}^{n} s_i + (1-p). \prod_{i=1}^{n} (1-s_i)} & \text{Otherwise} \end{cases}$$

   where p is the *a-priori* probability that e and e' are corresponding entities. Naturally, if we have no knowledge of the *a-priori* probability of an event X, then we assume symmetry between True and False, i.e., $p = 1/2$. An exceptional case arises if there is two individual similarities that are defined and completely opposite, this means at least one similarity equals 1 and at least one similarity equals 0, then the probability equation returns an undefined result. To avoid this issue, we impose $\mathrm{GS}_{\mathrm{Pr}}$ to the *a-priori* probability. Appendix A represents a detailed description of this probability combination. According to the above formula, if all individual

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

similarities are greater than p value, then GS_{Pr} pulls the global similarity to 1. Conversely, if all individual similarities are lesser than p value, then GS_{Pr} pulls the global similarity to 0. Otherwise, if the individual similarities are distributed above and below p value, then GS_{Pr} results a trade-off.

GS_{Pr} clusters the global similarity near 0, *p-value* (0.5) and 1. This means its performance must be evaluated with the similarity measures that produce the satisfying matching result for low, moderate and high similarities, separately.

These different GS's methods allow us calculating a global similarity between two entities. Later in this chapter, the performances of these GSs will be evaluated and compared to existing hybrid approaches by matching real-world datasets. After computing the global similarities between compared entities, a final step is required to decide whether these entities correspond. The next section represents an algorithm to select the corresponding entities using the calculated global similarities.

5.4 Decision Algorithm

Matching two source and target datasets requires measuring the similarity for each source entity with all target entities of the same block in order to find the most suitable target entity to be chosen as a corresponding one. A decision algorithm is defined as an algorithm that specifies how two entities should be chosen as corresponding according to their Global Similarity. As mentioned in the state-of-the-art, various methods such as a threshold or the top-K enable this automatic selection [BBR11]. In our context, the use of threshold allows refining the set of corresponding entities in order to obtain the best results compared to reality. Several geospatial entity matching approaches use this technique [SGV06, SSL12].

Consider two entities datasets $A = \{a_1, ..., a_n\}$ and $B = \{b_1, ..., b_m\}$ where $GS(a_i, b_j)$ is the global similarity between each pair. Existing approaches that use threshold consider that two entities a_i and b_j are corresponding only if their similarity exceeds a given threshold. This concept may produce 1:n or n:n correspondences. For example, consider a threshold of 0.5, if $GS(a_1, b_1) = 0.6$ and $GS(a_1, b_2) = 0.7$, then a_1 corresponds to b_1 and b_2. In our taxonomy (see Section 3.5.2 in Chapter 3), we proved that the resemblances impose a negative impact on the quality of matching by decreasing the *Precision* when using approaches that produce 1:n or n:n correspondences. For instance, when a small threshold is chosen, even pairs of entities which have low similarities will be detected as correspondences. On one hand, these correspondences may include a lot of False Positive (FP), which decreases the *Precision*. On the other hand, this guarantees that most true

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

correspondences (i.e., correspondences of ground-truth) will be detected, which decreases False Negative (FN) and increases *Recall*. As the threshold increases, the 1:n algorithm becomes more selective because only pairs with high similarities will be detected. On one hand, this decreases FP. On the other hand, the true correspondences that have low similarities will no longer be detected, which increases FN and consequently decreases TP. Hence, increasing FN and decreasing TP will certainly decrease *Recall* according to its formula ($\frac{TP}{TP+FN}$). But, according to *Precision* formula ($\frac{TP}{TP+FP}$), decreasing FP will increase *Precision* and decreasing TP will decrease *Precision*. This means that *Precision* varies depending on the change ratio of FP and TP. In other words, if we avoid FP more than we lost TP, then *Precision* increases. Conversely, if we avoid FP less than we lost TP, then *Precision* decreases.

On these bases, we intend to propose a decision algorithm that produces 1:1 correspondences in order to reduce the impact on result's quality. Firstly, we create a matrix, entities of A are in rows and entities of B are in columns. For each pair (a_i, b_j), we compute their Global Similarity. Then, pairs with the highest similarities that exceeds a given *threshold* are considered as corresponding, while remaining entities are considered as singleton. In other words, two entities a and b are corresponding if their similarity is higher than (1) the similarity between a and each of the remaining b_j, (2) the similarity between b and each of the remaining a_i and (3) a given *threshold*. a and b are two corresponding entities, $a \equiv b$, iff

$$\begin{aligned}
&\forall\, a_i \in A - \{a\}, \forall\, b_j \in B - \{b\} \\
&\mathrm{GS}(a,b) > \mathrm{GS}(a,b_j) \wedge \\
&\mathrm{GS}(a,b) > \mathrm{GS}(a_i,b) \wedge \\
&\mathrm{GS}(a,b) > \mathit{threshold}
\end{aligned}$$

According to the 1:1 correspondences, only pairs with the highest similarities are selected. Such pairs have a high possibility to be corresponding. Remaining pairs that may include a lot of FP, are eliminated even if their similarities are higher than the threshold. Therefore, *Precision* increases and gets more stable regardless of the threshold. On the other hand, corresponding entities that do not have the highest similarity, which is a rare case, will no longer be selected with this decision algorithm. Therefore, a slight decrease is expected for *Recall*. On these bases, if the increasing rate of *Precision* is higher than the decreasing rate of *Recall*, then the *F-measure* increases. Otherwise, the *F-measure* decreases.

This decision algorithm can be also used with any other similarity measure rather than GS. The next section represents the experiments to evaluate spatial and terminological

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

similarity measures, and the combination between several similarity measures. Also, a comparison between the 1:n and 1:1 decision algorithms is given during the evaluation of spatial similarity measures.

5.5 Experimental Evaluation Using Real-World Datasets

The objective of this section is to experimentally assess the effectiveness and efficiency of our propositions on challenging real-world datasets. The effectiveness ensures the result's quality using the standard performance measures namely *Precision*, *Recall* and *F-measure*, while the efficiency is measured according to the execution time and amount of memory allocated by the matching approach (see Section 4.2.2 in Chapter 4). In a first stage, we intend to evaluate and select the best terminological similarity measures for each terminological attribute. Then, we evaluate the spatial similarity measure Normalized-Distance (ND). Finally, we evaluate the Global Similarity (GS) approach based on the results of the first and second evaluations. For each of these three evaluations, two experiments are distinguished:

1. **Test cases evaluation:** evaluates the approaches using the test cases of our benchmark PABench, which allows us discovering strong and weak points of an approach. These test cases contain only corresponding entities having a given situation without any noise data, which allow us discovering whether an approach is able to deal with the correspondences that have the given situation.

2. **Full datasets evaluation:** consists in a general evaluation by matching entities retrieved according to a standard LBS query (e.g. find POIs in a given city), which allows us analyzing the general behaviors of an approach. For this general evaluation, we extract a source and target datasets from GeoBench DB that contain entities located in Paris and referred to POIs of the following types: restaurant, hotel and museum. Each of these datasets contains approximately 600 entities. They have 378 correspondences and the *Farthest Distance* (see Section 3.2 in Chapter 3) is equal to 295 meters.

To ensure a fair comparison between all evaluations, we use the same blocking techniques for all approaches, which are POIs types alignment (see Section 4.1.3 in Chapter 4) and a *blocking distance* β. This latter equals the *Farthest Distance* of the matched datasets in order to guarantee that all corresponding entities will be compared to each other. Therefore, the blocking of distance and type means that an entity from a source dataset will be compared to an entity from target dataset only if the distance between them is

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

lesser than β and having the same type. Note that the blocking using the type would improve the matching result when source and target datasets contain entities of several types.

5.5.1 Evaluation and Selection of Terminological Similarity Measures

This section analyzes the performance of matching using terminological attributes. For each attribute, we evaluate the performances of matching using distinct terminological similarity measures applied separately. In our context, we consider four terminological attributes namely POI *name*, *address*, *phone number* and *website*. Note that the POI *type* attribute is used for blocking purposes, so it is not included in the matching's evaluation. Some attributes, such as *phone* and *website*, need to be normalized to the same format in order to facilitate the comparison. For example, for a *website* we consider only the domain name and we remove all remaining terms such as "http" and "www"; for a phone number we remove all non numerical character such as "+", "/" and "-". Regarding the terminological similarity measures, we consider five known measures, which are used in several geospatial entity matching approaches [SGV06, MJA13, OR07, SSL12], that quantify the similarity between two strings:

1. Equality: returns a Boolean that indicates whether two strings equal each other.
2. Including: returns a Boolean that indicates whether one string is included in the other.
3. Levenshtein: is an edit distance measure [Lev66]. It counts the minimum number of operations (e.g. addition, deletion and change of character) required to transform one string into the other. The number of operations is normalized to the maximum length between the two strings and the similarity equals the complimentary of the normalization to one.
4. Trigram: is a term weighting measure [Ull77]. It converts the words of each string into a continuous sequence of tri-grams, and then compute the average of common grams.
5. TFIDF cosine: is a term weighting measure [Jon72]. It converts each string into a vector of common words and assign a weight for each word. This weight is a statistical measure used to evaluate how important a word is to a string. The similarity equals the cosine of angle between the two weighted vectors.

For this evaluation, we use the 1:1 correspondences decision algorithm defined in Section 5.4 and we repeat experiments by varying the threshold value.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

5.5.1.1 Test Cases Evaluation For Terminological Similarity Measures

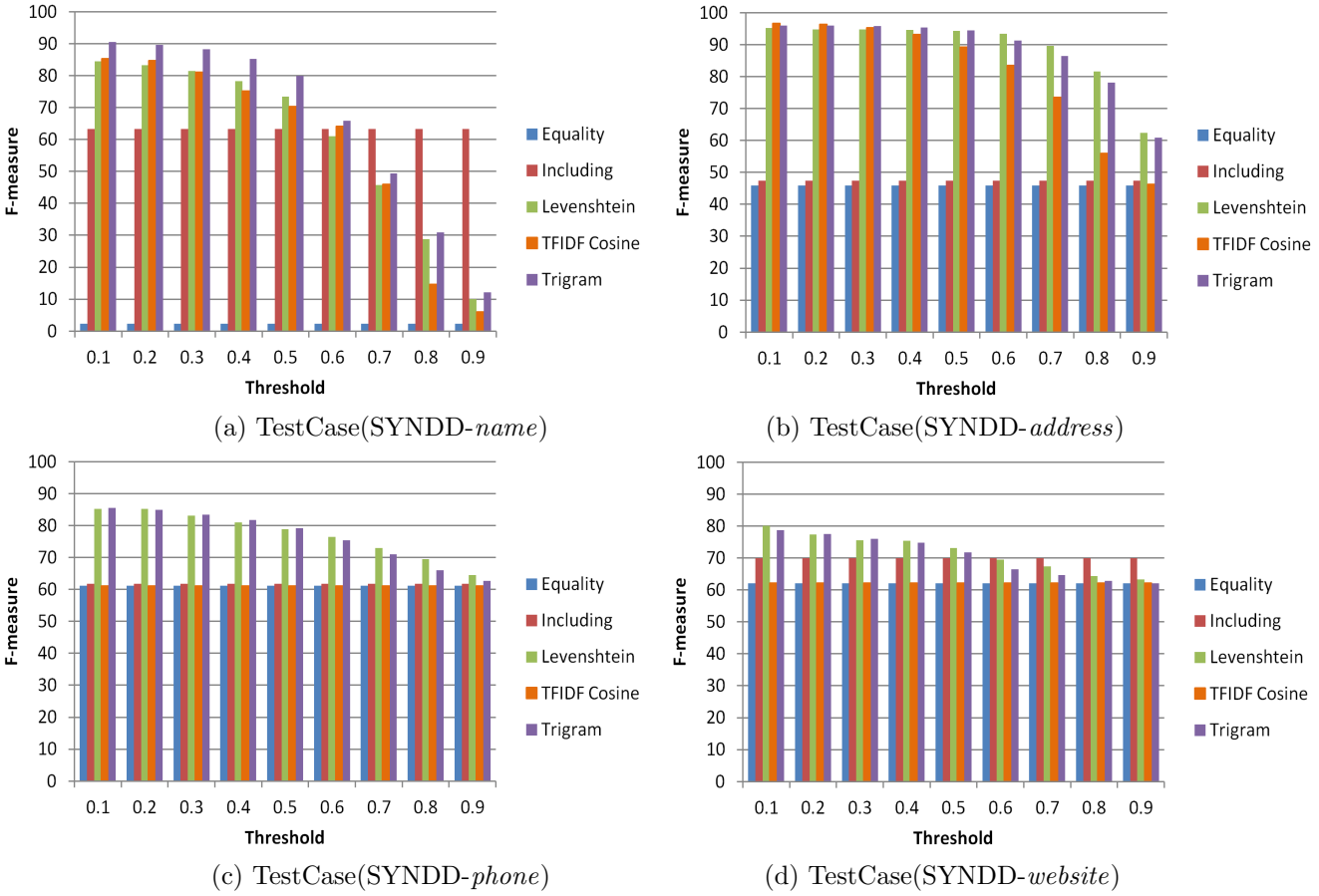
According to our taxonomy, there are three differences that may appear at a terminological attribute namely, Semantic Different Data (SEMDD), Syntactic different Data (SYNDD) and Missing Data (MD) (see Section 3.3.2 in Chapter 3). This section aims at evaluating the performance of matching using one single terminological attribute that has a given terminological difference. To do so, for each attribute and each difference, we generate a test case, denoted TestCase(*difference-attribute*), that combines all test cases of PABench that contain corresponding entities having the given difference for the given attribute, regardless of the differences of other attributes. For instance, TestCase(SYNDD-*phone*) is a test case for the *phone* attribute where the phone values of all corresponding entities have a SYNDD difference. Table 5.3 shows the number of correspondences and the *Farthest Distance* of all possible test cases in this experiment.

TestCase	**# of correspondences**	**Farthest Distance (km)**
SEMDD-*name*	0	N/A
SEMDD-*address*	0	N/A
SEMDD-*phone*	0	N/A
SEMDD-*website*	0	N/A
SYNDD-*name*	685	13.27
SYNDD-*address*	327	6.31
SYNDD-*phone*	300	13.75
SYNDD-*website*	256	13.75
MD-*name*	N/A	N/A
MD-*address*	546	23.97
MD-*phone*	352	12.57
MD-*phone*	511	12.57

TABLE 5.3: Test cases to evaluate string similarity measures applied to terminological attributes.

None of the four attributes has correspondences with SEMDD. This is due to the nature of information represented by these attributes. For instance, two phone numbers can never be semantically different. Concerning TestCase(MD-*name*), the *name* attribute is a primary attribute that always has value (see Section 3.1 in Chapter 3), which means that it can never has MD. So, TestCase(MD-*name*) and all test cases of SEMDD are eliminated from this experiment. In addition, the *address*, *phone* and *website* attributes are secondary and can have MD. But, a terminological similarity measure cannot compare null values in order to detect the correspondences. Consequently, none of the terminological similarity measures can handle the MD difference regardless of the attribute. On these bases, we can still evaluate and compare the terminological similarity measures applied to attributes having SYNDD. Figure 5.4 shows the *F-measure* of each terminological similarity measure applied to each terminological attribute. The x-axis

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés



(a) TestCase(SYNDD-*name*)

(b) TestCase(SYNDD-*address*)

(c) TestCase(SYNDD-*phone*)

(d) TestCase(SYNDD-*website*)

FIGURE 5.4: *F-measure*s of terminological similarity measures in the test cases experiment, applied to terminological attributes having SYNDD.

represents the values of threshold and the y-axis represents the percentage of *F-measure*.

Figure 5.4a shows the performances of matching using the *name* attribute. Equality and Including are not affected by thresholds because they produce a Boolean similarity; they achieve a *F-measure* of 2.3% and 63.2%, respectively. Although that the *name* values have SYNDD, but the Equality measure is still able to detect few correspondences because it can handle the accent cases. For instance, "Imperial Palace" and "Impérial Palace" are two POI names that have SYNDD, but the Equality measure consider them as equal. Among all measures, Trigram achieves the highest *F-measure* up to 90.5% for 0.1 threshold, followed by TFIDF Cosine and Levenshtein that achieve 85.3% and 84.5% *F-measure* for 0.1 threshold, respectively.

Figure 5.4b shows the performances of matching using the *address* attribute. Equality and Including achieve a *F-measure* up to 45.9% and 47.3%, respectively. TFIDF Cosine, Trigram and Levenshtein are equivalent and achieve a highest *F-measure* up to 96% for

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

0.3 threshold. Except that Trigram and Levenshtein are more resistant to the threshold than TFIDF.

Figure 5.4c shows the performances of matching using the *phone* attribute. Equality, Including and TFIDF Cosine are equivalent regardless of the threshold; they achieve approximately the same *F-measure* up to 61%. Trigram and Levenshtein are more effective; both achieve a *F-measure* up to 85% for 0.2 threshold.

Figure 5.4d shows the performances of matching using the *website* attribute. A highest *F-measure* up to 80% is achieved by Levenshtein for 0.1 threshold, followed by Trigram that achieves 78.6% for 0.1 threshold. Including achieves a *F-measure* up to 70%, while Equality and TFIDF Cosine are equivalent and achieve a *F-measure* of 62% regardless of the threshold. Matching using the *website* attribute give the lowest *F-measure*, this is due to the high diversity of values offered by LBS providers. For example, consider a POI of type hotel, one provider may offer the original website of this hotel and a second provider may offer the link of the hotel at Hotels.com, Booking.com or Facebook.com, which impacts the result.

Table 5.4 resumes the *F-measure*s of all measures and represents their efficiencies. The columns F, E and M refers to *F-measure*, execution time in second and amount of memory in MB.

	name			***address***			***phone***			***website***		
	F	**E**	**M**	**F**	**E**	**M**	**F**	**E**	**M**	**F**	**E**	**M**
Equality	2.3	11	5.5	45.9	4	1.5	61.1	4	1.25	62.1	3	1.25
Including	63.2	11	5.5	47.3	4	1.5	61.7	4	1.25	69.9	3	1.25
Levenshtein	84.5	11	5.5	**95.2**	4	1.5	**85.2**	4	1.25	**80**	3	1.25
Trigram	**90.5**	25	5.5	**95.9**	7	1.5	**85.5**	5	1.25	78.6	4	1.25
TFIDF Cosine	85.3	20	5.5	**96.5**	5	1.5	61.1	4	1.25	62.1	3	1.25

TABLE 5.4: Performance of terminological similarity measures in the test cases experiment, applied to terminological attributes having SYNDD.

For each attribute, the terminological similarity measures consume the same amount of memory to detect the correspondences. Concerning the execution time, Equality, Including and Levenshtein are the fastest and equivalent followed by TFIDF Cosine, while Trigram is the slowest measure. Note that the efficiency differs from one attribute to another according to the number of correspondences of the test case and the degree of inconsistency between values.

To conclude, all similarity measures have achieved their highest *F-measure* for low thresholds (0.1-0.2). This means that the attributes' values are very inconsistent and the corresponding entities have low similarities. The SYNDD difference of the *address*

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

attribute can be handled using Levenshtein, Trigram and TFIDF Cosine by achieving 96% *F-measure*. But, the SYNDD of the other attributes are hard to be handled regardless of the similarity measures. In the next section, we analyze whether the behaviors of these measures change when matching full datasets retrieved by a standard LBS query and contain noise data.

5.5.1.2 Full Datasets Evaluation For Terminological Similarity Measures

This evaluation analyzes the behaviors of terminological similarity measures by matching geospatial entities retrieved from several LBS providers based on a standard LBS query. Recall that in this evaluation we match two datasets containing entities located in Paris and referred to POIs of the following types: restaurant, hotel and museum. Each of these datasets contains approximately 600 entities, in which there 378 correspondences. Table 5.5 shows statistics for each terminological attribute in the full datasets with respect to the differences. Concerning the *name* attribute, 35.5% of correspondences have exactly the same values, while the remaining have SYNDD. Concerning the *address* and *phone* attributes, 96% of correspondences have SYNDD, while the remaining have MD. Note that there are no correspondences having exactly the same values for addresses or phone numbers due to the different formatting of LBS providers. However, the normalizing of values should resolve a part of SYNDD. Finally, 36% of correspondences have exactly the same values for the *website* attribute, while 24.9% have SYNDD and 39.1% have MD. On the other hand, the datasets of this evaluation contain singleton entities that may have resemblances with corresponding entities, which may impact the matching result.

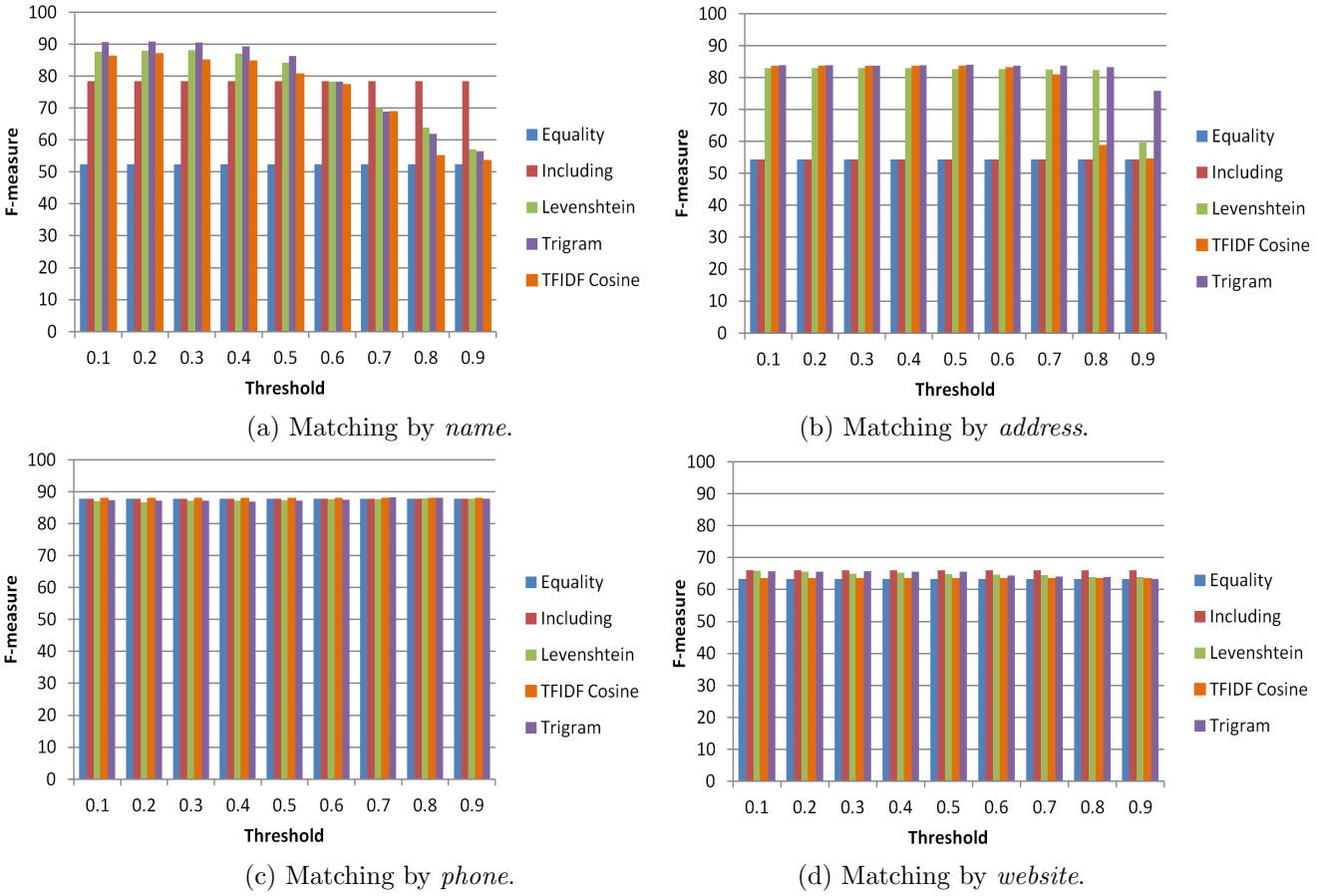
	∅	SYNDD	MD
name	35.7%	64.3%	0
address	0	96%	4%
phone	0	96%	4%
website	36%	24.9%	39.1%

TABLE 5.5: Statistic of terminological attributes in the full datasets evaluation with respect to differences.

Figure 5.5 shows the *F-measure* of each terminological similarity measures applied to each terminological attribute. The x-axis represents the values of threshold and the y-axis represents the percentage of *F-measure*.

Figure 5.5a shows the performances of matching using the *name* attribute. Equality and Including are not affected by threshold because they produce Boolean similarity; they achieve a *F-measure* of 52% and 78%, respectively. Among all measures, Trigram achieves the highest *F-measure* up to 91% for 0.2 threshold, followed by Levenshtein

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés



(a) Matching by *name*.

(b) Matching by *address*.

(c) Matching by *phone*.

(d) Matching by *website*.

FIGURE 5.5: *F-measures* of terminological similarity measures in the full datasets experiment.

that achieves a *F-measure* up to 88% for 0.3 threshold, while TFIDF Cosine achieves a *F-measure* up to 87% for 0.2 threshold. According to Trigram, Levenshtein and TFIDF Cosine, the best *F-measures* are obtained for low thresholds (0.2-0.3) and this means that most detected corresponding entities have low similarities for their *name* attributes.

Figure 5.5b shows the performances of matching using the *address* attribute. Both Equality and Including achieve 54% *F-measure*. Trigram and TFIDF Cosine achieve a *F-measure* up to 84% for 0.5 and 0.4 thresholds, respectively. Finally, Levenshtein achieves a *F-measure* up to 83% for 0.4 threshold. The addresses' values of corresponding entities have moderate similarities (0.4 - 0.5).

Figure 5.5c shows the performances of matching using the *phone* attribute. All similarity measures achieve approximately the same *F-measure* up to 88% regardless of the threshold. Although that 96% of correspondences have SYNND, but most detected corresponding entities have high similarities for their *phone* attributes even with Equality and Including, thanks to the normalizing phase.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

Figure 5.5d shows the performances of matching using the *website* attribute. Equality and Including achieve a *F-measure* of 63% and 65%, respectively. Both Trigram and Levenshtein achieve a *F-measure* up to 66% for 0.1 threshold, then it slightly decreases to 64% for 0.9 threshold. Finally, TFIDF Cosine achieves a *F-measure* up to 63% regardless of the threshold. The stability of measure with respect to the thresholds means that most detected corresponding entities have very high similarities for their website attributes. But, the *F-measure* achieves only 66%, this is due to the high number of correspondences that have MD (39%) and can never be detected.

Matching by comparing the *name* attribute gives the best *F-measure* (91%), followed by the *phone* (88%), *address* (83%) and *website* (66%), respectively. The reason is that corresponding entities having MD difference for their secondary attributes can neither be compared nor detected. Although 96% of corresponding entities have phones and addresses, but the *F-measure* obtained by matching the phones is better than the *F-measure* obtained by matching the addresses . This means that the addresses are more inconsistent than the phone numbers. The lowest *F-measure* is obtained for the *website* attribute because only 61% of corresponding entities have been compared, which decreases *Recall*.

Table 5.6 resumes the F-measures of all measures and represents their efficiencies. The columns F, E and M refers to *F-measure*, execution time in second and amount of memory in MB. The efficiency varies from one attribute to another due to the complexity of compared values. For each attribute, all similarity measures are equivalent in terms of usage memory, while Equality, Including and Levenshtein are equivalent in terms of execution time and faster than TFIDF Cosine followed by Trigram. Matching using the *name* and *website* require less execution time than the *phone* followed by the *address*.

	name			***address***			***phone***			***website***		
	F	**E**	**M**	**F**	**E**	**M**	**F**	**E**	**M**	**F**	**E**	**M**
Equality	52.4	7	5.5	54.4	10	5.7	**87.8**	8	5.5	63.4	7	6
Including	78.4	7	5.5	54.4	10	5.7	**87.8**	8	5.5	**66**	7	6
Levenshtein	88	7	5.5	**83**	10	5.7	**87.9**	8	5.5	**65.9**	7	6
Trigram	**91**	15	5.5	**84**	20	5.7	**88.2**	13	5.5	**65.8**	8	6
TFIDF Cosine	87	13	5.5	**83.5**	16	5.7	**87.8**	11	5.5	63.4	8	6

TABLE 5.6: Performance of terminological similarity measures in the full datasets experiment.

To conclude, concerning the efficiency, similar to the test cases experiment, Equality, Including and Levenshtein are faster than TFIDF Cosine followed by Trigram. Concerning the effectiveness, Trigram achieves the highest *F-measure* up to 91% for the *name* attribute in both test cases and full datasets experiments. Levenshtein, Trigram and

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

TFIDF Cosine achieve the highest *F-measure* for the *address* attribute in both experiments. But, in the full datasets experiment, the result's quality has relatively decreased compared to the test cases evaluation (83% vs 96%). Concerning the *phone* attribute, in the test cases experiment, Trigram and Levenshtein achieve the highest *F-measure* up to 85%. But, in the full datasets experiment all measures achieve an equivalent F-measure up to 87%. This means that the phone number values are less inconsistent in the general experiment, which allows the other measures to produce better results. Finally, concerning the *website* attribute, in the test cases experiment, Levenshtein achieves the highest *F-measure* up to 80%. But, in the full datasets experiment, Including, Levenshtein and Trigram achieve the highest *F-measure* up to 66%. This means that the website values are more inconsistent in the full datasets experiment, which decreases the result's quality. Anyhow, the best result is produced by matching the *name* attribute using Trigram measure in both experiments; a highest *F-measure* up to 91% is achieved.

Matching entities of LBS providers using a single terminological similarity measure is not sufficient to resolve the inconsistencies' issue. Such measures need to be combined with other measures to improve the result's quality. As mentioned earlier, for the combination, we need to select the most appropriate similarity measure for each attribute in order to achieve the highest performance by each GS (see Section 5.3). Based on the results of the test cases and full datasets experiments, we need to select, for each attribute, the terminological similarity measures that produce the best result for a low (i.e., 0.1), moderate (i.e., 0.5) and high (i.e., 0.9) thresholds. Table 5.7 resumes the selection of similarity measures for terminological attributes.

