Thèse de doctorat de l'Université Gustave Eiffel

École doctorale n°532, Mathématiques et Sciences et Technologies de l'Information et de la Communication (MSTIC)

Spécialité de doctorat : Sciences et technologies de l'information géographique

Unité de recherche : UMR LASTIG

De données hétérogènes et fragmentaires provenant de sources multiples à un référentiel géohistorique d'adresses

Charly BERNARD [1]

Thèse présentée et soutenue le 26 février 2026

Composition du Jury

Nathalie AUSSENAC-GILLES Directrice de recherche, IRIT, Université de Toulouse, CNRS	Rapporteure
Marlène VILLANOVA Professeure, LIG, Université Grenoble Alpes	Rapporteure
Fatiha SAÏS Professeure, LISN, Université Paris-Saclay, CNRS	Présidente du jury
Claire LAGESSE Maîtresse de conférences, ThéMA, Université Marie et Louis Pasteur	Examinatrice
Stéphane BACIOCCHI Ingénieur de recherche, CRH, EHESS-CNRS	Invité

Encadrement de la thèse

Julien PERRET Directeur de recherche, LASTIG, Univ. Gustave Eiffel, IGN-Géodata Paris	Directeur de thèse
Nathalie ABADIE Chargée de recherche IGN, LASTIG, Univ. Gustave Eiffel, IGN-Géodata Paris	Co-directrice de thèse
Bertrand DUMENIEU Ingénieur de recherche, CRH, EHESS-CNRS	Encadrant de thèse

1. Charly Kévin Patrick BERNARD

Remerciements

Cette partie du manuscrit a été pour ma part le plus difficile à faire du fait qu'elle constitue la partie la plus personnelle. Cette thèse de doctorat, débutée début 2023 a fait l'objet, de l'utilisation de graphes de connaissances. Il n'y a pas de raison que cette section y échappe : les remerciements sont illustrés avec des graphes de mes connaissances que je remercie. De même, puisque mes travaux portent sur la représentation des évolutions temporelles, j'ai choisi de définir une règle en organisant les remerciements de manière chronologique : en commençant par les connaissances les plus anciennes — ma famille — puis les amis, et enfin les collègues rencontrés durant mon parcours professionnel.

Encadrement scientifique

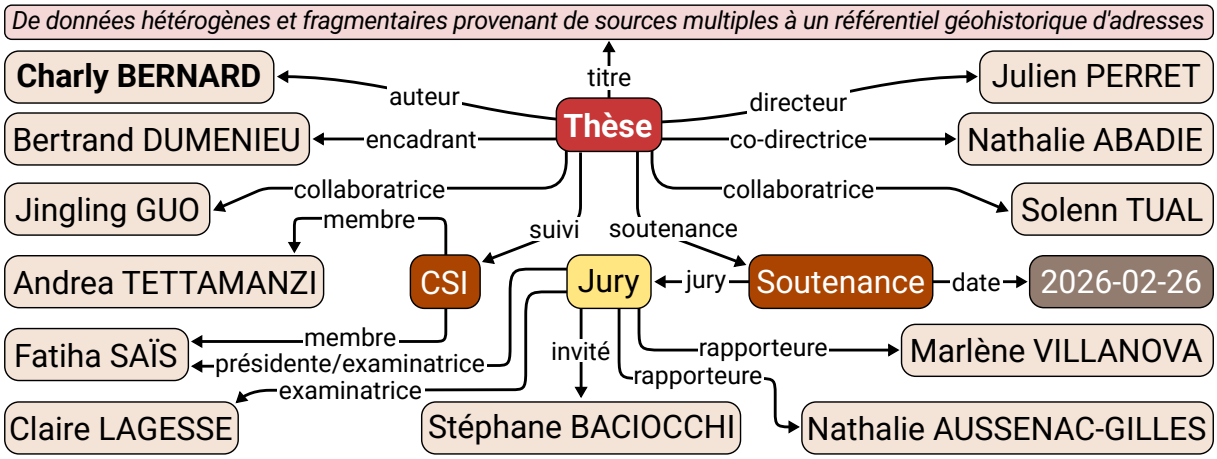


FIGURE 1 – Graphe des personnes directement liées à la thèse.

Tout comme la langue française adore le faire, je vais commencer par ajouter une exception à la règle que je viens de donner en commençant par remercier ceux qui sont directement liés à mes travaux de thèse (voir figure 1). En premier lieu, je ne peux nier l'excellence de l'encadrement que j'ai eu sur ces trois années et ce grâce aux trois membres de l'équipe d'encadrement dotés de qualités sur tous les plans. Tout d'abord, sur le plan scientifique et professionnel, ils m'ont aidé d'un soutien précieux grâce à leur expertise, leurs conseils apportés, leur disponibilité sans oublier la confiance qu'ils m'ont accordée. Sur le plan humain aussi : leur sympathie, leur gentillesse et leur bienveillance ont été des variables importantes pour le bon déroulement de ma thèse. Chacun des échanges que j'ai pu avoir avec les membres de l'équipe a toujours été très agréable. De plus, la sérénité apportée par l'équipe m'a permis de travailler dans un environnement serein et stimulant, au point que la période 2023-2026 est passée rapidement avec un brin de déception à

l'idée que cette aventure se clôture. Je ne peux donc que remercier chaleureusement mon équipe d'encadrement pour ces trois années passées avec eux et j'espère retravailler avec eux dans le futur. En plus de ce remerciement collectif, je souhaite également le faire sur un plan individuel pour chaque membre de mon équipe d'encadrement.

Je tiens tout d'abord à remercier Julien Perret, mon directeur de thèse. Je le remercie de m'avoir donné l'opportunité de participer au projet SoDUCo à mon arrivée au LASTIG, un projet dont les sujets de recherche m'ont intéressé et qui m'a également permis de travailler sur les thématiques abordées dans ce manuscrit, des thématiques dont l'intérêt a fortement contribué à me stimuler tout au long de ma thèse. Je lui suis également reconnaissant d'avoir constitué l'équipe d'encadrement dont j'ai pu souligner l'excellence. Tout au long de la thèse, il a su instaurer un cadre de travail me laissant une grande autonomie, tout en restant disponible et présent lorsque cela était nécessaire. Je souhaite ensuite remercier ma co-directrice de thèse, Nathalie Abadie. Bien que nous nous soyons côtoyés dans le cadre du projet SoDUCo, c'est véritablement au moment de la thèse que notre collaboration s'est concrétisée. Je la remercie grandement pour le temps qu'elle a consacré à m'accompagner tout au long de ce travail, aussi bien sur les aspects scientifiques que sur des dimensions parfois moins visibles mais tout aussi importantes, comme les démarches administratives (missions, préparation de la soutenance, et bien d'autres). Son expertise a été particulièrement précieuse sur les questions liées aux graphes de connaissances, au développement d'ontologies et plus largement aux technologies du Web sémantique. Enfin, je souhaite remercier Bertrand Dumenieu pour son encadrement et ses conseils tout au long de ce travail. Il m'avait déjà beaucoup aidé lors du développement de l'outil collaboratif de validation sur lequel je travaillais dans le cadre du projet SoDUCo. Ses remarques et ses conseils ont été particulièrement précieux pour réfléchir aux différentes manières de représenter et de modéliser les évolutions spatio-temporelles des entités géographiques.

Durant ces années de thèse, le travail est généralement solitaire. J'ai toutefois eu la chance de partager ma trajectoire doctorale avec la docteur Solenn Tual, nos thèses s'étant déroulées simultanément et sous la supervision de la même équipe d'encadrement. Bien que nos travaux aient été distincts, nous avons eu l'occasion de collaborer sur certains sujets complémentaires, tout comme nos contributions lors de ces travaux communs. J'ai pu partager de nombreux moments de travail avec une personne sérieuse et appliquée. Je n'oublie pas non plus les nombreux échanges très intéressants qu'on a eus sur différents sujets qu'on affectionnait. Toutes ces raisons me mènent à la remercier chaleureusement. Enfin, bien que ce fut bref sur la fin de thèse, je voudrais aussi remercier Jingling Guo, étudiante que j'ai encadrée avec Nathalie dans le cadre de son stage de fin d'étude en Master 2 Langue et informatique de Sorbonne Université (l'annexe A est une synthèse de ses travaux). Ce premier encadrement fut d'autant plus agréable qu'il s'est fait avec une étudiante sérieuse et investie.

En parallèle, j'aimerais remercier les membres de jury à commencer les rapporteures du jury : Nathalie Aussenac-Gilles et Marlène Villanova. Je les remercie grandement d'avoir accepté de relire mon manuscrit, d'avoir pris de leur temps de me faire part de leurs retours et interrogations concernant mes travaux et les reconstitutions qui en ont été faites dans le présent manuscrit. Ceux-ci m'ont permis d'apporter des améliorations à ce sujet. Je souhaite également remercier les deux examinatrices : Claire Lagesse et Fatiha Saïs. Je leur suis reconnaissant d'avoir accepté ce rôle et d'avoir consacré du temps à la lecture de ce manuscrit. Je les remercie également pour les questions et remarques qu'elles ont formulées lors de la soutenance, qui ont permis d'ouvrir des perspectives et d'enrichir la discussion autour de ces travaux. Puisque je remercie Fatiha Saïs, je voudrais en profiter pour la remercier également, aux côtés d'Andrea Tettamanzi, pour le suivi annuel de

mes travaux et les recommandations qu'ils ont apportées lors des fameux comités suivis individuels (CSI) de thèse. Je voudrais remercier Stéphane Baciocchi, que j'ai connu lors de ma participation au projet SoDUCo, et qui a accepté de former le dernier membre du jury en tant qu'invité. Pour terminer, je souhaiterais remercier Emmanuel Fritsch de m'avoir suggéré de me lancer dans cette aventure lors d'un de mes entretiens annuels ainsi que de m'avoir laisser une certaine flexibilité dans mon organisation du travail.

Famille

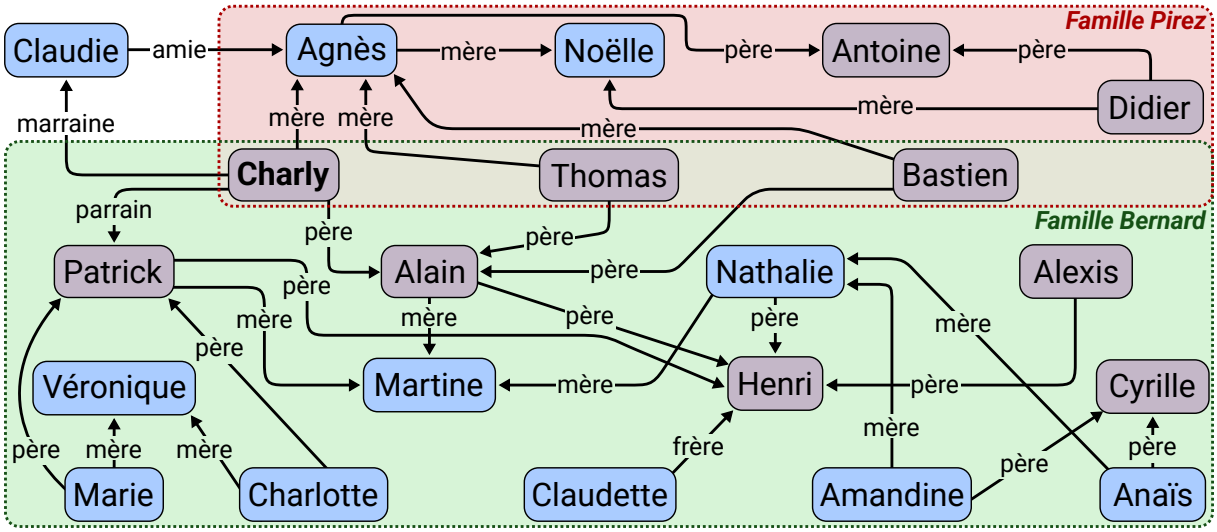


FIGURE 2 – Graphe des membres de ma famille que je remercie.

Ensuite, je voudrais remercier ma famille (dont les liens que je partage avec les membres sont indiqués dans la figure 2), en commençant logiquement par mes parents, Agnès et Alain, qui m'ont d'une part donné la vie et d'autre part éduqué. Bien que j'en aie eu conscience dès ma jeunesse, je mesure la chance que mes frères et moi avons eue de grandir dans une famille qui nous a offert à la fois une grande liberté : peu de contraintes, pas de punitions, la possibilité de sortir à peu près quand on voulait. De même, ils nous offert une stabilité dans tous les domaines notamment grâce au cadre de vie agréable aux Gambys, un hameau où régnait une belle ambiance, où je garde d'excellents souvenirs et où c'est toujours un plaisir d'y revenir. Je souhaite remercier ma mère, qui a été présente à chaque étape de notre enfance et de notre scolarité : nous guidant, nous faisant réviser quand il le fallait et veillant à notre bien-être au quotidien. Sa gentillesse et attention constante ont été un pilier dans mon éducation et celle de mes frères. Je souhaite également remercier mon père, notamment pour le goût du sport qu'il m'a transmis, avec qui je partage de simples discussions sur ce domaine mais aussi autour de la musique. Je lui suis aussi reconnaissant pour l'influence qu'il a eue au début de mes études supérieures. Par un heureux hasard, il avait pris connaissance de la géomatique et m'avait suggéré que ce domaine pouvait m'intéresser, ce qui m'a conduit à intégrer l'ENSG-Géomatique et donc à mener la thèse que ce manuscrit présente.

Je remercie aussi mes deux petits frères, Thomas et Bastien. C'est aussi grâce à eux que j'ai eu une excellente enfance. J'ai pu partager de nombreux moments avec eux, notamment sportifs. Je repense à la chance que nous avions d'avoir un terrain nous permettant de pratiquer différentes disciplines, principalement du football, du vélo mais aussi

du murket – ce sport que nous avions inventé. Je pense particulièrement à Thomas, avec qui j'ai peu de différence d'âge. Sans vraiment comprendre pourquoi, il arrivait que nous soyons confondus lorsque nous étions jeunes alors que nous étions bien différents. Cela ne signifiait pas que nous partagions peu de centres d'intérêt communs, bien au contraire. Nous avons vécu beaucoup de choses ensemble, même si ce fut moins dense ces dernières années. Parmi nos dernières expériences communes, je pense à notre participation à Raidy To Go 2012 sous la bannière *Les Gadlains* dont je remercierai plus loin certains autres membres de l'équipe. Même si l'écart d'âge est plus important avec Bastien, cela ne nous a pas empêché de bien nous entendre. Nous partageons encore aujourd'hui de nombreux moments. Avoir des goûts musicaux proches aide beaucoup car on a pu partager de nombreux concerts notamment celui de SCH au Vélodrome en juillet 2023. De plus, j'ai pu (re)découvrir divers endroits en France et participer à divers événements. Enfin, il ne faut pas les oublier, car ils m'ont aussi accompagnés tout au long de ma vie et ont aussi leur part dans mes souvenirs. Je parle bien sûr des différents animaux de compagnie que nous avons eus. J'ai une pensée pour ceux qui nous ont quittés, à savoir les chats Gus, Eirness, Esbrouf, Fifa et Lupin. Je pense également à ceux qui m'ont tenu compagnie lors de mes retours chez mes parents pendant ma thèse (et le font encore) : les chats Nino, Rosco, Sango et Uno sans oublier le chien Mario. Ils apportent de la légèreté et du réconfort dans notre quotidien.

En outre, j'aimerais remercier mes grands-parents maternels, Noëlle et Antoine. Même si faire plus d'une heure de trajet pour les voir dans le Brionnais me paraissait long quand j'étais petit, c'était toujours un plaisir de passer du temps avec eux pendant les vacances. Je garde en mémoire les jeux télé que nous regardions ensemble, les délicieux plats de ma grand-mère — surtout le gratin dauphinois avec le rosbif — et l'amour du football de mon grand-père, aujourd'hui décédé, dont j'ai hérité. Aussi, je remercie mes grands-parents paternels, Martine et Henri. Il était plus facile de les voir que mes grands-parents maternels, et je garde de très bons souvenirs de mes visites dans leur épicerie du centre de Mâcon. Des moments qui m'ont marqué au point où, enfant, je rêvais de faire ce métier. Je souhaite également remercier ma grande tante Claudette, chez qui j'ai de merveilleux souvenirs de vacances passées l'été à Toulon, dans le sud de la France.

Plus globalement, j'aimerais inclure dans mes remerciements l'ensemble de ma famille que ce soit du côté Bernard ou du côté Pirez à commencer par mes oncles et tantes, ainsi que mon parrain et ma marraine : Patrick, Véronique, Nathalie, Cyrille, Didier et Claudie. Je remercie aussi mes cousines, toutes issues du côté paternel : Marie, Charlotte, Amandine et Anaïs. Bien qu'il soit factuellement un demi-oncle que je considère plutôt comme un cousin, je remercie aussi Alexis. Je garde en mémoire les fêtes d'anniversaires, les repas de Noël et même les repas hebdomadaires du mardi soir que nous avons partagés ensemble.

Amis

La plupart de mes amis que j'ai connus le sont grâce à mes études. J'en garde uniquement des souvenirs positifs et, avant de les remercier, je souhaite adresser mes remerciements à l'ensemble des instituteurs et professeurs que j'ai eus au cours de ma scolarité (voir figure 3, qui présente une frise chronologique de mon parcours scolaire), c'est-à-dire à ceux des écoles publiques de Feillens, de Saint-Cyr-sur-Menthon et de Saint-Genis-sur-Menthon, du collège George Sand de Pont-de-Veyle, des lycées Lamartine de Mâcon et Lalande de Bourg-en-Bresse, ainsi qu'à ceux de l'ENSG-Géomatique (Géodata Paris depuis peu). Les citer individuellement serait trop long, mais je tiens à ce qu'ils sachent que

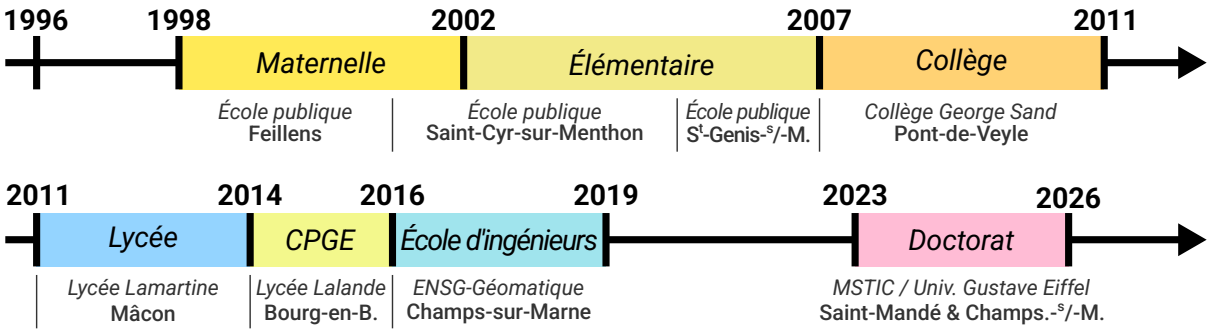


FIGURE 3 – Frise chronologique de mon parcours scolaire.

je n'en ai oublié aucun et que je leur suis reconnaissant. Au-delà de la scolarité, je souhaite également remercier tous les encadrants et entraîneurs qui m'ont accompagné dans la pratique du sport, notamment à l'US Saint-Cyr, à la JS Canton de Pont-de-Veyle, au TC Menthon et aux Jeunes de Mâcon Sports, pour leur engagement et leur encadrement tout au long de mes années de jeunesse.

Mis à part une cicatrice au-dessus du nez, je n'ai de souvenirs mémorables de mon passage à l'école maternelle de Feillens où je n'ai pas gardé d'attache contrairement aux écoles de Saint-Cyr-sur-Menthon et de Saint-Genis-sur-Menthon. Je commercerais par remercier Thibaut M., le premier ami que j'ai eu là-bas grâce à qui j'ai pu m'intégrer dans ce nouvel environnement, avec qui j'ai passé de belles années jusqu'au collège. Dans ces remerciements, je n'oublie pas non plus Calvin L., Gaëtan D., Guillaume P. et Quentin C. De même, j'inclus Basile L. avec qui je m'étais lié d'amitié durant mes années au collège George Sand à Pont-de-Veyle. Avant de parler de mes études supérieures, je voudrais remercier ceux que j'ai connus au lycée Lamartine de Mâcon en commençant ceux de ma classe de Seconde : Thibault C., Valentin D. et Jimmy B. Puis, je souhaiterais faire de même envers mes camarades de classe des deux dernières années du lycée qui ont été mémorables pour ma part en particulier le groupe des *fils* : Pierre G., Pierre-Antoine F., Antoine G., Aubin P., Romain K., Jimmy B., Ludovic F. Bien sûr, je n'oublie pas deux personnes qui m'ont marqué durant ces années à savoir Pascaline N. dont je garde beaucoup de souvenirs notamment en cours de mathématiques en première ainsi que Joseph N. dont le hasard de la vie a fait que nous sommes aujourd'hui deux agents de l'IGN. Une fois le baccalauréat en poche, j'ai rencontré de nouvelles personnes lors de mes années de prépa au lycée Lalande de Bourg-en-Bresse. Malgré la densité du travail, j'ai adoré cette période, en grande partie grâce à mes compagnons de chambre à l'internat : Alexandre L., Zakarya A. et Valentin B. De même, j'ai grandement apprécié travailler avec Alban G. et Thomas G., deux étudiants sérieux et fort sympathiques avec qui j'ai partagé de longues heures de travail, en particulier lors de notre TIPE.

Les concours aux grandes écoles passés et arrivé dans la région parisienne, j'aimerais remercier ceux qui ont rendus iconiques les trois années passées à l'École nationale des sciences géographiques (ENSG-Géomatique). Parmi ceux que j'ai connus ici, j'aimerais commencer par remercier Hugo N. avec qui j'ai passé un bon nombre d'heures de cours à ses côtés notamment en travaillant sur le mythique *projet Rugby* que je n'oublierai pas. De même que je n'oublierai pas le temps passé à communiquer par message durant ma thèse et à dire n'importe quoi, ce qui m'a permis de décompresser lors de mon temps de travail (*hihi*). Je tiens également à remercier Jocelyn Le M., toujours là à proposer des activités, notamment des voyages, que je n'aurais sûrement jamais réalisés sans ses propositions. Je tiens toutefois à ne surtout pas le remercier de m'avoir fait assister à des purges

footballistiques des horribles FC Nantes et Paris FC alors que j'en ai vu d'autres (étant un supporter de l'AS Saint-Étienne et du FBBP01). Par la suite, je remercie Robin D. et Matthieu D. avec qui j'ai commencé à découvrir la région Île-de-France lors de nos temps libres. Leurs connaissances respectives dans les transports et d'autres domaines divers m'ont toujours impressionné tout comme la bonne humeur et l'avenance de Robin ainsi que le second degré et le courage de Matthieu (avec deux T). Je voudrais aussi remercier Guillaume V. avec qui j'ai passé une excellente semaine en début de thèse à Séville sous une chaleur écrasante et durant laquelle j'ai encore assisté à une purge footballistique du Real Betis. Je n'oublie pas ta bonne humeur et ton rire qui se transmet lors de ta présence. Je n'oublie pas non plus Martin B. avec qui j'ai été en colocation durant près d'une année à la fin de nos études. Malgré quelques petits désagréments liés au logement (on comprendra ce qu'on voudra), ce fut une période agréable et j'ai beaucoup apprécié cette période alors que je suis de nature solitaire. De plus, je tiens à ne pas oublier Alice G. avec qui je n'avais trop eu l'occasion d'échanger durant nos années ENSG mais avec qui j'ai ensuite partagé quelques voyages (notamment celui en Égypte durant lequel la première version de ce manuscrit a été clôturée) et de nombreux moments. Je souhaite aussi la remercier pour le temps et l'énergie pour avoir conçu le *Mario* en peluche qui m'a fait énormément plaisir. Il y a d'autres personnes que je ne peux omettre ce serait trop long de leur adresser des remerciements individuels. Toutefois, je ne les oublie pas : Luc B. (avec qui j'ai été collègue au LASTIG pendant sa thèse), Nicolas L., Michaël G., Lucas V., Paul R., Florent C., Victor D. ainsi que mon parrain Bruce T. et mon filleul Madec G. Je n'oublie pas docteur Hiyam D. avec qui j'ai effectué mon premier projet de recherche (*Grand ciel bleu*) ainsi que la majorité des projets en 2e année. Ce fut agréable de partager avec elle ces moments durant mais aussi en dehors des périodes de travail et je l'en remercie. Je garde en mémoire sa bonne humeur et son côté beauf. Ensuite, je souhaite remercier mes camarades de promotion du Master 2 Carthagéo. Je garde de très bons souvenirs de cette année, qui fut sans doute l'une des meilleures pour moi notamment grâce la bonne ambiance qui y régnait. La faible charge de travail y ayant également contribué. Elle a marqué pour nous tous la fin de nos études et notre entrée dans le monde professionnel. J'espère que la sous-section 2.3.1 est en accord avec tout ce que nous y avions appris cette année-là (je me devais d'y inclure une référence bibliographique en guise de clin d'œil). Puisque je l'ai fait précédemment, je ne remercierai pas de nouveau Martin B., Matthieu D., Jocelyn Le M., Hugo N. et Guillaume V. J'aimerais toutefois inclure d'autres personnes dans ce cadre, à commencer par Hassen C. grâce à qui j'ai appris à lever le nez quand l'intervenant donne ses indications mais aussi à faire attention lorsqu'on se déplace à vélo. Par la suite, j'aimerais remercier Justine B., ma voisine de table lors de nos cours à l'Institut de géographie qui apportait toujours sa bonne humeur au sein du groupe et avec qui j'ai eu la chance de travailler sur quelques projets collectifs. En dernier lieu, je souhaite remercier Camille B., une brillante étudiante avec qui j'ai eu l'opportunité de faire équipe l'ensemble des projets. Travailler avec elle a été un vrai plaisir, autant pour les sessions de travail que pour les moments partagés en dehors des cours.

En plus, je voudrais parler d'autres personnes que j'ai connues dans le cadre de mes études supérieures. Je commence par Sarra B., que j'ai connue à l'ENSG, que je remercie notamment pour sa gentillesse rendant agréable les moments partagés. Ensuite, je garde un excellent souvenir de mon stage de fin d'études à la direction Innovation & Recherche de la SNCF où j'ai appris à connaître des personnes que j'ai appréciées, plus particulièrement Archana P. Je la remercie pour sa gentillesse et son énergie débordante (un peu trop des fois mais c'est pour ça qu'on l'adorait) qui faisait que chaque jour était imprévisible et

agréable à la fois. Ce stage à la SNCF m'a aussi mené, au gré de mes expériences avec mes équipes de football, l'AS Pro 500 et Écoréseau FC, à rejoindre le collectif Möketa en septembre 2022, juste avant de commencer ma thèse. Faire partie de cette équipe de football de joueurs de divers horizons m'a permis de me défouler et de m'amuser en parallèle de la thèse. Je remercie les membres du collectif avec qui j'ai joué et joue encore car je prends plaisir à jouer malgré les défaites. La bonne ambiance et le fair-play général qui y règne dans l'équipe me font oublier les déconvenues sportives même si, heureusement, c'est en train de changer.

Laboratoire et collègues

Enfin, lors de mon arrivée au LASTIG, j'ai eu la chance de rencontrer de nombreuses personnes avec qui j'ai plus ou moins échangé, et ce fut toujours un plaisir de travailler au sein de ce laboratoire. Le roulement important des membres – arrivées et départs – rend difficile de tous les citer, mais je souhaite tout de même exprimer ma gratitude envers les personnes qui ont marqué mon parcours.

Tout d'abord, je tiens tout particulièrement à remercier l'équipe dirigeante du laboratoire ainsi que les membres de l'équipe STRUDEL, dont j'ai fait partie durant les six années passées au laboratoire. Ensuite, je tiens à remercier Karine, chargée de gestion administrative du LASTIG, pour sa réactivité et son sérieux dans la gestion des dossiers administratifs notamment pour la participation aux différentes conférences (réservations de train et d'hôtel, inscriptions, ordres de mission, etc.).

Je souhaite également remercier la promotion dite de la *pince à linge* (une partie des doctorants du laboratoire présents sur la période 2023–2026) : Solenn, Martin, Florent, Chahine et Bérénice, avec qui j'ai évolué et partagé simultanément les différentes étapes de nos doctorats respectifs.

Je tiens aussi à remercier celles et ceux avec qui j'ai partagé le bureau K317 sur le site de Saint-Mandé à l'IGN lors des deux premières années de ma thèse : Laura, Bérénice, Sofiane, Pierre, Christelle, Matilde, Sami, Assia, Rebecca, Vera et Paul. L'ambiance y était à la fois sérieuse et conviviale, rendant l'environnement de travail particulièrement agréable. Durant la dernière année de thèse, j'ai partagé le bureau C406 au bâtiment Carnot de la Cité Descartes à Champs-sur-Marne avec Amir que je remercie aussi.

Je souhaite aussi remercier plusieurs collègues du laboratoire avec qui je n'ai pas directement travaillé, mais avec qui j'ai eu de nombreuses discussions sur des sujets variés, qui m'ont souvent permis de déconnecter et de partager des moments conviviaux : Raphaël, Sami et Christian.

Pour terminer, je tiens à remercier les membres de l'équipe SoDUCo, avec qui j'ai travaillé durant les trois années précédant ma thèse. Ce projet a largement contribué à l'émergence et à la construction de ce travail doctoral. Je remercie en particulier mes encadrants de thèse (Julien, Nathalie et Bertrand), ainsi que Solenn, mais aussi Joseph, Pascal, Julie, Stéphane, Edwin et Melvin, et plus largement l'ensemble des personnes impliquées dans le projet.

Derniers remerciements

J'aimerais également remercier les communautés contributives de projets tels que Wikimedia ou OpenStreetMap, auxquelles je participe et qui reposent sur l'engagement de nombreuses personnes œuvrant à la constitution et à l'amélioration de bases de données

ouvertes. Ces ressources m'ont d'ailleurs été utiles à plusieurs reprises dans le cadre de mes travaux. J'adresse également une pensée particulière à Benoît P., avec qui j'ai partagé quelques moments de contribution.

Pour conclure ces remerciements, je souhaite adresser une pensée à celles et ceux que j'aurais pu oublier. Je tiens également à remercier une nouvelle fois toutes les personnes mentionnées précédemment pour leur soutien, leur présence et les moments de vie que nous avons partagés à différentes périodes de ma vie.

Abstract

Addresses serve as indirect spatial references, allowing a location to be identified without resorting to precise geographic coordinates. Geocoding is the process of associating coordinates with localized data using address labels. It is therefore based on a reference database of addresses associated with coordinates, and on an algorithm designed to find the address label that best matches a candidate address label in the reference database. However, traditional tools mainly use contemporary data, which often leads to failures for historical addresses. proposed settlement, in the absence of field data. The entire approach is tested on heterogeneous and fragmentary data representing Parisian addresses from the 18th century to the present day, in the Faubourg Saint-Antoine district.

Addresses serve as indirect spatial references, allowing a location to be identified without resorting to precise geographic coordinates. Geocoding is the process of associating coordinates with localized data using address labels. It is therefore based on a reference database of addresses associated with coordinates, and on an algorithm designed to find the address label that best matches a candidate address label in the reference database. However, traditional tools mainly use contemporary data, which often leads to failures for historical addresses.

This thesis therefore proposes an approach for constructing a geohistorical address repository, integrating these heterogeneous and fragmentary data from various historical sources. The main objectives are to represent addresses and the geographical entities to which they refer, taking into account their spatio-temporal dimension, and to integrate, standardize, match, and merge these fragmentary data. To do this, the thesis relies on a knowledge graph and ontology, *PeGazUs*, dedicated to describing geographic entities and their spatio-temporal evolution. It also proposes a method for populating the graph, which, in addition to integrating data from various sources, makes it possible to generate new knowledge in order to reconstruct the state of the territory at any given moment and for each geographical entity considered, and thus track their evolution over time. Finally, it proposes an original evaluation approach to assess the proposed population algorithm in the absence of ground truth. The entire approach is tested on heterogeneous and fragmentary data representing Parisian addresses from the 18th century to the present day in the Faubourg Saint-Antoine neighborhood.

Résumé

Les adresses jouent le rôle de références spatiales indirectes, permettant de localiser un lieu sans recourir à des coordonnées géographiques précises. Le géocodage est l'opération qui consiste à associer des coordonnées à des données localisées à l'aide de libellé d'adresses. Il repose donc sur un référentiel d'adresses associées à des coordonnées, et sur un algorithme visant à retrouver, pour un libellé d'adresse candidat, celui qui lui correspond le mieux dans le référentiel. Mais les outils classiques utilisent essentiellement des données contemporaines, ce qui conduit souvent à des échecs pour des adresses historiques.

Les adresses sont présentes en grande quantité dans différents documents historiques, tels que des plans, des annuaires, ou des documents administratifs, dont l'accès est aujourd'hui facilité par la mise en ligne des archives par les institutions patrimoniales. Parallèlement, des outils d'extraction automatique, notamment basés sur des réseaux de neurones, permettent d'extraire un grand nombre d'informations à partir de ces documents anciens. Ces informations constituent des données hétérogènes et fragmentaires fournissant des représentations des libellés d'adresses et de leurs coordonnées le plus souvent incomplètes, discontinues dans le temps et dans l'espace. Il est essentiel de structurer et intégrer ces données hétérogènes et fragmentaires dans un référentiel géohistorique cohérent, permettant la représentation spatio-temporelle des adresses et leur valorisation à des fins de géocodage.

Cette thèse propose donc une approche pour construire un référentiel géohistorique d'adresses, intégrant ces données hétérogènes et fragmentaires issues de sources historiques variées. Les objectifs principaux sont de représenter les adresses et les entités géographiques auxquelles elles se réfèrent, en tenant compte de leur dimension spatio-temporelle, et d'intégrer, normaliser, apparier et fusionner ces données fragmentaires. Pour ce faire, la thèse s'appuie sur un graphe de connaissances et une ontologie, *PeGazUs*, dédiés à la description des entités géographiques et de leurs évolutions spatio-temporelles. Elle propose également une méthode de peuplement du graphe, qui, en plus d'intégrer les données provenant de sources variées, permet de générer de nouvelles connaissances afin de reconstituer l'état du territoire à chaque instant et pour chaque entité géographique considérée, et ainsi suivre leur évolution au fil du temps. Elle propose enfin une approche d'évaluation originale pour évaluer l'algorithme de peuplement proposé, en l'absence de vérité terrain. L'ensemble de l'approche est testé sur des données hétérogènes et fragmentaires représentant des adresses parisiennes allant du XVIIIe siècle à nos jours, sur le quartier du Faubourg Saint-Antoine.

Table des matières

Remerciements **i**

Abstract **ix**

Résumé **x**

Liste des figures **xv**

Liste des tables **xix**

Liste des listings **xx**

Liste des algorithmes **xxii**

Liste des préfixes RDF **xxiii**

Publications **xxv**

Introduction **1**

1 Enjeux et objectifs **5**
1.1 Contexte 5
1.2 Motivations pour créer un référentiel géohistorique d'adresses 7
1.3 Objectifs de la thèse et verrous scientifiques 10

2 Sources et informations géohistoriques **15**
2.1 Notions préliminaires sur les données géohistoriques 15
2.2 Catégories de sources 16
2.3 Typologie des sources utilisées 17
2.3.1 Sources iconographiques 17
2.3.2 Sources textuelles 20
2.3.3 Sources numériques et structurées 22
2.4 Évaluation de la fiabilité des sources 23
2.5 Caractéristiques des informations géohistoriques 23
2.5.1 Les entités géographiques 24
2.5.2 Imperfection des données géohistoriques 24
2.6 Histoire des systèmes d'adressage à Paris 27

3 Représentation des dynamiques spatio-temporelles **30**
3.1 Modèle spatio-temporel 31
3.1.1 Modèles à base de snapshots 32

3.1.2 Modèle *Space-Time Composites* . . . 33
3.1.3 Modèle *Time-stamping* . . . 33
3.1.4 Modèle *Spatio-Temporal Object* . . . 34
3.1.5 *Three-Domain Model* (TDM) . . . 34
3.1.6 Modèles de trajectoires fondés sur l'identité . . . 35
3.1.7 Modélisation fondée sur les événements . . . 35
3.1.8 Bilan des modèles spatio-temporels . . . 36
3.2 Représentation et suivi des changements . . . 36
3.2.1 Modéliser les changements . . . 36
3.2.2 Catégorisation des changements . . . 40
3.3 Modèles de données pour représenter l'information spatio-temporelle . . . 42
3.3.1 Les bases de données relationnelles . . . 42
3.3.2 Les modèles de graphes . . . 42
3.3.3 Les graphes de connaissances . . . 42
3.4 Ontologies pour la représentation des évolutions des entités géographiques . 44
3.4.1 Ontologies TSN et TSN-Change . . . 44
3.4.2 Ontologie HHT : Historical Hierarchical Territory . . . 46

4 Une ontologie pour représenter les adresses et leurs évolutions 50
4.1 Méthodes de création d'ontologies . . . 50
4.2 Développement de l'ontologie . . . 51
4.2.1 Adresse . . . 52
4.2.2 Entité géographique et évolution . . . 57
4.2.3 Sources . . . 66

5 Intégration de données géohistoriques 69
5.1 Intégration de données . . . 69
5.1.1 Cadre général de l'intégration de données . . . 69
5.1.2 Transformation des données . . . 71
5.1.3 Mise en correspondance des données . . . 72
5.1.4 Fusion des données . . . 74
5.2 Identité d'une entité géographique . . . 75
5.2.1 Enjeux de l'identité d'une entité géographique . . . 75
5.2.2 Approches existantes . . . 76
5.2.3 Choix du critère d'identité . . . 77
5.3 Construction du référentiel . . . 79
5.3.1 Préparation des données . . . 79
5.3.2 Alignement et détection des changements . . . 79

6 Méthode de construction du graphe 83
6.1 Représentation des données issues des sources selon l'ontologie . . . 83
6.2 Intégration des ressources dans un graphe de faits . . . 86
6.2.1 Nature des liens pour les observations similaires . . . 86
6.2.2 Identification des observations représentant la même entité . . . 88
6.2.3 Propagation de liens aux attributs, relations, changements et événements . . . 90
6.3 Déduction des dates d'apparition et de disparition des entités et des relations 94
6.4 Reconstitution de l'évolution des attributs . . . 96
6.4.1 Représentation élémentaire de l'évolution des attributs . . . 96

6.4.2 Reconstitution de l'évolution des attributs à partir de leurs versions élémentaires . . . 100

7 Mise en œuvre, résultats et évaluation 106
7.1 Le Faubourg Saint-Antoine comme cas d'étude . . . 106
7.1.1 Sources cartographiques . . . 107
7.1.2 Sources textuelles . . . 108
7.1.3 Sources numériques . . . 109
7.2 Mise en application de la méthode de peuplement . . . 111
7.2.1 Choix technologiques pour la mise en œuvre . . . 111
7.2.2 Transformation des données selon l'ontologie . . . 111
7.2.3 Création du graphe de faits . . . 114
7.3 Vérification de la cohérence du graphe . . . 116
7.3.1 Contraintes sur les ressources . . . 117
7.3.2 Contraintes sur les propriétés . . . 119
7.4 Évaluation de l'ontologie . . . 119
7.4.1 Modelet des adresses . . . 120
7.4.2 Modelet des évolutions temporelles . . . 121
7.4.3 Modelet des sources . . . 122
7.5 Évaluation de la méthode de peuplement de l'ontologie . . . 123
7.5.1 Stratégie d'évaluation . . . 123
7.5.2 Construction du jeu de données de référence et des jeux de données fragmentaires . . . 124
7.5.3 Méthodes d'évaluation du peuplement de l'ontologie . . . 126
7.5.4 Résultats de l'évaluation . . . 131

8 Discussion et perspectives 135
8.1 Représentation des incertitudes dans l'ontologie . . . 135
8.1.1 Incertitudes et provenance des données . . . 135
8.1.2 Enrichissement collaboratif . . . 136
8.1.3 Représentation des imperfections géométriques . . . 137
8.2 Limites de la mise en œuvre du peuplement de l'ontologie . . . 137
8.3 Critiques autour du critère d'identité . . . 139
8.4 Utiliser les relations spatiales entre entités géographiques pour améliorer la localisation . . . 139
8.5 Vers une exploitation plus fine des événements . . . 141
8.6 Extraction d'information dans les sources textuelles . . . 142
8.7 Limites de l'évaluation proposée . . . 143
8.7.1 Limitation du périmètre expérimental . . . 143
8.7.2 Absence d'évaluation avec des recouvrements temporels . . . 143
8.7.3 Absence de test de robustesse face à l'ajout de données contradictoires 144
8.7.4 Nombre limité d'expérimentations pour l'évaluation de robustesse . 144

Conclusion 146

A Exploitation approfondie des documents textuels 151
A.1 Méthodes pour l'extraction d'informations textuelles . . . 151
A.1.1 Méthode PURE (BERT) . . . 151
A.1.2 Méthode Text2KGBench . . . 152
A.1.3 Méthode hybride BERT + LLM . . . 152

A.2 Données et vérité terrain . . . 153
A.3 Résultats . . . 154
A.4 Discussion . . . 155

B Questions de compétence 156
B.1 Questions du modelet des adresses . . . 156
B.1.1 Liste d'adresses situées le long d'une voie donnée . . . 156
B.1.2 Coordonnées d'une adresse à partir de son libellé . . . 157
B.1.3 Liste d'adresses situées dans une zone géographique . . . 157
B.2 Questions du modelet des entités géographiques et de leurs évolutions . . . 158
B.2.1 Entités géographiques existent à une date donnée . . . 158
B.2.2 Temps valide d'une adresse . . . 159
B.2.3 Changements de géométrie affectant une voie . . . 159
B.2.4 Évolution de la géométrie d'une voie . . . 160

C Algorithmes pour la construction de graphes et la mise en œuvre 161
C.1 Algorithmes de la méthode de construction du graphe . . . 161
C.1.1 Initialisation des entités et propagation des liens . . . 161
C.1.2 Inférence des dates d'apparition et de disparition . . . 163
C.1.3 Reconstitution de l'évolution des attributs . . . 164
C.2 Mise en œuvre de la méthode . . . 166

Bibliographie 169

Liste des figures

1.1 Mise en correspondance des informations dans différentes sources décrivant le *10 rue Mâcon, Paris*. La carte est extraite de l'atlas de Jacoubet datant des années 1830 et les extraits d'annuaires sont issus de l'édition 1793 de l'*Almanach national de France* (Testu, 1793) et de l'édition 1844 de l'*Annuaire de la typographie parisienne et départementale* (Prétot, 1844). . 8

1.2 Extraits de l'édition 1813 de l'*Almanach des marchands et fabricans en cuirs, de la ville de Paris* (Lemoine, 1813). Les entrées de l'annuaire surlignées concernent des adresses faisant référence à des voies n'existant plus en 2026. 9

1.3 Résultat d'opérations de géocodage de l'adresse *10 rue Mâcon, Paris* via différents outils. ■ Localisation réelle de l'adresse au milieu du XIXe siècle, ■ Localisation de l'adresse selon un géocodeur. 10

1.4 Représentation de différentes rues de Mâcon dans Paris à deux périodes distinctes : au début du XIXe siècle dans le plan de Jacoubet à gauche (Jacoubet, 1836) et à la fin de ce même siècle dans l'atlas administratif de 1870 à droite (Service du plan de Paris et Service des eaux et de l'assainissement, 1870). 10

1.5 Illustration des différentes formes de fragmentation des données géographiques : spatiale, temporelle et attributaire. 11

1.6 Illustration des différentes formes d'hétérogenéité des données géographiques : spatiale, temporelle et attributaire. 12

2.1 Typologie des sources géohistoriques. 17

2.2 Trois rendus cartographiques d'un même territoire (autour de la fontaine Saint-Michel à Paris) à partir des mêmes données OpenStreetMap, mais avec des styles graphiques différents. Crédit : © OpenStreetMap contributors, sous licence ODbL. 18

2.3 Extraits de deux plans de la ville de Paris illustrant le quartier autour de la rue Mâcon. 19

2.4 Exemple de quatre documents textuels. 21

2.5 Exemple d'une même adresse (10 rue Mâcon, Paris) représentée dans différents modèles et formats : CSV, GeoJSON et RDF (Turtle). 23

2.6 Illustration des aspects de l'imperfection des données autour de la rue Mâcon dans l'atlas de Jacoubet. ■ Incertitude, ■ Imprécision, ■ Ambiguïté, ■ Incomplétude. 26

2.7 Les trois systèmes de numérotation des habitations utilisés à Paris. → : sens de numérotation. 27

3.1 Processus pour la reconstruction d'un référentiel géohistorique d'adresses. . 30

3.2 Comparaison des principaux modèles spatio-temporels. 31

3.3 Comparaison des principaux modèles de changement. 37
3.4 Représentation de la relation *Paris est la capitale du département de la Seine entre 1790 et 1968* avec l'approche 4d-fluent. 40
3.5 Typologie des processus spatio-temporels utilisée par Dumenieu (2015). . . 40
3.6 Typologie simplifiée des changements proposée par Bernard *et al.* (2017). . 41
3.7 Représentation simplifiée de l'ontologie TSN. 44
3.8 Représentation simplifiée de l'ontologie TSN-Change. 45
3.9 Représentation simplifiée de l'ontologie *HHT*. 47

4.1 Représentation simplifiée de l'ontologie *locn*. 53
4.2 Représentation simplifiée de l'ontologie des adresses. 55
4.3 Représentation pour l'adresse numérotée « *46 rue Gérard, 75013 Paris* ». Adresse, Entité géographique, Segment d'adresse (dont le segment final). 57
4.4 Représentation d'une adresse descriptive. Adresse, Entité géographique, Segment d'adresse (dont le segment final). 58
4.5 Schéma de la partie de l'ontologie décrivant une entité géographique. . . . 60
4.6 Exemple de modélisation pour la *rue Gérard*. `Landmark`, `Attribute`, `AttributeVersion`, `LandmarkRelation`, `skos:Concept`. 60
4.7 Schéma de la partie de l'ontologie décrivant la dynamique des entités géographiques. L'élément lié à la classe `Change` par le prédicat `appliedTo` décrit l'union de `Attribute`, de `Landmark` et de `LandmarkRelation`. 61
4.8 Exemple de représentation de l'évolution conjointe de deux voies : création de la *rue du Père Guérin* par démembrement de la *rue Gérard*. `Landmark`, `Attribute`, `AttributeVersion`, `Change`, `Event`, `TimeInstant`. . 63
4.9 Schéma du modèle central d'entités temporelles de l'ontologie OWL-Time. 65
4.10 Schéma représentant les 13 relations de l'algèbre des intervalles d'Allen avec le vocabulaire de OWL-Time. 66
4.11 Représentation de la partie de l'ontologie permettant de décrire des entités temporelles (`TemporalEntity`). 66
4.12 Représentation de l'intervalle temporel `it` défini par *entre 20 nivôse an IV et la fin du XIXe siècle*. Nivôse est le quatrième mois du calendrier républicain. `TimeInterval`, `TimeInstant`, `xsd:dateTimeStamp`, `time:TRS`, `time:TemporalUnit`. 67

5.1 Processus d'intégration de données provenant d'un ensemble de sources hétérogènes $S = \{s_1, \ldots, s_6\}$. Chaque source s_i est transformée en un référentiel normalisé n_i à l'aide d'un mapping m_i conforme à un vocabulaire commune, produisant l'ensemble $N = \{n_1, \ldots, n_6\}$. Ces derniers sont ensuite intégrés pour construire le référentiel global G. Par souci de lisibilité, seules les correspondances entre les éléments de n_1, n_2 et ceux de G sont représentées. 71
5.2 Évolution des emprises des quais situés au sud du Palais du Louvre à Paris de 1790 à 2013. 75
5.3 Schéma indiquant la manière dont sont générés les changements simples à partir du tri temporel des versions pour la méthode de Charles et Hernandez (2023). Version d'une unité territoriale, Changement inféré, → *before*, → *after*. 81

6.1 Initialisation du graphe de faits intégrant les entités réelles résultant du croisement des sources dont les données sont contenues des graphes de factoïdes. Les entités du graphe de faits sont reliées à ses observations dans les sources avec le lien `hasTrace`. ■ Graphe de factoïdes, ■ Graphe de faits. 88
6.2 Exemple de génération de liens de correspondance pour des entités de type `ltype:StreetNumber`. ■ Graphe de factoïdes, ■ Graphe de faits, ■ `Landmark`, ■ `LandmarkRelation`, ■ `skos:Concept`. 91
6.3 Exemple de génération de liens de correspondance pour des attributs. ■ Graphe de factoïdes, ■ Graphe de faits, ■ `Landmark`, ■ `Attribute`, ■ `skos:Concept`. 92
6.4 Exemple de génération de liens d'équivalences pour des changements. ■ Graphe de factoïdes, ■ Graphe de faits, ■ `Landmark`, ■ `Attribute`, ■ `Change`, ■ `skos:Concept`. 93
6.5 Reconstitution des états spatio-temporels de la rue Gérard à partir de ses observations. 95
6.6 Agrégation des versions des différents attributs de la rue Gérard. 96
6.7 Création de l'ensemble de changements Γ et de versions B (frise du bas) à partir de versions et de changements. La frise du haut montre l'agencement temporel des éléments de C et V pour l'attribut "géométrie" de la rue Gérard. ■ Version avec un temps valide, ■ Version élémentaire, → Version décrite dans un changement, ■ Changement, ■ Changement fictif, ■ Changement élémentaire. 98
6.8 Affiliation des changements et des versions pour l'attribut géométrie de la rue Gérard. ■ Version avec un temps valide, ■ Version élémentaire, ■ Changement, ■ Changement élémentaire, → `hasTrace`. 99
6.9 Situation où la propagation de l'affiliation des versions est nécessaire. ■ Version avec un temps valide, ■ Version élémentaire, ■ Changement élémentaire, ■ Changement fictif. 100
6.10 Traitement des versions non tracées selon les changements auxquels il est lié. ■ Version avec un temps valide, ■ Version élémentaire, ■ Changement, ■ Changement élémentaire. 102
6.11 Résultat de la suppression de versions lacunaires pour l'attribut géométrie de la rue Gérard. ■ Version avec un temps valide, ■ Version élémentaire, ■ Changement, ■ Changement élémentaire, → `hasTrace`. 103
6.12 Reconstitution finale de l'évolution temporelle après fusion des versions similaires pour l'attribut géométrie de la rue Gérard. ■ Version avec un temps valide, ■ Version élémentaire, ■ Changement, ■ Changement élémentaire, → `hasTrace`. 104
6.13 Reconstitution des historiques des différents attributs de la rue Gérard : existence de la voie, relations spatiales (`Within` pour Gentilly et Paris), nom et géométrie. La frise combinée en bas représente la succession d'états obtenue par combinaison des cinq historiques. ■ État de non-existence, ■ État d'existence, ■ Version d'attribut, ■ État d'entité, ■ Changement. . . 105

7.1 Localisation du quartier du Faubourg Saint-Antoine au sein de Paris. . . . 107
7.2 Diversité des transformations urbaines dans le quartier du Faubourg Saint-Antoine entre le début du XIXe siècle (plan de Jacoubet) et 2026 (OpenStreetMap). 108

7.3 Exemple de données issues de deux sources cartographiques (atlas de Jacoubet et atlas municipal de 1888) autour de la place des Vosges. Les points indiquent les numéros d'immeuble et les lignes les voies de communication. 109
7.4 Localisation des adresses dont la cible est située le long de la rue du Pas de la Mule. 121
7.5 Schéma expliquant la stratégie d'évaluation. 123
7.6 Construction de *lignes de vie* d'adresses autour de la *rue du Pas de la Mule*. Les lignes rouges représentent les liens entre versions successives non similaires, tandis que les lignes vertes correspondent à des versions successives similaires. 125
7.7 Construction de *lignes de vie* de l'adresse *2 rue du Pas de la Mule*. → : *a une géométrie différente de*, → : *a une géométrie similaire à*. 125
7.8 Lignes de vie de l'adresse *2 rue du Pas de la Mule* obtenues de trois manières : construction du jeu de référence R, peuplement du graphe G avec la méthode proposée, et peuplement du graphe G' avec la même méthode mais avec l'ajout de données fragmentaires. → : *a une géométrie différente de*, → : *a une géométrie similaire à*. 127
7.9 Comparaison de deux lignes de vie de l'adresse *2 rue du Pas de la Mule* obtenues avec la méthode de peuplement proposée. La situation de référence est obtenue à partir des observations initiales, tandis que la situation perturbée est générée à partir des mêmes observations, auxquelles est ajouté un changement fictif c_α. → : *a une géométrie différente de*, → : *a une géométrie similaire à*. 132
7.10 Outil de visualisation montrant l'évolution des entités géographiques, ici focalisé sur la *Rue de Cotte*. La géométrie affichée est valide de 1827 à 1836. 133
7.11 Outil de visualisation montrant un instantané du secteur autour de la *Rue de Cotte* en 1896. La carte de fond est le plan municipal de 1888. 134

8.1 Exemple d'évolution cadastrale où la parcelle initiale numérotée 1170 est subdivisée en plusieurs entités qui conservent le même identifiant. 140
8.2 Estimation de la position de la cible de l'adresse dont le label est « *au centre de la capitale, entre le Palais-Royal et les Tuileries, à proximité des principaux théâtres* » au début du XIXe siècle. 140
8.3 Proposition d'une typologie des liens de filiation en fonction des événements. 142
8.4 Proposition d'une chaîne de traitement pour un test d'ablation. 143
8.5 Comparaison de deux lignes de vie de l'adresse *2 rue du Pas de la Mule* obtenues avec la méthode de peuplement proposée. La situation de référence est obtenue à partir des observations initiales, tandis que la situation perturbée est générée à partir des mêmes observations, auxquelles est ajouté une observation c_δ décrivant une version fictive contradictoire avec l'évolution attendue. → : *a une géométrie différente de*, → : *a une géométrie similaire à*. 144

A.1 Extraction d'entités et de relations au sein de la phrase « Cette avenue était une partie de la route de Fontainebleau », issue du dictionnaire d'Hillairet traitant de l'avenue d'Italie. 152
A.2 Schéma expliquant les étapes de la génération des triplets RDF avec des LLM à partir d'une ontologie et d'un corpus de textes. 153

Liste des tables

1 Préfixes spécifiques à l'ontologie PeGazUs. xxiii
2 Préfixes de vocabulaires généraux RDF. xxiii
3 Préfixes d'autres ontologies utilisées. xxiv

2.1 Typologie de l'imperfection des informations géographiques historiques proposée par de Runz (2008). 25

7.1 Extrait d'un fichier de la BAN (Base Adresse Nationale) pour la ville de Paris, incluant les numéros d'adresse, compléments éventuels, noms de voies, codes postaux, communes et coordonnées géographiques (latitude, longitude). 110
7.2 Principales propriétés de l'ontologie *PeGazUs*, avec les types de sujets et d'objets correspondants. 119
7.3 Scores de cohérence structurelle et temporelle mesurant l'impact des données fragmentaires sur la reconstruction de l'évolution des adresses. 132

A.1 Exemple simplifié du jeu de données avec les phrases historiques associées aux rues. 154
A.2 Comparaison des résultats obtenus par les trois méthodes selon les métriques utilisées par Mihindukulasooriya *et al.* (2023). Les meilleurs scores pour chaque métrique sont indiqués en gras. 154

Liste des listings

4.1 Requête SPARQL relevant les changements de nom. 62

6.1 Représentation RDF en Turtle de la rue Gérard du plan municipal de 1888 selon l'ontologie *PeGazUs*. Le préfixe `ex:` désigne un espace de nom des ressources, et la ressource `ex:1887_1889` décrit l'intervalle temporel correspondant à la période [1887; 1889]. 84

6.2 Représentation RDF en Turtle d'un événement associé à la rue Gérard (formation en 1851) affirmée par le dictionnaire d'Hillairet. Le préfixe `ex:` désigne un espace de nom des ressources, et la ressource `ex:1851` décrit l'instant correspondant à l'année 1851. 85

7.1 Requête SPARQL fusionnée permettant de sélectionner les identifiants OSM des numéros d'immeubles de Paris (dont l'identifiant wikidata est Q90), leur voie d'appartenance et l'arrondissement, leur label et leur géométrie WKT. 110

7.2 Exemple de description d'adresses en JSON compact pour des données issues de la BAN. Les attributs des `landmarks` et les provenances des éléments ne sont pas tous indiqués dans cet exemple. 113

7.3 Exemple de description de l'événement « *Par arrêté municipal du 30 août 1978, sa portion orientale, de la rue Bobillot à la rue du Moulin-des-Prés, prend le nom de rue du Père-Guérin.* » en JSON provenant de Wikipédia. 114

7.4 Requête SPARQL sélectionnant les adresses et les coordonnées associées pour lesquelles la cible est située le long de la rue du Pas de la Mule. . . . 120

7.5 Requête SPARQL sélectionnant les adresses et les coordonnées associées pour lesquelles la cible est située dans la zone définie par `?area`. 121

7.6 Requête SPARQL identifiant la source d'une information donnée. 122

7.7 Requête SPARQL sélectionnant les informations descriptives d'une source. 123

7.8 Requête SPARQL identifiant le nombre de version d'attributs de type nom pour des cibles d'adresses. `ex:facts` désigne le graphe nommé des faits. . 128

B.1 Requête SPARQL pour identifier les adresses le long d'une voie donnée. . . 156

B.2 Extrait du résultat en CSV de la requête SPARQL du Listing B.1 identifiant les adresses le long de la rue du Pas de la Mule. Le résultat renvoie 13 lignes. 156

B.3 Requête SPARQL pour obtenir les coordonnées d'une adresse à partir de son libellé. 157

B.4 Résultat en CSV de la requête SPARQL du Listing B.3 identifiant les coordonnées associées à l'adresse dont le libellé est « 14 Place des Vosges, 75004 Paris 4e Arrondissement ». 157

B.5 Requête SPARQL pour lister les adresses situées dans une zone géographique donnée. 157

B.6 Extrait du résultat (3 sur 62 lignes) de la requête SPARQL du Listing B.5 retournant les adresses dont les coordonnées associées sont situées dans la zone définie. 158
B.7 Requête SPARQL pour identifier les voies existantes à une date donnée. . 158
B.8 Extrait du résultat de la requête SPARQL du Listing B.7 identifiant les entités géographiques existant en 1860. L'exécution renvoie 1355 résultats au total. 159
B.9 Requête SPARQL pour déterminer la période de validité d'une adresse. . . 159
B.10 Requête extrayant les événements de changement de géométrie d'une voie. 159
B.11 Extrait du résultat CSV de la requête SPARQL du Listing B.10. Les géométries (`LINESTRING`, `MULTILINESTRING`, `POLYGON`) sont abrégées pour la lisibilité. 160
B.12 Requête SPARQL pour retracer l'histoire des géométries successives d'une voie donnée. 160

Liste des algorithmes

C.1 Initialisation des entités du graphe des faits 161
C.2 Propagation des liens de provenance vers relations, attributs, changements et événements 162
C.3 Inférence des dates d'apparition et de disparition d'une entité ou relation. La fonction *estimation_contrainte_temporelle()* est introduite dans l'algorithme C.4. 163
C.4 Fonction *estimation_contrainte_temporelle()* : calcule et génère les triplets temporels pour un événement à partir des contraintes observées. 163
C.5 Initialisation des changements et des versions élémentaires pour un attribut a 164
C.6 Affiliation des changements et versions aux factoïdes pour un attribut a . . 164
C.7 Suppression des versions lacunaires 165
C.8 Fusion des versions élémentaires similaires successives 165
C.9 Construction du graphe de faits et reconstitution des attributs 166
C.10 Fonction *versions_noms_similaires()* : test de similarité entre versions de type nom 167
C.11 Fonction *versions_geometries_similaires()* : test de similarité entre versions de type géométrie 167

Liste des préfixes RDF

Dans cette thèse, lorsqu'une ressource n'est associée à aucun préfixe, cela signifie que l'espace de nom de cette ressource est celui associé au préfixe `addr`.

Préfixe	URI associée
Ontologie PeGazUs	
addr	http://rdf.geohistoricaldata.org/def/address#
atype	http://rdf.geohistoricaldata.org/id/codes/address/attributeType/
ctype	http://rdf.geohistoricaldata.org/id/codes/address/changeType/
ltype	http://rdf.geohistoricaldata.org/id/codes/address/landmarkType/
lrtype	http://rdf.geohistoricaldata.org/id/codes/address/landmarkRelationType/

Tableau 1 – Préfixes spécifiques à l'ontologie PeGazUs.

Préfixe	URI associée
Vocabulaires généraux	
dcterms	http://purl.org/dc/terms/
foaf	http://xmlns.com/foaf/0.1/
geo	http://www.opengis.net/ont/geosparql#
geof	http://www.opengis.net/def/function/geosparql/
locn	https://www.w3.org/ns/locn
ogc	http://www.opengis.net/rdf#
owl	http://www.w3.org/2002/07/owl#
prov	http://www.w3.org/ns/prov#
rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#
rdfs	http://www.w3.org/2000/01/rdf-schema#
sh	http://www.w3.org/ns/shacl#
skos	http://www.w3.org/2004/02/skos/core#
time	http://www.w3.org/2006/time#
wd	http://www.wikidata.org/entity/
wdt	http://www.wikidata.org/prop/direct/
xsd	http://www.w3.org/2001/XMLSchema#

Tableau 2 – Préfixes de vocabulaires généraux RDF.

Préfixe	**URI associée**
Autres ontologies	
hht	https://w3id.org/HHT#
hhtChange	https://w3id.org/HHT/Change#
osm	https://www.openstreetmap.org/
osmkey	https://www.openstreetmap.org/wiki/Key:
osmrel	https://www.openstreetmap.org/relation/
osmrdf	https://osm2rdf.cs.uni-freiburg.de/rdf/member#
tsn	http://purl.org/net/tsn#
tsnChange	http://purl.org/net/tsnchange#

TABLEAU 3 – Préfixes d'autres ontologies utilisées.

Publications

Conférences internationales

1. Charly Bernard, Solenn Tual, Nathalie Abadie, Bertrand Duménieu, Julien Perret, Joseph Chazalon.
   PeGazUs : A knowledge graph based approach to build urban perpetual gazetteers.
   Dans *International Conference on Knowledge Engineering and Knowledge Management (EKAW 2024)*, November 2024, Amsterdam, Netherlands. pp. 364–381.
   DOI : 10.1007/978-3-031-77792-9_22 — HAL : hal-04721538
2. Charly Bernard, Nathalie Abadie, Bertrand Duménieu, Julien Perret.
   Reconstructing the Temporal Evolution of Geographic Entities from Fragmentary Knowledge.
   Dans *7th International Conference on Knowledge Graphs and Semantic Web (KGSWC 2025)*, November 2025, Leipzig, Germany. pp. 192–207.
   DOI : 10.1007/978-3-032-13109-6_14 HAL : hal-05308831

Conférences nationales

1. Charly Bernard, Nathalie Abadie, Julien Perret, Bertrand Duménieu.
   Création d'un référentiel géo-historique d'adresses à partir de sources multiples.
   Dans *GAST — Gestion et l'Analyse de données Spatiales et Temporelles*, January 2024, Dijon, France.
   HAL : hal-04490732
2. Charly Bernard, Nathalie Abadie, Bertrand Duménieu, Julien Perret.
   PeGazUs : une méthode de reconstitution de l'évolution des entités géographiques à partir de données hétérogènes et fragmentaires.
   Dans *Ingénierie des Connaissances*, July 2025, Dijon, France.
   HAL : hal-05131104

Communications

1. Charly Bernard, Nathalie Abadie, Bertrand Duménieu, Julien Perret.
   Évaluation d'un graphe de connaissances géo-historique d'adresses créé à partir de sources multiples.
   Dans *SemWeb.Pro 2024*, November 2024, Paris, France.
   HAL : hal-05304293
2. Charly Bernard, Solenn Tual, Nathalie Abadie, Bertrand Duménieu, Julien Perret, Joseph Chazalon.
   PeGazUs : Une approche fondée sur les graphes de connaissances pour construire des gazetiers urbains perpétuels.

Dans *Journée du GdR IASIS — Détection de changement dans les données géo-spatiales*, March 2025, Paris, France.
HAL : hal-05304303

Posters

1. Charly Bernard, Nathalie Abadie, Julien Perret, Bertrand Duménieu.
Ontologie et méthodologie de peuplement d'un graphe de connaissances géo-historique d'adresses.
Dans *Symposium MaDICS*, May 2024, Blois, France.
HAL : hal-04601077

Introduction

Motivations

L'adresse constitue une information fondamentale dans l'étude des évolutions territoriales. Elle joue un rôle clé dans la localisation, la gestion administrative, la valorisation patrimoniale et l'analyse historique de données sur les espaces urbains. Or, l'adresse n'est pas une information stable car elle évolue au gré des transformations urbaines, des changements toponymiques ou des réorganisations administratives. Modéliser et reconstituer l'évolution des adresses dans le temps représente ainsi un défi majeur, à la fois conceptuel et technique. En effet, une adresse renvoie à des *entités géographiques* (bâtiments, voies, parcelles, etc.) et décrit des *relations* entre elles, qu'elles soient spatiales (inclusion, voisinage) ou thématiques (fonction, hiérarchie). L'évolution des adresses dépend donc directement de celle de ces entités et de leurs relations, dont il s'agit également de reconstituer la dynamique. Or, en l'absence de jeux de données décrivant de manière complète et continue les évolutions des entités géographiques, la reconstitution de l'évolution des adresses doit s'appuyer sur des sources multiples, hétérogènes, souvent incomplètes et fragmentaires, qu'il convient d'intégrer et de réconcilier.

Objectifs de la thèse

Cette thèse vise à proposer une approche méthodologique pour la construction d'un *référentiel géohistorique d'adresses*, c'est-à-dire un cadre cohérent permettant de représenter des entités géographiques issues de sources variées et leurs évolutions. Elle s'attache à répondre à trois questions principales :

- Comment représenter les adresses de manière à prendre en compte leur dimension spatio-temporelle ?
- Comment intégrer, normaliser, apparier et fusionner des données hétérogènes et fragmentaires issues de différentes sources dans ce modèle spatio-temporel ?
- Comment évaluer la robustesse et la pertinence d'une telle méthode de construction, en l'absence de vérité terrain ?

Contributions principales

Pour répondre à ces objectifs, la thèse s'appuie sur trois contributions majeures :

- une **ontologie**, appelée *PeGazUs*, dédiée à la description d'entités géographiques et d'adresses, intégrant leurs évolutions spatio-temporelles ;
- une **méthode de peuplement d'un graphe de connaissances géohistorique d'adresses** à partir de données hétérogènes et fragmentaires, fondée sur cette ontologie ;

— une **stratégie d'évaluation** de la méthode proposée, incluant un test de robustesse par l'ajout de données fragmentaires synthétiques.

Contexte de la recherche

Les travaux s'inscrivent dans le champ de la modélisation spatio-temporelle et de l'intégration de données géohistoriques, à l'intersection de la géomatique, du web sémantique et des humanités numériques.

Ils reposent sur l'utilisation de *graphes de connaissances*, c'est-à-dire une base de connaissances sous forme de graphes. On peut les définir comme des structures de données organisées sous forme de triplets (sujet–prédicat–objet), permettant de représenter des entités, leurs propriétés et les relations qu'elles possèdent. Ces graphes offrent un cadre adapté à l'intégration de données hétérogènes et fragmentaires. L'interrogation et la mise à jour de ces graphes sont réalisées à l'aide de requêtes *SPARQL* (*SPARQL Protocol and RDF Query Language*). Ce langage standard du Web sémantique permet, d'une part, d'extraire des sous-ensembles du graphe afin de filtrer et de croiser des informations structurées et, d'autre part, d'ajouter ou de supprimer des triplets.