	Low (0.1)	**Moderate (0.5)**	**High (0.9)**
name	Trigram	Trigram	Including
address	Trigram, **TFIDF Cosine**	Trigram, **TFIDF Cosine**	Trigram, **Levenshtein**
phone	Trigram, **Levenshtein**	Trigram, **TFIDF Cosine**	TFIDF Cosine, **Levenshtein**
website	Levenshtein	**Levenshtein**, Including	Including

TABLE 5.7: Selection of the most appropriate terminological similarity measure for each attribute at low, moderate and high thresholds.

Concerning the *name* attribute, Trigram is the most appropriate measure for 0.1 and 0.5 thresholds, while for 0.9 threshold, Including is the best. Concerning the *address* attribute, Trigram or TFIDF Cosine are the most appropriate measures for 0.1 and 0.5 thresholds, while for 0.9 threshold, Trigram or Levenshtein are the best. Concerning the *phone* attribute, Trigram or Levenshtein are the most appropriate for 0.1 threshold, Trigram or TFIDF Cosine are the most appropriate for 0.5 threshold, while for 0.9 threshold, Levenshtein or TFIDF Cosine are the best. Finally, concerning the *website* attribute, Levenshtein is the most appropriate for 0.1 threshold, Levenshtein or Including are the most appropriate for 0.5 threshold, while for 0.9 threshold, Including is the best.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

Concerning the attributes that have more than one option, the measure highlighted in bold is the more efficient one in terms of execution time. The next section evaluates and compares spatial similarity measures.

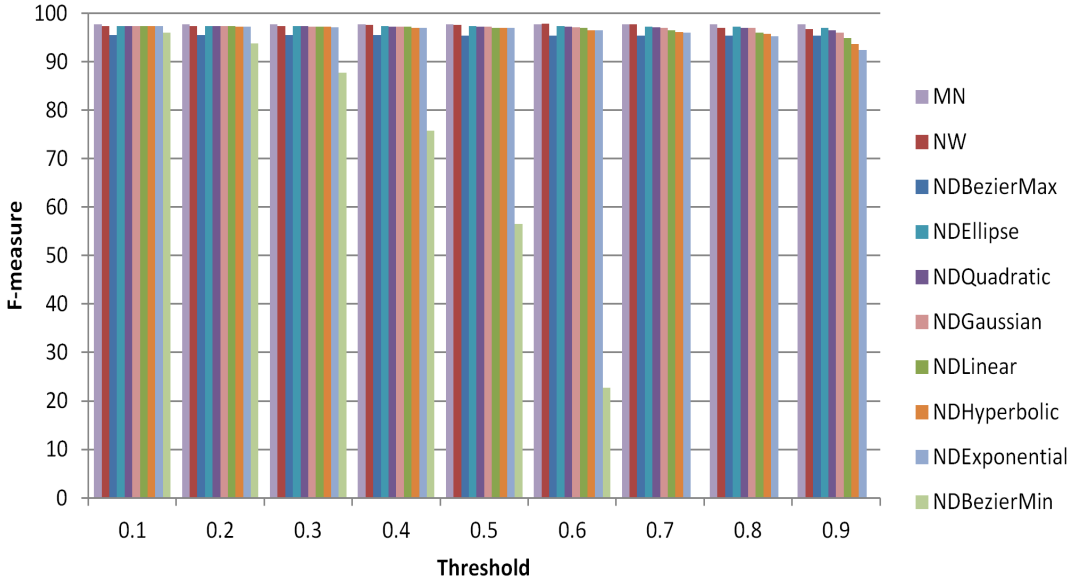
5.5.2 Evaluation and Selection of Spatial Similarity Measures

In our context, geospatial entities have only one spatial attribute that represents the location coordinates of an entity. Section 5.2 described and analyzed eight functions that can be used as Normalized-Distance (ND) to measure the similarity of the spatial attribute. Our objective is to discover whether these NDs produce a realistic quantification of spatial similarity between geospatial entities. In this section, we evaluate the matching using the eight NDs and compare them to two existing spatial similarity measures namely Mutually-Nearest (MN) and Normalized-Weights (NW) (see Section 2.2.3.1 in Chapter 2). MN considers two entities as corresponding only if they are mutually nearest to each other; it does not require a threshold. In contrast, ND and NW quantify the distance and produce a similarity between 0 and 1 to make a decision. NW considers two entities as corresponding only if their spatial similarity exceeds a given threshold, which produces 1:n. Hence, in order to have a fair comparison, we apply the decision algorithm of NW to all NDs. Note that for $ND_{BezierMax}$ and $ND_{BezierMin}$, we calculate the points of the cubic Bezier curve during the pre-matching phase. Then, to calculate the similarity during the matching, we search for the ordinate of the point that has the nearest abscissa to a given distance. This process requires more execution time compared to other NDs.

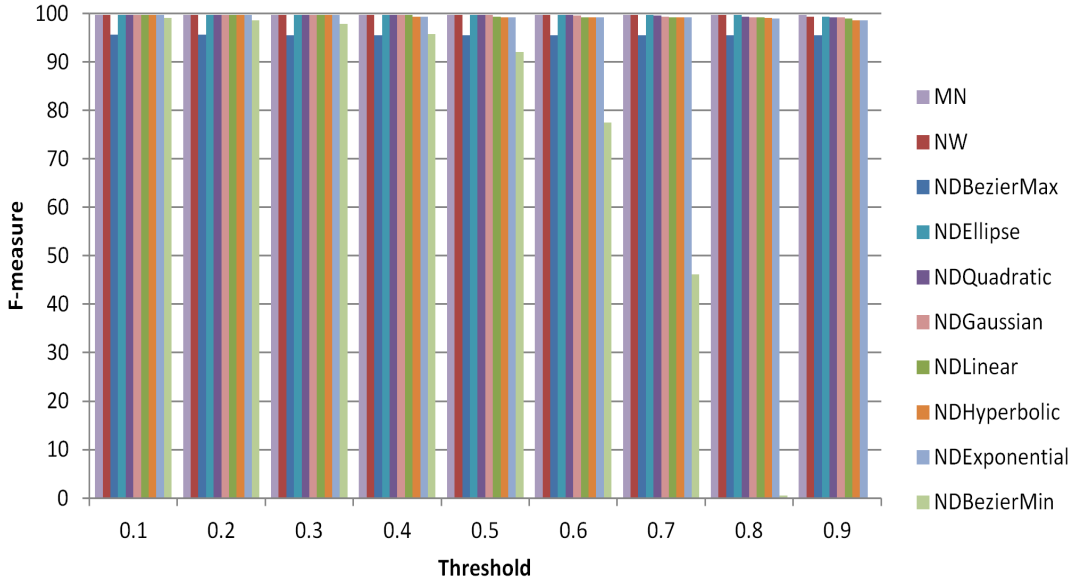
5.5.2.1 Test Cases Evaluation For Spatial Similarity Measures

This section represents a characterized evaluation of spatial similarity measures in order to discover their weaknesses and strengths. These measures use only the location coordinates attribute to detect the corresponding entities. This attribute may have a Different Location (DL) or an Equipollent Positions (EP) differences. Hence, our goal is to evaluate the spatial similarity measures against these two differences separately. To do so, let TestCase(DL) be the combination of all test cases of PABench that have DL regardless of the differences of the primary and secondary terminological attributes (see Section 4.2.2 in Chapter 4). Similarly, let TestCase(EP) be the combination of all test cases of PABench that have EP. Each of these two test cases contains approximately 390 correspondences. The *Farthest Distance* between the source and target datasets of TestCase(DL) equals 4.67 km, while for TestCase(EP) it equals 23.97 km.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés



(a) TestCase(DL)



(b) TestCase(EP)

FIGURE 5.6: *F-measures* of spatial similarity measures namely MN, NW and NDs.

Figures 5.6a and 5.6b shows the *F-measure*s of all methods for TestCase(DL) and TestCase(EP), respectively. The x-axis represents the values of threshold and the y-axis represents the percentage of *F-measure*. Concerning TestCase(DL), $ND_{BezierMax}$ and $ND_{BezierMin}$ are less effective than the other measures. The former achieves a *F-measure* up to 95% and it is unaffected by the threshold. $ND_{BezierMax}$ produces very high similarities even for entities separated by far distance (see Table 5.2) that have a high possibility to not be corresponding. Therefore, selecting such entities as corresponding increases FP and decreases *Precision*, which decreases *F-measure*. The latter produces very low similarities. It achieves a *F-measure* up to 96% for 0.1 threshold, then decreases to 23%

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

for 0.7 threshold. For a threshold greater than 0.7, $ND_{BezierMin}$ is unable to detect any correspondences, which means that the similarity values of all corresponding entities are less than 0.7. Remaining measures are more effective, they are equivalent for a threshold between 0.1 and 0.5 and achieve a *F-measure* up to 97.5%. For a threshold greater than 0.6, *F-measure* slightly decreases for ND_{Linear}, $ND_{Hyperbolic}$ and $ND_{Exponential}$. Similarly for TestCase(EP) in Figure 5.6b, $ND_{BezierMax}$ has a lower *F-measure* than the other NDs, while $ND_{BezierMin}$ sharply decreases with respect to thresholds. Remaining measure are equivalent and achieve a highest *F-measure* up to 99.6% with a soft decrease for ND_{Linear}, $ND_{Hyperbolic}$ and $ND_{Exponential}$ for a threshold greater than 0.5.

Table 5.8 resumes the effectiveness and shows the efficiency of MN, NW and all NDs against TestCase(DL) and TestCase(EP). The columns P, R, F, E and M refers to *Precision*, *Recall*, *F-measure*, execution time in second and amount of memory in MB, respectively. The below table shows the effectiveness of each measure where the highest *F-measure* is achieved. Note that, we consider the average of 10 executions to calculate the execution time and memory usage. All NDs are equivalent in terms of execution time, they consume 3 seconds for TestCase(DL) and 4 seconds for TestCase(EP), except for $ND_{BezierMax}$ and $ND_{BezierMin}$ that consume 7 seconds for each test case. MN and NW consume up to 7-8 seconds to detect the correspondences. But, MN, $ND_{BezierMax}$ and $ND_{BezierMin}$ have the lowest memory usage as of 1.2 MB for each test case, while remaining NDs require 1.5 MB for each. Concerning NW, it requires more memory (4-4.7 MB) than the other measures. This is due to the normalization phase of the probability matrix (see Section 2.2.3.1 in Chapter 2).

	TestCase(DL)					**TestCase(EP)**				
	P	**R**	**F**	**E**	**M**	**P**	**R**	**F**	**E**	**M**
MN	98.4	96.9	97.7	8	1.2	99.7	99.5	99.6	8	1.2
NW	98.9	96.6	97.8	8	4.7	100	99.5	99.7	7	4
$ND_{BezierMax}$	95.6	95.4	95.5	7	1.2	95.6	95.6	95.6	7	1.2
$ND_{Ellipse}$	97.4	97.2	97.3	3	1.5	99.7	99.5	99.6	4	1.5
$ND_{Quadratic}$	97.4	97.2	97.3	3	1.5	99.7	99.5	99.6	4	1.5
$ND_{Gaussian}$	97.4	97.2	97.3	3	1.5	99.7	99.5	99.6	4	1.5
ND_{Linear}	97.4	97.2	97.3	3	1.5	99.7	99.5	99.6	4	1.5
$ND_{Hyperbolic}$	97.4	97.2	97.3	3	1.5	99.7	99.5	99.6	4	1.5
$ND_{Exponential}$	97.4	97.2	97.3	3	1.5	99.7	99.5	99.6	4	1.5
$ND_{BezierMin}$	97.6	94.3	95.9	7	1.2	99.7	98.5	99.1	7	1.2

TABLE 5.8: Characterized evaluation of spatial similarity measures.

According to these results, we conclude that $ND_{BezierMax}$ and $ND_{BezierMin}$ do not produce a realistic quantification of spatial similarities and consume more time than the other NDs. In contrast, remaining NDs namely $ND_{Ellipse}$, $ND_{Quadratic}$, $ND_{Gaussian}$, ND_{Linear},

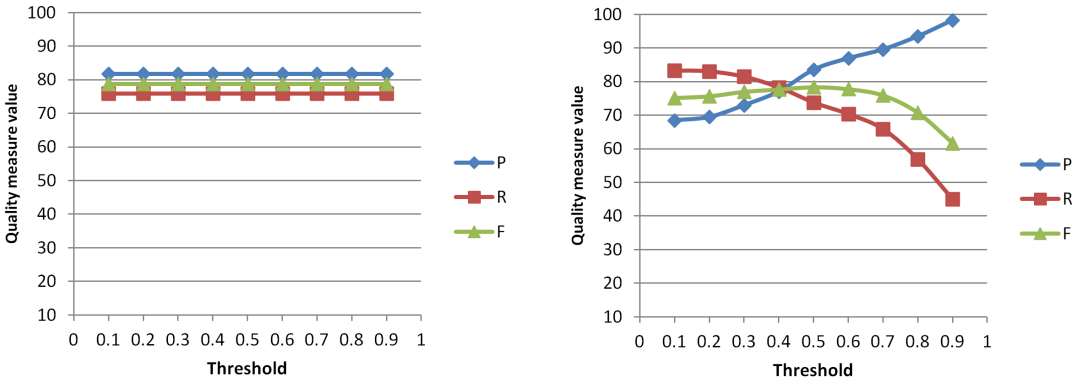
Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

$ND_{Hyperbolic}$ and $ND_{Exponential}$ can detect the correspondences that have spatial differences in equivalent rate to the context-based methods namely MN and NW; these NDs are faster than MN and NW. In the next section, we analyze whether these methods are still able to handle the spatial differences against a standard LBS query that contains noise data.

5.5.2.2 Full Datasets Evaluation For Spatial Similarity Measures

This evaluation analyzes the behaviors of NDs, MN and NW by matching geospatial entities retrieved from several LBS providers based on a standard LBS query. The full datasets of this evaluation contain 378 correspondences, in which 14% have DL, 12% have EP and remaining correspondences are separated by a distance less than 10 meters (i.e., no spatial difference is considered). Recall that these datasets contain singleton entities that may have resemblances with corresponding entities such as Superposition, which affects the results quality.

Figure 5.7 shows the effectiveness of MN and NW and remaining NDs. The x-axis represents the values of threshold and the y-axis represents the percentage of quality measures. The curves P, R and F refer to *Precision*, *Recall* and *F-measure* respectively.



(a) MN (b) NW

FIGURE 5.7: Effectiveness of NW and MN in the full datasets evaluation.

Figure 5.7a shows the effectiveness of MN; it is not affected by the threshold and achieves a 78.7% *F-measure* with 81.7% *Precision* and 75.9% *Recall*. This method has an advantage that it is stable and require less parameterization than the other methods.

Figure 5.7b shows the effectiveness of NW; it achieves a highest *F-measure* up to 78.3% for 0.5 threshold with 83.5% *Precision* and 73.8% *Recall*. The highest *Recall* achieves 83.3% for 0.1 threshold, then it decreases as the threshold increases and achieves 45% for

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

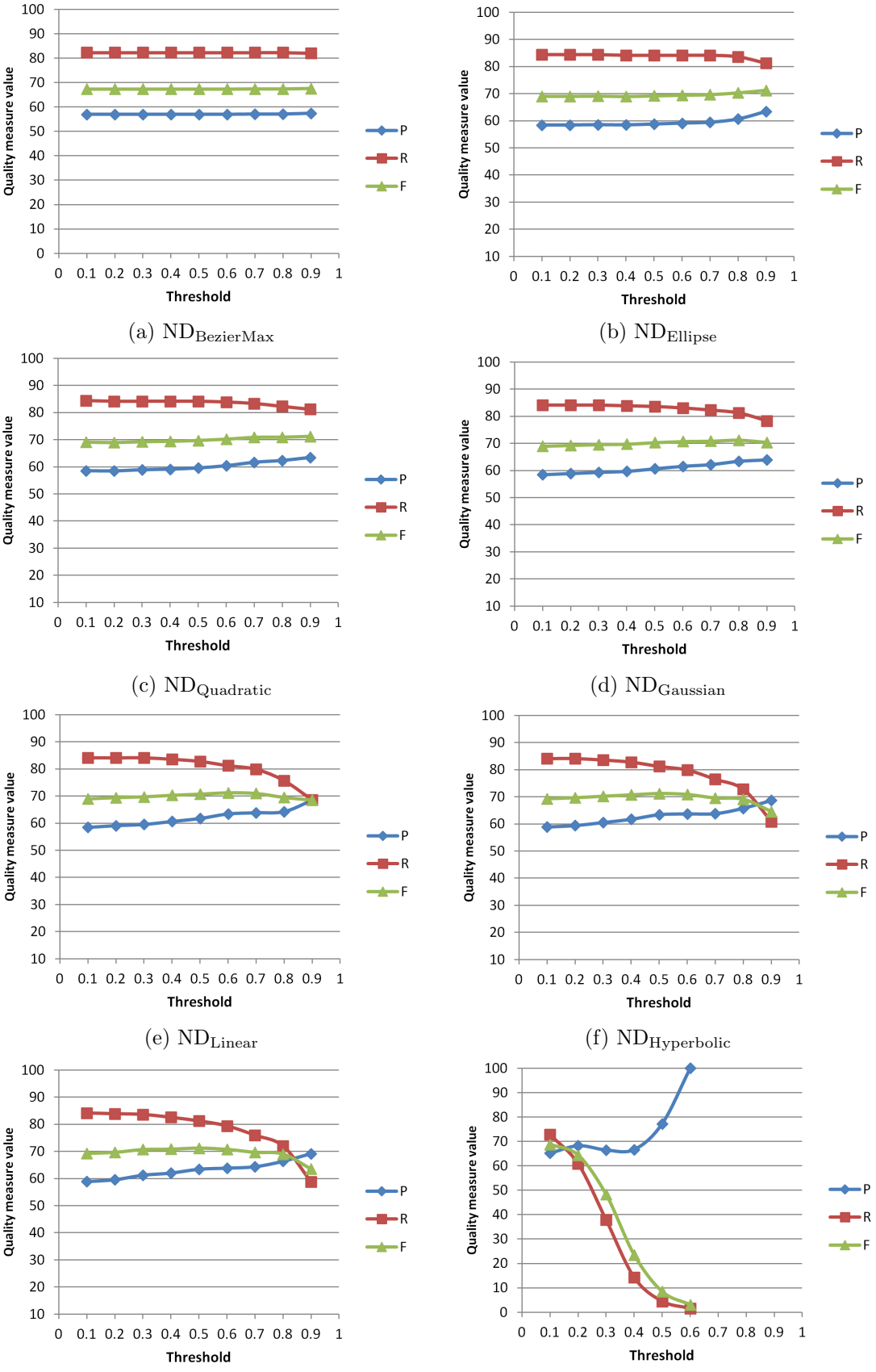
0.9 threshold. The lowest *Precision* achieves 68.5% for 0.1 threshold. Then it increases as the threshold increases until it achieves 98.3% for a 0.9 threshold. This means that as the threshold increases, NW avoids FP more than it loses TP, which increases *Precision*. The reason is that NW produces 1:n correspondences (see Section 2.2.3.1 in Chapter 2), which impacts the results quality as demonstrated in Section 3.5.2 of Chapter 3. Hence, for a low threshold the results include a lot of FP. However, these 1:n detected correspondences are refined as the threshold increases.

Figure 5.8 shows the effectiveness of all NDs. Figures 5.8a and 5.8h refers to $ND_{BezierMax}$ and $ND_{BezierMin}$, respectively. Similarly to the test cases experiment, these two methods are less effective than the others. $ND_{BezierMax}$ is stable regardless of the threshold due to the high similarity values that it produces; it achieves a *F-measure* up to 67.5%. $ND_{BezierMin}$ produces low similarities; it achieves a highest *F-measure* up to 68.7% for a 0.1 threshold then sharply decreases to 0% for a threshold greater than 0.7. A special case concerning *Precision* of $ND_{BezierMin}$; it decreases for 0.3 and 0.4 thresholds, while *Precision*s of other methods increases as the threshold increases. As mentioned earlier, *Precision* equals $\frac{TP}{TP+FP}$ and varies depending on the change ratio of FP and TP. In this case, we lost TP more than we avoid FP, which decreases *Precision*.

Figures 5.8b to 5.8g show the effectiveness of the six remaining NDs. These functions achieve a highest *F-measure* up to 71.2% with 63.4% *Precision* and 81.2% *Recall*, but at different thresholds. This is due to the variation of values produced by distinct functions. For example, $ND_{Ellipse}$ has the slowest decreasing rate among these NDs (see Table 5.2) and it achieves the highest *F-measure* with a 0.9 threshold. In contrast, $ND_{Exponential}$ has the fastest decreasing rate among these NDs and it achieves the highest *F-measure* with a 0.5 threshold. That is, the slower the decreasing of the function is, the higher the threshold of best result becomes. As shown in each figure of ND, the highest *Recall* is always obtained for a 0.1 threshold. For ND_{Linear}, $ND_{Hyperbolic}$ and $ND_{Exponential}$, *Recall* decreases sharply as the threshold becomes greater than 0.5. For $ND_{Quadratic}$ and $ND_{Gaussian}$, *Recall* decreases slightly as the threshold becomes greater than 0.6. Finally, *Recall* is almost stable for $ND_{Ellipse}$ regardless of the threshold. In contrast, *Precision* slightly increases as the threshold becomes greater than 0.6 for $ND_{Ellipse}$, $ND_{Quadratic}$ and $ND_{Gaussian}$, while it increases sharply as the threshold becomes greater than 0.3 for ND_{Linear}, $ND_{Hyperbolic}$ and $ND_{Exponential}$. As shown, these six NDs can achieve the same highest *F-measure*, but $ND_{Ellipse}$ has an advantage because its *Precision*, *Recall* and *F-measure* are more stable than the others.

However, the context-based method MN and NW outperforms all NDs. This may be due to the similarity values of NDs or to the decision algorithm. We repeat this experiment

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés



(a) $ND_{BezierMax}$ (b) $ND_{Ellipse}$

(c) $ND_{Quadratic}$ (d) $ND_{Gaussian}$

(e) ND_{Linear} (f) $ND_{Hyperbolic}$

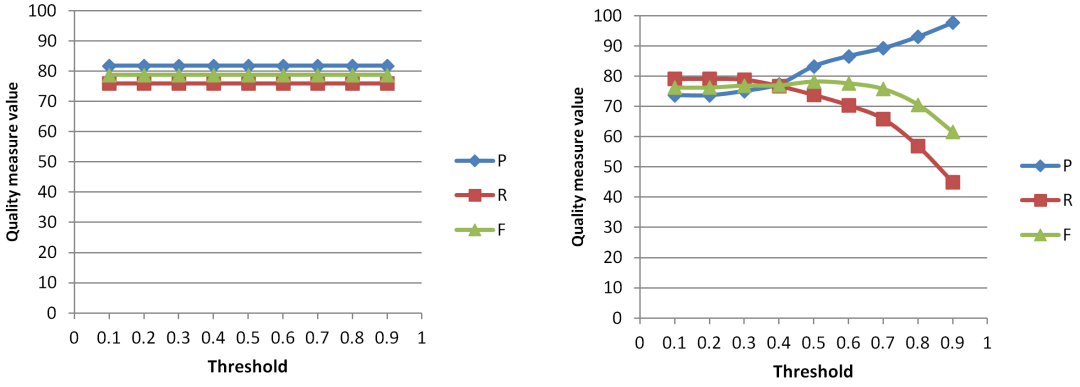
(g) $ND_{Exponential}$ (h) $ND_{BezierMin}$

FIGURE 5.8: Effectiveness of NDs in the full datasets evaluation with 1:n decision algorithm.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

by replacing the 1:n decision algorithm of NW by the 1:1 correspondences decision algorithm described in Section 5.4.

Figure 5.9 shows the effectiveness of MN and NW with the new decision algorithm. The result of MN is not affected by the new decision algorithm as show in Figure 5.9a, because this method does not require a threshold. Concerning NW in Figure 5.9b, compared to the previous experiment (see Figure 5.7b), *Precision* slightly increases and *Recall* slightly decreases for a threshold lesser than 0.4. However, it achieves the same highest *F-measure* up to 78.3% for 0.5 threshold as the previous experiment.



(a) MN (b) NW

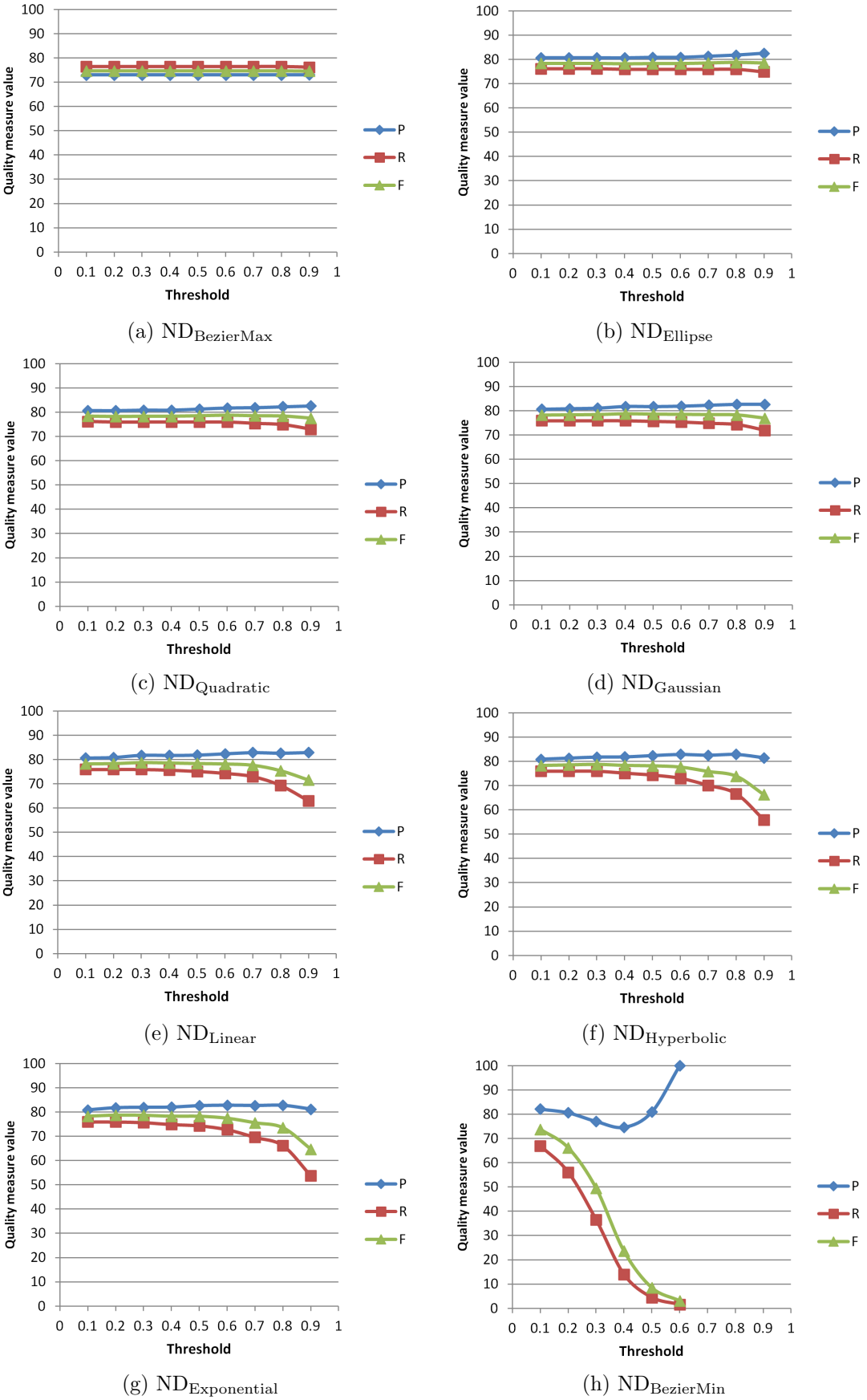
FIGURE 5.9: Effectiveness of NW and MN in the full datasets evaluation.

Figure 5.10 shows the effectiveness of all NDs with the new decision algorithm; a remarkable improvement is achieved. Concerning $ND_{BezierMax}$ in Figure 5.10a, *F-measure* is always stable regardless of the threshold. But, it increases to 74.6% compared to 67.5% in the previous experiment (see Figure 5.8a). Also, Figure 5.10h shows that the *F-measure* of $ND_{BezierMin}$ increases to 73.7% compared to 68.7% in the previous experiment (see Figure 5.8h).

Remaining functions achieve a highest *F-measure* up to 78.7% with 81.8% *Precision* and 76% *Recall*, but at different thresholds. These NDs become equivalent to the context-based measures ND and NW, thanks to the 1:1 correspondences decision algorithm. As shown in each figure of ND, the highest *Recall* is always obtained for a 0.1 threshold. For $ND_{BezierMax}$ and $ND_{Ellipse}$, *Recall* is almost stable regardless of the threshold. For $ND_{Quadratic}$ and $ND_{Gaussian}$, *Recall* decreases slightly as the threshold becomes greater than 0.7. Finally, for ND_{Linear}, $ND_{Hyperbolic}$, $ND_{Exponential}$ and $ND_{BezierMin}$, *Recall* decreases sharply as the threshold becomes greater than 0.4. On the other hand, *Precision* is almost stable for all these NDs, i.e., they avoid FP and lost TP with the same rate as the threshold increases.

As shown, $ND_{BezierMax}$ and $ND_{BezierMin}$ are still unable to outperform the context-based. Concerning remaining NDs, an appropriate decision algorithm allows to refine the result

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés



(a) $ND_{BezierMax}$ (b) $ND_{Ellipse}$

(c) $ND_{Quadratic}$ (d) $ND_{Gaussian}$

(e) ND_{Linear} (f) $ND_{Hyperbolic}$

(g) $ND_{Exponential}$ (h) $ND_{BezierMin}$

FIGURE 5.10: Effectiveness of NDs in the full datasets evaluation with 1:1 decision algorithm.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

and improve the quality, while MN and NW use the neighbor entities to improve their qualities. On these bases, we conclude that these NDs are able to quantify the similarity of spatial attribute.

Table 5.9 shows the efficiency of MN, NW and NDs. $ND_{BezierMax}$ and $ND_{BezierMin}$ consume 34 seconds, while remaining NDs have approximately the same execution time (7-8 seconds). These NDs outperforms MN and NW that consume 13 and 42 seconds respectively. But, MN has the lowest memory usage as of 1.5 MB, followed by $ND_{BezierMax}$ and $ND_{BezierMin}$ that require 3.7 MB, while remaining NDs require 4.7 MB each and NW takes up to 14 MB.

	Execution time (sec)	**Memory usage (MB)**
MN	13	**1.5**
NW	42	14
$ND_{BezierMax}$	34	3.7
$ND_{Ellipse}$	**7.9**	4.7
$ND_{Quadratic}$	**7.5**	4.7
$ND_{Gaussian}$	**7.7**	4.7
ND_{Linear}	**8**	4.7
$ND_{Hyperbolic}$	**8.2**	4.7
$ND_{Exponential}$	**8.1**	4.7
$ND_{BezierMin}$	34	3.7

TABLE 5.9: Efficiency of NDs, MN and NW.

To conclude, similar to the test cases experiment, $ND_{Ellipse}$, $ND_{Quadratic}$, $ND_{Gaussian}$, ND_{Linear}, $ND_{Hyperbolic}$ and $ND_{Exponential}$ are equivalent in terms of execution time and outperforms the others. But MN requires less memory to detect the correspondences. Both MN and NDs outperform NW in terms of efficiency. Concerning the effectiveness, although all measures, except $ND_{BezierMax}$ and $ND_{BezierMin}$, can achieve an equivalent *F-measure*, MN and $ND_{Ellipse}$ are more effective than the others because they are more stable regardless of the threshold, except that MN does not quantify the similarity. Note that for some applications, it is preferred to use a method that produces the highest *Recall* or *Precision*. In all previous experiments, NW always achieves the highest *Recall* for 0.1 threshold and highest *Precision* for 0.9 threshold, compared to other methods. However, in the full datasets experiment, the result's quality has relatively decreased compared to the characterized evaluation (78.7% vs 97.7-99.7%). This is due to the noise entities in the full datasets experiment, these entities may have resemblances that affect the performance of the matching approach. Hence, matching entities of LBS providers using only a spatial similarity measure is not sufficient to resolve the inconsistencies' issue. Such measure needs to be combined with other measures to improve the result's quality.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

As mentioned earlier, for the combination, we need to select the most appropriate similarity measure for each attribute in order to achieve the highest performance by each GS (see Section 5.3). Based on the results of the test cases and full datasets experiments, we need to select the spatial similarity measures that produce the best result for a low (i.e., 0.1), moderate (i.e., 0.5) and high (i.e., 0.9) thresholds. All NDs produce equivalent *F-measure* for 0.1 and 0.5 thresholds, while for 0.9 threshold $ND_{Ellipse}$ is the most appropriate. The next section evaluates the performances of combining several similarity measures based on the proposals of Global Similarity (GS).

5.5.3 Evaluation of Hybrid Approaches Applied to Spatial and Terminological Attributes

The two previous sections evaluated the results of matching using a single attribute. This section analyzes the performance of matching using the similarities of all available attributes. Section 5.3 described five methods, namely GS_{Min}, GS_{Max}, GS_{ExAvg}, GS_{Avg} and GS_{Pr}, that can be used as GS to combine the similarities of the several attributes. As mentioned earlier, GS_{Pr} must be evaluated by combining the measures that produce the best results for low, moderate and high similarities, separately. We denote GS_{PrMin} as the probability combination using the measures that produce the best results with the lowest similarities. Also, GS_{PrMax} and GS_{PrAvg} are defined similarly. Therefore, GS_{Min} and GS_{PrMin} are evaluated with the selected measures for low similarities, GS_{Max} and GS_{PrMax} are evaluated with the selected measures for high similarities and, GS_{ExAvg}, GS_{Avg} and GS_{PrAvg} are evaluated with the selected measures for moderate similarities. For all these methods, we use the 1:1 correspondences decision algorithm defined in Section 5.4.

5.5.3.1 Test Cases Evaluation For Hybrid Approaches

This section analyzes the performances of GS according to the test cases of PABench where 96 test cases are distinguished. Unfortunately, corresponding entities of GeoBench DB do not cover all these test cases because some situations rarely occur; there are 48 test cases having 0 correspondences and 15 test cases having less than 10 correspondences. Hence, this evaluation is limited to 33 test cases that have at least 10 pairs of corresponding entities. Table 5.10 shows the effectiveness of all GSs against each of the 33 test cases. Although experiments are repeated by varying the threshold, but we represent only the highest *F-measure* achieved by each method due to space limitations.

As we can see, all test cases can be handled; for most test cases, a highest *F-measure* up to 100% is achieved, except for TestCase #55 and #95 where the highest *F-measure*

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

# TestCase	Situation	# of corres.	GS_{Min}	GS_{PrMin}	GS_{Max}	GS_{PrMax}	GS_{Med}	GS_{Avg}	GS_{PrAvg}
4	(∅, ∅, MD)	42	**100**	**100**	**100**	94.9	**100**	**100**	**100**
12	(∅, SEMDD, MD)	25	98	98	**100**	**100**	**100**	**100**	98
17	(∅, SYNDD, ∅)	152	98.3	99	99.3	99.7	**100**	**100**	**100**
19	(∅, SYNDD, SYNDD)	92	95.5	**100**	**100**	99.5	**100**	**100**	99.5
20	(∅, SYNDD, MD)	112	96.8	**99.6**	98.7	82.7	**99.6**	**99.6**	**99.6**
23	(∅, SYNDD, (SYNDD, MD))	24	97.9	97.9	**100**	88.4	**100**	**100**	75
27	(∅, (SEMDD, SYNDD), SYNDD)	26	96	**100**	**100**	**100**	**100**	**100**	**100**
28	(∅, (SEMDD, SYNDD), MD)	33	**100**	**100**	**100**	68	**100**	**100**	**100**
31	(∅, (SEMDD, SYNDD), (SYNDD, MD))	12	**100**	**100**	**100**	95.7	**100**	**100**	95.7
35	(DL, ∅, SYNDD)	22	97.7	**100**	**100**	**100**	**100**	**100**	**100**
36	(DL, ∅, MD)	48	**100**	**100**	**100**	94.5	**100**	**100**	**100**
39	(DL, ∅, (SYNDD, MD))	23	**100**	**100**	**100**	**100**	**100**	**100**	**100**
43	(DL, SEMDD, SYNDD)	19	97.3	**100**	**100**	**100**	**100**	**100**	**100**
44	(DL, SEMDD, MD)	38	**100**	**100**	**100**	95.9	**100**	**100**	**100**
49	(DL, SYNDD, ∅)	30	98.3	**100**	**100**	**100**	**100**	**100**	**100**
51	(DL, SYNDD, SYNDD)	41	92.11	98.8	**100**	97.5	**100**	**100**	94.9
52	(DL, SYNDD, MD)	101	97.5	**100**	99.5	75.3	99.6	99.5	**100**
55	(DL, SYNDD, (SYNDD, MD))	49	95.8	97.9	97.9	93.5	**99**	**99**	89.9
59	(DL, (SEMDD, SYNDD), SYNDD)	31	94.9	**100**	**100**	**100**	**100**	**100**	**100**
60	(DL, (SEMDD, SYNDD), MD)	65	99.2	**100**	**100**	68.7	99.2	99.2	99.2
63	(DL, (SEMDD, SYNDD), (SYNDD, MD))	19	88.9	97.3	**100**	97.3	**100**	**100**	84.9
67	(EP, ∅, SYNDD)	25	**100**	**100**	**100**	**100**	**100**	**100**	**100**
68	(EP, ∅, MD)	58	**100**	**100**	**100**	94.6	**100**	**100**	**100**
75	(EP, SEMDD, SYNDD)	27	94.12	**100**	**100**	**100**	**100**	**100**	**100**
76	(EP, SEMDD, MD)	50	99	**100**	**100**	98	**100**	**100**	**100**
79	(EP, SEMDD, (SYNDD, MD))	21	97.6	**100**	**100**	**100**	**100**	**100**	**100**
81	(EP, SYNDD, ∅)	24	**100**	**100**	**100**	**100**	**100**	**100**	**100**
83	(EP, SYNDD, SYNDD)	34	**100**	**100**	**100**	**100**	**100**	**100**	98.5
84	(EP, SYNDD, MD)	135	95.4	98.5	**99.2**	75.6	98.5	98.1	97.7
87	(EP, SYNDD, (SYNDD, MD))	38	97.3	**100**	**100**	93	**100**	**100**	88.2
91	(EP, (SEMDD, SYNDD), SYNDD)	30	96.6	**100**	**100**	**100**	**100**	**100**	**100**
92	(EP, (SEMDD, SYNDD), MD)	61	96.6	98.3	**100**	68.8	**100**	**100**	98.3
95	(EP, (SEMDD, SYNDD), (SYNDD, MD))	21	95	95	97.6	89.5	**99.2**	**99.2**	**99.2**

TABLE 5.10: Effectiveness of GSs in the test cases experiment.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

equals 99% and 99.2%, respectively. In order to compare the methods, for each one, we count the test cases for which a given method achieves the highest *F-measure* relative to other methods. On this base, the worst methods are GS_{Min} and GS_{PrMax} that handle 10 and 13 test cases out of 33, respectively. Followed by GS_{PrAvg}, GS_{PrMin} and GS_{Max} that handle 21, 24 and 28 test cases, respectively. Finally, the most stable methods are GS_{ExAvg} and GS_{Avg} that handle 30 out of 33 test cases. However, these two last methods can still achieve a satisfying high *F-measure* up to 98% and 99% for the three remaining TestCases #52, #60 and #84. Concerning the efficiency, the execution time and the amount of memory for all methods do not exceed 1 second and 1 MB, respectively, because the source and target datasets of most test cases are small. The next section analyzes whether these methods are still able to handle the differences against a standard LBS query that contains noise data.

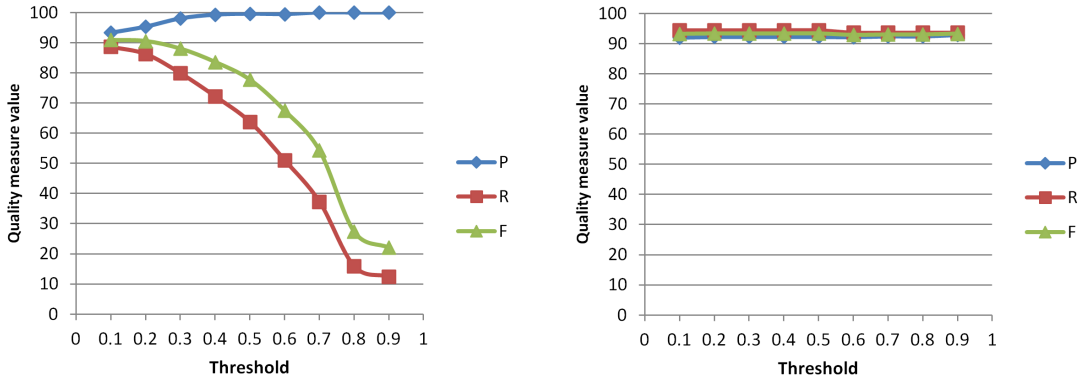
5.5.3.2 Full Datasets Evaluation For Hybrid Approaches

In the previous full datasets evaluations of individual spatial and terminological similarity measures, the highest *F-measure* was 91% by measuring the similarity of the *name* attribute using the Trigram measure. This section analyzes whether the GSs can improve the result's quality and compare their performances to a learning-based approach [SGV06]. Figure 5.11 shows the effectiveness of all seven GSs. The x-axis represents the values of threshold and the y-axis represents the percentage of quality measures. The curves P, R and F refer to Precision, Recall and F-measure respectively.

Figure 5.11a shows the performance of GS_{Min}. Among all GSs, GS_{Min} achieves the lowest *F-measure* of 90.9% for 0.1 threshold, then decreases sharply to achieve 22.1% *F-measure* for 0.9 threshold. GS_{Min} is unable to improve the result's quality; it is equivalent to Trigram applied to the *name* attribute (91%). Note that according to our blocking (i.e., distance and POI types), there are 26522 comparisons between source and target entities. Trigram measure applied to the *name* attribute produces the minimum similarity for 95% of these 26522 comparisons, which explains the equivalence between GS_{Min} and Trigram.

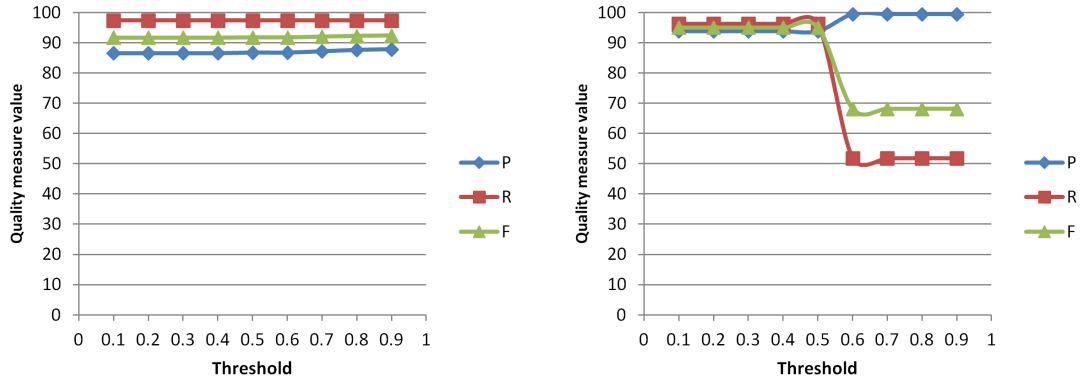
Figures 5.11b and 5.11c show the performances of GS_{PrMin} and GS_{Max}, respectively. Both are stable regardless of the threshold and achieve a *F-measure* up to 93.3% and 92.4%, respectively. These two methods slightly improve the result's quality. Concerning GS_{Max}, the measure $ND_{Ellipse}$ produces the maximum similarity for 87.4% of the 26522 comparisons.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés



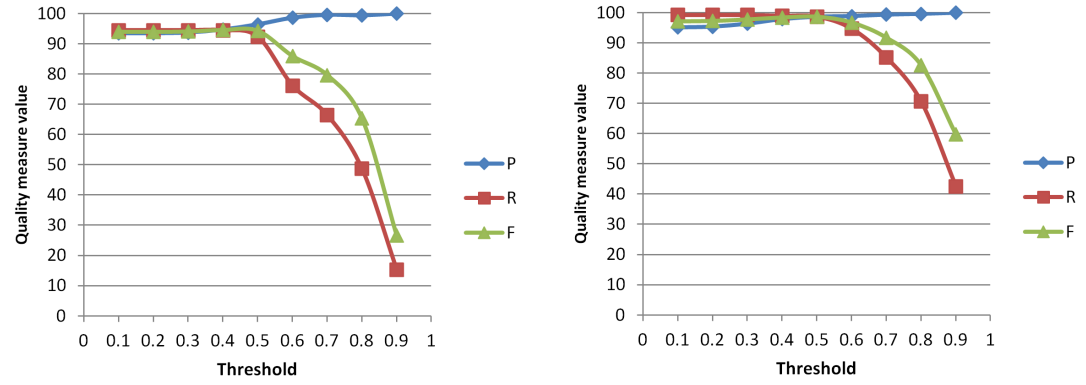
(a) GS_{Min}

(b) GS_{PrMin}



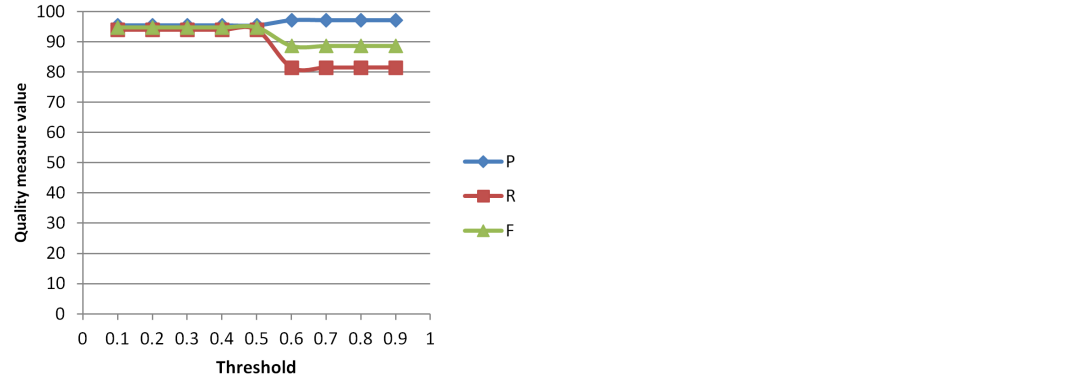
(c) GS_{Max}

(d) GS_{PrMax}



(e) GS_{ExAvg}

(f) GS_{Avg}



(g) GS_{PrAvg}

FIGURE 5.11: Effectiveness of GSs in the full datasets experiment.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

Figure 5.11d shows the performances of GS_{PrMax}. It achieves a *F-measure* up to 95% for a threshold between 0.1 and 0.5, then it decreases to 68.2% for a threshold between 0.6 and 0.9. A noticeable improvement is obtained with this method (95% vs 91%).

Figure 5.11e shows the performances of GS_{ExAvg}. It achieves a *F-measure* up to 94.5% for 0.4 threshold, then it sharply decreases to 26.6% for 0.9 threshold. This method also improve the result's quality (94.5% vs 91%).

Figure 5.11f shows the performance of GS_{Avg}. Among all GSs, GS_{Avg} achieves the highest *F-measure* of 98.7% for 0.5 threshold. *Precision* is almost stable regardless of the threshold and achieves 96%-100%, while *Recall* achieves 99% between 0.1 and 0.5 thresholds, then it decreases to achieve 41% for 0.9 threshold, which decreases *F-measure* to 59%. This method perfectly improves the result's quality (98.7% vs 91%).

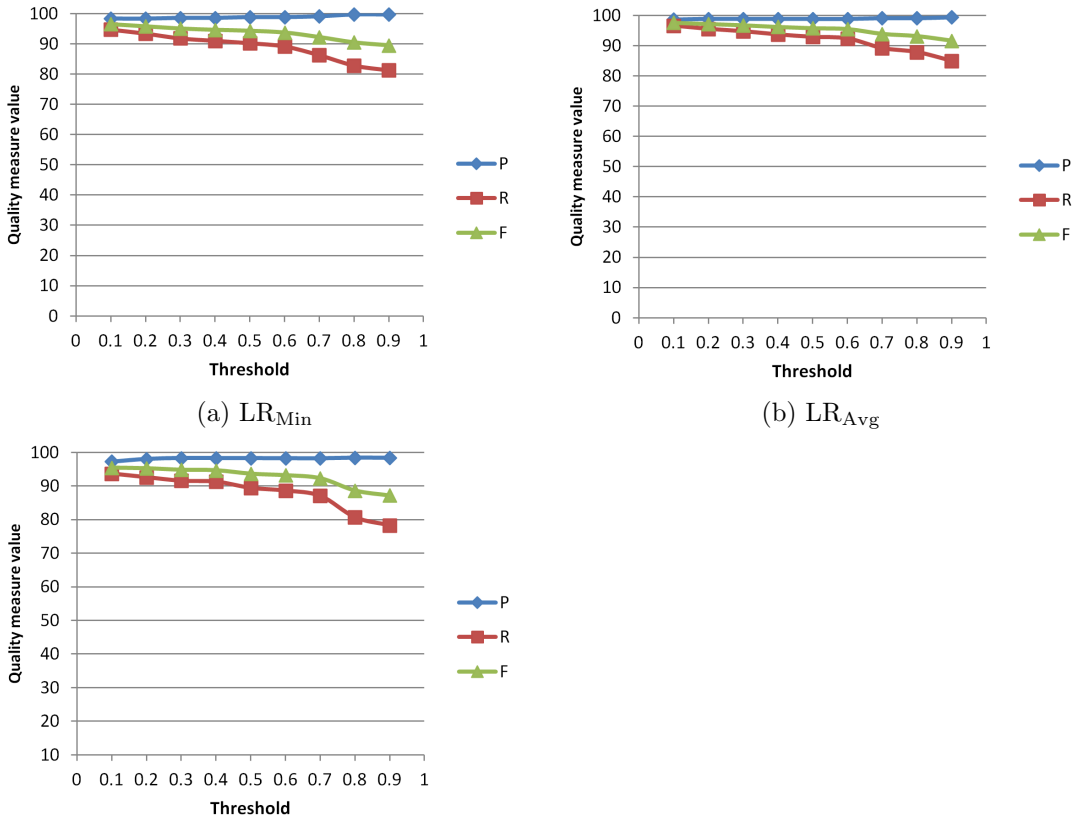
Figure 5.11g shows the performances of GS_{PrAvg}. It achieves a *F-measure* up to 94.7% for a threshold between 0.1 and 0.5, then it decreases to 88.6% for a threshold between 0.6 and 0.9. A noticeable improvement is obtained with this method (94.7% vs 91%).

We compare the performances of GSs to a learning-based approach proposed by Sehgal et al. [SGV06]. This method consists of a weighted average combination; it uses the Logistic Regression (LR) algorithm in order to learn the weight of each similarity measure. Three experiments are done for this approach namely LR_{Min}, LR_{Avg} and LR_{Max}, they use the similarity measures that produce the best result for a low, moderate and high threshold, respectively. Concerning the learning data, we use 50% of the full datasets for learning purposes, and we evaluate the matching using the remaining entities. Note that we apply the same blocking and decision algorithm as GSs in order to guarantee a fair comparison. Figure 5.12 shows the effectiveness of all LRs.

*Precision*s of all LRs are almost stable regardless of the threshold, while the highest *Recall*s are obtained for a 0.1 threshold, then decrease as the threshold increases. LR_{Min}, LR_{Avg} and LR_{Max} achieve a highest *F-measure*s for 0.1 threshold and up to 96.5%, 97.6% and 95.4%, respectively. All these LRs improve the quality of matching compared to individual similarity measures (91%). The combination of similarity measures that produce the best result for moderate similarities always outperforms the combination of those who produce the best result for low and high similarities. However, GS_{Avg} achieves 98.7% *F-measure* and outperforms all LRs. This is due to the small datasets used for learning; using more data may improve the result of LRs.

Table 5.11 resumes the effectiveness and efficiencies of all GSs and LRs. The columns P, R, F, E and M refers to *Precision*, *Recall*, *F-measure*, execution time in seconds and amount of memory in MB for each measures. All methods are equivalent in terms of amount of memory. But GS_{Max} and GS_{PrMax} are the fastest, they consume 26 seconds,

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés



(a) LR_{Min} (b) LR_{Avg}

(c) LR_{Max}

FIGURE 5.12: Effectiveness of LRs in the full datasets experiment.

followed by LR_{Max} that consumes 30 second, then GS_{Min} and GS_{PrMin} that consume 36-37 seconds. GS_{Med}, GS_{Avg} and GS_{PrAvg} consume more time, they require 39-40 seconds. Finally, LR_{Min} and LR_{Avg} are the slowest by consuming 50 seconds. Although the same datasets are used for this evaluation, but the distinct selected similarity measures affect the execution time.

	P	**R**	**F**	**E**	**M**
GS_{Min}	93.3	88.6	90.9	36	5
GS_{PrMin}	92.3	94.4	93.3	37	5
GS_{Max}	87.8	97.4	92.4	**26**	5
GS_{PrMax}	93.8	96.3	95	**26**	5
GS_{ExAvg}	94.7	94.4	94.5	39	5
GS_{Avg}	**98.7**	**98.7**	**98.7**	40	5
GS_{PrAvg}	95.4	93.9	94.7	40	5
LR_{Min}	98.4	94.7	96.5	50	5
LR_{Avg}	**98.7**	96.6	97.6	51	5
LR_{Max}	97.3	93.6	95.4	30	5

TABLE 5.11: Performance of GSs and LRs in the full datasets experiment.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

To conclude, according to the test cases and full datasets experiments, GS_{Avg}, that computes the average of individual similarities, seems the most resistant to the differences of test cases and improves the result's quality of the full datasets evaluation to achieve 98.7% *F-measure*. Recall that the global similarity of GS_{Avg} is computed by applying $ND_{Ellipse}$, Trigram, TFIDF Cosine, TFIDF Cosine and Levenshtein measures to *location, name, address, phone* and *website* attributes, respectively.

5.6 Conclusion

This chapter deals with the matching of geospatial entities that come from several LBS providers. In a first stage, we proposed a generalization for spatial similarity measure, namely Normalized-Distance (ND), that it is not context depending. Eight distinct function are described, namely cubic Bezier as logarithmic scale, Ellipse, Quadratic, Gaussian, Linear, Hyperbolic, Exponential and cubic Bezier as exponential scale, that can fit the requirements of ND. Concerning the terminological attributes, we collected five known string similarity measures, namely Equality, Including, Levenshtein, Trigram and TFIDF Cosine, that are used in several geospatial entity matching approaches. In addition, we proposed to combine the individual similarities of several attributes in order to compute a Global Similarity (GS) for two entities. Several propositions have been discussed such as pessimistic, optimistic, extreme average, average and probability, that can be used for GS. Finally, a decision algorithm that produces 1:1 correspondences is proposed to reduce the impact of inconsistencies. For all these methods, we presented a test cases experiments by matching the characterized datasets of PABEnch and full datasets experiments by matching two datasets collected using a standard LBS query.

Concerning the string similarity measure, they have been applied to each terminological attributes separately. The result of the test cases experiment is not satisfying; *F-measure* varies between 80% and 96% depending on the attributes. Also, the result of the full datasets experiment is not satisfying; the highest achieved *F-measure* is 91% by measuring the similarity of the *name* attribute using Trigram measure.

Concerning the spatial similarity measures, context-based methods MN and NW, and all NDs, except $ND_{BezierMax}$ and $ND_{BezierMin}$, can resolve the spatial differences, namely Different Location and Equipolent Positions, that appear between corresponding entities. Concerning the full datasets experiment, we proved that using 1:1 correspondences decision algorithm improve the result's quality, which makes NDs equivalent to the context-based spatial similarity measures in terms of effectiveness, except for $ND_{BezierMax}$ and $ND_{BezierMin}$. But, these NDs are more efficient than MN and NW. Among all NDs, $ND_{Ellipse}$ is the most effective because it is the most stable regardless of the threshold.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

However, the results of spatial similarity measures in the full datasets evaluation are not sufficient (78% *F-measure*).

According to the evaluation of spatial and string similarity measures, we selected the most appropriate measure for each attribute to fit each method of GS. The combination of similarities of several attributes using the several propositions of GS improve the results quality, except for the pessimistic combination. The highest achieved *F-measure* is 98.7% using an average combination. This combination is computed by applying $ND_{Ellipse}$, Trigram, TFIDF Cosine, TFIDF Cosine and Levenshtein measures to *location, name, address, phone* and *website* attributes, respectively. This method is used in the next chapter in order to create a prototype for online integration of LBS providers, in which we consider the visualization and the uncertainty of integrated entities.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

Chapter 6

Visualization of Geospatial Integration

Contents

6.1 Data Fusion and Uncertainty 128
6.1.1 Merging Corresponding Entities 129
6.1.2 Uncertainty Level Computation 131
6.2 Uncertainty Visualization: Proposals and Assessment . . . 133
6.2.1 Selection of Visual Variables to Portray Uncertainties 135
6.2.2 Selection of Uncertainties Information to Portray on Map or on Demand 138
6.3 Impact of Visualizing Uncertainty in LBS: A Use Case for Tourists 143
6.3.1 Experimental Protocol and Simulator 143
6.3.2 Results and Analysis 147
6.4 Multi-providers LBS Prototype 150
6.4.1 Overview and Architecture 150
6.4.2 Interfaces and Navigation 152
6.5 Conclusion 153

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

As mentioned previously, data integration consists in schema matching, entity matching and data fusion. The two previous chapters (see Chapters 4 and 5) deal with schema and entity matching. In this chapter, we will focus on the representation of resulting integrated data.

In the first step, a basic data fusion algorithm is proposed in order to merge corresponding entities detected during the entity matching phase. But, the merging outcome may not be 100% reliable. A merged entity may therefore include some uncertainties for its spatial and terminological attributes. We decide to compute these uncertainties in order to transmit the quality of merged entities to users. These uncertainties do not express the correctness of integrated entities compared to reality because we do not have access to real data. Instead, they express the completeness and consistency of information offered by distinct LBS providers.

The second step concerns the delivery of geospatial integrated entities to end-users. We mainly focus on the visualization of integrated entities including their uncertainties. We start by discussing the specifications of representing POIs. Then, we consider the visualization of uncertainty information; two experiments based on cognitive tests were conducted in collaboration with the partners of UNIMAP[1] such as EVS laboratory[2], Only Lyon tourist office[3], Saint-Etienne tourist office[4] and Rhône-Alpes tourisme[5] who develops the Tourist Information System in Rhône-Alpes (SITRA) that is also known as Apidae. These experiments evaluate several proposals to find the best appropriate way to represent the uncertainty of integrated geospatial data [BCD+14, BDC+14, SCC+14].

We also evaluate the impact of such uncertainty on users' decisions to select POIs for tourism purposes [CFC+16]. This experiment consists in checking whether uncertainty information is taken into account by users and in observing how it has been used.

Finally, we propose a multi-providers LBS prototype. It includes our solution for matching, merging and visualizing geospatial data.

6.1 Data Fusion and Uncertainty

Chapter 5 discusses the matching process in order to detect corresponding geospatial entities coming from different LBS providers. This section proceeds with merging of corresponding entities into new integrated entities. For this purpose, data fusion algorithms

[1] Partners of UNIMAP project: http://liris.cnrs.fr/unimap/participants.html
[2] EVS laboratory: http://umr5600.ish-lyon.cnrs.fr
[3] Only Lyon tourism office: http://www.lyon-france.com
[4] Saint-Etienne Tourisme: http://saint-etiennetourisme.com
[5] Rhône-Alpes tourisme: http://www.apidae-tourisme.com

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

serve to resolve the conflicted issue between corresponding entities by identifying the true values for attributes amongst multiple observed values [DGH+14]. Ideally, an integrated entity should include more complete and accurate information. But, corresponding entities may have different locations and the data fusion algorithm may still produce an incorrect position. Similarly, the names or the phone numbers of corresponding entities may differ, and the choice of the correct values relies on the data fusion algorithm. Thus, we propose a basic data fusion algorithm that consists of the following: first, in merging corresponding entities by choosing for each attribute one value among all available, and second, in computing an uncertainty score for the newly integrated entity.

6.1.1 Merging Corresponding Entities

A data fusion algorithm should be able to handle multiple observed values coming from different sources. Chapter 5 proposes a 2-join approach for matching two entity sets, but in reality we intend to match more than two entity sets. Several existing methods such as the serial and hierarchical joins [SKS+10, KKH+10], allow the matching of $N \geq 2$ datasets based on 2-join algorithms. Thus, after matching all available entity sets together, our goal is to merge the detected corresponding entities.