Des travaux existants s'appuient également sur des graphes de connaissances pour modéliser des données géohistoriques, en proposant à la fois une structure conceptuelle pour organiser les informations et une méthode pour alimenter cette structure à partir d'un ensemble de données. Notre approche se distingue par la prise en compte de données hétérogènes et fragmentaires dont l'ensemble est incomplet, imprécis et peut contenir des contradictions.

Le cas d'étude porte sur la ville de Paris, territoire sélectionné pour la richesse et la diversité de ses sources historiques ainsi que pour la variété des modes de référencement des adresses dans ces sources. Du fait de l'importance des sources produites entre la fin du XVIIIe siècle et aujourd'hui, nous retenons cette période d'étude.

Structure du manuscrit

Ce manuscrit s'articule autour de huit chapitres.

Le **chapitre 1** présente le contexte dans lequel s'inscrit cette thèse afin de mettre en avant l'importance de constituer un référentiel géohistorique d'adresses. Du fait de la difficulté de la construction d'une telle base, le chapitre approfondit les objectifs de la thèse et met en avant les verrous scientifiques associés. En parallèle, du fait que le cas d'application concerne la ville de Paris qui a connu différents systèmes d'adressage, nous présentons ces derniers et chacune de leurs spécificités afin de mettre en avant la complexité de ce que peut être une adresse. Enfin, ce chapitre positionne la thèse dans le domaine de l'intégration de données, en abordant les problématiques de transformation, de normalisation, de liage et de fusion de données hétérogènes. Dans cette optique, puisque les données traitées décrivent des entités géographiques, il est essentiel de définir leur identité afin de préciser la manière dont s'opère leur liage au sein du référentiel.

Ensuite, le **chapitre 2** décrit les sources et les informations géohistoriques, servant de base à la construction de notre référentiel. Il débute par la définition des termes entrant en jeu dans cette thèse et l'introduction des concepts associés. Puis, nous présentons les différents types de sources que nous utilisons (cartographiques, textuelles, numériques) avant de nous interroger à propos de la fiabilité de ces dernières. Ce chapitre se termine sur

les caractéristiques des informations géohistoriques contenues dans les sources présentées, en particulier nous abordons le problème de leurs imperfections.

Le **chapitre 3** présente un état de l'art sur la réprentation des dynamiques spatio-temporelles structuré autour de plusieurs axes. Il aborde d'abord les modèles spatio-temporels, puis les approches de représentation et de suivi des changements des entités géographiques. Le chapitre s'intéresse ensuite aux différents supports de l'information spatio-temporelle, tels que les bases de données, les graphes et les graphes de connaissances. Il se poursuit par une analyse des travaux visant à proposer un schéma de données adapté à la modélisation de l'évolution des entités géographiques. Nous nous attachons notamment à décrire en détail les approches *TSN* et *HTT*, qui proposent chacune une ontologie et une méthode de construction de l'évolution des entités géographiques. Enfin, il examine les méthodes de construction d'un référentiel géohistorique intégrant ces dimensions spatio-temporelles.

Le **chapitre 4** introduit **PeGazUs** (*PErpetual GAZetteer of approach-address UtteranceS*), une ontologie qui vise à fournir un modèle de représentation spatio-temporelle des entités géographiques et des adresses, adapté à l'intégration de données fragmentaires et hétérogènes. Nous l'avons développée à l'aide de la méthode SAMOD qui conseille de diviser la problématique en plusieurs parties appelées modelets. Ainsi, pour le développement de l'ontologie, nous présentons les trois modelets suivants :

- **adresses** : ce modelet décrit la manière dont les adresses, dans leur hétérogénéité, doivent être représentées ;
- **entités géographiques et évolutions spatio-temporelles** : cette partie définit comment chaque entité doit être modélisée tout en y intégrant la dimension temporelle ;
- **sources** : on décrit ici la manière dont chaque information (adresse, entité géographique) doit être reliée à sa source de provenance.

Le **chapitre 5** s'intéresse à l'intégration de données géohistoriques nécessaires au peuplement du graphe de connaissances. Après avoir défini un modèle de représentation et une méthode de construction, il devient indispensable de traiter la diversité et l'hétérogénéité des sources décrivant les entités géographiques dans le temps. Nous présentons d'abord les principes généraux de l'intégration de données, depuis la transformation et la normalisation des informations jusqu'à leur appariement et leur fusion. Nous abordons ensuite la notion d'identité des entités géographiques, en discutant les différentes approches existantes et le critère retenu pour assurer la cohérence du référentiel construit. Enfin, nous détaillons les méthodes permettant de construire un référentiel géohistorique fiable, incluant l'alignement des données issues de sources multiples ainsi que la détection des changements spatio-temporels.

L'ontologie une fois introduite, celle-ci est associée à une méthode de peuplement d'un graphe géohistorique des adresses dans le **chapitre 6**. Cette approche débute par la représentation des données issues des sources anciennes en cohérence avec l'ontologie. Puis, nous présentons la manière dont les informations similaires issues de différentes sources doivent être liées entre elles afin d'initialiser un graphe de faits. Ce graphe résulte de la réconciliation des données issues de toutes les sources utilisées pour le peuplement du graphe. Nous terminons ce chapitre par la génération de nouveaux faits à partir de la combinaison des faits existants. Nous pouvons déduire, à partir des faits existants, les états spatio-temporels de chaque entité ainsi que les relations qu'elles entretiennent entre elles. À un niveau plus fin, nous reconstituons l'évolution des attributs des entités géographiques.

Le **chapitre 7** présente la mise en œuvre et l'évaluation de la méthode présentée précé-

demment ainsi que de l'ontologie développée. En préambule, nous présentons les données et les sources avec lesquelles nous appliquons la méthode de peuplement. De plus, nous expliquons la manière dont les données sont extraites des sources. Concernant l'évaluation, nous commençons par l'ontologie conformément à la méthode SAMOD. Puis, il convient de vérifier la cohérence des graphes construits à partir des différentes sources de données anciennes, c'est-à-dire de vérifier que les triplets construits (avant et après application de la méthode de peuplement spatio-temporel) soient en cohérence avec l'ontologie définie. Pour terminer, nous évaluons la méthode de peuplement en proposant une stratégie à base de données synthétiques qui d'une part consiste à comparer les données générées à un jeu de données de référence et qui d'autre part permet de tester la robustesse de la méthode face à l'ajout de données fragmentaires. La stratégie mise en place, nous présentons les résultats de l'évaluation et les analysons.

Enfin, le **chapitre 8** conclut ce manuscrit en ouvrant la discussion sur les limites et les perspectives des travaux menés. Il débute par la question de la représentation des incertitudes au sein de l'ontologie, en abordant celle de la provenance des données, des imperfections géométriques et des mécanismes d'enrichissement collaboratif. Le chapitre se poursuit par une analyse critique des limites identifiées lors de la mise en œuvre de la méthode de peuplement spatio-temporel, ainsi qu'une réflexion approfondie sur le critère d'identité des entités et son influence sur la représentation des dynamiques spatio-temporelles dans le référentiel. Enfin, il explore plusieurs pistes d'amélioration, notamment une meilleure exploitation des relations entre entités pour affiner la localisation, une prise en compte plus fine des événements dans la reconstitution des évolutions, et une exploitation à grande échelle des documents textuels, dont la richesse informationnelle offre un potentiel conséquent pour l'enrichissement du graphe de connaissances.

Chapitre 1

Enjeux et objectifs

Ce premier chapitre a pour objectif de poser le cadre général de cette thèse. Après avoir introduit, dans la section 1.1, le contexte et les enjeux liés à la construction d'un référentiel géohistorique des adresses et des entités géographiques qui les composent, nous explicitons dans la section 1.2 les motivations de la création d'un tel référentiel. La section 1.3 présente ensuite les objectifs de la thèse ainsi que les verrous scientifiques associés. Cette thèse propose une approche générale pour représenter les évolutions d'entités géohistoriques au moyen d'un graphe de connaissances et une méthodologie de peuplement du graphe à partir de sources historiques hétérogènes et fragmentaires. Sa conception est guidée par une classe d'entités géographiques particulières : les adresses. Elles présentent un intérêt particulier pour les travaux sur le passé, car elles permettent d'ancrer géographiquement un grand nombre d'informations historiques granulaires : données socio-économiques, démographiques ou encore corpus documentaires.

1.1 Contexte

Face à la disponibilité croissante des sources d'archives numérisées et aux progrès récents des approches d'extraction d'informations dans des documents textuels ou iconographiques, de nouvelles perspectives s'ouvrent quant à l'exploitation de ces sources, en particulier pour la production de données structurées décrivant les aspects du passé qu'elles documentent.

Accessibilité aux sources historiques Les sources d'archives numérisées ont, en effet, fait l'objet de vastes campagnes menées par les GLAM (Galeries, Bibliothèques, Archives et Musées) pour rendre accessibles en ligne leurs fonds auparavant consultables uniquement sur place. Des projets comme *Gallica*[1] (bibliothèque numérique de la Bibliothèque nationale de France et de ses partenaires), *Europeana*[2] ou encore des portails de consultation des archives nationales, départementales et communales accessibles en ligne, tels que celui de la ville de Paris[3], prouvent cette dynamique de mise à disposition de sources patrimoniales numérisées. Ces projets offrent aujourd'hui un accès facilité à des millions de documents : registres d'état civil, plans cadastraux, annuaires, cartes anciennes, journaux, documents administratifs ou correspondances privées. L'ouverture croissante des données patrimoniales s'accompagne de la mise à disposition de formats et de services

1. https://gallica.bnf.fr/
2. https://www.europeana.eu/fr
3. https://archives.paris.fr

(API, points d'accès SPARQL, serveurs IIIF), permettant une réutilisation automatisée et interopérable de ces contenus.

Progrès des technologies d'extraction d'informations Parallèlement, avec les progrès des réseaux de neurones profonds, les approches d'extraction d'informations dans des documents ont connu des avancées importantes. La mise à disposition de ces corpus patrimoniaux numérisés ouvre de nouvelles possibilités pour leur traitement automatique et l'extraction d'informations à grande échelle (Terras, 2015). Les méthodes de reconnaissance de texte ont progressé. Les systèmes de reconnaissance optique de caractères (*Optical Character Recognition* , OCR) récents, comme ceux utilisant *Tesseract* (Smith, 2007) sont capables de traiter des documents historiques présentant des irrégularités telles que des déformations, des caractères partiellement effacés ou des impressions de faible qualité. De plus, la transcription automatique de manuscrits historiques est possible grâce à l'émergence de la reconnaissance d'écritures manuscrites (*Handwritten Text Recognition*, HTR), notamment au travers de plateformes comme *Transkribus*[4] (Kahle *et al.*, 2017). Ces technologies s'accompagnent d'améliorations en segmentation de documents complexes comme des tableaux ou des cartes. Le traitement automatique du langage naturel (TALN) permet quant à lui d'identifier les entités (personnes, lieux, organisations, dates, adresses) et leurs relations dans des corpus textuels volumineux, notamment grâce aux approches modernes de reconnaissance d'entités nommées (Lample *et al.*, 2016; Seow *et al.*, 2025). Enfin, la vision par ordinateur appliquée aux sources iconographiques facilite l'extraction d'informations géographiques dans des cartes et plans anciens ou encore dans des photographies anciennes, par exemple pour la vectorisation automatique d'éléments cartographiques (Dhar et Chanda, 2006; Sarzotti *et al.*, 2025).

Impacts sur la recherche en histoire La disponibilité croissante de documents historiques, combinée aux progrès des technologies d'extraction d'informations, entraîne des changements pour la recherche en histoire. En effet, il est désormais possible de produire une quantité importante d'informations géohistoriques, susceptibles d'être valorisées par leur structuration. Cependant, cette production n'est pas aisée : les documents disponibles sont hétérogènes, tant par leur nature que par leur qualité, ce qui exclut l'existence d'une méthode unique d'extraction ou de traitement. De plus, les données produites demeurent souvent imparfaites, ce qui appelle à les interpréter avec prudence et à en évaluer la qualité. Enfin, les données extraites de documents anciens proviennent de sources de seconde main, nécessairement partielles et subjectives. Pour permettre aux historiens de tenir compte des éventuels effets de sources dans leurs analyses, il est indispensable de toujours conserver la trace de l'origine des donnés produites. Dans ce contexte, plusieurs initiatives récentes témoignent de l'intérêt croissant pour la valorisation de ces informations à grande échelle. Parmi elles, le projet *SoDUCo*[5] (Social Dynamics in Urban Context) réunit des chercheurs en sciences géographiques, en histoire urbaine, en morphogenèse des réseaux et en informatique. Il a pour objectif de développer des méthodes et des outils permettant l'extraction d'informations à partir de documents historiques, notamment à travers des dispositifs d'annotation de texte, de géoréférencement et de vectorisation de plans, ainsi que des outils de validation des données extraites. Ces travaux visent à constituer des corpus géohistoriques structurés pour l'étude des transformations de la structure spatiale urbaine et des dynamiques sociales et professionnelles de la population parisienne au fil du temps.

4. https://www.transkribus.org/
5. https://soduco.geohistoricaldata.org/

1.2 Motivations pour créer un référentiel géohistorique d'adresses

Qu'est-ce qu'une adresse ? Dans les métadonnées accompagnant la publication des données d'adresses de la ville de Paris [6], on peut lire que les adresses sont décrites comme

> « [des] données géographiques ponctuelles. L'adresse est constituée d'un numéro de voirie et du nom d'une rue dans une commune. Elle permet ainsi de localiser géographiquement un lieu. »

Cette courte définition révèle l'ambiguïté de la notion d'adresse, à la fois entité géographique et énoncé textuel structuré, appuyé sur une hiérarchie d'entités géographiques (la rue, la commune), désignant un lieu.

L'adresse est une réponse au besoin de désigner de manière univoque et non ambiguë un lieu pour s'y rendre ou y faire parvenir quelque chose (Davis et Fonseca, 2007). Dans les systèmes de numérotages créés en Europe à partir de la fin du XVIIIe siècle, les adresses – qu'on définit dans ce cas comme des *adresses numérotées* – sont conçues comme des énoncés autonomes décrivant l'emplacement d'un bâtiment à l'intérieur d'une hiérarchie spatiale urbaine, permettant une compréhension rapide du cheminement à faire jusqu'à lui. Le numéro d'adresse incarne ce double rôle d'identifiant et de guide de cheminement. Bien qu'abstrait, il est généralement matérialisé sur le terrain par une plaque indicatrice. Il indexe l'immeuble, tout en permettant de connaître sa position relative dans la rue. L'adresse-cheminement (ou *adresse descriptive*) préexistait aux systèmes de numérotage mathématisant l'espace. Les bâtiments étaient désignés pour leur nom ou leur position relative à des points d'intérêts connus ou des amers visuels, par exemple « *maison de la Belle-Jardinière, quai aux Fleurs, coin de la rue de la Cité* » ou « *au centre de la capitale, entre le Palais-Royal et les Tuileries, à proximité des principaux théâtres* ». Cette forme d'adressage persiste, soit pour pallier l'absence de numérotage, soit comme système d'usage courant, par exemple pour localiser un lieu par rapport à d'autres repères au sein de communautés où ces points de référence sont connus. En Afrique subsaharienne, dans des lieux où le nommage des rues n'est pas systématique et encore moins le numérotage des bâtiments, Dumedah (2021) indique que les habitants utilisent des références géographiques informelles pour décrire des lieux.

Les différents systèmes d'adressage contemporains ou anciens possèdent des points communs : les adresses qu'ils produisent contiennent des références à des points de repère liés par des relations spatiales (explicitées ou non). De plus, l'énoncé de l'adresse doit faire comprendre à son destinataire (des humains ou des machines pour les plus récentes) à quel lieu géographique elle fait référence. L'adresse est donc d'abord une référence spatiale indirecte : elle localise un lieu non pas par des coordonnées géographiques, mais par un ensemble structuré de relations hiérarchiques et spatiales avec d'autres entités géographiques nommées (Morita, 1980; Coetzee *et al.*, 2011).

Toutefois, sans doute à la faveur des systèmes d'information géographique, les adresses ne sont aujourd'hui plus seulement des désignations indirectes de l'espace mais sont considérées comme des entités géographiques localisées, c'est-à-dire des références spatiales directes. Stockées en bases de données numériques, elles sont munies de coordonnées géographiques pointant - selon les spécifications techniques - une boite aux lettres, un portail, un escalier d'accès, etc. C'est par exemple le cas dans la Base Adresse Nationale (BAN), où chaque enregistrement correspond à une adresse géolocalisée. Puisque le terme

6. https://opendata.paris.fr/explore/dataset/adresse_paris/information/

d'adresse est souvent employé indifféremment pour désigner une référence spatiale directe ou indirecte, ce qui crée une ambiguïté qu'il convient de lever pour ce manuscrit :

> Un **énoncé d'adresse** est la référence spatiale indirecte d'un lieu. C'est un énoncé structuré d'un cheminement à l'intérieur d'une hiérarchie spatiale, non ambigu au sein de cette hiérarchie, composé d'une suite ordonnée de repères spatiaux dont la conceptualisation et la désignation sont connues et partagées.
> Une **entité adresse** est une référence spatiale directe localisant un énoncé d'adresse à la surface de la Terre avec des coordonnées géographiques. Selon le modèle de représentation de l'adresse, ces coordonnées peuvent être accessibles immédiatement (comme un point-adresse de la BAN) ou nécessitent un traitement (comme une adresse sur un plan ancien géoréférençable).
> En l'absence d'ambiguïté, on utilisera par simplification le terme **adresse** pour désigner **l'énoncé d'adresse**.

L'adresse comme ancre géohistorique En tant qu'identifiant de lieux, les adresses permettent de croiser de nombreuses informations, comme le montre la figure 1.1. En outre, pour peu que l'on dispose d'un référentiel d'adresses, il devient possible de réaliser des opérations de géocodage. Ce procédé vise à associer des coordonnées à un énoncé d'adresse donné, en recherchant l'énoncé le plus similaire au sein du référentiel (Goldberg *et al.*, 2007). Ainsi spatialisées, les informations associées à cette adresse peuvent faire l'objet d'analyses spatiales à l'aide d'outils et d'approches utilisés dans le domaine de la géomatique - logiciels de Systèmes d'Information Géographique (SIG), bibliothèques de statistiques spatiales, etc. Ainsi, les adresses jouent un rôle de point d'ancrage fondamental pour l'interconnexion et la mise en contexte de données socio-économiques de toutes sortes, ce qui en fait un objet d'étude stratégique pour l'analyse des évolutions urbaines et sociales dans le temps.

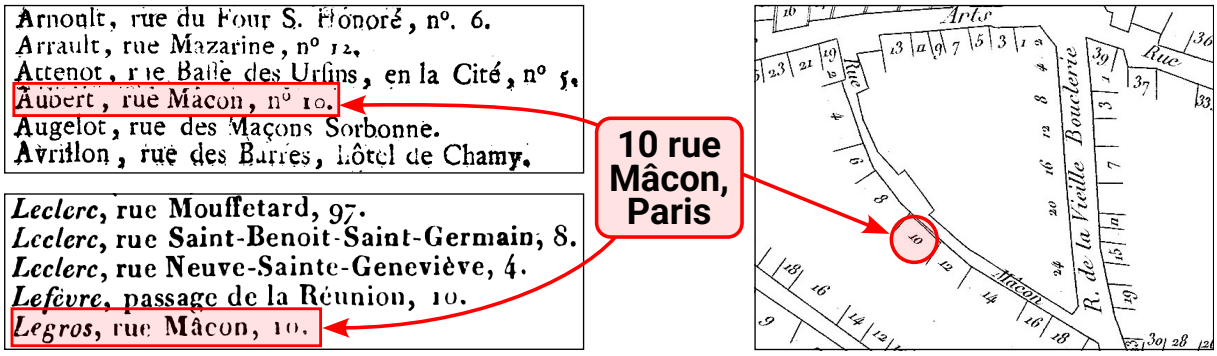
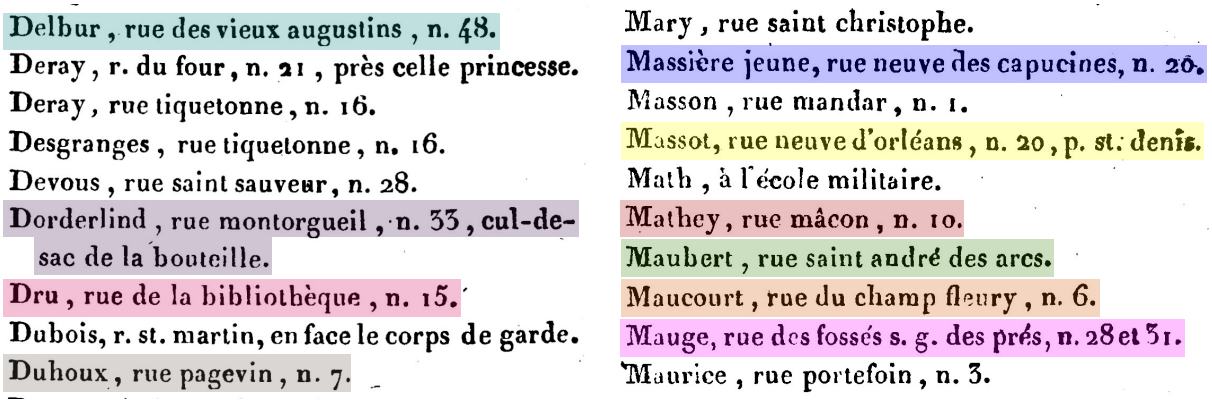


FIGURE 1.1 – Mise en correspondance des informations dans différentes sources décrivant le *10 rue Mâcon, Paris*. La carte est extraite de l'atlas de Jacoubet datant des années 1830 et les extraits d'annuaires sont issus de l'édition 1793 de l'*Almanach national de France* (Testu, 1793) et de l'édition 1844 de l'*Annuaire de la typographie parisienne et départementale* (Prétot, 1844).

Dans les bases de données géographiques, le terme *entité géographique* — parfois appelée *objet géographique* ou *geographic feature*, terme référencé par l'Open Geospatial Consortium (OGC) — désigne une abstraction d'un phénomène réel associé à une localisation par rapport à la Terre (Percivall, 2003). Une entité géographique est donc la

représentation informatique d'une entité localisée à la surface de la Terre, tel qu'une voie, un bâtiment, une parcelle ou une commune, caractérisée par une ou plusieurs géométries et des attributs descriptifs, qui permettent de représenter sa forme, sa localisation et ses propriétés.

Or, les entités du monde réel et leurs relations sont susceptibles d'évoluer dans le temps : une voie peut changer de nom, une parcelle être divisée, une commune fusionner avec une autre. Cette dimension évolutive complique l'usage des adresses comme points d'ancrage, en particulier lorsqu'elles proviennent de sources historiques.



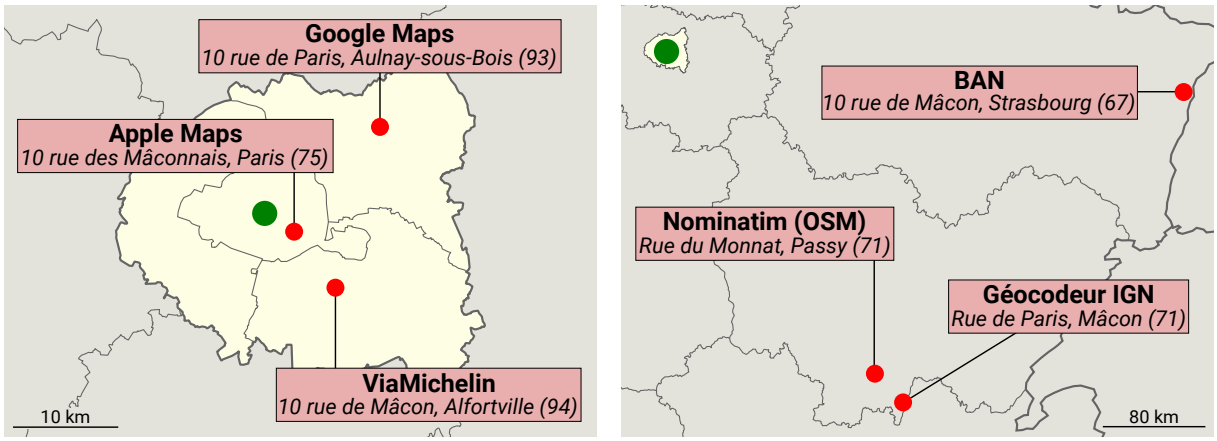
(a) Entrées des corroyeurs (b) Entrées des cordonniers-bottiers

FIGURE 1.2 – Extraits de l'édition 1813 de l'*Almanach des marchands et fabricans en cuirs, de la ville de Paris* (Lemoine, 1813). Les entrées de l'annuaire surlignées concernent des adresses faisant référence à des voies n'existant plus en 2026.

Supposons que l'on souhaite localiser l'ensemble des commerçants de la ville de Paris mentionnés dans un annuaire du XIXe siècle. Pour cela, il est nécessaire de géocoder les adresses associées à ces derniers. La figure 1.2 présente différentes entrées d'annuaires ; chacune est associée à une adresse. Si l'on utilise des outils de géocodage existants, comme ceux de la Base Adresse Nationale (BAN) [7], de l'IGN [8], de Google Maps, d'Apple Maps, de ViaMichelin ou de Nominatim [9] basé sur OpenStreetMap, les coordonnées renvoyées pour une grande partie des adresses sont éloignées de ce qui serait attendu. Ces erreurs s'expliquent par le fait que ces outils sont conçus pour les adresses actuelles et ne prennent pas en compte l'évolution des territoires, ce qui entraîne de nombreux échecs, se traduisant par des réponses incorrectes ou l'absence de réponse (Cura *et al.*, 2018).

Reprenons l'exemple de l'adresse *10 rue Mâcon, Paris*. Aucun des géocodeurs cités précédemment ne renvoie de coordonnées correctes pour cette adresse, comme l'illustre la figure 1.3. Cela résulte de la disparition des entités géographiques citées par l'énoncé de l'adresse : la rue Mâcon a été supprimée dans les années 1850, de même que ses numéros d'immeuble, dont le n° 10. En outre, du fait des évolutions urbaines, une adresse est limitée dans le temps : un même énoncé utilisé à deux temporalités différentes peut référencer des lieux distincts. Ainsi, la mention *10 rue Mâcon, Paris* décrit un bâtiment du centre de Paris au début du XIXe siècle, tandis qu'elle fait référence à un autre bâtiment dans le quartier de Bercy entre le milieu du XIXe siècle et les années 1990, en raison de l'existence d'une nouvelle rue de Mâcon (voir figure 1.4). La renumérotation des immeubles constitue

7. https://adresse.data.gouv.fr/
8. https://geoservices.ign.fr/documentation/services/services-geoplateforme/geocodage
9. https://nominatim.org/



(a) Paris et sa petite couronne (b) Nord-est de la France

FIGURE 1.3 – Résultat d'opérations de géocodage de l'adresse *10 rue Mâcon, Paris* via différents outils. ■ Localisation réelle de l'adresse au milieu du XIXe siècle, ■ Localisation de l'adresse selon un géocodeur.

également une autre source d'erreur, la position associée au numéro 10 pouvant varier au fil du temps le long de la voie. Enfin, les variations orthographiques ou les ambiguïtés toponymiques peuvent également compliquer l'identification des entités géographiques mentionnées dans les énoncés d'adresses. Les adresses sont donc des énoncés qui font référence à un lieu **à un moment donné**, ce qui rend nécessaire la prise en compte de la dimension temporelle pour toute tentative de géolocalisation historique.

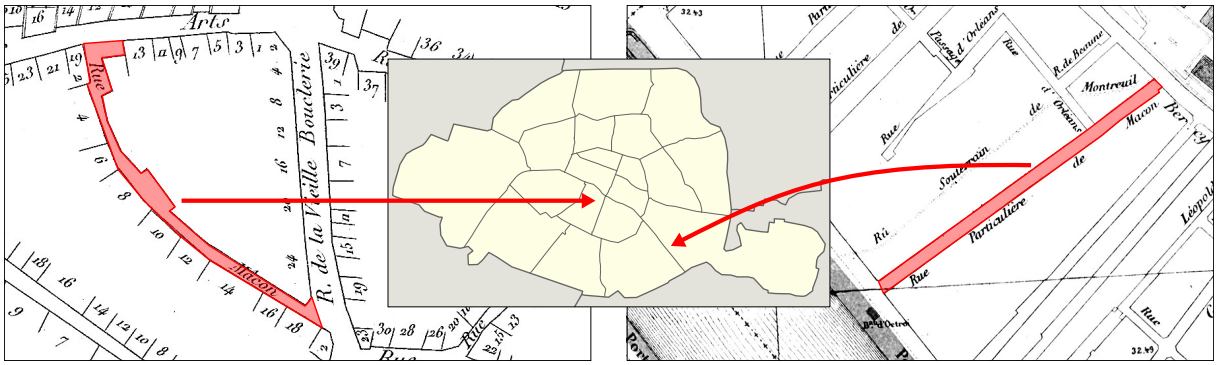


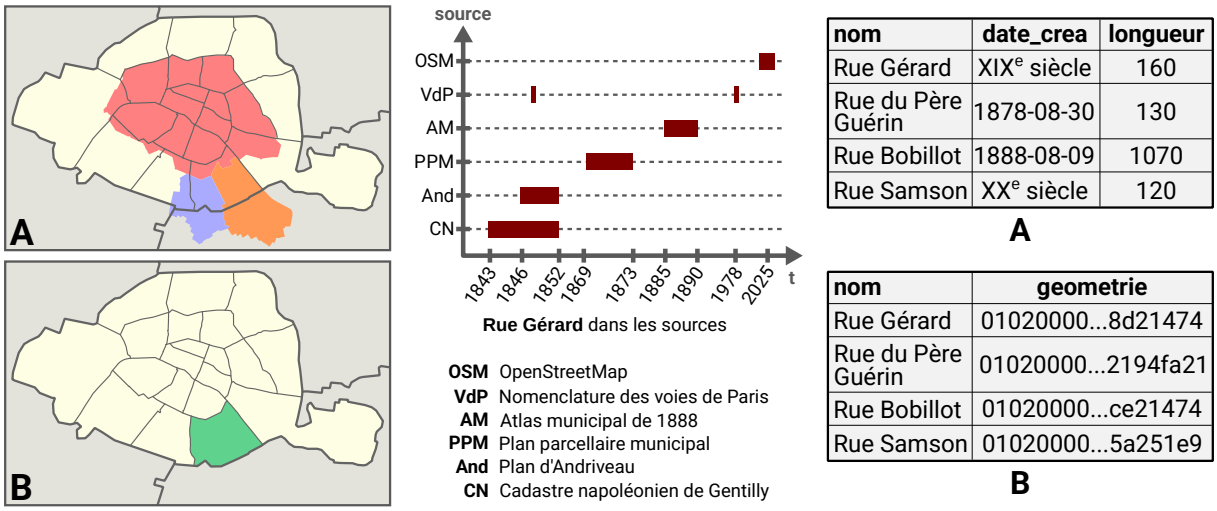
FIGURE 1.4 – Représentation de différentes rues de Mâcon dans Paris à deux périodes distinctes : au début du XIXe siècle dans le plan de Jacoubet à gauche (Jacoubet, 1836) et à la fin de ce même siècle dans l'atlas administratif de 1870 à droite (Service du plan de Paris et Service des eaux et de l'assainissement, 1870).

Créer un référentiel géohistorique d'adresses permet donc de rassembler et consolider ces informations, d'offrir un point d'ancrage spatial et temporel fiable à de nombreux autres jeux de données et corpus documentaires, et d'ouvrir de nouvelles perspectives notamment pour l'étude des dynamiques urbaines.

1.3 Objectifs de la thèse et verrous scientifiques

Cette thèse vise à proposer et mettre en œuvre une méthodologie pour créer un référentiel géohistorique d'adresses à partir des sources historiques disponibles. Costa (2012)

définit un référentiel géohistorique comme un noyau d'informations géographiques dont l'objectif de localiser de manière directe ou indirecte des données sur des thèmes divers et dotées de références spatiales. Un tel référentiel vise à rendre compte de l'état du territoire de façon aussi exacte et continue que possible sans qu'il y ait forcément de source d'informations unique décrivant tout ou partie de ce territoire sur la période concernée. Par exemple, si aucune source ne décrit le territoire en 1928, le référentiel doit permettre de reconstruire son état à cette date grâce aux sources existantes pour des dates proches. Une difficulté majeure de la création de référentiels géohistoriques d'adresses réside dans le fait que les informations disponibles sur les adresses et leur généalogie sont fragmentaires et hétérogènes.



nom	date_crea	longueur
Rue Gérard	XIXe siècle	160
Rue du Père Guérin	1878-08-30	130
Rue Bobillot	1888-08-09	1070
Rue Samson	XXe siècle	120

A

nom	geometrie
Rue Gérard	01020000...8d21474
Rue du Père Guérin	01020000...2194fa21
Rue Bobillot	01020000...ce21474
Rue Samson	01020000...5a251e9

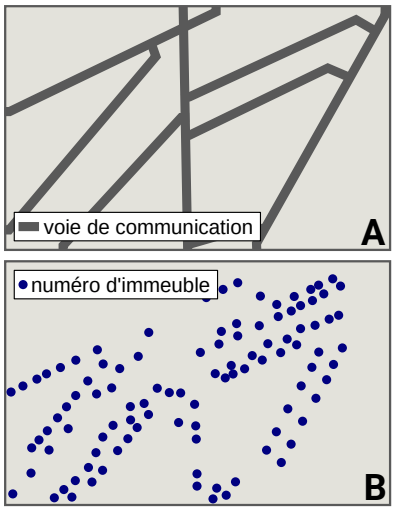
B

(a) Dimension spatiale (b) Dimension temporelle (c) Dimension attributaire

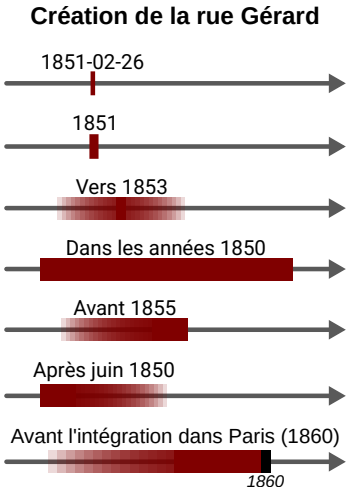
FIGURE 1.5 – Illustration des différentes formes de fragmentation des données géographiques : spatiale, temporelle et attributaire.

Ici, nous qualifions de données fragmentaires des ensembles d'informations incomplets et discontinus, qui ne couvrent qu'une partie de la réalité observée. Ces données présentent des lacunes temporelles, spatiales ou attributaires. La fragmentation spatiale correspond à des situations où l'information disponible dépend des limites administratives ou des sources mobilisées. Par exemple, pour décrire l'actuel 13e arrondissement de Paris avant son rattachement à la capitale en 1860, il est nécessaire de mobiliser des sources provenant de plusieurs territoires distincts (Paris, Gentilly et Ivry-sur-Seine), chacun documentant une portion différente de l'espace (figure 1.5a). La fragmentation temporelle se manifeste lorsque les sources disponibles ne couvrent pas l'ensemble de la période étudiée. Ainsi, même en accumulant plusieurs documents décrivant une même entité, certaines périodes peuvent rester sans information, comme dans le cas de la rue Gérard pour laquelle aucune source ne documente certaines intervalles temporels (figure 1.5b). Enfin, la fragmentation attributaire apparaît lorsque différentes sources décrivent des propriétés complémentaires d'une même entité. Par exemple, une source peut fournir le nom, la date de création et la longueur d'une voie, tandis qu'une autre ne documente que sa géométrie, conduisant à une description partielle dans chaque source prise isolément (figure 1.5c).

Par ailleurs, les données hétérogènes sont définies comme un ensemble d'informations qui diffèrent par leur nature, leur structure ou leur origine. Cette hétérogénéité peut prendre plusieurs formes. Premièrement, elle peut être spatiale lorsque différentes sources décrivent des objets de nature différente sur un même territoire. Par exemple, une source



(a) Dimension spatiale



(b) Dimension temporelle

nom	date_crea	longueur
Rue Gérard	XIXe siècle	160
Rue du Père Guérin	1878-08-30	130
Rue Bobillot	1888-08-09	1070
Rue Samson	XXe siècle	120

A

nom_voie	t_creation	long_km
R GÉRARD	C19	0.160
R. DU PÈRE GUÉRIN	18780830	0.130
R. BOBILLOT	18880809	1.070
R. SAMSON	C20	0.120

B

(c) Dimension attributaire

FIGURE 1.6 – Illustration des différentes formes d'hétérogenéité des données géographiques : spatiale, temporelle et attributaire.

peut documenter les voies de communication tandis qu'une autre décrit les numéros d'immeubles associés aux bâtiments, conduisant à des représentations complémentaires mais structurellement différentes de l'espace (figure 1.6a). Ensuite, l'hétérogénéité peut être temporelle lorsque les informations relatives au temps sont représentées selon des conventions variées. Les sources peuvent mobiliser des granularités différentes (année, mois, jour), utiliser des expressions temporelles imprécises (par exemple *avant*, *vers*) ou encore définir les dates relativement à un événement donné (figure 1.6b). Enfin, l'hétérogénéité peut être attributaire lorsque différentes sources décrivent les mêmes entités et propriétés mais selon des conventions distinctes. Les attributs peuvent porter des noms différents (par exemple `t_creation` ou `date_crea`, `long_km` ou `longueur`), utiliser des unités différentes (mètres ou kilomètres) ou encore adopter des syntaxes variées pour représenter une même valeur (par exemple *R GERARD* ou *rue Gérard*) (figure 1.6c).

Une des particularités des données issues d'archives par rapport aux données géographiques classiques est que la localisation des entités géographiques est rarement fournie sous la forme de coordonnées dans les sources historiques, à l'exception notable des cartes anciennes. On peut, en revanche, y trouver les noms des entités géographiques, leurs relations spatiales ou encore des indications sur leur évolution dans le temps. La représentation et l'intégration des données dans un référentiel géohistorique d'adresses doivent donc être centrées sur l'identité des entités géographiques, comme c'est le cas dans les index géographiques (également appelés *gazetteers*), leur dimension spatiale traditionnellement utilisée en priorité dans les systèmes d'information géographique (SIG), ne pouvant être mobilisée de façon systématique ici. Une difficulté supplémentaire tient à la nature des informations disponibles dans les sources : certaines décrivent l'état des entités réelles sur une période donnée, tandis que d'autres rendent compte des événements qui les font évoluer (nouvelle numérotation, changement de nom de rue, etc.). Le référentiel doit être capable d'intégrer ces deux types d'informations très différentes et d'exploiter leur complémentarité pour reconstituer l'évolution historique des entités géographiques concernées.

Du fait du caractère fragmentaire, hétérogène et parfois imparfait des informations issues des sources historiques, le référentiel géohistorique doit reposer sur un modèle capable de tolérer ces imperfections et de générer de nouvelles connaissances, notamment par le biais d'inférences. La mise en œuvre d'un tel référentiel implique des méthodes

spécifiques d'intégration et de mise en cohérence des données, afin de permettre leur exploitation conjointe. Cela suppose non seulement l'alignement et le liage d'informations issues de sources diverses, mais également l'intégration et l'inférence d'informations relatives à l'évolution des entités géographiques — telles que leur apparition, leur fusion, leur séparation ou, plus simplement, la continuité de leur identité au fil du temps.

La principale contribution scientifique de cette thèse réside dans la méthodologie proposée pour produire un référentiel d'adresses anciennes. Cette méthodologie vise à garantir le suivi temporel des adresses, la cohérence des données, leur traçabilité, leur interopérabilité et leur ouverture, tout en étant suffisamment générique et reproductible pour être appliquée à d'autres contextes urbains, pas seulement au cas d'application traité dans cette thèse. Une fois généré, le référentiel d'adresses peut servir à des usages concrets : il peut servir à indexer spatialement des sources d'archives pour des institutions patrimoniales, ou constituer un référentiel géohistorique ouvert, réutilisable par les chercheurs en humanités numériques.

Pour atteindre cet objectif, plusieurs verrous scientifiques doivent être levés :

— concevoir une ontologie pour représenter les adresses, leurs états successifs et les événements constitutifs de leurs évolutions spatio-temporelles. Il est nécessaire de définir la notion d'adresse au préalable ;
— peupler l'ontologie à partir de données hétérogènes et fragmentaires issues de sources diverses ;
— inférer des connaissances sur les événements à partir des données sur les états et réciproquement.

Les adresses postales standardisées n'apparaissent qu'au XVIIIe siècle, et les documents historiques contiennent de nombreuses formes d'adresses qui diffèrent des standards actuels. La conception d'une ontologie pour représenter des entités géographiques et leurs évolutions spatio-temporelles implique de formaliser les concepts géographiques et leurs relations, ainsi que les changements qu'ils subissent au fil du temps. Il s'agit de gérer la diversité des types d'entités (adresses, voies, parcelles, bâtiments) et de représenter leurs transformations spatiales et temporelles de manière cohérente. De plus, la phase de peuplement de l'ontologie consiste à intégrer des données hétérogènes et fragmentaires. Celles-ci proviennent de documents historiques et de bases de données modernes, ce qui engendre une grande diversité de données : il est nécessaire d'intégrer ces données de manière cohérente. Enfin, les données disponibles peuvent concerner tantôt l'état des entités géographiques à un instant donné, tantôt les événements qui les font évoluer. Il est nécessaire de tirer parti de ces deux types d'informations complémentaires.

Nous prendrons Paris comme terrain d'étude, entre le XVIIIe siècle et aujourd'hui car cette ville a été marquée par de nombreuses évolutions territoriales. Par exemple, la ville s'est constamment étendue au fil du temps notamment en 1860 où son territoire a plus que doublé. De plus, différentes rénovations ont eu lieu à différents instants et à différents endroits. Ces dernières ont mené à la destruction ou à la construction de bâtiments et de voies de communication. Les transformations menées par le baron Haussmann sous le Second Empire constituent un exemple d'une telle rénovation. Avec l'expansion du tissu urbain, il est devenu de plus en plus nécessaire avec le temps d'être davantage précis dans la manière dont on fait référence à des lieux, notamment pour les commerces et les habitations de particuliers. Ainsi, des systèmes d'adressage ont été mis en place mais trois systèmes se sont succédé à Paris, ce qui implique qu'une adresse dans un des systèmes ne fait pas forcément référence au même lieu dans un autre système. Enfin, Paris étant une ville historiquement importante du fait de son statut, elle est très documentée ce qui fait qu'il y a une abondance de données historiques, cartographiques, textuelles et

administratives à son sujet.

Conclusion

En résumé, ce chapitre a permis de présenter le contexte et les objectifs de la thèse, en soulignant les enjeux liés à la reconstitution d'un référentiel géohistorique d'adresses. Nous avons montré que la problématique dépasse la simple collecte de données, puisqu'elle suppose de prendre en compte leur hétérogénéité, leur évolution dans le temps et la question complexe de l'identité des entités géographiques. Ces éléments constituent le socle conceptuel et méthodologique sur lequel s'appuient les chapitres suivants, qui détailleront d'une part l'état de l'art des approches existantes, et d'autre part la méthode proposée pour répondre à ces enjeux.

Chapitre 2

Sources et informations géohistoriques

L'étude des transformations urbaines et de l'évolution des territoires repose sur l'exploitation de sources variées, qui diffèrent par leur nature, leur forme, leur contenu ou leur précision. Leur analyse nécessite une compréhension fine des concepts associés aux informations géohistoriques ainsi que des catégories de sources disponibles.

La section 2.1 présente les notions préliminaires sur les données géohistoriques permettant de les appréhender. La section 2.2 présente une catégorisation des sources exploitables pour la construction du référentiel, contenant des informations géographiques, selon leur niveau de proximité avec le terrain réel. La section 2.3 propose quant à elle une typologie détaillée de ces sources, en distinguant les sources iconographiques, les sources textuelles et les sources numériques structurées. Concernant la section 2.4, elle souligne la nécessité de s'interroger sur la fiabilité des sources et des informations qu'elles fournissent. Dans la continuité de cette analyse, la section 2.5 illustre les caractéristiques des informations géohistoriques contenues dans ces sources, en précisant d'une part les types d'entités géographiques concernées, et d'autre part les imperfections inhérentes à ces données. Enfin, la section 2.6 décrit l'histoire des systèmes d'adressage à Paris, montrant comment les pratiques de localisation et de numérotation ont évolué et influencé la représentation des lieux dans les sources historiques. Cette présentation constitue le cadre conceptuel sur lequel s'appuie la méthodologie développée dans cette thèse pour reconstituer l'évolution des voies et des adresses.

2.1 Notions préliminaires sur les données géohistoriques

Parmi les définitions proposées par le Larousse (2025) pour le terme *source*, on trouve notamment : « origine d'une information » ou encore « documents, textes originaux auxquels on a recours ; origine d'un récit, d'une œuvre ». Ces définitions mettent en évidence le rôle fondamental de la source comme point de départ de la connaissance, qu'elle soit brute, documentaire ou narrative. Il est important de distinguer plusieurs notions complémentaires. Tout d'abord, le terme *document* désigne tout support qui conserve ou transmet une information, qu'il soit matériel (manuscrit, carte, photographie, registre) ou immatériel (enregistrement sonore, fichier numérique). Cette définition s'inspire de Briet (2024) qui souligne que c'est la fonction attribuée à un objet qui le transforme en document. Une *source* n'est pas un objet en soi : tout document peut devenir une source dès lors qu'il

est mobilisé pour répondre à une question de recherche selon la problématique adoptée. Enfin, la *donnée* constitue une abstraction construite à partir de l'information contenue dans les documents ou les sources, permettant de représenter et de traiter de manière structurée des éléments du monde réel tels que des adresses ou des dates de construction.

Ce chapitre porte plus particulièrement sur les sources de l'information géographique historique, c'est-à-dire sur les éléments permettant de reconstituer la localisation, la forme, les propriétés, l'organisation et l'évolution des entités géographiques au cours du temps. Denègre et Salgé (1996) définissent une *information géographique* comme toute donnée localisée dans un même espace et relative aux composantes de la Terre et de l'Homme, à leurs interactions et à leurs évolutions respectives, quels que soient les domaines concernés (physique, social, économique, écologique ou culturel). Dans le cadre de données historiques, il est important d'y ajouter l'aspect temporel, pris en compte depuis longtemps en géomatique bien que peu modélisé dans les SIG (Goodchild et Knowles, 2008). On peut ainsi définir une *information géohistorique* comme une information géographique enrichie d'une dimension temporelle, c'est-à-dire le fait qu'elle concerne des entités spatiales disparues ou possédant une épaisseur temporelle donnée. En parallèle, von Lünen (2013) indique que les informations sont des éléments utiles aux historiens pour répondre à des questions historiques et relater des faits. Souvent dispersées dans des documents isolés, ces informations sont au cœur du travail de l'historien qui doit sélectionner celles qui lui paraissent pertinentes et les croiser afin d'en inférer de nouvelles connaissances. Par exemple, si deux sources cartographiques décrivent Paris à des instants proches, leur comparaison peut révéler des évolutions : apparition ou disparition de voies, changements dans la toponymie, etc. Ce processus de confrontation des sources constitue ainsi un moyen de produire de nouvelles connaissances historiques à partir d'indices épars.

La présente thèse s'inscrit dans une démarche proche, puisqu'elle vise à rassembler les informations géohistoriques issues de différentes sources, à identifier celles qui sont pertinentes pour l'étude de l'évolution des entités géographiques concernées, puis à les intégrer afin de déduire de nouvelles connaissances sur leur évolution. La différence réside toutefois dans l'approche proposée : nous cherchons à automatiser ce travail d'intégration et d'inférence afin de traiter des corpus volumineux et hétérogènes de manière systématique.

2.2 Catégories de sources

Ce sont les sources et les informations qu'elles contiennent qui permettent la construction d'un référentiel géohistorique. Pour en comprendre la nature et en évaluer la fiabilité, il est essentiel de distinguer les catégories de sources mobilisées selon leur niveau de traitement et leur proximité avec les faits (Sahut, 2014; Université du Québec à Montréal, 2025). Une *source primaire* correspond à un document contenant des informations originales et directes, telles que des données brutes issues de laboratoires, des témoignages de première main ou des documents produits à partir d'observations ou de mesures inédites. Giroux et Tremblay (2009) décrivent une telle source comme une information collectée directement par la personne qui la rapporte, sans intermédiaire. Elle peut également désigner un document original, produit au moment même des événements qu'il décrit ou analyse. Dans le domaine de l'histoire urbaine et géographique, on peut citer comme sources primaires les plans de Paris ou bien des annuaires commerciaux d'adresses. Une *source secondaire* repose sur l'analyse ou l'interprétation de sources primaires. Elle regroupe notamment les travaux de synthèse, de recoupement et de reformulation. À titre d'exemple, un atlas historique de Paris, ou une dictionnaire de l'histoire des rues constituent des sources secondaires. Enfin, une *source tertiaire*, parfois confondue avec la source secondaire, com-

bine des éléments issus des sources primaires et secondaires, en les citant, les commentant ou les critiquant, sans nécessairement introduire de nouvelles données (Fjordback Søndergaard *et al.*, 2003). Les inventaires cartographiques, les catalogues d'archives, ou encore les bases de données en ligne dérivées de la numérisation et de la compilation de sources (par exemple la Base Adresse Nationale) relèvent de cette catégorie.

Cependant, cette classification en catégories n'est pas rigide, la frontière entre les différentes classes n'est pas toujours nette. En effet, selon la question de recherche et le contexte d'utilisation, un même document peut être considéré comme source primaire ou secondaire. Par exemple, un plan de Paris constitue une source primaire pour l'étude du tissu urbain à la date de sa réalisation, mais peut être considéré comme secondaire s'il s'appuie sur des levés antérieurs réalisés par d'autres géomètres ou s'il compile des informations provenant de cartes précédentes.

Enfin, le caractère « primaire » d'une source ne garantit pas son objectivité ou son exactitude. En réalité, toute source est produite dans un contexte particulier, avec ses biais, ses choix et ses omissions (Cerutti et Grangaud, 2014). Quant aux sources secondaires et tertiaires, elles peuvent apporter une mise en perspective critique, bien qu'introduisant une distance supplémentaire vis-à-vis des faits.

2.3 Typologie des sources utilisées

Dans le cadre de la reconstitution de l'évolution des entités géographiques, les sources exploitées sont particulièrement hétérogènes, notamment dans leur forme. Barzun et Graff (1985) distinguent deux types de sources géohistoriques : les enregistrements et les reliques (voir figure 2.1). La différence entre les deux types réside dans l'intention de fournir une information spatiale. Tandis que les enregistrements regroupent des sources intégrant intentionnellement ce type d'informations, les reliques sont des sources où l'information existe mais pas de manière intentionnelle. C'est le cas de sources archéologiques ou de divers éléments qu'on peut retrouver sur le terrain : par exemple, le tissu urbain actuel est une relique du passé. Dans notre travail, nous utiliserons uniquement des enregistrements. On peut les diviser en trois sous-types : sources iconographiques, sources textuelles et sources numériques structurées.

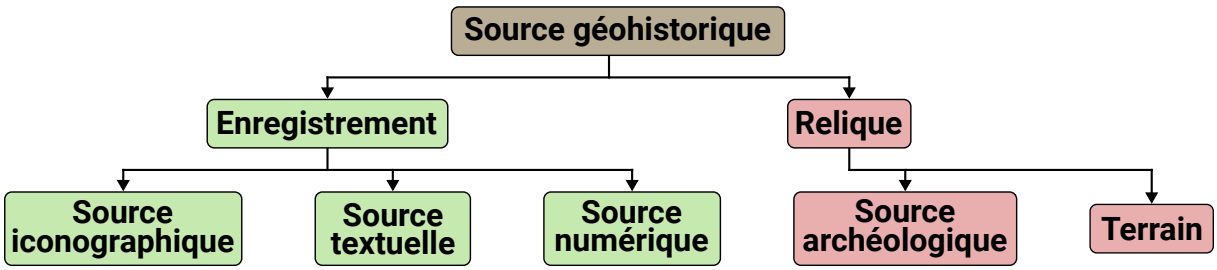


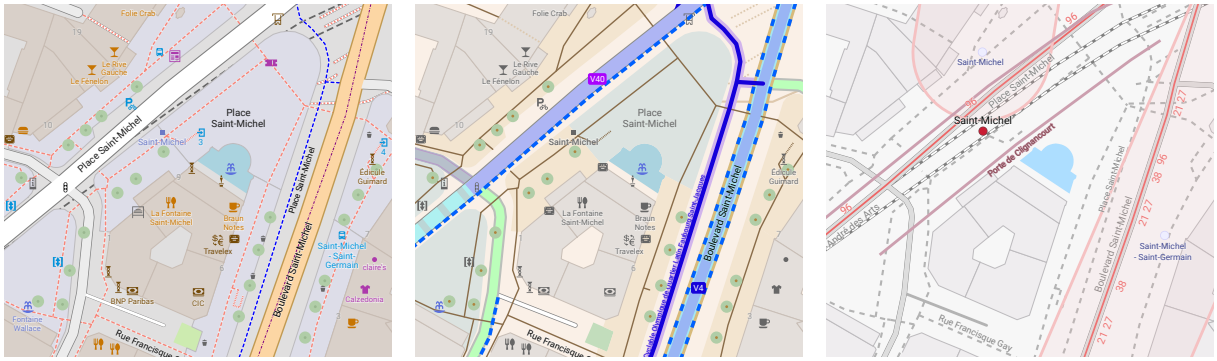
FIGURE 2.1 – Typologie des sources géohistoriques.

2.3.1 Sources iconographiques

Les sources iconographiques regroupent l'ensemble des documents visuels. Elles incluent notamment les photographies (aériennes, de terrain ou d'archives), les images satellites et les cartes géographiques (notamment topographiques et cadastrales). D'autres

sources peuvent être typées comme iconographiques et fournir des informations géohistoriques en représentant un espace ou un événement : affiches, gravures, illustrations, peintures ainsi que certains supports audiovisuels. Ces documents offrent une représentation directe ou interprétée de la réalité, et constituent des sources utiles pour l'analyse spatiale, historique ou sociale d'un territoire.

Parmi les sources iconographiques citées, les cartes géographiques constituent un corpus intéressant pour la représentation des dynamiques spatio-temporelles. En effet, les cartes topographiques et cadastrales sont des représentations visuelles d'un territoire à un instant donné. En plus de comprendre la dimension spatiale en représentant explicitement les emprises ou les positions des entités géographiques, elles peuvent inclure la dimension temporelle lorsqu'elles sont datées. La dimension thématique est également présente : les éléments graphiques et textuels permettent de transmettre des informations supplémentaires au-delà de la simple localisation des entités géographiques. Ainsi, la couleur, la texture, l'épaisseur ou le style des lignes, la taille et la forme des symboles, ainsi que le texte associé à une entité, fournissent des indications sur ses caractéristiques. Le texte peut associer un nom ou un identifiant à une entité géographique, tandis que la couleur ou le symbole peuvent indiquer la fonction ou la catégorie de cette même entité, l'épaisseur d'une ligne peut signaler l'importance d'une voie de communication, et la texture ou le hachurage peuvent renseigner sur l'occupation ou l'utilisation du sol (bâtiment, jardin, champ, etc.). D'autres éléments graphiques, comme les flèches, les cercles, les motifs ou les icônes, peuvent indiquer des relations, des hiérarchies ou des caractéristiques particulières des entités représentées.

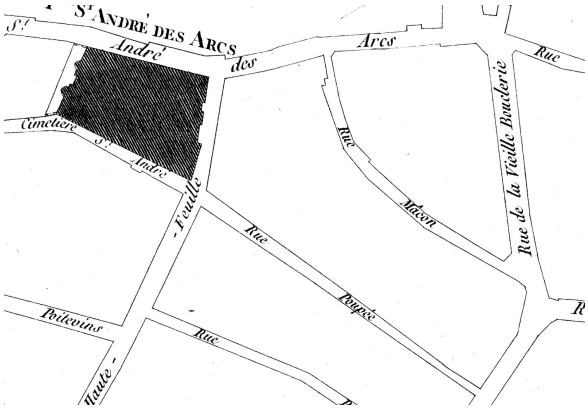


(a) Fond de plan standard (b) Fond de plan CyclOSM (c) Fond de plan *Transport*

FIGURE 2.2 – Trois rendus cartographiques d'un même territoire (autour de la fontaine Saint-Michel à Paris) à partir des mêmes données OpenStreetMap, mais avec des styles graphiques différents. Crédit : © OpenStreetMap contributors, sous licence ODbL.

Toutefois, les cartes présentent une grande diversité dans leur représentation du terrain, en raison notamment des contraintes inhérentes au support papier, qui limite la quantité d'informations pouvant être restituées. Ces limites imposent des choix dans la symbologie adoptée, en fonction à la fois des objectifs poursuivis par la carte et de la vision du territoire proposée par son auteur (Buisson, 2009). La figure 2.2 illustre les conséquences du choix graphique puisqu'à partir d'un même ensemble de données, la représentation qu'on se fait d'un territoire peut différer énormément. La diversité graphique inclut plusieurs dimensions : les éléments graphiques et textuels, le niveau de détail ou de généralisation cartographique, et la sélection de l'information à représenter, c'est-à-dire ce qui est inclus ou omis sur la carte. Zanin et Trémélo (2003) indiquent que la réalisation d'une carte implique de choisir le niveau cartographique, la projection, l'échelle et le degré

de généralisation, autant de facteurs qui influencent la manière dont les informations sont représentées. Cette diversité de représentation peut compliquer les processus d'extraction automatique de données, car il n'existe pas d'unité graphique ou de sémiologie commune : les mêmes types d'entités peuvent être représentés différemment selon la carte, ce qui nécessite une adaptation des méthodes de reconnaissance des informations extraites avant de passer à leur interprétation et leur harmonisation.



(a) Planche n°45 de l'atlas de Verniquet (Verniquet, 1795)



(b) Atlas Vasserot (Vasserot, 1825), reconstitué par le projet ANR ALPAGE (Noizet, 2009)

FIGURE 2.3 – Extraits de deux plans de la ville de Paris illustrant le quartier autour de la rue Mâcon.

La figure 2.3 illustre le fait que la vision de l'auteur de la carte, ainsi que les objectifs poursuivis lors de sa production, influencent aussi la lecture du territoire. D'un côté, la figure 2.3a est un extrait de l'*Atlas National de la Ville de Paris* d'Edme Verniquet dit *Atlas de Verniquet* (Verniquet, 1795) qui décrit le territoire parisien à l'époque de la Révolution française. Les contours des voies de communication ainsi que ceux des îlots urbains y sont tracés de manière très détaillée. Mis à part les biens nationaux qui sont aussi très détaillés, les bâtiments n'y sont pas représentés. De l'autre, la figure 2.3b est un extrait de ce qu'on appelle l'*atlas de Vasserot*, un ensemble des plans d'îlots dressés entre 1810 et 1836 par l'architecte Philibert Vasserot. Faisant office de cadastre du parcellaire parisien, il n'a donc pas le même objectif que celui de Verniquet puisqu'il tend à représenter les bâtiments et les parcelles de manière détaillée. Cependant, bien que ces deux atlas ne soient pas contemporains, ils offrent des informations complémentaires du même territoire.

Enfin, la qualité des cartes peut être variable selon les sources, les techniques de production ou les objectifs du document. Les précisions géométriques, la complétude ou la fréquence de mise à jour diffèrent d'une carte à l'autre, ce qui rend indispensable une évaluation critique de leur contenu avant utilisation. Dans l'éventail des cartes disponibles, ce sont principalement les plans de ville et les plans cadastraux qui présentent un intérêt particulier pour l'étude des dynamiques urbaines. Ces plans peuvent représenter des adresses, mais celles-ci sont parfois incomplètes, absentes ou interpolées, ce qui limite leur fiabilité pour une analyse fine. En effet, l'interpolation est une opération consistant à déduire la localisation de certaines adresses lorsque la numérotation des voies est continue, sans qu'elles soient explicitement relevées sur le terrain, ce qui introduit une part d'incertitude dans leur interprétation.

Par ailleurs, il existe aujourd'hui des cartes dynamiques, notamment interactives, proposées par des services tels que Google Maps, Géoportail ou OpenStreetMap. Ces cartes

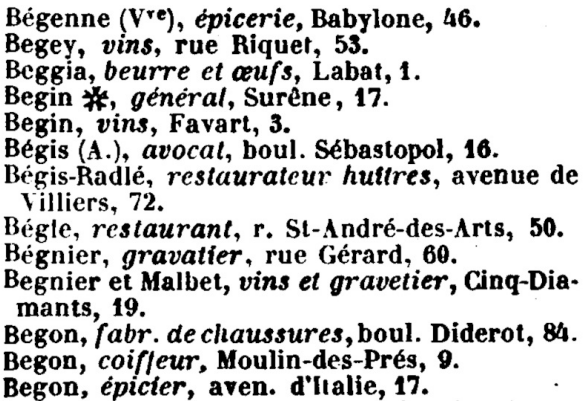
permettent d'ajouter des informations complémentaires qu'une carte classique ne fournit pas. Elles offrent des fonctionnalités multi-échelles, permettant de zoomer ou dézoomer, et la possibilité de choisir l'emprise en naviguant dans la carte. De plus, l'interaction avec les entités géographiques qui y figurent permet d'afficher des données détaillées. Cependant, dans le cadre de cette étude, il n'est pas pertinent de traiter ces cartes dynamiques. Elles constituent avant tout une représentation visuelle de données numériques et structurées (voir sous-section 2.3.3) : l'objectif n'est pas d'extraire le rendu cartographique, mais d'exploiter directement les données géographiques sous-jacentes.

L'extraction des données à partir de ces documents s'effectue principalement par vectorisation. La méthode la plus simple est la vectorisation manuelle, et elle est privilégiée lorsque le corpus de documents est de taille réduite (Ostafin *et al.*, 2020). Toutefois, elle est coûteuse en temps et variable en qualité, notamment selon la personne chargée de la vectorisation. Afin de dépasser ces limites, des travaux ont proposé des approches automatiques ou semi-automatiques qui doivent s'adapter à la grande hétérogénéité des documents en termes de forme et de qualité (Chen *et al.*, 2024). Les informations ainsi extraites sont localisées dans le système de coordonnées propre au document dont elles proviennent. Elles doivent être associées à des coordonnées géographiques par un processus de géoréférencement. Cette étape repose le plus souvent sur l'identification de points d'ancrage, c'est-à-dire de points caractéristiques présents sur la carte et dont la position géographique est connue afin de procéder à une transformation (Dupouey *et al.*, 2007). Toutefois, cette opération se complique lorsque le document représente un territoire ayant fortement évolué, rendant l'identification de tels points difficile ou incertaine. Par ailleurs, certains paramètres font que la transformation entre l'image et le système de coordonnées géographiques ne correspond pas à une simple fonction géométrique, ce qui impose l'utilisation de modèles de transformation plus complexes. Par exemple, il existe des déformations inhérentes au support ou à cause de la numérisation. De plus, il existe des problèmes liés à la construction initiale de la carte. Des tracés peuvent être imprécis ou non conformes à une projection cartographique stricte et il peut y avoir une non-uniformité des plans, c'est-à-dire que l'échelle varie selon les endroits de la carte.

Malgré les imperfections inhérentes au support cartographique et aux choix de représentation qui en découlent, les cartes anciennes constituent une ressource essentielle. Elles offrent en effet des représentations complémentaires des états des entités géographiques à des dates données, assorties de potentielles relations entre ces entités, dont les positions sont explicitement exprimées. Nous désignons chacun de ces ensembles de données géohistoriques dotés d'un même temps valide par le terme de *snapshot*, c'est-à-dire un ensemble d'observations rassemblées dans une source donnée durant la même période (Dumenieu, 2015).

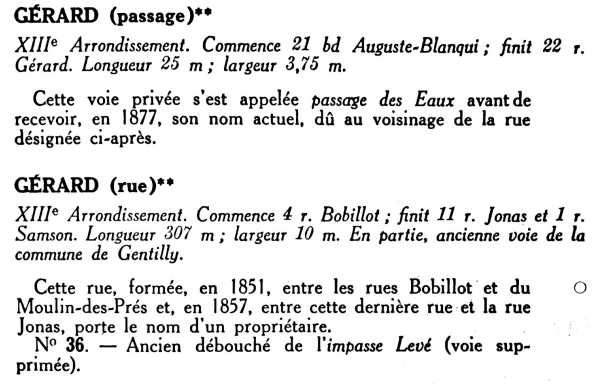
2.3.2 Sources textuelles

Les documents textuels sont des sources écrites. Ils peuvent contenir des références spatiales mais contrairement aux documents iconographiques, elles sont généralement indirectes. La position ou la géométrie d'un objet n'y sont pas représentées directement, mais les lieux y sont principalement mentionnés par le biais de leur nom. Leur localisation est généralement exprimée de manière relative à d'autres lieux également référencés dans le texte, par l'indication de distances (précises ou approximatives), d'itinéraires ou de relations hiérarchiques entre entités. Pour extraire et spatialiser ces références, de nombreux travaux ont proposé des approches spécifiques (Tahrat *et al.*, 2013; Moncla, 2015; Poirier, 2016; Fize *et al.*, 2017). En effet, la nature indirecte des références spatiales dans les textes



Bégenne (V^re^), *épicerie*, Babylone, 46.
Begey, *vins*, rue Riquet, 53.
Beggia, *beurre et œufs*, Labat, 1.
Begin ✱, *général*, Surêne, 17.
Begin, *vins*, Favart, 3.
Bégis (A.), *avocat*, boul. Sébastopol, 16.
Bégis-Radlé, *restaurateur huîtres*, avenue de Villiers, 72.
Bégle, *restaurant*, r. St-André-des-Arts, 50.
Bégnier, *gravatier*, rue Gérard, 60.
Begnier et Malbet, *vins et gravetier*, Cinq-Diamants, 19.
Begon, *fabr. de chaussures*, boul. Diderot, 84.
Begon, *coiffeur*, Moulin-des-Prés, 9.
Begon, *épicier*, aven. d'Italie, 17.

(a) Extrait de l'annuaire commercial Firmin-Didot de 1894 (Firmin-Didot et Firmin-Didot, 1894)



GÉRARD (passage)**

XIIIe Arrondissement. Commence 21 bd Auguste-Blanqui ; finit 22 r. Gérard. Longueur 25 m ; largeur 3,75 m.

Cette voie privée s'est appelée *passage des Eaux* avant de recevoir, en 1877, son nom actuel, dû au voisinage de la rue désignée ci-après.

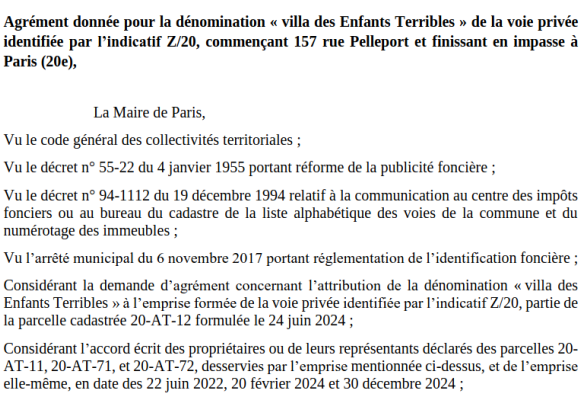
GÉRARD (rue)**

XIIIe Arrondissement. Commence 4 r. Bobillot ; finit 11 r. Jonas et 1 r. Samson. Longueur 307 m ; largeur 10 m. En partie, ancienne voie de la commune de Gentilly.