Merging entities is a common task in applications such as crisis management, data-warehousing or mash-up creation. Basic approaches have been proposed for data fusion such as considering values from the most recent up-to-date source or taking the average, maximum or minimum for numerical values [BN08, DN09]. Several advanced data fusion algorithms are proposed based on the voting strategy (see Section 2.2.4 in Chapter 2). In our context, the values of attributes may be inconsistent between each other, which imposes an issue for voting. For instance, the two following values "Hotel Leyal" and "Hotel Leyla" have Syntactic Different Data (SYNDD) and may be considered as two different values and separately receive one vote, while in reality, there is one single value that should get two votes. To resolve this issue, we adapted the voting strategy by proposing a basic data fusion algorithm that uses the similarity measures [DGH+14]. For a given attribute that has different values proposed by several providers, our intuition is to choose the value which is the most similar to the others.

Algorithm 1 shows our process of merging for this selection. It takes as input (i) a set of n corresponding entities ($n \geq 2$), namely $CORR$, resulting from the entity matching phase and (ii) a list of corresponding attributes, namely ATT, resulting from the schema matching phase. This algorithm returns a new integrated entity for which the values of attributes are selected amongst the values of the corresponding entities. To do so, each attribute has several potential values, for each of these values we associate a *voting*

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

Algorithm 1 Data fusion algorithm to merge a set of n corresponding entity

Input: A set of n corresponding entities $CORR$
Input: A list of corresponding attributes ATT
Output: An integrated entity $integrated_entity$

```
1: function merging()
2: for all att ∈ ATT do
3:    max ← 0
4:    for all e ∈ CORR do
5:       votingResult ← 0
6:       CORR' ← CORR − {e}
7:       for all e' ∈ CORR' do
8:          votingResult ← votingResult + Similarity(e.att.val, e'.att.val)
9:       end for
10:      if votingResult > max then
11:         max ← votingResult
12:         integrated_entity.att.val ← e.att.val
13:      end if
14:   end for
15: end for
16: integrated_entity.CORR ← CORR
17:
18: return integrated_entity
19: end function
```

result that it equals the sum of similarities of the given value with each of the remaining values. Therefore, the sum of similarities replaces the addition of votes for values. For instance, consider the three corresponding entities $e_1 \in E_1, e_2 \in E_2$ and $e_3 \in E_3$ and one corresponding attribute $att \in ATT$, the *voting result* of $e_1.att.val$ equals the similarity between $e_1.att.val$ and $e_2.att.val$ plus the similarity between $e_1.att.val$ and $e_3.att.val$. Note that for each attribute we use the appropriate similarity measure according to the result of Chapter 5. Finally, we select the value that has the highest score. In addition, we save the corresponding source entities $CORR$ of each integrated entity (line 16 of Algorithm 1); these source entities will be used later for visualization purposes. This merging process is repeated for each set of corresponding entities in order to create a set of integrated entities.

As an example, consider the following three values for the *name* attribute: *Eiffel Tower*, *Tour Eiffel* and *Eiffel*. Using Trigram measure, the similarity between *Eiffel Tower* and *Tour Eiffel* equals 0.63, the similarity between *Eiffel Tower* and *Eiffel* is 0.54, while it equals 0.6 for *Tour Eiffel* and *Eiffel*. The *voting result* of the value *Eiffel Tower* is therefore equal to 1.17 (0.63 + 0.54), the *voting result* of *Tour Eiffel* equals 1.23 and the one for *Eiffel* is 1.14. Hence, the chosen value for the *name* attribute is the one with the highest *voting result*, i.e., *Tour Eiffel*.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

One distinguished case concerns two values having the same *voting result* or an attribute that has only two potential values to be merged, i.e., there are only two corresponding entities or the other correspondences that have Missing Data (MD) for this attribute. Intuitively, we must select the value offered by the provider that has a better quality. Currently, our decision algorithm chooses one value randomly since we do not have any information about the quality of providers. Several works have been proposed to estimate the quality of geospatial data [TKA07, Gup08]. Our data fusion algorithm can also serve for this purpose, which improves the selection in the case of two values. For each provider, we count the number of times its values are chosen as the most similar to others when there are three or more potential values. After a long term, the providers' quality can be ranked according to these counts.

The result of data fusion strongly relies on the result of entity matching. In other words, a mistake made by the entity matching process, such as considering two different entities as corresponding, will certainly affect the data fusion algorithm and will result in incorrect integrated entities. For this reason, we make a decision to compute and inform the users about the uncertainty of integrated entities. The next section discusses such uncertainty.

6.1.2 Uncertainty Level Computation

Because the result of the merging process may not be 100% reliable, we propose to inform users whether the values of integrated entities are trustworthy. To do so, we decided to monitor both spatial and terminological uncertainties, that refers to the spatial and terminological information, respectively, as well as we consider a global uncertainty that groups these two uncertainties.

These uncertainties express the completeness and consistency of information offered by distinct LBS providers. In other words, if LBS providers offer the same value for a given attribute, then this value is chosen for the integrated entity and it is consider as a certain information. Otherwise, if providers offered inconsistent or missing values for the same attribute, then our merging process chooses only one value and we consider it as uncertain information. On these bases, given an integrated entity, we assign an *uncertainty score* between 0 and 1 for the spatial, terminological and global information, separately. These scores are deduced from the similarities between the values of source entities. A low *uncertainty score* means that the information of corresponding entities are consistent and complete between each others. On the contrary, a high *uncertainty score* means that the information of corresponding entities are inconsistent or missing. Consider an integrated entity e created by merging n corresponding entities

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

(i.e., $e.CORR = \{e_1, ..., e_n\}$). For each of the three uncertainties (i.e., spatial, terminological and global), we first average the similarities of its related attributes of all possible source entity pairs. The *uncertainty score* is therefore equal to the complimentary of its calculated average to one. Note that for each attribute, we use the appropriate similarity measure obtained in Chapter 5.

For the spatial *uncertainty score*, we average the spatial similarities of all entity pairs. Recall that there is only one primary spatial attribute for the location coordinates. Let *spatial_uncertainty* be the attribute containing the score of spatial uncertainty:

$$e.spatial_uncertainty = 1 - \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \mathrm{ND}_{\mathrm{Ellipse}}(d(e_i, e_j))}{C_n^2}$$

For the terminological *uncertainty score*, we average the terminological similarities of all terminological attributes of all entity pairs, including the attributes that have Missing Data (MD) (i.e., their similarity score equals 0). The interest of computing the MD is to consider the completeness of information. During the matching experiments in Chapter 5, we considered four common terminological attributes namely $name, address, phone$ and $website$. Let *terminological_uncertainty* be the attribute containing the score of terminological uncertainty:

$$e.terminological_uncertainty = 1 - \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \Big[\begin{array}{r}\mathrm{Trigram}(e_i.name, e_j.name) + \\ \text{TFIDF Cosine}(e_i.address, e_j.address) + \\ \text{TFIDF Cosine}(e_i.phone, e_j.phone) + \\ \mathrm{Levenshtein}(e_i.website, e_j.website)\Big]}{4C_n^2}$$

For the global *uncertainty score*, we average the global similarities of all entity pairs. Let *global_uncertainty* be the attribute containing the score of global uncertainty:

$$e.global_uncertainty = 1 - \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \mathrm{GS}_{\mathrm{Avg}}(e_i, e_j)}{C_n^2}$$

For example, consider three corresponding entities e_1, e_2 and e_3 where $\mathrm{GS}_{\mathrm{Avg}}(e_1, e_2) = 0.5$, $\mathrm{GS}_{\mathrm{Avg}}(e_1, e_3) = 0.8$ and $\mathrm{GS}_{\mathrm{Avg}}(e_2, e_3) = 0.9$. In this example, the average of the global similarities equals $\frac{0.5+0.8+0.9}{3} = 0.74$ and the global *uncertainty score* equals 0.26 $(1 - 0.74)$. Spatial and terminological uncertainties' score are calculated similarly. This process quantifies the spatial, terminological and global uncertainties.

To facilitate the understanding of these uncertainties, we convert the spatial, terminological and global uncertainty scores into three levels: certain, moderately certain,

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

uncertain. The first range [0, 0.33] is associated to the certain level, the middle range]0.33, 0.66] includes the moderately certain values, and the uncertain level stands for highest uncertainty values in the range]0.66, 1]. These ranges have been fixed arbitrarily for this experiment, but it is possible to adjust them using statistical methods [MU49] to better express the uncertainties they represent.

Finally, at the end of this step, a set of integrated entities is produced. Each integrated entity is associated to three uncertainty scores for spatial, terminological and global information, as well as the source entities of LBS providers. The next section considers the visualization of such uncertainties' levels to users in the context of LBS.

6.2 Uncertainty Visualization: Proposals and Assessment

The goal of this section is to propose an appropriate solution to deliver geospatial integrated entities for end-users [BCD+14, BDC+14, SCC+14]. The inconsistency between corresponding entities of several providers includes the icons used to visualize the POIs on maps. Traditionally, symbols of icons are designed to represent POIs types (e.g., park, lake and mall), so users can easily distinguish places on the map. Several existing works propose a standardization of signs to represent tourist information on a map. Five decades ago, Joly et de Brommer realized a project for a standardization of maps symbols [JdB64]. Also, the World Tourism Organization[6] (WOT) has defined a standardization for tourist signs including POIs [Org01]. Despite these standardization efforts, each LBS provider uses its own set of icons to represent POIs. Figure 6.1 shows three different legends to represent hotels collected from Share Icon[7], Icon Archive[8] and Icons DB[9], respectively. In addition, some providers use distinct colors to categorize the POIs. For instance, Figure 6.2 shows the legend of POIs' icons used by Mapquest provider, the POIs of type *Food*, *Bars* and *Coffee* are colored in orange, *Gas* and *Parkings* are colored in blue, while *Hotels* are colored in red. Some works have already discussed the integration of icons. Karam proposes a framework namely MPLoM for cartographic integration in order to integrate symbols [Kar11]. In this thesis, we do not consider the integration of icons. Instead, we use the symbols proposed by WOT for the icons of different types of POI.



FIGURE 6.1: Three different legends to represent an hotel on maps.

[6] World Tourism Organization: http://www2.unwto.org
[7] Share Icon hotel legend: https://www.shareicon.net/hotel-accomodation
[8] Icon Archive hotel legend: http://www.iconarchive.com/show/ios7-icons-by-icons8/Hotel.html
[9] Icons DB hotel legend: http://www.iconsdb.com/royal-azure-blue-icons/hotel-icon.html

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

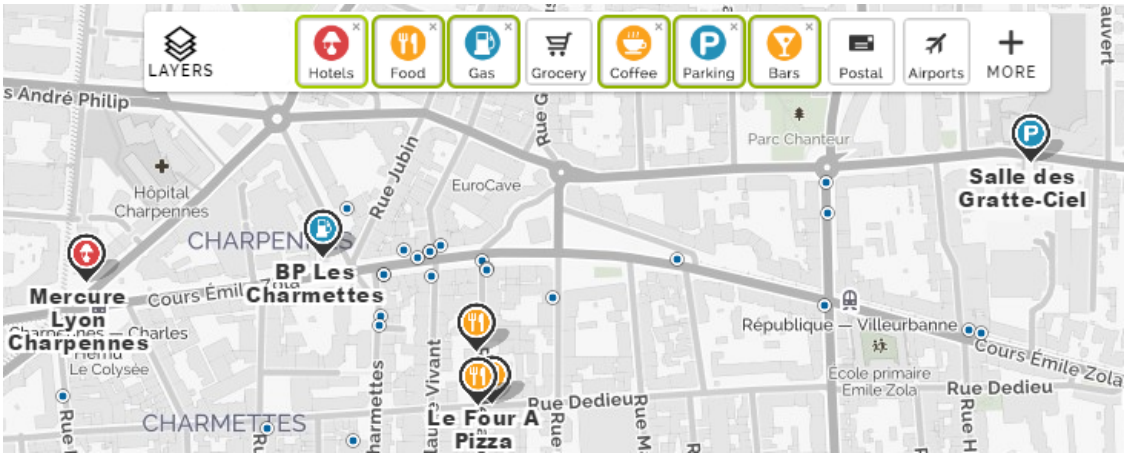


FIGURE 6.2: Legend of Mapquest provider: uses color to categorize the POIs.

As previously discussed, the integrated entities may not be 100% reliable and include three kinds of uncertainty for the spatial, terminological and global information. Therefore, users may need to check source information when observing strange outcome from the merging process.

We distinguish between two profiles of users namely "optimizers" and "maximizers" [SWM+02]. The optimizers choose quick optimal solutions when confronted with choices. In other word, they trust the integration result and consider only the entities that have a certain level. In contrary, the maximizers would check out every piece of information available, every property of every source provider which could generate the uncertainty. For instance, a user may have prior knowledge about a POI and the integrated entity shows incorrect information. A strange outcome may also be observed due to common rules. For example, an integrated entity that refers to a POI in Paris city is considered strange if the first two digits of its phone number begins by "04" that is the prefix of phone numbers in another region. Thus, our approach consists of visualizing (i) the integrated entities, (ii) the source information of each integrated entity to help the maximizers validate the result of integration and (iii) the levels of uncertainties to help the optimizers filtering the entities.

These requirements generate a large amount of information for each POI that might become an issue to entirely visualize. To meet these requirements, our approach is to use the *Details on Demand* technique proposed by Shneiderman [Shn96]. This technique provides more details on the data only after the user requests them. This gives the user a better overview on the data because of the reduced amount of information, but still enables her or him to grasp the "details on demand". Current LBS providers use interactive maps tools to allow users searching and visualizing POIs. These tools consist of a background map made of raster images or vector objects [Mac04]. Then, icons are placed on the map to show POIs locations. A click on an icon displays a panel

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

containing the details such as terminological information. The integrated terminological information and source information can be visualized for users on demand.

To create interactive maps that offer integrated entities, the following questions must be answered:

- How should spatial, terminological or global uncertainties be represented? What are the most appropriate visual variables for each kind of uncertainty?
- Which information visualization scenario should be given priority? In other words, which uncertainty should be represented directly on the map or by demand?
- Is the representation of such information accepted and useful for end-users?

In collaboration with EVS partner, we conducted several tests to answer these questions.

6.2.1 Selection of Visual Variables to Portray Uncertainties

To represent the uncertainties on interactive maps, we first need to find the appropriate visual variables for these uncertainties. Thomson et al. distinguishes three kinds of uncertainty for geospatial data [THM+05]:

1. Spatial uncertainty: refers to the truthfulness of spatial attributes.
2. Terminological uncertainty: refers to the truthfulness of terminological attributes.
3. Time uncertainty: refers to the truthfulness of temporal attributes; it is specific for spatio-temporal objects that are not considered in this thesis.

Authors pair each of these uncertainties with nine different categories namely: Accuracy, Precision, Completeness, Consistency, Lineage, Timing, Credibility, Subjectivity and Interrelatedness. So far, both completeness and consistency of information suit the definition of uncertainty in our study. The other categories cannot be considered because:

1. The Accuracy concerns the difference between observation and reality, but the information about reality is not available.
2. The Precision concerns the exactness of measurement; but we do not have any information about the quality of devices used by providers to collect data such as a GPS device.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

3. The Lineage concerns the conduit through which info passed, but we do not know how providers treat the information during the construction of their databases. Recall that there are several strategies to construct and update POI databases (see Section 2.1 in Chapter 2).

4. The Timing concerns the temporal gaps between occurrence, info collection and use; but not all LBS providers offer the information of when they collected or updated their data.

5. The Credibility concerns reliability of info source, but we assume that all providers offer the same good quality of data.

6. The Subjectivity concerns the amount of interpretation or judgment included; but such information are not offered but providers.

7. The Interrelatedness concerns the source independence from other information; the information about providers' relations and using of external sources is not always available.

MacEachren et al. made an empirical study to characterize the kind of visual variables that is appropriate for representing (i) those different categories for each kind of uncertainty and (ii) the uncertainty in an abstract context [MRH+05, MRO+12]. According to authors (see Section 2.4 in Chapter 2), *Location*, *Size* and *Fuzziness* visual variables are the most appropriate to portray spatial uncertainty. Whilst, *Smiley*, *Filled bar associated with Slider* and *Thermometer* are interesting to portray terminological uncertainty. Finally, *Fuzziness*, *Location* and *Color value* (i.e., color hue) are well suited to portray an abstract uncertainty. Note that each of these visual variables has three icons to represent three levels of uncertainty: certain, moderate and uncertain. An analysis of the results obtained by MacEachren et al. leads us to select the most appropriate visual variables useful in our context [Sec14], which are then evaluated using cognitive tests.

6.2.1.1 Experiment Set-up

Figure 6.3 illustrates a pre-selection of visual variables to represent the spatial, terminological and global uncertainties. Concerning the spatial uncertainty, we decided to compare *Location* with *Size associated to Fuzziness*. We chose *Location* symbol because it is intuitively implicated with space. We aggregated *Size* and *Fuzziness*; the taller the sign is, the fuzzier the sign is. We did this combination because independently, an order would be created between the signs with large or distinct signs seen before the others. Concerning the terminological uncertainty, we investigated the proposals of [MRO+12].

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

For our application, *Smiley* symbol is too connected with a score relative to the quality of a POI obtained from the opinions of different users. Then if the smiley is happy, it will be interpreted as a good POI for the public (e.g., a "good" restaurant) and this is not what we want to represent. Concerning *Filled bar associated with Slider*, we think it is too difficult to correctly perceive the differences between its three levels because only one small element of the slider is modified. For the previous reasons, only the *Thermometer* remained for our study, and it is compared to a new visual variable: *Frequency*, that we proposed based on graphic representations created to show uncertain chaotic behaviors of signals in Electronics Science. Finally, for the global uncertainty (i.e., spatial and terminological information together) we selected the visual variables of the abstract uncertainty. *Fuzziness* and *Color value* were selected, while the *Location* visual variable was eliminated because it was too closely related to the spatial uncertainty.

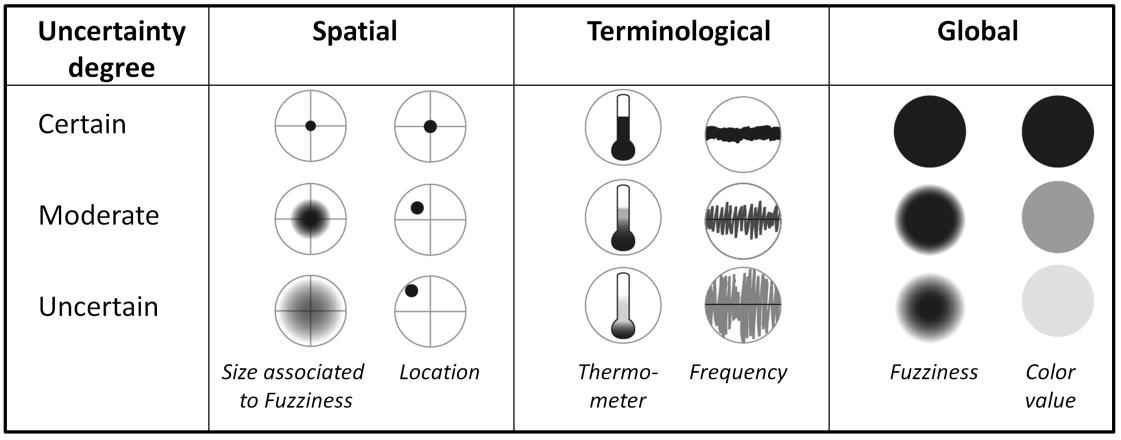


FIGURE 6.3: Visual variables that may fit the uncertainties of integrated entities.

6.2.1.2 Assessment and Result

An intern, who has followed a research master in cognitive science, conducted a perceptual test to determine which semiotic solution is best perceived and understood for each uncertainty [Sec14]. This test is based on the statistical Student t-test [Stu08] with a threshold of acceptability of risk up to 5% (i.e., if the significance level p is lower than 5% then the results are significant). For each visual variable (*Size associated to Fuzziness*, *Location*, *Thermometer*, *Frequency*, *Fuzziness* and *Color value*), a couple of icons (certain vs. moderate, certain vs. uncertain, moderate vs. uncertain) are presented on a map to 36 non cartography expert participants including 14 men and 22 women, aged from 18 to 30 years old. Participants were required to indicate as quickly as possible which icon represents the certain level. We are interested in response time (ms) spent by participants to choose an icon as well as the correctness of their answers (True or False). These experiments were repeated for each uncertainty with a counterbalance in

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

the order of proposals to avoid learning effect. These experiments enabled us to evaluate the comprehension level of variables' icons between each other. Table 6.1 shows the results of Student t-test for both response time and correctness.

	Spatial		**Terminological**		**Global**	
	Size associated to Fuzziness	*Location*	*Thermometer*	*Frequency*	*Fuzziness*	*Color value*
Correctness (%)	**98.1**	90.7	**99.5**	93.5	95.8	**100**
t-Student	2.6		3.65		-2.71	
Threshold of risk p (%)	1		< 1		1	
Time (ms)	**783**	1147	**741**	893	862	**708**
t-Student	-7.91		-3.9		2.24	
Threshold of risk p (%)	< 1		< 1		3.1	

TABLE 6.1: Results of Student t-test to evaluate the comprehension level of visual variables for spatial, terminological and global uncertainties, separately.

The Student t-test reveals significant effects in both results ($p < 5\%$). Concerning the spatial uncertainty, *Size associated to Fuzziness* outperforms *Location* in terms of both response time and correctness. This means that changing the size of a sign is better perceived than shifting it for a short distance. For the terminological uncertainty, the *Thermometer* is better perceived than the *Frequency*. Probably, the performance of participants is reduced for the *Frequency* due to the variation of symbols representing each level of uncertainty. Finally, *Color value* outperforms *Fuzziness* for the global uncertainty; according to Näsänen et al, increasing the blur of an icon would result in longer time to perceive [NO03].

To conclude, *Size associated to Fuzziness*, *Thermometer* and *Color value* are the most appropriate representatives of the uncertainty for the spatial, terminological and global uncertainties, respectively. These three visual variables are then selected for the next step of assessment.

6.2.2 Selection of Uncertainties Information to Portray on Map or on Demand

As mentioned above, portraying the whole uncertainty information may overload the interface of the interactive maps. Our approach proposes instead to portray the uncertainty of the most important information (i.e., spatial, terminological or global) according to users, and providing more details on demand.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

6.2.2.1 Proposal and Simulator

In collaboration with EVS partner, we proposed five different proposals to portray the uncertainties in interactive maps. Figures 6.4, 6.5 and 6.6 refer to the first three proposals where the global uncertainty is not considered. Proposal #1 portrays the spatial uncertainty directly on the map, whereas the terminological uncertainty is shown on demand when a user clicks on the integrated entity (Figure 6.4). Conversely, proposal #2 portrays the terminological uncertainty directly on the map, whereas the spatial uncertainty is shown on demand (Figure 6.5). In proposal #3, both spatial and terminological uncertainties are portrayed on the map (Figure 6.6). Proposal #4 is shown in Figure 6.7, global uncertainty is displayed on the map. In this case, spatial and terminological uncertainties are shown on demand. Finally, proposal #5 portrays all global, spatial and terminological uncertainties on the map (Figure 6.8). Note that when uncertainties are displayed directly on the map, their visual variables are combined with POIs' icons but without affecting the symbols and colors of icons.

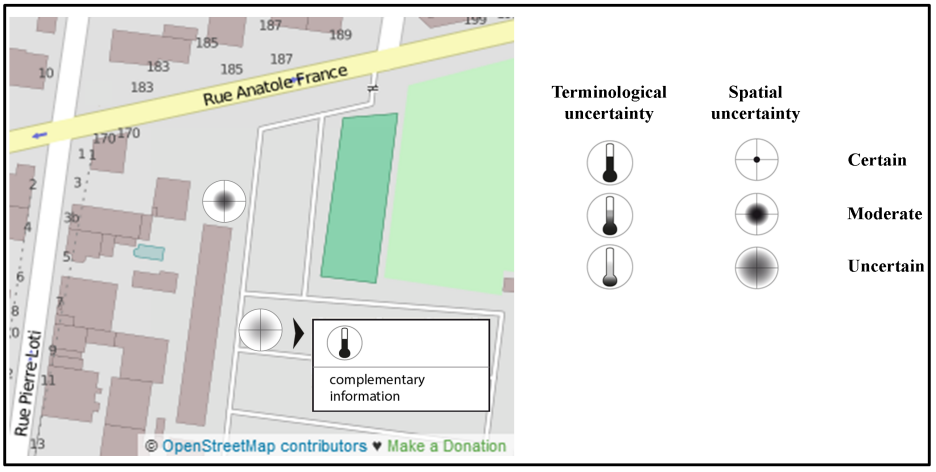


FIGURE 6.4: Proposal #1: Spatial uncertainty is displayed on the map.

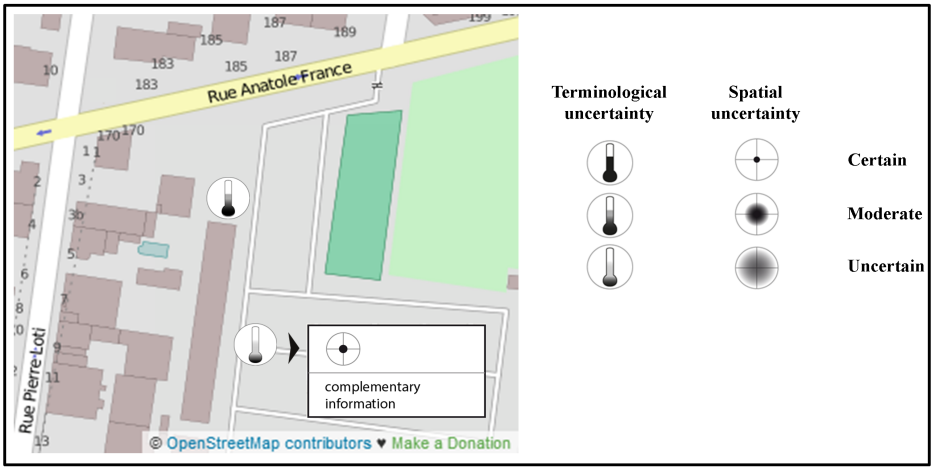


FIGURE 6.5: Proposal #2: Terminological uncertainty is displayed on the map.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

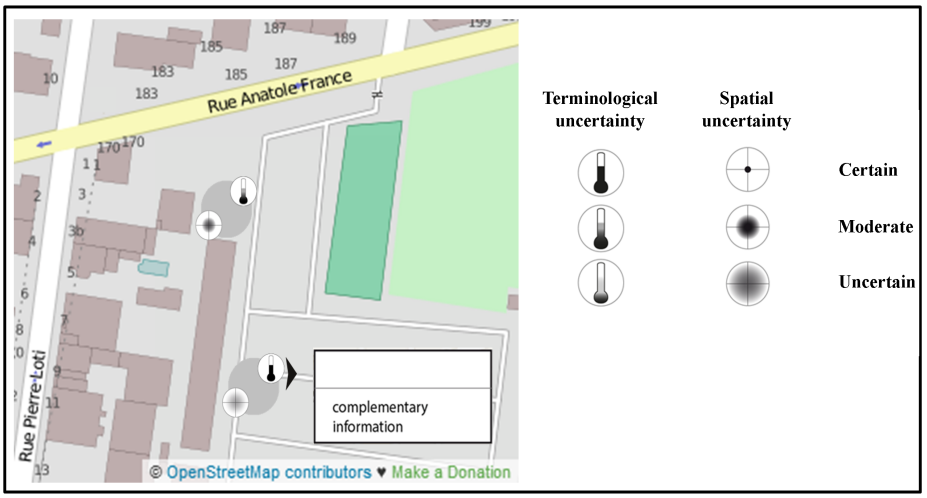


FIGURE 6.6: Proposal #3: Spatial and terminological uncertainties are both portrayed.

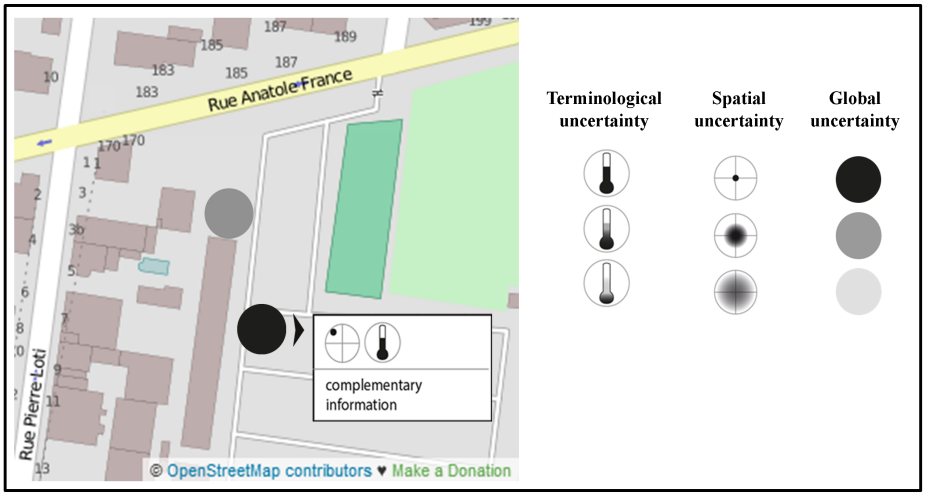


FIGURE 6.7: Proposal #4: Global uncertainty is portrayed on the map.

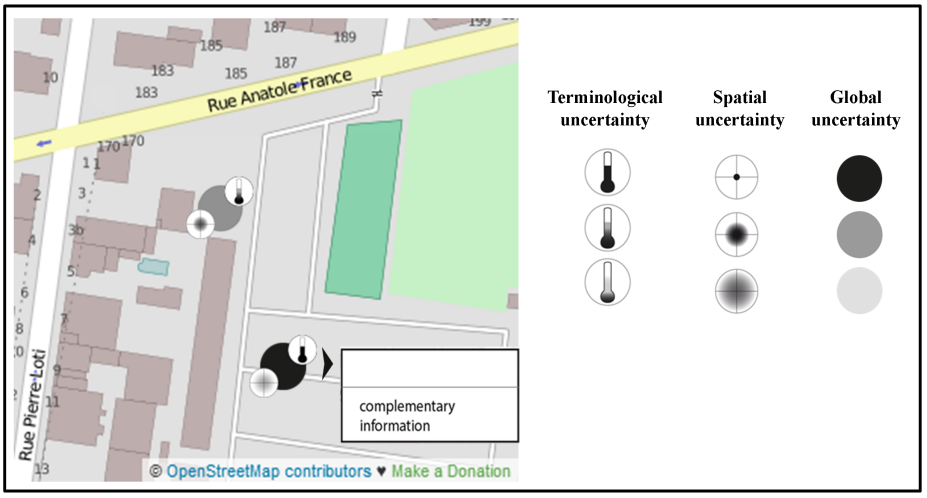


FIGURE 6.8: Proposal #5: All uncertainties are portrayed together on the map.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

We implemented a simulator[10] in order to evaluate the five proposals. For each proposal, a set of POIs is manually generated to represent all possible combinations of uncertainties. These POIs are located randomly on the map near each other (see Figure 6.9) and the terminological information are not considered at all since this simulator focuses only on the uncertainty representation.