Cette rue, formée, en 1851, entre les rues Bobillot et du Moulin-des-Prés et, en 1857, entre cette dernière rue et la rue Jonas, porte le nom d'un propriétaire.

N° **36.** — Ancien débouché de l'*impasse Levé* (voie supprimée).

(b) Extrait du dictionnaire d'Hillairet de 1960 décrivant l'historique du passage et de la rue Gérard (Hillairet, 1963).



Agrément donnée pour la dénomination « villa des Enfants Terribles » de la voie privée identifiée par l'indicatif Z/20, commençant 157 rue Pelleport et finissant en impasse à Paris (20e),

La Maire de Paris,

Vu le code général des collectivités territoriales ;

Vu le décret n° 55-22 du 4 janvier 1955 portant réforme de la publicité foncière ;

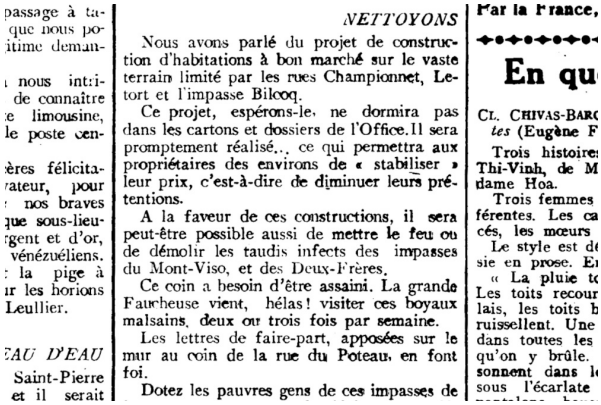
Vu le décret n° 94-1112 du 19 décembre 1994 relatif à la communication au centre des impôts fonciers ou au bureau du cadastre de la liste alphabétique des voies de la commune et du numérotage des immeubles ;

Vu l'arrêté municipal du 6 novembre 2017 portant réglementation de l'identification foncière ;

Considérant la demande d'agrément concernant l'attribution de la dénomination « villa des Enfants Terribles » à l'emprise formée de la voie privée identifiée par l'indicatif Z/20, partie de la parcelle cadastrée 20-AT-12 formulée le 24 juin 2024 ;

Considérant l'accord écrit des propriétaires ou de leurs représentants déclarés des parcelles 20-AT-11, 20-AT-71, et 20-AT-72, desservies par l'emprise mentionnée ci-dessus, et de l'emprise elle-même, en date des 22 juin 2022, 20 février 2024 et 30 décembre 2024 ;

(c) Extrait d'un agrément de dénomination de la ville de Paris concernant de la villa des Enfants Terribles au 26 août 2025 (Ville de Paris, 2025).



passage à ta-
que nous po-
itime deman-

nous int:i-
de connaître
e limousine,
le poste cen-

ères félicita-
rateur, pour
nos braves
que sous-lieu-
gent et d'or,
vénézuéliens.
la pige à
ir les horions
Leullier.

EAU D'EAU
Saint-Pierre
et il serait

NETTOYONS

Nous avons parlé du projet de construction d'habitations à bon marché sur le vaste terrain limité par les rues Championnet, Letort et l'impasse Bilcoq.

Ce projet, espérons-le, ne dormira pas dans les cartons et dossiers de l'Office. Il sera promptement réalisé... ce qui permettra aux propriétaires des environs de « stabiliser » leur prix, c'est-à-dire de diminuer leurs prétentions.

A la faveur de ces constructions, il sera peut-être possible aussi de mettre le feu ou de démolir les taudis infects des impasses du Mont-Viso, et des Deux-Frères.

Ce coin a besoin d'être assaini. La grande Faucheuse vient, hélas ! visiter ces boyaux malsains, deux ou trois fois par semaine.

Les lettres de faire-part, apposées sur le mur au coin de la rue du Poteau, en font foi.

Dotez les pauvres gens de ces impasses de

Par la France,

En qu

CL. CHIVAS-BARC
tes (Eugène F

Trois histoire
Thi-Vinh, de M
dame Hoa.

Trois femmes
férentes. Les ca
cés, les mœurs

Le style est d
sie en prose. E

« La pluie t
Les toits recour
lais, les toits h
ruissellent. Une
dans toutes les
qu'on y brûle.
sonnent dans l
sous l'écarlate

(d) Extrait du 22 avril 1922 de l'hebdomadaire *L'Aurore : contre tous les abus et toutes les iniquités*, évoquant un projet de construction à Paris (Longeard, 1922).

FIGURE 2.4 – Exemple de quatre documents textuels.

rend nécessaire le recours à des méthodes d'extraction et de géocodage adaptées pour les convertir en données géographiques exploitables.

Les documents textuels peuvent décrire des états spatio-temporels d'entités géographiques — ils attestent alors de leur existence à un instant donné et en établissant des relations avec d'autres objets — ou bien relater des événements, c'est-à-dire des évolutions affectant un lieu à un moment donné (changement de nom, de statut, de configuration, etc.). Ils présentent une grande hétérogénéité structurelle. Parmi ces documents, les annuaires constituent des répertoires qui listent des entités nommées — particuliers, organisations, commerçants — associées à une adresse. Ils fournissent des attestations d'état à une date donnée. La figure 2.4a en illustre un extrait. Chaque entrée décrit un commerçant associé à une adresse qui définit son lieu de travail à l'instant de la publication de l'ouvrage.

Les dictionnaires topographiques recensent des lieux (quartiers, rues, villes) assortis de caractéristiques telles que leur localisation, leur longueur ou leur fonction, et comportent souvent une description historique. Ils combinent ainsi données d'état et événements. Les dates mentionnées dans ces descriptions peuvent être précises (par exemple : *9 janvier 1896*), floues (*vers 1652*), ou exprimées de manière relative (*après le percement du boulevard de Stalingrad*). Un exemple de ce type de document est montré par la figure 2.4b.

Par ailleurs, les textes juridiques et administratifs, comme les bulletins de lois ou les délibérations municipales (voir figure 2.4c), ont une valeur officielle et documentent explicitement des transformations territoriales, telles que des créations ou modifications de voies, de limites administratives ou de statuts. D'autres textes, dont la destination première n'est pas la description d'entités géographiques, peuvent également constituer des sources précieuses. Par exemple, les articles de presse (voir figure 2.4d) peuvent mentionner des projets ou événements en se référant à des lieux, témoignant ainsi de leur existence et pouvant fournir des informations supplémentaires à leur propos.

2.3.3 Sources numériques et structurées

Les données structurées, issues de bases administratives ou de jeux de données collaboratifs, présentent une certaine rigueur formelle, notamment grâce à l'utilisation de standards de représentation, de nomenclature et de structuration des données. Cette normalisation facilite leur traitement automatique, leur interopérabilité, ainsi que leur réutilisation ou leur croisement avec d'autres sources. Ces bases, telles que la Base Adresse Nationale (BAN), les cadastres numériques, Wikidata ou OpenStreetMap, offrent des données explicites sur des entités géographiques, avec des identifiants stables, des coordonnées précises et des attributs normalisés, c'est-à-dire des caractéristiques dont le nom et la définition ont été fixés par une convention ou une communauté. Par exemple, dans OpenStreetMap, l'attribut `ref:INSEE` définit le code INSEE d'une unité administrative alors que, théoriquement, d'autres variantes comme `INSEE_code` ou `ref:FR:INSEE` pourraient être utilisées. De même, dans Wikidata, définir une propriété correspond à un choix communautaire, assurant que tous les contributeurs utilisent le même attribut pour la même information.

Néanmoins, la diversité des producteurs, des finalités et des échelles d'observation conduit à une hétérogénéité des modèles, des schémas et des formats de diffusion, malgré l'adoption de standards. Les données peuvent être organisées sous forme de tableaux, de bases relationnelles, de points d'accès SPARQL, ou encore de fichiers géographiques (GeoJSON, shapefiles). La figure 2.5 montre cette diversité de modèles et de formats pour décrire la même information. Cette pluralité nécessite souvent des opérations de transformation, c'est-à-dire toute opération impliquant toute modification des données pour les rendre utilisables ou compatibles avec un autre système, avant toute intégration ou comparaison inter-sources (voir sous-section 5.1.2). La figure 2.5 illustre cette diversité de modèles et de formats permettant de décrire une même information. Par ailleurs, les métadonnées — essentielles pour documenter la provenance, la validité temporelle ou la qualité des données — sont parfois absentes ou incomplètes, ce qui complique leur interprétation et leur réutilisation.

Au-delà de ces considérations techniques, l'intégration de sources structurées soulève également des questions sémantiques. Les schémas ou les vocabulaires de représentation sous-jacents peuvent varier fortement d'une base à l'autre, tant dans les définitions des entités (voie, bâtiment, etc.) que dans les types de relations ou les catégorisations employées. Ces divergences compliquent l'alignement conceptuel des jeux de données et nécessitent des méthodes de médiation ontologique pour garantir la cohérence des données mises en correspondance. Sans un tel effort d'harmonisation sémantique, le croisement de sources numériques et structurées peut produire des incohérences, des redondances ou des pertes d'information.

```
id,numero,voie,commune,lat,lng
1,10,Rue Mâcon,Paris,48.853080,2.343612
```

```
{
 "type": "Feature",
 "id": "ban:67890",
 "geometry": { "type": "Point",
  ↪  "coordinates":
  ↪  [2.343612,48.853080] },
 "properties": { "numero": "10",
  ↪  "voie": "Rue Mâcon",
  ↪  "commune": "Paris" }
}
```

```
:adresse67890 a :Adresse ;
  :numero "10" ;
  :voie "Rue Mâcon" ;
  :commune wd:Q90 ;
  geo:asWKT "POINT(2.343612
   ↪  48.853080)"^^geo:wktLiteral .
```

FIGURE 2.5 – Exemple d'une même adresse (10 rue Mâcon, Paris) représentée dans différents modèles et formats : CSV, GeoJSON et RDF (Turtle).

2.4 Évaluation de la fiabilité des sources

La fiabilité d'une source repose sur un ensemble de critères qu'il est essentiel d'identifier et d'analyser avec rigueur. Akoka *et al.* (2020) indiquent que les critères à évaluer pour la crédibilité d'une source dépendent de sa catégorie (voir section 2.2) même si des indicateurs communs existent. Ils affirment que le critère principal concerne l'auteur du document : sa légitimité dépend à la fois de son expertise dans le domaine traité, de sa réputation, mais aussi des idées qu'il défend et des intérêts qu'il pourrait poursuivre. Tout comme l'auteur, l'éditeur constitue également un indicateur pertinent, en particulier à travers sa ligne éditoriale, son positionnement institutionnel ou commercial, et sa reconnaissance dans le champ concerné. On peut aussi ajouter le profil du lectorat visé par le document, son mode ainsi que son volume de diffusion. Enfin, pour des sources primaires, étant donné qu'elles contiennent des informations directes sur un sujet, la date de production joue un rôle important.

Le contexte de production de la source influence également sa fiabilité. Par exemple, le rythme de publication joue un rôle important : un quotidien, soumis à l'urgence de l'actualité, offre généralement un niveau de vérification moindre qu'une revue mensuelle ou annuelle, qui bénéficie de délais plus longs pour l'analyse et la vérification des faits. La nature même du document — qu'il soit académique, journalistique, technique, scientifique ou administratif — détermine aussi le degré d'exigence méthodologique attendu.

La fiabilité d'une source est un élément crucial lors de l'intégration de données provenant de sources multiples. En effet, lorsqu'il s'agit de concilier des observations similaires issues de documents différents, des conflits peuvent apparaître et il est alors nécessaire de déterminer quelles informations sont les plus crédibles. Plusieurs travaux se sont intéressés à l'estimation de la fiabilité des sources ou des informations qu'elles contiennent afin de corriger ces problèmes (Li *et al.*, 2014; Tian *et al.*, 2020).

2.5 Caractéristiques des informations géohistoriques

Les informations géographiques issues de sources historiques se distinguent de celles qui sont produites et utilisées aujourd'hui. Tandis que les données géographiques contem-

poraines sont normalisées et associées à une géométrie précise (par exemple dans les systèmes d'information géographique), les documents anciens ne répondent pas à ces standards. La représentation des entités y est souvent hétérogène, variant selon les périodes, les contextes de production et les objectifs des auteurs. Les principales informations géographiques portent sur des entités géographiques dont la définition est donnée en section 1.2. Leur support et leur mode de production les rendent plus susceptibles de présenter des imperfections.

2.5.1 Les entités géographiques

Selon Smith (1995), il existe deux catégories d'entités géographiques selon comment leurs frontières sont perçues. D'une part, les entités *bona fide* correspondent à des objets délimités par des discontinuités physiques visibles de l'espace géographique, indépendantes de toute convention humaine. Elles incluent par exemple les montagnes, rivières, lacs, côtes ou forêts, qui existent indépendamment de toute décision humaine, même si leur limite peut parfois être imprécise. D'autre part, les entités *fiat* sont définies par des décisions humaines, légales ou administratives, comme les limites d'une commune. Néanmoins, la frontière entre les deux types reste floue, certaines délimitations d'entités *fiat* pouvant s'appuyer totalement ou partiellement sur des éléments existants, tels que des cours d'eau ou des voies de communication (Smith et Varzi, 2000). L'incertitude provient ici des variations possibles selon les documents ou contextes qui en fixent la définition.

Les adresses, dont nous souhaitons suivre l'évolution, sont principalement de type *fiat*. Leurs traces sont mentionnées dans des enregistrements, car leur délimitation correcte nécessite des informations intentionnelles. À l'inverse, les entités *bona fide* peuvent souvent être décrites à partir de reliques, ce qui explique pourquoi notre travail se concentre uniquement sur les enregistrements.

On peut introduire une troisième catégorie, utilisée par Plewe (2002), celle des entités *motivées*. Elles désignent des concepts construits à partir de phénomènes complexes, par des processus d'agrégation ou de catégorisation. C'est le cas des montagnes, des forêts ou encore des régions : elles n'ont pas de limites officielles et leur définition dépend fortement du contexte et des choix de modélisation. La conceptualisation même de ces entités peut ainsi être source d'*imperfections* dans l'information géographique. Ainsi, la nature même des entités géographiques conditionne la qualité et les limites des informations qui les décrivent. C'est pourquoi il est nécessaire d'examiner à présent les caractéristiques propres aux informations géohistoriques, afin de comprendre comment ces imperfections se manifestent et se propagent dans les sources.

2.5.2 Imperfection des données géohistoriques

Les informations historiques présentent fréquemment un certain flou, tant dans leur expression que dans leur interprétation. Dans le cas des énoncés d'adresses, présentes dans des sources textuelles, c'est la modalité de représentation de l'information — le langage naturel — qui introduit des imprécisions et des ambiguïtés. Il n'est pas d'usage de donner des coordonnées exactes pour référencer un lieu, et les formulations alternatives (« face à l'hôtel de ville », « au coin de la rue ») restent relatives et peu univoques. Dans le cas des sources cartographiques, l'imprécision résulte plutôt de la qualité du levé, de la gravure, de la conservation du papier ou du géoréférencement, ainsi que des choix de représentation et des déformations inhérentes aux documents. Prendre en compte cette diversité et cette incertitude est ainsi essentiel pour exploiter les informations géographiques historiques.

Dimension	Sous-catégorie	Définition
Incertitude	-	Doute sur la validité ou l'existence d'une connaissance
Imprécision	Vague	Information définie subjectivement ou de manière flexible, sans limite précise
	Approximation	Information aux limites mal définies
Ambiguïté	Conflit	Plusieurs classifications ou interprétations contradictoires pour un même objet
	Non-spécificité	Définition ou relation pouvant être interprétée de plusieurs façons selon le contexte ou l'échelle
Incomplétude	Absence	Valeurs manquantes pour décrire une entité
	Lacune	Données partielles ou couverture incomplète d'une structure ou d'un ensemble d'entités

TABLEAU 2.1 – Typologie de l'imperfection des informations géographiques historiques proposée par de Runz (2008).

Comme le souligne de Runz (2008), les informations issues de sources anciennes doivent être considérées comme *imparfaites*. Cette imperfection constitue un cadre général qui regroupe différentes dimensions affectant la qualité de l'information, parmi lesquelles l'incertitude, l'imprécision, l'ambiguïté ou encore l'incomplétude (voir tableau 2.1).

L'incertitude est particulièrement liée à la validité d'une connaissance : un objet existe-t-il réellement à cet endroit ? à cette période ? et avec quel degré de probabilité peut-on affirmer son existence ? Par exemple, dans la figure 2.6, les numéros 33 et 31 sont très proches, à tel point qu'il y a une incertitude sur l'existence de deux entités distinctes : existe-il une seule entité portant le numéro d'immeuble 33-31 ou bien y a-t-il deux entités respectivement numérotées 31 et 33 ? Klir et Yuan (1995) introduisent une classification des incertitudes reprises par Fisher (1999) et Fisher *et al.* (2005) pour l'adapter aux données géographiques. Plewe (2002) souligne que l'incertitude en information géohistorique ne se limite pas à l'espace. Elle concerne également les dimensions temporelles et thématiques des phénomènes. Il propose ainsi l'*Uncertain Temporal Entity Model* (UTEM), qui invite à considérer que tout modèle d'incertitude géohistorique doit intégrer conjointement l'espace, le temps et les thèmes. Cette incertitude peut être décrite selon plusieurs aspects. Sa *dimensionnalité* renvoie aux caractéristiques de l'entité concernées (spatiale, temporelle, thématique). Sa *cause* réside dans les conditions de production de l'information (méthodes de levé, transcription, compilation). Son *type* précise la nature du problème posé par l'incertitude. Enfin, sa *forme* décrit la manière dont l'incertitude se manifeste, notamment sous la forme d'une région de localisation possible, d'un intervalle temporel ou d'une hypothèse non vérifiable.

Avec l'incertitude, une autre dimension de l'imperfection est l'*imprécision*, renvoyant au caractère vague ou approximatif de la sémantique utilisée pour décrire une information. Elle ne concerne pas la validité d'une connaissance, mais les limites de sa définition. Bunel (2021) souligne qu'il est essentiel de distinguer la précision ou l'imprécision d'une donnée de la certitude ou de l'incertitude qui l'accompagnent. Par exemple, les rives d'un lac ou le niveau d'eau observable sur une photographie aérienne varient selon les saisons ; une montagne ne dispose pas d'une délimitation nette, et son extension peut être comprise tantôt comme son sommet, tantôt comme l'ensemble de son massif. Tandis que l'incertitude qualifie les doutes sur l'existence ou la validité d'une connaissance (« le lac avait-il bien cette limite à cette période ? »), l'imprécision concerne la manière dont cette connais-

sance est exprimée et délimitée. L'imprécision peut se manifester sous différentes formes selon de Runz (2008). Elle peut relever du caractère *vague*, lorsque l'information est définie à partir de connaissances subjectives ou flexibles (« Vincennes est proche de Paris »), sans seuil clair ni frontière explicite. Elle peut également être liée à l'*approximation*, lorsque les limites d'un objet sont mal définies ou floues comme le montre la figure 2.6 puisqu'au niveau des intersections des rues, aucune limite n'est clairement définie. De même, les numéros d'immeuble n'ont pas d'emprise précise. Troisième dimension de l'imperfection, l'*ambiguïté* se manifeste lorsqu'il existe un doute sur la manière de définir une entité ou un phénomène. Elle survient quand un même élément peut être interprété de plusieurs façons et peut prendre la forme d'un *conflit*, lorsque plusieurs classifications contradictoires s'appliquent à un même objet. de Runz (2008) prend l'exemple d'une zone mixte située à la frontière entre un quartier pavillonnaire et un secteur d'immeubles et considère qu'il peut être difficile à la catégoriser de façon univoque. Dans la figure 2.6, il y a une conflit concernant l'appartenance du numéro 20, situé le long de la rue Saint-Séverin, à cette dernière car il se peut qu'il appartienne à l'autre rue, située à l'intersection. Une autre dimension de l'ambiguïté peut être celle de la *non-spécificité*, lorsque la définition d'une relation ou d'un objet laisse place à plusieurs interprétations possibles. Par exemple, si l'on considère l'adresse « 10 rue Mâcon, Paris », on peut se demander quelle est exactement la relation entre le numéro « 10 » et la voie « rue Mâcon » : le premier est-il simplement situé le long de la seconde, ou bien appartient-il à celle-ci dans une relation de type inclusion ? Enfin, l'*incomplétude* renvoie à l'absence d'information ou à l'existence de lacunes dans les données disponibles. Pouvant être une simple *absence*, lorsque certaines valeurs manquent pour décrire une entité (par exemple l'absence du nom d'une voie dans une liste), elle peut aussi prendre la forme d'une *lacune*, lorsque les données existantes ne couvrent qu'une partie d'une réalité. Comme l'illustre figure 2.6, certains numéros de la rue de la Vielle Boucherie ont été omis ; cette lacune ne constitue pas pour autant un témoignage de leur non-existence.

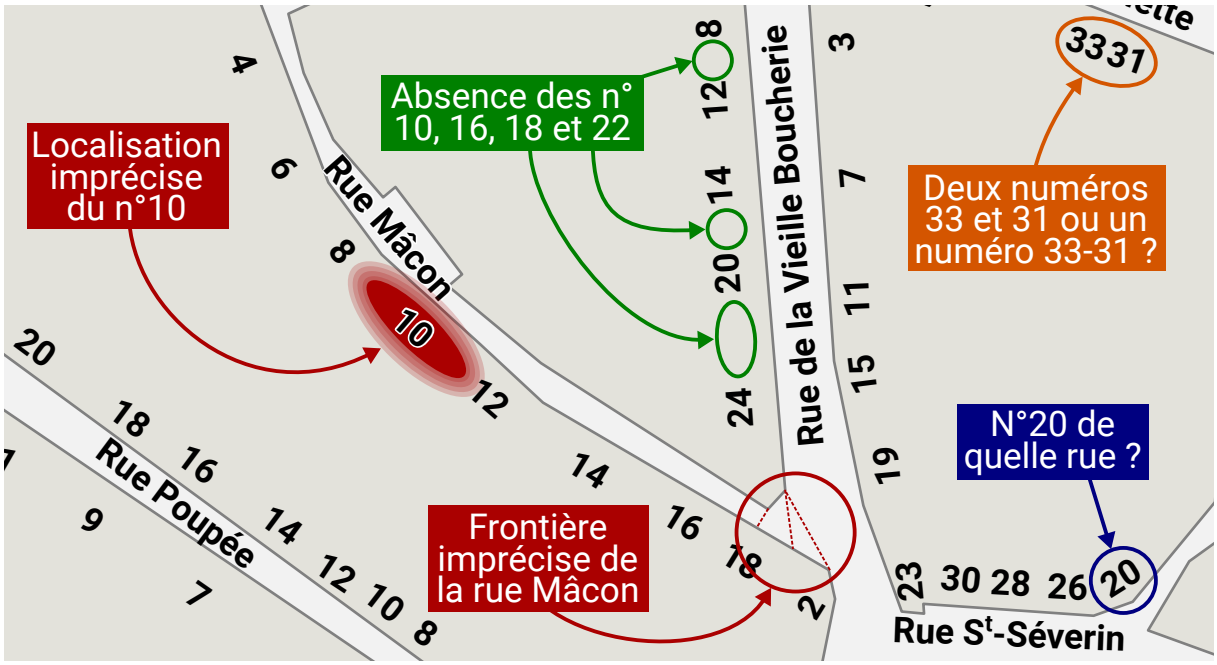
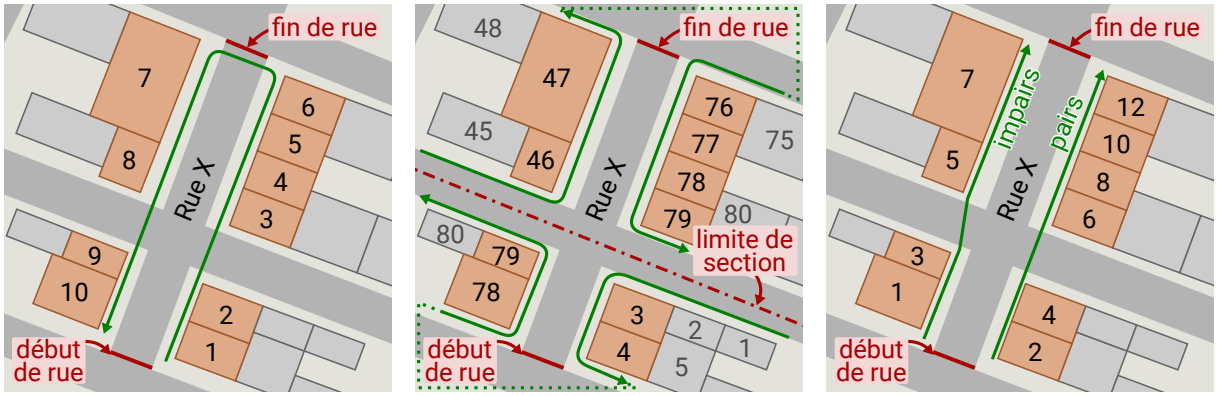


FIGURE 2.6 – Illustration des aspects de l'imperfection des données autour de la rue Mâcon dans l'atlas de Jacoubet. ■ Incertitude, ■ Imprécision, ■ Ambiguïté, ■ Incomplétude.

Ces différentes dimensions de l'imperfection — incertitude, imprécision, ambiguïté et

incomplétude — constituent des caractéristiques fondamentales des informations géographiques historiques. Des travaux récents en géomatique illustrent le besoin de formaliser ces imperfections pour pouvoir les exploiter. Par exemple, Cadorel *et al.* (2022a) proposent une méthode de représentation floue pour traiter les descriptions spatiales vagues présentes dans les annonces immobilières. Les mentions de localisation y sont souvent exprimées de façon approximative (« près de l'avenue des Champs-Élysées »), parfois combinées à des types de lieux ou à des relations spatiales implicites. En s'appuyant sur la théorie des ensembles flous, ils définissent différents niveaux de précision (toponyme seul, type de lieu, relation spatiale, combinaison) et parviennent à modéliser cette imprécision. Un tel cadre est transposable aux sources historiques, où les localisations sont fréquemment exprimées sous une forme vague ou relative (« au coin de la rue »). De leur côté, Feliachi et Abadie (2017) se concentrent sur la précision planimétrique des géométries associées aux ressources du Web de données. Ils distinguent plusieurs manières de positionner un objet dans l'espace (au barycentre du bâtiment, à son entrée, ou le long de l'axe de la rue la desservant) et proposent des indicateurs pour comparer ces localisations dans le cadre d'une application de mise en correspondance automatique de données.

2.6 Histoire des systèmes d'adressage à Paris



(a) Système de Kreenfelt (b) Système révolutionnaire (c) Système impérial

FIGURE 2.7 – Les trois systèmes de numérotation des habitations utilisés à Paris. → : sens de numérotation.

Notre étude des adresses et de leur évolution temporelle se concentre sur la ville de Paris. Pour comprendre les données issues des différentes sources, il est nécessaire d'examiner la manière dont celles-ci représentent les adresses, et plus généralement, comment les systèmes d'adressage qui se sont succédé ont été institués.

Jusqu'à la fin du XVIIIe siècle, les habitations parisiennes n'étaient pas identifiées par des numéros, mais par des enseignes, souvent figurées par des objets, des animaux ou des symboles (Pronteau, 1966). Ces enseignes pouvaient être accompagnées de précisions topographiques ou de repères de voisinage. Ainsi, dans Berty *et al.* (1866), on trouve l'adresse « *Maison du Pied de Griffon [...] dont l'un formait le coin oriental de la rue du Chantre, et l'autre faisait front sur la rue de Beauvais* ».

L'apparition des numéros de rue à Paris à partir du milieu du XVIIIe siècle, puis leur généralisation et leur normalisation au XIXe, ont profondément transformé la manière de référencer les lieux. Ces évolutions successives conditionnent aujourd'hui la lecture et l'interprétation des sources géohistoriques relatives aux adresses, et expliquent les divergences

observées entre documents anciens, plans cadastraux et bases numériques contemporaines.

À partir de la fin du XVIIIe siècle, trois systèmes de numérotation des immeubles se sont succédés à Paris : les systèmes de Kreenfelt, sectionnaire (ou révolutionnaire) et impérial. Le dernier est toujours en vigueur de nos jours. Le système de Kreenfelt, nommé parfois à tort système royal alors qu'il est d'initiative privée et non initiée par le Roi (Depaulis, 2016), est mis en place dès 1779 par Marin Kreenfelt, éditeur de l'*Almanach de Paris* (Kreenfelt et Mangeot, 1789). Il consiste en la numérotation des points d'entrée des bâtiments de toutes les rues. La technique appliquée est celle du fer à cheval, c'est-à-dire que la numérotation débute à l'une des extrémités de la voie, se fait de manière croissante vers l'autre extrémité sur un des côtés puis revient de l'autre côté, comme le montre la figure 2.7a. Dans l'édition de 1789 de L'Almanach de Paris, on peut trouver les adresses suivantes : « *Beudet, M. Passage Montpensier, 13* », « *Liottier, Ag^{t} de Change, R. du Bouloy, 35* », « *Vouves, ^{lle}, R. de Seve, 129* ».

Sous la Révolution française, un autre système fait son apparition et est fondé sur une numération selon les sections et non plus les voies. Ces sections sont des subdivisions de la ville de cette époque. L'ordre de numérotation est peu clair et très peu documentée. Des travaux de recherche comme ceux de Fleury (1995) ou Waquet (2018) ont tenté de décoder ce système. Il en découle que certaines règles générales de numérotation peuvent être dégagées, telles que le fait de commencer la numérotation à une extrémité de la section (au nord-est ou au nord-ouest) ou de suivre l'ordre des contours des îlots d'immeubles. Cependant, ces règles restent générales et ne s'appliquent pas strictement à toutes les sections.

La figure 2.7b fournit un exemple simple de numérotation au sein des deux sections et on remarque qu'il est possible d'avoir deux habitations au sein d'une même rue ayant le même numéro. Le manque de logique claire de ce système de numérotation s'explique en partie par le fait que la numérotation n'était pas faite pour le repérage des immeubles et maisons, mais plutôt pour prélever l'impôt foncier. Dans la version de 1801 de l'annuaire de Duverneuil et La Tynna, on trouve des adresses avec cette numérotation où on trouve le nom de la rue, le numéro de l'adresse dans la section et le nom de cette dernière bien que celle-ci n'y soit pas toujours systématique : « *Carré, R. de la Heaumerie, 14* », « *Vallant, R. d'Enfer, 71. Observatoire* », « *Laugier, R. Neuve Egalité, 290 - Bonne Nouv.* » (Duverneuil et de La Tynna, 1801).

En 1805, un système proche de celui de Kreenfelt remplace le système révolutionnaire : il s'agit du système impérial. La numérotation est faite à la rue comme dans le système de Kreenfelt mais la technique appliquée diffère. Les rues y sont orientées : le point de départ de la rue et donc de sa numérotation est, en principe, l'extrémité la plus proche de la Seine. Dans le cas où la voie est parallèle au fleuve traversant Paris, le point de départ de la numérotation est situé à l'extrémité la plus proche de l'amont de la Seine (Mermet et Varet-Vitu, 2015). Mis à part quelques exceptions, la numérotation est croissante et parallèle de part et d'autre de la rue : les numéros pairs se trouvent sur le côté gauche dans le sens de progression vers la fin de la rue, tandis que les numéros impairs occupent le côté droit (voir la figure 2.7c).

Conclusion

Les sources géohistoriques sont variées — iconographiques, textuelles, structurées ou numériques — et chacune apporte un éclairage spécifique sur l'évolution des entités géographiques. Complémentaires sur les plans spatial, temporel et thématique, elles permettent de reconstituer l'évolution des entités en combinant les informations qu'elles contiennent.

En raison des caractéristiques propres à chaque source, les informations qu'elles véhiculent sont toutefois soumises à des imperfections liées au support, à la vision de leur auteur, à leur fiabilité, ainsi qu'à la nature même des entités géographiques représentées.

L'étude de l'histoire des systèmes d'adressage à Paris, présentée dans la section précédente, illustre concrètement ces enjeux : elle montre comment les pratiques d'adressage, influençant la manière dont les sources se réfèrent aux lieux.

Chapitre 3

Représentation des dynamiques spatio-temporelles

La construction d'un référentiel géohistorique repose sur deux éléments clés, illustrés par la figure 3.1. Le premier élément concerne le choix d'un **modèle spatio-temporel** et d'un **modèle de données** adaptés. Une fois ces modèles définis, il est alors possible de construire le **modèle logique**, qui correspond à la formalisation du modèle spatio-temporel dans le modèle de données choisi. Le modèle spatio-temporel définit la manière dont les entités et leurs évolutions sont représentées dans le temps et l'espace, tandis que le modèle de données (base de données, graphe de connaissances, etc.) permet de formaliser ces représentations sous la forme d'un schéma de données ou d'une ontologie. Dans ce cadre, il est particulièrement important d'expliciter les changements entre états successifs des entités, une dimension rarement abordée dans les modèles existants mais indispensable pour suivre précisément l'évolution des territoires. Le second élément porte sur la méthode de peuplement du référentiel à partir de sources multiples, qui sera abordée en détail dans le chapitre 5.

Ce chapitre présente donc dans un premier temps, via la section 3.1, les différents modèles spatio-temporels existant dans la littérature. Ensuite, la section 3.2 examine la représentation et le suivi des changements. Par la suite, les modèles de données adaptés à la gestion des données spatio-temporelles sont présentés dans la section 3.3. Pour terminer, la section 3.4 aborde deux exemple d'ontologies permettant de représenter et de raisonner sur l'évolution des entités géographiques.

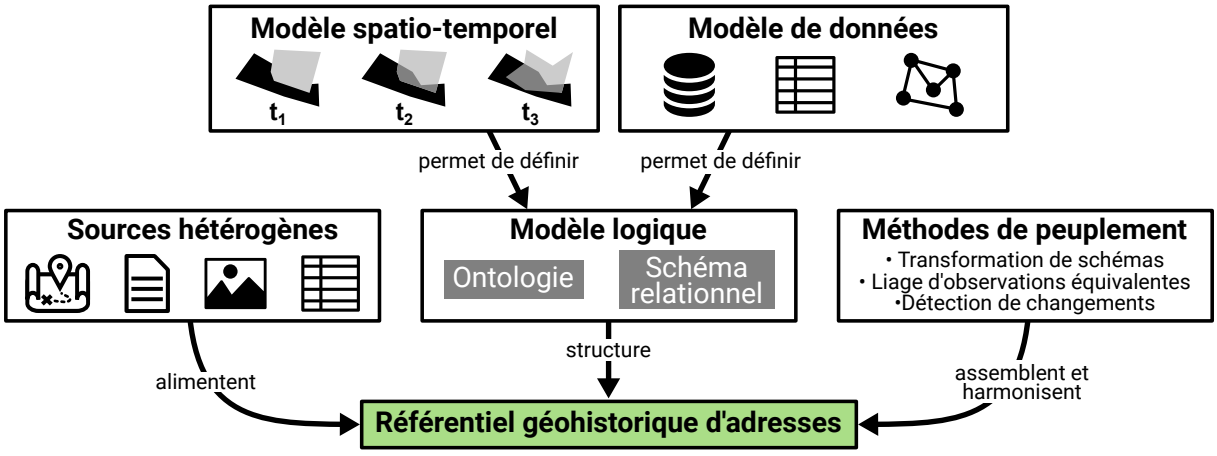
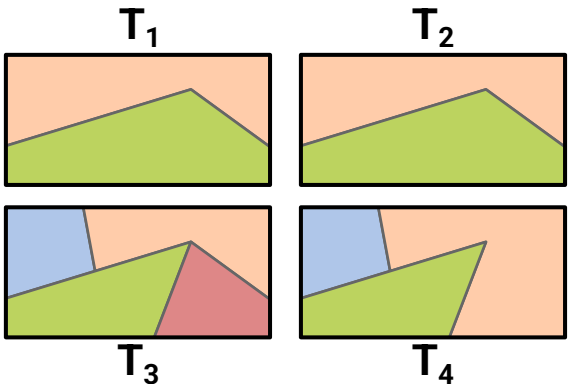


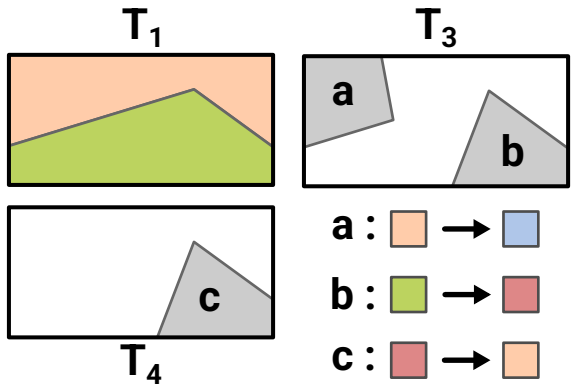
FIGURE 3.1 – Processus pour la reconstruction d'un référentiel géohistorique d'adresses.

3.1 Modèle spatio-temporel

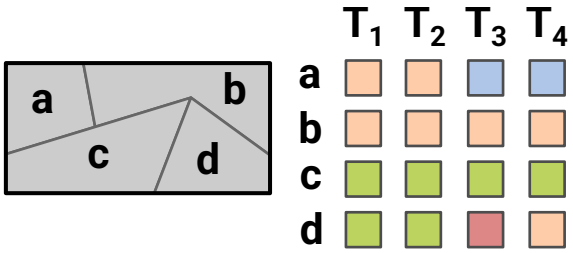
Pour représenter nos données de manière cohérente, il est essentiel de choisir un modèle spatio-temporel capable de capturer les dynamiques spatio-temporelles des entités étudiées, à travers les observations d'états et d'événements disponibles. La figure 3.2 synthétise l'ensemble des modèles présentés où chacun est illustré par rapport à la situation de référence donnée par la figure 3.2a.



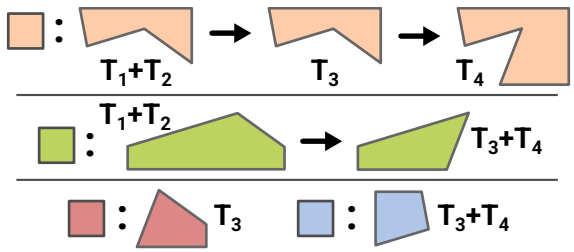
(a) Modèle *snapshot*.



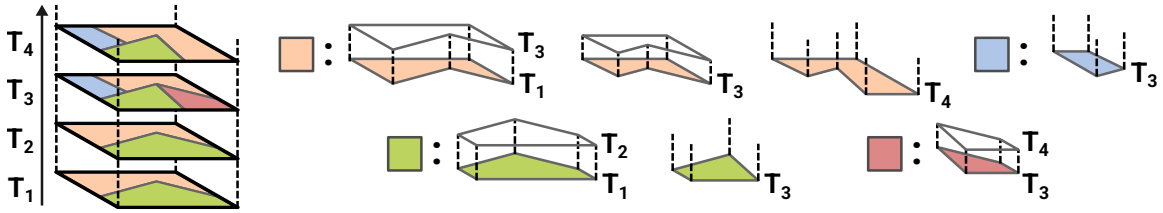
(b) Modèle *snapshot* avec stockage différentiel.



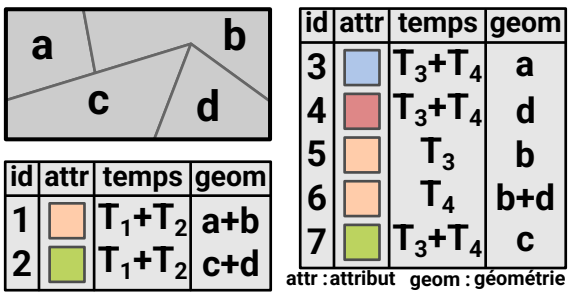
(c) Modèle *Space-Time Composites*.



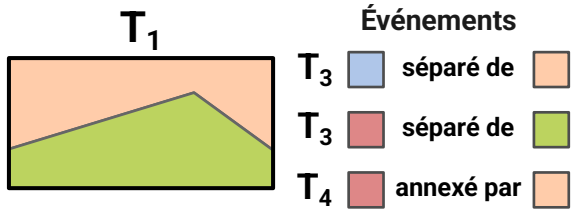
(d) Modèle *Time-stamping*.



(e) Modèle *Spatio-Temporal Object*.



(f) Modèle *Three-Domain Model*.



(g) Modèle basé sur les événements.

FIGURE 3.2 – Comparaison des principaux modèles spatio-temporels.

3.1.1 Modèles à base de snapshots

3.1.1.1 Principes et limites

Les premières approches spatio-temporelles s'appuient sur la méthode des *snapshots* ou instantanés, introduite notamment par Armstrong (1988). Elles consistent à modéliser l'état du territoire à différents instants, chaque état étant représenté comme un *snapshot* distinct et traité de manière homogène. Ces modèles sont adaptés lorsque les données de travail sont issues de sources cartographiques. Par exemple, Dumenieu (2015) utilise ce genre de modèle en entrée de sa méthode de construction d'un graphe spatio-temporel d'évolution d'objets urbains à partir de sources cartographiques anciennes. De son côté, Costes (2016) fait de même pour intégrer, qualifier et analyser les réseaux géo-historiques à partir de sources multiples et hétérogènes. Cette méthode, assez simple, présente néanmoins plusieurs limites. Premièrement, les *snapshots* sont ordonnés temporellement mais aucun lien explicite n'est établi entre les entités d'un état et celles de l'état qui le suit, ce qui impose de comparer deux *snapshots* consécutifs pour détecter une éventuelle une évolution entre représentations successives d'une même entité. De plus, il existe des verrous concernant le tri temporel des *snapshots* car leurs temps valides peuvent être incertains ou se chevaucher. Par ailleurs, le temps est traité comme un attribut quelconque, au même titre qu'un nom ou une longueur. Ensuite, aucune règle temporelle n'est imposée, et l'écart entre deux *snapshots* peut varier selon les sources disponibles. Dans le cas des plans historiques de Paris, cette approche reviendrait simplement à associer une valeur temporelle à chaque carte et à modéliser chacune de manière homogène. Enfin, d'autres problèmes sont liés au volume d'informations stockées, dont une grande partie constitue des doublons puisque entre deux *snapshots* successifs, ce n'est souvent qu'une minorité d'entités qui évoluent, d'autant plus si l'intervalle temporel qui les sépare est faible. Si on prend l'exemple de la figure 3.2a qui modélise l'état d'un territoire sur quatre périodes distinctes, les données sont dupliquées notamment sur les instants T_1 et T_2.

3.1.1.2 Modélisation différentielle

Afin de pallier ces limites liées aux doublons, des approches présentent une optimisation de l'information stockée. Hazelton (1991); Hazelton *et al.* (1992); Langran (1992) proposent de ne constituer qu'un seul *snapshot* représentant l'état initial. À partir ce *snapshot*, les autres peuvent être reconstruits par le biais de changements qui vont décrire comment une entité à un instant donné dérive de sa version de base. La figure 3.2b illustre une déclinaison optimisée de la configuration introduite dans la figure 3.2a. Rien ne change à l'instant T_2 donc son état est similaire à T_1 qui contient l'état de base. Au contraire, pour reconstituer l'état à T_3, on reprend l'état de base auquel sont ajoutées les différences associées à l'instant (la géométrie a et b changent de territoire). À T_4, il suffit de reprendre l'état généré à T_3 et on applique le même principe qu'à l'étape précédente. Cette approche réduit considérablement le volume de données à stocker. Néanmoins, cette méthode présente plusieurs limites. Pour obtenir l'état du territoire à un instant donné, il faut repartir de l'état initial et appliquer tous les changements successifs, ce qui peut devenir lourd à mesure que l'on s'éloigne dans le temps de l'état initial. De plus, le modèle est très peu flexible, ce qui s'explique par le fait qu'une fois construit, il est difficile d'y ajouter un *snapshot* intermédiaire. Si le modèle est construit avec un ensemble de *snapshots* $S = \{s_0, s_1, ..., s_n\}$ et qu'on souhaite inclure s_k, un nouveau *snapshot* situé temporellement entre s_i et s_{i+1}, la ligne de vie des changements d'une entité géographique doit être révisée. Enfin, cette approche est plus adaptée aux réorganisations de l'espace qu'aux

changements radicaux, comme la démolition ou la reconstruction totale d'une entité.

3.1.2 Modèle *Space-Time Composites*

Le modèle *Space-Time Composites* a été proposé par Langran et Chrisman (1988) et repris par Langran (1992). Il constitue une amélioration des approches précédentes où le territoire est conçu comme étant un ensemble d'entités homogènes spatialement et uniformes temporellement. Dans ce modèle, ni les *snapshots*, ni les entités géographiques ne sont les éléments de base. L'unité fondamentale est l'entité spatio-temporelle composite (ou simplement *composite*) qui représente une portion du territoire sur une période donnée durant laquelle l'ensemble de ses attributs restent invariants. On peut considérer cette approche comme partitionnelle dans le sens où elle repose sur une organisation de l'espace-temps en cellules homogènes, ce qu'illustre la figure 3.2c représentant le modèle présenté pour le cas de la figure 3.2a. Ainsi, pour chaque composite (a, b, c et d), on connaît à chaque instant l'appartenance à un territoire modélisé chacun par une couleur. Cependant, ce modèle comporte aussi des limites. Tout d'abord, le territoire est vu ici comme un ensemble d'entités homogènes. Or dans notre cas, nous avons besoin de suivre une diversité d'entités et qui aussi peuvent occuper la même portion de territoire. Une rue et le quartier dans laquelle elle se situe partagent une partie de l'espace. De plus, il existe un risque de surfragmentation car plus les entités évoluent spatialement, plus les composites sont importantes. Cela implique des mises à jour délicates, notamment si l'on souhaite ajouter de nouvelles données. Enfin, ce modèle exige de disposer d'un espace de stockage conséquent. Ainsi, cette méthode fonctionne bien pour de petites zones, mais devient coûteuse sur de grands territoires (niveau régional, national). Ensuite, il y a une perte d'identité des objets, ce qui rend difficile le suivi d'une entité dans le temps. Il n'existe pas de lien de filiation entre des entités successives : par exemple, si un territoire A est divisé en B et C, aucun lien n'est automatiquement établi entre eux. Enfin cette approche a le même inconvénient principal que les deux précédentes puisqu'elle ne permet pas de représenter explicitement des changements.

3.1.3 Modèle *Time-stamping*

Pour contrer le problème lié à la perte d'identité et mieux suivre l'évolution de chaque entité, le modèle *Time-stamping* a été proposé par Hunter et Williamson (1990). Contrairement aux méthodes basées sur les *snapshots*, il permet de suivre l'évolution des entités, en se centrant sur elles et en considérant leur évolution comme une succession d'états. La ligne de vie d'une entité correspond alors à la succession de ces entités composites, ce qui fait écho à la notion de succession d'états spatio-temporels proposée par Hornsby et Egenhofer (2000) (voir sous-section 5.2.2). Dans ce modèle, le temps est traité comme un attribut, au même titre que les autres propriétés des entités, mais il définit la validité de l'état de l'entité géographique. Un nouvel état est créé dès qu'un attribut de l'entité évolue.

Cette approche est pratique pour reconstituer des états, car on connaît l'état de chaque entité à chaque instant. La figure 3.2d adapte le modèle à l'exemple de la figure 3.2a. Chaque entité, identifiée par une couleur, est associée à une liste d'états qui ont chacun un temps valide. Cette méthode est appliquée dans la gestion des versions des données d'OpenStreetMap : chaque objet évolue indépendamment et un changement géométrique ou attributaire implique la création d'une nouvelle version, copie de la précédente avec les mises à jour.

Cependant, le modèle présente des limites. Il n'est pas très adapté pour représenter explicitement les changements ou leurs causes. Si on reprend la figure 3.2d, on connaît l'enchaînement de chaque état de chaque entité mais il est impossible de savoir ce qui a provoqué la transition. En outre, cette approche peut entraîner une duplication d'information : si une entité conserve sa géométrie mais voit ses attributs évoluer fréquemment, la géométrie est copiée dans chaque version.

3.1.4 Modèle *Spatio-Temporal Object*

Le modèle *Spatio-Temporal Object* proposé par Worboys (1992) considère les entités géographiques comme des entités spatio-temporelles. Chacune est définie dans un espace à trois dimensions : deux dimensions spatiales et une dimension temporelle. Ainsi, une entité n'est plus seulement décrite par sa géométrie dans le plan, mais par une empreinte qui évolue dans le temps.

Ce modèle repose sur l'hypothèse de changements soudains. Une entité garde une forme stable pendant un intervalle de temps donné, puis subit une transformation ponctuelle qui conduit à une nouvelle empreinte spatiale. Les évolutions continues, comme une frontière qui se déplace progressivement ou une rivière qui s'élargit, ne sont pas représentées car le modèle est conçu pour décrire des phénomènes discrets. La figure 3.2e adapte le modèle à l'exemple de la figure 3.2a. L'entité représentée en vert (■) illustre ce principe : elle possède une première forme, valable à partir de T_1, puis une seconde forme qui la remplace à T_3. Le modèle relie ainsi ces deux empreintes spatiales dans l'espace-temps, constituant un objet spatio-temporel discontinu.

Ce modèle est particulièrement adapté aux contextes où les changements sont enregistrés comme des événements ponctuels, comme dans les données cadastrales ou historiques. Il permet de suivre la succession des états stables d'une entité mais ne prend pas en compte les phénomènes continus.

3.1.5 *Three-Domain Model* (TDM)

Peuquet et Wentz (1994) introduit *Triad Framework*, un cadre conceptuel pour penser la modélisation spatio-temporelle. L'idée consiste à voir tout phénomène géographique comme étant la composition des dimensions spatiale (*où*), temporelle (*quand*) et thématique (*quoi*). Yuan (1994) utilise ce cadre pour proposer le *Three-Domain Model* (TDM) qui reprend les avantages des approches précédentes tout en éliminant leurs redondances en représentant indépendamment chaque dimension. Le territoire est partitionné spatialement en unités élémentaires, temporellement en périodes élémentaires, et les attributs sont combinés pour former une base de données.

Le résultat est la combinaison d'un ou plusieurs éléments de chaque partition. La figure 3.2f illustre l'adaptation de l'exemple de la figure 3.2a au modèle présenté dans cette section. Il existe une partition spatiale $G = \{a, b, c, d\}$, temporelle $T = \{T_1, T_2, T_3, T_4\}$ et sémantique $S = \{$■, ■, ■, ■$\}$. Dans ce modèle, les évolutions spatiales et attributaires sont modélisées indépendamment. Il offre également la possibilité de réaliser des requêtes complexes combinant ces trois dimensions. En couplant deux dimensions, on peut répondre à une question dans l'autre dimension :

- **quoi + où → quand** : « À quel moment l'entité orange (■) possédait-elle pour partie de son territoire l'emprise de la géométrie d ? »
- **quoi + quand → où** : « Quelle était l'emprise de l'entité verte (■) à l'instant T_2 ? »

— **où** + → **quoi** : « Quelle entité se trouvait sur l'emprise de la géométrie a à T_4 ? »

Cette approche est intéressante car elle modélise séparément les dimensions tout en permettant leur combinaison pour reconstruire les évolutions. Elle se révèle particulièrement adaptée lorsque l'on souhaite reconstituer l'évolution des entités à partir de données hétérogènes et fragmentaires, où la diversité des dimensions représentées est importante. Une des limites de ce modèle réside toutefois dans la non-explicitation des changements.

3.1.6 Modèles de trajectoires fondés sur l'identité

La principale limite des approches précédemment mentionnées porte sur de la non représentation explicite des changements, ce qui peut limiter l'étude de dynamiques spatio-temporelles des territoires. Toutefois, il en existe d'autres qui prennent en compte cet aspect. Hornsby et Egenhofer (1997) et Hornsby et Egenhofer (2000) les modélisent en utilisant le concept de l'identité, essentiel pour suivre l'évolution d'une entité. L'approche consiste à construire des lignes de vie de chaque entité, c'est-à-dire à reconstituer une succession d'états spatio-temporels où des changements tendent à décrire la manière dont deux états successifs évoluent (création, disparition...) Cependant, ces approches présentent une limite importante car elles ne prennent pas en comptent les trois dimensions fondamentales. En effet, elles se concentrent uniquement sur les dimensions spatiale et temporelle et ne gèrent pas explicitement l'évolution des attributs thématiques. Autrement dit, des changements de nom, de statut administratif ou d'autres propriétés de l'entité ne sont pas intégrés dans son suivi spatio-temporel. Or la dimension thématique reste importante car, comme il a été dit dans la section 1.3, le référentiel que nous visons à construire n'est pas centré sur la seule dimension spatiale.

3.1.7 Modélisation fondée sur les événements

Les modèles fondés sur les événements se concentrent sur l'événement lui-même et non nécessairement sur l'état des entités. Un événement est considéré comme un élément à part entière, qui peut être temporalisé, spatialisé et thématisé. Cette approche permet de suivre le processus qui produit un changement et d'enregistrer les relations et attributs spécifiques à l'événement. Peuquet et Wentz (1994) ont proposé une approche initiale : on part d'un état initial et on décrit de manière chronologique les événements appliqués sur la zone d'étude.

Cette méthode est similaire à l'optimisation présentée en sous-section 3.1.1, mais elle explicite les événements en tant qu'éléments indépendants plutôt que de se contenter d'optimiser les snapshots. Chaque événement est lié à des changements spécifiques comme le montre la figure 3.2g. À partir de la situation de référence fournie par la figure 3.2a, on observe trois évolutions entre T_1 et T_4, qui sont décrite par ce modèle par trois événements associés chacun à des instants et décrivant la manière dont les entités évoluent. Cependant, cette approche a ses limites. La reconstruction d'un état dépend de l'état initial et de tous les événements successifs, ce qui implique que toutes les informations doivent être correctes et complètes. Cette exigence la rend moins fiable pour des données fragmentaires ou imparfaites. Par exemple, si un événement est oublié, la reconstitution des états ultérieurs sera incorrecte.

D'autres approches basées sur les événements existent, comme l'Event Oriented Spatio-Temporal Data Model abrégé en ESTDM (Peuquet et Duan, 1995). Avec des données fragmentaires et lacunaires, il est nécessaire de pouvoir décrire les évolutions de manière indépendante, comme le fait le Three-Domain Model, mais sur l'ensemble des attributs.

Toutefois, lorsque la dimension spatiale n'est pas toujours disponible dans les données, combiner les géométries pour reconstruire la position des entités devient difficile. De plus, un problème similaire à celui rencontré dans les différentes variantes du modèle à base de *snapshots* se pose aussi ici. En effet, tous ces modèles supposent l'existence d'un état de référence bien construit, sur lequel les événements viennent s'appliquer. Or, dans le cas de données fragmentaires et lacunaires, cet état initial peut être difficile à établir, ce qui rend ces approches peu adaptées.

3.1.8 Bilan des modèles spatio-temporels

Les modèles présentés offrent des approches variées pour représenter des phénomènes spatio-temporels. Cependant, ils montrent rapidement leurs limites lorsqu'il s'agit de traiter des données géohistoriques, caractérisées par leur fragmentation, leur hétérogénéité et l'incertitude des informations. En particulier, aucun de ces modèles ne permet de gérer pleinement le flou qui peut affecter simultanément les dimensions spatiale, temporelle et thématique.

Notre démarche doit s'inspirer de ces travaux tout en visant à dépasser leurs contraintes et leurs limites. On retient :

— la nécessité de combiner descriptions par états et par événements, afin de capturer à la fois les situations stabilisées des entités et leurs évolutions ;
— l'importance de ne pas dépendre d'un état de référence unique, difficile à établir à partir de données fragmentaires ;
— la pertinence de modéliser séparément chaque dimension (temps, espace, attributs), tout en permettant leur articulation.

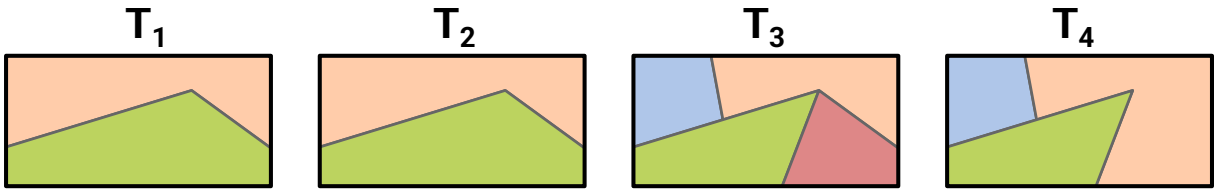
De plus, contrairement aux approches centrées sur la géométrie, nous considérons la dimension spatiale comme une parmi d'autres, ce qui permet de gérer plus efficacement les situations où l'information géographique est absente ou incertaine.

3.2 Représentation et suivi des changements

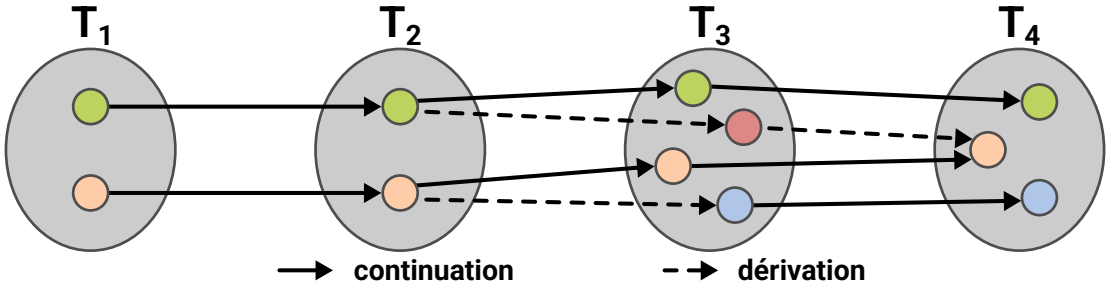
Dans le cadre de cette thèse, un changement est défini comme l'explicitation d'une ou plusieurs différences entre deux états successifs d'une entité géographique. Avec une définition aussi large, il y pourrait potentiellement y avoir une infinité de changements, tout comme il existe une infinité de différences. Pour rendre la notion opérationnelle, il est nécessaire de cadrer cette définition. Selon Claramunt *et al.* (1998), le changement est constitué par l'altération des propriétés d'un objet persistant ou d'une région de l'espace. Le changement est ainsi vu comme étant la représentation de différences dans l'état d'une entité ou d'un phénomène observé à des moments différents. Les changements dans les états des entités géographiques sont le résultat d'événements ayant lieu à des instants définis ou non. L'état d'une entité géographique correspond à une forme stable de cette entité dans un intervalle de temps. Ainsi, on considère qu'un changement est une évolution appliquée à une ou plusieurs entités géographiques, qu'elle soit attributaire ou spatiale.

3.2.1 Modéliser les changements

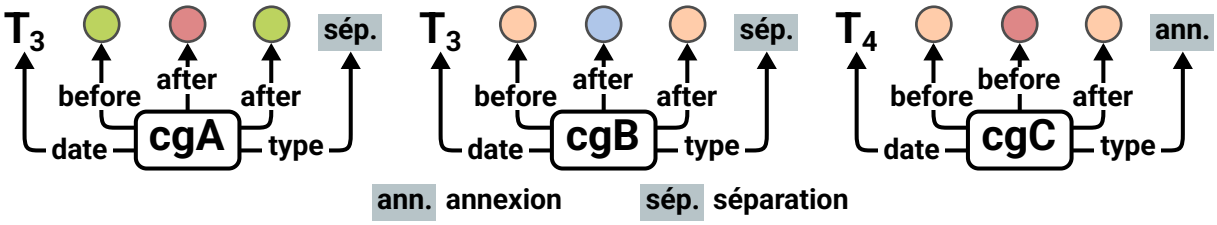
Peu d'approches proposent une modélisation explicite des changements, malgré leur importance pour l'étude des dynamiques spatio-temporelles des territoires. Or, dans le cadre de projets collaboratifs tels que Wikipédia ou OpenStreetMap, l'évolution des données est soigneusement tracée, car ces environnements reposent sur une approche cumu-



(a) Situation de référence (*modèle snapshot*).



(b) Modèle de représentation des évolutions spatio-temporelles à partir de liens de filiation.



(c) Modèle de représentation des évolutions spatio-temporelles avec le *Change-Bridge*.

FIGURE 3.3 – Comparaison des principaux modèles de changement.

lative où chaque contribution enrichit progressivement la base de connaissances. De manière similaire, notre démarche vise à agréger des fragments d'information issus de sources multiples et hétérogènes, en mettant l'accent sur la conservation et la reconstitution de l'ensemble des évolutions afin de garantir la qualité et la cohérence des données Hartig (2009). Dans Wikipédia, l'évolution d'une page se modélise par la succession de ses versions éditées. Ce suivi est possible car les éditions se font sur les pages de manière indépendante, la modification d'une page n'a pas d'influence sur une autre. Sur OpenStreetMap, l'approche est différente car il est possible d'éditer plusieurs éléments à la fois et la modification d'une entité peut impliquer un changement sur une autre. Pour suivre l'évolution d'un objet géographique au fil des éditions, OpenStreetMap crée une nouvelle version de cet objet à chaque fois qu'il est modifié. Cependant, une session d'édition ne concerne généralement pas un seul objet, mais un ensemble d'objets modifiés par un même utilisateur. Pour regrouper ces modifications, OSM introduit un groupe de modifications (`changeset`), qui rassemble l'ensemble des nouvelles versions d'objets générées au cours d'une même session d'édition. La limite de cette approche est que les changements entre deux versions successives n'y sont pas explicités, ce qui implique qu'il faille faire des comparaisons systématiques afin d'en déduire des évolutions. Pour éviter ce genre d'opérations, Keßler *et al.* (2011) proposent une ontologie dérivée de Provenance (Hartig, 2009) qui permet de décrire les changements effectués sur les éléments au sein d'un changeset. L'ontologie est intéressante car elle explicite l'ensemble des changements élémentaires appliqués à un élément (ajout, suppression, mise à jour d'un attribut, changement de géométrie...) et permet de les regroupes dans une collections de changements, ce qui peut être pratique dans une approche cumulative où nous avons des fragments d'informations

qui vont s'ajouter en permanence.

Cependant, au-delà de l'explicitation des modifications élémentaires, certaines méthodes vont plus loin en modélisant les continuités et ruptures entre entités au fil du temps. C'est le cas des approches fondées sur les liens de filiation, qui permettent de représenter explicitement la succession des états d'un même objet et d'analyser les dynamiques spatio-temporelles associées. Dans la continuité, d'autres travaux introduisent des mécanismes intermédiaires, tels que les change bridges, qui visent à relier les états successifs d'un objet en explicitant systématiquement les transformations sous-jacentes.

3.2.1.1 Liens de filiation

Del Mondo (2011), tout comme Dumenieu (2015), Costes (2016) et Leborgne *et al.* (2019), relient les observations entre des *snapshots* consécutifs afin de déduire les dynamiques spatio-temporelles. Pour cela, ils s'appuient sur l'établissement de liens de filiation entre entités.

Le défi consiste à identifier automatiquement les liens entre les observations issues de sources successives, à déduire le type de relations de filiation permettant d'expliquer le passage des observations d'une source à la suivante et à gérer l'incertitude associée à des données imparfaites. Afin de distinguer les informations issues directement des sources de celles générées par chacune des méthodes proposées, les liens de filiation sont enregistrés dans un graphe annexe, qui regroupe l'ensemble de ces relations. Ce graphe peut être interprété comme un « graphe de changements », puisqu'il décrit de manière détaillée l'évolution des entités.

Par exemple, si l'on considère un ensemble de *snapshots* comme celui de la figure 3.3a, on peut représenter les observations qu'ils contiennent dans un graphe comme l'illustre la figure 3.3b. Par la suite, les observations entre sources consécutives sont liées pour décrire des filiations comme la continuation ou la dérivation. En fonction de la combinaison des liens, on peut en déduire des évolutions. Les deux observations à T_2 sont chacune liées à un lien de continuation et un autre de dérivation à des observations de T_3 : on peut ainsi en déduire des séparations qui on lieu à T_3.

Certaines évolutions impliquent plusieurs entités géographiques initiales et résultantes, ce qui complique leur représentation. Au lieu d'utiliser des liens simples qui relient deux entités, Dumenieu (2015) propose d'utiliser des liens plus complexes qui permettent de relier deux ensembles d'entités, offrant ainsi une formalisation adaptée à ces cas complexes. Par ailleurs, les imperfections des données rendent parfois les filiations incertaines ; il est alors possible d'associer des scores de confiance aux liens pour représenter cette incertitude.

Cependant, cette approche présente plusieurs limites pour notre cas d'étude. Elle suppose l'existence de *snapshots* complets, ce qui n'est pas le cas lorsque l'on travaille avec des sources fragmentaires et hétérogènes. Une conséquence directe est qu'il devient difficile d'ajouter une nouvelle source s_k entre deux *snapshots* existants s_i et s_{i+1}, car cela nécessite de réorganiser les liens existants pour les reconstruire. En pratique, l'approche n'est donc pas incrémentale.

3.2.1.2 Change-Bridge

Le *Change-Bridge*, proposé par Kauppinen et Hyvönen (2004), s'appuie sur un modèle à base de graphes de connaissances. Il offre une approche pour expliciter et détailler les évolutions spatio-temporelles, en s'attaquant à la question centrale de la représentation des changements affectant les entités, leurs relations et les concepts dans des graphes de

connaissance successifs. À partir du constat que les entités géographiques subissent des changements et que leurs relations évoluent, les auteurs proposent d'organiser ces évolutions en versions successives d'un graphe (notées $G_1, G_2, \ldots, G_n$), chacune décrivant un état particulier du monde à un instant donné. Cette approche ne concerne pas uniquement les ressources, mais également les concepts et propriétés mobilisés pour les décrire qui peuvent également évoluer dans le temps — par exemple, la notion de *commune française*, apparue à la Révolution, se substituant à celle de paroisse.

Pour relier ces différentes versions, les auteurs introduisent la notion de *Change Bridge*, une relation de correspondance explicite entre un ensemble de ressources R_i d'un graphe G_i et un ensemble R_{i+1} dans G_{i+1}. Cette correspondance est exprimée dans un graphe tiers, garantissant l'intégrité des versions de base. Les changements sont décrits par l'intermédiaire d'une ressource dont la classe détermine le type de changement et une date y est associée pour la temporaliser. Dans un changement, il y a des ressources qui sont remplacées par d'autres. Ainsi, un changement indique les ressources à remplacer en étant reliées à elles via le prédicat `before` tandis que `after` indique les ressources qui remplacent les précédentes.

Si on reprend l'exemple initial de la figure 3.3a où il existe trois changements (deux à l'instant T_3 et un à T_4), on peut les représenter chacun comme le montre la figure 3.3c. À T_3, les deux entités qui existaient à T_2 sont séparées en deux et donc le modèle *Change-Bridge* permet d'indiquer pour chaque changement de type *séparation* ayant lieu à la date T_3 que les entités initiales reliées par le prédicat `before` deviennent les entités reliées par `after`.