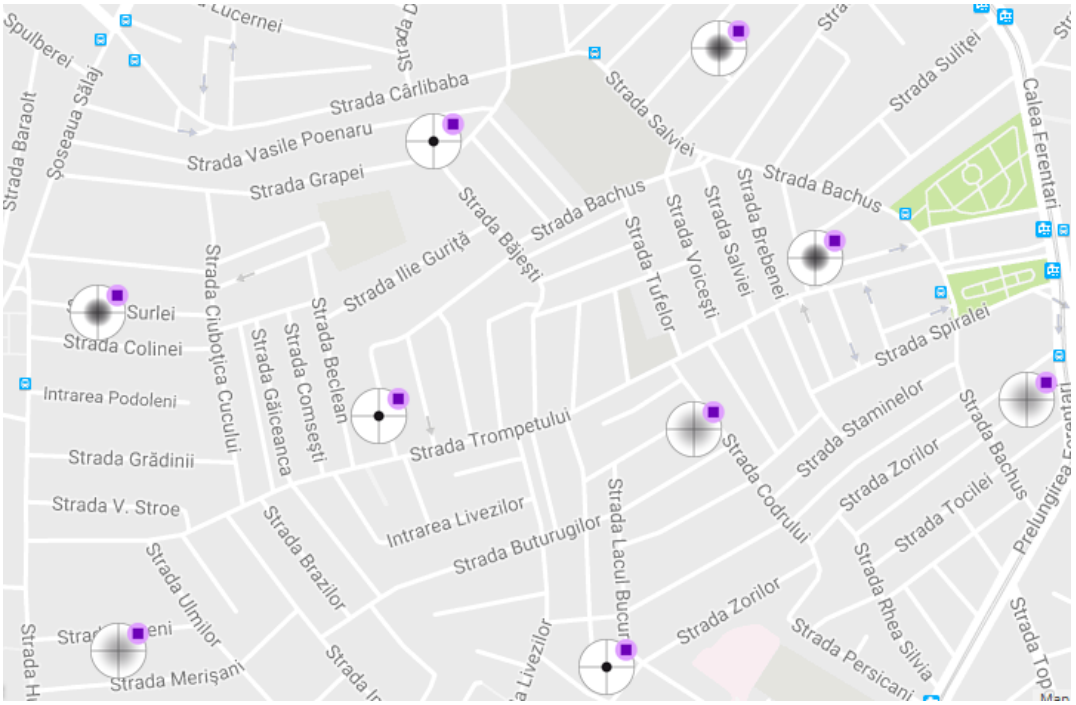


FIGURE 6.9: Simulator: example of POIs location with proposal #1.

6.2.2.2 Assessment of Proposals

A perceptual test of two stages was conducted by the same intern [Sec14]. This test included 25 non cartography expert participants including 14 men and 11 women, aged from 22 to 59 years old. In the first stage, our partner in EVS laboratory implemented a cognitive test to determine which uncertainty is the most important to the user in an abstract context, i.e., without considering any map or legend. The meaning of the three distinct uncertainties (i.e., spatial, terminological and global) was explained to the participants. Then, participants ranked five possible combinations of these uncertainties. Table 6.2 shows the number of times the uncertainties are placed in the first position.

Results are analyzed based on Chi-squared test χ^2 in order to ensure the significance of responses and to reject any random behaviors [Pla83], always with a threshold of acceptability of risk up to 5%. Significantly ($\chi^2 = 20.8$, $p < 1\%$), participants placed

[10] Simulator that implements five proposals for visualizing uncertainty: http://liris-unimap01.insa-lyon.fr/prototype_semio/test2

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

Uncertainties	**Number of times ranked in first position**
Spatial	13
Terminological	1
Spatial and Terminological	3
Global	7
Global, Spatial and Terminological	1

TABLE 6.2: Number of times the uncertainties are placed in the first position.

the spatial uncertainty as the most important (for 13 participants) followed by the global uncertainty (for 7 participants). This means that the uncertainty of spatial information (i.e., location of POIs) was the most important for users when they search for POIs.

In the second stage, our partner in EVS laboratory conducted a cognitive test in order to evaluate the five proposals using our simulator[10] on a desktop computer. Participants were asked to give an appreciation for each proposal on a scale of 1 (not satisfied) up to 7 (totally satisfied) [SIN92, Ber98]. The retrieved data were subjected to ANalysis Of VAriance (ANOVA), based on Fisher's test [Fis25], with repeated measures that revealed significant effect on the type of the proposal ($F = 3.19$, $p = 1.6\%$). Figure 6.10 represents the average score on the appreciation scale with respect to proposals. As shown, it appears that portraying only the spatial uncertainty (i.e., proposal #1) or only the global uncertainty (i.e., proposal #4) slightly outperforms portraying only the terminological uncertainty (proposal #2), and portraying several uncertainties such as proposals #3 and #5 does not increase the preferences.

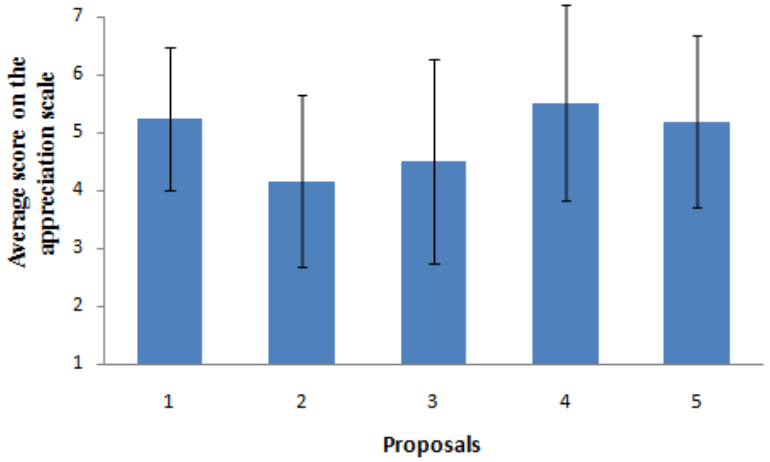


FIGURE 6.10: Average score on the appreciation scale according to proposals.

The above results were not sufficient to select one proposal because the averages of proposals #1 and #4 were approximately equal. Therefore, participants were asked to choose among the five proposals the one that is the most relevant for the context of LBS. Table 6.3 shows the number of times the proposals are chosen as the most relevant.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

Proposals	**Number of times chosen as most relevant**
Proposal #1	7
Proposal #2	0
Proposal #3	4
Proposal #4	9
Proposal #5	5

TABLE 6.3: Number of times the proposals are chosen as the most relevant.

A Chi-squared test χ^2 is conducted on the distribution of preferred proposals for participants. Result shows a trend effect towards significance ($\chi^2 = 117.64, p = 5.6\% > 5\%$) for proposal #4 (global uncertainty) followed by proposal #1 (spatial uncertainty).

According to the previous experiments, proposal #4 is selected as the best and most appropriate method to inform users about uncertainty of integrated geospatial data. This proposal consists in portraying the global uncertainty directly on the map using a *Color value* visual variable. Whilst, the spatial and terminological uncertainties are shown on demand in the tool-tip of complementary information. *Size associated to Fuzziness* and *Thermometer* visual variables are used to portray the spatial and terminological uncertainties, respectively. In the next section, we intend to evaluate whether the visualization of uncertainties is really taken into account by tourists to search POIs.

6.3 Impact of Visualizing Uncertainty in LBS: A Use Case for Tourists

After discovering the best and most appropriate solution to represent the geospatial uncertainties in LBS context, one crucial assumption still remains to be checked: is portraying uncertainty useful information for tourists? This section evaluates how portraying uncertainty information impacts tourists' behavior when searching for POIs [CFC+16]. More specifically, we intend to assess whether uncertainty information is accepted and also taken into consideration by non geography expert users in the context of a tourist trip. A collaboration between UNIMAP partners has lead us to elaborate an experimental protocol for this test, which is performed by another intern who has also followed a research master in cognitive science [Cat15].

6.3.1 Experimental Protocol and Simulator

The experimental protocol consists first in checking whether available uncertainty information is taken into account by participants, and second in observing how it has been

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

used: as a tool to filter data or as additional information to process?

Compliant with previous background material, the adopted approach is to first gather preliminary knowledge about tourists' needs and contexts of tourist trips. To do so, we elaborated a list of questions in order to conduct a series of interviews with partners of UNIMAP that are professionals in the domain of tourism in France such as Only Lyon tourist office, Saint-Etienne tourist office and Rhône-Alpes Tourisme. The purpose is to identify tourists' archetypes, contexts of trip planning and to check how tourists' strategies are identified by tourist offices. Thanks to this preliminary work, three tourist trip missions are included in this experiment:

1. M1: it consists of a tourist who is planning for a future trip and wants to book an hotel,
2. M2: it consists of a tourist who is looking on site for a restaurant during a trip,
3. M3: it consists of a tourist who is looking on site for a monument during a trip.

To reach the goals of this experiment, we imagine three different groups of participants that are all familiar with LBS:

1. G1: The first group is composed of 8 men and 7 women ($N_1 = 15$), aged from 21 to 29 years old (Average = 24.13, Standard deviation = 2.3). This group is used as a control group to measure a basis response time for choices.
2. G2: The second group is composed of 7 men and 8 women ($N_2 = 15$), aged from 21 to 28 years old (Average = 23.47, Standard deviation = 2.17). This group is another control group from the cognitive load point of view.
3. G3: The third group is composed of 7 men and 8 women ($N_3 = 15$), aged from 22 to 53 years old (Average = 27, Standard deviation = 8.08). The third group is used to check whether uncertainty information is used by participants.

For each mission, a participant of a given group should select a POI among several ones. To do so, each participant studies four maps: one training map with no expectations just provided for participants to get accustomed to such maps, one for the first mission M1 containing nine POIs of type *Hotel*, one for the second mission M2 containing nine POIs of type *Restaurant*, and last one for the third mission M3 containing nine POIs of type *Monument*.

For each mission, the same POIs are used for all groups but with different uncetainties' levels:

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

1. in G1: POIs are presented without uncertainty information and participants do not have access to the source entities of providers.

2. in G2: POIs are presented with the same uncertainty level and participants have the possibility of accessing the source entities of providers.

3. in G3: POIs are presented with varying levels for uncertainties and participants have the possibility of accessing the source entities of providers.

We implemented a simulator[11] for the use of a multi LBS providers including our solution for visualizing integrated entities. The goal is to provide participants with an interface as close as possible to real LBS. The interface is also adapted to the contexts of the experiment to avoid any bias; it provides only the necessary elements for the purposes of the specific given missions. Datasets of fake POIs have been generated for each mission. In order to avoid the bias of any knowledge on the tested POIs, the test takes place in the city of Bucharest and we preliminarily checked this city was unknown by all participants. POIs' main terminological attributes are the most frequently provided ones including *name, address* and *phone*. Preliminary tests made by tourist offices indicates that tourists consider the price as the main criterion when choosing an hotel and a restaurant, while opening hour is the main criterion when choosing a place to visit such as monument or museum. Therefore, we added a fake *price* attribute for hotels and restaurants POIs, and a fake *opening hours* attribute for monuments. Concerning the first mission M1 (i.e., looking for hotels), the POIs are located on the map near each other. Whilst, for the second and third missions (i.e., on site looking), participant's location is indicated on the map and POIs are surrounding at the same distance on a 600 meters radius circle in order to avoid any decision based on the distance to POIs. Recall that integrated entities are represented using the proposal #4 (see Section 6.2.2), i.e., global uncertainty portrayed directly on the map. Figure 6.11 shows the interface of this prototype. The "Pre-built Map Selector" panel allows participant to switch the datasets of the three missions exclusively. In this figure, the second mission M2 (i.e., choosing a restaurant on site) for the third group G3 (i.e., integrated entities include several levels of uncertainties) is selected. The Legend panel shows the icons of the selected mission to represent the POIs and their uncertainty levels. Beside these two panels, there is an interactive map built using Google Map's background. As shown, the location of participant is in the middle of the map surrounded by nine restaurants of different uncertainty levels at the same distance. A click on an icon shows its terminological information in a tool-tip as shown for *Caru' cu Bere* restaurant.

[11] Simulator to evaluate the impact of uncertainty visualizing: http://liris-unimap01.insa-lyon.fr/prototype_semio/test3

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

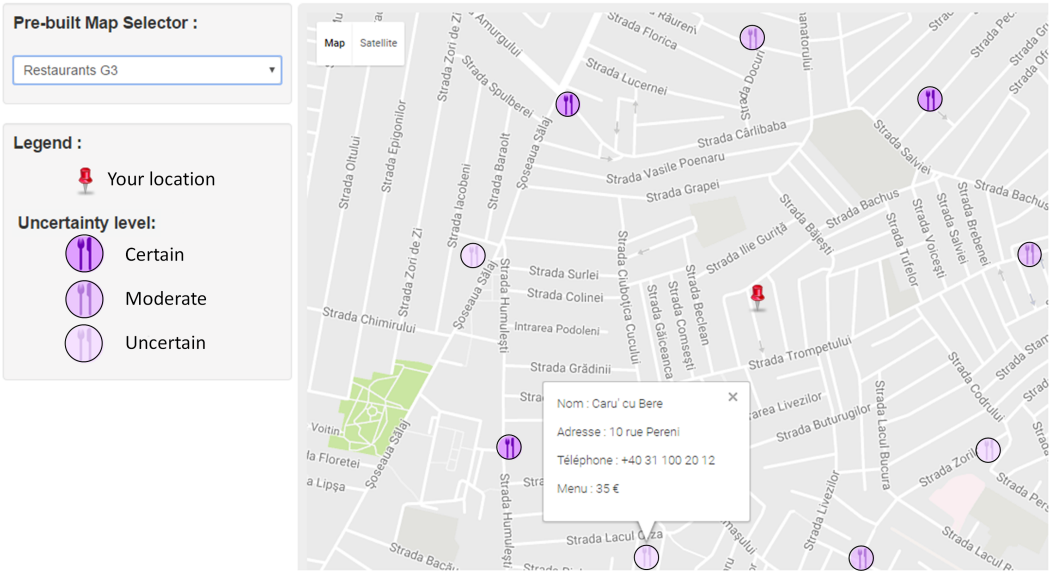


FIGURE 6.11: Interface of a first prototype for multi LBS providers.

Participants of G2 and G3 have the possibility to access the source information of providers switching the map to another "source mode" on demand. A double click on an icon shows markers referring to the locations of source entities and a table comparing the whole information (see Figure 6.12).

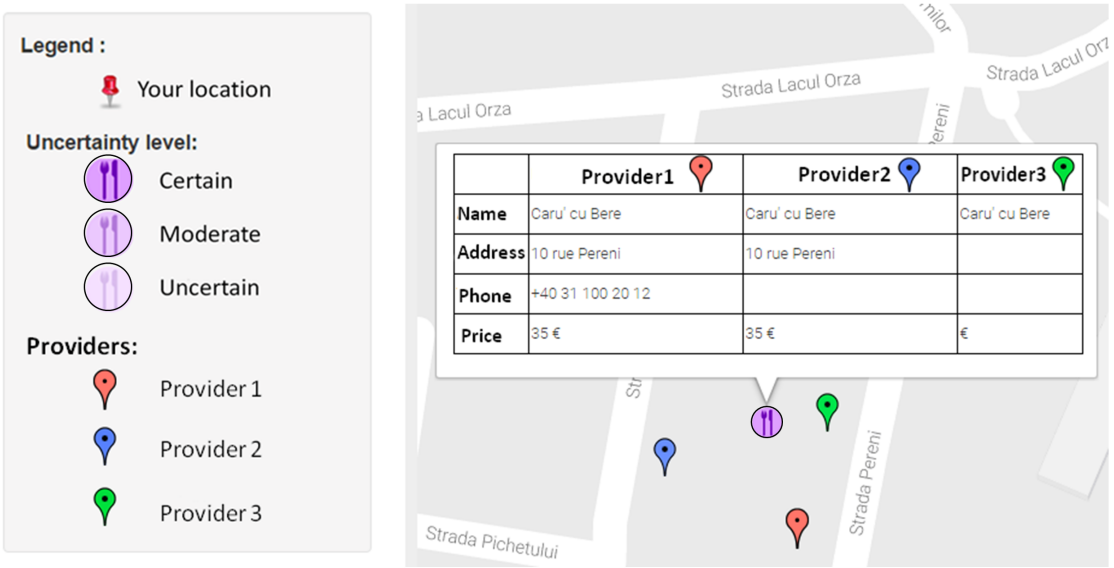


FIGURE 6.12: "Source mode" to compare the source entities on demand.

For all participants, the experiment was performed on a desktop computer. This was done to avoid the bias generated by different devices, such as the small screen of smart phones. Different kinds of data have been collected. We are interested in response time spent by participants in choosing a POI in each mission as well as their choice criteria, especially for G3 who have an additional varying uncertainty information to handle.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

6.3.2 Results and Analysis

We first checked whether data follow a normal distribution using a Shapiro-Wilk test [SW65] with a threshold of 5% (i.e., if p is less than the chosen threshold, then there is evidence that the data tested are not from a normally distributed population):

- in M1: $W = 0.89$; $p < 1\%$
- in M2: $W = 0.91$; $p < 1\%$
- in M3: $W = 0.88$; $p < 1\%$

where W and p values are Shapiro-Wilk test statistics.

As we found collected data are not following a normal distribution ($p < 5\%$), according to cognitive science procedure, a Kruskal-Wallis H test [KW52] is applied to compare the three groups, mission by mission, and then applying a Mann-Whitney test [MW47] to compare the groups two by two and mission by mission.

Means and standard deviations of the responses' times for the three missions are given in Figure 6.13. With 5% significance threshold, Kruskal-Wallis H test yielded significant effects between the three groups:

- in M1: $H(2, N = 45) = 23.9$; $p < 1\%$
- in M2: $H(2, N = 45) = 14.3$; $p < 1\%$
- in M3: $H(2, N = 45) = 14$; $p < 1\%$

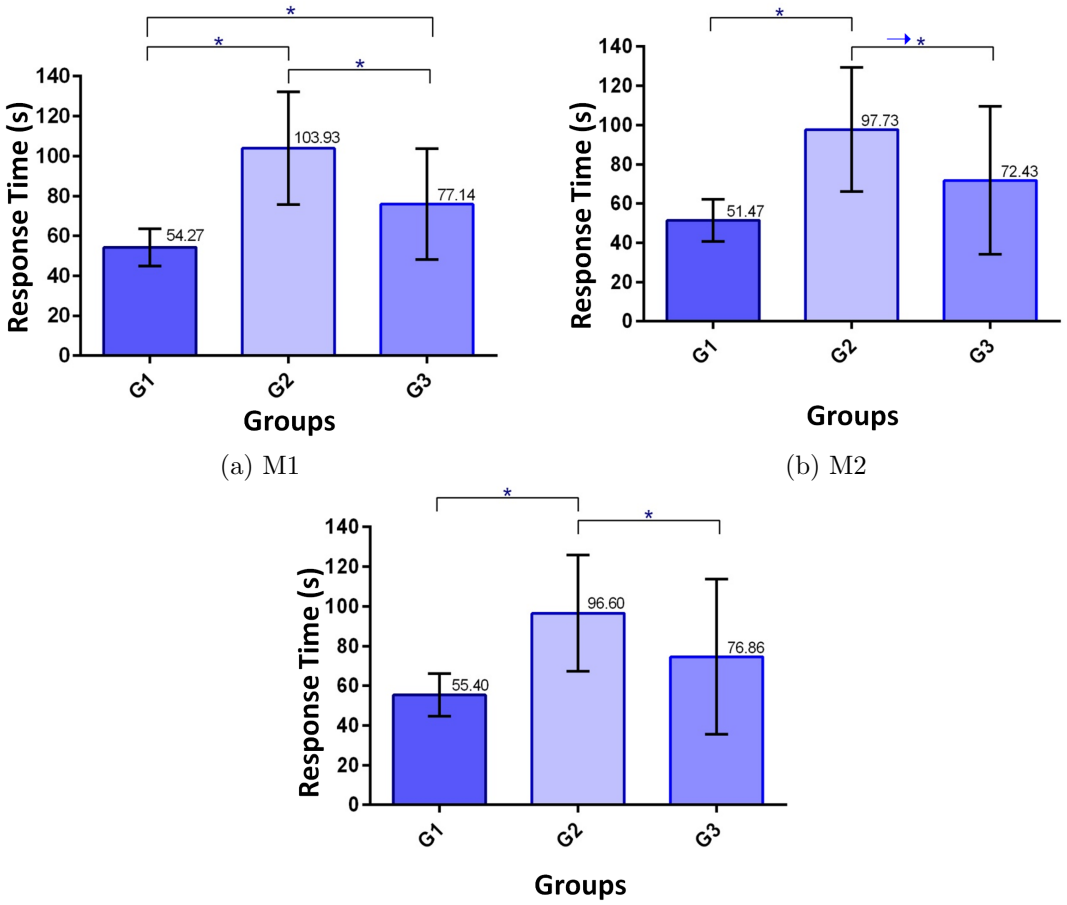
where H and p values are Kruskal-Wallis test statistics.

With 5% significance threshold, Mann-Whitney test yielded the results of Table 6.4 where we can observe:

- in M1: significant effects between each group.
- in M2: a significant effect between G1 and G2, and not significant but as a trend towards significance between G2 and G3.
- in M3: a significant effect between G1 and G2, and between G2 and G3.

As we can observe in Figure 6.13 and Table 6.4, whatever the mission, response time is significantly shorter for G1 compared to G2, and also shorter for G3 compared to G2

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés



(a) M1 (b) M2

(c) M3

FIGURE 6.13: Mean response times for the three missions. Black bars represent standard deviations. "*" indicates significant differences between groups, and "→*" a trend towards significance ($p = 5.6\%$).

	Variables	p (%)
M1	G1-G2	**<1**
	G1-G3	**<1**
	G2-G3	**<1**
M2	G1-G2	**<1**
	G1-G3	9
	G2-G3	5.6
M3	G1-G2	**<1**
	G1-G3	50
	G2-G3	**2.8**

TABLE 6.4: Mann-Whitney test results with 5% significance level: response time comparison of groups two by two. Significant results are indicated by bold, a trend towards significance are underlined.

(significantly or as a trend towards significance). G1 seems to have the shortest response time because of the lower amount of available information to process; no uncertainty and no source providers information. G2 showing the longest response time for each mission makes us able to state that the additional uncertainty visualization and source providers

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

information imply a heavier cognitive load for the user. As G2 and G3 have the same amount of available information to process, we can conclude that this is the varying uncertainty information which made G3 participants faster than G2. Participants of G2 did not use uncertainty as a criterion for their choices as the level is the same for all POIs. If we now look at the response time difference between G1 and G3 that is significant in M1, the response time trends to be similar. So, the additional cognitive load implied by source providers information seems to be reduced by portraying varying levels of uncertainty.

Another result could be observed concerning response time of group G3, whatever the mission, standard deviation is always wide. This phenomenon could be explained by the behavior of two participants profiles namely "optimizers" and "maximizers" [SWM+02]. The optimizers use the uncertainty's level as a filter to select POIs. This can shorten the time to complete a mission. The maximizers would deeply browse every POI to check out every information available, every property of every source provider which could generate this uncertainty. In this case, time to complete the mission would be longer than other groups' time.

As only G3 has varying uncertainty level portrayed, we asked participants of this group how the criteria (i.e., uncertainty level, price and opening hours) are used for their choices:

- in M1: 100% of participants say that the uncertainty is the main criterion.
- in M2: 100% of participants say that also the uncertainty is the main criterion.
- in M3: 80% of participants say that the uncertainty is the main criterion whereas 20% declare that their choices are made based on the *opening hours* criterion.

We could observe that the three missions' objectives were not imposing the use of such information, but users used it as a major criterion for their choices. Whatever the mission, almost all of participants used uncertainty information as the main criterion for their choices.

As a final remark, despite our best precautions, participants indicated that POI's localization is an additional criterion to justify a choice, for instance because of the presence on the map of an avenue or a dead-end street. It seems to be very difficult to fully wipe out every map bias when not using a fake map; all participants have their own sensibilities and preferences. Moreover, when someone is using a map, he/she processes local information, e.g., looking for a specific place, but also contextual information, e.g., what is surrounding? like parks, public transportation, etc. This information constitutes

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

additional criteria taken into account when making a choice. It seemed to us difficult to fully avoid this bias as we also wanted the experiment as close as possible to real life conditions.

6.4 Multi-providers LBS Prototype

This section represents a multi-LBS providers prototype[12] for the UNIMAP[13] project including our solution for matching and merging geospatial entities, and visualizing their uncertainty. The purpose of this prototype is to offer a more complete result to the users by integrating the data of multi-LBS providers. We first describe an architectural overview of the prototype, then we represent its interface and a navigation scenario.

6.4.1 Overview and Architecture

With the development of the internet and web services, the GIS community can benefit from the experiences and technical progress to create spatial data infrastructures and geo-web services. Our prototype can be seen as a mediator LBS; Figure 6.14 schematizes the communication between a user and the prototype. As we can see, a user searches for POIs of a given type in a given area by sending a request to the UNIMAP prototype. This latter queries the LBS providers' web services by calling their APIs. Once the prototype receives the datasets of source entities from all providers, a process is in charge to integrate the datasets in real-time. The integration's result is then sent to the user in order to visualize and explore the integrated entities of multi-LBS providers.

Figure 6.15 details the phases inside UNIMAP prototype. The first phase takes the user's request as input and generates an appropriate query for each LBS. Then, calls the LBS providers APIs servers in parallel in order to collect the entities that suit the user's request; each provider returns one dataset. Once all datasets are received, the third phase matches them together in order to find the corresponding entities. In this prototype, a serial join algorithm is used to match several entity sets (see Section 2.2.2 in Chapter 2). Note that the distinct schemas and types architecture of providers have been matched manually (see Sections 4.1.2 and 4.1.3 in Chapter 4). Recall that corresponding entities are detected using GS_{Avg} approach with a 0.5 decision threshold. This threshold has been fixed according to experiments performed in Chapter 5. The next phase is in charge of merging the corresponding entities and computing their spatial, terminological and global uncertainties' scores (see Section 6.1 in Chapter 6). Finally,

[12] LBS prototype for UNIMAP project: http://liris-unimap01.insa-lyon.fr/prototype_unimap
[13] UNIMAP project: http://liris.cnrs.fr/unimap

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

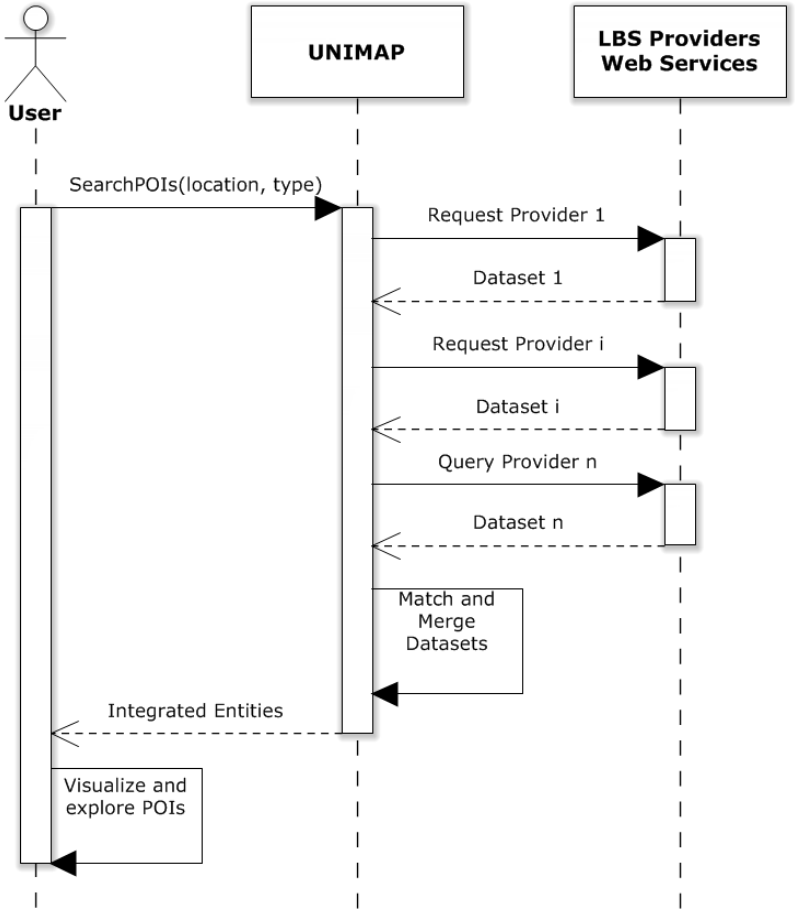


FIGURE 6.14: Sequence diagram for UNIMAP LBS framework.

integrated entities are sent to the user and represented on an interactive map. A user is free then to explore them. Note that integrated entities are represented according to their uncertainties level using the proposal #4 presented in Section 6.2.2 of Chapter 6, i.e., global uncertainty is portrayed directly on the map, while spatial and terminological uncertainties are portrayed on demand.

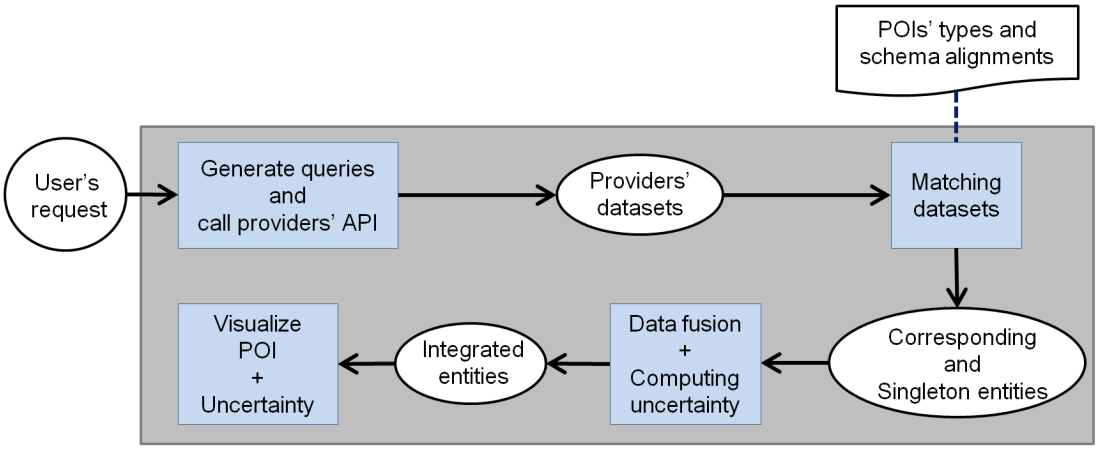


FIGURE 6.15: Process flow of UNIMAP prototype.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

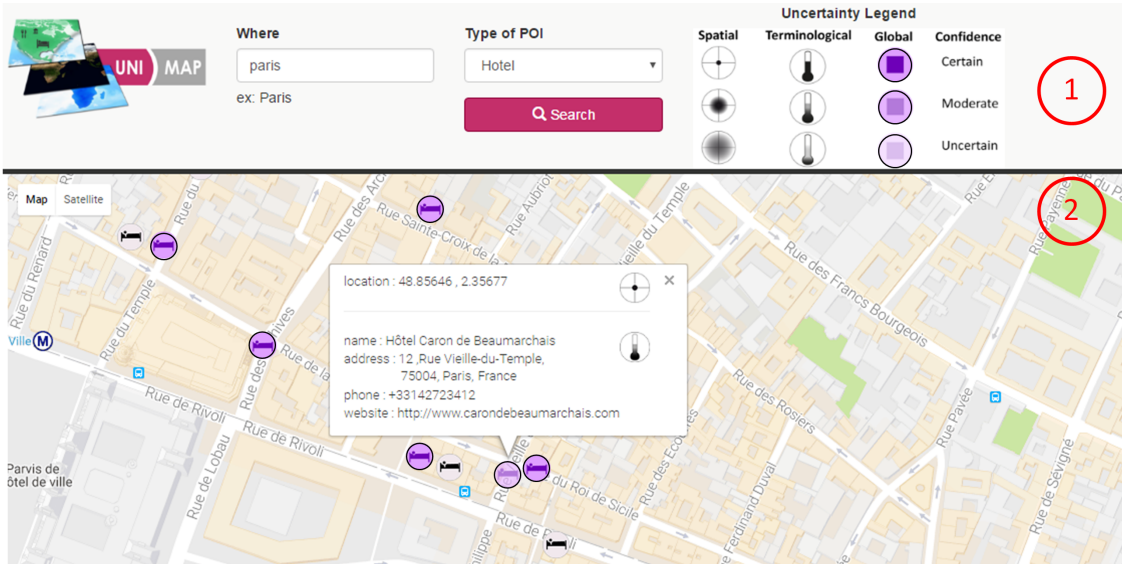


FIGURE 6.16: Interface of UNIMAP prototype.

This prototype is implemented as a web application using Javascript and PHP; it integrates on real-time the POIs of three LBS providers namely Google Maps, Nokia Here Maps and Mapquest. The current version uses the free APIs packages of LBS providers (see Table 2.1 in Chapter 2), which creates some limitations such as the allowed number of queries per day and the number of returned POI per query.

6.4.2 Interfaces and Navigation

The prototype interface is composed of two panels as shown in Figure 6.16. The first panel includes the query's fields in which users indicate the location and type of POIs they required. Also, it contains the legend for the visualization solution used to portray global, spatial and terminological uncertainties. The second panel consists of an interactive map inheriting Google Maps background and features (e.g., zoom in/out, satellite/map view, etc).

When a user starts navigating, the prototype detects and centers the map at a user's location. The user specifies the targeted location and selects the POI type using the query's fields "Where" and "Type of POI", then launch the search by clicking on the the "Search" button. First, the prototype collects all appropriate POIs from the three LBS providers. Secondly, collected entities are matched and merged together in real-time, corresponding and singleton entities are then displayed to the user. By default, integrated entities and their global uncertainties are displayed on the map. For example, the map panel of Figure 6.16 illustrates the POIs of type hotel in Paris. Uncolored icons refer to singleton entities, while colored icons indicate the degree of global uncertainty for integrated entities. The user can click on a POI to display the tool-tip that contains

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

the full POI information. The tool-tip of a singleton entity shows the information of the POI as they are offered by the LBS provider. Concerning integrated entities, the tool-tip contains the integrated information, spatial uncertainty and terminological uncertainty of the integration as shown for *Hôtel Caron de Beaumarchais* in Figure 6.16. At the right top corner of the tool-tip, the *Size associated to Fuzziness* icon indicates that the spatial information are certain, while the *Thermometer* icon indicates that the terminological information are moderately certain for the selected POI.

In addition, the prototype allows users to portray the providers' original entities of an integrated POI. The user can right-click an integrated entity to check and compare the source entities. This case is called the Source Mode where all POIs are hidden, except the selected one. Figure 6.17 shows the source entities of an integrated entity; the hotel icon shows the location of the integration result, while markers refer to the location of original entities. As we can see, there are only two original entities, which means that the third provider has a Not Found Entity (NFE) for the selected POI. A click on a marker displays a tool-tip containing the terminological information of its provider. Also, the user can compare the terminological information of all providers by clicking on the hotel icon to display a comparative table as shown in Figure 6.18. In this example, the two entities have the similar name and address. In contrary, the phone number has two different values , while the website is given by Here provider and is missed by Mapquest. A right-click on any icon hides the source entities and re-displays the integrated entities.

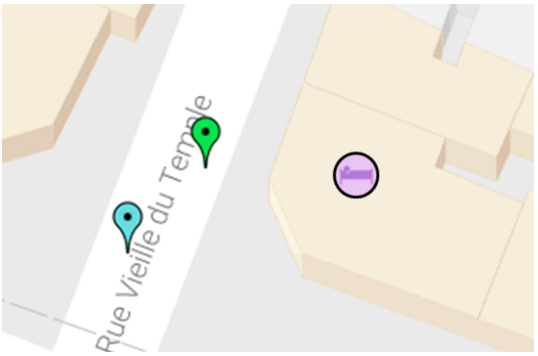


FIGURE 6.17: Source mode of an integrated entity.

6.5 Conclusion

This chapter reveals the merging and visualizing of integrated entities by considering their uncertainties. In a first stage, a basic data fusion algorithm is proposed; it considers the value that is most similar to the others and estimates its uncertainty. This

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

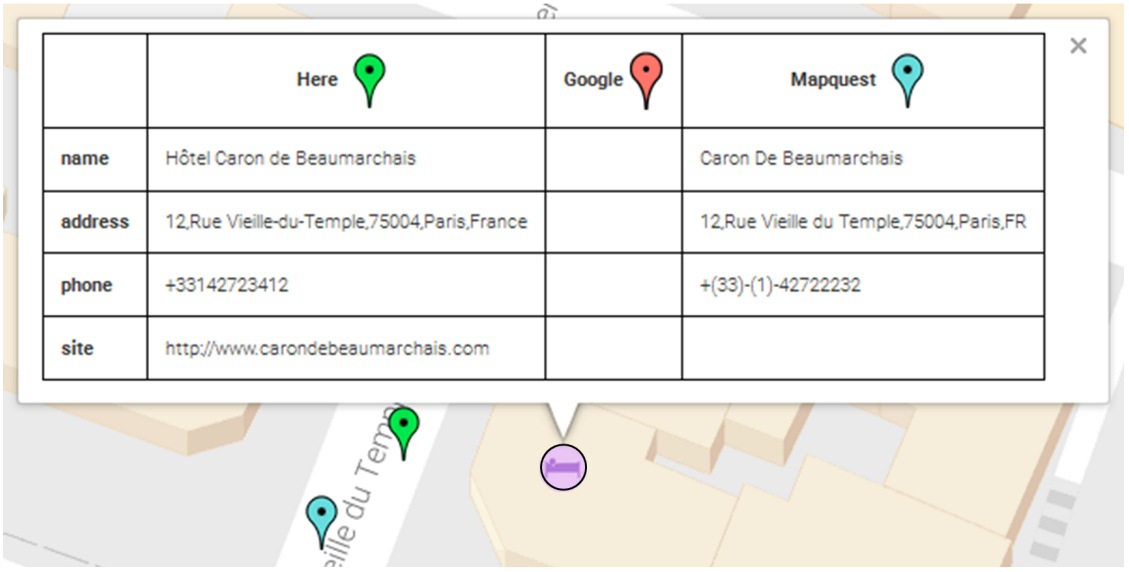


FIGURE 6.18: A comparative table of an integrated entity.

generic algorithm can be replaced by another generic approach possibly based on different assumptions. In a second stage, existing works of uncertainty visualization have been analyzed where three uncertainties are considered: spatial, terminological and global. Then, a study is conducted by our partner to select and evaluate the best appropriate visual variables to represent the uncertainties of integrated data. In addition, cognitive test indicates that users prefer visualizing the global uncertainty directly on the map, while the spatial and terminological uncertainties are displayed on demand. Another cognitive test is conducted to study how additional uncertainty information impacts users' behaviors for different missions in a tourist trip. Result shows that almost all users utilize uncertainty as a main criterion, when available, to choose their POIs. Finally, we proved the feasibility and the benefits of our research by implementing a prototype for multi-LBS providers. This latter includes our solution for matching and merging inconsistent entities collected from different sources of unknown quality. As well as the visualization of integrated entities including their uncertainties. This prototype is a limited application used as a demonstration of researches made within the framework of the UNIMAP. The next chapter concludes this dissertation.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

Chapter 7

Conclusions and Perspectives

Contents

7.1 Contribution Summary 156
7.1.1 Taxonomy and Benchmarking 156
7.1.2 Spatial Similarity Measure and Geospatial Entity Matching . . 156
7.1.3 Visualization and SHS Evaluation 157
7.2 Short-term Perspectives 158
7.2.1 Enrichment of Taxonomy and Benchmark 158
7.2.2 Improving Geospatial Entity Matching 159
7.2.3 Different Estimations of Uncertainty 159
7.2.4 Experimenting New Visual Mode 160
7.2.5 Considering the Geographical Context 160
7.3 Long-term Perspectives 161
7.3.1 Enhancing Geospatial Entity matching 161
7.3.2 Combining Blocking Algorithm 161
7.3.3 Visualization and Navigation 162
7.4 Final Words 162

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

The amount of geospatial data and the number of LBS providers have been growing at a dramatic pace in recent years. The contradiction between geographic entities, which come from multiple sources and describe the same reality casts doubts on the validity of the POIs offered. Geospatial entities referring to the same POI may include incomplete, inconsistent, inaccurate or even wrong data from one provider to another at spatial, terminological and legend levels. This thesis is a contribution to improve the correctness and the completeness of geospatial data coming from multiple LBS providers. Our proposal is applied to a use case of POIs for tourists.

7.1 Contribution Summary

Contributions in this thesis can be summarized as follows: (i) designing a benchmark to evaluate geospatial entity matching approaches, (ii) proposing entity matching and merging algorithms and (iii) finding a suitable map-design to visualize and represent integrated POIs.

7.1.1 Taxonomy and Benchmarking

In this thesis, we highlighted the absence of a benchmark to compare and evaluate geospatial entity matching approaches. Thus, we proposed a taxonomy that characterizes the inconsistencies between LBS providers at four levels: schema, terminology, spatial and availability. We studied the impact of the identified differences on the results' quality of a matching approach. We believe that the proposed taxonomy will allow researchers to better evaluate their matching approaches, identify the capabilities of their approaches and also guide performance improvements in existing geospatial entity matching approaches. Based on this taxonomy, we designed a benchmark, called PABench [BDF+15], that serves to evaluate and compare geospatial entity matching approaches. The evaluation datasets of PABench contain real-world entities collected from existing LBS providers using a tool called GeoBench [MMBD14].

7.1.2 Spatial Similarity Measure and Geospatial Entity Matching

The integration of geospatial data consists of three main phases: schema matching, entity matching and data fusion [PS00, SKS+10]. Current LBS providers have simple schemas. Hence, the schema matching task is done manually and it is out of the scope of this thesis.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

Concerning the entity matching, our approach consists of measuring and combining the similarities of several attributes in order to improve the detection of corresponding entities. First, we studied several mathematical functions in order to propose a generalization for spatial similarity measure. This estimates the similarity between two geospatial entities by comparing their location coordinates. We also evaluated the performance of several existing terminological similarity measures in order to find the appropriate similarity measure for each terminological attribute. Then, we implemented and compared several methods to combine spatial and terminological similarities. Finally, a threshold-based decision algorithm is proposed; it reduces the impact of inconsistencies on the matching results. Experimental evaluation using both full datasets and characterized test cases of PABench are done to assess and compare our proposition to some existing methods. Extensive results show that (i) combining spatial and terminological similarities instead of one type of information can improve the performance of matching and (ii) our combination method outperforms some existing works in terms of both efficiency and effectiveness.

Regarding the data fusion, we proposed an algorithm to merge corresponding entities detected during the entity matching phase. For each attribute, we chose the value that is the most similar to the others. Because the result of the merging process may not be 100% reliable, we chose to inform users whether the values of integrated entities are trustworthy. To do so, our merging algorithm estimates both spatial and terminological uncertainties, these refer to the spatial and terminological information, respectively, as well as a global uncertainty that groups these two uncertainties. It is important to note that these uncertainty scores are deduced from the similarities between the values of source entities.

7.1.3 Visualization and SHS Evaluation

We investigated the solutions to represent integrated POIs with their uncertainties [MRH+05]. A cognitive study has been conducted with UNIMAP partners, that allows both selecting the appropriate visual variable for each dimension of uncertainty and finding the most useful information of uncertainty to display in a LBS interactive mapping application [BCD+14, BDC+14, SCC+14].

We have proposed and studied different representations of uncertainty in a geospatial integration context. The uncertainty scores are converted into three confidence degrees namely: certain, moderate certain and uncertain, that have been evaluated among many users. Solutions have been implemented into a first application simulator to demonstrate the feasibility and the benefits in a scenario. This experiment indicates that users

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

prefer visualizing the global uncertainty directly on the map, while the spatial and terminological uncertainties are displayed on demand.

Another experiment in the context of tourist trips has specified how additional uncertainty information impacts the users' behaviors when using multi-providers LBS [CFC+16]. We can state that adding varied uncertainty visualization impacts user choices and time to make them. We could observe that adding source providers information increases user's cognitive load but this cognitive overload seems to be reduced by visualizing varied uncertainty levels. We can also state that uncertainty information is taken into account in user decision. This experiment-driven research work leads us to globally conclude that visualizing uncertainty is a useful additional feature, required by potential users, to design LBS that integrate POIs from different providers in the context of tourism.

Finally, we proved the feasibility and the benefits of our contributions by implementing a prototype for the UNIMAP project. This composite prototype is developed to automatically generate a multi-provider LBS in real-time and without any human intervention. After having summarized our contributions, we will represent the limitations we faced and the short and long term perspectives that should be considered.

7.2 Short-term Perspectives

In this section, we discuss the current perspectives that can be addressed in the near future in order to improve our proposals.

7.2.1 Enrichment of Taxonomy and Benchmark

Our taxonomy formalizes the differences that may appear between two entities. But, we previously discussed and analyzed the impact of complex combinations of differences that may appear between an entity from one provider and several entities from another provider. An interesting work consists of extending the taxonomy by defining and formalizing these complex combinations. After this, GeoBench's database should be extended with entities that have complex combinations in order to create evaluation datasets. These datasets allow us to assess the performance of the matching approach against these complex combinations.

Additionally, our benchmark PABench specifies a list of test cases to evaluate geospatial entity matching approaches. However, the evaluation datasets collected from GeoBench's database do not cover all these test cases. As for June 2016, 48 out of 96 defined test

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

cases rarely occur. These datasets may be enhanced through two steps. The former consists in comparing and adding more real entities. The latter consists in automatically generating entities to cover the situations of differences that occur only rarely between entities. For instance, it is possible to use existing geospatial entities and apply modifications (e.g., abbreviation, synonyms) to the values of spatial and terminological attributes [IRV13]. Moreover, we suggest to exhaustively collect the POIs of a given geographic area from several providers. These POIs help extract statistics and frequencies of the situations of differences found in reality.

7.2.2 Improving Geospatial Entity Matching

It is possible to add and evaluate more methods to combine similarities, in order to have better matching's result. In addition, our matching algorithm can be extended to consider the problem of matching POIs that change their locations by time, such as events or a restaurant in a cruise ships [XDZ09]. For such case, it is possible to extend the spatial similarity measures to compare sequences of location's time-line instead of fixed locations. Yet, one problem remains concerning the POIs that are superposed. But, this problem may be resolved if LBS providers offer additional geospatial information. Actually, the APIs of all providers mentioned in this thesis offer the latitude and longitude of POIs' locations, but the altitude is not provided despite the fact that some providers own this information. For example, the mobile application of Google Maps allows users to visualize the POIs of each floor of a mall apart. The availability of such information would improve the quality of data integration and visualization.

In addition, we proposed a 2-join geospatial entity matching approach that is limited to match two datasets. For our multi-provider prototype, we used serial join to match three datasets based on our 2-join algorithm. A disadvantage of the serial join is that the integration's result strongly depends on the order of matched datasets [SKS+10, KKH+10]. Hence, an interesting research aspect is to compare existing works that allow the matching of three or more datasets. Criteria such as result's quality and performance should be considered in this comparison.

7.2.3 Different Estimations of Uncertainty

Our data fusion algorithm is basic and can be replaced by another generic approach, possibly based on different assumptions [LDOS11, DGH+14]. One important task of our data fusion algorithm is to estimate the uncertainty of a chosen value among all available values. In this thesis, the range values of the uncertainty levels have been chosen arbitrarily for our cognitive experiments. Some work is still needed to adjust

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

these intervals to better express the uncertainties they represent. The estimation of uncertainties has become an important issue in scientific computation. It consists of estimating the dispersion of the output of a model due to the uncertainties associated with the inputs of the model, which are the LBS providers' data in our case. Machine learning methods can be used for this purpose. For instance, Monte Carlo methods [MU49] can be seen as approximation methods, in the statistical sense of the term, that are used in many fields such as risk management, biology, mathematics and finance. Such methods are probability based that aim to evaluate a deterministic quantity by using random processes. In other words, a statistical analyzing of random input and output data allows us to classify certain and uncertain outputs.

7.2.4 Experimenting New Visual Mode

Our cognitive experiments for uncertainty visualization have been done using desktop computers. But, LBS are widely used on mobile devices that do not have the same properties as desktop computers. Hence, our solution may be extended for mobile users and additional experiments will be needed to ensure the feasibility and benefits. We could also include contextual information (e.g., surrounding events, accessibility using public transportation, impact of the environment, etc.) in the experiments in order to expand further significant results. In addition, the results of some experiments show a trend effect towards significance due to lack of participants. Including more participants in future experiments will certainly help obtain reliable results.

7.2.5 Considering the Geographical Context

Our taxonomy could be improved by relating the differences between entities to the context of geographical zone, such as rural or urban area, or even the type of district (e.g., poor, rich, industrial, etc.). Characterizing such relations allows us to improve the geospatial entity matching approaches according to the geographical context. For instance, after identifying the differences of distinct geographical zones, experiments are needed to find the best matching algorithm that is able to handle the differences of each zone. On this base, when a user seeks a LBS query, the matching algorithm is selected depending on the geographical zone of the query.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

7.3 Long-term Perspectives

This section provides a discussion on some of the possible future directions for the continuation of this research work.

7.3.1 Enhancing Geospatial Entity matching

This study considers the integration of punctual geospatial entities, but it may be extended for the integration of multidimensional geospatial data (e.g., polygons, lines and 3D). To do so, we should extend our taxonomy to define the relations between multidimensional entities (e.g., topological relations), as well as defining their inconsistencies and resemblances. The extended taxonomy helps us understanding the relations between these entities in order to propose a suitable spatial similarity measure and extend the geospatial entity matching approaches.

The quality of geospatial entity matching may be improved using *Social and Human Science* (SHS) common rules [Col13]. For example, some POIs' types are correlated to POIs' locations such as post offices that are often located in the main streets and luxury stores that are grouped into downtown. First, a social study is needed to define a list of SHS rules that characterizes the POIs. Then, the geospatial entity matching approach can be configured based on those contextual rules to corresponding entities. After enhancing the evaluation datasets of PABench, we can assess such rule-based approach to determine whether or not SHS rules improve the quality of geospatial entity matching.

7.3.2 Combining Blocking Algorithm

During the entity matching process, a blocking task is needed to group the most likely matching entities into blocks, which reduces the search space for entity matching from the Cartesian product to a small subset. Peter Christen proposes a survey that describes and evaluates several blocking methods [Chr12] namely: (1) Standard Blocking, (2) Sorted Neighborhood, (3) Q-gram Indexing, (4) Suffix-Array Based, (5) Canopy Clustering and (6) String-Map Based. All of these methods use terminological common similarity measures to find the possible matches. Our proposed matching approach considers only a spatial blocking algorithm using a spatial similarity measure. An interesting research aspect is to study the possibility and the performance of using spatial similarity measure with existing blocking algorithms. Also, we can study and evaluate the combination of existing blocking algorithms with spatial blocking in order to produce a better clustering

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

of entities to be matched, which may consequently improves the performance of matching approaches.

7.3.3 Visualization and Navigation

An interesting research aspect is to consider the users' profiles and device types (e.g., computer or smart phone) for the visual representation and the navigation process in multi-provider LBS. For example, Robal et al. have developed domain user profiles ontology for web usage that could be adapted for LBS applications too [RHK07]. Authors provide a method for logging user preferences input via a graphical user interface (GUI) and classify them via the ontology's reasoner into concepts that can be used to improve the results of LBS queries. For instance, if users indicate in the GUI that they do not have a car, then the ontology's reasoner would create a concept that these users need public transportation. Hence, once users search for POIs, we can classify or highlight the POIs depending on their closeness to bus or metro stations, which would help facilitate a users choice.

7.4 Final Words

The number of LBS is increasing daily and the integration of such services can improve the completeness and correctness of information offered to users. This thesis is a contribution in this field applied to a touristic use case. The research on this issue is carried out within the framework of a multidisciplinary project. This work could not have been achieved without the collaboration of partners in Informatics, Geomatics and Human and Social Sciences, and the practitioners from tourist offices.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

Appendix A

Probability-based Combination of Individual similarities

The Global Similarity (GS) consists of combining the individual similarities of two compared entities. According to the requirements of GS (see Section 5.3 in Chapter 5), it must combine the individual similarities of two compared entities without any considerations to the neighbor entities, i.e. without any context-depending. On this base, we intend to propose a method namely $\mathrm{GS_{pr}}$ that combines several invidual similarities using probability concept. Let X be the event that e and e' entities correspond to each other with $\Omega = \{\mathrm{True}, \mathrm{False}\}$. We denote a_i a pair of corresponding attributes (e.g., "POI name" vs. "place_name") between e and e'. For each a_i that do not have a Missing Data difference, we calculate an individual similarity s_i using simple or hybrid similarity measure that returns a numeric value between 0 and 1. Firstly, we impose p as the *a-priori* probability of X to be True. Naturally, if we have no knowledge of the *a-priori* probability of X, then we assume symmetry between True and False, i.e. $p = 1/2$. Secondly, suppose that each of the n similarities is a probability, denoted $Pr(a_i) = s_i$, that predict the truthfulness of X. For example, if a_i is a pair of corresponding attributes describing the POI name, then $Pr(a_i)$ is the probability that X is true given the names of two entities are similar. We assume that these probabilities are pair wise independent (i.e. $Pr(a_i \cap a_j) = Pr(a_i) \times Pr(a_j)$) because they measure the similarities of no correlated attributes (i.e. independent information). For instance, the similarity between two POI names remains unaffected regardless of the similarities of locations or phone numbers. This means that each probability is just as likely to be right regardless of whether the other is right or wrong. Using these n probabilities, we calculate one global combined probability [com16]. This latter is considered as the global similarity that helps us in deciding whether the two entities do correspond. To do so, we need to calculate the probability that X is true given the similarities of $a_1, a_2, ...$ and

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

a_n. Based on the stated assumptions, the combined probability is given by the following formula:

$$Pr(\text{X = True given } n \text{ Similarities}) = \frac{p.Pr(a_1 \cap ... \cap a_n)}{p.Pr(a_1 \cap ... \cap a_n) + (1-p).Pr(\overline{a_1} \cap ... \cap \overline{a_n})} = \frac{p.\prod_{i=1}^{n} Pr(a_i)}{p.\prod_{i=1}^{n} Pr(a_i) + (1-p).\prod_{i=1}^{n}(1-Pr(a_i))}$$

We denote a defined probability when the similarity of two corresponding attributes equals 0 or 1 (i.e. $Pr(a_i) = 0$ or $Pr(a_i) = 1$). A specific case is distinguished, if the set of n similarities contains both defined probability (i.e. 0 and 1), then the above formula is unable to calculate a combined probability because there are two defined and opposed probabilities. To resolve the issue of this case, we set the combined probability to the *a-priori* p value.

Definition A.1. Let $\mathrm{GS_{pr}}$ be the probability-based combination of n independent similarities between two entities:

$$\mathrm{GS_{Pr}}: \mathbb{E} \times \mathbb{E} \to [0,1]$$

$$(e, e') \to \mathrm{GS_{Pr}}(s_1, ..., s_n) = \begin{cases} p & \text{if } \exists s_i, s'_i \backslash s_i = 0 \wedge s'_i = 1 \\ \frac{p.\prod_{i=1}^{n} s_i}{p.\prod_{i=1}^{n} s_i + (1-p).\prod_{i=1}^{n}(1-s_i)} & \text{Otherwise} \end{cases}$$

where p is the *a-priori* probability that e and e' are corresponding entities.

On this base, if all independent probabilities are greater than p value, then $\mathrm{GS_{Pr}}$ pulls the global similarity to 1. Conversely, if all independent probabilities are lesser than p value, then $\mathrm{GS_{Pr}}$ pulls the global similarity to 0. Otherwise, if the independent probabilities are distributed above and below p value, then $\mathrm{GS_{Pr}}$ results a trade-off.

Consider the following example, let e and e' be two entities. The similarity between their locations, denoted s_1, equals 0.75 and the similarity between their names, denoted s_2, equals 0.6. Assuming that *a-priori* probability of correspondence equals $p = 0.5$, then the Global Similarity between e and e' equals to the probability that e and e' are two corresponding entities given s_1 and s_2:

$$\mathrm{GS}(e, e') = \frac{0.5 \times 0.75 \times 0.6}{0.5 \times 0.75 \times 0.6 + (1-0.5) \times (1-0.75) \times (1-0.6)} = 0.81$$

Notice that the resulting global similarity is not equal to the *a-priori* value of 0.5 that we assumed by imposing symmetry between True and False. Now, suppose that we have one additional similarity between the phone numbers of e and e', denoted s_3, that equals

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

0.3. The Global Similarity of e and e' becomes:

$$\mathrm{GS}(e, e') = \frac{0.5 \times 0.75 \times 0.6 \times 0.3}{0.5 \times 0.75 \times 0.6 \times 0.3 + (1 - 0.5) \times (1 - 0.75) \times (1 - 0.6) \times (1 - 0.3)} = 0.65$$

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

Bibliography

[ACH+91] Esther M. Arkin, L. Paul Chew, Daniel P. Huttenlocher, Klara Kedem, and Joseph S. B. Mitchell. An efficiently computable metric for comparing polygonal shapes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 13(3):209–216, 1991.

[ADMR05] David Aumueller, Hong Hai Do, Sabine Massmann, and Erhard Rahm. Schema and ontology matching with COMA++. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 906–908, 2005.

[ARS09] Arvind Arasu, Christopher Ré, and Dan Suciu. Large-scale deduplication with constraints using dedupalog. In *Proceedings of the 25th International Conference on Data Engineering, ICDE 2009, March 29 2009 - April 2 2009, Shanghai, China*, pages 952–963, 2009.

[ATV08] B. Alexe, W. C. Tan, and Y. Velegrakis. Stbenchmark: towards a benchmark for mapping systems. *Proceedings of the VLDB*, 1(1):230–244, 2008.

[BBR11] Zohra Bellahsene, Angela Bonifati, and Erhard Rahm. *Schema Matching and Mapping*. Springer-Verlag, Heidelberg, Germany, 2011. ISBN: 978-3-642-16517-7.

[BBS05] Mikhail Bilenko, Sugato Basu, and Mehran Sahami. Adaptive product normalization: Using online learning for record linkage in comparison shopping. In *Proceedings of the 5th IEEE International Conference on Data Mining (ICDM 2005), 27-30 November 2005, Houston, Texas, USA*, pages 58–65, 2005.

[BCD+14] Bilal Berjawi, Elisabeth Chesneau, Fabien Duchateau, Franck Favetta, Claire Cunty, Maryvonne Miquel, and Robert Laurini. Representing uncertainty in visual integration. In *The 20th International Conference on Distributed Multimedia Systems: Research papers on distributed multimedia systems, distance education technologies and visual languages and*

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

computing, Pittsburgh, PA, USA, August 27-29, 2014., pages 365–371, 2014.

[BDC+14] Bilal Berjawi, Fabien Duchateau, Elisabeth Chesneau, Franck Favetta, Geoffrey Seccia, Claire Cunty, Maryvonne Miquel, and Robert Laurini. Uncertainty visualization of multi-providers cartographic integration. *Journal of Visual Languages and Computing*, 25(6):995–1002, 2014.

[BDF+15] Bilal Berjawi, Fabien Duchateau, Franck Favetta, Maryvonne Miquel, and Robert Laurini. PABench: Designing a Taxonomy and Implementing a Benchmark for Spatial Entity Matching. In *The Seventh International Conference on Advanced Geographic Information Systems, Applications, and Services*, pages 7–16, Lisbon, Portugal, February 2015.

[BDK+05] Catriel Beeri, Yerach Doytsher, Yaron Kanza, Eliyahu Safra, and Yehoshua Sagiv. Finding corresponding objects when integrating several geo-spatial datasets. In *ACM International Workshop on Geographic Information Systems*, pages 87–96, 2005.