Cette première contribution présente toutefois des limites. Elle propose une manière de représenter des changements (qui doivent être instanciés manuellement) mais ne fournit pas de mécanisme pour les détecter automatiquement. Dans une seconde publication, Kauppinen *et al.* (2008) approfondissent cette problématique en l'appliquant à un cas concret, celui de l'évolution des municipalités finlandaises entre 1865 et 2007. Cette étude se concentre sur la modélisation explicite des changements géographiques au fil du temps, tels que la création, la fusion, la division, le changement de nom, ou encore la modification d'appartenance territoriale. Les auteurs proposent un modèle simple pour représenter chaque changement sous forme d'instance caractérisée par plusieurs attributs : un identifiant unique, une date, un type de lieu (par exemple commune), une description textuelle, un type de changement (selon une taxonomie spécifique), et éventuellement des propriétés comme `from`, `to` ou `movepart`, permettant de préciser quels territoires sont transférés entre quelles entités. Les entités géographiques sont modélisées dans deux versions de l'ontologie : une version *actuelle* et une version *historique*. Dans cette dernière, chaque entité est associée à un intervalle de validité temporelle, reflétant l'application du changement à une date donnée. Ainsi, chaque modification engendre la création d'une nouvelle version de l'entité concernée. Cette approche s'inspire du modèle *4d-fluent* (Welty et Fikes, 2006), dans lequel une entité est vue comme la somme de ses *parties temporelles*, chacune correspondant à un état sur une période donnée. Les relations entre entités sont alors définies entre ces parties temporelles. Par exemple, pour représenter que Paris est la préfecture du département de la Seine entre 1790 et 1968 comme le montre la figure 3.4, la relation serait portée non pas entre les entités *Paris* et *Seine* elles-mêmes, mais entre leurs versions temporellement valides sur cette période.

Bien que cette approche fournisse un cadre cohérent pour représenter des évolutions spatio-temporelles, elle présente plusieurs limites. Elle reste centrée sur le cas des municipalités, dont les propriétés sont relativement peu nombreuses et bien structurées, et repose sur des données complètes et préalablement harmonisées. Les changements sont

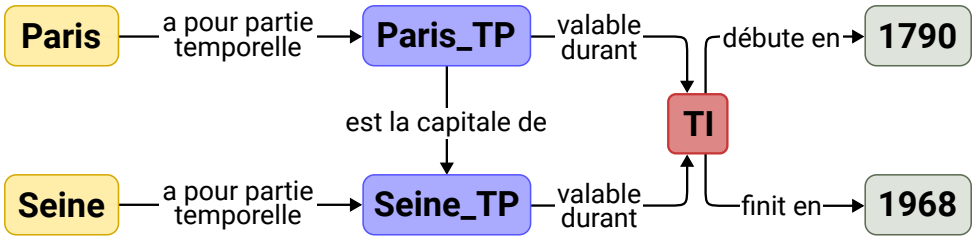


FIGURE 3.4 – Représentation de la relation *Paris est la capitale du département de la Seine entre 1790 et 1968* avec l'approche 4d-fluent.

renseignés manuellement et injectés dans une structure conforme à l'ontologie cible, ce qui rend la méthode difficile à appliquer lorsque les données sont fragmentaires, issues de sources hétérogènes, ou fondées sur des états d'entités plutôt que sur des événements. Cette dépendance à la complétude rend l'approche peu robuste face aux omissions ou révisions. Par exemple, la découverte a posteriori d'un changement non pris en compte nécessite une mise à jour rétroactive complexe, susceptible de propager des erreurs et de déstabiliser les identifiants des entités dans l'ensemble du graphe.

En somme, les travaux de Kauppinen posent les bases conceptuelles intéressantes pour articuler les différentes versions d'un graphe d'entités géographiques évolutives. Toutefois, leur mise en oeuvre reste cantonnée à des situations où les données sont suffisamment complètes, homogènes et bien structurées, ce qui relève plus de l'exception que de la règle lorsqu'il est question d'information géographique issue de sources historiques hétérogènes.

3.2.2 Catégorisation des changements

En fonction des types d'entités géographiques en jeu ainsi que des objectifs auxquels doit répondre un référentiel spatio-temporel, il est nécessaire d'établir une typologie des événements du monde réel affectant ces entités, avant de présenter la typologie des changements correspondants dans les données.

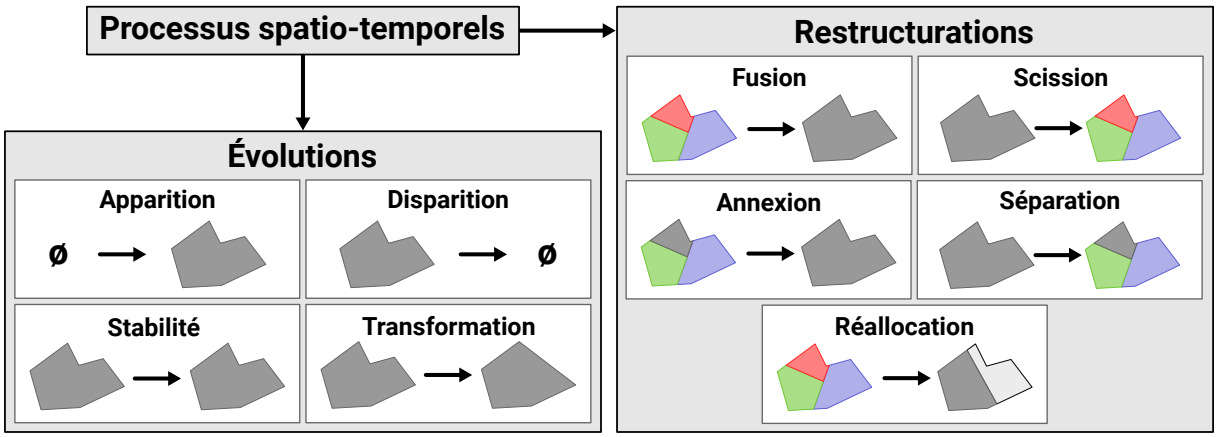


FIGURE 3.5 – Typologie des processus spatio-temporels utilisée par Dumenieu (2015).

Dumenieu (2015) propose une typologie des processus spatio-temporels s'appuyant sur celle de Claramunt et Thériault (1995). L'objectif est de décrire la manière dont évolue une entité au sein de différentes sources, tout en notant qu'une entité peut rester inchangée. Il distingue les évolutions, traitant des processus impliquant une seule entité,

des restructurations qui en impliquent plusieurs (voir figure 3.5). Quatre types d'évolutions sont possibles. L'apparition et la disparition décrivent respectivement la première et la dernière observation d'une entité dans le monde réel. La stabilité permet de décrire une continuation entre deux observations tandis que la transformation décrit une dérivation, une modification de son emprise spatiale. Côté restructuration, il existe cinq processus. La fusion décrit le processus indiquant l'assemblage de plusieurs entités impliquant la formation d'une nouvelle alors que la scission est le processus symétrique, ce qui veut dire une entité est divisée en plusieurs nouvelles. L'annexion et la séparation sont des processus proches des derniers cités. D'une part, l'annexion décrit l'absorption d'entités par une autre, la différence avec la fusion est qu'aucune nouvelle entité n'est créée. On peut concevoir l'annexion comme un cas particulier de la fusion où une des entités qui fusionne existe toujours. La séparation est la symétrie de l'annexion dans le sens pour l'entité qui est divisée en plusieurs autres continue aussi d'exister. Enfin, la réallocation est un processus plus complexe car il décrit une réorganisation d'un ensemble d'entités en un nouveau. Bien que cette typologie permette de décrire certaines dynamiques spatiales, elle présente des limites pour notre cas, car elle ne prend pas en compte la dimension thématique, cruciale dans notre travail. En effet, nous nous intéressons aux changements qui s'opèrent dans les données, tels que des modifications de nom ou d'identifiant, qui reflètent les événements du monde réel affectant les entités géographiques.

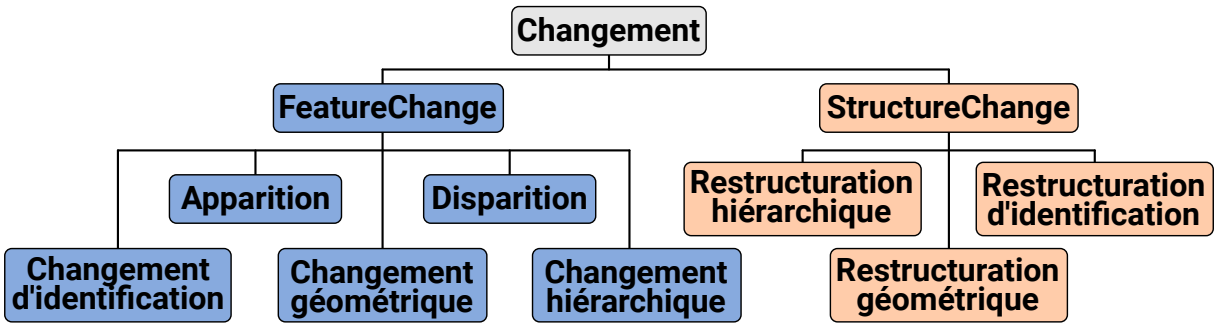


FIGURE 3.6 – Typologie simplifiée des changements proposée par Bernard *et al.* (2017).

Bernard *et al.* (2017) s'appuie sur la typologie des événements proposée par Plumejeaud *et al.* (2011) pour définir une typologie des changements qui s'opèrent dans les données. Sa taxonomie distingue notamment les changements impliquant une seule entité (*FeatureChange*) de ceux en impliquant plusieurs (*StructureChange*), comme le montre la figure 3.6. Elle propose une typologie plus générique où les processus utilisés par Dumenieu (2015) qui sont strictement spatiaux sont repris mais contenus dans des concepts plus globaux. Du côté des *FeatureChange*, on retrouve l'apparition et la disparition ainsi que les changements géométriques, c'est-à-dire la transformation. Les changements d'identification (nom, identifiant, acronyme) y sont aussi inclus tout comme les changements hiérarchiques (nouvelle position dans hiérarchie de la nomenclature). Pour les *StructureChange* qui traitent des changements affectant simultanément plusieurs entités, on retrouve les restructurations géométriques précédemment citées auxquelles s'ajoutent les restructurations hiérarchiques et d'identification.

Enfin, on peut citer une autre taxonomie de changement qui est encore plus générique que la précédente puisqu'elle ne s'applique pas uniquement aux entités géographiques. En plus d'introduire une manière de modéliser les changements, Kauppinen et Hyvönen (2004) en proposent une taxonomie en distinguant quatre types : *resourceChange* (changement des individus), *hierarchyChange* (modification dans la hiérarchie des classes), *property-*

Change (évolution des propriétés associées aux classes), et *typeChange* (requalification entre instances et classes). La catégorie *resourceChange* s'avère particulièrement pertinente pour modéliser les phénomènes de fusion, de scission ou de transfert d'identité entre entités.

3.3 Modèles de données pour représenter l'information spatio-temporelle

3.3.1 Les bases de données relationnelles

Les premières approches proposées pour représenter des données spatio-temporelles reposaient généralement sur l'utilisation de bases de données relationnelles (Armstrong, 1988; Langran, 1992). Le modèle relationnel permet la représentation des entités et de leurs évolutions au moyen de tables et clés, mais il impose un cadre rigide peu adapté à l'intégration de données hétérogènes et fragmentaires. En effet, la diversité des sources et des formats, ainsi que la variabilité des informations disponibles (par exemple, le fait que certaines entités sont dépourvues de géométrie explicite), rendent difficile leur représentation cohérente dans un schéma logique relationnel.

3.3.2 Les modèles de graphes

À l'inverse, les modèles de graphes offrent une structure plus flexible, capable de relier des entités au moyen de relations explicites et de suivre des évolutions complexes dans le temps. En géomatique, les graphes ont initialement servi à décrire des relations entre des entités géographiques, à représenter des structures spatiales comme des réseaux routiers ou hydrographiques (Jiang *et al.*, 2000; Jiang et Claramunt, 2004), voire à résoudre des problèmes spécifiques comme le calcul d'itinéraires à l'aide d'algorithmes de plus court chemin avec, par exemple, l'algorithme de Dijkstra (Dijkstra, 1959). Les graphes ont ensuite été utilisés pour représenter des liens de filiation entre des observations d'entités, afin de suivre l'évolution des objets spatiaux dans le temps (Spéry *et al.*, 2001; Del Mondo, 2011; Dumenieu, 2015; Costes, 2016). Dans ce contexte, un graphe peut être considéré comme un réseau reliant des versions d'entités géographiques entre elles par des relations de continuation ou de filiation, ce qui permet de représenter des trajectoires ou des successions d'états.

Néanmoins, les modèles de graphes utilisés dans les travaux cités ici présentent certaines limites. Ils restent souvent peu expressifs sémantiquement et ne permettent pas de définir précisément le type de relation selon le contexte. Cette faible expressivité limite l'analyse détaillée des dynamiques spatio-temporelles et la capacité à intégrer des informations thématiques ou causales sur les changements.

3.3.3 Les graphes de connaissances

L'utilisation d'un modèle de graphe de connaissances s'inscrit dans une approche de modélisation qui permet de représenter des versions d'entités et leurs relations de manière souple, évolutive et interopérable. Le modèle standard en vigueur pour représenter des graphes de connaissances sur le Web est le modèle RDF [1]. Tout d'abord, ce modèle permet de s'affranchir des contraintes liées aux formats et aux schémas de données en reposant

1. https://www.w3.org/RDF/

sur une structure unique : le triplet *Sujet – Prédicat – Objet*. Cette formalisation simple et universelle facilite la représentation d'informations de natures très diverses. De plus, le recours à des vocabulaires standardisés, représentés selon les modèles standards RDFS [2], OWL [3], ou SKOS [4] partagés et ouverts favorise l'interopérabilité sémantique. L'ajout de liens explicites permet de dépasser les logiques de silos en reliant des jeux de données publiés sur le Web et en naviguant de l'un à l'autre. Enfin, ce modèle offre une tolérance élevée aux données imparfaites, notamment celles issues de sources d'archives — incomplètes, imprécises, incohérentes ou dépourvues de géométrie — tout en fournissant des langages et outils adaptés pour détecter les incohérences, corriger et enrichir ces informations. Ce choix facilite en outre la publication des données selon les principes FAIR et leur intégration dans une démarche de science ouverte.

Le modèle RDF est largement répandu pour représenter de grandes bases de connaissances sur le Web comme DBPedia [5] ou Yago [6]. Selon Weikum *et al.* (2021), qui proposent une synthèse des concepts, méthodes et applications liés à la construction et à la curation de grandes bases de connaissances, une base de connaissances se distingue par plusieurs caractéristiques fondamentales. Son contenu est centré sur un domaine d'intérêt spécifique, et la qualité des informations qu'elle contient tend vers une précision de niveau quasi-humain, avec une mise à jour continue au fil du temps. Contrairement à une base de données relationnelle traditionnelle, une base de connaissances ne repose pas sur un schéma figé. Elle permet au contraire d'ajouter dynamiquement de nouveaux types, attributs et relations à mesure que le périmètre couvert s'élargit. Cette capacité d'extension est essentielle dans les domaines où la couverture est, par nature, ouverte et incomplète. La base est donc conçue pour rester perpétuellement en construction.

Pour répondre à ces objectifs, un certain nombre de requis doivent être satisfaits. La base doit tout d'abord garantir l'exactitude des entités qu'elle contient, lesquelles doivent correspondre à des éléments réels et pertinents dans le domaine visé. En parallèle, elle doit viser une couverture aussi large que possible des entités d'intérêt. L'enrichissement progressif de la base repose sur son expressivité : on cherche à densifier la base en multipliant les déclarations pertinentes sur les entités. Par ailleurs, la structure de la base doit faire preuve d'agilité pour s'adapter à l'évolution des données, des schémas ou des besoins auxquels elle doit répondre. L'efficacité est également un enjeu central, notamment face aux volumes massifs de données à traiter — des millions d'entités et des milliards de relations pouvant être impliquées (Zou, 2020; Zhou *et al.*, 2021). À cela s'ajoute une exigence de traçabilité : il doit être possible de retracer, pour chaque assertion de la base, son origine, qu'il s'agisse d'une extraction automatique, d'une inférence ou d'une contribution humaine.

Un atout majeur des graphes de connaissances réside dans leur capacité à intégrer et articuler des données hétérogènes et fragmentaires, grâce à la flexibilité du modèle RDF. Ce type de représentation permet d'unifier des sources variées, aux structures et granularités différentes, tout en préservant leur provenance, grâce à une approche d'intégration légère, par liage. C'est pourquoi, dans de nombreux domaines, des approches reposent sur ce modèle pour organiser, relier et analyser des données complexes et incomplètes (Shi *et al.*, 2017; Wilcke *et al.*, 2017; Asprino *et al.*, 2023).

L'ensemble de ces arguments plaide en faveur de l'utilisation de graphes de connais-

2. https://www.w3.org/TR/rdf-schema/
3. https://www.w3.org/OWL/
4. https://www.w3.org/TR/skos-reference/
5. https://www.dbpedia.org/
6. https://yago-knowledge.org/

sances pour un référentiel géohistorique de suivi de l'évolution des entités géographiques.

3.4 Ontologies pour la représentation des évolutions des entités géographiques

Pour représenter de manière homogène et interopérable des données géohistoriques, hétérogènes et fragmentaires sur l'évolution des adresses, nous choisissons d'utiliser un modèle de graphes de connaissances. Ces derniers permettent non seulement de structurer l'information, mais également d'expliciter les changements et de suivre les dynamiques des entités au fil du temps. Dans cette section, nous présentons deux ontologies permettant de représenter ces deux derniers aspects : **TSN** avec son extension **TSN-Change** et **HHT** et son extension **HHT-Change**, qui intègrent la notion de changements dans la représentation des entités spatio-temporelles.

3.4.1 Ontologies TSN et TSN-Change

Bernard *et al.* (2017) introduisent deux ontologies complémentaires : TSN et TSN-Change. L'ontologie TSN a pour objectif de décrire n'importe quelle nomenclature territoriale statistique, tandis que TSN-Change vise à modéliser les changements qui surviennent dans une telle nomenclature au fil du temps.

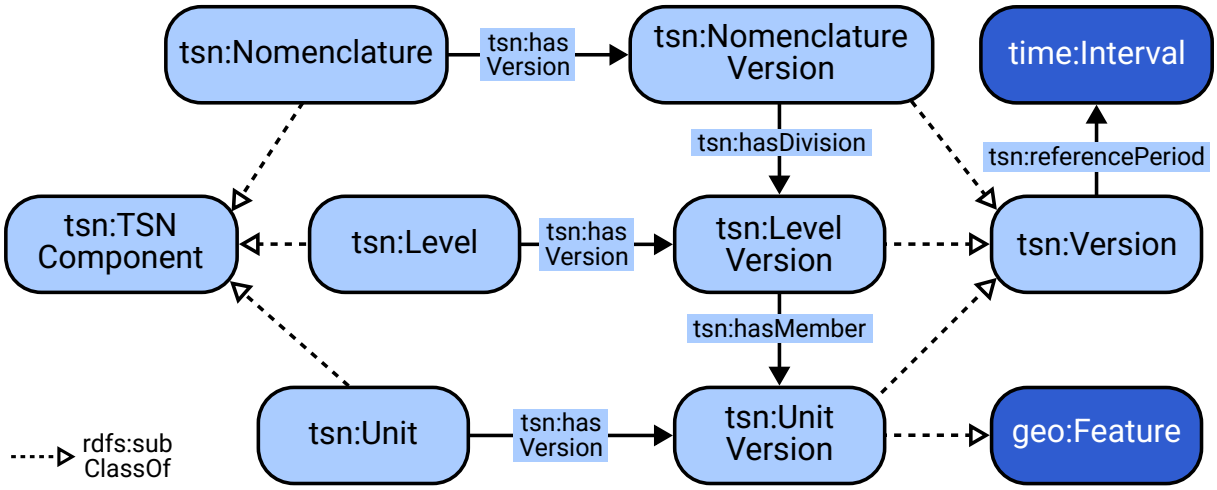


FIGURE 3.7 – Représentation simplifiée de l'ontologie TSN.

TSN L'ontologie TSN se distingue des approches antérieures associées à la représentation des nomenclatures, souvent limitées à un contexte particulier, comme l'ontologie géographique de l'INSEE pour le Code Officiel Géographique (COG) ou l'*Administrative Geography and Civil Voting Area Ontology* publiée par l'Ordnance Survey pour les territoires britanniques. Ces modèles spécialisés, bien que précis, manquent du niveau d'abstraction nécessaire pour représenter de manière générique tout type de nomenclature. TSN repose sur trois concepts principaux, tous sous-classes de la super-classe `TSNComponent` : `Nomenclature`, qui représente une nomenclature donnée (par exemple la NUTS ou l'ontologie géographique de l'INSEE) ; `Level`, qui décrit un niveau hiérarchique dans la nomenclature (comme le niveau communal, départemental ou régional) ; et `Unit`, qui désigne

une unité territoriale particulière (par exemple la France, Paris ou le 8e arrondissement de Paris).

L'aspect temporel est intégré grâce à l'approche *4D-fluent* combinée à OWL-Time. Celle-ci implique que les relations ne soient pas établies entre les entités elles-mêmes, mais entre leurs parties temporelles. Par exemple, l'appartenance de Paris au niveau communal en 2003 est représentée comme une relation entre `Paris 2003` et `niveau communal 2003`. Pour gérer cette temporalité, la classe `Version` a été introduite, avec ses sous-classes `NomenclatureVersion`, `LevelVersion` et `UnitVersion`, représentant respectivement les parties temporelles des concepts de base de l'ontologie. Ainsi, une instance de `NomenclatureVersion` correspond à une version spécifique d'une nomenclature, comme la publication 2024 de la NUTS ou le COG au 1er janvier 2025. Chaque version de nomenclature est associée à ses niveaux hiérarchiques (`LevelVersion`) via la propriété `tsn:hasLevel`, et ces niveaux sont à leur tour reliés aux versions d'unités territoriales (`UnitVersion`) par la propriété `tsn:hasUnit`.

Pour illustrer le fonctionnement de TSN, on peut imaginer l'adaptation du COG à cette ontologie, y compris pour des périodes anciennes, comme le milieu du XIXe siècle.

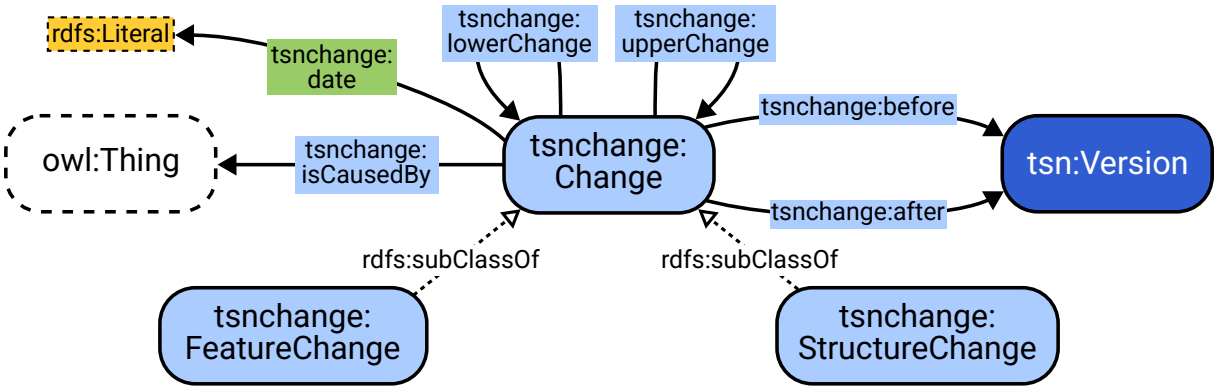


FIGURE 3.8 – Représentation simplifiée de l'ontologie TSN-Change.

TSN-Change Si l'ontologie TSN permet de décrire une nomenclature territoriale statistique, les données représentées à l'aide de ce vocabulaire ne représentent que des états spatio-temporels d'entités géographiques. Or, ces entités évoluent, et il est nécessaire de disposer d'un vocabulaire permettant de représenter explicitement les changements survenant entre deux versions successives d'une nomenclature. Dans leur analyse, Noy et Klein (2004) distinguent les changements du monde réel, qui affectent directement les entités décrites, des changements de conceptualisation, qui modifient la manière dont ces entités sont représentées dans les données.

L'ontologie TSN ne permet pas de décrire les changements entre différentes versions, ce qui a conduit à la création de TSN-Change afin de combler cette lacune. Deux approches principales sont envisageables pour représenter ces évolutions. La première consiste à fixer une version de référence, dite *version model* (Westfechtel et Conradi, 1998), et à stocker les différences (*deltas* observées avec les versions ultérieures. Cette méthode pose toutefois un problème pratique : pour reconstituer l'état d'une entité à une date donnée, il faut réappliquer, depuis la version initiale, l'ensemble des changements intervenus. Ainsi, si l'on décrit Paris à son origine et qu'en 1860 sa géométrie change, obtenir l'état de 1861 nécessite de recalculer toutes les modifications depuis la version initiale. La seconde approche consiste à conserver toutes les versions, chacune représentée par un graphe

RDF, et à décrire explicitement les changements entre deux versions consécutives. C'est le principe de la méthode *Materialization Using Versions and Deltas* proposée par Stefanidis *et al.* (2014).

Dans TSN, aucune relation n'existe par défaut entre les versions successives d'une même instance. Par exemple, si l'on sait que `COG_1859` et `COG_1860` sont deux versions d'une même nomenclature, rien n'indique laquelle est antérieure à l'autre. Cette absence de lien se retrouve aussi au niveau des versions d'unités territoriales, ce qui constitue un obstacle pour décrire les évolutions. TSN-Change introduit des relations spécifiques, comme `tsnchange:hasNextVersion` et son inverse `tsnchange:hasPreviousVersion`, afin de relier explicitement deux versions successives d'une même ressource et permettre ainsi la représentation des transitions temporelles.

L'ontologie permet de composer des changements simples pour former des changements complexes, ce qui facilite la représentation de scénarios d'évolution plus élaborés. En pratique, un changement est rarement isolé : il se produit souvent conjointement avec d'autres modifications, parfois à différents niveaux hiérarchiques. L'agrandissement de Paris en 1860 en est un exemple typique : il s'accompagne de la disparition ou de la réduction territoriale de plusieurs communes limitrophes, ce qui induit nécessairement plusieurs changements interdépendants au niveau des données. Ce type de situation justifie la création de concepts agrégeant plusieurs changements dépendants. TSN-Change prend également en compte cette hiérarchisation en représentant des changements multi-niveaux, reliés entre eux par les prédicats `lowerChange` et `upperChange`. Dans le cas de l'agrandissement de Paris, la modification au niveau communal entraîne aussi un remaniement des arrondissements, ce qui se traduit pas des changements multi-niveaux dans les données représentant cet événement.

Enfin, lorsque deux ressources successives (`<R1 hasNextVersion R2>`) présentent des différences, TSN-Change reprend le concept de Change Bridge de Kauppinen et Hyvönen (2004). Ce mécanisme, à la différence d'un simple lien direct entre les deux versions comme le proposent Harbelot *et al.* (2015), permet de mieux modéliser les changements complexes, de type *n-m* comme une réallocation, cas difficilement représentable avec une relation directe.

3.4.2 Ontologie HHT : Historical Hierarchical Territory

La deuxième ontologie présentée est HHT (*Historical Hierarchical Territory*). Elle a pour objectif de fournir un cadre formel pour la représentation des territoires dans leurs différentes dimensions, qu'elles soient administratives, judiciaires, religieuses ou fiscales, en y associant à la fois leurs propriétés et les relations qui les lient. Elle vise également à représenter les évolutions spatio-temporelles des unités territoriales, en permettant de retracer les transformations, continuités ou réorganisations survenues au fil du temps. Ce travail s'inscrit dans le cadre du projet ANR ObARDI consacré à l'étude des territoires sous l'Ancien Régime (Gay *et al.*, 2021).

L'approche HHT s'inspire des ontologies TSN et TSN-Change, tout en s'en distinguant sur un point essentiel : elle permet à chaque unité territoriale de disposer de sa propre temporalité d'évolution. Dans TSN et TSN-Change, la détection des changements repose sur la comparaison entre deux versions d'une nomenclature, ce qui implique que l'ensemble des changements inférés est considéré comme survenant simultanément. Les dynamiques temporelles dépendent ainsi des dates de publication des versions, un postulat qui s'avère inadapté pour décrire l'évolution des territoires historiques, d'autant que ces évolutions sont rapportées par des sources variées et hétérogènes.

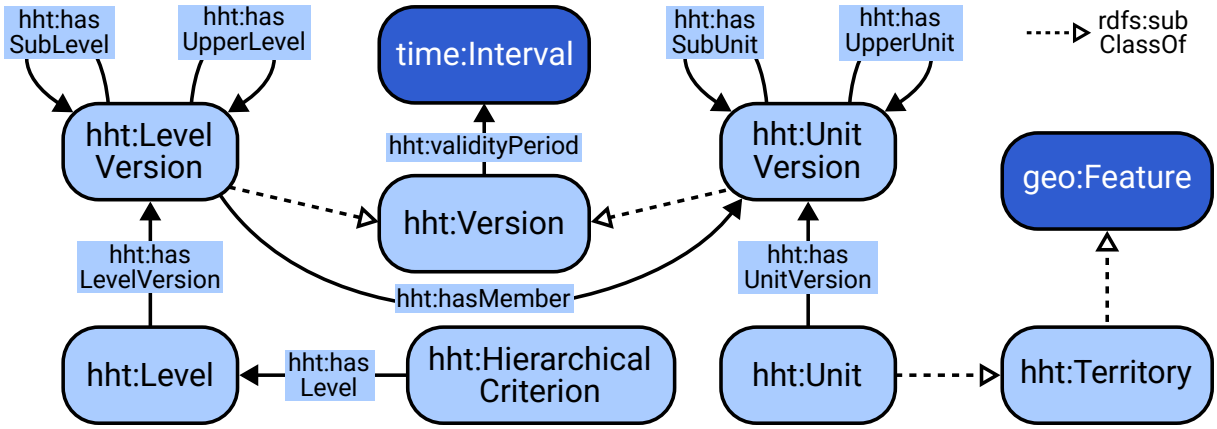


FIGURE 3.9 – Représentation simplifiée de l'ontologie *HHT*.

Dans HHT, chaque territoire est décrit par un ensemble de propriétés — nom, géométrie, type ou catégorie hiérarchique, relations hiérarchiques avec d'autres territoires — qui peuvent évoluer indépendamment. La question de l'identité territoriale est abordée à travers la notion de *trajectoire de vie*, définie comme l'ensemble des versions successives d'un territoire et de leurs propriétés. Aucune de ces propriétés n'est considérée comme essentielle à elle seule pour établir l'identité : une simple modification, comme un changement de nom, ne suffit pas à signifier qu'un territoire perd son identité. Par exemple, si l'on adoptait le nom comme critère unique, un changement d'appellation impliquerait conceptuellement la substitution d'un territoire par un autre, ce qui ne reflète pas la réalité historique. Dans le cadre de HHT, le critère d'identité retenu repose sur la combinaison d'un changement de nom et d'une modification d'au moins une autre propriété mentionnée.

Enfin, la représentation est contrainte par la nature des données historiques, qui offrent rarement une description précise des emprises spatiales des territoires, ce qui nécessite de gérer un certain degré d'incertitude dans leur localisation et leur forme.

Ontologie HHT À l'instar de TSN et TSN-Change, l'ontologie HHT représente à la fois les territoires et leur position dans une hiérarchie, ainsi que les évolutions qui les affectent, à travers deux modules principaux : `HHT` et `HHT-Change`, auxquels s'ajoute le module `HHT-Claim`.

La représentation des territoires repose sur la classe `hht:Unit`, définie par sa dimension spatiale et sa position au sein d'une hiérarchie. Cette classe possède une sous-classe, `hht:HistoricalTerritory`, qui décrit un territoire réclamé ou occupé par une personne, un groupe de personnes ou une institution définissant elle-même ses frontières. L'organisation hiérarchique est représentée par la classe `hht:Level`, représentant un niveau dans la hiérarchie, et par la classe `hht:HierarchicalCriterion`, qui précise la dimension dans laquelle s'inscrit la hiérarchie (administrative, judiciaire, religieuse ou fiscale).

Pour prendre en compte les évolutions temporelles, chaque concept de l'ontologie est associé à un concept décrivant ses versions, ces dernières étant modélisées comme des sous-classes de `hht:Version`. Une instance d'un concept est reliée à ses versions par la propriété `hht:hasVersion`, et chaque version est caractérisée par une période de validité, exprimée au moyen de la propriété `hht:validityPeriod`.

HHT-Change Les évolutions sont modélisées comme des changements conduisant à la création d'une nouvelle version d'un territoire. Trois types principaux sont distingués. La

classe `hht:FeatureChange` représente une modification simple d'une unité territoriale et relie deux de ses versions. La nature précise de ces modifications — qu'il s'agisse d'un changement de nom, de géométrie ou de relation avec une unité hiérarchiquement supérieure — est exprimée au moyen de ses sous-classes. La classe `hht:CompositeChange` permet de regrouper plusieurs changements simples de type `hht:FeatureChange` en une évolution plus complexe. Enfin, la classe `hht:UpdateKnowledge` ne décrit pas une évolution historique, mais correspond à une mise à jour des informations contenues dans le graphe de connaissances.

HHT-Change intègre également le concept de *Change-Bridge*, inspiré de TSN-Change, dans lequel un changement simple relie explicitement la version périmée et la version rendue effective. Cette dernière est par ailleurs incluse dans un `hht:CompositeChange`.

L'information temporelle n'est pas directement portée par les changements. La datation d'un changement se déduit de la fin de validité de la version qu'il rend caduque et/ou du début de validité de la version qu'il rend effective ; en théorie, ces dates devraient être similaires.

Le modèle initial présente toutefois une limite : le fait d'adopter l'approche *4D fluent* peut poser des difficultés ici, car elle suppose que les versions reliées partagent la même période de validité. Or, contrairement à TSN, les entités dans HHT possèdent chacune leur propre temporalité d'évolution. Cela complique la description des relations d'appartenance entre unités territoriales, car un changement survenant à un niveau hiérarchique peut entraîner la modification des versions d'unités situées aux niveaux inférieurs ou supérieurs.

Prenons l'exemple fictif suivant : on souhaite indiquer qu'un « 21e arrondissement » est situé à Paris entre 2005 et 2015. Cette opération requiert la création d'une version du 21e arrondissement et d'une version de Paris valables sur cette période. Mais comme les versions de Paris sont alors fragmentées par cet ajout, il faut aussi mettre à jour celles de tous les territoires en relation avec Paris durant cet intervalle. Il devient ainsi nécessaire de générer une version de la France afin de pouvoir affirmer que « Paris (2005–2015) est la capitale de France (2005–2015) ». Ce processus peut conduire à une sur-fragmentation de la trajectoire de vie des unités territoriales, chaque changement — même indépendant de l'unité considérée — induisant la création d'une nouvelle version dans le graphe.

Charles et Hernandez (2023) proposent donc de limiter l'impact de cette fragmentation en imposant des conditions plus strictes à la création de nouvelles versions d'une unité. Ainsi, il devient possible de décrire des relations entre unités territoriales par l'intermédiaire de leurs versions sans exiger que celles-ci possèdent nécessairement la même période de validité.

L'ontologie HHT a été conçue pour représenter des évolutions de territoires historiques, et nous avons souligné que l'un des problèmes majeurs résidait dans la faible disponibilité de géométries précises pour ces territoires. Pour contourner cette difficulté, HHT exploite la hiérarchisation des unités territoriales afin de déduire leurs emprises spatiales. Par exemple, pour déterminer l'emprise d'une région française, il suffit de calculer l'union des emprises des départements qui la composent ; de même, pour obtenir celle d'un département, on agrège les géométries des communes qui le constituent. Ce mécanisme permet de générer les géométries à partir de celles du niveau hiérarchique le plus bas. Dans l'organisation administrative française, ce niveau correspond aux communes, tandis que dans le cadre du projet ANR ObARDI, ce sont les paroisses qui sont utilisées comme niveau de référence. Les auteurs considèrent en effet que leurs emprises ont peu évolué durant la période étudiée (Bourel *et al.*, 2022). Dans l'ontologie, ce niveau paroissial est représenté par la classe `hht:ElementaryLevel`, sous-classe de `hht:Level`.

Conclusion

Ce chapitre a présenté les principaux modèles spatio-temporels proposés dans la littérature, ainsi que différentes approches de représentation et de suivi des changements affectant les entités géographiques. Parmi ces modèles, le *Three-Domain Model* apparaît comme le plus adapté pour la construction d'un référentiel géohistorique, dans la mesure où il permet de représenter de manière indépendante les trois dimensions fondamentales des entités : la géométrie, la temporalité et les attributs descriptifs.

Cependant, ce modèle comme les autres modèles présentés, ne fournit pas de mécanisme explicite pour représenter les changements entre états successifs, tels que les relations de filiation ou les transitions entre états successifs. L'approche *Change-Bridge* permet de formaliser des changements au niveau des entités et des attributs.

Pour combiner ces deux perspectives — représentation spatio-temporelle et explicitation des changements — il est nécessaire de choisir un modèle de données adapté. Compte tenu du caractère hétérogène et fragmentaire données issues des sources disponibles, un graphe de connaissances paraît le support le plus approprié, offrant à la fois souplesse de modélisation, une capacité d'intégration et des possibilités de raisonnements.

Dans ce cadre, les ontologies existantes, notamment HHT et TSN, fournissent des vocabulaires pour la représentation de l'évolution des entités géographiques, en intégrant notamment le principe du *Change-Bridge*. Le chapitre suivant s'appuie sur ces éléments pour définir une ontologie dédiée à la représentation des données géohistoriques, intégrant explicitement le concept de *Change Bridge* et s'inspirant des stratégies de modélisation mises en œuvre dans HHT et TSN. En outre, en raison de la fragmentarité et de l'hétérogénéité des données, l'approche adoptée privilégiera une modélisation fine des évolutions au niveau des propriétés plutôt qu'au niveau des entités elles-mêmes.

Chapitre 4

Une ontologie pour représenter les adresses et leurs évolutions

Une première étape pour construire un graphe de connaissances sur les adresses et leurs évolutions consiste à se doter d'une ontologie adaptée, fournissant le vocabulaire nécessaire à la représentation des données sur les états des adresses, les changements constatés entre ces états, les événements qui en sont la cause et les sources d'où proviennent ces données. L'objectif de ce chapitre est de présenter le développement de l'ontologie PeGazUs destinée à fournir ce vocabulaire.

Ce chapitre s'appuie en grande partie sur les travaux publiés par Bernard *et al.* (2024), dont il reprend et développe les principaux apports. Il est structuré comme suit : la section 4.1 présente les méthodes existantes de construction d'ontologies et justifie notre choix de la méthode SAMOD. La section 4.2 décrit ensuite le développement de l'ontologie en trois sous-parties : la représentation de l'adresse, la représentation des entités géographiques et de leur évolution (changements et événements), puis la prise en compte des sources.

Les ressources nécessaires pour reproduire notre travail autour de l'ontologie *PeGazUs* sont mises à disposition dans un dépôt GitHub accessible à l'adresse suivante : `https://github.com/charlybernard/pegazus-extension`. Ce dépôt contient :

- l'ontologie *PeGazUs* complète accompagnée d'une documentation ;
- les jeux de données utilisés pour tester et évaluer l'ontologie ;
- la chaîne de traitement automatique pour peupler l'ontologie à partir des jeux de données.

4.1 Méthodes de création d'ontologies

Comme le souligne Rawsthorne (2024), il existe un état de l'art assez important des méthodes de développement d'ontologies. Nous en donnons ici un résumé succinct, centré sur les éléments les plus pertinents pour notre problématique.

Nous cherchons à développer une ontologie de domaine dont le but est de représenter les adresses et leurs évolutions. Pour cela, différentes méthodes existent pour permettre le développement de l'ontologie. Notre choix de la méthode repose sur deux critères principaux. D'une part, notre problématique étant complexe, l'ontologie qui en résulte ne sera pas simple. Il est donc important de choisir une méthode permettant de la construire progressivement, par exemple en décomposant la problématique en plusieurs sous-problèmes plus faciles à traiter. D'autre part, nous voulons avoir des moyens pour évaluer l'ontologie afin de vérifier qu'elle réponde aux besoins définis.

En majorité, elles recommandent de commencer par une phase préparatoire. Celle-ci consiste en l'acquisition de connaissances et l'élaboration de spécifications ou définition des compétences via des questions. Nous avons examiné et comparé les méthodologies de création d'ontologies suivantes : SAMOD (Peroni, 2017), MOMo ((Shimizu *et al.*, 2023), OTK (Sure, 2003), HCOME (Kotis et Vouros, 2006), Ontology Development 101 (Noy et McGuinness, 2001) et DOGMA (Jarrar et Meersman, 2009). Les deux premières proposent de décomposer le domaine en sous-domaines afin de diminuer la complexité de l'analyse. Cette stratégie vise à produire autant de modules d'ontologie qu'il y a de sous-domaines et à les fusionner itérativement. OTK et SAMOD proposent en outre plusieurs méthodes d'évaluation de l'ontologie produite, là où la plupart des autres méthodologies se cantonnent à suggérer de procéder à une évaluation, sans plus de précisions sur la manière de le faire. Ainsi, en prenant en compte les deux critères pour le développement de notre ontologie, SAMOD paraît être la plus adaptée.

Nous avons donc choisi de suivre la méthodologie appelée *Simplified Agile Methodology for Ontology Development* (SAMOD) introduite par Peroni (2017). La modélisation de phénomènes géohistoriques est complexe et mobilise des thèmes différents de l'ingénierie des connaissances, la méthode SAMOD permet de les identifier et de les séparer pour faciliter la conception.

Cette méthode agile consiste, en effet, à séparer un problème de modélisation complexe en modelets qui sont des sous-problèmes, plus simples à traiter, valider et tester. La méthode conseille un travail commun entre des ontologues et des experts du domaine lié au sujet traité. Pour chaque modelet, il faut rédiger un argumentaire en langage naturel qui décrit le sous-problème à traiter. Cet argumentaire présente le contexte dans lequel s'inscrit le modelet et la raison de son existence. En parallèle, l'argumentaire introduisant des notions et termes importants en jeu, il doit être accompagné d'un glossaire. Ensuite, afin d'expliciter les objectifs auxquels doit répondre le modelet, un ensemble de questions informelles de compétence en langage naturel doit être fourni. Ces questions sont celles à laquelle la base de connaissances doit permettre de répondre. Dans le but d'illustrer ces dernières, un ensemble d'exemples de réponses attendues est associé à chaque question.

Une fois les modelets développés de manière indépendante, il est nécessaire de procéder à leur fusion afin d'obtenir un modèle cohérent et unifié. Par ailleurs, lors du développement des modelets, il convient de les lier à d'autres vocabulaires externes pour garantir l'interopérabilité et favoriser la réutilisation de connaissances existantes. Enfin, l'évaluation doit intervenir à plusieurs niveaux : outre le contrôle de chaque modelet en répondant à des questions de compétences, il est important de vérifier la cohérence globale à l'aide d'un raisonneur tel que Hermit, intégré dans le logiciel Protégé (Musen, 2015), et d'utiliser des outils de détection d'erreurs de modélisation comme OOPS ! (Poveda-Villalón *et al.*, 2014) afin d'identifier d'éventuelles incohérences ou omissions.

4.2 Développement de l'ontologie

L'ontologie **PeGazUs** (*PErpetual GAZetteer of approach-address UtteranceS*) constitue notre vocabulaire de base. Elle répond aux besoins identifiés pour constituer un référentiel géohistorique d'adresses à partir de sources multiples. En effet, il est essentiel de pouvoir représenter la diversité des adresses. Ces dernières étant des énoncés se référant à des entités géographiques, il convient de les représenter tout comme leur évolution dans le temps. Enfin, le peuplement du référentiel étant multisource, il est nécessaire pour assurer la traçabilité des informations en reliant chaque affirmation à sa source de provenance. Cette ontologie repose ainsi sur trois modelets principaux :

— **Adresses** : chaque instance de la classe `addr:Address` fait référence à un lieu connu ou non et décrit ce dernier par le biais d'une succession de relations entre entités géographiques ;
— **Entités géographiques et évolution spatio-temporelle** : fournit le vocabulaire nécessaire pour décrire une entité géographique, ses attributs, leurs différentes versions au cours du temps, les changements entre eux et les événements qui causent ces changements ;
— **Sources** : chaque affirmation dans le graphe doit pouvoir être reliée à sa source afin de permettre d'évaluer sa fiabilité, et de reconstituer l'historicité des connaissances intégrées dans le graphe.

4.2.1 Adresse

La modélisation d'une adresse s'appuie sur sa définition donnée dans la section 1.2 du chapitre 1. Il s'agit d'une désignation spatiale associée à l'énoncé structuré d'un cheminement à l'intérieur d'une hiérarchie spatiale, non ambigu au sein de cette hiérarchie, composé d'une suite ordonnée de repères spatiaux dont la conceptualisation et la désignation sont connues et partagées.

4.2.1.1 Schémas conceptuels et ontologies pour la représentation des adresses

Davis et Fonseca (2007) proposent un schéma générique des adresses pour les bases de données relationnelles. Ils découpent chaque adresse en composantes géographiques (ou repères) liées entre elles par des relations spatiales. Cependant, cette modélisation se cantonne aux adresses numérotées. De même, Cai *et al.* (2005) proposent de représenter les adresses via trois groupes d'éléments : des informations au niveau du bâtiment, l'élément de localisation au sein de la municipalité, la position au sein d'une région. La conséquence est que les entités géographiques référencées dans une adresse sont contraintes et il est impossible d'inclure explicitement des relations spatiales entre chacune. La modélisation n'est donc pas adaptée aux adresses descriptives (Cura *et al.*, 2018). Par exemple, pour une adresse descriptive dont le libellé est « *situé sur le blv. de Clichy, dans la partie comprise entre la place Blanche et la rue Fontaine* », la définition classique n'est pas compatible car aucune municipalité n'y est indiquée, il y a strictement plus d'un repère de type voie de communication et une relation spatiale est définie en langage naturel par le terme *entre*. Coetzee *et al.* (2011) proposent des bonnes pratiques pour définir des modélisations des adresses comprises par les humains et les machines. Toutefois, ces conseils se limitent là aussi aux adresses numérotées et ont vocation à créer des standards. Dans le domaine de la géomatique, les besoins de représentation sont limités aux adresses numérotées actuelles et ont vocation de se rapprocher des standards comme la norme ISO 19133 :2005 qui propose un schéma conceptuel d'adresse (International Organization for Standardization, 2005). Ainsi, la majorité des travaux sur cette thématique utilise ou développe des schémas proches des standards (Coetzee et Cooper, 2007; American Planning Association, 1950; Zandbergen, 2008; Borkar *et al.*, 2000).

La littérature s'accorde sur le fait qu'une adresse définit des références à des points de repère reliés entre eux par des relations spatiales. Les schémas proposés se concentrent davantage sur les types de repères à utiliser que sur les relations spatiales qui existent entre eux, ces dernières étant implicitement considérées comme des relations d'inclusion spatiale dans une hiérarchie de types de repères a priori connue. Si on reprend l'exemple de l'adresse numérotée suivante : « *73 avenue de Paris, Saint-Mandé* », il est compris de tous que la cible de l'adresse est située au numéro 73 appartenant à l'avenue de Paris, dans

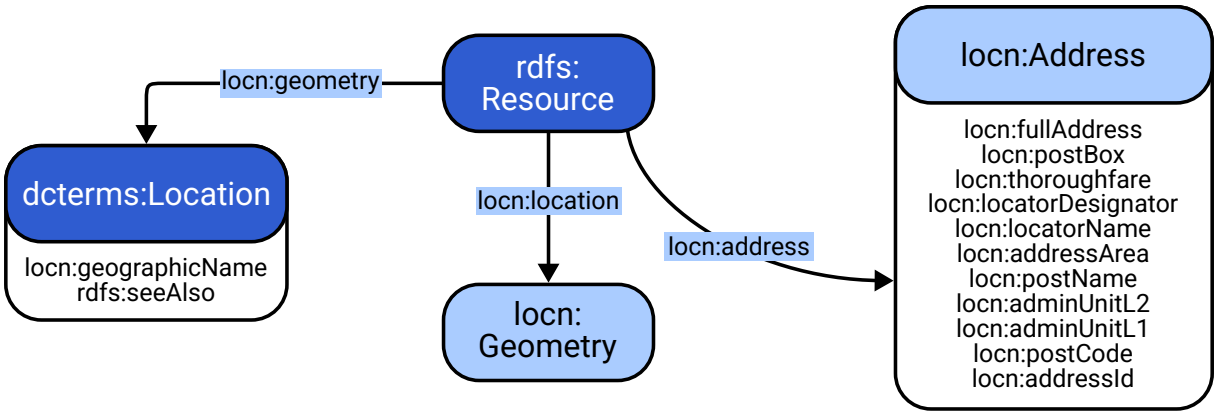


FIGURE 4.1 – Représentation simplifiée de l'ontologie *locn*.

la commune de Saint-Mandé. Or dans notre cas, les adresses anciennes peuvent utiliser des relations spatiales très hétérogènes. Prenons l'adresse descriptive suivante : « *Sous les colonnades du Théâtre Français, en face la petite rue du Rempart* ». Dans cet énoncé, l'adresse désigne un lieu non référencé explicitement et présente une structure différente des adresses numérotées actuelles, caractérisée par des relations diverses avec d'autres entités, comme *sous* et *en face*. En outre, il se peut que d'autres adresses aient des libellés avec des structures similaires aux adresses numérotées mais décrivant des entités assez différentes. Prenons l'exemple des adresses sectionnaires de Paris (voir section 2.6). Pour une adresse dont le libellé est « rue des Martyrs n°133 » (Léry, 1917), le numéro n'est pas associé à la rue mais à une section, ici non mentionnée. La cible de cette adresse serait un objet géographique situé le long de la rue des Martyrs et près du numéro 133, dans une section (ici non identifiée). Le numéro dépend donc de la section. Puisqu'une rue peut traverser plusieurs sections, il arrive, dans ce système de numérotage, qu'il y ait des adresses ayant exactement le même libellé mais ne décrivant pas la même localisation. À titre d'exemple, durant la Révolution française, il a existé à Paris cinq numéros 1 le long de la rue Saint-Denis (Waquet, 2018).

Dans le domaine des graphes de connaissances, une bonne pratique consiste à réutiliser, autant que faire se peut, les ontologies existantes. Il en existe qui permettent de représenter les adresses comme *locn*[1] (figure 4.1). Cependant, celle-ci a été conçue pour représenter des adresses numérotées modernes, en s'appuyant sur le schéma fédéré d'adresses publié dans le cadre de la directive européenne INSPIRE (Temporary MIWP 2021–2024 subgroup 2.3.1, 2024). En outre, une instance de la classe `locn:Address` ne permet pas de représenter l'énoncé de l'adresse sous forme structurée, mais seulement via une accumulation de `Literal`. Les entités géographiques mentionnées dans l'énoncé de l'adresse ne peuvent être décrites que par une chaîne de caractères et non par une ressource, ce qui ne permet pas de distinguer deux entités ayant le même label, et encore moins de gérer et représenter leur évolution temporelle. Ainsi, si on construit un référentiel d'adresses de toute la France, un simple label ne permettait pas de distinguer la *rue de l'Église* de Paris de celles des autres communes. Enfin, les types de repères y sont limités (rue, unité administrative...) et les relations spatiales ne sont pas indiquées, ce qui implique de les déduire avec toutes les difficultés que cela pose et que nous avons mentionnées ci-dessus.

En résumé, les schémas et ontologies existants permettent de représenter différents aspects des adresses, mais ils ne répondent pas complètement à nos besoins. Au vu de ces limites, il est nécessaire de proposer une ontologie qui épouse la définition de l'adresse que

1. http://www.w3.org/ns/locn

avons introduite dans la section 1.2. Cela implique, tout d'abord, de décrire explicitement les relations entre les ressources correspondant aux entités géographiques mentionnées dans une adresse, même lorsque ces relations sont seulement implicites. De plus, il est indispensable de conserver l'ordre dans lequel les relations apparaissent dans l'énoncé de l'adresse, car cet ordre porte une information structurante essentielle.

4.2.1.2 Représentation des relations entre entités géographiques

Les adresses étant une suite ordonnée de relations entre des points de repères, il est nécessaire de pouvoir décrire précisément ces relations. La question de la représentation des relations spatiales, notamment entre entités géographiques, a largement été traitée dans la littérature scientifique en géomatique. O'Keefe (1996) propose une approche générique, où les relations spatiales sont définies comme la composition de trois propriétés fondamentales : le lieu, la direction et la distance — indépendamment du type d'entités concernées. Pour décrire une relation entre deux entités A et B, le lieu définit la localisation de A, la direction et la distance sont respectivement la direction et la norme du vecteur $\overrightarrow{AB}$. Toutefois, la connaissance exacte des distances et des directions n'est jamais aisée, ce qui est particulièrement vrai dans les données historiques. Dans le cas des adresses, cette difficulté s'explique par le fait qu'une adresse ne vise pas à fournir la position géométrique précise des entités géographiques mentionnées dans son énoncé, ni leur distance les unes par rapport aux autres, mais plutôt à identifier et à relier ces entités dans un contexte donné. Autrement dit, ce n'est pas la position absolue d'une entité géographique qui importe dans une adresse, mais sa position relative à d'autres entités, telle qu'elle est exprimée dans l'énoncé de l'adresse. C'est la raison pour laquelle que O'Keefe (1996) suggère d'utiliser les prépositions spatiales pour décrire à la fois les notions de distance et de direction, et représenter des relations spatiales entre des entités. Une telle approche permet également d'envisager l'introduction d'un certain degré d'incertitude dans la description de la relation spatiale, sans nécessairement recourir à une formalisation mathématique stricte, à l'instar de l'approche qualitative proposée par Frank (1992). Ces prépositions sont de natures différentes : directionnelles (*au-dessous*, *en bas de*, *en haut de*, *sous*), omnidirectionnelles (*à*, *autour de*, *le long de*) et métriques (*près de*, *sur X kilomètres*). Quant à Touya *et al.* (2012), ils indiquent qu'il existe deux types de relations spatiales. D'une part, les relations quantitatives se caractérisent par des mesures avec des valeurs chiffrées. Par exemple, pour définir la proximité entre deux bâtiments, on la caractérise par leur distance exprimée en mètres et l'orientation donnée par un angle. D'autre part, les relations binaires associent des entités géographiques à un prédicat comme *est parallèle à* ou *est inclus dans*. En parallèle, ils proposent un schéma conceptuel de description des relations. Celles-ci sont définies par un prédicat qui établit les liens entre les entités géographiques concernées, dans un ordre explicite. Par exemple, la relation *A est dans B* illustre que les relations spatiales ne sont généralement pas symétriques, puisque *B est dans A* n'est pas équivalent à l'énoncé initial. Il est donc indispensable de préciser l'ordre des entités impliquées dans une relation spatiale afin d'en préserver la signification. Autrement dit, la structure de ces relations repose sur une asymétrie sémantique qu'il convient de représenter de manière explicite. Touya *et al.* (2012) utilisent les prédicats `member1` et `member2` pour donner un rôle à chaque entité dans la relation. En reprenant l'exemple cité précédemment (*A est dans B*), une telle relation serait décrite par les triplets suivants : `<rel member1 A>`, `<rel member2 B>` et `<rel isDescribedBy within>`. Tenbrink et Kuhn (2011) proposent un modèle similaire qui s'appuie sur les travaux de Levinson (1996) et utilisent les prédicats `locatum` et `relatum` pour distinguer

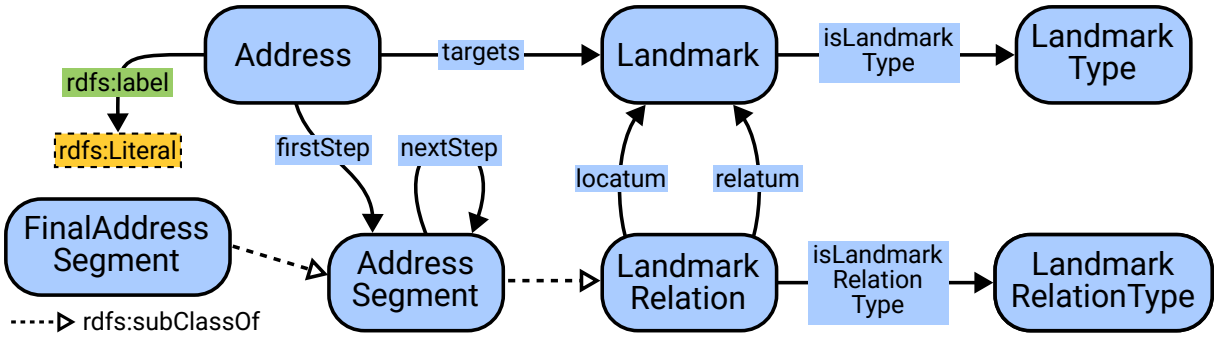


FIGURE 4.2 – Représentation simplifiée de l'ontologie des adresses.

les rôles des entités dans les relations. Dans ce modèle, `locatum` désigne l'entité localisée, c'est-à-dire celle dont on cherche à exprimer la position. Le `relatum`, quant à lui, joue le rôle de repère spatial : il sert de point de comparaison ou de référence pour situer le `locatum`.

Ces approches ne traitent que des relations binaires alors qu'il existe des relations n-aires comme celles dont le prédicat est *à l'intersection de* ou *entre*. Toutefois, elles ne sont pas incompatibles puisque rien n'interdit l'existence de plusieurs `relatum` dans une relation.

4.2.1.3 Proposition d'un modelet d'adresses

Notre ontologie doit permettre de représenter des adresses sous forme structurée. Ceci permet en effet d'interroger le graphe de connaissances à l'aide de requêtes pour connaître les entités géographiques impliquées dans une adresse et leurs relations. Ceci facilitera, en outre, la représentation de l'évolution de ces entités géographiques dans le modelet suivant.

Conformément à la méthode SAMOD, nous avons recensé les questions de compétences auxquelles l'ontologie doit permettre de répondre :

1. Quelles sont les adresses répertoriées le long d'une rue donnée ?
2. Quelles sont les coordonnées correspondant à la cible d'une adresse donnée ?
3. Quelles sont les adresses localisées dans une zone donnée ?

Pour le premier point, une question telle que « Quelles sont les adresses répertoriées rue d'Anjou-Saint-Honoré en 1864 ? » devrait permettre d'obtenir la liste des adresses distinctes (c'est-à-dire une combinaison unique de numéros, rue, et éventuellement section en cas de numérotation sectionnaire) recensées à cette date dans la rue concernée. Concernant le deuxième point, une question comme « Quelles sont les coordonnées de l'adresse « 2 Rue du Chat qui Pêche » en 1821 ? » devrait renvoyer les coordonnées correspondantes, par exemple : (48.853046, 2.34608), exprimées dans le système de référence de coordonnées WGS84. Enfin, pour le troisième point, une question du type « Quelles sont les adresses localisées dans la zone rectangulaire d'emprise définie par Xmin, Xmax, Ymin, Ymax ? » devrait retourner la liste des cibles d'adresses dont les coordonnées se situent à l'intérieur de l'emprise géographique spécifiée.

L'ontologie, présentée en figure 4.2, introduit trois concepts principaux : `Address`, `AddressSegment` et `Landmark`. Le concept d'`Address` (adresse) est l'élément central. Associée à un énoncé par une valeur littérale à l'aide de la propriété `rdfs:label`, une ressource de cette classe désigne un lieu géographique. Pour cela, elle cible un élément

`Landmark` qui est le lieu en question. Ce lien se fait par le biais du prédicat `targets`. La cible peut être connue ou non mais c'est cette ressource sur laquelle va se porter l'information géographique associée à l'adresse. Le concept de `Landmark` (repère) désigne une entité géographique, dont la nature est précisée par le type de repère (`LandmarkType`). Parmi les types existants, on peut citer les unités administratives, les voies, les numéros d'immeuble et les bâtiments. L'adresse est un énoncé composé d'une suite ordonnée de relations spatiales, ce qui se modélise ici par les éléments `AddressSegment`. Un `AddressSegment` (segment d'adresse) représente une portion du cheminement d'une adresse en reliant des repères à l'aide d'une relation. Afin d'expliciter cette suite, l'élément de classe `Address` est lié via la propriété `firstStep` (a pour première étape) au premier élément de type `AddressSegment` de la liste qui lui-même est lié via la propriété `nextStep` (a pour prochaine étape) à un élément de même type et ainsi de suite. Pour avoir un lien direct entre une adresse et l'ensemble de ses segments, le prédicat `hasStep` est introduit, avec les axiomes suivants : $\texttt{firstStep} \circ \texttt{nextStep} \sqsubseteq \texttt{hasStep}$ et $\texttt{hasStep} \circ \texttt{nextStep} \sqsubseteq \texttt{hasStep}$. Pour décrire la fin de cette suite, le dernier élément est de type `FinalAddressSegment` (segment d'adresse final), qui est une sous-classe de `AddressSegment`. Comme expliqué précédemment, un élément de type `AddressSegment` décrit une étape du cheminement via une relation entre repères. La classe `AddressSegment` est ainsi une sous-classe de `LandmarkRelation` qui décrit des relations entre des entités géographiques (cette classe est détaillée dans la sous-sous-section 4.2.2.2). Une ressource de cette classe lie un repère qu'on appelle *locatum* à un ou plusieurs repères qu'on nomme *relatum*. Pour indiquer la nature de la relation (*est située dans*, *longe*, *est à l'intersection de*), on relie la ressource à une autre de la classe `LandmarkRelationType` via le prédidat `isLandmarkRelationType`. Généralement, le locatum est la cible de l'adresse car chaque étape du cheminement a pour objectif de donner des précisions sur la localisation de la cible. Un contre-exemple fictif où il existe un segment d'adresse dont le locatum n'est pas la cible peut être une adresse dont le libellé serait « *à côté de l'école Jean Moulin qui fait face à la mairie* ». La cible est le locatum du segment `<cible à_côté_de école_Jean_Moulin>` mais pas celui liant la relation « face à » à la mairie puisque c'est l'école qui est concernée.

Pour illustrer la capacité de ce modèle à s'adapter à différents types d'adresses, nous présentons deux exemples, chacun étant représenté dans une figure : une adresse numérotée et une adresse descriptive. Dans le premier exemple illustré par la figure 4.3, l'adresse « *46 rue Gérard, 75013 Paris* » fait référence au numéro 46 constituant l'entité cible `RepA` qui est un numéro d'immeuble (`<RepA isLandmarkType ltype:HouseNumber>`). Les coordonnées associées à l'adresse correspondent donc à celles de cette entité.

De plus, l'adresse décrit un ensemble de relations avec les entités géographiques mentionnées. Bien que ces relations y soient implicites, notre modèle les explicite. La première, décrite par `SegAdr1`, indique que `RepA` appartient à la rue Gérard [2] définie par la voie de communication `RepB` (`<RepB isLandmarkType ltype:Thoroughfare>`). Elle est suivie par `SegAdr2`, décrivant une relation d'inclusion (« *dans* ») de la cible dans la zone associée au code postal *75013*, laquelle est décrite dans cet exemple par la ressource `RepC` (`<RepC isLandmarkType ltype:PostalCodeArea>`). Enfin, le segment final `SegAdr3` est similaire au segment précédent à la différence que le *relatum* est `RepD` qui fait référence à la ville de Paris.

Dans le second exemple décrit par figure 4.4, l'adresse « *au centre de la capitale* », « *entre le Palais-Royal et les Tuileries* » et « *à proximité des principaux théâtres* » est moins structurée, formulée de manière descriptive. Elle présente une entité cible non référencée et les relations y sont explicites mais restent floues. La cible est située au

2. On peut considérer de manière quasi-équivalente que `RepA` est situé le long de la rue Gérard.

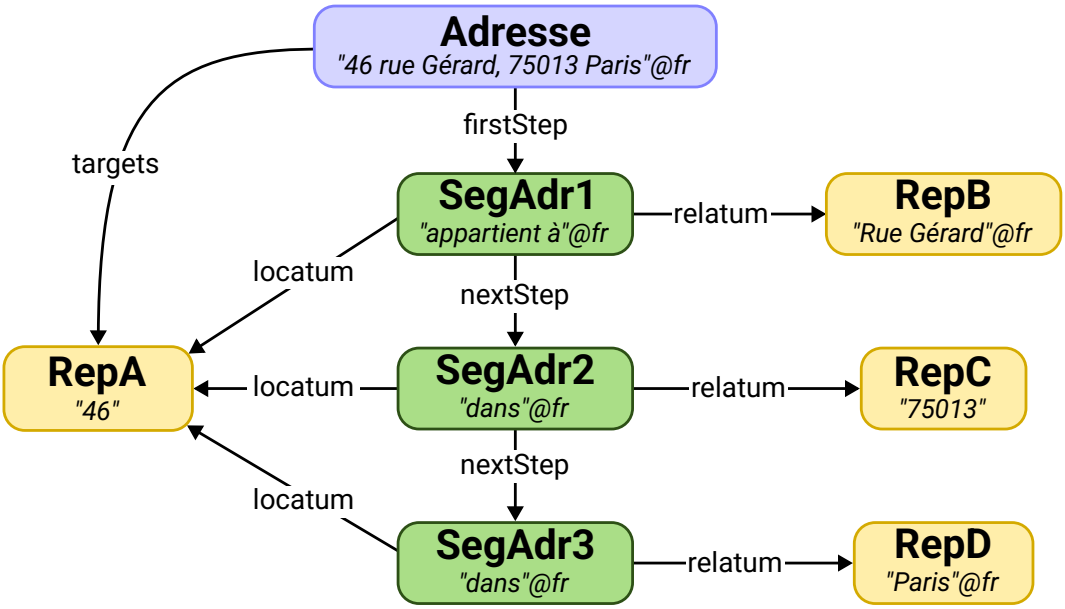


FIGURE 4.3 – Représentation pour l'adresse numérotée « *46 rue Gérard, 75013 Paris* ». Adresse, Entité géographique, Segment d'adresse (dont le segment final).

centre de la capitale, elle se trouve entre deux entités géographiques (le Palais Royal et les Tuileries), et enfin elle est à proximité d'un ensemble géographique plus large désigné par « *les principaux théâtres* ».

Ces exemples montrent que les entités géographiques et les relations spatiales constituent le cœur de la représentation des adresses. Cependant, comme les adresses et les entités géographiques évoluent dans le temps, il est nécessaire de représenter également ces dynamiques, ce qui fait l'objet de la section suivante.

4.2.2 Entité géographique et évolution

4.2.2.1 Objectifs

Nous présentons ici la manière dont nous représentons une entité géographique et son évolution à partir de données fragmentaires et hétérogènes. Notre approche doit autoriser des lacunes et ne pas imposer a priori des attributs fixes, de sorte que la représentation puisse être adaptée aux informations disponibles. Par exemple, lorsqu'aucune donnée sur la géométrie n'est disponible, celle-ci n'est tout simplement pas représentée. Cette flexibilité permet de tirer parti de toutes les informations accessibles tout en conservant la cohérence du modèle.

Par ailleurs, il est essentiel de pouvoir représenter les dynamiques des entités géographiques dans le temps. Pour cela, nous nous appuyons sur les modèles présentés au chapitre 3, en respectant les besoins liés à notre travail : il est nécessaire de séparer les dimensions spatiales, temporelles et thématiques, tout en permettant leur combinaison pour reconstruire les évolutions et expliciter les changements. Ainsi, notre approche rend possible la représentation fine des transformations géographiques, tout en restant compatible avec la diversité et l'hétérogénéité des sources.

Le modèle proposé est conçu pour satisfaire les questions de compétence suivantes :

— Quelles sont les voies existant à un instant donné ?

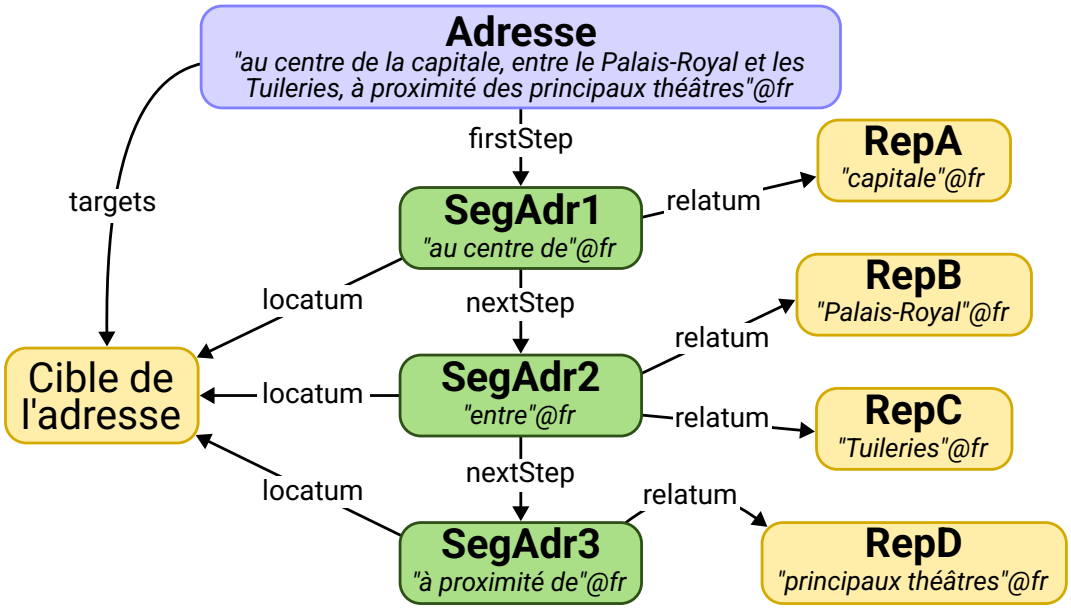


FIGURE 4.4 – Représentation d'une adresse descriptive. ■ Adresse, ■ Entité géographique, ■ Segment d'adresse (dont le segment final).

— Sur quel(s) intervalle(s) temporel(s) est valable une adresse ou une dénomination donnée ?
— Quel est l'historique d'un repère ? Autrement dit, quels sont les événements qui lui sont associés ?
— Quels sont les états et les événements qui manquent dans l'historique d'une adresse ?

4.2.2.2 Représentation des entités géographiques

Dans les bases de données relationnelles traditionnelles, une entité géographique est décrite par un ensemble fixe d'attributs, auquel on associe une - ou plusieurs - géométrie(s) (Güting, 1994; Devogele *et al.*, 1998). On définit ici une géométrie comme une description explicite de la dimension spatiale d'une entité géographique donnant une indication sur sa localisation, sa taille, sa forme (Collet, 2005). L'intégration de la dimension temporelle suppose de versionner simultanément l'ensemble des attributs et des géométries, ce qui réduit la granularité et la flexibilité du modèle : toute modification d'un attribut implique la création d'une nouvelle version complète de l'entité. Cette contrainte ne résulte pas du modèle relationnel en lui-même, qui permet de représenter des entités spatio-temporelles dont les propriétés peuvent être versionnées indépendamment, mais du modèle logique imposé par le standard Simple Feature Access (SFA) de l'OGC (Open Geospatial Consortium, 2011). Ce dernier définit une représentation des entités géographiques sous la forme de tables dont les colonnes correspondent aux attributs, conformément à la norme ISO 19109 (International Organization for Standardization, 2025).

Le modèle adopté par OpenStreetMap (OSM) illustre une approche plus adaptée à nos besoins. En effet, il permet la représentation d'entités très diverses — routes, bâtiments, unités administratives — de manière uniforme, en utilisant un système de `tags` (clé-valeur) pour décrire les attributs d'une entité, par exemple `<tag k="name" v="Rue Gérard"/>` (Minghini et Frassinelli, 2019). Cette approche est particulièrement flexible, car elle permet d'ajouter des attributs sans se limiter à un schéma fixe, mais elle devient

moins adaptée lorsque l'on souhaite représenter des attributs ayant plusieurs valeurs ou susceptibles d'évoluer dans le temps. Pour la dimension spatiale, OSM utilise une structure plus complexe : une entité peut être un `node` (point géographique), un `way` (chaîne ordonnée de nodes) ou une `relation` (collection de nodes et ways avec rôles spécifiques).

Dans la plupart des approches existantes, la temporalisation des entités géographiques repose sur leur versionnement global dans le temps. Ainsi, toute modification d'un attribut, y compris de la géométrie, conduit à la création d'une nouvelle version complète de l'entité. Cette conception découle directement des standards de modélisation géographique. Cette stratégie, bien que fonctionnelle dans de nombreux contextes, n'est pas adaptée à notre cas, car elle ne permet pas de suivre finement l'évolution individuelle de chaque attribut et complique le traitement des sources fragmentaires ou hétérogènes.

Pour le Web de données, GeoSPARQL propose un cadre pour représenter des entités géographiques et leurs relations. GeoSPARQL est un standard de l'OGC qui étend RDF et SPARQL afin de représenter et d'interroger des données géographiques (Car *et al.*, 2012). Il fournit une ontologie pour représenter les entités géographiques avec leur géométrie, des vocabulaires de relations topologiques comme RCC-8 (Randell *et al.*, 1992) et DE-9IM (Clementini *et al.*, 1993) ainsi que des formats de représentation des coordonnées (GML, WKT, GeoJSON, KML, DGGS). De plus, ce standard ajoute une extension SPARQL incluant des fonctions topologiques, telles que `geo:sfWithin` pour indiquer qu'une entité est située à l'intérieur d'une autre, ou `geo:sfIntersects` pour décrire qu'une entité intersecte une autre, ainsi que des règles d'inférence permettant le raisonnement spatial. Toutefois, ce standard présente certaines limites, notamment en ce qui concerne les relations : il ne gère pas directement les relations directionnelles ou métriques, telles que « A est au nord de B » ou « A est à 200 mètres de B », ni les relations sémantiques, comme « A est le chef-lieu de B ». En outre, pour pouvoir déduire les relations topologiques, il est nécessaire de disposer systématiquement des géométries des entités concernées. Enfin, comme il a été mentionné dans la sous-sous-section 4.2.1.2, nous traitons des relations plus complexes que binaires, que GeoSPARQL ne permet pas de représenter.

Les relations entre des entités couvertes par les ontologies TSN et HHT sont des relations hiérarchiques. Avec le vocabulaire proposé par HHT, il est possible de décrire les triplets suivants : `<ex:Paris hht:hasUpperUnit ex:IleDeFrance>` et `<ex:IleDeFrance hht:hasUpperUnit ex:France>`. TSN utilise un principe similaire mais avec des prédicats différents. Cette approche présente plusieurs limites. En premier lieu, les types de relations sont restreints du fait que seules des relations hiérarchiques binaires sont représentées. Elles ne permettent pas de modéliser explicitement d'autres types de relations, notamment celles impliquant strictement plus de deux entités (« A est situé entre B et C »). En outre, afin d'ajouter une dimension temporelle aux relations, HHT et TSN relient non pas directement les entités mais leurs versions valides sur une période donnée. Cette approche impose que les entités impliquées dans une relation partagent exactement la même période de validité. Dans notre contexte, cela tendrait à fragmenter excessivement les versions car chaque modification d'un attribut ou chaque évolution dans une relation nécessiterait la création d'une nouvelle version, ce qui complique la gestion et l'analyse de l'évolution des entités.

Ainsi, nous proposons de versionner les attributs et non pas les entités comme l'illustre la figure 4.5 et d'expliciter les relations entre entités, quelle que soit leur nature. `Landmark`, `LandmarkRelation`, `Attribute` et `AttributeVersion` sont les quatre concepts clés. Les deux premiers cités ont déjà été introduits dans la section précédente. Concernant la classe `Attribute`, elle permet de faire un pont entre les entités et leur dimension théma-

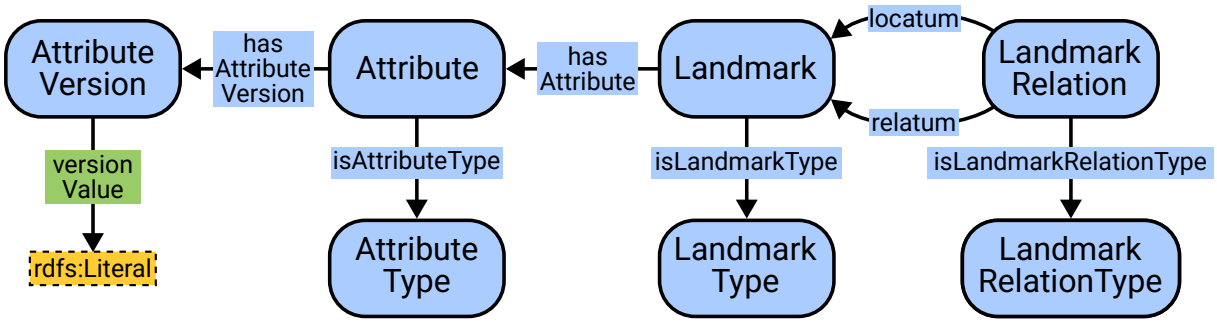


FIGURE 4.5 – Schéma de la partie de l'ontologie décrivant une entité géographique.

tique. Tout d'abord, cette classe décrit un attribut qui dépend d'une entité géographique et ce lien de dépendance est indiqué grâce à la propriété `hasAttribute` (sa propriété symétrique est `isAttributeOf`). De plus, il est associé à sa nature comme `atype:Name` ou `atype:Geometry` [3] qui décrivent respectivement la nom et la geométrie. La nature est définie par une ressource de classe `AttributeType` et reliée à l'attribut par la propriété `isAttributeType`. En parallèle, une instance de la classe `Attribute` agglomère l'ensemble des versions d'un attribut (`AttributeVersion`) d'une entité. Cette relation qui permet l'agglomération des versions d'attributs est décrite par le prédicat `hasAttributeVersion` (dont la symétrie est `isAttributeVersionOf`).

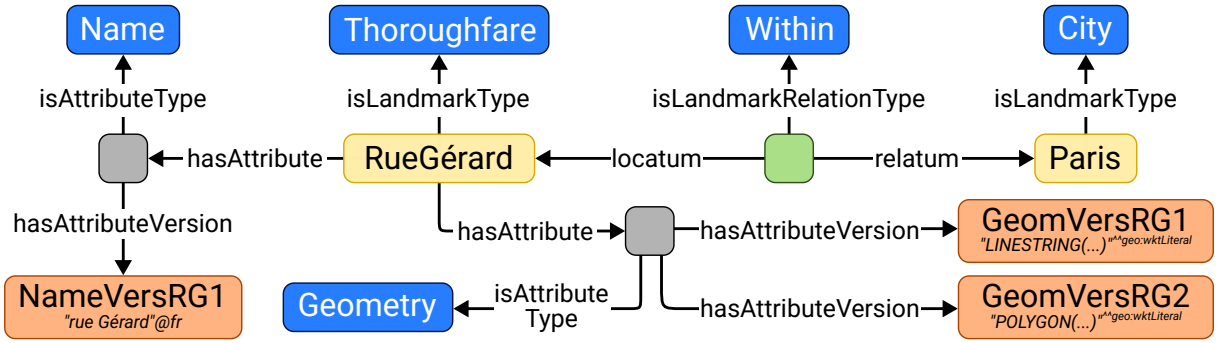


FIGURE 4.6 – Exemple de modélisation pour la *rue Gérard*. `Landmark`, `Attribute`, `AttributeVersion`, `LandmarkRelation`, `skos:Concept`.

Prenons l'exemple de la figure 4.6 qui montre la représentation de la rue Gérard selon l'ontologie. Cette dernière a eu plusieurs géométries et elles ne sont pas rattachées à l'entité *Rue Gérard* mais à son attribut dont le type est *Geometry*. De plus, l'ensemble des versions d'attribut décrivant le nom sont agglomérées via l'attribut de type *Name*. Les versions d'attributs sont des informations spécifiques à une entité et ne dépendent pas d'autres. Pour décrire de telles informations, il est possible faire appel à une ressource de classe `LandmarkRelation` qui, dans l'exemple, décrit une relation d'appartenance de la rue à Paris.

À ce stade, bien que l'ontologie sépare l'ensemble des dimensions d'une entité, rien n'est temporalisé. Il est impossible de savoir quand apparaît ou disparaît une entité, une relation et les versions d'attributs. Par exemple, si on reprend de la figure 4.6, il est impossible de savoir si les versions de géométrie sont contemporaines ou si l'une précède l'autre. C'est la raison pour laquelle on a besoin de s'intéresser à la représentation des dynamiques temporelles.

3. `atype` est le préfixe du `skos:ConceptScheme` qui définit les types des attributs.

4.2.2.3 Représentation des dynamiques temporelles

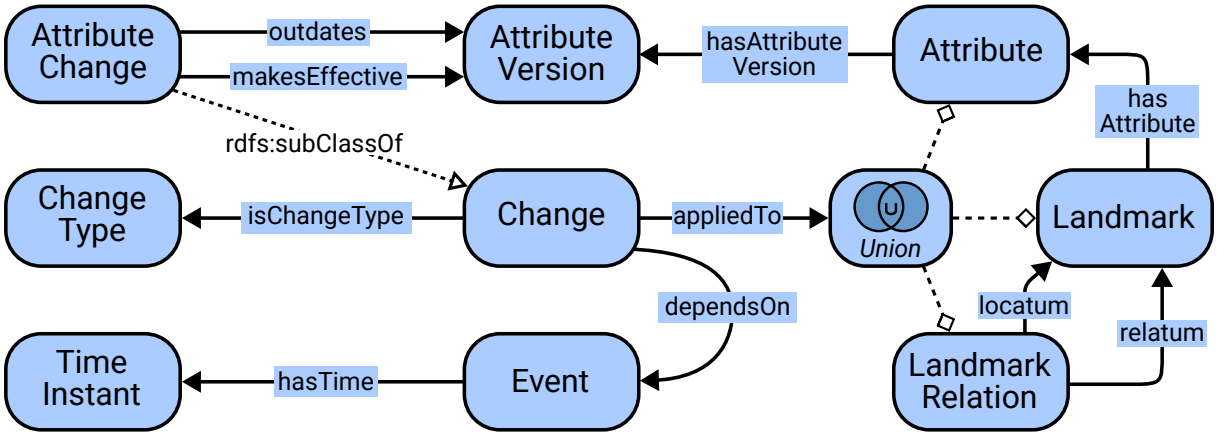


FIGURE 4.7 – Schéma de la partie de l'ontologie décrivant la dynamique des entités géographiques. L'élément lié à la classe `Change` par le prédicat `appliedTo` décrit l'union de `Attribute`, de `Landmark` et de `LandmarkRelation`.

L'état de l'art présenté dans la section 3.2 a montré la pertinence du concept de *Change-Bridge* pour représenter explicitement les changements. Celui-ci est notamment utilisé par Bernard *et al.* (2018) et Charles et Hernandez (2023) pour relier différentes versions d'entités. Dans notre cas, nous adaptons ce mécanisme de sorte à illustrer les dynamiques à différents niveaux : entités, attributs et relations entre entités. Par ailleurs, nous travaillons avec des données issues de sources historiques, dont les valeurs temporelles sont rarement parfaitement définies.

C'est la raison pour laquelle deux concepts sont introduits : `Event` et `Change`. Le premier décrit un événement, c'est-à-dire un phénomène qui a lieu dans le monde réel à un instant donné. Pour fixer une temporalité, une ressource de classe `Event` est associée à une ressource décrivant un instant, de classe `TimeInstant` via le prédicat `hasTime`. Dans certains cas, la date exacte de l'événement n'est pas connue, mais il est possible d'en estimer une borne temporelle. L'ontologie prévoit alors l'utilisation des prédicats `hasTimeAfter` et `hasTimeBefore` pour indiquer respectivement une date au plus tôt ou au plus tard. Elle permet également d'exprimer des relations temporelles entre événements, notamment des relations d'antériorité (`eventAfter`), de postériorité (`eventBefore`) et de simultanéité (`eventSameTime`). Si aucune information temporelle n'est disponible, aucune relation temporelle n'est spécifiée pour l'événement.

Un événement implique une ou plusieurs évolutions dans les données du graphe. Ces opérations élémentaires sont des changements (`Change`) liés à l'événement qui en est la cause par le prédicat `dependsOn` tandis que `appliedTo` indique à quelle ressource s'applique ce changement. Lorsqu'un changement concerne un attribut d'une entité géographique, il est de classe `AttributeChange`. Cette classe permet non seulement de spécifier l'attribut concerné, mais aussi d'indiquer les versions impliquées dans l'évolution, via les propriétés `outdates` et `makesEffective`. La propriété `outdates` signale que la version mentionnée devient obsolète du fait du changement, tandis que `makesEffective` désigne la version rendue valide ou active par ce même changement. Enfin, la nature du changement doit être spécifiée pour rendre compte du type d'évolution appliquée à une ressource. Un événement peut être décrit sans qu'aucun changement ne lui soit associé. En revanche, tout changement doit dépendre d'un événement.

Chaque ressource de type `Change` est reliée à une ressource de classe `ChangeType` au moyen de la propriété `isChangeType`. La séparation des différentes dimensions d'une entité géographique nous permet de proposer une typologie simple des changements, tout en laissant la possibilité d'en définir de nouveaux. La classe `ChangeType` est définie comme une sous-classe de `skos:Concept` et comporte plusieurs individus représentant les différents types de changement. Nous avons ainsi défini sept types de changements, applicables respectivement aux attributs, aux entités et aux relations :

— **Attributs** : `AttributeVersionAppearance` décrit l'apparition d'une ou plusieurs versions d'un attribut ; `AttributeVersionDisappearance` correspond à la disparition d'une ou plusieurs versions ; `AttributeVersionTransition` définit la transition entre deux ou plusieurs versions.
— **Entités** : `LandmarkAppearance` et `LandmarkDisappearance` décrivent respectivement l'apparition et la disparition d'une entité.
— **Relations** : `LandmarkRelationAppearance` et `LandmarkRelationDisappearance` adoptent une typologie analogue pour caractériser l'apparition et la disparition de relations entre entités.

Cette typologie n'autorise pas explicitement les réincarnations d'entités et de relations, mais reste suffisamment générique pour permettre de formuler des requêtes expressives sur le graphe. Par exemple, il est possible d'identifier les changements de nom d'une entité L par la requête fournie par le Listing 4.1.

```
SELECT ?L ?A ?C
WHERE {
 ?L a :Landmark ; :hasAttribute ?A .
 ?A :isAttributeType atype:Name .
 ?C a :AttributeChange ; :appliedTo ?A .
}
```

Listing 4.1 – Requête SPARQL relevant les changements de nom.

Enfin, si l'on considère un événement E impliquant un ensemble de changements $\{C_1, C_2, C_3\}$, où C_1 et C_2 sont de type `LandmarkDisappearance` et décrivent respectivement la disparition des entités L_1 et L_2, et où C_3 est de type `LandmarkAppearance` et décrit l'apparition de l'entité L_3, on peut en déduire que E correspond à une fusion de L_1 et L_2 en L_3.

Avec cette modélisation, nous pouvons découper des modifications impliquant une ou plusieurs entités géographiques en changements élémentaires. Prenons l'exemple avec l'événement « création de la rue du Père Guérin sur la portion orientale de la rue Gérard à la date du 30 août 1978 ». Dans le graphe de connaissances, conformément à l'ontologie, l'événement implique plusieurs changements représentés sur la figure 4.8 :

— `AppearanceChangeRPG` est un changement de type `LandmarkAppearance` qui décrit l'apparition de la rue du Père Guérin (`RuePèreGuérin`) ;
— un second (`NameChangeRPG`) de type `AttributeVersionAppearance` appliqué à l'attribut de type nom de la rue du Père Guérin (`NameAttributeRPG`) décrivant l'apparition d'une version ;
— le troisième changement, noté `GeomChangeRPG`, est similaire au précédent à la différence où il s'applique à l'attribut de type géométrie de la rue ;
— le dernier est de type `AttributeVersionTransition` (`GeomChangeRG`) et s'applique à l'attribut géométrique de la rue Gérard (`NameAttributeRG`), cette transition

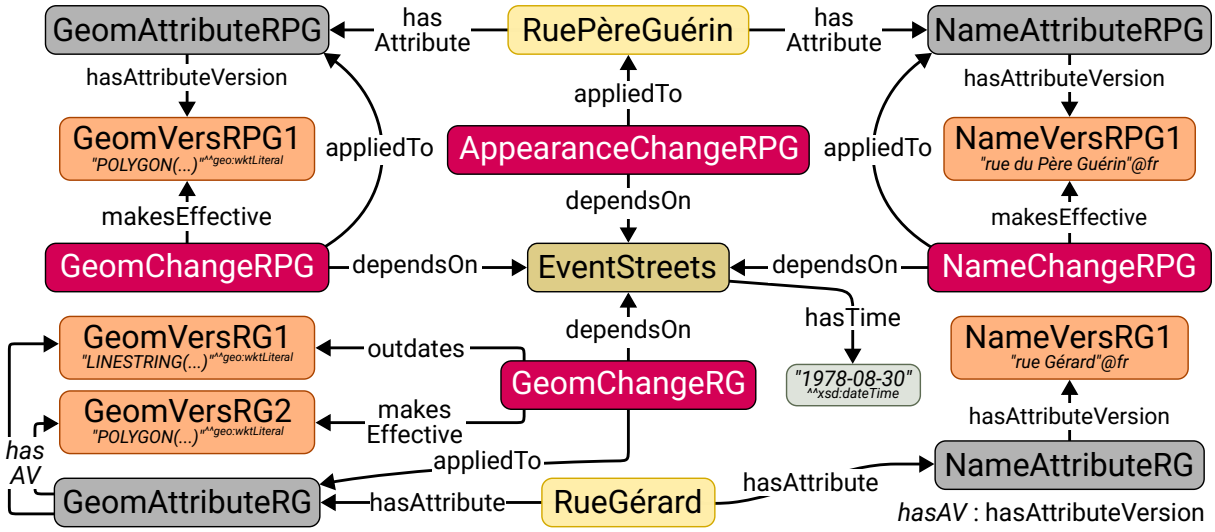


FIGURE 4.8 – Exemple de représentation de l'évolution conjointe de deux voies : création de la *rue du Père Guérin* par démembrement de la *rue Gérard*. `Landmark`, `Attribute`, `AttributeVersion`, `Change`, `Event`, `TimeInstant`.

indique que l'emprise spatiale de la voie allant de la rue Simonet à la rue Bobillot est remplacée à cette date par une nouvelle version.

4.2.2.4 Représenter l'information temporelle

Pour représenter des connaissances géohistoriques, un défi supplémentaire se pose : comment représenter les informations temporelles disponibles en tenant compte de leur diversité et de leurs imperfections ? En effet, il est essentiel de pouvoir décrire et manipuler à la fois des événements ponctuels et des périodes plus étendues de manière cohérente. De plus, certaines sources peuvent fournir des indications temporelles vagues.

Pour cela, nous utilisons deux grandeurs temporelles complémentaires : l'instant et l'intervalle. Un instant correspond à une durée très courte (tendant vers zéro) et peut être décrit par une valeur temporelle unique. Il permet de représenter des événements ou de situer précisément le moment où quelque chose se produit. Un intervalle, en revanche, correspond à une durée plus ou moins longue, définie (si possible) par deux instants indiquant son début et sa fin. Il permet de représenter un état ou une situation et de préciser la période durant laquelle cet état est valable. Cette distinction est essentielle pour traiter des données historiques, où l'information temporelle peut être incomplète, vague ou approximative, et où la précision relative des instants et des intervalles doit être explicitement prise en compte.

La manière dont le temps est représenté dans notre graphe de connaissances doit répondre à deux contraintes. D'une part, il faut qu'il soit possible de faire des raisonnements sur ces données. D'autre part, nous avons besoin de représenter des valeurs temporelles qui sont très diverses du fait de l'hétérogénéité des sources avec lesquelles nous travaillons : leur précision, leur calendrier et leur caractère plus ou moins flou peuvent varier. Certains instants peuvent avoir une précision annuelle, mensuelle ou quotidienne donc il faut pouvoir prendre en compte ce paramètre. En parallèle, le référentiel dans lequelle est décrite la valeur temporelle peut varier. En effet, bien que la majorité des informations temporelles contemporaines sont décrites dans le calendrier grégorien, il existe des dates, durant la Révolution française par exemple, qui sont situées dans un autre référentiel : le calendrier républicain (ou calendrier révolutionnaire français). Par exemple, pour faire référence à

l'instauration de ce calendrier qui a eu lieu le 6 octobre 1793 dans le calendrier grégorien, son équivalent dans le calendrier républicain est le 15 vendémiaire an II. Enfin, un autre paramètre à prendre en compte est la temporalité floue, qui ne peut être négligée, car certaines informations possèdent des dates imprécises, telles que *vers 1925*, *avant le 8 mars 1763* ou *entre 1895 et 1899*. De nombreux travaux (Barro *et al.*, 1994; Grieco *et al.*, 2003; Ohlbach, 2004) se sont penchés sur la représentation floue du temps et reposent sur le concept des sous-ensembles flous introduit par Zadeh (1965).

Avant de prendre en compte ces trois aspects, il convient de définir la manière dont est décrite la valeur temporelle, c'est-à-dire la manière dont on associe une valeur alphanumérique à un instant. La norme ISO 8601 (International Organization for Standardization, 2017) définit une représentation normalisée des instants qui prennent la forme *AAAA-MM-JJTHH :MM :SS,ss-/+FF :ff*, où *AAAA* désigne l'année sur quatre chiffres, *MM* le mois (de 01 à 12), *JJ* le jour du mois (de 01 à 31), la lettre *T* séparant la date de l'heure, *HH* les heures (de 00 à 23), *MM* les minutes, *SS* les secondes, *ss* la fraction décimale de seconde, le signe `+` ou `-` indiquant le décalage par rapport au temps universel coordonné (UTC), *FF* les heures de décalage et *ff* les minutes de décalage. Toutefois, cette norme est appliquée par le vocabulaire *Time Ontology in OWL* (plus connue sous le nom *OWL-Time*), une recommandation du W3C (Cox, 2017). Dans cette ontologie, la super-classe `TemporalEntity` désigne toute entité temporelle. Elle se décline en deux sous-classes principales. La première, `Instant`, représente un point unique dans le temps, sans durée discernable, où le début et la fin coïncident. La seconde, `Interval`, représente une période temporelle dotée d'une certaine étendue, avec un début et une fin distincts. Parmi les sous-classes d'`Interval`, `ProperInterval` correspond aux intervalles dont le début et la fin sont explicitement différents, et `DateTimeInterval` en est un cas particulier. En outre, ce vocabulaire permet de décrire la précision temporelle soit implicitement, via un littéral `xsd:dateTime` (par exemple avec une granularité à la seconde), soit explicitement à l'aide de `DateTimeDescription` (où l'on précise années, mois, jours, et type d'unité pour indiquer la précision souhaitée). De plus, l'ontologie autorise l'usage de différents systèmes de référence temporels (TRS) — Unix-time, temps géologique, calendriers lunisolaire ou dynastique, etc. — via les classes `TemporalPosition`, `TimePosition` ou `GeneralDateTimeDescription`, ainsi que l'usage d'`Era` pour les ères calendaires spécifiques (comme les ères japonaises ou les règnes historiques), afin de modéliser fidèlement le contexte temporel utilisé.

Enfin, pour décrire des relations topologiques entre instants et intervalles, ce vocabulaire utilise l'algèbre des intervalles d'Allen (Allen, 1990) illustrée par la figure 4.10. Néanmoins, la limite de OWL-Time concerne la représentation floue du temps. C'est la raison pour laquelle, il existe des travaux pour pallier cette limite. Métais *et al.* (2018) proposent une extension de l'algèbre des intervalles d'Allen intégrant des relations temporelles imprécises et personnalisables pour représenter à la fois des données temporelles précises et floues. Ils présentent également une ontologie, permettant un raisonnement flexible sur ces relations temporelles, illustrée par un prototype appliqué à la gestion des connaissances temporelles.

Dans un souci de simplicité de représentation des entités temporelles et d'intégration des caractéristiques du flou qui peuvent exister, nous proposons un vocabulaire proche de OWL-Time. La classe `TemporalEntity` incluent les sous-classes `TimeInstant` et `TimeInterval`. La classe `TimeInstant` est spécialisée en deux sous-classes. La première, `CrispTimeInstant`, est utilisée pour représenter un instant précisément daté à l'aide d'un littéral de type `xsd:dateTimeStamp` définissant l'horodatage, relié à la ressource par le prédicat `timeStamp`. Cet horodatage, à lui seul, n'indique pas le système de référence

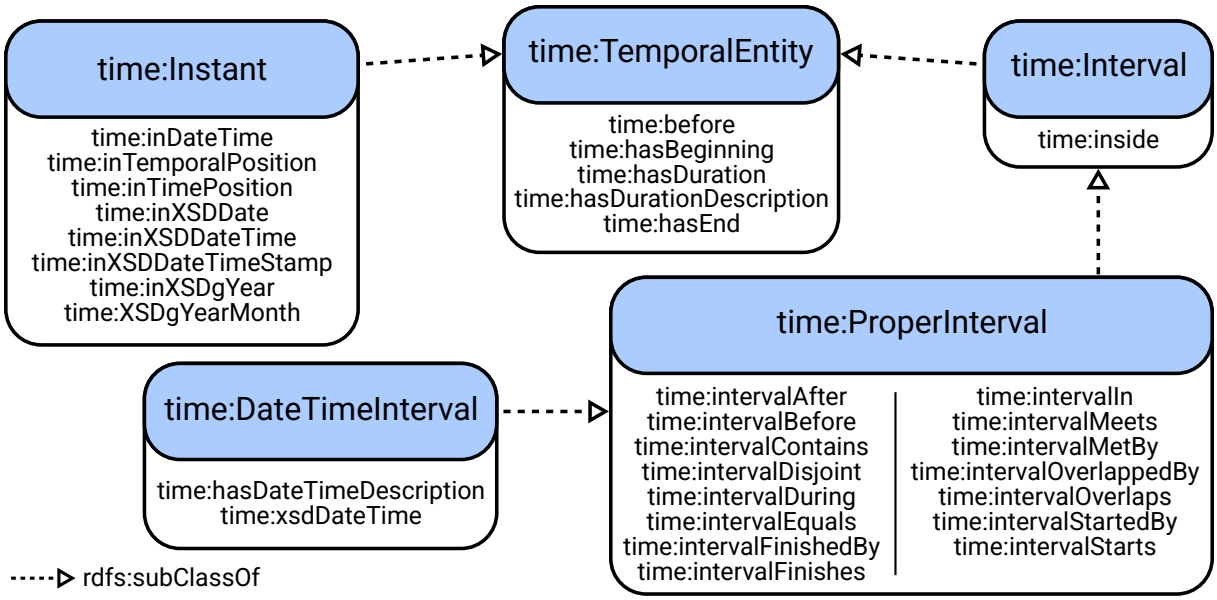


FIGURE 4.9 – Schéma du modèle central d'entités temporelles de l'ontologie OWL-Time.

temporel dans lequel il est valable. Pour lever cette ambiguïté, une ressource de cette classe est accompagnée du système de référence temporel, c'est-à-dire le calendrier utilisé (grégorien, julien, révolutionnaire), représenté par une instance de la classe `time:TRS` de l'ontologie OWL-Time et reliée à la ressource au moyen du prédicat `timeCalendar`. En outre, on lui ajoute une précision, soit une instance de la classe `time:TemporalUnit` (comme `time:Day` ou `time:Year`) relié via `timePrecision`. Si l'instant temporel est flou, `FuzzyTimeInstant` est employé à l'aide de bornes approximatives. Cette classe permet de spécifier des points comme le début (`hasFuzzyBeginning`), le pic (`hasFuzzyPeak`) et la fin (`hasFuzzyEnd`) d'une période estimée. Ces points sont des instants précis de classe `CrispTimeInstant`. En parallèle, lorsqu'il est nécessaire de représenter un intervalle temporel, une ressource de classe `TimeInterval` permet de le faire en indiquant ses bornes. Ces bornes sont des instances de `TimeInstant`, qui peuvent être soit précises, soit floues. Le début de l'intervalle est relié à la ressource via le prédicat `hasBeginning`, et la fin via `hasEnd`.

Pour illustrer l'utilisation des intervalles temporels flous, considérons l'intervalle défini par « *entre le 20 nivôse an IV (9 janvier 1796) et la fin du XIX*^{e} *siècle* » représenté par la figure 4.12. La borne de début est un instant précis : un `CrispTimeInstant` décrit dans le calendrier républicain avec une précision à la journée (`timePrecision`). La borne de fin est floue et est représentée par un `FuzzyTimeInstant`, défini à partir de trois points : le début (`hasFuzzyBeginning`) situé en 1870, le pic (`hasFuzzyPeak`) situé en 1885, et la fin (`hasFuzzyEnd`) située en 1900. Chacun de ces points est un `CrispTimeInstant` avec une précision annuelle et décrit dans le calendrier grégorien. Cet exemple montre comment combiner instants précis et flous pour représenter de manière flexible les périodes historiques dont les limites ne sont pas exactement connues.

En combinant les trois volets — la représentation des entités géographiques, celle des dynamiques temporelles et de l'information temporelle — il devient alors possible de représenter de manière flexible et cohérente les entités géographiques et leurs évolutions dans le temps. Un dernier point essentiel concerne la provenance des données. La construction du graphe repose sur un ensemble de sources historiques contenant des données formant un ensemble d'informations hétérogènes et fragmentaires. Il est donc important de pouvoir

Relation	Situation	Relation inverse
intervalBefore(X,Y)	X Y	intervalAfter(Y,X)
intervalMeets(X,Y)	X Y	intervalMetBy(Y,X)
intervalOverlaps(X,Y)	X Y	intervalOverlappedBy(Y,X)
intervalStarts(X,Y)	X Y	intervalStartedBy(Y,X)
intervalDuring(X,Y)	X Y	intervalContains(Y,X)
intervalFinishes(X,Y)	X Y	intervalFinishedBy(Y,X)
intervalEquals(X,Y)	X Y	intervalEquals(Y,X)

FIGURE 4.10 – Schéma représentant les 13 relations de l'algèbre des intervalles d'Allen avec le vocabulaire de OWL-Time.

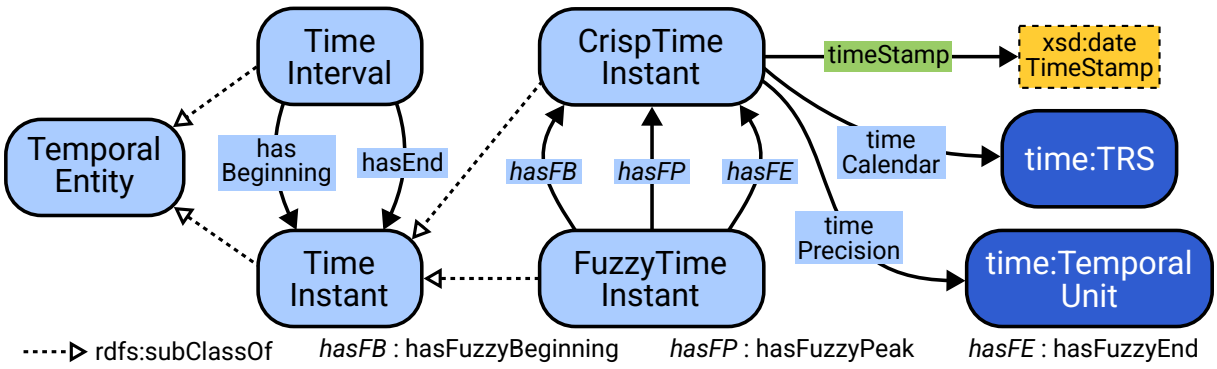


FIGURE 4.11 – Représentation de la partie de l'ontologie permettant de décrire des entités temporelles (`TemporalEntity`).

tracer l'origine de chaque information afin d'assurer la qualité et la fiabilité du graphe. C'est une condition essentielle pour garantir la confiance des utilisateurs - notamment des historiens - dans les données produites. C'est la raison pour laquelle le dernier modelet présenté traite spécifiquement de la représentation des sources et des liens de provenance des données.

4.2.3 Sources

Pour construire un référentiel géohistorique, les données intégrées proviennent de sources historiques et peuvent contenir des informations contradictoires ou erronées. C'est pourquoi il est important de pouvoir sourcer chaque affirmation dans le graphe de connaissances. Dans ce cadre, nous définissons des questions de compétence spécifiques afin de guider la conception du modèle de provenance. Les principales questions de compétence pour ce modèle sont :

1. de quelle source provient l'information `X` ?
2. quelle est la description de la source `Y` ?

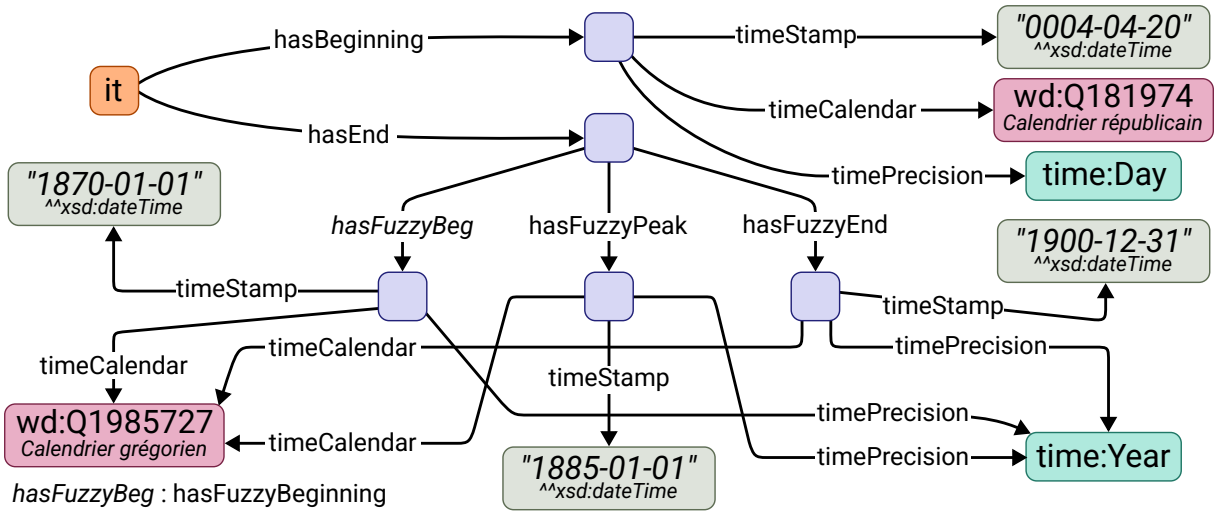


FIGURE 4.12 – Représentation de l'intervalle temporel `it` défini par *entre 20 nivôse an IV et la fin du XIXe siècle*. Nivôse est le quatrième mois du calendrier républicain. `TimeInterval`, `TimeInstant`, `xsd:dateTimeStamp`, `time:TRS`, `time:TemporalUnit`.

Les données incluses dans le graphe peuvent avoir deux origines : soit elles sont directement extraites de sources historiques, soit elles résultent de processus d'inférences. Il est donc essentiel de pouvoir indiquer la raison de l'existence de chaque donnée dans le graphe. Pour cela, nous utilisons l'ontologie *PROV-O*, qui fournit un modèle générique et standard pour décrire la provenance des données (Moreau et Groth, 2013). Dans ce modèle, une donnée, un document ou tout autre ressource numérique est représenté comme une instance de la classe `prov:Entity`, tandis qu'une action ou un processus qui produit ou utilise une entité est représenté comme une instance de la classe `prov:Activity`. La relation `prov:wasDerivedFrom` permet ainsi de lier une donnée à l'observation dont elle dérive. Cependant, l'ontologie reste volontairement générique et ne propose pas de modèle riche pour décrire les archives ou la gestion de fonds documentaires. C'est la raison pour laquelle nous complétons notre vocabulaire par l'utilisation de *RIC-O* (Records in Context Ontology), qui fournit un vocabulaire adapté à la représentation des archives et de leurs contextes institutionnels, juridiques et documentaires (Expert Group on Archival Description (EGAD), 2025). En particulier, le concept de `rico:Record` permet de décrire les documents archivistiques dans lesquels les entités du graphe sont attestées, et les relier à leur contexte par des relations telles que `rico:isOrWasDescribedBy`.

Reprenons l'exemple de la *rue Gérard* représentée dans notre graphe. Si son existence est attestée par le plan municipal de 1888, on la relie à la ressource décrivant ce plan, mais pas directement. En effet, l'entité géographique ne dérive pas de la source elle-même, mais d'un enregistrement résultant d'un ensemble d'opérations appliquées au document (numérisation, géoréférencement, vectorisation, etc.). Ainsi, on indique explicitement que la *rue Gérard* dérive de l'enregistrement qui la représente dans le plan municipal de 1888. On exprime ce lien par le triplet : `<rueGerard prov:wasDerivedFrom rueGerard1888>`, où `rueGerard1888` désigne l'enregistrement représentant la rue dans la source. Ainsi, pour $V = \{v_1, \ldots, v_n\}$ les n versions d'attributs de la rue Gérard dans le plan de 1888, il faut générer `<v_i prov:wasDerivedFrom rueGerard1888>` pour $i \in [\![1; n]\!]$. Par ailleurs, pour indiquer dans quelle source cette observation est attestée, on exprime la relation : `<rueGerard1888 rico:isOrWasDescribedBy plan1888>` où `plan1888` cor-

respond au document archivistique représentant le plan municipal. Cet exemple illustre la complémentarité entre *PROV-O*, qui relie les entités dérivées à leurs antécédents, et *RIC-O*, qui décrit précisément les sources archivistiques et leur contexte.

Conclusion

Dans ce chapitre, nous avons présenté le développement d'une ontologie destinée à modéliser les adresses descriptives et, plus largement, les entités géographiques et leurs dynamiques spatio-temporelles. Après avoir justifié le choix de la méthode SAMOD parmi les approches existantes, nous avons détaillé les différents volets de la création de notre ontologie : la structure des adresses, la représentation de l'évolution des entités dans le temps et l'espace, ainsi que l'intégration explicite des sources historiques. L'ontologie qui en résulte offre ainsi un cadre flexible pour représenter des données hétérogènes et fragmentaires, tout en garantissant la traçabilité de chaque information par ses liens de provenance.

Cette contribution constitue une étape essentielle vers la construction d'un référentiel géohistorique robuste, capable de prendre en compte la diversité des sources. Elle donne également les bases nécessaires à l'évaluation du modèle, telle que requise par la méthode SAMOD, qui implique de vérifier la capacité de l'ontologie à répondre aux questions de compétence définies lors de sa création.

Une fois l'ontologie développée, il faut maintenant définir une méthode pour le peuplement de l'ontologie à partir des données. Le chapitre suivant s'intéresse à l'intégration de données géohistoriques, en proposant un état de l'art des approches existantes pour alimenter le graphe de connaissances.

Chapitre 5

Intégration de données géohistoriques

L'ontologie pour représenter les adresses, les entités géographiques et leurs évolutions ainsi développée, nous avons ainsi un vocabulaire formel pour représenter les informations géohistoriques. Pour peupler cette ontologie et passer à un graphe de connaissances exploitable, il est nécessaire d'intégrer les données géohistoriques issues de sources variées et hétérogènes décrivant notre domaine d'intérêt. L'intégration de ces données constitue une étape importante pour construire un référentiel capable de représenter l'évolution des entités géographiques dans le temps. Les sources disponibles — qu'elles soient textuelles, iconographiques ou numériques — sont souvent fragmentaires, incomplètes ou imprécises. Ces caractéristiques imposent de mettre en œuvre des méthodes adaptées pour harmoniser, relier et enrichir les informations, tout en conservant leur traçabilité et leur provenance.

Ce chapitre propose un état de l'art sur l'intégration de données géohistoriques et s'organise en trois grandes parties. La section 5.1 expose les principes et méthodes généraux de l'intégration de données, depuis la transformation et la normalisation des informations jusqu'à leur appariement et leur fusion. La deuxième partie, présentée en section 5.2, traite de la notion d'identité d'une entité géographique, en abordant les enjeux, les approches existantes et le choix du critère d'identité retenu pour notre référentiel. Enfin, la section 5.3 présente des méthodes de construction d'un référentiel géohistorique, depuis la préparation des données jusqu'à l'alignement et la détection des changements.

5.1 Intégration de données

5.1.1 Cadre général de l'intégration de données

Comme indiqué en section 1.3, cette thèse s'inscrit dans un cadre plus large d'intégration de données, visant à rassembler, harmoniser, structurer et assurer la cohérence de données issues de sources diverses afin de les exploiter de manière unifiée. Lenzerini (2002) définit l'intégration comme étant une combinaison de données résidant dans différentes sources qui doit permettre en sortie une vue réconciliée des informations. Plus récemment, Doan *et al.* (2012) indiquent qu'un système d'intégration de données efficace doit être en mesure de fournir un accès uniforme à des sources multiples et variées, de s'adapter à l'augmentation du nombre de sources, d'harmoniser des données hétérogènes reposant sur des schémas différents. Dans le domaine de la géomatique, Mustière (2014) définit l'intégration des données géographiques comme l'ensemble des processus permet-

tant de traiter les similitudes et les différences entre plusieurs jeux de données hétérogènes décrivant un même espace, afin de les représenter dans un modèle commun.

Afin de formaliser ce processus et de disposer d'un point d'appui commun pour comparer et relier ces données, nous introduisons la notion d'*observation* — également appelée *enregistrement* ou *attestation* — qui représente une mention de l'existence d'une entité géographique dans une source donnée. Ainsi, dans l'exemple de la figure 1.1, l'adresse *10 rue Mâcon, Paris* fait l'objet de trois observations dans trois sources différentes.

Approches classiques d'intégration Plusieurs approches classiques existent pour traiter l'intégration de données hétérogènes. Les approches ETL (*Extract, Transform, Load*) consistent à extraire les données de chaque source, à les transformer afin de les uniformiser, puis à les insérer dans un référentiel commun. Elles sont particulièrement utilisées dans le contexte du *big data*, car elles se prêtent bien à des transformations complexes et à l'agrégation de grands volumes d'informations (Bansal et Kagemann, 2015; Mallek *et al.*, 2020). Toutefois, ces approches visent principalement à produire une base consolidée, et non une véritable vue réconciliée des données. Or, l'objectif de cette thèse est de conserver et d'exploiter l'ensemble des informations issues des différentes sources, de manière à pouvoir reconstituer l'évolution des entités géographiques. Une autre approche consiste à définir un référentiel commun dit *schéma global* vers lequel chaque source est mappée, c'est-à-dire qu'entre chaque source et le schéma global, il existe un ensemble de relations établissant des correspondances entre leurs éléments similaires (Lenzerini, 2002). Les sources contiennent alors les données brutes, tandis que le schéma global offre une vue intégrée et virtuelle de ces sources sous-jacentes. Cependant, cette approche ne propose pas une homogénéisation préalable des informations : chaque élément du schéma global reste lié à l'ensemble des observations issues des sources, ce qui rend complexe la reconstruction d'évolutions à partir de données hétérogènes. L'homogénéisation des observations demeure donc une étape nécessaire, qui ne peut être éludée si l'on souhaite aboutir à une représentation cohérente des entités géographiques et de leurs évolutions.

Vers l'utilisation de graphes de connaissances Dans la littérature récente, ces limites ont conduit à l'essor des approches fondées sur les graphes de connaissances. Ceux-ci permettent de représenter explicitement les entités, leurs attributs, leurs relations et leurs évolutions temporelles, tout en intégrant des données issues de sources variées et fragmentaires (Hogan *et al.*, 2021). Les graphes de connaissances constituent ainsi un cadre flexible et extensible pour l'intégration, car ils autorisent l'hétérogénéité des formats des données d'origine, facilitent l'alignement sémantique et offrent des mécanismes puissants de raisonnement et d'interopérabilité.

Vue d'ensemble de l'intégration Ainsi, le processus global d'intégration qui paraît pertinent pour notre cas est celui illustré par la figure 5.1. À partir d'un ensemble de sources $S = \{s_1, \ldots, s_n\}$, on tend à créer un référentiel global G. Pour cela, chaque source s_i est associée à un *mapping* (ou mise en correspondance) m_i permettant de convertir les observations de s_i vers une représentation conforme au modèle logique standard de G. Nous ne discutons pas ici du choix du modèle logique, qui a été détaillé dans le chapitre 3 avec l'introduction de l'ontologie PeGazUs. Enfin, l'étape finale consiste à relier chaque observation à l'entité correspondante dans G afin de construire la représentation et l'évolution de cette dernière.

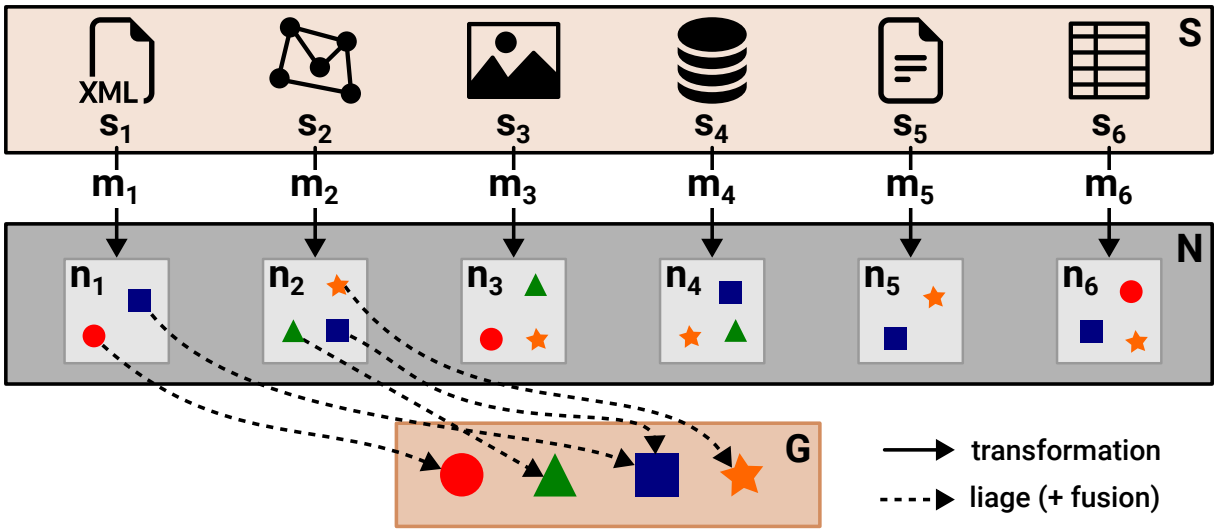


FIGURE 5.1 – Processus d'intégration de données provenant d'un ensemble de sources hétérogènes $S = \{s_1, \ldots, s_6\}$. Chaque source s_i est transformée en un référentiel normalisé n_i à l'aide d'un mapping m_i conforme à un vocabulaire commune, produisant l'ensemble $N = \{n_1, \ldots, n_6\}$. Ces derniers sont ensuite intégrés pour construire le référentiel global G. Par souci de lisibilité, seules les correspondances entre les éléments de n_1, n_2 et ceux de G sont représentées.

5.1.2 Transformation des données

La première étape du processus présenté en figure 5.1 consiste à transformer des données issues de sources hétérogènes. Cette transformation tend à une homogénéisation des données, c'est-à-dire qu'elle vise à convertir un ensemble de sources $S = \{s_1, \ldots, s_n\}$ dans un modèle commun, facilitant ainsi la mise en correspondance des différentes observations à partir d'un ensemble centralisé, le référentiel G. Dans la pratique, certaines sources fournissent des données non structurées ou semi-structurées comme des documents textuels (articles de journaux, actes administratifs) ou iconographiques (plans, photographies). Une étape préalable de reconnaissance et de structuration des données est alors nécessaire pour extraire et organiser les informations pertinentes, avant de pouvoir les normaliser et les intégrer dans le référentiel global.

La mise en cohérence des données hétérogènes peut s'envisager selon plusieurs dimensions. La première, dite *sémantique*, consiste à adapter les données à un modèle logique commun exploitable par les systèmes d'intégration, tout en les alignant sur un vocabulaire ou une ontologie de référence. Par exemple, une même entité géographique peut être décrite par plusieurs sources à l'aide d'attributs différents, dont certains renvoient à des concepts similaires ou proches. La mise en cohérence sémantique consiste alors à aligner ces attributs sur un vocabulaire ou un modèle commun. Ainsi, il est possible de disposer d'une représentation uniforme pour toutes les sources, facilitant le traitement et la comparaison des observations. L'un des autres objectifs de cette mise en cohérence vise à aligner, manuellement ou automatiquement, les concepts d'une source avec ceux d'un vocabulaire de référence ou d'une ontologie, c'est-à-dire à lier les schémas entre eux : chaque terme utilisé par les référentiels $n_1, \ldots, n_m$ est relié aux éléments correspondants du référentiel global G (Rahm et Bernstein, 2001; Pereira Nunes *et al.*, 2013). Dans le cadre de la création d'un référentiel d'entités géographiques, le concept *nom* pourrait être aligné avec l'attribut `name` d'OpenStreetMap et la propriété `rdfs:label` dans Wikidata.

Cette étape peut toutefois se heurter à des difficultés, car les concepts ne sont pas toujours équivalents et un concept d'une source peut correspondre à plusieurs, ou à aucun, concept du vocabulaire de référence.

Enfin, la dimension syntaxique consiste en premier lieu à nettoyer et harmoniser les données pour supprimer les variations superficielles. Cela inclut la gestion des espaces, de la casse, des signes diacritiques (dont les accents), ainsi que l'harmonisation des unités (par exemple, convertir en mètre des longueurs exprimées en kilomètre), des formats de dates (par exemple, au format ISO) ou de géométries (au format WKT). Par exemple, le fichier TOPO [1] de la DGFIP (Direction Générale des Finances Publiques), qui regroupe des entités topographiques françaises, utilise une syntaxe particulière avec uniquement des majuscules, sans accents, et des abréviations, comme `AV DES CHAMPS-ELYSEES`. La normalisation syntaxique permet ainsi de rendre ces informations facilement comparables et exploitables pour la mise en correspondance automatique et l'analyse. La langue constitue également un aspect de cette dimension syntaxique, dans la mesure où les jeux de données sont produits dans des langues différentes, bien que l'anglais y soit souvent dominant. La gestion du multilinguisme implique alors d'harmoniser les données en les exprimant dans une langue commune. Cela nécessite notamment de traduire certains termes (par exemple *Parigi* en italien pour désigner *Paris*, ou *Gérard Street* en anglais pour *Rue Gérard*). Néanmoins, la traduction seule peut être insuffisante car des termes distincts dans une langue peuvent être confondus dans une autre. Par exemple, les termes français *esplanade* et *promenade* peuvent tous deux être traduits par *promenade* en anglais, alors qu'ils renvoient à des réalités différentes. La normalisation linguistique nécessite donc également un travail de désambiguïsation des concepts.

Pour appliquer les opérations d'harmonisation précédemment citées, plusieurs méthodes existent. La première consiste à créer des fichiers de mapping qui décrivent les correspondances entre les éléments d'un modèle source et ceux du modèle cible, puis à utiliser un outil tiers pour automatiser la conversion. Par exemple, le logiciel Ontotext Refine [2] permet de transformer des fichiers TSV, CSV, XLS, JSON ou XML en un graphe de connaissances RDF en appliquant des opérations de conversions définies à l'aide d'un fichier de mapping à chaque élément des fichier sources. Ces fichiers de mapping peuvent être simples (tables CSV indiquant les correspondances entre sources et cibles), plus structurés (fichiers JSON ou YAML) ou basés sur des standards comme RML ou R2RML (Das *et al.*, 2012) pour définir des transformations vers RDF plus complexes. Alternativement, il est possible d'écrire directement un code pour effectuer la conversion, ce qui offre une plus grande flexibilité pour des transformations particulières ou complexes. Toutefois, cette méthode reste *ad hoc* et nécessite plus d'efforts de développement et de maintenance.

5.1.3 Mise en correspondance des données

La mise en correspondance — ou liage — de données consiste à identifier parmi l'ensemble des observations fournies par différentes sources celles qui se réfèrent à la même entité réelle. Les opérations de transformation présentées dans la section précédente peuvent simplifier ce processus. Dans le contexte des entités géographiques et des adresses, cela implique par exemple de reconnaître qu'une voie citée sous des noms différents dans plusieurs documents correspond à la même entité du réseau routier dans le monde réel.

1. https://data.economie.gouv.fr/explore/dataset/topo-fichier-des-entites-topographiques/
2. https://www.ontotext.com/products/ontotext-refine/

Dans la littérature, on retrouve généralement plusieurs étapes pour réaliser cette mise en correspondance automatique de données. Papadakis *et al.* (2021) proposent un panorama des approches de résolution d'entités, structurées en quatre générations successives selon leur évolution dans le temps. Les deux premières générations ciblent principalement des données relationnelles structurées et s'appuient sur des mises en correspondances de schémas (*schema matching*), l'expertise humaine et des connaissances externes pour maximiser la précision. Tandis que les approches de la première génération se concentrent sur la véracité des correspondances (en identifiant les doublons et les entités équivalentes à partir de schémas bien définis), la deuxième conserve ces principes mais introduit une dimension de passage à l'échelle, afin de gérer de très grands volumes de données. La troisième génération ne se limite plus aux données relationnelles structurées afin de prendre en compte la diversité croissante des formats et des sources de données modernes (web, semi-structurées). Les approches de cette génération traitent des données hétérogènes avec beaucoup de vocabulaire, des schémas multiples, des structures peu ou pas contraignantes. Enfin, la quatrième génération reprend les chaînes de traitement des générations précédentes en y intégrant un aspect incrémental, permettant d'améliorer itérativement la mise en correspondance grâce à des techniques telles que l'apprentissage profond ou la participation collaborative.

Bien que chaque génération ajoute une dimension supplémentaire (véracité, volume, variété, vélocité), la phase centrale de ces méthodes reste l'*Entity Matching*, où les observations sont comparées entre elles afin d'évaluer leur degré de similarité selon un ensemble de règles prédéfinies.

Saïs (2019) décrit trois grandes familles d'approches pour la mise en correspondance : numérique, logique et probabiliste. La première englobe celles fondées sur des règles de comparaison de propriétés. Cette approche repose sur l'hypothèse selon laquelle plus deux observations ont des valeurs similaires pour leurs propriétés équivalentes, plus elles sont susceptibles de représenter la même entité du monde réel. Ainsi, une approche naïve consisterait à s'appuyer sur l'égalité stricte des noms : deux observations sont considérées comme similaires si elles portent exactement le même label. Cependant, cette stratégie montre rapidement ses limites, car un même nom peut désigner des entités différentes, tandis que des désignations distinctes peuvent faire référence à une même entité. Par exemple, l'expression « *Le Louvre* » peut désigner, par métonymie, un quartier central parisien ou le musée situé dans ce secteur — que l'on peut également appeler « *musée du Louvre* » — ou le bâtiment historique qui l'abrite, le « *palais du Louvre* ». De plus, malgré certaines normalisations syntaxiques, des variations subsistent, telles que « *rue Mâcon* », « *rue de Mâcon* » ou encore « *rue Macon* ». Pour pallier ces limites, les comparaisons exactes sont remplacées par des mesures de similarité textuelle, telles que la distance de Levenshtein ou de Hamming, permettant d'associer un *score de similarité* à chaque paire d'observations. D'autres critères peuvent être pris en compte, comme la proximité géométrique entre observations (Walter et Fritsch, 1999; Samal *et al.*, 2004; Zhang *et al.*, 2005). Dans la pratique, il est rare de se limiter à une seule propriété : pour détecter toutes les observations décrivant la rue de Mâcon, il faut généralement combiner la similarité lexicale avec un critère de position géographique. Étant donné que la comparaison exhaustive présente une complexité quadratique ($O(n^2)$), elle est généralement précédée d'une étape de *blocking*, destinée à limiter le nombre de comparaisons en regroupant les observations potentiellement similaires pour les traiter prioritairement.

Parmi les principales techniques, on retrouve l'indexation par *n-grammes* et l'approche du *sorted neighborhood* (Christen, 2012). Les blocs peuvent par exemple être construits en fonction du type d'entité (voie, numéro d'immeuble, municipalité, etc.). Une deuxième

approche consiste à générer une *liste de candidats* ordonnée par score (*Top-k matching*), qui peut ensuite être validée manuellement (Christen, 2012). Une autre approche, plus automatisée, consiste à associer chaque entité à son *meilleur candidat* au-dessus d'un seuil de confiance (Nauman et Herschel, 2010; Getoor et Machanavajjhala, 2012). En géomatique, l'heuristique la plus pratiquée est la proximité spatiale (Feliachi *et al.*, 2017). On ne teste que les géométries situées dans une zone tampon de rayon fixé autour de l'observation à apparier. Cette méthode repose sur l'idée que deux observations d'entités géographiques localisées loin l'une de l'autre représentent peu probablement la même entité géographique du monde réel. Après la phase de comparaison, qui consiste à calculer des scores de similarité sur différentes propriétés, plusieurs stratégies de mise en correspondance sont possibles. Olteanu (2008) utilise la théorie des croyances afin d'apparier des entités entre sources grâce à un calcul de similarité fondé sur plusieurs critères (géométrie, topologie, attributs), permettant de sélectionner des candidats homologues pour chaque entité source. L'avantage de cette méthode est de permettre la représentation de données imparfaites et incomplètes, mais reste coûteuse car elle exige d'évaluer systématiquement l'ensemble des couples et de répéter le processus dans les deux sens. L'approche pose aussi la question de la compatibilité entre des sources qui ne décrivent pas les mêmes propriétés.

Dans la classification des approches de Saïs (2019), les approches logiques, s'appuient sur des mécanismes d'inférence liés au web sémantique. Enfin, les approches probabilistes sont fondées sur des techniques d'apprentissage. Celles-ci estiment la probabilité que deux entités représentent la même entité réelle, en s'appuyant sur des similarités à partir de modèles d'apprentissage.

5.1.4 Fusion des données

Les correspondances une fois identifiées, une étape ultérieure et optionnelle consiste à la *fusion* des entités similaires, afin de produire un profil consolidé ou d'alimenter un graphe de connaissances (Christen, 2012). Alors que la mise en correspondance vise à identifier les observations qui se réfèrent à la même entité réelle, la fusion consiste à construire une représentation unique et cohérente à partir de ces observations. Autrement dit, l'appariement produit des liens ou des regroupements d'observations, tandis que la fusion vise à générer une entité consolidée qui intègre, nettoie et harmonise les informations disponibles.

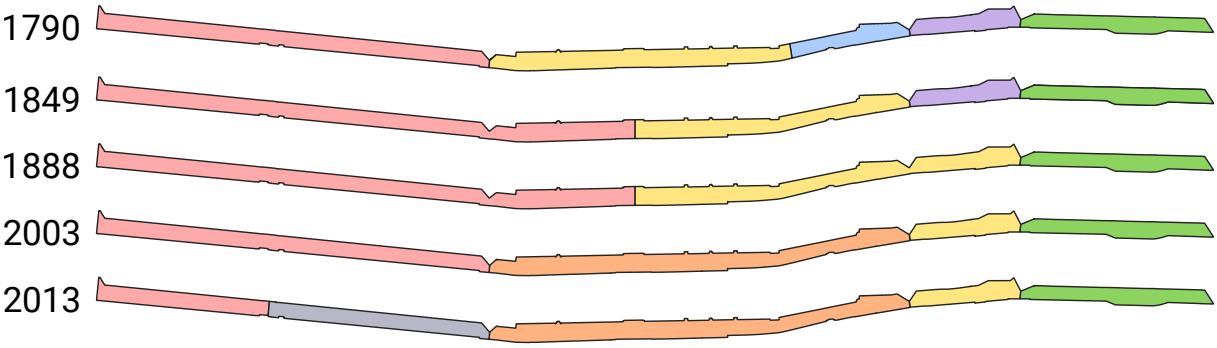
Prenons l'exemple de l'entité géographique correspondant au numéro *10* de la rue Mâcon à Paris. Deux observations issues de sources différentes peuvent être appariées. Une observation A où le numéro *10* localisé le long de la rue Mâcon dans une ancienne carte historique, fournit ainsi la position de l'entité à un instant donné. Une observation B où le numéro *10* est mentionné dans un annuaire historique, apporte des informations thématiques, telles que l'usage du bâtiment ou le propriétaire. L'appariement permet d'établir que ces deux observations décrivent la même entité géographique. La fusion consiste alors à produire une entité consolidée, intégrant les informations des deux sources. Dans le cadre de cette thèse, cette fusion est particulièrement utile pour reconstituer l'évolution historique des entités géographiques : en agrégeant les positions et les informations thématiques associées aux observations, il devient possible de retracer la trajectoire et l'évolution de l'entité *10* de la *rue Mâcon* au fil du temps.

5.2 Identité d'une entité géographique

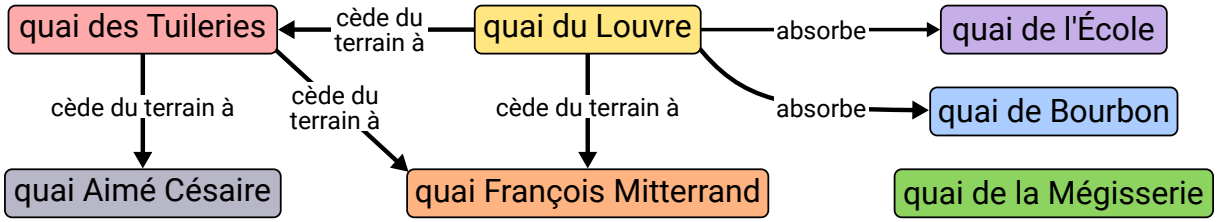
5.2.1 Enjeux de l'identité d'une entité géographique

Dans l'intégration de données, l'appariement permet d'identifier pour chaque entité réelle l'ensemble de ses observations dans les sources, c'est-à-dire de relier entre elles les observations qui décrivent la même entité. Ainsi, dans un contexte spatio-temporel, la question de l'identité et de la cohérence concernant la définition de ce qui constitue « la même entité » se pose, notamment afin de gérer les ambiguïtés et de préserver la cohérence de l'information lorsque des entités évoluent ou changent de nom au fil du temps. Ces aspects jouent un rôle central dans la thèse, car ils interviennent à la fois dans la définition de l'ontologie et dans son peuplement à partir de données hétérogènes et fragmentaires.

Les entités géographiques ne cessant d'évoluer, à partir de quand considère-t-on qu'une entité est créée ou a disparu ? Prenons l'exemple des quais situés au sud du Palais du Louvre au centre de Paris. La figure 5.2b met en évidence les relations d'absorption et de cession entre les quais. Comme le montre la figure 5.2a, ces derniers ont vu leur emprise changer au cours du temps, au point où l'emprise du quai du Louvre de 1790 n'a plus rien à voir avec celle de 2013. De plus, ce quai en 2013 possède la même emprise que le quai de l'École avait jusqu'en 1849. Par conséquent, on peut se demander si le quai du Louvre de 2013 est l'héritier du quai du Louvre de 1790 avec qui il partage le même nom ou bien du quai de l'École à la même année avec qui il partage la même géométrie. De même, le quai de Bourbon absorbé par le quai du Louvre au début du XIXe siècle possède un homonyme depuis 1814 sur l'île Saint-Louis et on peut aussi se demander s'il s'agit de la même entité dont la position a changé.