[Ber83] Jacques Bertin. Semiology of graphics: diagrams, networks, maps. 1983.

[Ber98] Nicole Berthier. Les techniques denquête, méthode et exercices corrigés. *Armand Colin, Paris*, 1998.

[BG11] Indrajit Bhattacharya and Lise Getoor. Query-time entity resolution. *CoRR*, abs/1111.0045, 2011.

[BKM06] Mikhail Bilenko, Beena Kamath, and Raymond J. Mooney. Adaptive blocking: Learning to scale up record linkage. In *Proceedings of the 6th IEEE International Conference on Data Mining (ICDM 2006), 18-22 December 2006, Hong Kong, China*, pages 87–96, 2006.

[BKS93] Thomas Brinkhoff, Hans-Peter Kriegel, and Ralf Schneider. Comparison of approximations of complex objects used for approximation-based query processing in spatial database systems. In *Proceedings of the Ninth International Conference on Data Engineering, April 19-23, 1993, Vienna, Austria*, pages 40–49, 1993.

[BKSS04] Catriel Beeri, Yaron Kanza, Eliyahu Safra, and Yehoshua Sagiv. Object fusion in geographic information systems. In *(e)Proceedings of the Thirtieth International Conference on Very Large Data Bases, Toronto, Canada, August 31 - September 3 2004*, pages 816–827, 2004.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

[BM03] Mikhail Bilenko and Raymond J. Mooney. Adaptive duplicate detection using learnable string similarity measures. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 39–48, 2003.

[BMR11] Philip A. Bernstein, Jayant Madhavan, and Erhard Rahm. Generic schema matching, ten years later. *PVLDB*, 4(11):695–701, 2011.

[BN08] Jens Bleiholder and Felix Naumann. Data fusion. *ACM Computing Surveys*, 41(1):1:1–1:41, 2008.

[BP12] Sergey Brin and Lawrence Page. Reprint of: The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, 56(18):3825–3833, 2012.

[Cas06] Mara Luisa Casado. Some basic mathematical constraints for the geometric conflation problem. In *Proceedings of the 7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*, pages 264–274, Lisboa, Instituto Geogrfico Portugus, 2006.

[Cat15] Léonor Ferrer Catala. Évaluation des usages des services cartographiques en ligne et de la représentation de l'incertitude de points d'intérêts touristiques. Technical report, Université Lumière Lyon 2, June 2015. Available at https://liris.cnrs.fr/unimap/download/Memoire-de-stage-FERRER-CATALA-L.pdf.

[CD99] Jean-Paul Chilès and Pierre Delfiner. *Geostatistics: modeling spatial uncertainty*. Wiley series in probability and statistics. Wiley, New York, 1999.

[CFC+16] Léonor Ferrer Catala, Franck Favetta, Claire Cunty, Bilal Berjawi, Fabien Duchateau, Maryvonne Miquel, and Robert Laurini. Visualizing integration uncertainty enhances user's choice in multi-providers integrated maps. In *Proceedings of the International Working Conference on Advanced Visual Interfaces, AVI 2016, Bari, Italy, June 7-10, 2016*, pages 298–299, 2016.

[Chr08] Peter Christen. Febrl: A freely available record linkage system with a graphical user interface. In *Proceedings of the Second Australasian Workshop on Health Data and Knowledge Management - Volume 80*, HDKM '08, pages 17–25, Darlinghurst, Australia, Australia, 2008. Australian Computer Society, Inc.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

[Chr12] Peter Christen. A survey of indexing techniques for scalable record linkage and deduplication. *IEEE Trans. Knowl. Data Eng.*, 24(9):1537–1555, 2012.

[Col13] Maxime Colomb. Étude des mécanismes de repérage en milieu urbain à travers la prise de points de repères géographiques et du nommage des lieux de la ville, applique à saint etienne. Technical report, Université Jean Monnet de Saint-Étienne, Septembre 2013. Available at http://liris.cnrs.fr/unimap/download/Maxime2.pdf.

[com16] The combination of probabilities, 2016. Available at http://www.mathpages.com/home/kmath2x6x7/kmath2x6x7.htm.

[CRF03] William Cohen, Pradeep Ravikumar, and Stephen Fienberg. A comparison of string metrics for matching names and records. In *KDD Workshop on Data Cleaning and Object Consolidation*, volume 3, pages 73–78, 2003.

[CTKS03] Ching-Chien Chen, Snehal Thakkar, Craig A. Knoblock, and Cyrus Shahabi. Automatically annotating and integrating spatial datasets. In *Advances in Spatial and Temporal Databases*, pages 469–488, 2003.

[DB14] Fabien Duchateau and Zohra Bellahsene. Designing a benchmark for the assessment of schema matching tools. In *Open Journal of Databases (OJDB)*, volume 1, pages 3–25. RonPub, Germany, 2014.

[DBHS10] Xin Dong, Laure Berti-Equille, Yifan Hu, and Divesh Srivastava. Global detection of complex copying relationships between sources. *PVLDB*, 3(1):1358–1369, 2010.

[DBS09] Xin Luna Dong, Laure Berti-Equille, and Divesh Srivastava. Integrating conflicting data: The role of source dependence. *PVLDB*, 2(1):550–561, 2009.

[DDH01] AnHai Doan, Pedro M. Domingos, and Alon Y. Halevy. Reconciling schemas of disparate data sources: A machine-learning approach. In *Proceedings of the 2001 ACM SIGMOD international conference on Management of data, Santa Barbara, CA, USA, May 21-24, 2001*, pages 509–520, 2001.

[DDL00] AnHai Doan, Pedro M. Domingos, and Alon Y. Levy. Learning source description for data integration. In *WebDB (Informal Proceedings)*, pages 81–86, 2000.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

[Dev97] Thomas Devogele. *Processus d'intégration et d'appariement de Bases de Données Géographiques; Application à une base de données routières multi-échelles. (Integration and Matching processes for geographic databases; application to road-network multi-scale databases)*. PhD thesis, Versailles Saint-Quentin-en-Yvelines University, France, 1997.

[DGH+14] Xin Luna Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Kevin Murphy, Shaohua Sun, and Wei Zhang. From data fusion to knowledge fusion. *PVLDB*, 7(10):881–892, 2014.

[DHM05] Xin Dong, Alon Y. Halevy, and Jayant Madhavan. Reference reconciliation in complex information spaces. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, Baltimore, Maryland, USA, June 14-16, 2005*, pages 85–96, 2005.

[DN09] Xin Luna Dong and Felix Naumann. Data fusion - resolving data conflicts for integration. *PVLDB*, 2(2):1654–1655, 2009.

[DORB14] Nilesh N. Dalvi, Marian Olteanu, Manish Raghavan, and Philip Bohannon. Deduplicating a places database. In *23rd International World Wide Web Conference, WWW '14, Seoul, Republic of Korea, April 7-11, 2014*, pages 409–418, 2014.

[Doy00] Yerahmiel Doytsher. A rubber sheeting algorithm for non-rectangular maps. *Computers & Geosciences*, 26(9):1001–1010, 2000.

[DPS98] Thomas Devogele, Christine Parent, and Stefano Spaccapietra. On spatial database integration. *International Journal of Geographical Information Science*, 12(4):335–352, 1998.

[DR02] Hong Hai Do and Erhard Rahm. Coma - a system for flexible combination of schema matching approaches. In *Proceedings of 28th International Conference on Very Large Data Bases (VLDB)*, pages 610–621, 2002.

[DSS12] Xin Luna Dong, Barna Saha, and Divesh Srivastava. Less is more: Selecting sources wisely for integration. *PVLDB*, 6(2):37–48, 2012.

[EEV02] Mohamed G. Elfeky, Ahmed K. Elmagarmid, and Vassilios S. Verykios. Tailor: A record linkage tool box. In *Proceedings of the 18th International Conference on Data Engineering, San Jose, CA, USA, February 26 - March 1, 2002*, pages 17–28, 2002.

[ERT13] Jérôme Euzenat, Maria-Elena Rosoiu, and Cssia Trojahn. Ontology matching benchmarks: generation, stability, and discriminability. *Web*

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

Semantics: Science, Services and Agents on the World Wide Web, 21, 2013.

[ES07] Jérôme Euzenat and Pavel Shvaiko. *Ontology matching*. Springer-Verlag, Heidelberg, Germany, 2007. ISBN: 3-540-49611-4.

[Eva97] Beverley J. Evans. Exploratory cartograpic visualisation dynamic display of spatial data-reliability: Does it benefit the map user? *Computers & Geosciences*, 23(4):409 – 422, 1997.

[Fis25] Ronald A. Fisher. *Statistical methods for research workers*. Edinburgh Oliver & Boyd, 1925.

[GE00] Mark Gahegen and Manfred Ehlers. A framework for the modeling of uncertainty between remote sensing and geographic information. *ISPRS Journal of Photogrammetry and Remote Sensing*, 55:176–188, 2000.

[GLM04] Nils Gesbert, Thérèse Libourel, and Sébastien Mustière. Apport des spécifications pour les modèles de bases de données géographiques. *Revue Internationale de Géomatique*, 14(2):239–257, 2004.

[GRB03] Dominique Gruyer, Cyril Royère, and Véronique Berge-Cherfaoui. Utilisation de la théorie des croyances pour l'association dans le suivi de cibles. *Technique et Science Informatiques*, 22(7-8):983–1010, 2003.

[Gup08] Stephen C. Guptill. *Fundamentals of Spatial Data Quality* - edited by RODOLPHE DEVILLERS AND ROBERT JEANSOULIN. *Transactions in GIS*, 12(1):161–162, 2008.

[GZK03] Matej Gombosoi, Borut Zalik, and Sebastian Krivograd. Comparing two sets of polygons. *International Journal of Geographical Information Science*, 17(5):431–443, 2003.

[HRO06] Alon Y. Halevy, Anand Rajaraman, and Joann J. Ordille. Data integration: The teenage years. In *Proceedings of the 32nd International Conference on Very Large Data Bases, Seoul, Korea, September 12-15, 2006*, pages 9–16, 2006.

[IRV13] Ekaterini Ioannou, Nataliya Rassadko, and Yannis Velegrakis. On generating benchmark data for entity matching. *Journal on Data Semantics*, 2(1):37–56, 2013.

[JdB64] Fernand Joly and Stphane de Brommer. Projet de normalisation de symboles de cartes thmatiques. association cartographique internationale,

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

symposium technique 524 p. *Communication enregistre l'U.G.I. sous le n1073, Bulletin du Comit Franais de Cartographie, Institut gographique national*, page 30, 1964.

[Jon72] Karen Sprck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21, 1972.

[Kar11] Roula Karam. *Multi-providers location based services for mobile-tourism: a use case for location and cartographic integrations on mobile devices.* PhD thesis, LIRIS INSA, Lyon, 2011.

[KFKL10] Roula Karam, Franck Favetta, Rima Kilany, and Robert Laurini. Integration of similar location based services proposed by several providers. In *Networked Digital Technologies*, pages 136–144, 2010.

[KKH+10] Toralf Kirsten, Lars Kolb, Michael Hartung, Anika Groß, Hanna Köpcke, and Erhard Rahm. Data partitioning for parallel entity matching. *The Computing Research Repository (CoRR)*, abs/1006.5309, 2010.

[KR08] Hanna Köpcke and Erhard Rahm. Training selection for tuning entity matching. In *Proceedings of the International Workshop on Quality in Databases and Management of Uncertain Data, Auckland, New Zealand, August 2008*, pages 3–12, 2008.

[KR10] Hanna Köpcke and Erhard Rahm. Frameworks for entity matching: A comparison. *Data & Knowledge Engineering*, 69:197–210, February 2010.

[KSA+12] Craig A. Knoblock, Pedro A. Szekely, José Luis Ambite, Aman Goel, Shubham Gupta, Kristina Lerman, Maria Muslea, Mohsen Taheriyan, and Parag Mallick. Semi-automatically mapping structured sources into the semantic web. In *The Semantic Web: Research and Applications - 9th Extended Semantic Web Conference, ESWC 2012, Heraklion, Crete, Greece, May 27-31, 2012. Proceedings*, pages 375–390, 2012.

[KSG07] Hyunmo Kang, Vivek Sehgal, and Lise Getoor. Geoddupe: A novel interface for interactive entity resolution in geospatial data. In *International Conference on Information Visualisation*, pages 489–496, 2007.

[KTR09] Hanna Köpcke, Andreas Thor, and Erhard Rahm. Comparative evaluation of entity resolution approaches with FEVER. *PVLDB*, 2(2):1574–1577, 2009.

[KTR10] Hanna Köpcke, Andreas Thor, and Erhard Rahm. Evaluation of entity resolution approaches on real-world match problems. *PVLDB*, 3(1):484–493, 2010.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

[KW52] William H. Kruskal and Allen W. Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621, 1952.

[Lau01] Robert Laurini. *Information systems for urban planning - a hypermedia co-operative approach*. Taylor & Francis, 2001.

[LDL+12] Xian Li, Xin Luna Dong, Kenneth Lyons, Weiyi Meng, and Divesh Srivastava. Truth finding on the deep web: Is the problem solved? *PVLDB*, 6(2):97–108, 2012.

[LDOS11] Xuan Liu, Xin Luna Dong, Beng Chin Ooi, and Divesh Srivastava. Online data fusion. *PVLDB*, 4(11):932–943, 2011.

[Lev66] V.I. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966.

[LSPP14] George Lamprianidis, Dimitrios Skoutas, George Papatheodorou, and Dieter Pfoser. Extraction, integration and analysis of crowdsourced points of interest from multiple web sources. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information*, GeoCrowd '14, pages 16–23, New York, NY, USA, 2014. ACM.

[LT92] Robert Laurini and Derek Thompson. *Fundamentals of spatial information systems*. The A.P.I.C. series. Academic press, London, San Diego, New York, 1992.

[Mac95] Alan M. MacEachren. How maps work. *New York, London: The Guilford Press*, 1995.

[Mac04] Alan M. MacEachren. *How Maps Work - Representation, Visualization, and Design*. Guilford Press, 2004.

[MAK10] Srivatsa Maddodi, Girija V. Attigeri, and A. K. Karunakar. Data deduplication techniques and analysis. In *3rd International Conference on Emerging Trends in Engineering and Technology, ICETET 2010, Goa, India, November 19-21, 2010*, pages 664–668, 2010.

[Mas06] Atsushi Masuyama. Methods for detecting apparent differences between spatial tessellations at different time points. *International Journal of Geographical Information Science*, 20(6):633–648, 2006.

[MBR01] Jayant Madhavan, Philip A. Bernstein, and Erhard Rahm. Generic schema matching with cupid. In *VLDB*, pages 49–58, 2001.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

[MJA13] Grant McKenzie, Krzysztof Janowicz, and Benjamin Adams. Weighted multi-attribute matching of user-generated points of interest. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL'13, pages 440–443, New York, NY, USA, 2013. ACM.

[MMBD14] Anthony Morana, Thomas Morel, Bilal Berjawi, and Fabien Duchateau. Geobench: a geospatial integration tool for building a spatial entity matching benchmark. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Dallas/Fort Worth, TX, USA, November 4-7, 2014*, pages 533–536, 2014.

[Mor74] Joel L. Morrison. A theoretical framework for cartographic generalization with the emphasis on the process of symbolization. *International Yearbook of Cartography*, 14(1974):115–27, 1974.

[MRANIG14] Ludovic Moncla, Walter Renteria-Agualimpia, Javier Nogueras-Iso, and Mauro Gaio. Geocoding for texts with fine-grain toponyms: An experiment on a geoparsed hiking descriptions corpus. In *Proceedings of the 22Nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '14, pages 183–192, New York, NY, USA, 2014. ACM.

[MRH+05] Alan M. MacEachren, Anthony Robinson, Susan Hopper, Steven Gardner, Robert Murray, Mark Gahegan, and Elisabeth Hetzler. Visualizing geospatial information uncertainty: What we know and what we need to know. *Cartography and Geographic Information Science*, 32(3):139–160, 2005.

[MRO+12] Alan M. MacEachren, Robert E. Roth, James O'Brien, Bonan Li, Derek Swingley, and Mark Gahegan. Visual semiotics & uncertainty visualization: An empirical study. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2496–2505, 2012.

[MSW00] Michael Minami, J Sakala, and J Wrightsell. Using arcmap. *Environmental Systems Research Institute, Inc., Redlands, California, USA*, 2000.

[MT94] A.M. MacEachren and D.R.F. Taylor. *Visualization in Modern Cartography*. Modern Cartography; 2 Dissertation Series; 10. Elsevier Science & Technology Books, 1994.

[MU49] Nicholas Metropolis and Stan Ulam. The monte carlo method. *Journal of the American statistical association*, 44(247):335–341, 1949.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

[MW47] Henry B. Mann and Donald R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60, 1947.

[NO03] Risto Näsänen and Helena Ojanpää. Effect of image contrast and sharpness on visual search for computer icons. *Displays*, 24(3):137–144, 2003.

[OR07] Ana-Maria Olteanu-Raimond. A multi-criteria fusion approach for geographical data matching. In *5th International Symposium on Spatial Data Quality*, June 2007.

[Org01] World Tourism Organisation. *Tourism Signs and Symbols: A Status Report and Guidebook*. Published by World Tourism Organization, 2001.

[PD96] Jignesh M. Patel and David J. DeWitt. Partition based spatial-merge join. In *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, Montreal, Quebec, Canada, June 4-6, 1996.*, pages 259–270, 1996.

[Pla83] Robin L. Plackett. Karl Pearson and the Chi-Squared Test. *International Statistical Review / Revue Internationale de Statistique*, 51(1):59–72, 1983.

[POI13] Points of interest standards working group. Technical report, Open Geospatial Consortium, 2013. Available at https://portal.opengeospatial.org/files/?artifact_id=54800.

[PR10] Jeff Pasternack and Dan Roth. Knowing what to believe (when you already know something). In *COLING 2010, 23rd International Conference on Computational Linguistics, Proceedings of the Conference, 23-27 August 2010, Beijing, China*, pages 877–885, 2010.

[PS00] Christine Parent and Stefano Spaccapietra. Database integration: The key to data interoperability. In *Advances in Object-Oriented Data Modeling*, pages 221–253. 2000.

[PSAR09] Rafael Ponce-Medellin, Juan Gabriel González Serna, Rocio Vargas A., and Lirio Ruiz. Technology integration around the geographic information: A state of the art. *CoRR*, abs/0911.0909, 2009.

[PZLL11] Christelle Pierkot, Esteban Zimányi, Yuan Lin, and Thérèse Libourel. Advocacy for external quality in GIS. In *GeoSpatial Semantics - 4th International Conference, GeoS 2011, Brest, France, May 12-13, 2011. Proceedings*, pages 151–165, 2011.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

[RAUB11] Juan J. Ruiz, F. Javier Ariza, Manuel A. Ureña, and Elidia B. Blázquez. Digital map conflation: a review of the process and a proposal for classification. *International Journal of Geographical Information Science*, 25(9):1439–1466, 2011.

[RB01] Erhard Rahm and Philip A. Bernstein. A survey of approaches to automatic schema matching. *VLDB J.*, 10(4):334–350, 2001.

[RHK07] Tarmo Robal, Hele-Mai Haav, and Ahto Kalja. Making web users' domain models explicit by applying ontologies. In *Advances in Conceptual Modeling - Foundations and Applications, ER 2007 Workshops CMLSA, FP-UML, ONISW, QoIS, RIGiM,SeCoGIS, Auckland, New Zealand, November 5-9, 2007, Proceedings*, pages 170–179, 2007.

[Rij79] C. J. Van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, Newton, Massachusetts, USA, 2nd edition, 1979. ISBN: 0408709294.

[RLA+15] Faizan Ur Rehman, Ahmed Lbath, Imad Afyouni, Abdullah Murad, Md. Abdur Rahman, Bilal Sadiq, Akhlaq Ahmad, and Saleh M. Basalamah. Semantic multimedia-enhanced spatio-temporal queries in a crowdsourced environment. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems, Bellevue, WA, USA, November 3-6, 2015*, pages 100:1–100:4, 2015.

[RLM+15] Faizan Ur Rehman, Ahmed Lbath, Abdullah Murad, Md. Abdur Rahman, Bilal Sadiq, Akhlaq Ahmad, Ahmad M. Qamar, and Saleh M. Basalamah. A semantic geo-tagged multimedia-based routing in a crowdsourced big data environment. In *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference, MM '15, Brisbane, Australia, October 26 - 30, 2015*, pages 759–760, 2015.

[RM08] Ana-Maria Olteanu Raimond and Sébastien Mustière. Data matching - a matter of belief. In *International Symposium on Spatial Data Handling (SDH)*, pages 501–519, 2008.

[Saa85] Alan Saalfeld. A fast rubber-sheeting transformation using simplicial coordinates. *The American Cartographer*, 12(2):169–173, 1985.

[Saa88] Alan Saalfeld. Conflation automated map compilation. *International Journal of Geographical Information System*, 2(3):217–228, 1988.

[SCC+14] Geoffrey Seccia, Claire Cunty, Elisabeth Chesneau, Bilal Berjawi, and Franck Favetta. Évaluer des modes de représentation cartographique de

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

l'incertitude. In *International Conference on Spatial Analysis and GEOmatics (MAGIS SAGEO'2014)*, page 5, November 2014.

[SE90] J. Star and J.E. Estes. *Geographic Information Systems: An Introduction.* Prentice Hall series in geographic information science. Prentice Hall, 1990.

[Sec14] Geoffrey Seccia. Élaboration d'un protocole de test pour l'évaluation de représentation cartographiques relatives à l'incertitude de points d'intérêt touristiques. Technical report, Université Lumière Lyon 2, June 2014. Available at https://liris.cnrs.fr/unimap/download/Master-report-juin-2014-Geoffrey.pdf.

[SGV06] Vivek Sehgal, Lise Getoor, and Peter Viechnicki. Entity resolution in geospatial data integration. In *ACM International Symposium on Geographic Information Systems*, pages 83–90, 2006.

[Sha76] Glenn Shafer. *A mathematical theory of evidence*, volume 1. Princeton university press Princeton, 1976.

[Shn96] Ben Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the 1996 IEEE Symposium on Visual Languages, Boulder, Colorado, USA, September 3-6, 1996*, pages 336–343, 1996.

[SIN92] François DE SINGLY. Lenquête et ses méthodes: le questionnaire. *Paris: Nathan*, 128, 1992.

[SKS+10] Eliyahu Safra, Yaron Kanza, Yehoshua Sagiv, Catriel Beeri, and Yerach Doytsher. Location-based algorithms for finding sets of corresponding objects over several geo-spatial data sets. *International Journal of Geographical Information Science*, 24(1):69–106, 2010.

[SKSD06] Eliyahu Safra, Yaron Kanza, Yehoshua Sagiv, and Yerach Doytsher. Integrating data from maps on the world-wide web. In *Proceedings of the 6th InternationalSymposium on Web and Wireless Geographical Information Systems, W2GIS 2006, Hong Kong, China, December 4-5, 2006,*, pages 180–191, 2006.

[SSL12] Tatjana Scheffler, Rafael Schirru, and Paul Lehmann. Matching points of interest from different social networking sites. In *KI 2012: Advances in Artificial Intelligence - 35th Annual German Conference on AI, Saarbrücken, Germany, September 24-27, 2012. Proceedings*, pages 245–248, 2012.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

[Stu08] Student. The probable error of a mean. *Biometrika*, pages 1–25, 1908.

[SV04] Jochen H. Schiller and Agnès Voisard. *Location-Based Services*. Morgan Kaufmann, 2004.

[SW65] Samuel Sanford Shapiro and Martin B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965.

[SWM+02] Barry Schwartz, Andrew Ward, John Monterosso, Sonja Lyubomirsky, Katherine White, and Darrin R. Lehman. Maximizing versus satisficing: Happiness is a matter of choice. *Journal of Personality and Social Psychology*, 83(5):1178–1197, 2002.

[SX08a] Shashi Shekhar and Hui Xiong. Euclidean distance. In *Encyclopedia of GIS*, page 291. 2008.

[SX08b] Shashi Shekhar and Hui Xiong. Manhattan distance. In *Encyclopedia of GIS*, page 631. 2008.

[THM+05] Judi Thomson, Elizabeth Hetzler, Alan MacEachren, Mark Gahegan, and Misha Pavel. A typology for visualizing uncertainty. In *Proceedings of SPIE Conference on Visualization and Data Analysis*, volume 5669, pages 146–157, 2005.

[TKA07] Snehal Thakkar, Craig A. Knoblock, and José Luis Ambite. Quality-driven geospatial data integration. In *ACM International Symposium on Geographic Information Systems, Washington, USA, 7-9 November*, page 16, 2007.

[TKM02] Sheila Tejada, Craig A. Knoblock, and Steven Minton. Learning domain-independent string transformation weights for high accuracy object identification. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 350–359, 2002.

[TR07] Andreas Thor and Erhard Rahm. Moma - A mapping-based object matching system. In *Third Biennial Conference on Innovative Data Systems Research (Online Proceedings)*, pages 247–258, 2007.

[Ull77] Julian R. Ullmann. A binary n-gram technique for automatic correction of substitution, deletion, insertion and reversal errors in words. *Comput. J.*, 20(2):141–147, 1977.

[vE00] Corné van Elzakker. Use and users of maps on the web. *Cartographic Perspectives*, (37):34–50, 2000.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

[Vol06] Steffen Volz. An iterative approach for matching multiple representations of street data. In *Proceedings of the JOINT ISPRS Workshop on Multiple Representations and Interoperability of Spatial Data, Hannover*, pages 101–110, 2006.

[WMK+09] Steven Euijong Whang, David Menestrina, Georgia Koutrika, Martin Theobald, and Hector Garcia-Molina. Entity resolution with iterative blocking. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2009, Providence, Rhode Island, USA, June 29 - July 2, 2009*, pages 219–232, 2009.

[XDZ09] Kexin Xie, Ke Deng, and Xiaofang Zhou. From trajectories to activities: a spatio-temporal join approach. In *Proceedings of the 2009 International Workshop on Location Based Social Networks*, pages 25–32. ACM, 2009.

[XFC12] Sen Xu, Soren Flexner, and Vitor Carvalho. Geocoding billions of addresses: Toward a spatial record linkage system with big data. *GIScience in the Big Data Age*, 13, 2012.

[Zai10] Katrin Simone Zaiß. *Instance-based ontology matching and the evaluation of matching systems*. PhD thesis, Heinrich Heine University Düsseldorf, 2010.

[ZCB87] Rami Zwick, Edward Carlstein, and David V. Budescu. Measures of similarity among fuzzy concepts: A comparative analysis. *Int. J. Approx. Reasoning*, 1(2):221–242, 1987.

[ZCSK13] Ying Zhang, Yao-Yi Chiang, Pedro Szekely, and Craig A. Knoblock. A semantic approach to retrieving, linking, and integrating heterogeneous geospatial data. In *Joint Proceedings of the Workshop on AI Problems and Approaches for Intelligent Environments and Workshop on Semantic Cities*, AIIP '13, pages 31–37, New York, NY, USA, 2013. ACM.

[ZCV10] Katrin Zaiss, Stefan Conrad, and Sven Vater. A benchmark for testing instance-based ontology matching methods. In *Proc. EKAW Posters and Demos*, 2010.

[ZSM05] Meng Zhang, Wei Shi, and Liqiu Meng. A generic matching algorithm for line networks of different resolutions. In *Workshop of ICA Commission on Generalization and Multiple Representation Computering Faculty of A Coruña University-Campus de Elviña, Spain*, 2005.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

Résumé long en français

Services de Géolocalisation

Un Système d'Information Géographique (SIG) est une application informatique conçue pour réaliser une large gamme d'opérations sur l'information géographique [SE90, LT92]. L'information géographique peut représenter un point quelconque sur le globe et a de multiples descriptions. Ainsi, un SIG a pour fonction de capturer, stocker, manipuler, analyser, visualiser et exporter toutes sortes d'informations géographiques. Ces dernières années, les evolutions dans le domaine de la technologie de l'information, ont amené une prolifération de nouveaux produits tels que les services de géolocalisation (LBS) [SV04]. Ces derniers sont des services informatiques issus du SIG et capables d'offrir des services très spécialisés. Ces services fournissent généralement des informations utiles liées à une adresse ou une localisation géographique donnée via téléphone mobile, tablette ou PC de bureau. Les LBS comprennent des services permettant d'identifier l'emplacement d'une personne ou d'un objet, la navigation de voiture, le repérage des véhicules et les services météorologiques personnalisés. En outre, ils peuvent intégrer des services par exemple commerciaux sous forme de coupons ou de publicité destinés aux clients en fonction de leur emplacement actuel. Les LBS ont dû faire face à plusieurs problèmes depuis leur premier lancement, tels que la fiabilité des dispositifs GPS et les limitations des appareils mobiles.

Cette thèse ne considère que les LBS appliquées au domaine du tourisme. Elle se concentre plus particulièrement sur les LBS qui offrent des informations sur les points d'intérêt (POI), tels que la localisation du restaurant le plus proche ou la découverte de musées dans une ville donnée. Les LBS fournissent des informations utiles sur n'importe quel endroit, ce qui rend ces services de grand intérêt pour les utilisateurs et les communautés. Par exemple, les GPS sémantiques utilisent les POIs pour aider les personnes à identifier leurs emplacements pendant leur navigation [RLA+15, RLM+15]. Dans le secteur du tourisme, qui est devenu une ressource économique majeure pour de nombreux pays,

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

les fournisseurs LBS (par exemple, Google Maps[1], OpenStreetMap[2], Bing Maps[3] etc.) permettent aux touristes de rechercher, rapidement et à distance, des POIs tels que des monuments, des parcs et des hôtels.

Techniquement, les POIs sont modélisés sous la forme d'entités qui sont décrites par des informations spatiales telles que les coordonnées de localisation et les informations terminologiques comme le nom de lieu, le téléphone et le site Web. Habituellement, des outils cartographiques interactifs sont proposés par les fournisseurs de LBS pour faciliter le processus de découverte des POIs. Ces outils se composent d'une carte de base faite d'images matricielles ou objets vectoriels [Mac04]. Des légendes ou des icônes sont ensuite placées sur la carte pour afficher la localisation des POIs. Un clic sur une légende affiche une fenêtre contenant les informations terminologiques de lieu. Traditionnellement, les légendes sont conçues pour représenter les types de POI (par exemple, un parc, un lac et un centre commercial), afin que les touristes puissent facilement distinguer les endroits sur la carte.