(a) Évolution des géométries des quais entre 1790 et 2013.



(b) Représentation en graphe des évolutions montrant quelles voies absorbent ou cèdent du terrain à d'autres voies.

FIGURE 5.2 – Évolution des emprises des quais situés au sud du Palais du Louvre à Paris de 1790 à 2013.

5.2.2 Approches existantes

Pour décrire l'identité, Hallot (2012) s'appuie sur la définition philosophique proposée par Audi (1999), selon laquelle l'identité est le concept qui relie une chose à elle-même. Comme le mentionne Drapeau Contim (2010), Gottfried Wilhelm Leibniz pose le principe d'identité des indiscernables qui stipule que deux éléments A et B sont considérés comme identiques si, et seulement si, pour toute propriété P, on observe l'égalité $P(A) = P(B)$. Appliquée à l'évolution temporelle, ce principe conduit à une contradiction : un élément A observé à deux instants différents, t_1 et t_2, devrait être considéré comme différent de lui-même, dans la mesure où il existe nécessairement une propriété P telle que $P(A_{t_1}) \neq P(A_{t_2})$. Cette approche soulève le célèbre paradoxe du bateau de Thésée. Un objet qui change progressivement toutes ses propriétés au fil du temps — jusqu'à ne plus posséder aucune de ses propriétés originelles — peut-il encore être considéré comme le même objet ? Ce paradoxe peut être illustré par l'exemple de la figure 5.2 où le quai du Louvre en 1790 ne partage plus aucune propriété (mis à part son nom) avec sa version de 2013.

Ce constat a donné lieu à plusieurs conceptions philosophiques de l'identité à travers le temps. Carosella *et al.* (2008) présentent trois approches. La première, dite *absolue*, postule que tout changement de propriété entraîne une rupture d'identité. Dans cette perspective, le quai des Tuileries en mai 2022 et cette même rue en juin 2022 sont deux entités distinctes, dès lors que leur emprise au sol a évolué. Une seconde approche repose sur la notion de *sortal*, c'est-à-dire une classe d'appartenance qui fournit un critère stable d'identification. Selon cette vision, deux entités peuvent coexister dans l'espace et le temps, mais ne peuvent être confondues que si elles relèvent du même sortal. Enfin, le *quadri-dimensionnalisme* propose de considérer les entités comme des entités étendus dans les quatre dimensions de l'espace-temps. L'identité d'un objet n'est alors pas réduite à un instant, mais résulte de l'intégralité de son « parcours spatio-temporel ».

Dans le domaine de la géomatique, cette problématique est traitée par le modèle *Identity-based change* (*changement basé sur l'identité*) proposé par Hornsby et Egenhofer (2000). Le principe repose sur une succession de phases, appelées états spatio-temporels, que traverse une entité au cours de sa vie. La transition entre deux phases successives d'une entité est définie par un changement d'état. Trois types d'états spatio-temporels sont distingués : l'existence, la non-existence simple (l'entité n'existe pas et n'a jamais existé), et la non-existence avec histoire (l'entité n'existe plus mais conserve une mémoire ou une trace). De ces trois états résultent neuf changements d'état possibles : la continuité de la non-existence sans histoire, la création, le souvenir, la destruction, la continuité de l'existence, l'élimination, l'oubli, la réincarnation et la continuité de la non-existence avec histoire. Chaque changement représente une transition précise dans la trajectoire de vie d'une entité.

Pour la suite, il est nécessaire de distinguer deux notions fondamentales : les *états* (spatio-temporels) et les *événements*. Un état correspond à une configuration particulière d'une entité géographique à un moment donné, définie par sa localisation, ses attributs et sa période de validité. Un événement correspond à une occurrence située à un instant donné, susceptible d'impliquer (ou non) un ou plusieurs changements d'état. Ainsi, l'inauguration officielle d'une rue constitue un événement, mais n'induit pas de changement d'état puisque l'entité préexistante (la rue) ne voit pas son statut modifié. En revanche, la fusion de deux rues en une nouvelle génère trois changements d'état : la destruction des deux rues initiales et la création de la rue issue de leur fusion. Un changement d'état témoigne donc directement de l'évolution d'une entité, tandis que l'événement joue le rôle de déclencheur ou de contexte dans lequel ces évolutions prennent place.

Le modèle proposé par Hallot (2012), fondé sur le concept de *Identity-based change*,

considère qu'une entité évolue à travers une série d'états successifs. Ces états permettent de structurer la dynamique d'existence de l'objet, y compris au-delà de sa présence physique. Dans le but d'être en cohérence avec la réalité physique, il met en place trois postulats : un retour en arrière est impossible (on ne peut accepter qu'une entité passe d'un état d'existence à un état de non existence sans histoire par exemple), aucun déplacement instantané n'est possible (une entité a une unique position spatiale à un instant donné) et une entité ne peut ressusciter.

De plus, ce modèle postule qu'un objet possède une identité dès lors qu'il est conçu, voire simplement imaginé, ou dès qu'il entre en relation, spatiale ou autre, avec un autre objet. Par exemple, une voie de communication n'est pas considérée comme créée uniquement lorsqu'elle existe physiquement sur le terrain, mais dès sa planification. Cette perspective soulève néanmoins plusieurs questions, notamment comment distinguer un projet avancé d'une simple idée de représentation, comme celle d'un bâtiment. Autrement dit, le simple fait de concevoir une entité, même sans réalité physique tangible, implique-t-il qu'elle doive être considérée comme existante ? Certaines entités, bien qu'elles aient disparu matériellement, continuent d'exister dans les mémoires collectives ou les représentations sociales, et restent mobilisées pour décrire ou structurer l'espace. C'est le cas, par exemple, de l'enceinte de Thiers ou de la commune de Belleville (absorbée par Paris en 1860), qui, malgré leur disparition institutionnelle ou physique, conservent une signification géographique, politique ou historique. Pour les entités qui n'existent plus, il est nécessaire de leur conserver une identité si l'on souhaite maintenir leurs relations avec d'autres entités. Une telle persistance symbolique suppose de distinguer clairement la phase d'existence d'un objet de sa phase de non-existence. Dans cette dernière, l'objet ne possède plus d'extension spatiale effective, mais il est souvent possible de lui en associer une, par exemple en considérant l'étendue qu'il occupait à la fin de sa phase d'existence. Ce principe s'applique aisément à des entités relativement stables, comme la Yougoslavie, mais devient plus complexe dans le cas d'objets mobiles ou éphémères tels qu'une grande roue, un marché ou un festival, dont la localisation varie dans le temps. La question du début et de la fin de l'identité d'un objet peut également être abordée à partir de ses premières et dernières mentions dans les sources disponibles. Ces repères documentaires permettent d'encadrer temporellement la trajectoire identitaire de l'entité, indépendamment de sa matérialité. Enfin, certaines entités peuvent être qualifiées de non-présentes : elles ne disposent pas de manifestation observable à un instant donné, sans que cela remette en cause leur existence. Par analogie, on peut penser au Soleil que l'on ne voit pas la nuit, ou à un marché forain temporairement absent, mais dont la réalité demeure attendue et reconnue.

Avec ces approches, on peut supposer que les entités ont des durées de vie infinies qui possèdent divers états d'existence encadrés par des états de non-existence.

5.2.3 Choix du critère d'identité

Dans le cadre de la reconstitution d'entités géographiques, plusieurs règles peuvent être définies pour déterminer si un changement observé constitue un changement d'identité. Une approche possible, dite *multi-attributaire*, combine plusieurs propriétés comme le nom et l'emprise géographique de l'entité. Charles et Hernandez (2023) considèrent que si une entité géographique voit son nom et une autre de ses caractéristiques évoluer simultanément, elle disparaît conceptuellement et est remplacée par une nouvelle entité. Bernard *et al.* (2017) utilisent cette approche multi-attributaire en associant un score de similarité entre deux états spatio-temporels successifs d'une potentielle même entité.

Ce score combinant plusieurs caractéristiques, telles que le nom, la géométrie ou l'identifiant, chaque critère étant pondéré en fonction de son importance. Ainsi, il est considéré que deux états décrivent une même entité si leur score dépasse un seuil prédéfini, reflétant ainsi la continuité de l'entité malgré des modifications partielles. Ainsi, si l'on prend l'exemple de la *place du Trône*, renommée *place de la Nation* le 2 juillet 1880, l'approche multi-attributaire considérerait que l'entité reste la même, car bien que le nom ait changé, d'autres propriétés (notamment sa localisation et sa géométrie) restent inchangées. Autrement dit, le passage de l'état « *place du Trône* » à l'état « *place de la Nation* » est interprété comme un simple événement affectant certains attributs de l'entité, mais ne constitue pas la création d'une nouvelle entité distincte. Cette logique illustre comment la combinaison de plusieurs critères permet de maintenir la continuité de l'identité malgré des modifications partielles.

Cette combinaison d'attributs montre ses limites lorsque les données disponibles sont hétérogènes et fragmentaires. En effet, les entités peuvent être décrites de manière incomplète selon les sources : certains attributs, comme la géométrie, ne sont parfois représentés que sur certains plans, tandis que d'autres informations peuvent manquer ou varier d'une source à l'autre. Pour cette raison, nous choisissons de retenir uniquement le nom et la place de l'entité dans une hiérarchie comme critère principal d'identité. Pour cette raison, nous choisissons de retenir uniquement le nom et la place de l'entité dans une hiérarchie comme critère principal d'identité. Autrement dit, deux entités possédant le même nom et partageant la même relation d'inclusion ou d'appartenance vis-à-vis d'une entité de niveau supérieur sont considérées comme identiques (par exemple, deux numéros appartenant à la même rue, ou deux rues situées dans le même arrondissement). Le nom étant systématiquement présent dans toutes les sources disponibles, il constitue un point de repère fiable. Ainsi, dès qu'une entité change de nom, nous considérons qu'elle disparaît et qu'elle est succédée par une nouvelle entité. De notre point de vue, *place de la Nation* remplace *place du Trône* qui garde la même emprise géographique. De même, les entités sont généralement décrites en relation avec une entité d'appartenance. Pour les numéros, ils sont systématiquement associés à une voie, une place ou un quartier (*46 rue Gérard* ou *2 place du Trône*). Pour des entités telles que les rues, quartiers ou communes, l'entité englobante peut être indiquée explicitement (*la rue Gérard à Paris* ou *le 13e arrondissement de Paris*) ou implicite et déduite du contexte de la source. Dans un dictionnaire recensant les voies de Paris, la mention *rue du Temple* renvoie implicitement à la voie située à Paris.

Cette approche soulève néanmoins des questions sur ce qu'il faut considérer comme un véritable changement de nom. Certaines variantes orthographiques ou évolutions lexicales ne constituent pas nécessairement un changement d'identité, par exemple *boulevart des Invalides* vs *boulevard des Invalides*, *quay des Thuilleries* vs *quai des Tuileries*, ou *rue des Bons Enfans* vs *rue des Bons Enfants*. D'autres cas sont plus ambigus, comme la transformation de *boulevard de Rochechouart* en *boulevard Marguerite de Rochechouart*. Bien que le nom ait été légèrement modifié, le boulevard continue de désigner le même référent, Marguerite de Rochechouart, ce qui maintiendrait l'identité de l'entité. L'interprétation peut toutefois varier selon que l'on privilégie la continuité sémantique du nom ou la stricte nomenclature officielle. Des cas plus nets incluent la *cité Bauer* devenue *cité Anne-Marie Bauer* (changement de référent) ou la *place André Trannoy* devenue *place André et Françoise Trannoy*. Des changements de désignation, tels que *rue du Gué*, *impasse du Gué* ou *passage du Gué*, sont également considérés comme des changements d'identité selon cette approche.

Une autre question autour du critère d'identité concerne l'évolution des relations spa-

tiales. En particulier, une voie peut changer de commune ou de périmètre territorial sans modifier son nom. Dans ce cas, considère-t-on qu'il s'agit de la même entité qui change uniquement de territoire ou bien qu'une nouvelle remplace la précédente ? Cette problématique est illustrée par lors des extensions territoriales de Paris (notamment en 1860 dans dans les années 1920). La rue Gérard en est un exemple puisque la voie de Gentilly intègre le territoire de la capitale en 1860.

5.3 Construction du référentiel

5.3.1 Préparation des données

Un état de l'art général concernant la transformation des données a déjà été présenté dans la sous-section 5.1.2. Nous détaillons ici plus spécifiquement les questions liées à la préparation (incluant l'harmonisation et la normalisation) des données géographiques et temporelles, en particulier dans le cadre de la représentation de l'évolution des territoires.Les sources disponibles pour décrire les territoires et leur évolution sont extrêmement variées, différant par leur granularité, leur précision spatiale et temporelle, ainsi que par la manière dont les attributs sont représentés. Il est donc nécessaire de représenter les données de manière homogène pour pouvoir les exploiter de façon cohérente, tout en conservant les spécificités de chaque source afin de ne pas perdre d'information.

En géomatique, les approches recommandent de choisir une ontologie pivot qui intègre les concepts nécessaires pour représenter l'ensemble des informations et les relier entre elles (Abadie, 2012; Euzenat et Shvaiko, 2013). Bernard *et al.* (2018) utilisent leur ontologie TSN comme pivot. Des chaînes de traitement permettent de construire autant de graphes qu'il existe de versions de nomenclature, conformément à l'ontologie TSN, en s'appuyant sur l'outil GeoTriples (Kyzirakos *et al.*, 2014). Cet outil convertit les données géographiques de leur format original vers du RDF et applique un mapping R2RML vers les concepts de l'ontologie, générant ainsi un graphe RDF des entités et de leurs attributs pour chaque version.

5.3.2 Alignement et détection des changements

Après avoir préparé et harmonisé les données, il est possible de procéder à l'alignement et à la détection des changements. Cette étape vise à établir des liens entre des éléments similaires appartenant à différentes versions et à identifier les évolutions qu'ils ont subies au fil du temps.

Avant d'effectuer l'alignement, il est nécessaire de définir un critère d'identité déterminant ce qui doit être considéré comme identique d'une source à l'autre. Selon les travaux, ce critère peut varier. Dumenieu (2015) et Costes (2016) privilégient la géométrie, ce qui conduit à des critères de similarité spatiale et mobilise des méthodes spécifiques d'alignement de données spatiales (Devogele, 1997; Olteanu, 2008; Li et Goodchild, 2011), incluant des mesures géométriques (distances, orientation des polylignes) et topologiques (position relative et voisinage). Les approches basées sur les ontologies HHT et TSN combinent plusieurs attributs, avec un poids fort accordé au nom, pour définir des critères de similarité hybrides mêlant caractéristiques sémantiques et spatiales.

Les méthodes de Dumenieu (2015) et Costes (2016) ne distinguent pas explicitement l'appariement de la détection des changements : les liens de continuation et de dérivation sont créés simultanément, et la détection des évolutions résulte de la combinaison de l'ensemble de ces liens.

Tual *et al.* (2023) proposent une méthode appliquée aux annuaires du commerce parisien du 19e siècle avec des graphes de connaissance. L'objectif est de suivre les transformations urbaines au travers des activités professionnelles des habitants, à partir d'entrées décrivant le nom, l'adresse et l'activité à différentes périodes. La détection d'entrées similaires repose sur la combinaison des propriétés suivantes : *Nom et Activité*, *Nom et Adresse*, *Activité et Adresse*. La distance token-wise la plus faible (équivalant au plus fort score de similarité) est retenue. Si cette distance est supérieure à un seuil, un lien de similarité (`owl:sameAs`) est généré. Ainsi, il devient possible de suivre les entrées au fil du temps, sans toutefois mettre en œuvre une détection explicite des changements.

Bernard *et al.* (2018), qui utilisent aussi les graphes de connaissance, génèrent un graphe décrivant à la fois les liens calculés entre des éléments similaires de différentes versions et les changements détectés, conformément au modèle TSN-Change. Le processus commence par l'intégration des données dans une base PostGIS, afin de pouvoir exploiter les informations spatiales et attributaires de manière homogène. La détection d'entités similaires repose sur une approche multicritères de l'identité : pour deux versions d'entités géographiques issues de deux nomenclatures, un score de similarité est calculé en combinant différentes caractéristiques, telles que le nom, la géométrie et l'identifiant, chaque critère étant pondéré différemment. Deux entités sont considérées comme identiques si ce score dépasse un seuil prédéfini. Les liens ainsi établis permettent ensuite de générer des Change-Bridge, à partir desquels les changements survenus entre les versions sont déduits automatiquement. Cette approche distingue explicitement la détection des changements de l'appariement des entités, en offrant une représentation détaillée des évolutions.

Charles et Hernandez (2023) proposent une méthode d'alignement incrémental des données, construisant progressivement un graphe. Cette approche s'appuie sur des critères hiérarchiques, déclinés en niveaux, pour intégrer les unités territoriales de manière incrémentale. Chaque entité est introduite par une version initiale valide pour la période étudiée ; lorsqu'une nouvelle version entre en contradiction avec des versions existantes, celles-ci sont fragmentées afin de reconstituer une trajectoire temporelle cohérente. Par exemple, si une version v_α couvrant [1870, 1890] doit être insérée dans une ligne temporelle comportant v_1, v_2, v_3 valides respectivement sur [1790, 1850], [1850, 1880] et [1880, 1900], alors v_α contredit v_2 et v_3, et la ligne de vie est réajustée en v_1, v_a, v_b, v_c, v_d avec des périodes de validité [1850, 1870], [1870, 1880], [1880, 1890], [1890, 1900]. Cette procédure garantit une cohérence temporelle tout en intégrant les ajouts successifs. Une fois que la génération de versions est constituée, la détection des changements peut être faite. Elle se fonde sur un algorithme basé sur des règles, combinant l'utilisation de règles SHACL et de requêtes SPARQL. Le processus se déroule en cinq étapes principales illustrées de manière condensée par la figure 5.3. Les versions d'une même unité géographique sont triées chronologiquement. Un lien `hasNextVersion` est ajouté entre deux versions successives v_1 et v_2, dont les périodes de validité sont respectivement $I_1 = [deb_1, fin_1]$ et $I_2 = [deb_2, fin_2]$. Pour chaque paire de versions successives, un changement de classe `hht:FeatureChange` est introduit (comme le montre la figure 5.3), selon la relation temporelle entre I_1 et I_2, selon l'algèbre d'Allen :

- si $\texttt{meets}(I_1, I_2)$ signifiant que $fin_1 = deb_2$, un changement c est créé tel que $\texttt{hht:before}(c, v_1)$ et $\texttt{hht:after}(c, v_2)$;
- si $fin_1 < deb_2$, l'unité a disparu puis réapparu. Deux changements c_1 et c_2 sont alors créés de classes respectives `hht:Disappearance` et `hht:Reappearance`, avec $\texttt{hht:before}(c_1, v_1)$ et $\texttt{hht:after}(c_2, v_2)$;
- si une version v_1 n'a pas de prédécesseur, i.e. $\nexists v_0, \texttt{hasNextVersion}(v_0, v_1)$, un changement de classe `hht:Appearance` est créé, tel que $\texttt{hht:after}(c, v_1)$;

— si une version v_1 n'a pas de successeur, i.e. $\nexists v_2, \texttt{hasNextVersion}(v_1, v_2)$, un changement de classe `hht:Disappearance` est créé, tel que $\texttt{hht:before}(c, v_1)$.

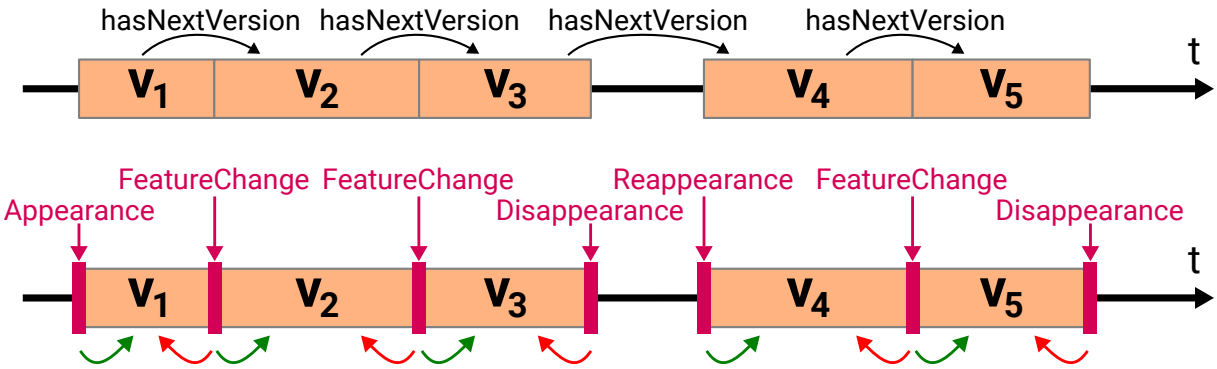


FIGURE 5.3 – Schéma indiquant la manière dont sont générés les changements simples à partir du tri temporel des versions pour la méthode de Charles et Hernandez (2023). ■ Version d'une unité territoriale, ■ Changement inféré, → *before*, → *after*.

Chaque changement est qualifié en comparant les versions associées. On identifie les propriétés ayant changé et la nature de l'évolution. Par exemple, si les géométries diffèrent entre deux versions, un changement de géométrie est détecté. Des comparaisons plus fines permettent d'identifier les pertes, gains ou réallocations de portions territoriales. Des changements composites (`hht:CompositeChange`) sont générés en regroupant plusieurs `FeatureChange`, notamment pour les modifications de géométrie :

— Les changements de géométrie non encore rattachés à un composite sont regroupés.
— Les changements déjà inclus dans un composite sont également regroupés.
— Chaque `FeatureChange` inclus dans un `CompositeChange` est lié à la même entité.

Malgré leurs apports, les méthodes présentent plusieurs limites importantes. Tout d'abord, les données d'entrée sont essentiellement structurées et souvent tabulaires, sous forme de shapefiles correspondant à des versions de nomenclature publiées à intervalles réguliers pour Bernard *et al.* (2018). Cette structuration facilite l'intégration dans l'ontologie, mais limite la prise en compte de données plus hétérogènes ou incomplètes. De plus, les ontologies TSN et HTT sont conçues principalement pour représenter les évolutions des territoires administratifs où la hiérarchie des unités est importante. Elles sont donc moins adaptées à la représentation d'autres types d'entités géographiques, comme les voies de communication ou les numéros d'immeuble où la dimension hiérarchique n'est pas incluse.

L'utilisation du concept de Change-Bridge pour comparer des états spatio-temporels d'une même entité comporte également des limites. En effet, cette approche n'est pas compatible avec l'intégration de données décrivant des évolutions : les changements sont considérés comme des inférences déduites des données existantes, mais ne constituent pas une source d'information à part entière. De plus, les données utilisées doivent être complètes. Le peuplement en déduit des évolutions à partir des données disponibles, or, en présence de données lacunaires, une absence d'information ne signifie pas nécessairement qu'un changement ou une entité est absent. Par conséquent, les états doivent être homogènes : si une version omet une propriété, un changement sera alors faussement déduit. Par exemple, une mention textuelle de Paris en 1906 sans sa géométrie ne signifie pas que cette dernière n'existe pas ou a changé. Par ailleurs, ces approches décrivent des états à intervalles réguliers et en déduisent des changements entre deux versions successives, mais ne permettent pas d'intégrer explicitement les changements eux-mêmes. Il est donc

impossible d'assigner une date précise à un changement. La taxonomie des changements proposée est assez dense et multi-niveau, ce qui la rend peu générique et contraint à mettre à jour l'ontologie pour intégrer de nouveaux types d'attributs.

Enfin, plusieurs changements peuvent s'appliquer simultanément à une même ressource, ce qui complique leur représentation. Pour la détection des entités similaires, l'approche multicritères combine différents critères pondérés. Cependant, la définition de ces poids reste délicate. Cette méthode est aussi limitée car elle suppose que chaque version possède tous les critères à comparer. Or, dans des documents textuels, une information décrivant une entité ne contient souvent pas la description complète de ses attributs, notamment sa géométrie.

Conclusion

Nous avons présenté dans ce chapitre un état de l'art au sujet de l'intégration des données en donnant des pistes pour intégrer des données géohistoriques hétérogènes dans un référentiel structuré. Nous avons d'abord rappelé le cadre général de l'intégration de données, en abordant les étapes successives de transformation, de mise en correspondance et de fusion des informations, indispensables pour produire un graphe de connaissances cohérent.

Par la suite, la question de l'identité des entités géographiques a été abordée, afin d'exposer les enjeux autour de cette notion, de montrer ce que la littérature affirme à son sujet puis de justifier le choix d'un critère d'identité adapté à notre référentiel.

Enfin, nous avons discuté des méthodes existantes à propos de la construction du référentiel géohistorique, de la préparation et l'harmonisation des données à l'alignement et la détection des changements. Ces méthodes permettent de reconstituer l'évolution des entités géographiques dans le temps mais ne sont pas parfaitement adaptés à notre besoin, ce qui justifie le développement d'une méthode de peuplement utilisant les données fragmentaires et hétérogènes dont nous disposons.

Chapitre 6

Méthode de construction du graphe

Ce chapitre présente la méthode que nous proposons pour peupler le graphe géo-historique d'adresses à partir de données hétérogènes et fragmentaires issues de sources multiples. Il s'agit de la méthode introduite dans Bernard *et al.* (2025a), qui constitue une évolution de notre première approche présentée dans Bernard *et al.* (2025c). Les deux méthodes reposent sur des principes communs, décrits dans les sections qui suivent. La différence intervient à la section 6.4 dans laquelle seule l'approche mise à jour sera abordée.

La construction du graphe repose sur une approche incrémentale et inférentielle, guidée par l'ontologie développée et validée conformément aux recommandations d'évaluation d'ontologies de la méthode SAMOD.

La section 6.1 introduit la démarche qui commence par la représentation des données sources selon l'ontologie, en adoptant une approche par *factoïdes* (Pasin et Bradley, 2015). Chaque information est contextualisée dans un graphe de factoïdes spécifique à sa source, contenant des versions d'entités (parcelles, adresses, voies, etc.) valides sur des intervalles temporels définis.

Ensuite, la section 6.2 traite de l'étape d'intégration qui consiste en la détection des ressources dans les graphes factoïdes que l'on considère représenter la même entité du monde réel, puis les lient au sein d'un graphe unifié appelé « graphe de faits ».

À partir de l'évolution des entités dans le temps, la méthode infère les changements et événements affectant ces repères (création, suppression, division, fusion, etc.) et les représente explicitement dans le graphe (voir section 6.3). Parallèlement, la section 6.4 traite de la partie de la méthode au sujet des attributs. Pour chaque attribut associé aux entités (nom, géométrie, etc.), son historique est reconstruit en agrégeant les versions issues des différentes sources, permettant de faire apparaître les évolutions des attributs et les événements qui en sont à l'origine.

6.1 Représentation des données issues des sources selon l'ontologie

La première étape du processus de construction consiste à représenter les données issues des sources historiques avec l'ontologie PeGazUs. Chaque source est représentée indépendamment afin de respecter ses spécificités sémantiques, temporelles et structurelles. Pour cela, nous adoptons une *approche par factoïdes* (Pasin et Bradley, 2015), initialement développée dans le cadre de la prosopographie numérique.

Un *factoïde* est une unité élémentaire d'information, issue d'une source, décrivant un objet du monde réel dans un certain contexte. Cette approche permet de représenter

séparément chaque observation faite par une source sur un objet, sans présupposer de la véracité partielle ou totale des informations. Chaque source contribue à une vision partielle et datée d'une entité géographique et c'est la combinaison des factoïdes qui permet la vision totale de chacune.

Dans notre approche, chaque source est associée à un graphe nommé, appelé *graphe de factoïdes*. Ce graphe contient :

— une ou plusieurs versions d'entités décrites par la source pendant un intervalle de validité temporel ;
— les attributs observés ou extraits (nom, numéro, géométrie, etc.) ;
— les relations explicites entre entités (appartenance, inclusion, continuité, etc.) ;
— les métadonnées sur la source (type, auteur, précision, date de publication, etc.).

Les graphes de factoïdes utilisent l'ontologie PeGazUs pour avoir une représentation uniforme des adresses. Chaque mention d'une entité géographique dans une source devient une instance de la classe `Landmark`. En fonction des informations associées à la mention, on peut y ajouter des attributs et leurs versions associées, ainsi que de potentielles relations entre entités. De plus, les mentions peuvent refléter deux aspects complémentaires de l'entité : soit son état spatio-temporel, soit un événement associé. Dans le premier cas, on associe à la mention un intervalle temporel, qui correspond par défaut au temps de validité de la source utilisée. Par exemple, dans l'atlas municipal de 1888 (dont on suppose que le temps valide est l'intervalle [1887-1889]), la rue Gérard est décrite par une géométrie polygonale et un nom. Les triplets associés sont donnés par le Listing 6.1.

```
ex:RueGerard1888 a :Landmark ;
    :isLandmarkType ltype:Thoroughfare ; :hasTime ex:1887_1889 ;
    :hasAttribute ex:RueGerard1888AttrGeom, ex:RueGerard1888AttrName ;
    prov:wasDerivedFrom ex:AtlasParis1888RueGerardEntry .

ex:RueGerard1888AttrGeom a :Attribute ;
    :hasAttributeVersion ex:RueGerard1888AttrVersGeom .

ex:RueGerard1888AttrName a :Attribute ;
    :hasAttributeVersion ex:RueGerard1888AttrVersName .

ex:RueGerard1888AttrVersGeom a :AttributeVersion ;
    :versionValue "POLYGON((...))"^^geo:wktLiteral ;
    prov:wasDerivedFrom ex:AtlasParis1888RueGerardEntry .

ex:RueGerard1888AttrVersName a :AttributeVersion ;
    :versionValue "rue Gérard"@fr ;
    prov:wasDerivedFrom ex:AtlasParis1888RueGerardEntry .

ex:AtlasParis1888RueGerardEntry a prov:Entity ;
    rico:isOrWasDescribedBy ex:AtlasParis1888 .

ex:AtlasParis1888 a rico:Record ;
    rdfs:label "Atlas municipal (...) de la Ville de Paris"@fr .
```

Listing 6.1 – Représentation RDF en Turtle de la rue Gérard du plan municipal de 1888 selon l'ontologie *PeGazUs*. Le préfixe `ex:` désigne un espace de nom des ressources, et la ressource `ex:1887_1889` décrit l'intervalle temporel correspondant à la période [1887; 1889].

Dans le second cas, une mention est interprétée comme un événement ponctuel daté, générant un changement appliqué à l'entité ; la nature de ce changement dépend du type d'événement. Par exemple, selon le dictionnaire d'Hillairet, la rue Gérard a été ouverte en 1851. Dans le graphe de factoïdes associé à ce dictionnaire, cette information est représentée par un événement (`ex:Event1851`) associé à cette date, ainsi que par un changement (`ex:RueGerard1851App`) appliqué à l'entité correspondante. Le changement est typé comme une apparition de la voie (`ctype:LandmarkAppearance`) et dépend directement de l'événement. Le Listing 6.2 montre les triplets générés pour cette mention, illustrant comment l'événement et le changement sont liés à l'entité tout en conservant son attribut de nom.

```
ex:RueGerard1851 a :Landmark ; :isLandmarkType ltype:Thoroughfare ;
    :hasAttribute ex:RueGerard1851AttrName ;
    prov:wasDerivedFrom ex:HillairetRueGerardEntry .

ex:RueGerard1851AttrName a :Attribute ;
    :hasAttributeVersion ex:RueGerard1851AttrVersName .

ex:RueGerard1851AttrVersName a :AttributeVersion ;
    :versionValue "rue Gérard"@fr ;
    prov:wasDerivedFrom ex:HillairetRueGerardEntry .

ex:Event1851 a :Event ;
    :hasTime ex:1851 ; prov:wasDerivedFrom ex:HillairetRueGerardEntry .

ex:RueGerard1851App a :Change ; :isChangeType ctype:LandmarkAppearance ;
    :dependsOn ex:Event1851 ; :appliedTo ex:RueGerard1851 .

ex:HillairetRueGerardEntry a prov:Entity ;
    rico:isOrWasDescribedBy ex:DictionnaireHillairet ;
    rdfs:label "Entrée 'Rue Gérard' du Dictionnaire historique des rues de
    ↪ Paris"@fr .

ex:DictionnaireHillairet a rico:Record ;
    rdfs:label "Dictionnaire historique des rues de Paris"@fr .
```

Listing 6.2 – Représentation RDF en Turtle d'un événement associé à la rue Gérard (formation en 1851) affirmée par le dictionnaire d'Hillairet. Le préfixe `ex:` désigne un espace de nom des ressources, et la ressource `ex:1851` décrit l'instant correspondant à l'année 1851.

Ainsi, l'ensemble des données provenant des sources historiques sont représentées conformément à l'ontologie PeGazUs, qu'importe leur nature. Répartir les données dans des graphes nommés, selon les sources dont elles sont issues, garantit l'intégrité de chaque source : une fois le graphe nommé créé, son contenu n'est plus modifié.

Cette répartition en graphes nommés permet également de gérer la qualité des sources : certaines peuvent être exclues si elles sont jugées peu fiables, ou d'autres incluses si elles apportent des informations pertinentes. Les étapes suivantes exploitent les données contenues dans chaque graphe de factoïdes en les reliant et en les confrontant. Comme dans de nombreuses approches d'intégration de données, l'objectif est de créer un graphe final de faits qui réconcilie l'ensemble des graphes issus des sources à l'aide d'une chaîne de

traitement que l'on présente dans les sections suivantes.

6.2 Intégration des ressources dans un graphe de faits

Une fois les graphes de factoïdes construits à partir des différentes sources, nous avons à présent un ensemble d'observations décrivant un territoire à différentes temporalités. L'étape suivante consiste à initialiser le graphe de faits en créant l'ensemble des entités du monde réel décrites par une ou plusieurs observations auxquelles elles doivent être rattachées. Cette phase implique l'identification de ressources similaires au sein des graphes de factoïdes et leur rattachement à une ressource racine unique dans le graphe de faits. La sous-section C.1.1 présente de manière algorithmique la méthode mise en œuvre pour cette étape.

6.2.1 Nature des liens pour les observations similaires

Comme vu dans le chapitre 5, une partie importante de l'intégration consiste à détecter et relier entre elles les données décrivant des éléments similaires, tels que les entités géographiques, les relations, les changements, les événements et les attributs. Afin de relier des entités considérées similaires, Tual *et al.* (2023) proposent d'utiliser un lien `owl:sameAs` en comparant deux à deux toutes les données disponibles. Cette approche ne nous paraît pas adaptée pour plusieurs raisons. Tout d'abord, dans le cadre de traitements de gros volumes de données, le nombre de comparaisons croît de manière quadratique, ce qui peut rapidement augmenter les temps de calcul, même si certaines optimisations sont possibles. De plus, plus le nombre d'entités similaires est élevé, plus le nombre de liens nécessaires augmente fortement. Par exemple, si quatre entités A, B, C, D sont considérées comme similaires, il faudrait créer (explicitement ou par inférence) les liens `owl:sameAs(A,A)`, `owl:sameAs(A,B)`, `owl:sameAs(A,C)`, `owl:sameAs(A,D)`, `owl:sameAs(B,C)`, ..., soit un total de 16 liens. Conceptuellement, ces quatre entités sont des observations : elles reflètent un point de vue particulier porté par une source sur une entité réelle, avec sa propre temporalité et ses attributs. Dans ce cas, si l'on reliait ces observations directement aux entités réelles correspondantes, le nombre de liens à créer se limiterait ici à quatre, ce qui est beaucoup plus parcimonieux. En outre, la symétrie et la transitivité du prédicat `owl:sameAs` entrent en contradiction avec le principe de préservation de l'intégrité de chaque source, cité précédemment. En effet, si l'on établit un lien `owl:sameAs(A,B)` entre deux observations A et B d'une même entité, les triplets associés à A deviennent également associés à B et inversement, entraînant une sorte de fusion des observations. Or, ce comportement est indésirable dans notre approche, car nous souhaitons conserver l'intégrité et l'individualité de chaque observation.

Plus précisément, l'utilisation de `owl:sameAs` pose les difficultés suivantes :

— elle induit une équivalence forte et symétrique entre les ressources, alors que les représentations issues des sources peuvent être partielles, contradictoires ou contextuelles ;

— elle entraîne une propagation non contrôlée des assertions entre les ressources, ce qui peut polluer le graphe en agrégeant à tort des données incompatibles ;

— elle empêche la représentation de relations asymétriques entre une entité et ses observations historiques. Nous devons en effet établir un lien orienté permettant de tracer les versions documentées d'une même entité sans induire une substitution complète entre les identifiants.

Au-delà des aspects techniques, l'utilisation du prédicat `owl:sameAs` présente un problème conceptuel : ce lien implique une équivalence stricte entre deux ressources, ce qui n'est pas approprié pour relier des observations fragmentaires ou partielles représentant des entités réelles. Nous proposons, comme beaucoup d'approches d'intégration de données, de créer un graphe unifié qui est le *graphe de faits* contenant les représentations des entités du monde réel, et de relier les observations directement à ces ressources plutôt qu'entre elles. Toutefois, l'utilisation de `owl:sameAs` pose également un problème dans ce contexte pour la même raison. La distinction entre une représentation d'entité réelle et ses représentations documentaires rejoint les critiques formulées à l'encontre de `owl:sameAs` dans le Web des données. Comme l'ont montré Halpin *et al.* (2010), il est essentiel de distinguer :

— les cas où deux ressources décrivent la même chose dans des contextes différents (non remplaçables) ;
— ceux où une ressource représente une autre sans en être équivalente ;
— ceux où deux ressources sont très proches mais pas identiques (par exemple : la commune de Paris vs. le département de Paris).

La propriété que nous voulons introduire se positionne dans la première catégorie : elle doit indiquer qu'une ressource source R_S constitue une trace documentaire de la ressource réelle R_F, sans que ces deux ressources soient substituables ou fusionnables.

Nous avons examiné plusieurs alternatives à `owl:sameAs`, mentionnées dans la littérature ou intégrées dans des vocabulaires établis :

— relations de succession ou de remplacement : `dc:replaces` de Dublin Core (Coyle et Baker, 2009), `wdt:P1365` et `wdt:P155` (respectivement *remplace* et *remplacé par*) de Wikidata,
— relations d'équivalence ou de correspondance : `skos:exactMatch`, `skos:closeMatch`, `skos:broadMatch`, `skos:narrowMatch` de SKOS (Miles *et al.*, 2005) ou encore `foaf:isPrimaryTopicOf` de FOAF (Graves *et al.*, 2007).

Les relations de succession ou de remplacement ne sont pas compatibles avec le lien entre une entité et ses observations. Les relations décrivent généralement la succession temporelle entre entités, alors que nous souhaitons indiquer qu'une observation constitue une représentation spécifique d'une entité. Une approche plus adaptée serait d'utiliser une relation d'inclusion, liant chaque observation à l'entité qu'elle décrit. Les relations d'équivalence ou de correspondance sont conceptuellement plus cohérentes pour relier des entités similaires, mais elles ne sont pas parfaites pour notre cadre.

Aucune de ces propriétés ne satisfait cependant les exigences suivantes : asymétrie contrôlée, possibilité d'associer les entités réelles à l'ensemble des représentations documentaires qui les décrivent (c'est-à-dire à leurs observations dans les graphes de factoïdes), et capacité à centraliser les relations au moyen d'une observation commune, évitant la multiplication de liens entre les observations. Tous ces arguments nous mènent à l'introduction d'une propriété spécifique, `hasTrace` dont la propriété symétrique est `isTraceOf`, qui relie la ressource du graphe de faits aux ressources issues des graphes de factoïdes. Elle permet de :

— relier une représentation d'entité réelle R_F du graphe de faits à une ressource R_S issue d'un graphe source ;
— ne pas induire de symétrie, ce qui permet une dissociation claire entre la donnée agrégée sur une entité du monde réel et ses représentations documentaires ;
— propager de manière ascendante des informations (des sources vers les faits) mais pas descendante ni entre sources.

Autrement dit, `hasTrace` permet de considérer les ressources des graphes sources

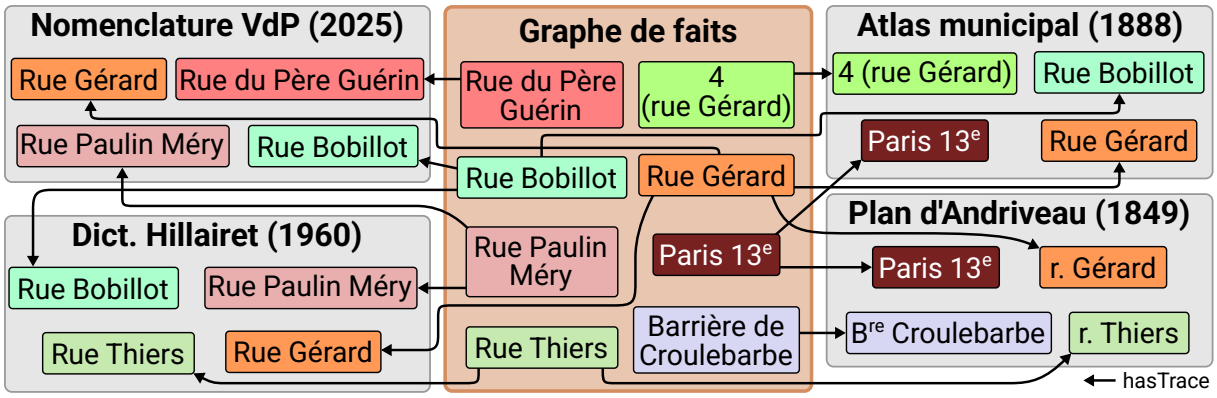


FIGURE 6.1 – Initialisation du graphe de faits intégrant les entités réelles résultant du croisement des sources dont les données sont contenues des graphes de factoïdes. Les entités du graphe de faits sont reliées à ses observations dans les sources avec le lien `hasTrace`. ■ Graphe de factoïdes, ■ Graphe de faits.

comme des témoignages ou des vues partielles sur une entité réelle unique. On peut ainsi agréger dans le graphe de faits les attributs validés, les changements observés et les événements historiques, tout en conservant la traçabilité vers les sources. Lorsque plusieurs ressources sont reliées à une même entité réelle, elles constituent autant d'observations qui fournissent des informations similaires ou, au contraire, différer sur les dimensions spatiales, attributaires ou temporelles. Leur comparaison permet alors d'identifier des évolutions des entités au cours du temps.

6.2.2 Identification des observations représentant la même entité

L'objectif de cette étape est de créer et d'initialiser dans le graphe de faits les entités correspondant au monde réel, à partir de l'ensemble des observations issues des graphes de factoïdes. Cette démarche présente des points communs avec celle proposée par (Charles, 2025), qui introduit l'ontologie DHFC pour distinguer explicitement ce qui est exprimé dans un document de l'interprétation faite de ces informations par un historien. Dans ce cadre, on peut rapprocher les graphes de factoïdes de ce que l'auteur nomme « graphes de source » et le graphe de faits de son « graphe de synthèse ».

Les constructions réalisées lors de cette étape sont décrites par l'algorithme C.1, tandis que le résultat attendu est illustré par la figure 6.1. Bien que les observations puissent concerner non seulement les entités mais aussi leurs attributs, relations et événements associés, il est essentiel de commencer par l'identification et la création des entités elles-mêmes. Ce graphe, initialement vide, est construit de manière incrémentale : les entités sont créées au fur et à mesure, et chaque observation qui leur est associée est reliée à l'entité correspondante. Cette étape permet d'établir un cadre structuré à partir duquel il sera possible de déduire et d'affirmer les liens pour les autres éléments du graphe, tout en conservant l'intégrité et l'individualité de chaque observation. De plus, cette approche offre la flexibilité d'intégrer de nouvelles sources au fur et à mesure, en enrichissant le graphe de faits sans modifier les graphes de factoïdes existants.

Considérons O_L l'ensemble des observations des entités géographiques, que nous définissons par

$$O_L := \{\, o \mid \exists G \in \mathcal{G}_S : \texttt{Landmark}(o) \in G \,\}$$

où o désigne une observation d'une source décrite par G faisant partie d'un ensemble de graphes de factoïdes $\mathcal{G}_S = \{G_\alpha, \ldots, G_\omega\}$. Autrement dit, O_L définit l'ensemble des ressources pour lesquelles il existe un triplet dans un graphe de factoïdes les définissant comme appartenant à la classe `Landmark`.

Soit un autre ensemble L représentant les entités géographiques réelles. Au départ, L est vide et il est rempli de manière incrémentale au fur et à mesure du parcours des observations de O. À terme, nous considérons l'ensemble des entités

$$L := \{l \mid \texttt{Landmark}(l) \in G_F\},$$

où G_F désigne le graphe nommé des faits.

Pour chaque observation $o \in O_L$, si une entité $l \in L$ existe déjà et que l'on considère que o la décrit, on crée le lien `hasTrace` entre l et o. Dans le cas contraire, on crée un nouvel élément l' dans L et on établit le lien $\texttt{hasTrace}(l', o)$. Ainsi, il existe une surjection de l'ensemble des observations O_L vers l'ensemble des entités du graphe des faits L, chaque observation étant associée à l'entité qu'elle décrit.

Lors de la création d'une nouvelle entité, on lui associe également deux changements décrivant respectivement son apparition et sa disparition, chacun dépendant d'un événement non temporalisé. Nous définissons l'ensemble des types de changements possibles :

$$\Theta_L := \{\texttt{ctype:LandmarkAppearance}, \texttt{ctype:LandmarkDisappearance}\}.$$

Cette représentation reflète l'idée que toute entité fait au moins l'objet d'une naissance et une disparition, même si les dates exactes ne sont pas connues. Pour les entités encore actuelles, on peut considérer que leur disparition se situe à un instant infini, ce qui permet de représenter leur existence comme continue jusqu'à aujourd'hui. On définit alors l'application $\chi_L : L \times \Theta_L \longrightarrow C_L$, telle que pour toute entité $l \in L$:

$$c_l^{\text{app}} := \chi_L(l, \texttt{ctype:LandmarkAppearance}), c_l^{\text{dis}} := \chi_L(l, \texttt{ctype:LandmarkDisappearance}),$$

où c_l^{app} désigne le changement impliquant l'apparition de l et c_l^{dis} le changement impliquant la disparition de l.

Chaque changement est associé à un événement non temporalisé via l'application $\epsilon : C_L \longrightarrow E_L$, tel que

$$e_l^{\text{app}} := \epsilon(c_l^{\text{app}}), \qquad e_l^{\text{dis}} := \epsilon(c_l^{\text{dis}}).$$

Au préalable, nous avons indiqué qu'il faut déterminer si une observation $o \in O_L$ peut être considérée comme décrivant une entité existante $l \in L$. Les critères précis de similarité ne sont pas encore détaillés, mais comme expliqué au chapitre 1, l'information systématiquement disponible dans les sources pour une entité est son nom. Lors de la représentation des données conformément à l'ontologie PeGazUs, nous conservons également de manière systématique le type de l'entité. Cette information est essentielle pour éviter des confusions, par exemple pour ne pas relier le cimetière du Père-Lachaise au quartier du Père-Lachaise, qui sont souvent désignés par la même appellation « Père-Lachaise » mais représentent des entités de types différents.

Notre critère principal d'identité repose sur la similarité des noms et de la place de l'entité dans la hiérarchie. Nous n'explicitons pas ici la manière dont cette similarité est calculée, plusieurs techniques peuvent être employées selon les objectifs et les données disponibles.

On définit une similarité entre une observation $o \in O$ et une entité $l \in L$ par :

$$\text{sim}(o, l) := \begin{cases} \text{sim_label}(o, l) & \text{si } \exists t, \texttt{isLandmarkType}(o, t) \wedge \texttt{isLandmarkType}(l, t), \\ 0 & \text{sinon.} \end{cases}$$

Ici, sim_label(o, l) désigne la similarité entre les labels simplifiés de o et l. Le lien `hasTrace` est alors généré lorsque la similarité dépasse un seuil minimal $\alpha \in [0; 1]$:

$$\text{sim}(o, l) \geq \alpha \implies \texttt{hasTrace}(o, l).$$

Un tel critère purement fondé sur le nom présente cependant des limites. Par exemple, deux numéros d'immeubles « 8 » situés dans des rues différentes seraient considérés comme identiques (le « 8 » de la rue Gérard et celui de la rue de Rivoli). De même, des voies avec des noms génériques comme « rue de l'Église » apparaissent dans de nombreuses communes. Pour gérer ces ambiguïtés, nous combinons le critère de nom avec les relations entre entités. Une observation o décrit une entité l du graphe des faits si les deux conditions vérifiées : leur nom est similaire et elles partagent une relation avec une même entité déjà unifiée.

Cela suppose de définir un ordre de résolution des liens d'identité. En effet, pour trouver les équivalences entre les numéros d'immeubles, il convient de le faire au niveau des voies de communication et pour ces dernières, il convient de le faire au niveau des quartiers et des communes. Nous retenons l'ordre suivant : commune ou quartier, voie de communication, puis numéro d'immeuble. Bien que cette supposition soit incorrecte pour d'autres cas d'usage, on suppose par ailleurs l'unicité des noms de communes dans la zone étudiée : ainsi, toutes les observations mentionnant *Paris* sont rattachées à l'entité du graphe des faits `Paris`.

Nous mettons à jour la similarité entre une observation $o \in O_L$ et une entité $l \in L$ pour tenir compte des relations spatiales. Soit o_{rel} et l_{rel} les entités reliées respectivement à o et l par une relation de type t. Nous utilisons la fonction relation(t, x, y) comme notation abrégée indiquant qu'il existe une relation de type t dont le locatum est x et le relatum est y. Formellement, cette notation correspond à :

$$\exists\, \text{rel} \Big(\texttt{isLandmarkRelationType}(\text{rel}, t) \ \wedge\ \texttt{locatum}(\text{rel}, x) \ \wedge\ \texttt{relatum}(\text{rel}, y) \Big).$$

On définit alors :

$$\text{sim}'(o, l) := \begin{cases} \text{sim_label}(o, l) & \text{si } \exists t, \texttt{isLandmarkType}(o, t) \wedge \texttt{isLandmarkType}(l, t) \\ & \quad \wedge \text{relation}(t, o, o_{\text{rel}}) \wedge \text{relation}(t, l, l_{\text{rel}}) \\ & \quad \wedge \texttt{hasTrace}(l_{\text{rel}}, o_{\text{rel}}), \\ 0 & \text{sinon.} \end{cases}$$

Ainsi, une observation o décrit une entité l de type voie si leurs labels simplifiés coïncident et si chacune est en relation `lrtype:Within` avec des entités déjà reliées par `hasTrace`. Le même principe s'applique aux numéros d'immeubles, en remplaçant le type de relation `lrtype:Within` par `lrtype:Belongs`. Reprenons l'exemple des numéros d'immeuble ayant le numéro 8. La figure 6.2 est une application de ce principe. Dans le graphe des faits, il existe ainsi plusieurs entités portant le nom *8*, la règle permet de relier l'observation décrivant le 8 de la rue Gérard à l'entité du graphe des faits, notamment grâce à une étape faite auparavant ayant permis la mise en relation entre l'entité du graphe des faits *Rue Gérard* et ses observations.

6.2.3 Propagation de liens aux attributs, relations, changements et événements

Une fois les entités du graphe de faits établies et les observations similaires rattachées aux entités qu'elles décrivent, il est possible, comme le montre l'algorithme C.2, de

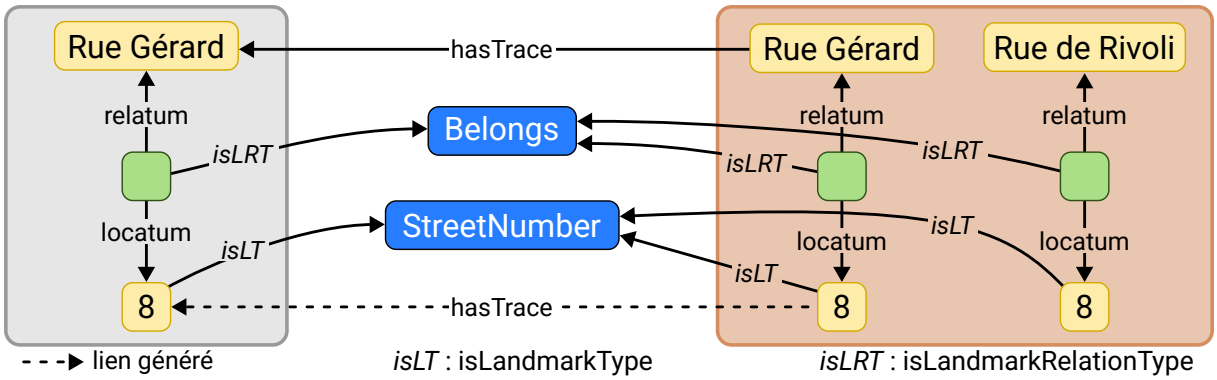


FIGURE 6.2 – Exemple de génération de liens de correspondance pour des entités de type `ltype:StreetNumber`. Graphe de factoïdes, Graphe de faits, `Landmark`, `LandmarkRelation`, `skos:Concept`.

propager ces équivalences aux ressources dépendantes : attributs, relations entre entités, changements et événements. Détecter des relations entre repères constitue une tâche complexe, car une relation peut être reliée à plusieurs ressources via la propriété `relatum`. L'objectif est similaire à la section précédente, il s'agit de faire en sorte que l'ensemble des observations de O_R soient reliées aux relations de l'ensemble R. Ici, O_R est défini par :

$$O_R := \{\, o \mid \exists G \in \mathcal{G}_S : \texttt{LandmarkRelation}(o) \in G \,\}$$

En parallèle, R est défini de la manière suivante :

$$R := \{r \mid \texttt{LandmarkRelation}(r) \in G_F\},$$

Pour qu'une observation o décrivant une relation soit tracée vers une relation r du graphe de faits, il faut que toutes les entités liées à o – dont l'ensemble est noté $\mathcal{O}_o$ – soient elles-mêmes tracées par des éléments de E. Formellement, $\mathcal{O}_o := \{o_l \mid \texttt{locatum}(o, o_l) \vee \texttt{relatum}(o, o_l)\}$. Il faut donc s'assurer que : $\forall o_l \in \mathcal{O}_o,\ \exists l \in L,\ \texttt{hasTrace}(l, o_l)$.

Ainsi, un élément $o \in O_R$ peut être considéré comme décrivant une relation $r \in R$ impliquant la création d'un lien $\texttt{hasTrace}(o, r)$ si les trois conditions suivantes sont satisfaites :

1. **Similarité des types :** leurs natures respectives doivent être compatibles, c'est-à-dire

$$\exists t, \texttt{isLandmarkRelationType}(o, t) \wedge \texttt{isLandmarkRelationType}(r, t).$$

2. **Correspondance des locatums :** le locatum de o doit être une trace du locatum de r :

$$\exists l, \texttt{locatum}(o, o_l) \wedge \texttt{locatum}(r, l) \wedge \texttt{hasTrace}(o_l, l).$$

3. **Correspondance des relatums :** pour chaque relatum de o, il existe un relatum de r dont il est la trace. Autrement dit,

$$\forall o_l \in \{o_l \mid \texttt{relatum}(o, o_l)\},\ \ \exists l \in \{l \mid \texttt{relatum}(r, l)\} \text{ tel que } \texttt{hasTrace}(o_l, l).$$

De manière analogue à la création des entités du graphe de faits, si une observation $o \in O_R$ n'est pas encore associée à une relation dans R, il convient de créer une nouvelle relation $r \in R$ et de générer le lien $\texttt{hasTrace}(o, r)$. En parallèle, il faut générer deux

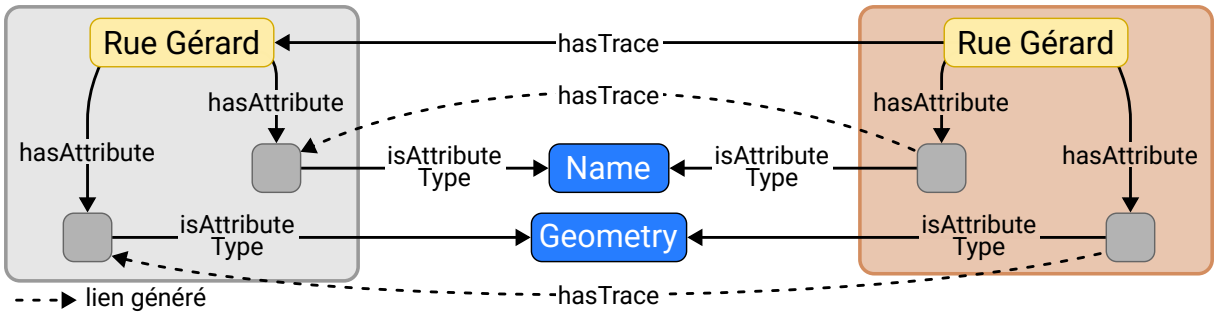


FIGURE 6.3 – Exemple de génération de liens de correspondance pour des attributs. Graphe de factoïdes, Graphe de faits, `Landmark`, `Attribute`, `skos:Concept`.

ensembles associés à R : l'ensemble C_R des changements associés à l'apparition et à la disparition des éléments de R et E_R, l'ensemble des événements correspondants.

Pour formaliser, on définit une application $\chi_R : R \times \Theta_R \longrightarrow C_R$. Elle associe le changement correspondant à chaque relation et chaque type de changement avec : $\Theta_R :=$ {`ctype:LandmarkRelationAppearance`, `ctype:LandmarkRelationDisappearance`}

Concernant les attributs, les changements et les événements, les règles à appliquer sont moins complexes. Tout d'abord, nous postulons qu'une observation o représentant un attribut d'une entité géographique décrit l'attribut du même type de l'entité dont elle est la trace. Soit O_A l'ensemble des observations décrivant des attributs, chacune étant nécessairement reliée à une observation d'entité $o_l \in O_L$ par une relation `hasAttribute`. Il est défini par :

$$O_A := \{\, o \mid \texttt{Attribute}(o) \wedge \exists o_l \in O_L,\ \texttt{Landmark}(o_l) \wedge \texttt{hasAttribute}(o_l, o) \,\}$$

L'étape précédente a permis d'associer o_l à une entité réelle $l \in L$ avec $\texttt{hasTrace}(o_l, l)$. Si l possède un attribut a de même type t que o, alors l'observation o doit être associée à l'attribut correspondant de l :

$$\begin{aligned} \exists t, \texttt{isAttributeType}(o, t) \wedge \texttt{isAttributeType}(a, t) \\ \wedge\, \texttt{hasAttribute}(o_l, o) \wedge \texttt{hasAttribute}(l, a) \\ \wedge\, \texttt{hasTrace}(o_l, l) \implies \texttt{hasTrace}(o, a) \end{aligned}$$

Si o n'est pas reliée à l'attribut a du graphe de faits, alors il convient de générer cette ressource et de la relier à l'entité correspondante de $l \in L$. Cette génération de liens `hasTrace` est illustrée par la figure 6.3 qui montre une application des règles présentées pour la rue Gérard.

Contrairement au cas des attributs, l'association des observations de changements à des changements du graphe de faits ne peut pas toujours être garantie. Nous définissons l'ensemble des observations de changements comme suit :

$$O_C := \{\, o \mid \texttt{Change}(o) \wedge \exists x \in (O_L \cup O_R \cup O_A),\ \texttt{appliedTo}(o, x) \,\}.$$

Il n'est donc pas possible d'assurer, en général, que $\forall o \in O_C,\ \exists c \in C,\ \texttt{hasTrace}(o, c)$, où C désigne l'ensemble des changements du graphe de faits ($C_L \subset C$ et $C_R \subset C$). En particulier, le cas des changements appliqués aux attributs (`AttributeChange`) reste complexe et sera discuté dans une section ultérieure. En revanche, pour les changements appliqués aux entités et aux relations (`LandmarkChange` et `LandmarkRelationChange`), la correspondance est immédiate, puisque lors de la création de chaque entité et chaque

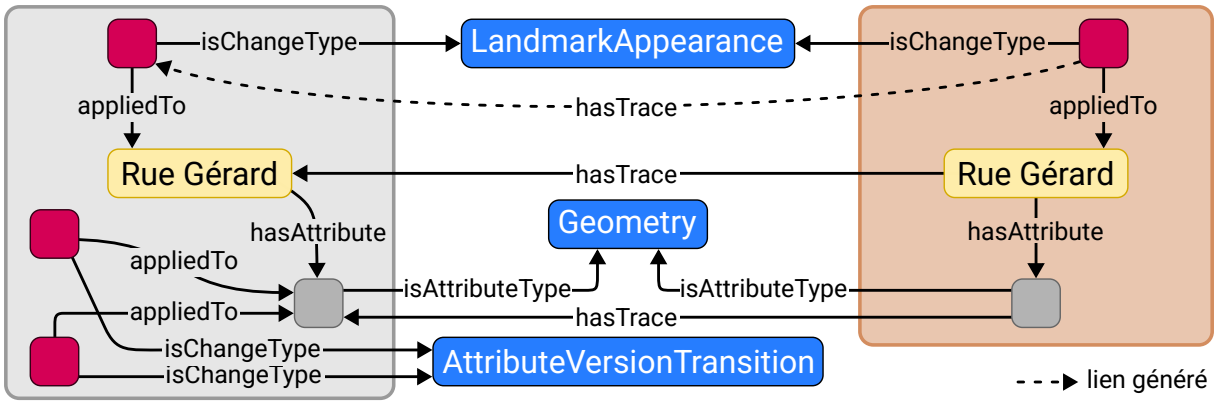


FIGURE 6.4 – Exemple de génération de liens d'équivalences pour des changements. Graphe de factoïdes, Graphe de faits, `Landmark`, `Attribute`, `Change`, `skos:Concept`.

relation dans le graphe de faits, des changements d'apparition et de disparition associés y sont aussi initialisés et sont uniques. Ainsi, il existe nécessairement un changement de même nature auquel les relier :

$$\exists t,\ \texttt{isChangeType}(o,t) \wedge \texttt{isChangeType}(c,t) \wedge \texttt{appliedTo}(o,o_l) \\ \wedge\ \texttt{appliedTo}(c,l) \wedge \texttt{hasTrace}(o_l,l) \implies \texttt{hasTrace}(o,c)$$

Ces règles relatives à la création de liens de provenance pour les changements sont appliquées à l'exemple illustré par la figure 6.4. Dans ce cas, en raison de la multiplicité des changements affectant l'attribut de type géométrie (changements de type `AttributeVersionTransition`), aucun lien `hasTrace` n'est généré pour ces derniers. Seul le changement portant directement sur la rue est concerné par un lien de provenance `hasTrace`.

Enfin, la tâche finale de cette étape consiste à relier les observations décrivant des événements – dont l'ensemble est noté $O_E = \{o \mid \texttt{Event}(o)\}$ – aux événements du graphe de faits. Un changement est supposé dépendre d'un seul événement. Ainsi, lorsqu'une observation de changement $o_c \in O_C$ dépend d'une observation d'événement $o_e \in O_E$, c'est-à-dire $\texttt{dependsOn}(o_c, o_e)$, et que cette observation o_c est associée à un changement du graphe de faits $c \in C$ tel que $\texttt{hasTrace}(c, o_c)$, lequel dépend d'un événement e (noté $\texttt{dependsOn}(c, e)$), alors on peut en déduire que l'observation d'événement o_e est associée à l'événement du graphe de faits e, c'est-à-dire :

$$\forall (o_c, o_e, c, e) \in O_C \times O_E \times C \times E, \\ \Big(\texttt{dependsOn}(o_c, o_e) \wedge \texttt{hasTrace}(c, o_c) \wedge \texttt{dependsOn}(c, e)\Big) \Rightarrow \texttt{hasTrace}(o_e, e)$$

Par ailleurs, les événements peuvent être considérés comme à l'origine de plusieurs changements. Pour un événement du graphe de faits e donné, il peut exister plusieurs changements $c_1, c_2, \ldots, c_n$ tels que :

$$\forall e\ \exists c_1, \ldots, c_n \quad \texttt{dependsOn}(c_i, e), \quad \text{avec } i \in [\![1; n]\!].$$

Autrement dit, la relation $\texttt{dependsOn}(c, e)$ peut être multiple pour un événement $e \in E$ fixé, tandis qu'un changement c ne dépend que d'un seul événement. Enfin, une même observation d'événement peut être associée à plusieurs événements du graphe de faits. Dans

ce cas, ces événements doivent être considérés comme équivalents, ce que l'on exprime formellement par :

$$\forall e_1, e_2, o_e \in E \times E \times O_E, \quad \texttt{hasTrace}(o_e, e_1) \ \wedge \ \texttt{hasTrace}(o_e, e_2) \ \Rightarrow \ \texttt{owl:sameAs}(e_1, e_2).$$

6.3 Déduction des dates d'apparition et de disparition des entités et des relations

Après avoir propagé les liens de correspondance vers les attributs, relations, changements et événements, nous cherchons désormais à inférer de nouvelles connaissances à partir de l'ensemble des observations reliées aux ressources du graphe des faits. La méthode est synthétisée dans la sous-section C.1.2 (voir algorithme C.3 et algorithme C.4). En premier lieu, on peut déduire ou estimer les dates d'apparition et de disparition des entités et de leurs relations. Les observations peuvent être explicitement datées ou non. Lorsque ce n'est pas le cas, la temporalité peut être déduite des sources qui les documentent. Chaque observation correspond alors à un état spatio-temporel associé à un intervalle temporel ou à un événement ponctuel.

Soit $x \in L \cup R$ une entité ou une relation, à laquelle on associe

$$obs(x) = \{o \mid \texttt{hasTrace}(x, o)\}$$

l'ensemble des observations qui la décrivent dans des sources historiques.

On suppose l'existence de deux changements c_{app} et c_{dis} décrivant respectivement l'apparition et la disparition de x, tels que

$$\texttt{appliedTo}(c_{app}, x) \quad \text{et} \quad \texttt{appliedTo}(c_{dis}, x).$$

Ces changements sont eux-mêmes associés à des événements e_{app} et e_{dis} :

$$\texttt{dependsOn}(c_{app}, e_{app}) \quad \text{et} \quad \texttt{dependsOn}(c_{dis}, e_{dis}).$$

Cela implique que l'événement d'apparition e_{app} précède t_{deb} et que l'événement de disparition e_{dis} succède à t_{fin}, c'est-à-dire :

$$\texttt{hasTimeBefore}(e_{app}, t_{deb}) \text{ et } \texttt{hasTimeAfter}(e_{dis}, t_{fin})$$

Sinon, l'observation décrit une évolution appliquée à x, associée à un instant t, avec *tPred* et (c, e) définis par $tPred \in \{\texttt{hasTime}, \texttt{hasTimeBefore}, \texttt{hasTimeAfter}\}$ et $(c, e) \in \{(c_{app}, e_{app}), (c_{dis}, e_{dis})\}$:

$$\texttt{appliedTo}(o_c, x) \wedge \texttt{dependsOn}(o_c, o_e) \wedge \texttt{hasTrace}(o_e, e) \wedge tPred(o_e, t) \Rightarrow tPred(e, t).$$

Ainsi, chaque événement e est associé à un ensemble de contraintes temporelles issues des observations via des propriétés temporelles. Ces contraintes peuvent être combinées afin d'estimer une valeur temporelle cohérente pour l'événement. Cette estimation des valeurs temporelles se fait de la manière suivante :

- si un seul triplet comportant le prédicat `hasTime` existe pour e, alors l'événement est simplement associé à cette valeur temporelle.
- si plusieurs triplets comportant le prédicat `hasTime` sont présents, on génère un instant flou qui concilie toutes les informations temporelles ;

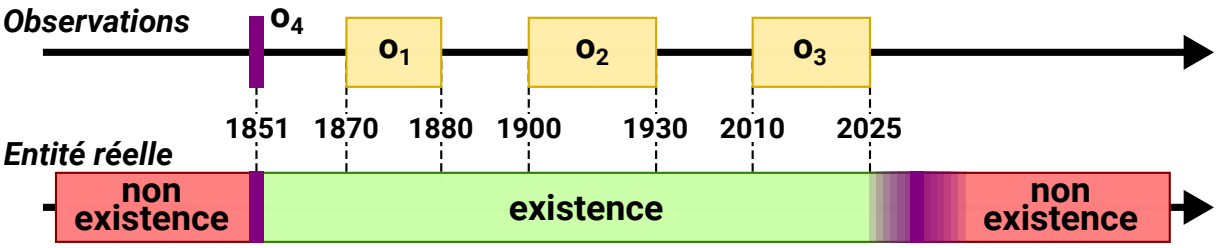


FIGURE 6.5 – Reconstitution des états spatio-temporels de la rue Gérard à partir de ses observations.