Motivation

Au cours des dernières années, la quantité de données et le nombre de fournisseurs ont connu une croissance spectaculaire. La multiplication de l'information géographique décrivant la même réalité met en doute la validité des POIs fournis [DPS98]. Les entités spatiales faisant référence au même POI peuvent inclure des données incomplètes, incohérentes, inexactes ou même erronées d'un fournisseur à un autre. En outre, certains POIs peuvent être inclus dans la base de données d'un fournisseur mais pas dans les autres ou ils peuvent être répétés en plusieurs exemplaires dans la même base de données. Cela est dû aux différentes politiques et stratégies utilisées par les fournisseurs LBS pour construire des bases de données, mettre à jour les informations et élaborer les résultats des requêtes. En conséquence, les utilisateurs peuvent obtenir, pour la même requête, des réponses différentes en changeant d'un fournisseur à un autre. Par exemple, plusieurs fournisseurs LBS peuvent donner des emplacements différents pour le même POI. La Figure A.1 montre les résultats de la recherche de l'*Hôtel Walser* dans la ville de Courmayeur, en Italie, auprè de trois fournisseurs différents, la Figure A.1a est extraite de GoogleMaps[4] où l'hôtel est situé sur le côté droit de l'autoroute (la rue jaune), la Figure A.1b provient de Nokia Here Maps[5] où l'hôtel est situé sur le côté

[1] Google Maps: http://maps.google.comm
[2] OpenStreetMap: http://openstreetmap.org
[3] Bing Maps: http://maps.bing.com
[4] *Walser Hotel* par GoogleMaps. [Accès: Juin 2016]
[5] *Walser Hotel* par Nokia Here Maps. [Accès: Juin 2016]

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

gauche de l'autoroute (la rue rouge) et la Figure A.1c est prise de OpenStreetMap[6] où l'*hôtel Walser* est totalement absent. En outre, les incohérences peuvent affecter les attributs terminologiques. La Figure A.2a et la Figure A.2b montrent les données terminologiques du restaurant L'*Ecluse* offertes par GoogleMaps[7] et Nokia Here Maps[8] respectivement. Les deux fournisseurs offrent la même adresse et le même numéro de téléphone, mais il existe de différences syntaxiques pour le nom de lieu ("La petite écluse des Grands-Augustins" vs. "L'Ecluse") et le site web ("lecluse-restaurant-paris.fr vs "Leclusebaravin.com").



(a) GoogleMaps (b) HereMaps (c) Openstreetmaps

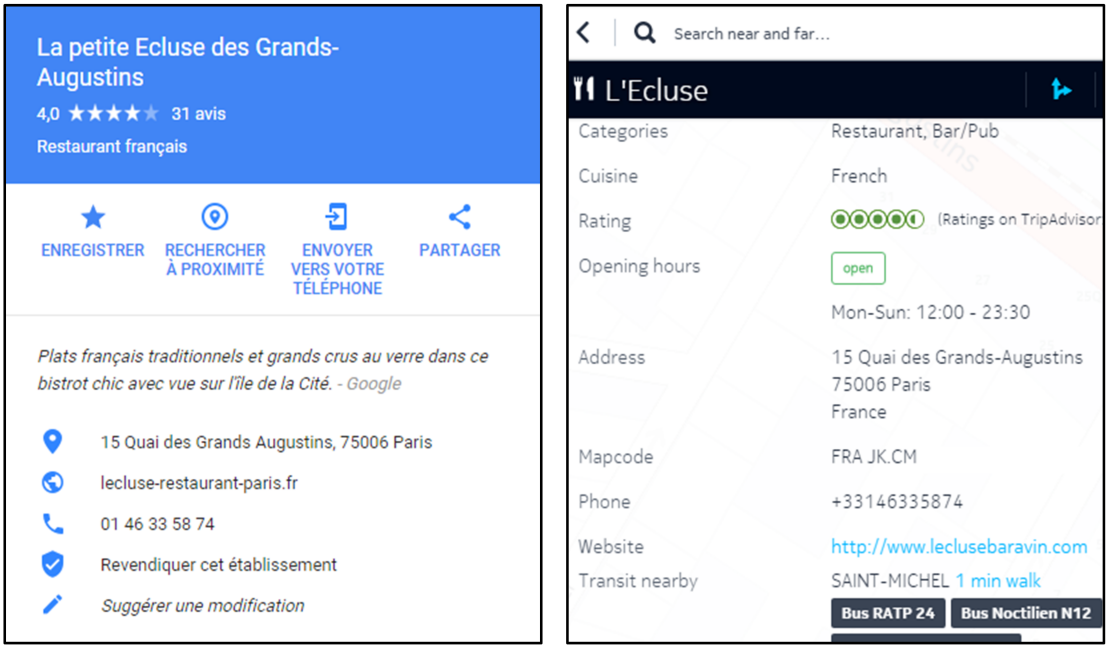
FIGURE A.1: *Hôtel Walser* dans la ville de Courmayeur, en Italie, est situé différemment par trois Fournisseurs LBS. [Accès: juin 2016]

Les exemples précédents montrent l'hétérogénéité des données géospatiales émanant de plusieurs fournisseurs LBS. L'hétérogénéité peut également affecter la représentation des POIs sur les cartes. Chaque fournisseur LBS utilise son propre ensemble de légendes pour représenter les POI. Par exemple, la Figure A.3 montre trois légendes différentes pour représenter les hôtels collectés de Share Icon[9], Icon Archive[10] et Icons DB[11], respectivement. De plus, quant à la zone réelle des lieux, les POIs pourraient être représentés d'un fournisseur à l'autre de différentes manières, par exemple le point (0D), la ligne (1D), la polyligne (2D) ou le volume (3D).

Dans le domaine de la recherche de bases de données, l'intégration des données a été largement proposée pour résoudre l'hétérogénéité des données provenant de plusieurs sources afin d'améliorer la qualité des données [HRO06]. Dans notre contexte, l'intégration de POIs sources des LBS existants permet de créer un meilleur service avec des informations plus complètes et plus précises en ce qui concerne le domaine touristique. Ceci est la thématique qu œur de cette thse. L'intégration géospatiale a été largement étudiée sous

[6] *Walser Hotel* par OpenStreetMap. [Accès: Juin 2016]
[7] *L'Ecluse Restaurant* par GoogleMaps. [Accès: Juin 2016]
[8] *L'Ecluse Restaurant* par Nokia Here Maps. [Accès: Juin 2016]
[9] Share Icon hotel legend: https://www.shareicon.net/hotel-accomodation
[10] Icon Archive hotel legend: http://www.iconarchive.com/show/ios7-icons-by-icons8/Hotel.html
[11] Icons DB hotel legend: http://www.iconsdb.com/royal-azure-blue-icons/hotel-icon.html

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés



(a) *L'Ecluse restaurant* par GoogleMaps (b) *L'Ecluse restaurant* par HereMaps

FIGURE A.2: Le restaurant L'*Ecluse* en France dispose de données terminologiques différentes, fournies par deux fournisseurs LBS. [Accès: Février 2016]



FIGURE A.3: Trois légendes différentes pour représenter un hôtel sur des cartes collectées respectivement par Share Icon, Icon Archive et Google Maps.

le terme de "map conflation", où les cartes vectorielles et raster sont intégrées [RAUB11]. Cette thèse porte sur l'intégration d'objets ponctuels géospatiaux; elle ne tient pas compte de l'intégration des cartes de base ou de l'intégration d'objets géographiques complexes tels que les rues ou les bâtiments. L'intégration d'objets ponctuels nécessite la mise en correspondance de structures de données hétérogènes et la conciliation de données géospatiales incohérentes. L'intégration de données décrivant une même réalité est alors sujette à l'incertitude en raison de différents types d'hétérogénéité [CD99]. Ignorer cette incertitude peut, au mieux, conduire à des prédictions légèrement incorrectes et au pire être complètement fatale à l'utilisation de données intégrées. Cette thèse prend également en compte les incertitudes des données intégrées afin d'offrir une représentation précise et significative des données intégrées.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

Les problématiques

L'objectif de cette recherche est d'améliorer la qualité des résultats des requêtes faites par les utilisateurs afin que les touristes puissent rechercher et trouver les POIs avec des informations plus précises et complètes. Les données de plusieurs fournisseurs LBS, c'est-à-dire les entités qui se réfèrent au même endroit devront être intégrées pour d'une part éviter la duplication et d'autre part obtenir des informations plus complètes. Les principaux problèmes liés à l'intégration de données issues de LBS sont détaillés ci-dessous.

1. Accès aux données : l'intégration LBS nécessite un accès aux données des fournisseurs. La plupart des fournisseurs LBS, en particulier des fournisseurs commerciaux, donnent un accès gratuit et limité à leurs bases de données. Ils offrent un accès sur demande à l'aide de services Web Application Programming Interface (API). Les APIs des fournisseurs sont entourés de limitations concernant le nombre de requêtes par jour et le nombre de POIs retournés par demande, etc. L'accès aux POIs via les fournisseurs LBS est soumis à des contraintes techniques et aléatoires, comme la publicité et les avis des utilisateurs sur les POIs. Cela affecte la disponibilité et le nombre de POIs, et donc leur intégration. En outre, les données des fournisseurs sont sujettes à des politiques d'utilisation. Ces politiques diffèrent d'un fournisseur à l'autre. Par exemple, les données gratuites sont disponibles à titre personnel ou pour des projets de recherche, mais pas pour des projets commerciaux.

2. Qualité des données de base : les fournisseurs de LBS ont des stratégies distinctes pour construire leur bases de données. Par exemple, le fournisseur Apidae-Tourisme[12], également connu sous le nom SITRA, emploie des experts en géographie pour visiter les lieux afin de vérifier et de valider les POIs. Par contre, le fournisseur OpenStreetMap permet aux utilisateurs d'ajouter, de modifier et de supprimer des POIs. Ces diverses stratégies conduisent à une qualité différente des sources de données. L'intégration de données de faible et de haute qualité risque de ne pas fournir une meilleure information.

3. Hétérogénéité du schéma : Le schéma d'une base de données décrit sa structure. En 2010, un groupe de travail du W3C[13] a été créé pour élaborer les spécifications normalisées pour la représentation des informations des POIs sur le Web. En 2013, ce groupe de travail a été transféré à Open Geospatial Consortium (OGC) en vertu de Points Of Interest Standards Working Group (POISWG) [POI13]. Le but de ce

[12] Fournisseur Apidae-Tourisme, aussi connu sous le nom Sitra: http://apidae-tourisme.com
[13] W3C working group for POI standard model: https://w3.org/2010/POI

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

groupe était de produire un standard de codage des données des POIs qui inclut un modèle de données abstrait et des implémentations de schéma JSON et XML de ce modèle de données. La création du modèle standard a été initiée[14], mais abandonnée en 2014. De nos jours, aucun schéma standard ne peut être utilisé pour construire une base de données des POIs; chaque fournisseur LBS utilise son propre modèle. Cela signifie que les données des fournisseurs distincts ont des étiquettes et des hiérarchies différentes.

4. Hétérogénéité des valeurs des attributs : les bases de données créées par différentes personnes et pour différents objectifs peuvent avoir des variations au niveau de l'instance. Cela se traduit par l'intégration des données même lorsqu'elles se rapportent au même objet. Les variations des valeurs des attributs des POIs se manifestent pour plusieurs raisons. En ce qui concerne les valeurs spatiales, les objets du monde réel peuvent être représentés différemment. Ils peuvent être différents en raison de la précision des appareils GPS, de l'erreur humaine lors de la localisation d'un lieu ou des stratégies distinctes pour localiser un POI. Par exemple, un parc peut être représenté comme un polygone qui couvre la zone du parc ou un point situé à la porte du parc. Les valeurs terminologiques peuvent être (1) syntaxiquement différentes, telles que l'utilisation d'abréviations et d'erreurs d'orthographe, ou (2) sémantiquement différentes comme les synonymes et les hyponymes. Les variations des valeurs au fil du temps peuvent également provoquer l'hétérogénéité des données si celles-ci ne sont pas mises à jour régulièrement. Un autre problème fréquent concernant l'hétérogénéité des données est le langage utilisé pour décrire les POIs. Par exemple, les noms de lieux des POIs offerts par Open-StreetMap sont décrits dans différentes langues. Par contre, Bing Maps utilise la langue du pays où se trouvent les POIs. Ainsi, les données de plusieurs fournisseurs peuvent être décrites dans différentes langues, ce qui provoque une hétérogénéité supplémentaire. Ces hétérogénéités distinctes créent des obstacles pour l'appariement d'entités et l'intégration LBS.

5. Évaluation de l'alignement des entités géospatiales : L'intégration des LBS nécessite d'associer les entités qui se rapportent au même POI du monde réel. Dans certains contextes, il existe des clés identificateurs qui peuvent être utilisées pour l'alignement d'entités. Par exemple, le numéro de sécurité sociale (NSS) peut être utilisé comme une clé pour identifier les citoyens dans plusieurs bases de données, parce que chaque personne a un NSS unique. Malheureusement, les fournisseurs LBS utilisent un identifiant interne spécifique dans leurs bases de données. Cela signifie que les POIs manquent d'un identifiant unique global. Dans ce cas, l'alignement d'entités peut être effectué en comparant les valeurs

[14] POI standard model project: https://github.com/opengeospatial/poi

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

des attributs décrivant les détails des POIs. Plusieurs approches ont été proposées pour résoudre l'alignement des entités géospatiales. Cela peut être accompli en mesurant les similitudes des valeurs entre les entités afin de déterminer si elles correspondent. Ces approches ont été évaluées dans différents contextes en utilisant différentes métriques et ensembles de données. Ces ensembles de données sont souvent de petite taille, mal caractérisés, choisis au hasard et non mis à la disposition des chercheurs, ce qui rendent incomparables les résultats des expériences des approches. Des benchmarks ont été proposés pour l'alignement d'entités dans des contextes tels que les publications et le commerce [KTR10]. Mais, toutes ces références ne tiennent pas compte de l'aspect spatial des entités. À notre connaissance, il n'existe pas de benchmarks pour l'évaluation des approches qui correspondent aux entités géospatiales.

6. Évaluation de la fusion des données : Une fois que les entités correspondantes de plusieurs sources sont identifiées, un algorithme est nécessaire pour les fusionner afin d'obtenir de nouvelles entités intégrées qui sont de meilleure qualité que les entités sources. Certains attributs peuvent avoir différentes valeurs d'un fournisseur à un autre. La fusion de données est un processus de décision qui permet de sélectionner la valeur qui semble la plus correcte. Par exemple, une solution de base consiste à choisir la valeur qui provient d'une source de haute qualité, ce qui introduit de nouveaux problèmes sur la classification de la qualité des fournisseurs LBS. Pour évaluer les valeurs sélectionnées par l'algorithme de fusion, nous devons les comparer à des valeurs correctes. Malheureusement, il n'y a aucune source de données sûre à 100% qui peut être utilisée comme vérité de base. La seule façon de réaliser cette évaluation est de vérifier manuellement les valeurs de la réalité, ce qui est impossible en raison du nombre croissant des POIs. Au lieu de cela, il est possible d'estimer la certitude des entités fusionnées d'après les sources d'information, mais pas à la réalité. Inversement, si un attribut a plusieurs valeurs contradictoires, la valeur choisie ne peut pas être considérée comme fiable. Un algorithme de fusion devrait alors pouvoir évaluer l'incertitude des entités intégrées en fonction de la contradiction des entités sources. Cette incertitude peut résulter de l'information spatiale, des informations terminologiques ou des deux types d'information à la fois. L'incertitude des données exige la présence d'informations supplémentaires qui doivent être fournies aux utilisateurs via des applications LBS.

7. Visualisation des données : Lors de l'examen des entités intégrées, des informations supplémentaires sur l'incertitude des données et les sources de données doivent être fournies aux utilisateurs afin qu'ils puissent estimer la qualité des données intégrées et comparer les valeurs des différentes sources afin de prendre des décisions éclairées lors du choix des POI. Il est alors nécessaire de fournir toutes

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

les informations supplémentaires sans surcharger les cartes, en particulier pour les utilisateurs de téléphones mobiles. Des solutions de représentation des données doivent être évaluées pour s'assurer que les utilisateurs peuvent comprendre les informations supplémentaires sans aucune contradiction avec les informations de base. Une autre question concerne la visualisation des entités intégrées. Comme nous l'avons mentionné, les légendes sont utilisées pour représenter des POIs sur des cartes. Cependant, chaque fournisseur LBS a son propre ensemble de légendes. Il est nécessaire de résoudre l'hétérogénéité à ce niveau également.

Questions de recherche

L'objectif principal de cette thèse est de créer des LBS unifiés en intégrant les données de plusieurs fournisseurs. Après avoir énuméré les problématiques d'intégration des données géospatiales, cette section représente les questions que cette thèse aborde.

1. Comment les fournisseurs de LBS se distinguent-ils les uns des autres? En d'autres termes, quels types d'hétérogénéité et d'incohérence existent-ils entre les fournisseurs LBS ? Et comment les données LBS peuvent-elles être intégrées? (Problèmes : hétérogénéité du schéma et hétérogénéité des valeurs des attributs).

2. Les techniques d'alignement des entités géospatiales sont-elles suffisamment efficaces? Comment mettre en œuvre une évaluation et une comparaison équitables des approches d'appariement d'entités géospatiales ? (Problème: évaluation de l'appariement d'entités géospatiales).

3. Comment les entités faisant référence au même objet du monde réel doivent-elles être fusionnées ? Et comment estimer l'incertitude résultant d'une telle fusion ? Quelle est la meilleure façon de livrer des POIs intégrés et de représenter leur incertitude dans le contexte touristique ? Comment les touristes interprètent-ils cette incertitude ? Influe-t-elle sur les choix des touristes lorsqu'ils recherchent des POIs ? (Problèmes: évaluation de la fusion des données et visualisation des données).

On note que les autres problèmes comme l'accès aux données et la qualité des sources de données ne sont pas prises en compte dans cette thèse. Bien que les APIs d'accès aux données soient entourées de plusieurs limites techniques, elles sont toujours suffisantes pour notre projet de recherche. De plus, plusieurs travaux existants traitent la qualité des données géospatiales [TKA07, PZLL11]. Mais dans cette thèse, nous utilisons les sources de données LBS telles qu'elles sont, sans tenir compte de leurs qualités.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

Méthodologie et contributions

Après avoir développé les problèmes et les questions de recherche, et en nous appuyant sur l'état de l'art, nous présentons en détail notre apport.

En ce qui concerne la première contribution, nous avons déjà mis en évidence la nécessité de créer un benchmark pour assurer une évaluation et une comparaison équitables des approches d'alignement d'entités géospatiales. La Figure A.4 montre le flux de processus pour créer un tel benchmark en utilisant des données réelles collectées de plusieurs fournisseurs LBS. Pour ce faire, nous commençons par créer une taxonomie qui décrit et formalise l'ensemble des incohérences et des hétérogénéités entre les fournisseurs LBS. Ensuite, un outil appelé GeoBench Aligner est implémenté; Il s'agit d'un processus semi-automatique qui recueille des données provenant de fournisseurs LBS déjà existants. La validation d'un utilisateur est nécessaire pour indiquer les incohérences entre les entités et décider si deux entités correspondent. La sortie de cet outil est une base de données, nommée GeoBench DB, caractérisée par la taxonomie et qui comprend des entités réelles pour lesquelles nous connaissons les lines de correspondance. Un deuxième outil, nommé PABench Extractor, utilise les données issues de GeoBench DB en entrée pour générer des séries de données décrivant une situation donnée. Ces bases de données caractérisées seront utilisées ultérieurement pour évaluer les approches d'alignement d'entités géospatiales. Notre benchmark, appelé PABench (POI Alignment Benchmark), est constitué des ensembles de données et d'une liste de paramètres ayant pour objectif d'évaluer la performance et la qualité des approches d'alignement. Ainsi, PABench permet d'évaluer ces approches dans différentes situations, ce qui permet de découvrir leurs points faibles et forts.

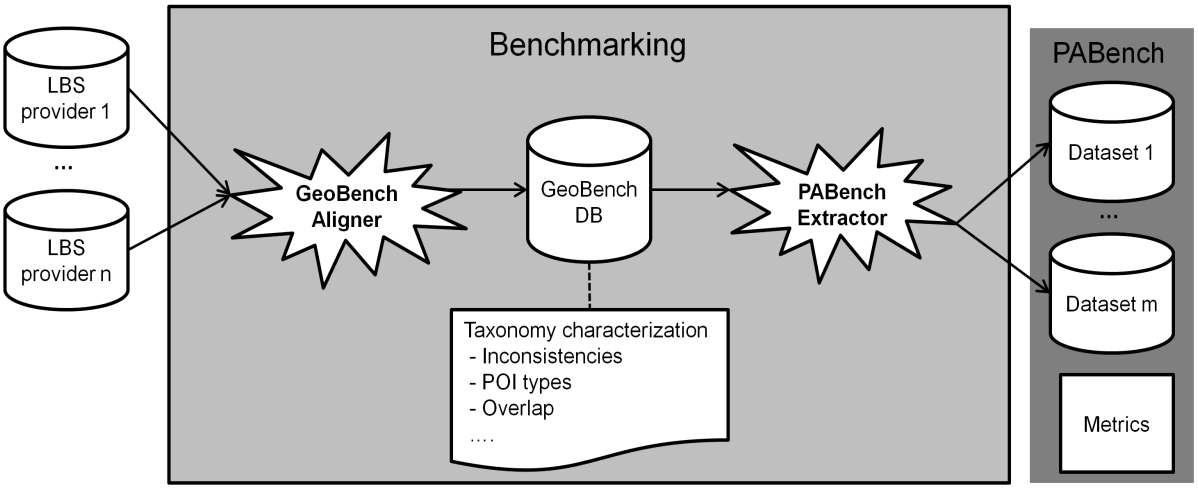


FIGURE A.4: Elaboration de benchmark.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

Nous nous concentrons ensuite sur le processus d'alignement pour détecter les entités de deux ensembles de données qui se réfèrent à la même place dans le monde réel. Par conséquent, la deuxième contribution consiste à élaborer un processus d'alignement et de fusion pour intégrer des entités de plusieurs fournisseurs. La Figure A.5 montre le flux de processus de cette contribution. Nous proposons une généralisation pour la mesure de similarité spatiale, appelée Normalized-Distance (ND). Cette mesure est calculée par la distance Euclidienne entre deux entités comparées. Puis, les informations spatiales et terminologiques des entités sont comparées en utilisant des mesures de similarité distinctes. Ensuite, nous proposons une méthode, nommée Global Similarity (GS), qui sert à combiner numériquement plusieurs similarités afin d'obtenir une similarité globale entre les entités comparées. En utilisant la similarité combinée, nous décidons si les deux entités correspondent. En ce qui concerne la phase de fusion, un algorithme de base est utilisé en s'appuyant sur les méthodes existantes dans la litérature. Finalement, ce processus produit un ensemble de données contenant des entités fusionnées où chaque entité est accompagnée d'un coefficient représentant son incertitude. Ce coefficient est estimé à l'aide des mesures de similarité initiales et combinées. Notre proposition est finalement comparée aux approches déjà existantes en utilisant PABench.

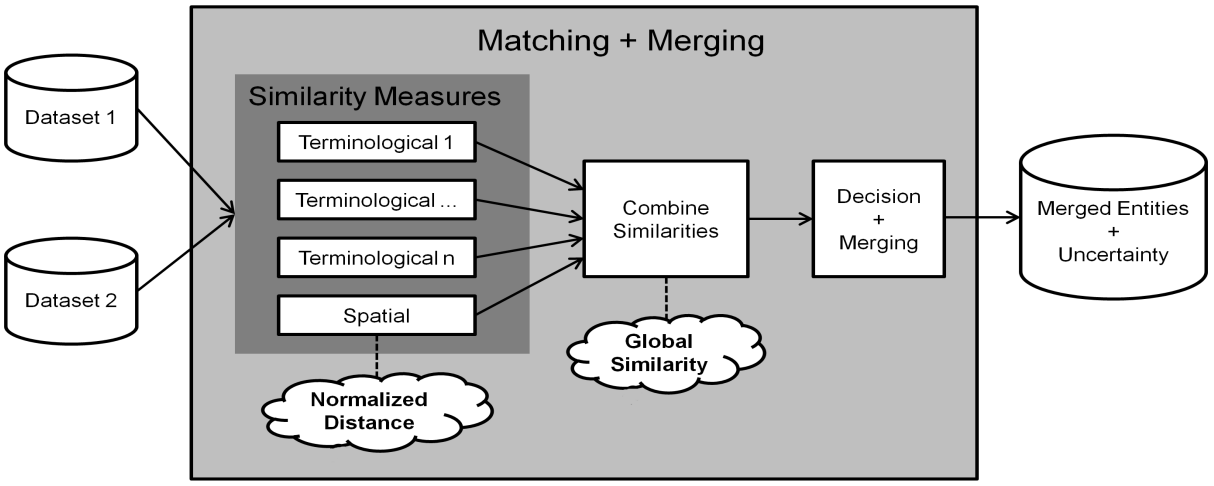


FIGURE A.5: Flux de processus de l'alignement et fusion.

La dernière contribution de cette thèse concerne la restitution des entités intégrées et leurs incertitudes aux utilisateurs finaux. La Figure A.6 montre le flux de processus de cette contribution. Après avoir étudié les solutions existantes pour représenter les POIs intégrés et les informations d'incertitude, nous sélectionnons et adaptons ce qui peut être utilisé dans notre contexte. Sur ces bases, plusieurs propositions sont élabor'ees et un prototype intégrant ces propositions est réalisé. Un premier test psycho cognitif est mené pour trouver la proposition la plus appropriée pour représenter les POIs intégrés et leur incertitude à trois niveaux, nommé l'incertitude spatiale, l'incertitude terminologique et l'incertitude globale. Un deuxième test psycho-cognitif est ensuite réalisé pour analyser

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

comment l'incertitude reflète sur les décisions des utilisateurs lors de la recherche de POI.

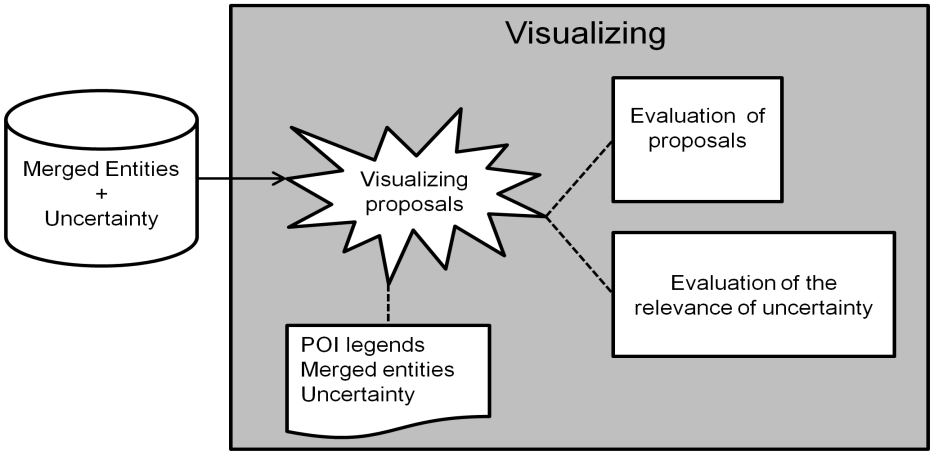


FIGURE A.6: Flux de processus de la visualisation.

Conclusion

La quantité de données géospatiales et le nombre de fournisseurs de LBS ont augmenté de façcon spectaculaire au cours des dernières années. La contradiction entre les entités géographiques provenant de sources multiples et qui décrivent la même réalité, remet en doute la validité des POIs offerts. Les entités géospatiales qui réfèrent au même POI peuvent inclure des données incomplètes, incompatibles, inexactes ou même fausses, d'un fournisseur à l'autre aux niveaux spatial, terminologique et représentatif

Cette thèse est une contribution pour améliorer l'exactitude et l'exhaustivité des données géospatiales provenant de plusieurs fournisseurs LBS. Notre proposition est appliquée à un cas d'utilisation des POIs pour les touristes. La recherche sur cette question se déroule dans le cadre d'un projet multidisciplinaire. Ainsi, ce travail n'a pas pu être réalisé sans la collaboration de partenaires en Informatique, Géomatique et Sciences Humaines et Sociales, et de praticiens des offices de tourisme.

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés

FOLIO ADMINISTRATIF

THESE DE L'UNIVERSITE DE LYON OPEREE AU SEIN DE L'INSA LYON

NOM : BERJAWI
(avec précision du nom de jeune fille, le cas échéant)

DATE de SOUTENANCE : 01/09/2017

Prénoms : Bilal

TITRE : Intégration des données hétérogènes issues de plusieurs fournisseurs de services géo-localisés : un cas d'utilisation sur les points d'intérêt touristique

NATURE : Doctorat

Numéro d'ordre : 2017LYSEI072

Ecole doctorale : INFOMATHS

Spécialité : Informatique

RESUME :

Les fournisseurs de services géo-localisés (LBS) offrent des données textuelles et spatiales complémentaires, parfois incohérentes et imprécises, représentant les différents points d'intérêt (POI) sur un territoire donné. Ainsi, une même requête lancée auprès de divers fournisseurs de services touristiques peut donner des résultats différents et parfois incohérents, pour les attributs terminologiques et/ou les attributs spatiaux. De plus, chaque fournisseur utilise sa propre convention graphique pour représenter les POIs. L'intégration de ces données spatiales hétérogènes dans un contexte dynamique, large échelle, utilisant des sources incomplètes et de qualités variables est actuellement un verrou technologique. Dans ce travail de thèse, nous cherchons une solution à cette intégration aussi bien au niveau des données que de leur représentation.

MOTS-CLÉS : Alignement d'entités géographique, Intégration géographique, SIG, Services géo-localisés, Qualité de données géospatiale, Benchmark d'alignement d'entités, Incertitude de données géospatiale, Cartography

Laboratoire (s) de recherche : LIRIS

Directeur de thèse: Maryvonne MIQUEL

Président de jury :

Composition du jury : Maryvonne MIQUEL, Franck FAVETTA, Thomas DEVOGELE, Mauro GAIO, Ahmed LBATH, Thérèse LIBOUREL, Chrisitne ZANIN, Philippe LAMARRE

Cette thèse est accessible à l'adresse : http://theses.insa-lyon.fr/publication/2017LYSEI072/these.pdf
© [B. Berjawi], [2017], INSA Lyon, tous droits réservés