— si aucun triplet comportant le prédicat `hasTime` n'existe, on génère des triplets `hasTimeAfter`(e, t) et `hasTimeBefore`(e, t) à partir des observations disponibles, où t correspond respectivement à l'instant initial parmi les débuts des intervalles observés et à l'instant initial parmi les fins des intervalles observés.

Plus précisément, lorsque l'on génère les triplets `hasTimeAfter` et `hasTimeBefore` à partir des observations :

— pour `hasTimeAfter`(e, t), on considère l'ensemble des valeurs t_i issues des couples $(\texttt{hasTimeAfter}, t_i)$ présents dans les contraintes et l'on prend $t = \max(t_i)$;
— pour `hasTimeBefore`(e, t), on considère l'ensemble des valeurs t_i issues des couples $(\texttt{hasTimeBefore}, t_i)$ présents dans les contraintes et l'on prend $t = \min(t_i)$.

Cette précision permet d'assurer que la borne inférieure de l'événement respecte toutes les observations antérieures, et que la borne supérieure respecte toutes les observations postérieures.

Pour illustrer cette partie par un exemple, considérons la rue Gérard, décrite par quatre observations triées temporellement comme le montre la figure 6.5. Trois d'entre elles concernent des états de la voie : l'observation o_1 avec un intervalle de temps valide $[1870, 1880]$, o_2 avec $[1900, 1930]$, et o_3 avec $[2010, 2025]$. La quatrième observation o_4 correspond à une évolution, indiquant l'apparition de la voie en 1851. Cette observation est associée au changement c_{app}, lui-même dépendant de l'événement d'apparition e_{app}, pour lequel on connaît la date exacte `hasTime`$(e_{app}, 1851)$. Pour l'événement d'apparition e_{app}, on dispose également des informations provenant des états observés : `hasTimeBefore`$(e_{app}, 1870)$, `hasTimeBefore`$(e_{app}, 1900)$ et `hasTimeBefore`$(e_{app}, 2010)$. La combinaison de ces contraintes conduit naturellement à confirmer que `hasTime`$(e_{app}, 1851)$. Concernant l'événement de disparition e_{dis} associé à la voie, les observations ne fournissent aucune date exacte. Toutefois, nous disposons des deux contraintes suivantes : `hasTimeAfter`$(e_{dis}, 1880)$, `hasTimeAfter`$(e_{dis}, 1930)$ et `hasTimeAfter`$(e_{dis}, 2025)$. On en déduit `hasTimeAfter`$(e_{dis}, 2025)$, c'est-à-dire que la date de disparition, si elle existe, est postérieure à 2025. La datation des événements d'apparition e_{app} et de disparition e_{dis} permet ainsi de reconstituer les trois états successifs de la voie, comme le montre la frise inférieure de la figure 6.5.

À ce stade, il est désormais possible de reconstituer les états spatio-temporels des entités et des relations. En effet, grâce aux événements datés, plus ou moins précisément, qui servent de transitions entre les phases de non-existence, d'existence et de disparition (selon la typologie de Hallot (2012)), on peut déterminer quelles entités et relations existaient à un instant donné t. Cependant, la dimension attributaire n'est pas encore traitée : on sait qu'une entité existait à un instant t, mais pas quelles étaient ses propriétés à cet instant (nom, géométrie, etc.).

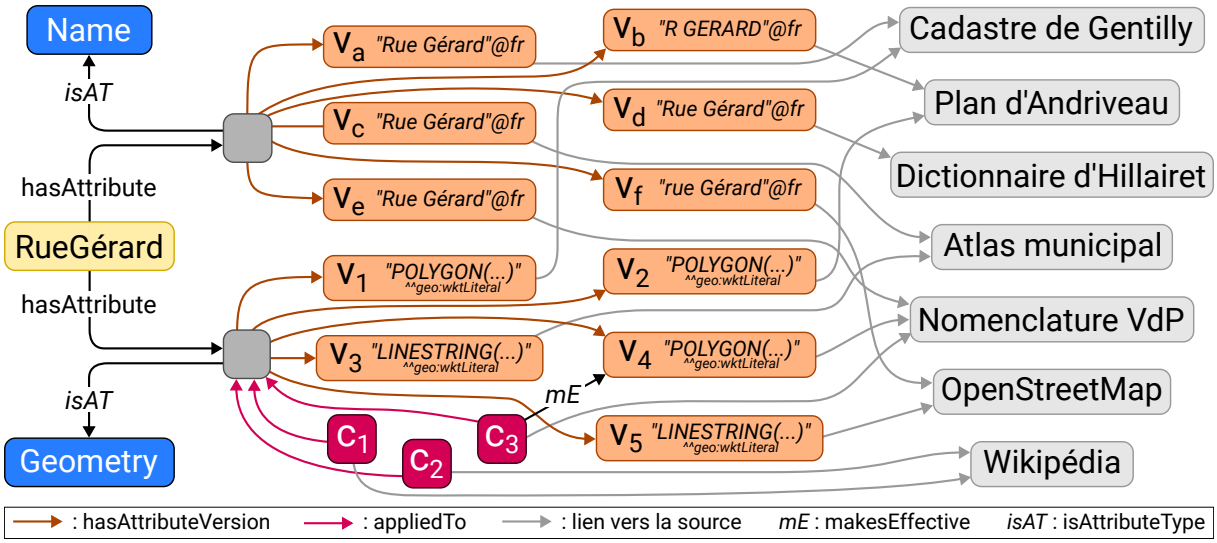


FIGURE 6.6 – Agrégation des versions des différents attributs de la rue Gérard.

6.4 Reconstitution de l'évolution des attributs

Pour chaque attribut $a \in A$ d'une entité du graphe de faits, nous disposons d'un ensemble d'observations issues de différentes sources : $obs(a) = \{o \mid \texttt{hasTrace}(a, o)\}$. Ces observations sont reliées à des changements qui décrivent les évolutions de l'attribut, et permettent ainsi de reconstituer, au fil du temps, les différentes versions de cet attribut. L'objectif est donc d'établir une chronologie cohérente de ses valeurs à partir des données hétérogènes et fragmentaires dont nous disposons. Par exemple, la figure 6.6 illustre pour la rue Gérard à Paris l'agrégation de l'ensemble des versions et des changements appliqués à chaque attribut provenant de diverses sources. On veut être capable d'une part de reconstituer l'évolution de la géométrie de la rue à partir des ensembles de versions $V = \{v_1, \ldots, v_5\}$ et de changements $C = \{c_1, c_2, c_3\}$.

Une première approche pour reconstituer l'évolution des attributs a été proposée par Bernard *et al.* (2025c). Cependant, Bernard *et al.* (2025a) affirment que cette méthode présente plusieurs limites. Tout d'abord, les changements inférés restent temporellement flous, l'incertitude pouvant être très élevée lorsque deux versions successives sont éloignées dans le temps. De plus, elle ne prend en compte que des versions pour en déduire des changements, alors que nous disposons des changements dans les données qui peuvent aussi inférer des versions. Enfin, la méthode rencontre des difficultés dans les cas de chevauchements temporels entre différentes versions, conduisant parfois à des inférences incohérentes (par exemple, à des intervalles $[t_i, t_j]$ avec $t_i > t_j$).

Cette étape est illustrée de manière algorithmique dans la sous-section C.1.3. Elle consiste en deux phases principales :

1. la représentation élémentaire de l'évolution des attributs ;
2. la reconstitution de l'évolution complète à partir de cette représentation élémentaire.

6.4.1 Représentation élémentaire de l'évolution des attributs

À ce stade, les attributs des entités géographiques agrègent différents factoïdes — des versions (`AttributeVersion`) et des changements (`AttributeChange`) — à partir

desquels il est possible de reconstituer l'évolution de chaque attribut en générant des faits. Pour un attribut a d'une entité géographique, notons $C = \{c_1, ..., c_n\}$ l'ensemble de ses changements et $V = \{v_1, ..., v_m\}$ pour celui de ses versions.

Tout d'abord, il convient de procéder à un découpage élémentaire. Cette tâche consiste à diviser la ligne de vie d'un attribut en un ensemble d'intervalles indivisibles. Ces intervalles représentent les périodes pendant lesquelles une même version de l'attribut est valable, sans discontinuité ni chevauchement. Formellement, si l'on considère un attribut a, le découpage élémentaire consiste à construire une partition de l'axe temporel en intervalles disjoints : $\mathcal{T}(a) = \{[t_0, t_1], [t_1, t_2], \ldots, [t_{n-1}, t_n]\}$ tels que, pour tout k, une version de a reste constante sur $[t_k, t_{k+1}]$. À chaque intervalle $[t_k, t_{k+1}]$ est associée une version élémentaire notée $\beta_k \in B$, représentant l'état de l'attribut a sur toute cette période. De même, à chaque instant de transition t_k est associé un changement élémentaire $\gamma_k \in \Gamma$, qui relie la version β_{k-1} à la version β_k. Ainsi, ce découpage fournit la base de la construction des versions et changements, comme illustré dans la figure 6.7.

Pour cela, deux phases de traitement sont nécessaires : l'initialisation de changements et de versions et l'affiliation de ces données initialisées aux factoïdes. En prenant l'exemple de la rue Gérard située à Paris dont une représentation multi-source est donnée par la figure 6.6, son attribut "géométrie" est composé de huit factoïdes :

- une version v_1 valable sur la période 1845-1849, d'après le cadastre napoléonien de Gentilly ;
- une version v_2 valable entre 1847 et 1851, d'après le plan Andriveau de la ville de Paris ;
- une version v_3 valable entre 1887 et 1889, d'après l'atlas municipal de Paris ;
- deux changements de géométrie c_1 et c_2 qui ont respectivement lieu en 1857 et 1979 selon Wikipédia ; les changements sont attestés, mais aucune information sur la géométrie avant ou après n'est donnée ;
- un changement de géométrie c_3 qui a lieu en 1979 et rend effectif la version v_4, d'après la nomenclature des voies de Paris ;
- une dernière version v_5 valable entre 2024 et 2025, d'après OpenStreetMap.

On remarque qu'il y a d'une part un chevauchement des temps valides de v_1 et v_2 que, d'autre part, qu'il existe des intervalles temporels conséquents pour lesquels nous ne disposons d'aucune information. Malgré ces lacunes et ces chevauchements, notre méthode permet tout de même de reconstruire l'évolution de la géométrie de la rue.

6.4.1.1 Initialisation des changements et des versions

Cette section consiste à associer à chaque attribut une succession de versions élémentaires indivisibles qui ne se chevauchent pas temporellement et séparées par des changements comme le montre la frise du bas de la figure 6.7. Ces initialisations se font à partir des factoïdes associés à l'attribut (les éléments des ensembles C et V) en commençant par générer des changements élémentaires puis en inférant une version entre chaque paire de changements successifs. Le processus est synthétisé par l'algorithme C.5.

Pour former un ensemble Γ de changements élémentaires, on définit $\Omega = C \cup C_V$ avec $C_V = \left\{\gamma_1^\alpha, \gamma_1^\beta, \ldots, \gamma_m^\alpha, \gamma_m^\beta\right\}$ où $\forall i \in [\![1; m]\!]$, $\texttt{makesEffective}(\gamma_i^\alpha, v_i) \wedge \texttt{outdates}(\gamma_i^\beta, v_i)$ avec $v_i \in V$. C_V est un ensemble de changements fictifs appliqués aux versions ayant un temps valide. Pour le cas de la rue Gérard, toutes les versions, exceptée v_4, ont un temps valide donc sont associées chacune à deux changements fictifs indiquant leur mise en effectivité et leur péremption. Pour v_1, il y a deux changements γ_1^α et γ_1^β associés respectivement aux valeurs temporelles 1845 et 1849. L'attribut "géométrie" de la rue est ainsi associé à $\Omega = \left\{c_1, c_2, c_3, \gamma_1^\alpha, \gamma_1^\beta, \gamma_2^\alpha, \gamma_2^\beta, \gamma_3^\alpha, \gamma_3^\beta, \gamma_5^\alpha, \gamma_5^\beta\right\}$.

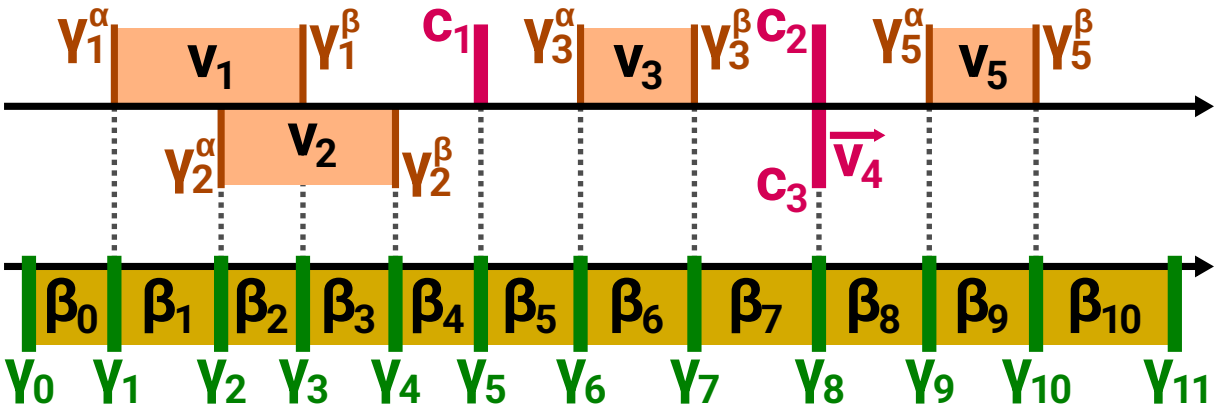


FIGURE 6.7 – Création de l'ensemble de changements Γ et de versions B (frise du bas) à partir de versions et de changements. La frise du haut montre l'agencement temporel des éléments de C et V pour l'attribut "géométrie" de la rue Gérard. ■ Version avec un temps valide, ■ Version élémentaire, → Version décrite dans un changement, ■ Changement, ■ Changement fictif, ■ Changement élémentaire.

À chaque changement $\omega_i \in \Omega$ est associé un instant temporel $t_i \in T_\Omega$, avec $T_\Omega = \{t_1, \ldots, t_{n+2m}\}$ l'ensemble des valeurs temporelles liées aux éléments de Ω. Si $\omega_i \in C$, alors t_i est obtenu via la valeur temporelle de l'événement e_i dont il dépend, c'est-à-dire tel que $\texttt{dependsOn}(\omega_i, e_i) \wedge \texttt{hasTime}(e_i, t_i)$. Dans le cas contraire, $\omega_i \in C_V$, alors ω_i fait partie d'une paire $(\gamma_i^\alpha, \gamma_i^\beta)$ associée à une version v_i d'attribut a_i d'une entité l_i. Les instants (t_i^α, t_i^β) sont extraits du temps valide tv_i de l_i : $\texttt{hasAttribute}(l_i, a_i) \wedge \texttt{hasAttributeVersion}(a_i, v_i) \wedge \texttt{hasTime}(l_i, tv_i) \wedge \texttt{hasBeginning}(tv_i, t_i^\alpha) \wedge \texttt{hasEnd}(tv_i, t_i^\beta)$. Ainsi, si $\omega_i = \gamma_i^\alpha$, alors sa valeur temporelle associée est t_i^α et t_i^β s'il est égal à γ_i^β. Pour tout sous-ensemble $S \subseteq \Omega$ tel que $\forall \omega, \omega' \in S, t_\omega = t_{\omega'}$, on définit un changement élémentaire $\gamma_S \in \Gamma$ associé à l'instant $t_S = t_\omega$. Ainsi, Γ est l'ensemble des changements obtenus en regroupant les éléments de Ω ayant lieu au même instant temporel. De même que Ω, il existe un ensemble de valeurs temporelles T_Γ. Par exemple, pour la rue Gérard, les changements c_2 et c_3 ayant lieu en 1979 appartiennent à la même classe d'équivalence, et donc à un même élément de Γ. En agrégeant de cette manière les changements de Ω, on obtient $\Gamma = \{\gamma_1, \ldots, \gamma_{10}\}$.

L'ensemble final de changements élémentaires est $\Gamma = \{\gamma_1, \ldots, \gamma_p\}$ avec $p \leq Card(\Omega) = 2m + n$ où l'existence de chaque élément découle de la fusion d'un ou plusieurs éléments de Ω. On y ajoute cependant deux changements $\gamma_{-\infty}$ et $\gamma_{+\infty}$ associés à deux instants infinis respectivement négatifs et positifs (que sont γ_0 et γ_{11} dans la figure 6.7). Enfin, on trie temporellement ces éléments en reliant deux changements successifs avec le triplet $\texttt{hasNextChange}(\gamma_i, \gamma_j)$. Pour $\gamma \in \Gamma$, si $\exists \gamma' = \underset{x \in \Omega \setminus \{\gamma\}, t(x) - t(\gamma) > 0}{\arg\min} t(x) - t(\gamma)$ où $t(x) \in T_\Gamma$ est la valeur temporelle associé au changement x alors $\texttt{hasNextChange}(\gamma, \gamma')$.

Les changements maintenant initialisés et ordonnés temporellement, il est aisé d'en faire de même avec les versions. Une version est initialisée entre deux changements successifs générés lors de l'étape précédente. Ainsi, un ensemble de versions B est créé tel que : $\forall \gamma_i, \gamma_j \in \Gamma, \texttt{hasNextChange}(\gamma_i, \gamma_j) \implies \exists \beta \in B, \texttt{AttributeVersion}(\beta) \wedge \texttt{makesEffective}(\gamma_i, \beta) \wedge \texttt{outdates}(\gamma_j, \beta)$. Par la suite, on trie les versions de la même manière. Deux versions $\beta_i, \beta_j \in B$ reliées à un même changement γ se succèdent : $\texttt{outdates}(\gamma, v_i) \wedge \texttt{makesEffective}(\gamma, v_j) \implies \texttt{hasNextVersion}(v_i, v_j)$.

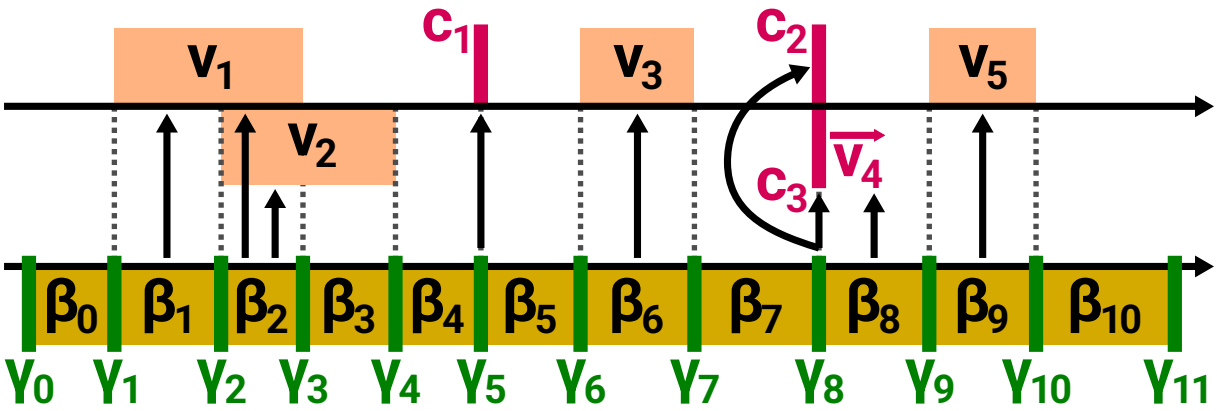


FIGURE 6.8 – Affiliation des changements et des versions pour l'attribut géométrie de la rue Gérard. ■ Version avec un temps valide, ■ Version élémentaire, ■ Changement, ■ Changement élémentaire, → `hasTrace`.

6.4.1.2 Affiliation des changements et des versions

Une fois les changements de Γ et les versions de B créés, il faut les affilier respectivement aux éléments existants de C et V : $\forall c \in C, \exists \gamma \in \Gamma, \texttt{hasTrace}(\gamma, c)$ (cela vaut aussi respectivement pour V et B). L'intérêt de l'affiliation est de lier les informations que l'on vient d'initialiser aux factoïdes. Ce processus est décrit par l'algorithme C.6.

L'affiliation des changements de C à ceux de Γ se fait lors de l'étape présentée dans la section précédente : si un changement γ de Γ est créé à partir d'un changement c de C alors $\texttt{hasTrace}(\gamma, c)$. Ainsi, pour la rue Gérard, les triplets générés sont $\texttt{hasTrace}(\gamma_5, c_1)$, $\texttt{hasTrace}(\gamma_8, c_2)$ et $\texttt{hasTrace}(\gamma_8, c_3)$. Afin de pouvoir faire de même pour les versions, il convient de relier tous les éléments de Ω aux changements élémentaires de Γ via la propriété `derives`, super-propriété de `hasTrace`. Afin de pouvoir faire de même pour les versions, il convient de relier tous les éléments de Ω aux changements élémentaires de Γ via la propriété `derives`, super-propriété de `hasTrace`. Ce prédicat permet de lier les changements inférés aux changements élémentaires dont ils dérivent et il est nécessaire car `hasTrace` relie uniquement des éléments aux factoïdes décrits dans les sources. Les changements fictifs (ceux appartenant à C_V) ne sont pas des factoïdes mais des éléments dérivés de factoïdes, générés lors du processus d'initialisation des ressources élémentaires. L'utilisation de `derives` permet ainsi de distinguer explicitement les changements décrits dans les sources de ceux qui sont inférés ou générés par le processus de reconstruction. Cette distinction est nécessaire pour les étapes ultérieures du processus, notamment dans la section 6.4. Formellement, il faut que $\forall \omega \in \Omega,\ \exists! \gamma \in \Gamma$ tel que $\texttt{derives}(\gamma, \omega)$, c'est-à-dire que chaque changement ω est dérivé par au moins un changement élémentaire γ. Dans le cas de la figure 6.7, $\texttt{derives}(\gamma_4, \gamma_2^{\beta})$, $\texttt{derives}(\gamma_6, \gamma_3^{\alpha})$ ou bien $\texttt{derives}(\gamma_5, c_1)$ font partie des dix triplets de l'exemple.

Pour les versions, l'affiliation est moins triviale car contrairement aux changements, il est possible qu'une version de V soit la trace de plusieurs versions de B. D'après la figure 6.8, β_1 et β_2 sont tracées par v_1 et β_2 et β_3 sont tracées par v_2. Pour tout $(v, \beta) \in V \times B$, la relation $\texttt{hasTrace}(\beta, v)$ est satisfaite si l'intersection de leurs intervalles temporels est non vide. Une manière de vérifier que cette intersection est non nulle consiste à vérifier que les deux versions partagent une borne temporelle en commun, ce qui revient à dire qu'elles possèdent chacune un changement de même type dont l'un dérive

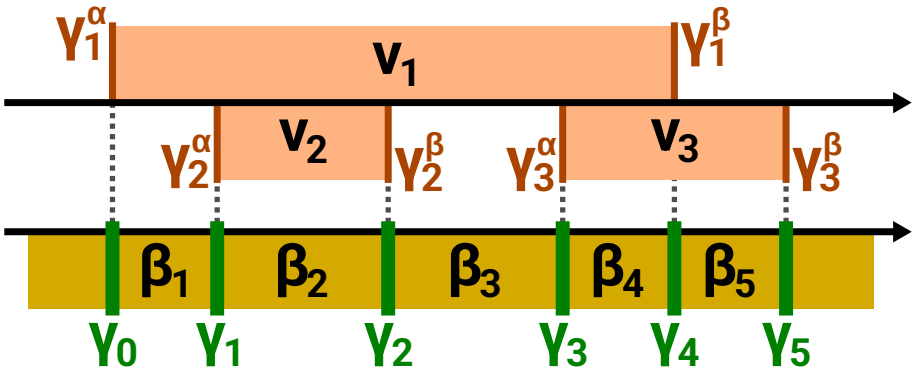


FIGURE 6.9 – Situation où la propagation de l'affiliation des versions est nécessaire. ■ Version avec un temps valide, ■ Version élémentaire, ■ Changement élémentaire, ■ Changement fictif.

de l'autre. Formellement, si c et γ sont des changements, on a :

$$\texttt{makesEffective}(c, v) \wedge \texttt{makesEffective}(\gamma, \beta) \wedge \texttt{derives}(\gamma, c) \implies \texttt{hasTrace}(\beta, v),$$

et de même pour `outdates` :

$$\texttt{outdates}(c, v) \wedge \texttt{outdates}(\gamma, \beta) \wedge \texttt{derives}(\gamma, c) \implies \texttt{hasTrace}(\beta, v).$$

Dans tous les cas, l'affiliation des versions se fait à partir de celles faites pour les changements. Concernant la rue Gérard, les triplets de type $\texttt{hasTrace}(\beta_i, v_j)$ sont générés pour les paires (β_1, v_1), (β_2, v_1), (β_2, v_2), (β_3, v_2), (β_6, v_3), (β_8, v_4) et (β_9, v_5).

Toutefois, si on prend l'exemple montré en figure 6.9, les premières conditions d'affiliation ne sont pas complètes. $\texttt{hasTrace}(\beta_1, v_1)$ en est déduit puisque $\texttt{derives}(\gamma_0, \gamma_1^\alpha)$ et γ_0 et γ_1^α rendent effectifs respectivement β_1 et v_1, alors . De manière similaire, $\texttt{hasTrace}(\beta_4, v_1)$ en est également déduit car $\texttt{derives}(\gamma_4, \gamma_1^\beta)$ et puisque γ_4 et γ_1^β périment respectivement β_4 et v_1 . À ce stade, il manque ainsi les triplets $\texttt{hasTrace}(\beta_2, v_1)$ et $\texttt{hasTrace}(\beta_3, v_1)$. C'est la raison pour laquelle il faut propager cette information de proche en proche. Autrement dit, on crée un triplet $\texttt{hasTrace}(\beta, v)$ pour une version β située « entre » une paire de versions (β_i, β_j) telles que $\texttt{hasTrace}(\beta_i, v)$ et $\texttt{hasTrace}(\beta_j, v)$.

On peut écrire : $\texttt{derives}(\beta_i, v) \wedge \texttt{derives}(\beta_j, v) \wedge \texttt{hasNextVersion}^+(\beta_i, \beta) \wedge \texttt{hasNextVersion}^+(\beta, \beta_j) \implies \texttt{hasTrace}(\beta, v)$. $\texttt{hasNextVersion}^+$ indique qu'il existe une chaîne de versions consécutives reliant β_i à β et β à β_j, c'est-à-dire que la relation `hasNextVersion` est appliquée de manière transitive.

6.4.2 Reconstitution de l'évolution des attributs à partir de leurs versions élémentaires

6.4.2.1 Suppression de versions lacunaires

Le découpage élémentaire présenté en sous-section 6.4.1 permet d'obtenir le découpage le plus fin des changements et versions qui existent dans l'ensemble des sources pour un attribut d'une entité géographique. À cette étape, on est capable de reconstituer l'évolution d'une entité géographique selon ce que disent les sources. Toutefois, pour des attributs, il existe des périodes durant lesquelles aucune information venant des sources n'est fournie. Dans le cas de la rue Gérard, les versions β_0, β_4, β_5, β_7 et β_{10} ne sont pas

tracées donc durant leur temps valide, on ne sait pas ce que vaut l'attribut "géométrie". Les supprimer en les fusionnant avec leur voisine (leur version précédente ou suivante) permettrait de combler les lacunes sur ces intervalles temporels. L'algorithme C.7 synthétise les traitements à faire pour cela. Différents critères doivent être pris en compte pour que les fusions se fassent sans perturber la cohérence des évolutions temporelles. Fusionner deux versions voisines implique de supprimer le changement qui les lie. Or, un changement tracé provient nécessairement d'une information sourcée : il n'est donc pas souhaitable de le supprimer. Ainsi, il sera impossible de fusionner deux versions voisines liées par un changement tracé. Par conséquent, les paires (β_4, β_5) et (β_7, β_8) de la figure 6.8 ne pourront être fusionnées.

Pour une version β non tracée, c'est-à-dire telle que $\nexists v \in V, \texttt{hasTrace}(v, \beta)$, trois cas peuvent se présenter en fonction de ses deux changements γ_{app} et γ_{dis} :

— aucun des deux changements n'est tracé :

$$\nexists c_{app}\, \texttt{hasTrace}(\gamma_{app}, c_{app}) \ \wedge \ \nexists c_{dis}\, \texttt{hasTrace}(\gamma_{dis}, c_{dis}) \tag{6.1}$$

— un seul des deux changements est tracé :

$$\begin{aligned} &\Big(\exists c_{app}\, \texttt{hasTrace}(\gamma_{app}, c_{app}) \ \wedge \ \nexists c_{dis}\, \texttt{hasTrace}(\gamma_{dis}, c_{dis})\Big) \\ &\vee \Big(\nexists c_{app}\, \texttt{hasTrace}(\gamma_{app}, c_{app}) \ \wedge \ \exists c_{dis}\, \texttt{hasTrace}(\gamma_{dis}, c_{dis})\Big) \end{aligned} \tag{6.2}$$

— les deux changements sont tracés :

$$\exists c_{app}, c_{dis}\, \texttt{hasTrace}(\gamma_{app}, c_{app}) \ \wedge \ \texttt{hasTrace}(\gamma_{dis}, c_{dis}) \tag{6.3}$$

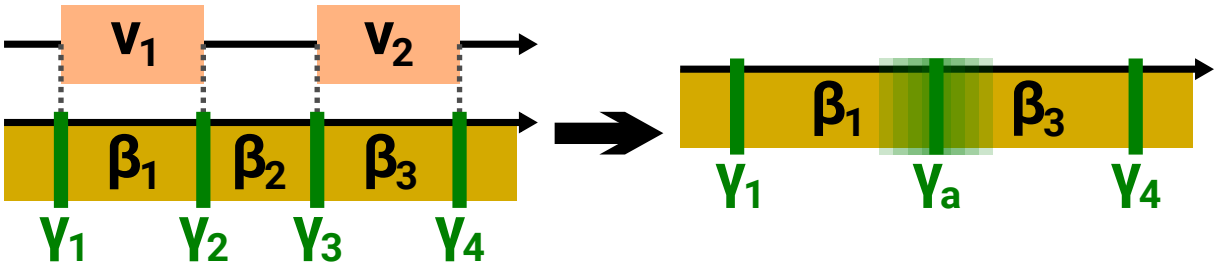
Le premier cas est illustré par la figure 6.10a. β_2 est une version dont les deux changements ne sont pas tracés donc on pourrait la fusionner avec β_1 ou β_3. Dans ce cas, il convient de supprimer β_2 et de fusionner ses deux changements. Autrement dit, il faut que le changement γ_2 qui rend obsolète β_1 (et rend effectif β_2) soit le même que γ_3 qui rend effectif β_3 (et périme β_2). Le processus se fait de la manière suivante :

$$\begin{aligned} &\texttt{AttributeVersion}(\beta) \ \wedge \ \texttt{makesEffective}(\gamma_M, \beta) \ \wedge \ \texttt{outdates}(\gamma_M, \beta_i) \\ &\wedge \ \nexists v_M\, \texttt{hasTrace}(\gamma_M, v_M) \ \wedge \ \texttt{outdates}(\gamma_O, \beta) \ \wedge \ \texttt{makesEffective}(\gamma_O, \beta_j) \\ &\wedge \ \nexists v_O\, \texttt{hasTrace}(\gamma_O, v_O) \implies \exists \gamma\, \texttt{AttributeChange}(\gamma) \ \wedge \ \texttt{outdates}(\gamma, \beta_i) \\ \wedge \ &\texttt{makesEffective}(\gamma, \beta_j) \ \wedge \ \nexists \beta\, \texttt{AttributeVersion}(\beta) \ \wedge \ \nexists \beta\, \texttt{makesEffective}(\gamma_M, \beta) \\ \wedge \ &\nexists \beta\, \texttt{outdates}(\gamma_O, \beta) \ \wedge \ \nexists \beta_i\, \texttt{outdates}(\gamma_M, \beta_i) \ \wedge \ \nexists \beta_j\, \texttt{makesEffective}(\gamma_O, \beta_j) \end{aligned} \tag{6.4}$$

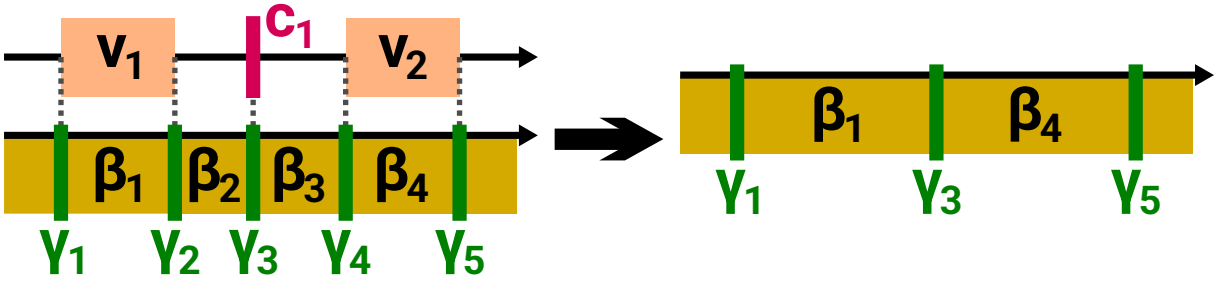
Étant donné que ce cas implique la fusion de deux changements γ_1 et γ_2 ayant des valeurs temporelles t_1 et t_2 distinctes (avec $t_1 < t_2$), le changement γ résultant de l'agrégation n'a pas d'instant t précis, on peut juste en déduire que $t_1 \leq t \leq t_2$.

Le deuxième cas, illustré par la figure 6.10b, prend place lorsqu'un seul des deux changements liés à une version est tracé. Dans ce cas, il convient de le fusionner avec sa voisine, avec laquelle il partage un changement non tracé. Dans l'exemple de la figure 6.10b, β_2 partage un changement non tracé avec β_1 tout comme β_3 avec β_4. La fusion se fait ainsi :

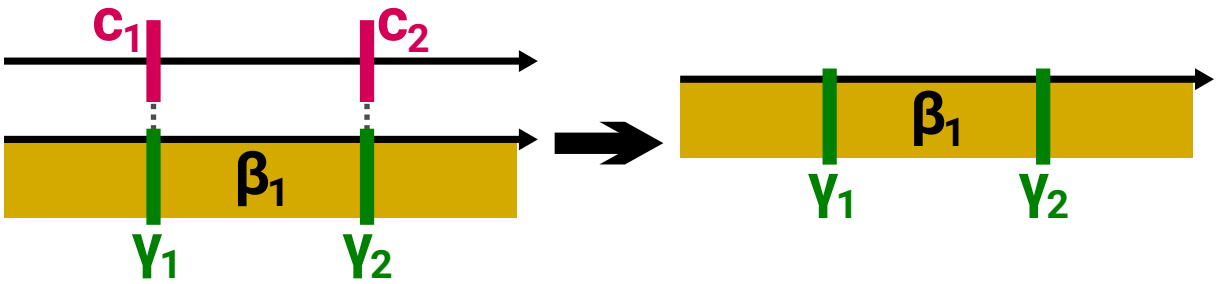
$$\begin{aligned} &\texttt{AttributeVersion}(\beta) \ \wedge \ \texttt{makesEffective}(\gamma_M, \beta) \ \wedge \ \texttt{hasTrace}(\gamma_M, c_M) \\ &\wedge \ \texttt{outdates}(\gamma_O, \beta) \ \wedge \ \texttt{makesEffective}(\gamma_O, \beta_i) \ \wedge \ \nexists\texttt{hasTrace}(\gamma_O, c_O) \\ \implies \ &\texttt{outdates}(\gamma_O, \beta_i) \ \wedge \ \nexists\texttt{AttributeVersion}(\beta) \ \wedge \ \nexists\texttt{makesEffective}(\gamma_M, \beta) \\ &\wedge \ \nexists\texttt{outdates}(\gamma_O, \beta) \ \wedge \ \nexists\texttt{makesEffective}(\gamma_O, \beta_i) \end{aligned} \tag{6.5}$$



(a) Version non tracée (β_2) délimitée par deux changements non tracés.



(b) Versions non tracées (β_2 et β_3) délimitées par un seul changement non tracé.



(c) Version non tracée (β_1) délimitée par deux changements tracés.

FIGURE 6.10 – Traitement des versions non tracées selon les changements auxquels il est lié. Version avec un temps valide, Version élémentaire, Changement, Changement élémentaire.

Pour que ce cas soit traité en totalité, il faut appliquer la formule en intervertissant les deux propriétés `outdates` et `makesEffective`.

Enfin, lorsque les deux changements d'une version sont tracés (comme le montre la figure 6.10c), c'est-à-dire lorsque :

$$\texttt{AttributeVersion}(\beta) \wedge \texttt{makesEffective}(\gamma_M, \beta) \wedge \texttt{hasTrace}(\gamma_M, c_M) \\ \wedge \texttt{outdates}(\gamma_O, \beta) \wedge \texttt{hasTrace}(\gamma_O, c_O)$$

alors il ne faut rien faire, car cela signifie qu'il existe une version non mentionnée dans les sources historiques utilisées.

Concernant l'attribut géométrique de la rue Gérard, les versions β_0 et β_{10} sont concernées par le premier cas. β_0 est supprimée et ses changements γ_0 et γ_1 sont fusionnés sous γ_a (voir figure 6.11). β_1 est ainsi rendue effective par γ_a dont la valeur temporelle est située entre $-\infty$ et 1845. Pour β_{10} qui est aussi supprimée, ses changements γ_{10} et γ_{11} fusionnent pour former γ_b. Ensuite, le deuxième cas s'applique aux versions β_4, β_5 et β_7. Tandis que β_4 est absorbée par β_3, les deux autres le sont par β_6. Enfin, aucune version n'est concernée par le dernier cas.

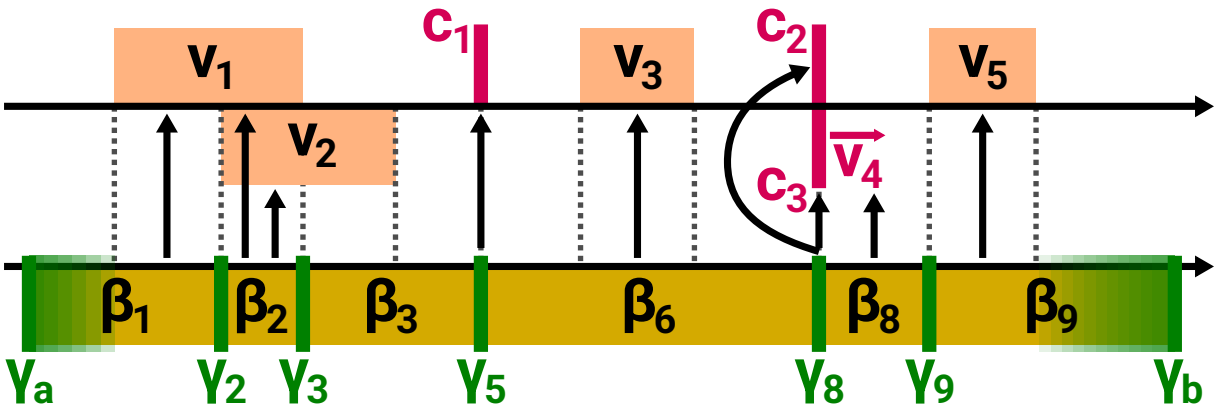


FIGURE 6.11 – Résultat de la suppression de versions lacunaires pour l'attribut géométrie de la rue Gérard. ■ Version avec un temps valide, ■ Version élémentaire, ■ Changement, ■ Changement élémentaire, → `hasTrace`.

6.4.2.2 Fusion des versions élémentaires similaires successives

L'étape finale consiste à fusionner les versions similaires en fonction de leur valeur. L'objectif de la section précédente était de générer une alternance de versions et de changements sans chevauchements temporels. Cette génération dépend de l'agencement initial. Le résultat présenté sur la frise chronologique en figure 6.11 dépend de la manière dont sont triées les données fournies par la frise du haut. Le but de cette dernière étape est de prendre en compte les valeurs associées aux versions : deux versions successives dont les valeurs sont similaires sont potentiellement à fusionner.

Avant de procéder aux étapes suivantes, il est nécessaire de comparer toutes les versions d'un attribut. Cette comparaison est indépendante des étapes précédentes, mais elle doit être déjà effectuée à ce stade. Concrètement, les versions sont comparées par leur valeur val, telle que $\texttt{versionValue}(v, val)$ afin de déterminer leur similarité. Si deux versions v_i et v_j sont jugées similaires, on instancie le triplet $\texttt{sameVersionValueAs}(v_i, v_j)$, sinon on instancie $\texttt{differentVersionValueFrom}(v_i, v_j)$. Les méthodes concrètes utilisées pour comparer les versions ne sont pas détaillées ici, car elles n'ont pas d'impact sur notre méthode. Elles peuvent dépendre du type d'attribut (par exemple nom ou géométrie) ou de paramètres personnalisables, tels que le score minimal de similarité pour considérer deux versions comme similaires. On pourrait se limiter à comparer uniquement les versions successives, c'est-à-dire comparer v_1 avec v_2, v_2 avec v_3, etc. Cependant, afin que la suppression éventuelle d'une version intermédiaire n'affecte pas la cohérence, toutes les versions sont comparées entre elles. Par exemple, si $\texttt{hasNextVersion}(v_1, v_2)$ et $\texttt{hasNextVersion}(v_2, v_3)$, on compare v_1 avec v_2 puis v_2 avec v_3. Si v_2 est ensuite supprimée, v_1 pointe directement vers v_3 ($\texttt{hasNextVersion}(v_1, v_3)$), et le résultat de la comparaison entre v_1 et v_3 est déjà disponible.

Une fois ces comparaisons effectuées, il est possible de procéder aux fusions si nécessaire. Comme pour l'étape précédente, on ne peut pas fusionner deux versions successives si elles sont séparées par un changement tracé. Les conditions pour fusionner deux versions β_i et β_j sont les suivantes :

— elles partagent un changement qui n'est pas tracé :

$$\exists\gamma \left(\neg\texttt{hasTrace}(\gamma, c) \wedge \texttt{makesEffective}(\gamma, \beta_i) \wedge \texttt{outdates}(\gamma, \beta_j)\right);$$

— les versions dont elles dérivent sont similaires, c'est-à-dire que leurs ensembles de

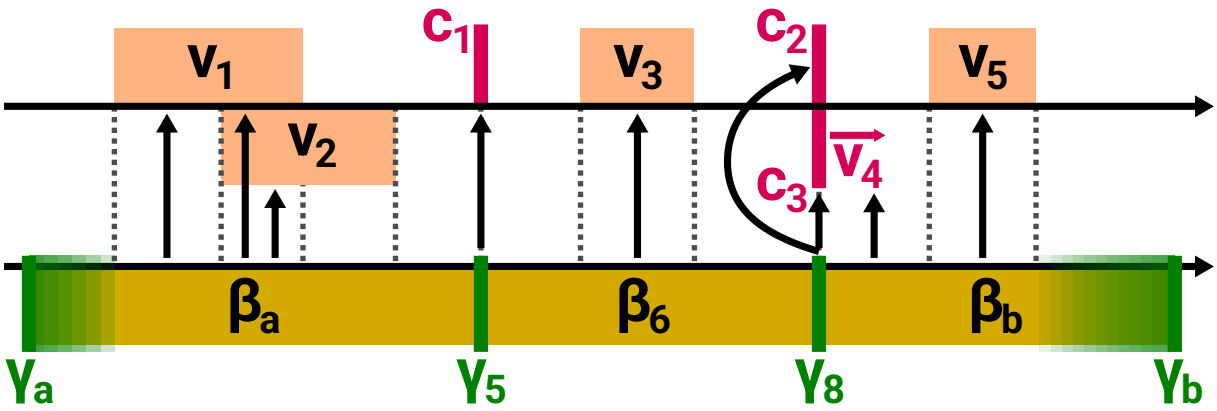


FIGURE 6.12 – Reconstitution finale de l'évolution temporelle après fusion des versions similaires pour l'attribut géométrie de la rue Gérard. ■ Version avec un temps valide, ■ Version élémentaire, ■ Changement, ■ Changement élémentaire, → `hasTrace`.

traces respectifs présentent au moins une correspondance de valeurs :

$$\forall\, v_i \in \texttt{Traces}(\beta_i),\ \exists\, v_j \in \texttt{Traces}(\beta_j) \quad \texttt{sameVersionValueAs}(v_i, v_j) \vee v_i = v_j,$$

où $\texttt{Traces}(\beta) = \{v \in V \mid \texttt{hasTrace}(\beta, v)\}$.

Lorsque ces conditions sont satisfaites, on génère un triplet temporaire de la forme $\texttt{toBeMergedWith}(\beta_i, \beta_j)$, où la propriété `toBeMergedWith` est transitive. Ainsi, pour chaque ensemble $\mathcal{B} \subseteq B$ de versions tel que : $\forall\, \beta_i, \beta_j \in \mathcal{B}$, $\texttt{toBeMergedWith}(\beta_i, \beta_j)$, une fusion est effectuée. Cette fusion conduit à la création d'une nouvelle version β délimitée par deux changements existants :

- le changement d'apparition c_{app} correspond à celui associé à la version la plus ancienne de l'ensemble, soit la version $b \in \mathcal{B}$ telle que : $\nexists\, x\ \texttt{hasNextVersion}(x, b)$;
- le changement de disparition c_{dis} correspond à celui associé à la version la plus récente de l'ensemble, soit la version $b \in \mathcal{B}$ telle que : $\nexists\, x\ \texttt{hasNextVersion}(b, x)$.

Si on reprend l'exemple de la rue Gérard, les comparaisons indiquent des similarités entre les versions v_1, v_2 et v_3 ainsi qu'entre v_4 et v_5. En appliquant les conditions de fusion, les versions tracées par v_1 et v_2 sont à fusionner ensemble, soit β_1, β_2 et β_3, via la version β_a (voir figure 6.12). Bien que similaire à v_2, la version v_3 n'est pas prise en compte ici car il existe le changement c_1 qui le sépare. Enfin, β_8 et β_9 sont à fusionner pour former β_b puisqu'elles sont respectivement tracées par v_4 et v_5 et que $\texttt{sameVersionValueAs}(v_4, v_5)$.

Grâce à cette approche, nous sommes désormais capables de reconstituer l'évolution détaillée de chaque attribut d'une entité géographique. En complément, il est possible de reconstituer l'historique complet d'une entité en combinant les historiques de toutes les ressources dont elle dépend. Il est ainsi possible de décrire son évolution comme étant la succession de versions d'entités comme les approches classiques le font.

Le schéma de la figure 6.13 illustre cette reconstitution à travers cinq historiques distincts :

- l'historique de la voie, composé d'états spatio-temporels représentant son existence ou sa non-existence ;
- l'historique de la relation `Within` indiquant que la rue est dans Gentilly ;
- l'historique de la relation `Within` indiquant que la rue est dans Paris ;
- l'historique de l'attribut de nom ;
- l'historique de l'attribut de géométrie.

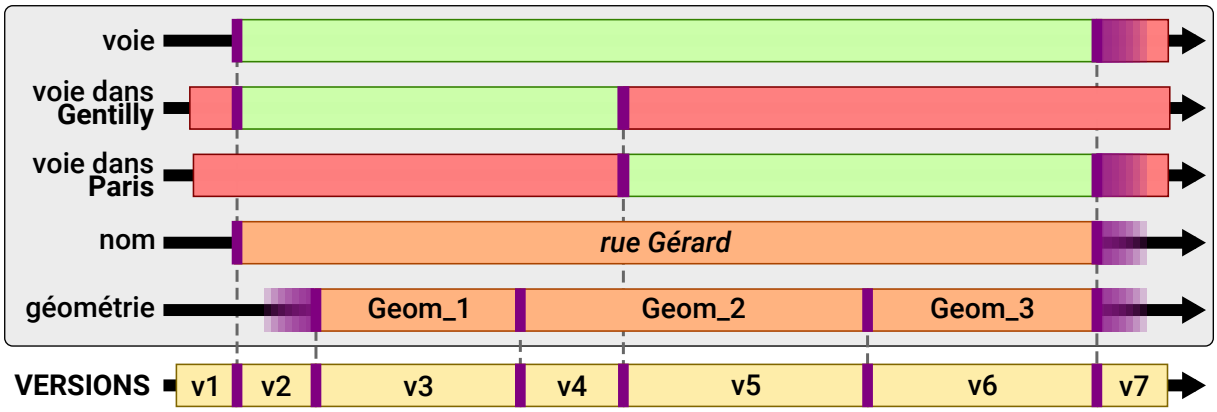


FIGURE 6.13 – Reconstitution des historiques des différents attributs de la rue Gérard : existence de la voie, relations spatiales (`Within` pour Gentilly et Paris), nom et géométrie. La frise combinée en bas représente la succession d'états obtenue par combinaison des cinq historiques. ■ État de non-existence, ■ État d'existence, ■ Version d'attribut, ■ État d'entité, ■ Changement.

Ces différents historiques peuvent être combinés pour produire une frise temporelle synthétique, où chaque état est obtenu par combinaison des informations issues des cinq historiques. Ainsi, les versions v_1 et v_7 indiquent que la rue n'existe pas pendant leur période de validité. La version v_2 correspond à une voie existante, portant le nom *Rue Gérard* et située dans Gentilly, mais sans géométrie connue. Les versions v_3 et v_4 sont similaires à v_2, mais diffèrent par leur géométrie respective (`Geom_1` et `Geom_2`). La version v_5 correspond à la rue nommée *Rue Gérard*, avec la géométrie `Geom_2`, et localisée dans Paris, tandis que v_6 est similaire à v_5 mais avec une géométrie `Geom_3`.

Cependant, pour garantir la fiabilité de ce processus, il est nécessaire d'évaluer la cohérence des données générées par la méthode que nous venons de présenter.

Conclusion

Ce chapitre a présenté la méthode de construction du graphe de connaissances géo-historique que nous proposons, depuis la représentation des données issues des sources jusqu'à la reconstitution de l'évolution temporelle des entités et de leurs attributs. L'ontologie fournit le cadre conceptuel garantissant la cohérence et la réutilisabilité du modèle, tandis que les différentes étapes de traitement — identification des observations équivalentes, intégration des entités, propagation des liens et déduction des changements — permettent de construire un graphe de faits temporellement et sémantiquement structuré.

Toutefois, la conception d'une ontologie ne suffit pas à garantir la qualité du graphe obtenu : il est également nécessaire d'évaluer la méthode de peuplement elle-même. Cette évaluation doit porter à la fois sur la capacité du processus à intégrer correctement des données hétérogènes et fragmentaires sur la pertinence des inférences produites (fusion d'entités, détection des changements, reconstitution des trajectoires d'attributs). Le chapitre suivant est ainsi consacré à l'évaluation expérimentale de la méthode proposée et de l'ontologie en y testant les questions de compétence, afin d'en mesurer la robustesse, la précision et les limites à partir d'un cas d'étude concret.

Chapitre 7

Mise en œuvre, résultats et évaluation

Ce chapitre présente la mise en œuvre de la méthode de reconstruction de l'évolution des entités géographiques à partir de sources hétérogènes et fragmentaires, appliquée au quartier du Faubourg Saint-Antoine, en lien avec l'ontologie *PeGazUs*. Nous y détaillons également l'évaluation de cette ontologie et de la méthode de peuplement présentée au chapitre précédent.

Le chapitre s'appuie majoritairement sur les travaux décrits par Bernard *et al.* (2025b). Ce chapitre commence par la présentation du Faubourg Saint-Antoine dans la section 7.1 ainsi que des différentes sources utilisées, qu'elles soient cartographiques, textuelles ou numériques. La section 7.2 détaille la mise en application de la méthode de peuplement, mise en œuvre à l'aide de scripts Python et de requêtes SPARQL.

Par la suite, la section 7.3 s'attache à vérifier la cohérence du graphe en distinguant les contraintes sur les ressources et celles sur les propriétés, avant de présenter l'évaluation de l'ontologie dans la section 7.4 à travers ses différents modelets. Enfin, la section 7.5 détaille l'évaluation de la méthode de peuplement elle-même, en présentant la stratégie d'évaluation et les résultats obtenus.

7.1 Le Faubourg Saint-Antoine comme cas d'étude

Nous testons la méthode de peuplement sur un ensemble de jeux de données issus de sources multiples décrivant les rues et numéros d'adresses du quartier du Faubourg Saint-Antoine, sur une période allant de la fin du XVIIIe siècle à nos jours. Ce quartier s'étend partiellement sur les 4e, 11e, 12e arrondissements de Paris comme le montre la figure 7.1.

Cette zone a été choisie afin de bénéficier d'une couverture par divers plans historiques de la ville de Paris. Des arrondissements périphériques, tels que le 13e, auraient pu être envisagés, mais leur emprise étant située en dehors de Paris avant 1860, nous ne disposons pas d'une description complète de leur territoire avant cette date (par exemple, ils ne figurent pas sur les plans de Jacoubet et Verniquet). Le Faubourg Saint-Antoine offre un ensemble de situations d'évolution urbaines variées (voir figure 7.2). Il existe des parties du quartier où la structure urbaine a peu évolué : la figure 7.2a illustre la cas de la place des Vosges (anciennement place Royale) qui constitue un exemple parfait de stabilité morphologique. De l'autre côté de la place de la Bastille, l'ancien hospice des Orphelins (ou ancien hôpital Trousseau) a été démoli au début du XXe siècle. Cette disparition a entraîné une transformation ponctuelle du quartier avec la création de nouveaux bâtiments et voies autour de l'actuel square Trousseau (voir figure 7.2b). Le sud du quartier, autour

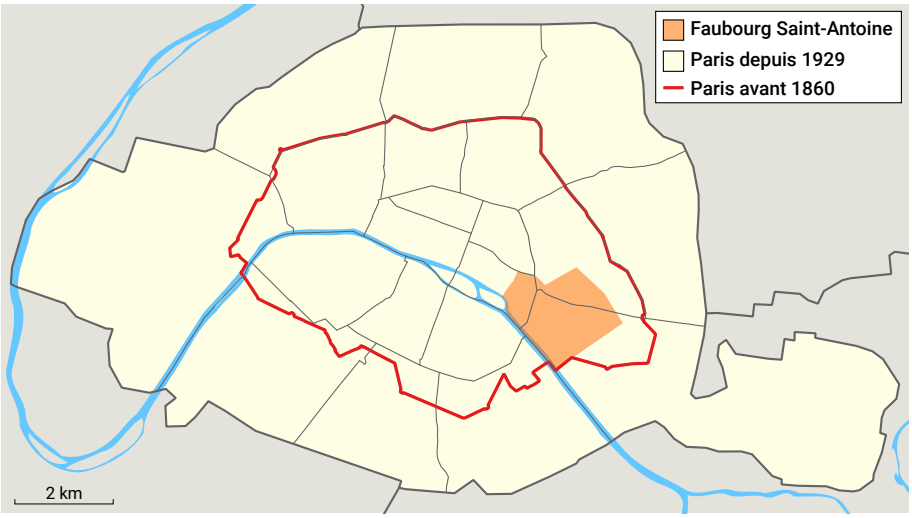


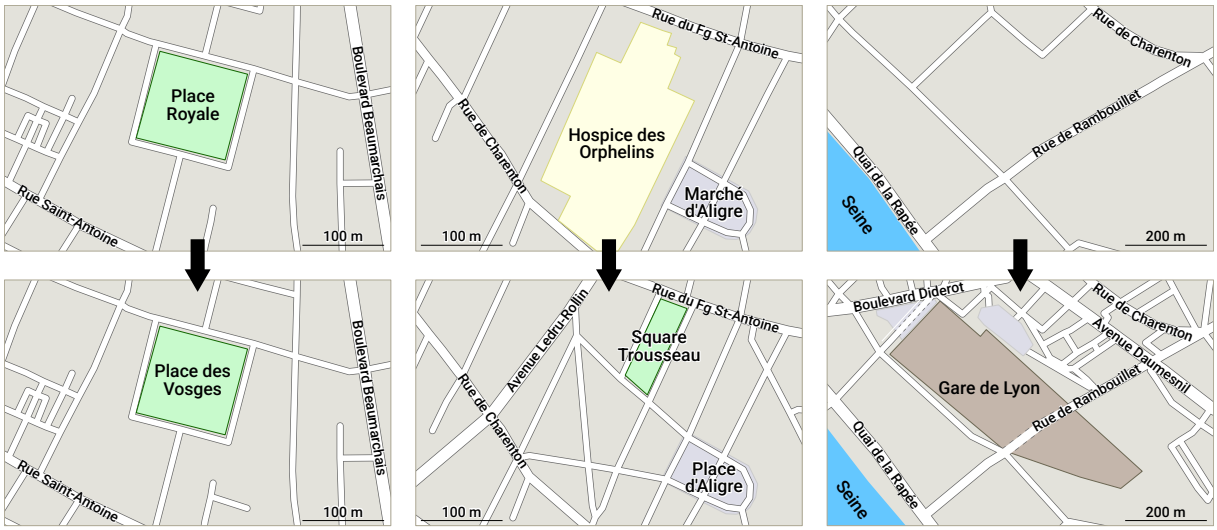
FIGURE 7.1 – Localisation du quartier du Faubourg Saint-Antoine au sein de Paris.

de la gare de Lyon, constitue un cas de rupture dans l'organisation spatiale du territoire. Contrairement aux transformations ponctuelles autour de l'ancien hôpital Trousseau, les évolutions sont le résultat de plusieurs phases d'aménagement, liées notamment au développement des infrastructures ferroviaires, qui se sont succédé entre le XIXe et le XXe siècle (voir figure 7.2c). Nous avons également souhaité une emprise suffisamment étendue — au-delà de quelques rues seulement — afin de garantir la diversité des situations d'évolution observables et de disposer d'un volume de données conséquent. Il existe des territoires dans Paris avec des évolutions conséquentes mais ces transformations se sont produites de manière soudaine et homogène, limitant ainsi la variété des cas d'évolution. Cette sélection nous permet donc de démontrer que la méthodologie proposée est applicable à un territoire de taille significative et ayant connu des évolutions nombreuses et variées sur la période étudiée.

7.1.1 Sources cartographiques

Les sources cartographiques utilisées comprennent principalement des plans anciens à grande échelle décrivant l'état du territoire parisien aux différentes époques de leur publication. Parmi ceux-ci, on compte l'*Atlas National de la ville de Paris* d'Edme Verniquet publié entre 1791 et 1799 (Verniquet, 1795), l'*Atlas Général de la ville et des faubourgs de Paris* de Théodore Jacoubet publié en 1836 (Jacoubet, 1836), le *Plan de Paris fortifié et des communes environnantes : 1849* par Eugène Andriveau-Goujon (Andriveau-Goujon, 1849), ainsi que l'édition de 1888 de l'*Atlas municipal des vingt arrondissements de la ville de Paris* (Alphand et Fauve, 1888). Ces documents fournissent des états spatio-temporels des entités géographiques représentées, notamment l'emprise et le nom des voies, ainsi que l'existence et la localisation des numéros d'immeubles. Un exemple est fourni en figure 7.3.

Les cadastres napoléoniens dont celui de Paris ainsi que ceux d'autres communes limitrophes, sont également utilisés. En effet, avant l'annexion de 1860, une partie du territoire aujourd'hui parisien appartenait à ces communes, notamment Gentilly, ce qui rend leur documentation cadastrale indispensable pour reconstituer l'évolution spatiale et administrative du territoire. Constitué d'états de sections, de plans parcellaires et de matrices cadastrales, ce corpus offre également une base historique, décrivant notamment les parcelles et leurs évolutions (fusion, scission, changement de nature ou de propriétaire).



(a) Un cas de stabilité morphologique autour de la place des Vosges.

(b) Un cas de transformation ponctuelle autour de l'ancien hospice des Orphelins.

(c) Un cas de ruptures majeures autour de la gare de Lyon.

FIGURE 7.2 – Diversité des transformations urbaines dans le quartier du Faubourg Saint-Antoine entre le début du XIXe siècle (plan de Jacoubet) et 2026 (OpenStreetMap).

Les données issues de ces sources cartographiques sont généralement fournies sous forme de fichiers GeoJSON, contenant des polylignes ou des polygones représentant les emprises des voies ainsi que des points localisant les numéros d'immeubles. Les étapes de numérisation, de vectorisation et de géoréférencement qui ont été appliquées sur ces sources papier pour produire les données numériques ne font pas partie du cadre de ce travail. Nous utilisons directement les données obtenues, telles qu'elles ont été produites par des travaux antérieurs appliquant ces étapes.

7.1.2 Sources textuelles

Les sources textuelles se divisent en sources officielles émanant de la Ville de Paris ou de l'État, et en autres sources historiques ou commerciales. Les publications administratives de la Ville de Paris, telles que les comptes rendus du Conseil ou le Bulletin Officiel, fournissent des informations sur les dates de création et de renommage des voies. D'autres documents historiques, comme le *Dictionnaire topographique, étymologique et historique des rues de Paris* de La Tynna (Duverneuil et de La Tynna, 1801), issu d'une série d'annuaires publiés annuellement au début du XIXe siècle sous le titre *Almanach du commerce de Paris*, ou la *Topographie historique du Vieux Paris* d'Adolphe Berty (Berty *et al.*, 1866), apportent des détails sur la localisation des voies, leurs relations spatiales, leurs anciens noms, ainsi que des informations sur certains bâtiments. Les annuaires de commerce, quant à eux, attestent de l'existence de numéros d'immeubles et de voies. Pour extraire automatiquement ces informations textuelles, des technologies telles que la reconnaissance optique de caractères (OCR) sont utilisées, suivies de la reconnaissance d'entités nommées (NER), particulièrement efficace sur des documents structurés. Pour les documents à descriptions plus libres et contextuelles, des grands modèles de langage (LLM) peuvent être employés afin d'améliorer l'extraction d'informations implicites.

À partir des données extraites par OCR, il est nécessaire d'identifier et d'extraire les informations textuelles pertinentes, puis de les mettre en cohérence avec l'ontologie



FIGURE 7.3 – Exemple de données issues de deux sources cartographiques (atlas de Jacoubet et atlas municipal de 1888) autour de la place des Vosges. Les points indiquent les numéros d'immeuble et les lignes les voies de communication.

PeGazUs. Cette étape pourrait être réalisée à l'aide de LLM, mais elle reste complexe en raison de plusieurs facteurs :

- **Diversité des sources** : les textes proviennent de documents hétérogènes, avec des structures, des formulations et un vocabulaire très variés, ce qui rend l'extraction directe difficile.
- **Complexité de l'ontologie** : pour être correctement intégrées, les données doivent suivre une structure complexe puisqu'elles doivent être représentées sous forme de triplets en respectant le vocabulaire de *PeGazUs*. Cette représentation impliquant parfois le recours à des nœuds blancs, leur manipulation peut s'avérer délicate pour les LLM qui peuvent aussi omettre certaines informations ou ne pas respecter strictement la structure attendue.

7.1.3 Sources numériques

Les sources numériques et structurées rassemblent des données facilement convertibles en triplets RDF conformes à l'ontologie PeGazUs à l'aide de scripts dédiés. Parmi celles-ci, on trouve les fichiers officiels tels que les dénominations des emprises des voies actuelles et caduques publiées en open data par la Ville de Paris, qui fournissent notamment l'historique non structuré des noms, les emprises valides, des relations spatiales entre voies, des dates d'apparition ou de dissolution, ainsi que des codes officiels (FANTOIR, Ville de Paris). Parmi les sources officielles de l'État, le fichier FANTOIR, remplacé depuis 2023 par le fichier TOPO, fournit des informations sur les dates d'existence des voies, tandis que la Base Adresse Nationale française (BAN) contient exclusivement les localisations géographiques et d'autres informations associées aux numéros d'immeubles. Un extrait de la BAN est présenté dans le tableau 7.1 [1].

1. Les positions des adresses (fournies par `lat` et `lon`) dans la BAN n'ont pas de règle unique définie. Il est indiqué pour chaque adresse la manière dont est fixée cette position (par rapport au point de livraison postal, à la porte d'entrée, au centre de la parcelle, etc).

num	rep	nom_voie	cp	nom_com	lat	lon
5		Boulevard Saint-Martin	75003	Paris 3e	2.361407	48.867933
134		Boulevard Raspail	75006	Paris 6e	2.309998	48.860110
4		Rue Regnard	75006	Paris 6e	2.338073	48.849885
46		Rue Gérard	75013	Paris 13e	2.352083	48.829464
10	b	Rue de Cotte	75012	Paris 12e	2.377551	48.849106
10		Rue des Mâconnais	75012	Paris 12e	2.388908	48.832740
48	bis	Rue de Rivoli	75004	Paris 4e	2.354480	48.856900
1		Place des Vosges	75004	Paris 4e	2.364309	48.855132
3		Boulevard de Charonne	75011	Paris 11e	2.392378	48.857381

Tableau 7.1 – Extrait d'un fichier de la BAN (Base Adresse Nationale) pour la ville de Paris, incluant les numéros d'adresse, compléments éventuels, noms de voies, codes postaux, communes et coordonnées géographiques (latitude, longitude).

Les données collaboratives issues d'OpenStreetMap (OSM) proposent les linéaires des voies ainsi que la localisation des numéros d'immeubles associés, accessibles via plusieurs outils d'extraction, dont l'API Overpass qui requiert une maîtrise du langage Overpass QL. Bast *et al.* (2021) introduisent une alternative avec *osm2rdf*, un outil qui permet de convertir les données d'OSM en triplets RDF. Ils présentent aussi un point d'accès SPARQL[2] facilitant l'accès aux relations spatiales implicites. Un exemple d'extraction d'informations concernant des numéros d'immeuble de Paris est fourni par le Listing 7.1.

```
SELECT DISTINCT
    ?houseNumberId ?houseNumberLabel ?houseNumberGeomWKT
    ?streetId ?streetName
    ?arrdtId ?arrdtName ?arrdtInsee
WHERE {
  ?selectedArea osmkey:wikidata "Q90" ; ogc:sfContains ?houseNumberId .
  ?houseNumberId osmkey:addr:housenumber ?houseNumberLabel ;
        geo:hasGeometry ?houseNumberGeom .
  ?houseNumberGeom geo:asWKT ?houseNumberGeomWKT .
  ?streetId osmkey:type "associatedStreet" ;
      osmrel:member ?member ;
      osmkey:name ?streetName.
  ?member osmrel:member_role "house" ;
      osmrel:member_id ?houseNumberId.
  ?arrdtId ogc:sfContains ?houseNumberId ;
      osmkey:name ?arrdtName ;
      osmkey:ref:INSEE ?arrdtInsee ;
      osmkey:boundary "administrative" ;
      osmkey:admin_level "9"^^xsd:int .
}
```

Listing 7.1 – Requête SPARQL fusionnée permettant de sélectionner les identifiants OSM des numéros d'immeubles de Paris (dont l'identifiant wikidata est Q90), leur voie d'appartenance et l'arrondissement, leur label et leur géométrie WKT.

Wikidata fournit des données structurées sur les voies, comprenant leur nom officiel, leur localisation, leur type, leurs dates de création et de suppression ainsi que leurs codes

2. https://qlever.cs.uni-freiburg.de/osm-planet

d'identification. Ces informations peuvent être extraites via le point d'accès SPARQL de Wikidata[3]. De plus, Wikipedia propose des descriptions historiques des voies et des références vers d'autres sources utiles. Ces jeux de données peuvent être intégrés à notre graphe de connaissances grâce à un *mapping* adapté, même si l'hétérogénéité des sources rend cette tâche complexe et chronophage. Alors que certains jeux de données correspondent à des *snapshots* du territoire à des dates précises, d'autres sources documentent l'évolution des voies dans le temps, comme les *Dénominations des emprises des voies actuelles*[4] et les *Dénominations caduques des voies*[5] publiées par la Ville de Paris et justifient donc la mise en œuvre de notre méthode de peuplement.

7.2 Mise en application de la méthode de peuplement

7.2.1 Choix technologiques pour la mise en œuvre

La mise en œuvre de la méthode de peuplement repose sur l'utilisation du langage Python pour convertir les données en triplets RDF conformes à PeGazUs et peupler ainsi des graphes de factoïdes puis de requêtes SPARQL pour implémenter les règles de peuplement du graphe de faits présentées au chapitre 6. Les procédures ont été conçues de manière à être les plus générales possible, afin de faciliter l'intégration de nouvelles sources de données sans nécessiter la réécriture complète des scripts. Cette approche est essentielle, car les sources exploitées sont hétérogènes et la génération de descriptions complexes sous forme de triplets rendrait coûteuse en temps de développement l'adaptation d'un script spécifique pour chaque type de données. De plus, la chaîne de traitement est développée de façon modulaire et paramétrable : de nombreux paramètres peuvent être ajustés en fonction des besoins. Les étapes exécutées dans Python sont indépendantes les unes des autres, ce qui permet de mettre à jour ou de modifier certaines procédures sans impacter l'ensemble du processus. Le code implémentant la méthode est disponible en ligne : https://github.com/charlybernard/pegazus-extension.

L'ensemble des opérations est centralisé au sein d'un notebook codé en Python, qui fait appel à un ensemble de fonctions dédiées à l'application successive des différentes étapes. Ce script permet non seulement de générer les triplets et de les intégrer dans le triplestore GraphDB via une API, mais aussi de construire dynamiquement des requêtes SPARQL utilisées pour effectuer des opérations de mise à jour.

7.2.2 Transformation des données selon l'ontologie

7.2.2.1 Justification d'un format intermédiaire de données

Pour des données structurées et homogènes — comme celles issues d'*OpenStreetMap* (OSM) ou de la *Base Adresse Nationale* (BAN) — la conversion initiale vers le modèle RDF est largement automatisable. Pour chaque source, nous définissons un *mapping* : une mise en correspondance permettant de générer automatiquement et de manière reproductible un ensemble de triplets décrivant chaque élément des données initiales conformément à l'ontologie *PeGazUs*.

Cependant, cette génération tend à créer des ressources indépendantes et distinctes qui référencent souvent les mêmes entités (par exemple, la même voie ou la même commune).

3. https://query.wikidata.org
4. https://opendata.paris.fr/explore/dataset/denominations-emprises-voies-actuelles
5. https://opendata.paris.fr/explore/dataset/denominations-des-voies-caduques

Ainsi, pour la BAN, il existera autant d'instances de « Rue Gérard » ou de « Paris » que d'adresses qui les mentionnent. Pour obtenir un graphe cohérent, il serait alors nécessaire de fusionner les entités similaires à l'aide de requêtes SPARQL — ou d'opérations de liage — afin d'identifier et de regrouper les ressources décrivant une même entité géographique.

Nous n'avons toutefois pas retenu cette approche pour deux raisons principales. Premièrement, la génération directe de triplets à partir d'un *mapping* RDF suppose une bonne maîtrise de l'ontologie, qui est complexe, et impose de créer un *mapping* spécifique à chaque source. Deuxièmement, cette méthode n'est réellement adaptée qu'à des données structurées de manière homogène — ce qui est le cas pour OSM et la BAN, mais beaucoup moins pour les adresses historiques, souvent issues de documents hétérogènes et partiellement structurés. Pour cette raison, nous avons adopté une approche plus générique : les données issues de chaque source sont d'abord converties dans un format JSON intermédiaire, à partir duquel un processus automatisé génère les triplets RDF. Ce format JSON offre une certaine flexibilité, permettant de représenter de manière uniforme des données de nature et de granularité variées, tout en facilitant leur transformation ultérieure en RDF conforme à notre ontologie. Ainsi, si on désire mettre à jour l'ontologie, il serait uniquement nécessaire mettre à jour ce processus automatisé.

7.2.2.2 Description du format intermédiaire

Nous avons conçu deux formats JSON : l'un pour décrire des états spatio-temporels, l'autre pour des événements. Un exemple du formats intermédiaire dédié aux états spatio-temporels est illustré par le Listing 7.2. Le format JSON est structuré en trois parties principales :

1. `landmarks`, qui répertorie l'ensemble des entités géographiques distinctes mentionnées dans la source, telles que les numéros, les voies, les codes postaux ou les communes. Il est possible de leur ajouter des attributs via la clé `attributes`, associée à un dictionnaire décrivant la version qui y est associée (voir Listing 7.2) ;

2. `relations`, qui décrit les liens entre ces entités, par exemple un numéro appartenant à une voie (`Belongs`) ou une voie située dans une commune (`Within`) ;

3. `addresses`, qui regroupe les adresses complètes en les définissant comme une suite ordonnée de relations convergeant vers une entité cible ; la suite ordonnée est spécifiée par la liste associée à la clé `segments`, et la cible par l'élément correspondant à la clé `target`.

De plus, pour chaque élément des trois parties, il est également possible d'associer un lien de provenance grâce à la clé `provenance`. Ce lien peut être spécifié sous la forme d'un dictionnaire : `"provenance": {"uri": "http://exemple.org/provenance"}`. Un fichier JSON décrit un ensemble de données issues d'une même source. L'ensemble des provenances est associé à la source décrite par la clé `source` comme l'illustre le Listing 7.2). Enfin, ce format permettant de décrire des états spatio-temporels, il faut indiquer les temps valides. Par défaut, chaque élément d'une source a un temps valide similaire à celui de la source qui est associé à la clé `time` comme montre le Listing 7.2. Pour les éléments qui ont des temps valides distincts de la source, un dictionnaire spécifique peut être défini pour un élément de `landmarks`, `relations` ou `addresses`.

```
{
  "landmarks":[
    {"id":"lm1","label":"58","type":"street_number", "attributes": {
        "name": {"value": "58"},
        "geometry": {"value":"POINT(2.351 48.828)", "datatype":"wkt_literal"}
    }},
    {"id":"lm2", "label":"60", "type":"street_number"},
    {"id":"lm3", "label":"Rue Gérard", "type":"thoroughfare", "lang":"fr"},
    {"id":"lm4", "label":"75001", "type":"postal_code_area"},
    {"id":"lm5", "label":"Paris", "type":"municipality", "lang":"fr"}
  ],
  "relations":[
    {"id":"rel1", "type":"belongs", "locatum":"lm1", "relatum":["lm3"]},
    {"id":"rel2", "type":"within", "locatum":"lm1", "relatum":["lm4"]},
    {"id":"rel3", "type":"within", "locatum":"lm1", "relatum":["lm5"]},
    {"id":"rel4", "type":"belongs", "locatum":"lm2", "relatum":["lm3"]},
    {"id":"rel5", "type":"within", "locatum":"lm2", "relatum":["lm4"]},
    {"id":"rel6", "type":"within", "locatum":"lm2", "relatum":["lm5"]}
  ],
  "addresses":[
    {"id":"addr1", "label":"58 rue Gérard, 75001 Paris", "lang":"fr", "target":"lm1",
    ↪  "segments":["rel1", "rel2", "rel3"]},
    {"id":"addr2", "label":"60 rue Gérard, 75001 Paris", "lang":"fr", "target":"lm2",
    ↪  "segments":["rel4", "rel5", "rel6"]}
  ],
  "time": {
    "start":{"stamp": "2024-01-01", "calendar": "gregorian",  "precision": "day"},
    "end":{"stamp": "2025-01-01", "calendar": "gregorian",  "precision": "day"}
  },
  "source": {
    "label": "BAN", "lang": "fr", "publisher": {"label": "DINUM / ANCT / IGN"}
  }
}
```

Listing 7.2 – Exemple de description d'adresses en JSON compact pour des données issues de la BAN. Les attributs des `landmarks` et les provenances des éléments ne sont pas tous indiqués dans cet exemple.

Le second format JSON, destiné à décrire les événements, est assez similaire au format des états spatio-temporels. Un exemple est illustré par le Listing 7.3. Dans ce cas, un fichier décrit l'ensemble des événements d'une source, et contient essentiellement deux parties :

— `source` décrit la source ;
— `events` fournit une liste d'événements présents dans la source.

Chaque événement possède les clés suivantes :

— `time` : instant de l'événement, avec le `stamp`, le `calendar` et la `precision` ;
— `label` : énoncé de l'événement ;
— `lang` : langue de l'énoncé ;
— `landmarks` : liste des entités impliquées dans l'événement ;
— `relations` : liste des relations impliquées dans l'événement ;
— `provenance` : lien de provenance de l'événement.

Pour décrire les changements dépendant de l'événement, on utilise pour les `landmarks` une liste de changements associée à la clé `changes`. Chaque changement appliqué à l'entité est de la forme `{"on":"landmark", "type":"appearance"}` (ou `"disappearance"`). Pour un changement lié à un attribut, la forme est `{"on":"attribute", "attribute":` *`type`*`, "makes_effective": [...], "outdates": [...]}`, avec au moins un des deux

derniers champs renseignés, *type* donne le type de l'attribut. Pour les `relations`, les changements sont définis de manière similaire.

```
{
  "source": { "label": "Wikipedia", "lang": "fr" },
  "events": [
    {
    "label": "Par arrêté [...] rue du Père Guérin", "lang": "fr",
    "time": {"stamp":"1978-08-30", "calendar":"gregorian", "precision":"day"},
      "landmarks": [
        {
          "id": 1, "label": "rue du Père Guérin", "lang": "fr",
          "type": "thoroughfare",
          "changes": [
            {"on":"landmark", "type":"appearance"},
            {"on":"attribute", "attribute":"geometry"},
            {"on":"attribute", "attribute":"name", "makes_effective":[{"value":"rue
            ↪  du Père Guérin", "lang":"fr"}]}
          ]
        }, {
          "id": 2, "label": "rue Gérard", "lang": "fr", "type": "thoroughfare",
          "changes":[{"on":"attribute", "attribute":"geometry"}]
        }
      ],
      "relations": [{
          "type":"touches", "locatum":1, "relatum":2,
          "changes":[{"on":"relation", "type":"appearance"}]
        }
      ],
      "provenance": {"uri": "https://fr.wikipedia.org/wiki/Rue_G%C3%A9rard_(Paris)"}
    }
  ]
}
```

Listing 7.3 – Exemple de description de l'événement « *Par arrêté municipal du 30 août 1978, sa portion orientale, de la rue Bobillot à la rue du Moulin-des-Prés, prend le nom de rue du Père-Guérin.* » en JSON provenant de Wikipédia.

7.2.2.3 Conversion des données vers le format intermédiaire

La conversion des données vers le format intermédiaire s'effectue via un script spécifique pour chaque source. Pour les données non structurées notamment les descriptions d'événements issus de sources textuelles, la génération du format intermédiaire a été réalisée manuellement. Cette méthode est beaucoup plus simple et compacte que la création manuelle des triplets RDF.

Une fois les données converties dans le format intermédiaire, les triplets peuvent être générés automatiquement à partir de ce format. Ils sont ensuite injectés dans le triplestore *GraphDB* via son API, en étant insérés séparément dans des graphes nommés distincts pour chaque source, comme indiqué précédemment.

7.2.3 Création du graphe de faits

À ce stade, nous avons donc un ensemble de factoïdes répartis dans des graphes nommés selon leurs sources d'origine, à partir desquels le graphe de faits sera construit. Les

graphes nommés de l'étape précédente sont figés et ne reçoivent aucune insertion supplémentaire ; ils servent uniquement à la construction du graphe de faits.

7.2.3.1 Initialisation des ressources et attribution des dates

L'application des règles présentées dans le chapitre 6 pour peupler le graphe de faits peut être effectuée soit par un moteur d'inférence soit par des requêtes SPARQL, selon le degré de contrôle requis. Dans notre cas, nous privilégions l'utilisation de requêtes SPARQL pour deux raisons principales. D'une part, cette approche permet un contrôle plus direct sur les liens générés, ce qui est essentiel pour éviter la propagation non souhaitée de certains liens. D'autre part, elle permet de générer de nouvelles ressources dans le graphe de faits. En effet, après l'application des règles, il peut arriver qu'une ressource R_S issue du graphe des factoïdes ne trace aucune ressource dans le graphe de faits, c'est-à-dire qu'il n'existe pas de R_F tel que `hasTrace(`R_F`,` R_S`)`. Dans ce cas, il est nécessaire de créer une nouvelle ressource R_F dans le graphe de faits. Or, un moteur d'inférence standard ne permet pas cette opération de génération. À l'inverse, les requêtes SPARQL autorisent des traitements conditionnels et la production explicite de nouveaux triplets. Ainsi, chaque opération pouvant être exprimée en SPARQL a été implémentée sous cette forme, de manière à maximiser la cohérence entre la génération des données et leur manipulation au sein du graphe. Enfin, certaines opérations nécessitent la création de triplets temporaires, qui ne sont pas destinés à apparaître dans le graphe de faits. Ces triplets sont insérés dans un graphe nommé temporaire.

La construction du graphe de faits est synthétisée dans la section C.2, notamment grâce à l'algorithme C.9. Elle commence par l'initialisation des ressources dans le graphe de faits, en les reliant à leurs observations dans les différents graphes nommés via la relation `hasTrace`. Ensuite, les dates d'apparition et de disparition des entités et relations sont déduites à partir des informations contenues dans ces graphes.

7.2.3.2 Reconstitution de l'évolution des attributs

Pour reconstituer l'évolution des attributs, il faut comparer les différentes versions d'un même attribut en générant des triplets avec les prédicats `sameVersionValueAs` ou `differentVersionValueFrom`. Pour cela, nous utilisons un ensemble de fonctions codées en Python qui évaluent la similarité des versions selon le type d'attribut et des paramètres personnalisables. Pour les attributs servant à décrire des identifiants, nous utilisons une égalité stricte, comme c'est le cas pour un identifiant de code postal. Notre graphe contient principalement des attributs de type nom et géométrie, qui ne peuvent être comparés sur un critère d'égalité stricte.

Comparaison les versions d'attributs de type nom. Pour les attributs de type *nom*, nous évaluons la similarité des versions d'attributs comme décrit à l'algorithme C.10. Pour chaque version v_i, un label simplifié s_i est généré à partir de sa valeur n_i, de la langue associée $lang_i$ et du type de l'entité appliqué à l'attribut. Cette étape est réalisée par la fonction *versions_noms_similaires()* présentée dans l'algorithme C.10. La méthode de simplification dépend du type de label : il peut s'agir d'une version d'attribut correspondant à une rue, une commune ou un numéro d'immeuble. Concrètement, la simplification consiste généralement à : supprimer les signes diacritiques, éliminer les prépositions, conjonctions et adverbes, remplacer les caractères non alphanumériques par des espaces et mettre tous les caractères en minuscule. Cette étape est suivie d'un tri alphanumérique des caractères, après suppression des espaces. Par exemple, pour une version dont la valeur

est « "rue du Père-Guérin"@fr », la valeur simplifiée avant tri sera « rue pere guerin », puis « eeeeginprrruu » après le tri alphanumérique. Cette règle est appliquée ici uniquement aux labels en français. Notre implémentation a toutefois été conçue de manière à pouvoir ajouter facilement d'autres règles de simplification selon la langue, bien que cela n'ait pas été nécessaire dans notre cas, toutes nos sources étant en français. Rien n'interdit donc en théorie d'appliquer cette approche à des sources multilingues, par exemple pour des voies parisiennes ayant des labels dans d'autres langues.

Comparaison les versions d'attributs de type géométrie. Côté géométrie, les problématiques sont différentes. La fonction *versions_geometries_similaires()* présentée dans l'algorithme C.11 qui permet de déterminer la similarité des deux versions d'attributs de type géométrie illustre cela. Du fait de l'hétérogénéité importante des géométries issues des différentes sources, déterminer la similarité entre deux versions v_1 et v_2 d'un attribut de type géométrie, associées respectivement aux valeurs g_1 et g_2 (telles que `versionValue`(v_1, g_1) et `versionValue`(v_2, g_2)), s'avère complexe. Les géométries peuvent être exprimées dans des systèmes de coordonnées distincts, et, dans le cas des voies, prendre la forme de polygones ou de polylignes. Dans ce travail, toutes les géométries ont été transformées pour être exprimées dans le système de référence de coordonnées Lambert 93. Cela permet notamment de créer des zones tampon à rayon exprimé en mètres, plutôt qu'en degrés, ce qui est essentiel pour des calculs métriques sur les distances et surfaces. Un autre enjeu provient des imprécisions de géoréférencement : deux géométries représentant la même entité peuvent présenter un décalage spatial significatif, dû aux différences de précision planimétrique des levés des plans, de systèmes de référence de coordonnées, de numérisation ou de géoréférencement des plans. Nous proposons différentes méthodes en fonction du type d'entité associée à la géométrie.

Pour les numéros d'immeubles, qui sont généralement des éléments ponctuels, on compare des points : si les géométries sont des polylignes ou des polygones, on prend leur centroïde. Deux versions sont considérées similaires si la distance entre leurs points respectifs est inférieure à un seuil prédéfini. Pour les voies de communication, les géométries sont le plus souvent des polylignes mais peuvent également être des polygones ; dans ce cas, on peut générer une polyligne à partir d'un polygone et appliquer des techniques inspirées de l'appariement de données linéaires. Une autre approche consiste à appliquer une zone tampon aux polylignes, fixe ou proportionnelle à leur longueur, de sorte que l'on compare ainsi des polygones et des polylignes de manière uniforme. En notant $S(g)$ la surface d'une géométrie, deux versions v_1 et v_2 sont considérées comme similaires si le rapport de recouvrement $\frac{S(g_1 \cap g_2)}{S(g_1 \cup g_2)}$ est supérieur ou égal à un coefficient minimal $\alpha_{\min} \in [0, 1]$.

L'ensemble des comparaisons est stocké dans un graphe nommé distinct des autres qui est utilisé pour la reconstitution de l'évolution des attributs.

7.3 Vérification de la cohérence du graphe

Conformément aux recommandations de la méthode *SAMOD*, nous avons procédé à plusieurs évaluations. Nous présentons dans cette section celle de la cohérence des différents graphes de connaissances générés. Autrement dit, il s'agit de s'assurer que les triplets générés à partir des différentes sources, ainsi que ceux produits à l'issue de l'application complète de la méthode de peuplement du graphe de faits, sont conformes à l'ontologie et cohérents.

L'ontologie définit le vocabulaire utilisé pour décrire les entités et leurs évolutions : elle spécifie les classes, les propriétés et les liens possibles entre elles. Cependant, elle

ne contraint pas la manière dont les ressources doivent être combinées. Autrement dit, l'ontologie permet de représenter la connaissance, mais pas de vérifier la validité ou la cohérence des graphes qui en découlent. C'est précisément le rôle des contraintes SHACL (Shapes Constraint Language) qui permettent d'exprimer des règles de validation garantissant que les triplets respectent la structure attendue, les types de relations autorisées, et les dépendances logiques entre les entités (Knublauch et Kontokostas, 2017).

Cette étape de validation repose donc sur la vérification d'un ensemble de contraintes SHACL, qui assurent la cohérence structurelle et sémantique du graphe. Plusieurs niveaux de validation sont ainsi considérés :

1. les factoïdes doivent être conformes à l'ontologie ;
2. le graphe de faits doit respecter les contraintes structurelles définies par l'ontologie ;
3. les versions et les changements doivent alterner correctement ;
4. les versions incomplètes doivent avoir été supprimées, comme indiqué en sous-sous-section 6.4.2.1 ;
5. les versions successives équivalentes doivent avoir été fusionnées, selon l'approche décrite en sous-sous-section 6.4.2.2.

La première étape de validation consiste à contrôler la qualité des données sources utilisées pour construire le graphe. Les triplets issus des graphes de factoïdes sont ainsi vérifiés à l'aide de règles SHACL et de requêtes SPARQL afin de garantir leur conformité à l'ontologie. Cette validation préliminaire permet de s'assurer que la structure des données initiales n'introduit pas d'incohérences dans le graphe de faits. La même procédure est ensuite appliquée au graphe de faits, pour confirmer que la méthode de peuplement produit un graphe conforme à l'ontologie et structurellement cohérent.

La cohérence temporelle est ensuite évaluée par des requêtes SPARQL qui vérifient que les versions et les changements s'alternent correctement pour chaque attribut représenté dans le graphe. Cette vérification permet également de s'assurer qu'aucune période temporelle n'est laissée sans version. Concrètement, chaque version doit être reliée à au moins une assertion issue des graphes de factoïdes. Si ce n'est pas le cas, nous examinons si les changements associés sont eux-mêmes correctement reliés, comme illustré en figure 6.10c. Dans ce cas, la version reste considérée comme cohérente.

Enfin, l'évaluation de la fusion des versions similaires successives, décrite en sous-sous-section 6.4.2.2, fait l'objet d'une vérification manuelle. Un outil de visualisation cartographique a été développé à cet effet [6] (voir figure 7.10 et figure 7.11). Il permet d'explorer l'évolution d'une entité géographique en affichant une frise chronologique représentant la séquence de ses versions et l'évolution de chacun de ses attributs, tandis qu'une carte associée visualise les géométries correspondantes. L'outil offre également la possibilité de visualiser l'état du territoire à un instant donné : la sélection d'une date fait apparaître sur la carte l'ensemble des entités géographiques existantes à cette époque, avec l'emprise spatiale qu'elles occupaient alors. Pour ce faire, il exécute des requêtes SPARQL sur le graphe peuplé, en se limitant au graphe nommé des faits, puis transforme les résultats obtenus afin d'afficher les informations correspondantes dans la frise et sur la carte.

7.3.1 Contraintes sur les ressources

Les règles de validation définies pour l'ontologie *PeGazUs* précisent les contraintes à respecter pour assurer la conformité du graphe aux attentes du modèle conceptuel. Elles

6. Disponible dans le dépôt : https://github.com/charlybernard/pegazus-extension

concernent notamment la complétude des relations obligatoires, les cardinalités des liens, et la compatibilité des types d'entités. Elles se répartissent comme suit.

Repères et relations associées. Les entités de classe `Landmark` représentent les repères géographiques de base du modèle. Chaque repère doit être associé à au moins une instance de `LandmarkType`, ce qui permet de spécifier sa nature (bâtiment, voie, parc, etc.). Lorsqu'un repère possède un attribut (`:hasAttribute`), celui-ci doit impérativement être une instance de la classe `Attribute`, garantissant ainsi la cohérence des liens entre les entités géographiques et leurs caractéristiques.

Les instances de `LandmarkRelation` et ses sous-classes doivent, quant à elles, être associées à une unique instance de `LandmarkRelationType`, précisant la nature de la relation (par exemple, « appartient à », « se situe près de », etc.).

Chaque relation impliquant un ou plusieurs points de repère, qu'elle soit représentée sous la forme d'une `LandmarkRelation` et de ses sous-classes, doit comporter exactement un `locatum` et au moins un `relatum`, ces deux éléments devant être des instances de la classe `Landmark`. Cette contrainte assure à la fois la cohérence sémantique des relations, en garantissant qu'elles sont systématiquement typées via une instance de `LandmarkRelationType`, et leur cohérence directionnelle, en identifiant explicitement le `locatum` (entité localisée) et les `relatums` (entités de référence) au sein de chaque relation.

Adresses et segments d'adresses. Chaque adresse doit être reliée à une et une seule instance de `Landmark`, assurant ainsi une correspondance explicite entre l'adresse et le repère géographique qu'elle désigne. La structure interne d'une adresse repose sur une séquence de segments d'adresse, représentant les étapes successives de son énoncé. Le premier segment, introduit par la propriété `:firstStep`, doit être une instance de la classe `AddressSegment` ou de sa sous-classe `FinalAddressSegment`. Chaque instance de `AddressSegment` peut à son tour référencer au plus un segment suivant à l'aide de la propriété `:nextStep`, lequel doit également appartenir à l'une de ces deux classes. Une instance de la classe `FinalAddressSegment` marque la fin de la séquence et, par définition, ne peut pas posséder de segment suivant. Ces règles assurent une représentation ordonnée cohérente de la composition des adresses.

Attributs et versions. Les ressources de type `Attribute` représentent les propriétés décrivant un repère ou une relation. Chaque attribut doit être lié à une unique instance de `AttributeType`, précisant la nature de l'information représentée (par exemple, un nom, une géométrie ou un type de voie). Les différentes valeurs que peut prendre un attribut au cours du temps sont représentées par des instances de la classe `AttributeVersion`, reliées à l'attribut via la propriété `:hasAttributeVersion`. Cette structuration garantit que toute information descriptive est clairement typée et que ses évolutions temporelles sont traçables.

Changements et dépendances. Les ressources de type `Change` (et ses sous-classes) permettent de représenter les transformations affectant les repères, leurs relations ou leurs attributs. Chaque changement doit être associé à une unique instance de `ChangeType`, indiquant la nature de la modification (création, suppression, fusion, etc.), ainsi qu'à une unique instance de `Event`, qui en fournit le contexte temporel. Une instance de `AttributeChange` s'applique toujours à une seule instance d'`Attribute`, et peut rendre caduque (`outdates`) ou effective (`makesEffective`) une ou plusieurs versions de cet attribut. De manière analogue, une instance de `LandmarkChange` porte sur une unique instance

de `Landmark`, tandis qu'une instance de `LandmarkRelationChange` concerne une seule instance de `LandmarkRelation`. Ces contraintes garantissent que chaque changement reste localisé, non ambigu et correctement ancré dans le temps.

Événements et temps. Les changements et les versions sont contextualisés par des ressources de type `Event`, qui doivent toutes être associées à une instance temporelle, soit `TimeInstant`, soit l'une de ses sous-classes `CrispTimeInstant` ou `FuzzyTimeInstant`. Une instance de type `CrispTimeInstant` doit comporter trois informations essentielles : une précision temporelle (`timePrecision`) définie comme une instance de `TemporalUnit`, un horodatage exact (`timeStamp`) exprimé au format `xsd:dateTimeStamp`, et une référence à un calendrier (`timeCalendar`) de classe `TRS`. Ces contraintes permettent de décrire de manière explicite le degré de précision et le système de référence temporel utilisé, assurant ainsi la cohérence chronologique de l'ensemble des événements du graphe.

7.3.2 Contraintes sur les propriétés

Nous proposons également de vérifier, à l'aide de règles SHACL que les ressources liées par les propriétés de l'ontologie sont bien déclarées comme instances des classes définies dans les axiomes portant sur leurs domaines et portées respectifs. Le tableau suivant résume les principales propriétés typées et les contraintes associées à leur domaine et à leur portée.

Propriété	**Sujet**	**Objet**
`firstStep`	`Address`	`AddressSegment`
`nextStep`	`AddressSegment`	`AddressSegment`
`isLandmarkType`	`Landmark`	`LandmarkType`
`isLandmarkRelationType`	`LandmarkRelation`	`LandmarkRelationType`
`isChangeType`	`Change`	`ChangeType`
`isAttributeType`	`Attribute`	`AttributeType`
`hasAttributeVersion`	`Attribute`	`AttributeVersion`
`outdates`	`AttributeChange`	`AttributeVersion`
`makesEffective`	`AttributeChange`	`AttributeVersion`
`locatum`	`LandmarkRelation`	`Landmark`
`relatum`	`LandmarkRelation`	`Landmark`
`targets`	`Address`	`Landmark`
`appliedTo`	`Landmark` / `Attribute` / `LandmarkRelation`	`Change`
`dependsOn`	`Change`	`Event`
`hasTime`	Toute ressource	`TemporalEntity`
`versionValue`	`AttributeVersion`	`Literal`

Tableau 7.2 – Principales propriétés de l'ontologie *PeGazUs*, avec les types de sujets et d'objets correspondants.

7.4 Évaluation de l'ontologie

Conformément à la méthode SAMOD, nous avons procédé à l'évaluation de l'ontologie PeGazUs. Tout d'abord, l'ontologie créée est soumise au raisonneur HermiT, intégré au

logiciel Protégé[7] (Musen, 2015), afin de tester sa cohérence et d'identifier d'éventuelles relations de subsomption entre les classes créées. Nous l'avons également analysée à l'aide de l'outil OOPS![8], qui permet de détecter les anomalies et les mauvaises pratiques de conception dans les ontologies (Poveda-Villalón *et al.*, 2014). Ceci nous a permis de corriger les erreurs détectées. Après avoir peuplé l'ontologie, nous avons traduit en SPARQL les questions de compétence listées lors de sa création et les avons testées modelet par modelet, afin de nous assurer que l'ontologie permet effectivement d'y répondre. Nous ne présentons pas ici l'intégralité des requêtes SPARQL ni leurs résultats, hormis deux ou trois exemples illustratifs. L'ensemble des requêtes et des résultats est disponible dans l'annexe B.

7.4.1 Modelet des adresses

La partie de l'ontologie sur les adresses est évaluée à partir des données que nous y avons intégrées à l'aide de la méthode de peuplement présentée précédemment. Nous vérifions que l'ontologie permet de répondre aux questions de compétence définies lors de la définition de ce modelet, transcrites en requêtes SPARQL. Ces questions sont les suivantes :

1. Quelles sont les adresses répertoriées le long d'une rue donnée ?
2. Quelles sont les coordonnées correspondant à la cible d'une adresse donnée ?
3. Quelles sont les adresses localisées dans une zone donnée ?

```
SELECT ?address ?addressLabel ?geom ?streetName
WHERE {
  ?lm a :Landmark ; :isLandmarkType :Thoroughfare ; rdfs:label ?streetName .
  FILTER(LCASE(STR(?streetName)) = LCASE("rue du Pas de la Mule") &&
  ↪ LANG(?streetName) = "fr")

  ?address a :Address ; rdfs:label ?addressLabel ; :hasStep ?addrSegment .
  ?addrSegment a :AddressSegment ; :relatum ?lm .

  ?address :targets ?target .
  ?target :hasAttribute [ a :Attribute ; :isAttributeType atype:Geometry ;
  ↪ :hasAttributeVersion ?attrVersion ] .
  ?attrVersion :versionValue ?geom .
}
```

Listing 7.4 – Requête SPARQL sélectionnant les adresses et les coordonnées associées pour lesquelles la cible est située le long de la rue du Pas de la Mule.

Le Listing 7.4 présente une requête SPARQL sélectionnant l'ensemble des adresses dont la cible est située le long de la rue du Pas de la Mule. Cette requête renvoie 13 adresses et leurs coordonnées, qui sont cartographiées en figure 7.4. Ce premier résultat montre que l'ontologie développée permet bien de répondre aux deux premières questions de compétence. Pour la troisième question, la requête SPARQL définie par le Listing 7.5 consiste à sélectionner toutes les adresses et leurs géométries, de manière similaire à la requête du Listing 7.4. La différence est que cette requête filtre les adresses pour rechercher

7. http://protege.stanford.edu/
8. https://oops.linkeddata.es/

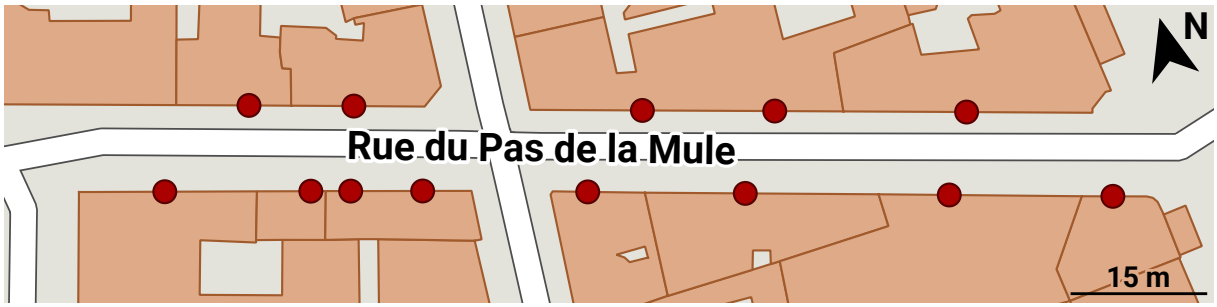


FIGURE 7.4 – Localisation des adresses dont la cible est située le long de la rue du Pas de la Mule.

celles dont la cible est située dans une zone géographique définie. En pratique, seules les adresses dont la géométrie est incluse dans cette zone sont retournées. Ce filtrage est rendu possible grâce aux fonctions spatiales de GeoSPARQL, qui permettent d'évaluer des relations spatiales entre les géométries des entités. Ces tests valident la capacité du modelet des adresses à répondre aux questions de compétences listées.

```
SELECT DISTINCT ?addr ?addrLabel ?geom
WHERE {
    BIND("<http://www.opengis.net/def/crs/OGC/1.3/CRS84> POLYGON ((2.367404
    ↪  48.856463, 2.364829 48.854656, 2.368562 48.854247, 2.367404
    ↪  48.856463))"^^geo:wktLiteral AS ?area)

    ?addr a :Address ; rdfs:label ?addrLabel ;
        :targets [:hasAttribute [:isAttributeType atype:Geometry ;
        :hasAttributeVersion [:versionValue ?geom]]] .
    FILTER (geof:sfWithin(?geom, ?area))
}
```

Listing 7.5 – Requête SPARQL sélectionnant les adresses et les coordonnées associées pour lesquelles la cible est située dans la zone définie par `?area`.

Par ailleurs, nous avons également testé la capacité de cette ontologie à représenter des adresses très diverses. En plus des adresses parisiennes, nous avons donc également peuplé l'ontologie avec des adresses japonaises et tchèques provenant d'OpenStreetMap, qui présentent des structures différentes des adresses françaises. Nous avons ainsi pu constater la généricité de l'ontologie qui permet bien de représenter ces adresses étrangères sans nécessiter de modifications.

7.4.2 Modelet des évolutions temporelles

Rappelons les questions de compétences pour ce modelet :

— Quelles sont les voies existant à un instant donné ?
— Sur quel(s) intervalle(s) temporel(s) est valable une adresse ou une dénomination de lieu donnée ?
— Quel est l'historique d'une entité ? Autrement dit, quels sont les événements qui lui sont associés ?
— Quels sont les états et les événements manquants dans l'historique d'une entité ?

Ces questions ont également été traduites en requêtes SPARQL pour tester la capacité du modelet à représenter et interroger les évolutions spatio-temporelles des entités. Concernant la première question, un instant t est fixé et la requête sélectionne les entités de type "Voie de communication" associées à un événement de création qui a pour instant $t_{\text{app}} \leq t$ et un événement de disparition qui a pour instant $t_{\text{dis}} \geq t$. Ensuite, pour la deuxième question, il faut se rappeler qu'une adresse fait référence à un ensemble d'entités géographiques reliées par un ensemble de relations. Sa validité temporelle peut alors être interprétée de deux manières. On peut considérer que l'adresse est valable dès lors que sa cible existe, auquel cas elle est valide sur l'intervalle $[t_{\text{app}}, t_{\text{dis}}]$ défini comme dans la question précédente. Sinon, elle est valable uniquement lorsque la cible et tous ses segments d'adresse existent simultanément, soit l'intersection de $[t_{\text{app}}, t_{\text{dis}}]$ avec les intervalles de validité des segments. Cette deuxième option revient à considérer que l'on recherche l'intervalle temporel sur lequel un énoncé particulier d'adresse est valable. La troisième question implique de sélectionner l'ensemble des événements reliés de près ou de loin à une entité géographique, c'est-à-dire ceux qui lui sont directement appliqués, ainsi que ceux appliqués à ses attributs et aux relations dont elle est le *locatum*. Les événements où l'entité apparaît en tant que *relatum* (rôle passif) sont exclus. Enfin, la dernière question vise à identifier les informations manquantes. Elle correspond à la sélection des événements sans valeur temporelle, c'est-à-dire les événements e tels que $\nexists t, \texttt{hasTime}(e, t)$. Pour les attributs, il s'agit de détecter les intervalles temporels dépourvus de version valable. Par exemple, sur la figure 6.7, il existe de nombreuses lacunes dont entre v_2 et v_4. On peut récupérer ces intervalles via des requêtes SPARQL bien que leur formulation soit assez complexe.

7.4.3 Modelet des sources

Les questions de compétences associées à ce modelet sont les suivantes :

1. De quelle source provient l'information X ?
2. Quelle est la description de la source Y ?

Ici, il n'est pas nécessaire d'introduire une représentation spécifique, car les informations de provenance sont déjà représentées dans les deux modelets précédents. En effet, les ressources créées sont associées à d'autres entités qui indiquent leur origine, plutôt qu'à la source documentaire elle-même. Par exemple, la ressource décrivant la *rue Gérard* issue d'OpenStreetMap est associée à l'entité `osmrel:11832935` [9], laquelle est elle-même reliée à la source *OpenStreetMap*.

Ainsi, pour répondre à la première question de compétence, la requête SPARQL permettant d'identifier la provenance d'une information donnée `?x` est indiquée par le Listing 7.6.

```
SELECT ?x ?source WHERE {
  ?x prov:wasDerivedFrom ?prov .
  ?prov rico:isOrWasDescribedBy ?source .
}
```

Listing 7.6 – Requête SPARQL identifiant la source d'une information donnée.

Quant à la seconde question, étant donné qu'on réutilise le vocabulaire et les recommandations de l'ontologie *RiC-O*, chaque source est représentée comme une instance de

9. `osmrel` est le préfixe de https://www.openstreetmap.org/relation/

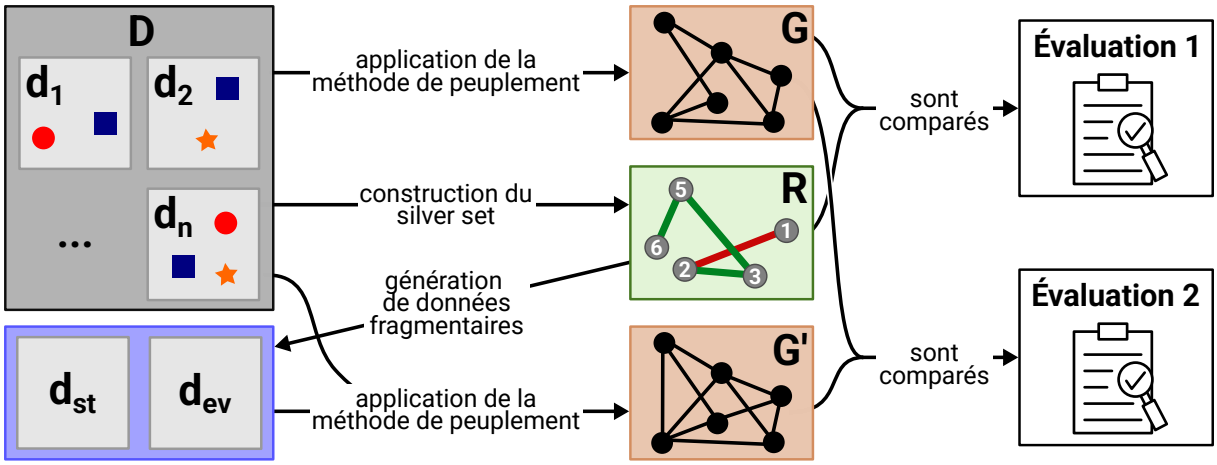


FIGURE 7.5 – Schéma expliquant la stratégie d'évaluation.

la classe (`rico:Record`) associée à des propriétés comme sont son titre, son auteur, sa date de création, son type, etc. Pour obtenir la description d'une source donnée, il suffit donc d'interroger le graphe afin de sélectionner les informations nécessaires, par exemple à l'aide d'une requête SPARQL fournie par le Listing 7.7.

```
SELECT ?title ?author ?date ?type WHERE {
  ?source a rico:Record ;
          rico:title ?title ;
          rico:hasCreator ?author ;
          rico:hasCreationDate ?date .
}
```

Listing 7.7 – Requête SPARQL sélectionnant les informations descriptives d'une source.

7.5 Évaluation de la méthode de peuplement de l'ontologie

Pour évaluer la méthodologie proposée dans le chapitre 6, nous proposons une approche à partir d'un cas d'usage réel. Les ressources nécessaires pour reproduire l'évaluation sont mises à disposition dans un dépôt GitHub accessible à l'adresse suivante : https://github.com/charlybernard/pegazus-evaluation.

7.5.1 Stratégie d'évaluation

Pour évaluer notre méthodologie de peuplement de l'ontologie PeGazUs, il serait idéal de comparer le résultat à une vérité terrain décrivant l'évolution des adresses et des entités géographiques qui les constituent sur une zone et un intervalle temporel donnés. Cependant, à notre connaissance, un tel jeu de données n'existe pas pour des adresses parisiennes anciennes. Une alternative serait de construire manuellement une vérité terrain à partir des données d'entrée fournies à la méthode, puis de la comparer à la sortie produite. Toutefois, cette approche s'avère peu praticable, compte tenu du volume et de la complexité des données impliquées.

Pour ces deux raisons, nous proposons une évaluation de la méthode de peuplement appliquée sur des données d'adresse en *snapshots* du quartier du Faubourg Saint-Antoine. Celle-ci permettent, en effet, de générer des jeux de données hétérogènes et fragmentaires, simulant ceux que l'on peut obtenir via des sources historiques et de construire automatiquement un jeu de données de référence grâce auquel nous pourrons évaluer les résultats de notre algorithme de peuplement. Nous proposons ainsi une stratégie dont la chaîne de traitement illustrée par la figure 7.5.

Soit $S = \{s_1, \ldots, s_n\}$ l'ensemble des sources, et $D = \{d_1, \ldots, d_n\}$ l'ensemble des jeux de données correspondants, où chaque jeu d_i est dérivé de la source s_i. L'ensemble D constitue les données d'entrée de notre méthode tandis que R désigne le jeu de données de type *silver set* utilisé comme référence pour l'évaluation. Le jeu R a été construit automatiquement à partir de D, indépendamment de la méthode de peuplement proposée, et une validation manuelle a été réalisée sur un échantillon pour vérifier que le processus automatique produit bien les évolutions attendues. Contrairement à un *gold standard* construit généralement entièrement manuellement, un *silver set* combine automatisation et vérification manuelle, ce qui en fait une référence fiable pour notre évaluation (Hahn *et al.*, 2010). Enfin, nous notons G le graphe obtenu en appliquant la méthode proposée aux données de D, et G' le graphe construit selon le même processus, mais enrichi des données fragmentaires générées à partir de R.

L'évaluation se déroule en deux volets. Le premier consiste à comparer l'évolution des adresses dans G avec celle du jeu de référence R afin de s'assurer que la méthode fait correctement ce pour quoi elle a été mise en place. Le second vise à tester la robustesse de la méthode face à l'ajout de nouvelles données fragmentaires, ceci se fait en comparant pour chaque adresse ses évolutions générées dans G et G'.

7.5.2 Construction du jeu de données de référence et des jeux de données fragmentaires

Construction du jeu de données de référence Le jeu de données de référence R est construit à partir des jeux de données issus des sources suivantes, ordonnées chronologiquement : s_1 : le cadastre de Paris de 1807 (période 1806-1808) ; s_2 : l'atlas Vasserot (1810–1836) ; s_3 : l'atlas Jacoubet (1827–1839) ; s_4 : l'atlas municipal de 1888 (1887-1889) ; s_5 : la Base Adresse Nationale (2024) ; et s_6 : OpenStreetMap (2025).

Pour construire R, nous représentons chaque adresse par un petit graphe orienté reliant chronologiquement ses observations à travers les jeux de données de D. Pour relier les observations d'une adresse a, nous définissons son identité comme la combinaison de son numéro et du nom de la rue associée. Soit o_i l'observation de a dans le jeu de données $d_i \in D$. La *ligne de vie* de a est une séquence ordonnée temporellement de ces observations, connectées par des liens $o_i \to o_j$ avec $i < j$, en sautant les jeux où a est absent (voir figure 7.6).

Par exemple, l'adresse dont l'énoncé est « *2 rue du Pas de la Mule* » est absente dans la source s_4, ce qui donne les liens suivants : $o_1 \to o_2$, $o_2 \to o_3$, $o_3 \to o_5$ et $o_5 \to o_6$. Pour chaque lien, nous évaluons si les géométries des observations successives sont similaires. Pour cela, nous testons leur proximité spatiale, suivant l'hypothèse selon laquelle deux adresses proches dans l'espace ont de fortes de chances de désigner le même lieu. Nous calculons donc la distance euclidienne entre leurs deux géométries et vérifions si elle est inférieure à un seuil prédéfini noté d_{max}. Si ce n'est pas le cas, nous considérons qu'un changement s'est produit (par exemple, le nom de la rue a été supprimé puis réattribué à une autre voie). Sinon, on considère qu'il s'agit de la même version.

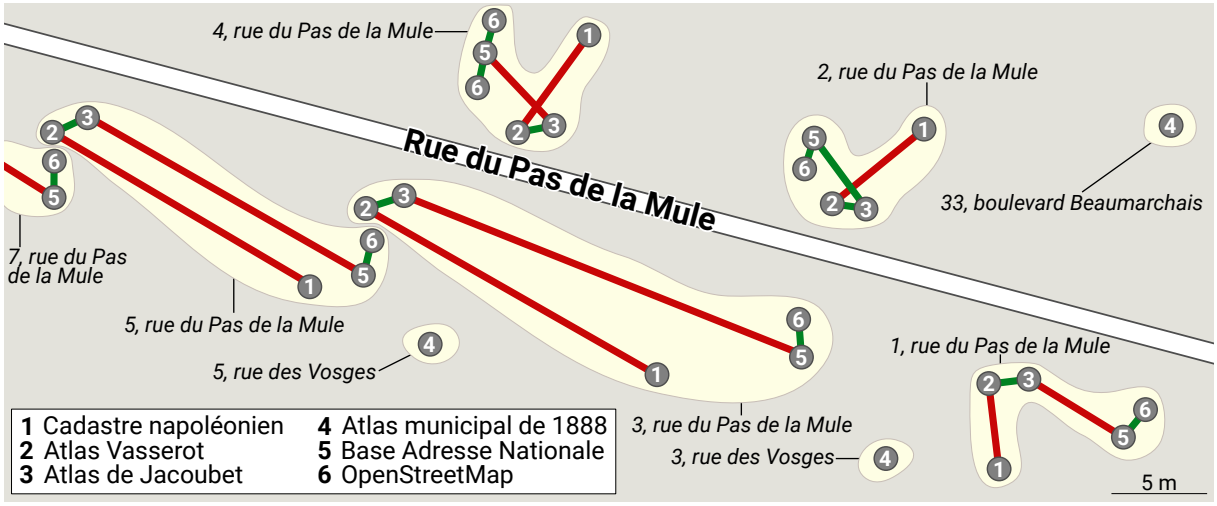


FIGURE 7.6 – Construction de *lignes de vie* d'adresses autour de la *rue du Pas de la Mule*. Les lignes rouges représentent les liens entre versions successives non similaires, tandis que les lignes vertes correspondent à des versions successives similaires.

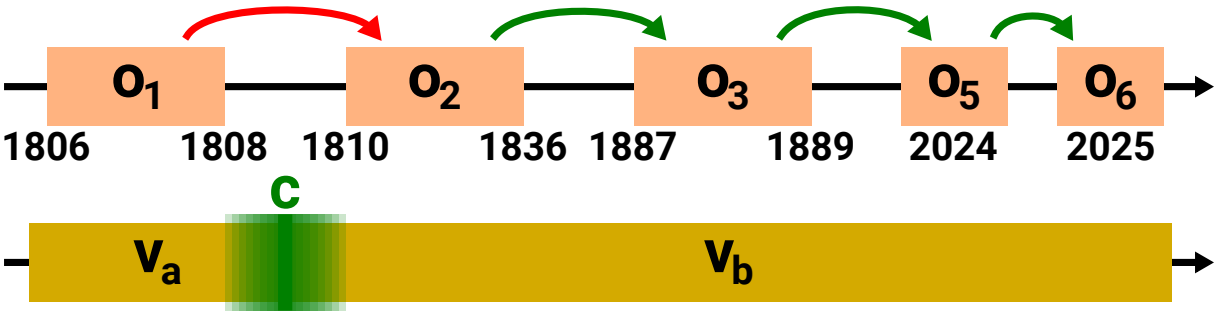


FIGURE 7.7 – Construction de *lignes de vie* de l'adresse *2 rue du Pas de la Mule*. → : *a une géométrie différente de*, → : *a une géométrie similaire à*.

À partir de ces *lignes de vie*, comme celle de l'exemple illustré par la figure 7.7, nous déduisons que l'adresse possède deux versions distinctes : v_a et v_b. La version v_a est dérivée de l'observation o_1, tandis que la version v_b résulte de la fusion des observations o_2, o_3, o_5 et o_6 en raison de la similarité de leur géométrie. Ces deux versions successives sont séparées par un changement c, lié à un événement réel dont la date se situe entre 1808 et 1810. Ainsi, les temps valides respectifs de v_a et v_b sont $]-\infty; x]$ et $[x; +\infty[$, avec $x \in [1808; 1810]$. Dans le jeu de référence R, nous associons donc à cette adresse les deux versions v_a et v_b, ainsi que le changement c qui les relie.

Bien que les sources de données soient hétérogènes, elles ne sont pas fragmentaires. Cela motive la seconde partie de l'évaluation, qui consiste à intégrer des données fragmentaires, à la fois des états et des événements.

Génération du jeu de données fragmentaires Nous générons deux jeux de données synthétiques destinés à évaluer la robustesse de la méthode de peuplement :

— d_{st}, représentant des états d'adresses à différents moments ;

— d_{ev}, représentant des événements affectant ces adresses.

Ces jeux de données sont construits de manière à introduire des observations fragmentaires cohérentes avec les sources disponibles à partir du jeu de référence R, tout en

étant indépendants du résultat produit par l'algorithme de peuplement. L'objectif est de tester la robustesse de la méthode : nous évaluons si l'ajout d'informations partielles mais cohérentes perturbe les lignes de vie reconstruites à partir des données originales. Autrement dit, nous vérifions dans quelle mesure la méthode parvient à intégrer des données fragmentaires supplémentaires sans altérer la cohérence du graphe existant.

Pour construire d_{st}, nous procédons de la manière suivante. Nous sélectionnons aléatoirement un sous-ensemble d'adresses dans R. Pour chaque adresse du sous-ensemble, une ou plusieurs de ses versions $v \in R$ sont choisies aléatoirement. Nous générons une observation fragmentaire dérivée de v, en définissant une géométrie et un temps valide. D'une part, la géométrie est représentée par un point situé à une distance inférieure au seuil de similarité de celles associées à v (afin de simuler une observation différente mais suffisamment similaire). D'autre part, le temps valide est un intervalle temporel strictement inclus dans la période de validité de v. Par exemple, pour l'adresse « *2 rue du Pas de la Mule* », une observation o_α dérivée de la version v_a peut être générée en créant une géométrie située dans un rayon inférieur au seuil d_{max} autour de celle de v_a, et un intervalle temporel tel que $[1797; 1809] \subset]-\infty; x]$ où $x \in [1808; 1810]$.

Le jeu d_{ev} est construit selon un principe similaire, mais s'applique à des événements. Il débute par la sélection aléatoire d'un sous-ensemble d'adresses dans R. Pour chaque adresse ainsi choisie, nous générons un événement associé à un instant t, tiré dans un intervalle cohérent avec la ligne de vie de l'adresse dans R. Plus précisément, l'événement doit impliquer un changement affectant une version de l'adresse, tel que sa mise en effectivité ou sa péremption. Ainsi, pour une version dont l'intervalle de validité est $[t_{\text{start}}, t_{\text{end}}]$, l'instant t associé à une mise en effectivité vérifie $t \leq t_{start}$ tandis que celui associé à une péremption vérifie $t \geq t_{end}$ Reprenons l'exemple de la « 2 rue du Pas de la Mule ». Trois événements potentiels peuvent être générés :

- un événement d'apparition situé avant la première observation connue, c'est-à-dire avec une date antérieure à 1806 ;
- un événement de disparition situé après la dernière observation connue, c'est-à-dire avec une date postérieure à 2025 ;
- un événement intermédiaire situé entre deux observations, c'est-à-dire avec une date comprise entre 1808 et 1810, pouvant correspondre à un changement de géométrie.

Pour chaque étape de génération de données, le choix aléatoire des adresses permet de simuler une variété de cas réalistes : certaines adresses reçoivent des états supplémentaires, d'autres des événements, d'autres encore les deux. Cette diversité est essentielle pour évaluer la capacité de la méthode à intégrer de manière cohérente des informations fragmentaires et hétérogènes.

7.5.3 Méthodes d'évaluation du peuplement de l'ontologie

Cette section présente les deux méthodes de l'évaluation de peuplement du graphe. La première consiste à mesurer la capacité de la méthode à reconstruire fidèlement les évolutions des adresses en comparant les lignes de vie des adresses générées dans G à celles créées dans le jeu de référence R à partir des mêmes données. La seconde vise à évaluer la robustesse de cette reconstruction face à l'ajout de données fragmentaires, en comparant les évolutions obtenues dans G et dans un graphe perturbé G', enrichi des nouvelles observations synthétiques. La figure 7.8 illustre ces deux évaluations à travers un exemple d'adresse : le jeu de référence R, le graphe reconstruit G, et le graphe perturbé G'. Elle met en évidence les deux comparaisons réalisées : entre R et G pour vérifier la

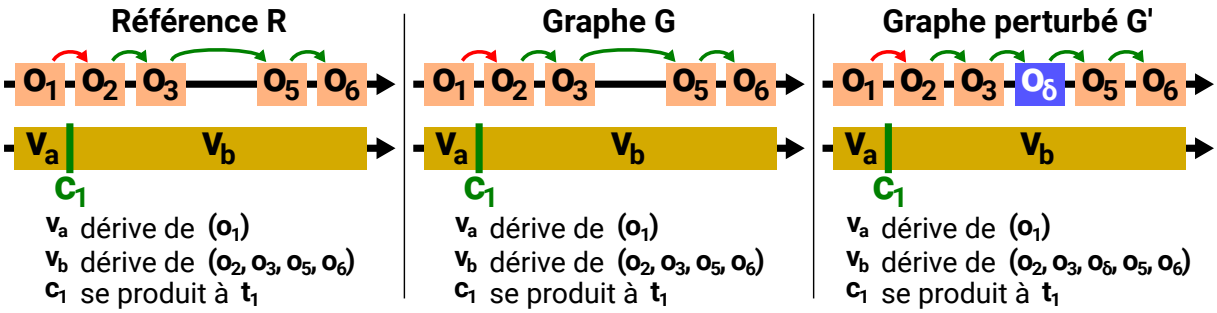


FIGURE 7.8 – Lignes de vie de l'adresse *2 rue du Pas de la Mule* obtenues de trois manières : construction du jeu de référence R, peuplement du graphe G avec la méthode proposée, et peuplement du graphe G' avec la même méthode mais avec l'ajout de données fragmentaires. $\rightarrow$: *a une géométrie différente de*, $\rightarrow$: *a une géométrie similaire à*.

fidélité de la reconstruction, et entre G et G' pour évaluer la stabilité des évolutions face à l'ajout de données supplémentaires.

7.5.3.1 Comparaison des évolutions des adresses dans le graphe reconstruit et le graphe de référence

L'objectif de cette étape d'évaluation est de vérifier si notre méthode de peuplement préserve l'ordre temporel des entités géographiques à travers les différentes observations successives. Une fois le jeu de référence R établi, nous construisons le graphe G en appliquant la méthode de peuplement avec les mêmes sources de données. Afin d'évaluer G, nous effectuons une comparaison entre R et G au niveau des adresses. Ainsi, pour chaque adresse du jeu de référence, nous vérifions si le graphe G décrit bien les évolutions telles qu'elles sont dans R.

Dans le graphe G, l'évolution d'une adresse est décrite à travers la cible qu'elle désigne, c'est-à-dire ici le numéro d'immeuble qui lui est associé. Par exemple, pour l'adresse « 2 rue du Pas de la Mule », l'entité géographique centrale est sa cible, c'est-à-dire le numéro d'immeuble *2*. L'évolution d'une adresse coïncidant avec celle de son numéro, la vérification consiste donc à s'assurer que l'évolution du numéro reconstruite dans G est cohérente avec celle spécifiée dans R.

Nous vérifions tout d'abord que les événements d'apparition et de disparition du numéro d'immeuble dans G sont temporellement compatibles avec les observations de R : la date d'apparition doit être antérieure à la première observation disponible, et la date de disparition postérieure à la dernière où le numéro est attesté. Ensuite, nous examinons l'évolution des attributs du numéro, en particulier le nom et la géométrie. Concernant le nom, le critère d'identité repose sur l'hypothèse qu'il n'existe qu'une seule version de l'attribut de type *nom*. Nous vérifions cette hypothèse dans G au moyen de la requête SPARQL fournie par le Listing 7.8 en s'assurant que la valeur de `?nb` est égale à 1 pour chaque adresse, confirmant qu'une seule version de nom a été reconstruite comme attendu.

Pour les géométries, la comparaison est moins triviale. Nous exigeons tout d'abord que le nombre de versions d'attributs géométriques dans G corresponde à celui observé dans R. Ainsi, dans le cas du numéro « *2 rue du Pas de la Mule* », où R comporte deux versions d'attribut de type géométrie (notées v_a et v_b) pour l'entité géographique, la méthode doit également en produire deux (notées v_α et v_β) dans G. Cependant, l'égalité du nombre de versions ne garantit pas la validité du processus de reconstruction. Nous exploitons pour cela l'information sur les observations d'origine de chaque version. Dans

```
SELECT ?addr (COUNT(?vers) AS ?nb) WHERE {
    BIND(ex:facts AS ?g)
    GRAPH ?g {
        ?addr :targets ?target .
        ?target :hasAttribute [:isAttributeType atype:Name ;
        ↪  :hasAttributeVersion ?vers]
    }
}
GROUP BY ?addr
```

Listing 7.8 – Requête SPARQL identifiant le nombre de version d'attributs de type nom pour des cibles d'adresses. `ex:facts` désigne le graphe nommé des faits.

R, v_a dérive de l'observation o_1, tandis que v_b dérive des observations o_2, o_3, o_4 et o_5. Pour que G soit correct, v_α et v_β doivent dériver respectivement des mêmes ensembles d'observations. En effet, si deux versions dérivent du même ensemble d'observations, alors leurs géométries sont nécessairement similaires, ce qui permet d'établir une bijection entre $\{v_a, v_b\}$ et $\{v_\alpha, v_\beta\}$. Cette correspondance permet d'identifier les paires de versions potentiellement similaires entre R et G. Si la correspondance est établie, nous comparons leurs intervalles de validité temporelle pour vérifier qu'ils coïncident également. Le cas échéant, nous considérons que l'évolution de a dans G n'a été correctement reconstruite.

Ainsi, la comparaison entre R et G repose sur la cohérence des événements temporels, sur l'unicité attendue des versions d'attributs de type *nom*, et sur une correspondance structurée entre versions d'attributs géométriques appuyée à la fois sur les observations d'origine et sur les intervalles de validité. Afin de quantifier la similarité entre le jeu de référence R et le graphe reconstruit G, nous introduisons un score de correspondance global pour chaque adresse. Pour qu'une adresse $a \in A$ soit considérée comme correctement reconstruite, trois conditions doivent être simultanément vérifiées :

1. la correspondance temporelle des événements d'apparition et de disparition du numéro d'immeuble, c'est-à-dire que les intervalles de validité dans G recouvrent ceux de R ;
2. l'unicité et la cohérence de l'attribut *nom*, c'est-à-dire qu'une seule version d'attribut de nom est associée à l'adresse dans G ; il n'y a pas de comparaison directe à R pour cet attribut, car le critère utilisé pour lier les observations et créer le jeu de référence reposait justement sur le nom. Cela implique que chaque adresse dans R possède systématiquement une seule version d'attribut de nom, rendant la comparaison directe inutile.
3. la correspondance des versions d'attributs géométriques, établie à l'aide de la bijection entre versions de R et G basée sur les observations d'origine et les intervalles de validité temporelle.

Nous définissons alors un score binaire pour chaque adresse a :

$$s(a) = \begin{cases} 1 & \text{si les trois conditions ci-dessus sont satisfaites,} \\ 0 & \text{sinon.} \end{cases}$$

Le score global sur l'ensemble des adresses A est alors :

$$S(R, G) = \frac{1}{|A|} \sum_{a \in A} s(a),$$

Ainsi, $S(R, G) = 1$ indique une reconstruction parfaitement cohérente avec le jeu de référence, et $S(R, G) = 0$ correspond à une absence totale de correspondance.

7.5.3.2 Test de robustesse à l'ajout de données fragmentaires

La principale limite de cette première évaluation est que la méthode de peuplement du graphe est appliquée à un ensemble de données ne décrivant que des *snapshots*. Ces données ne sont pas fragmentaires et il n'y a pas de chevauchement entre les *snapshots* : chacun possède des temps valides distincts, et les données représentent uniquement des états spatio-temporels mais aucun événement. L'objectif de cette seconde évaluation est de tester la robustesse de l'algorithme de peuplement à l'ajout de données fragmentaires : on évalue si l'introduction de telles données modifie les évolutions des adresses précédemment reconstruites.

Nous disposons déjà d'un graphe G construit à partir d'un ensemble de données D. Nous construisons ensuite deux graphes supplémentaires G' distincts afin d'évaluer la robustesse du processus de peuplement. Le premier, noté G_{st}, est obtenu à partir du même ensemble D auquel nous adjoignons uniquement l'ensemble de données fragmentaires d_{st}. Le second, noté $G_{\text{st+ev}}$, est construit à partir de D enrichi des deux ensembles de données fragmentaires d_{st} et d_{ev}. Ces données fragmentaires ont été générées de manière à ne pas entrer en conflit avec les observations de D. Ainsi, leur ajout ne devrait pas modifier l'évolution des adresses : pour chaque adresse, la ligne de vie reconstruite dans G doit être similaire à celle reconstruite dans G_{st} et dans $G_{\text{st+ev}}$.

Pour évaluer l'impact de l'ajout de données fragmentaires (d_{st} et d_{ev}) sur les *lignes de vie* des adresses reconstruites, on se concentre sur les attributs de type géométrie de la cible de l'adresse. La raison est que ce sont ceux pour lesquels l'ajout de données fragmentaires est susceptible de modifier l'évolution, et donc d'influer directement sur les résultats de la méthode de peuplement. On pourrait étendre cette comparaison aux attributs de type nom, mais comme rappelé précédemment, le nom est utilisé comme critère de liage des observations, ce qui implique qu'il existe toujours une seule version par adresse. Deux scores sont introduits pour faire cette évaluation. Le premier mesure la *cohérence structurelle* : il s'agit de vérifier que, pour chaque attribut, les versions générées après l'intégration des données fragmentaires demeurent similaires à celles du graphe de référence G. Autrement dit, le nombre de versions doit rester identique et chacune d'elles doit présenter une géométrie comparable à la version correspondante dans G (sans prise en compte ici de la dimension temporelle). Le second évalue la *cohérence temporelle* : il vérifie que les intervalles de validité associés à chaque version n'ont pas été modifiés par l'ajout de données fragmentaires.

Cohérence structurelle On s'assure que les versions générées dans G' sont structurellement cohérentes avec celles de G, c'est-à-dire que l'agencement des versions est similaire. Par exemple, si dans G, un attribut de géométrie possède la suite de versions $v_1 \to \cdots \to v_n$, alors dans G' on doit retrouver une suite $v'_1 \to \cdots \to v'_n$ associée au même attribut, telle que chaque paire (v_i, v'_i) soit équivalente. Cette condition formalise la correspondance des versions d'attributs géométriques, établie à l'aide d'une bijection entre versions de G et G' à partir des observations d'origine. Elle reflète ainsi l'agencement structurel des versions tel que défini précédemment, sans prendre en compte la dimension temporelle. Pour formaliser la notion de correspondance entre les ensembles de valeurs associées à un attribut a dans deux graphes G et G', nous introduisons d'abord une fonction $\text{hasSameTraces}(v, v')$, qui compare les traces associées à deux valeurs v et v' :

$$\text{hasSameTraces}(v, v') = \begin{cases} 1 & \text{si } \{o \in D \mid \texttt{hasTrace}(v, o)\} = \{o \in D \mid \langle \texttt{hasTrace}(v', o)\}, \\ 0 & \text{sinon.} \end{cases}$$

En notant $V_G(a)$ (respectivement $V_{G'}(a)$) l'ensemble des versions d'attribut de type géométrie de l'adresse a dans G (respectivement dans G') et en utilisant la fonction hasSameTraces, la correspondance entre les ensembles de valeurs $V_G(a)$ et $V_{G'}(a)$ s'exprime avec la fonction same :

$$\text{same}(V_G(a), V_{G'}(a)) = \begin{cases} 1 & \text{si } \forall v \in V_G(a),\ \exists v' \in V_{G'}(a), \text{hasSameTraces}(v, v') = 1, \\ 0 & \text{sinon.} \end{cases}$$

Ainsi, $\text{same}(V_G(a), V_{G'}(a)) = 1$ signifie que chaque valeur de $V_G(a)$ possède au moins une correspondance exacte dans $V_{G'}(a)$. Autrement dit, pour une adresse a, les versions de son attribut de géométrie dans G sont identiques à celles présentes dans G', si l'on fait abstraction de leurs intervalles de validité temporelle.

Le score structurel pour l'adresse a est alors défini par : $s_{\text{stru}}(a) = \text{same}(V_G(a), V_{G'}(a))$ et donc le score structurel global sur l'ensemble des adresses est :

$$S_{\text{stru}}(G, G') = \frac{1}{|A|} \sum_{a \in A} s_{\text{stru}}(a).$$

Par exemple, pour l'adresse « *2 rue du Pas de la Mule* », et comme décrit précédemment dans la section 7.5.3.1, supposons que nous ajoutions une observation supplémentaire o_α dont le temps valide est $[1797; 1809]$ et qui possède une géométrie similaire à o_1. Dans le graphe G' généré par la méthode de peuplement, l'attribut de type géométrie est alors représenté par deux versions : v'_α dérivant des observations $\{o_1, o_\alpha\}$ et v'_β la succédant, dérivant de $\{o_2, o_3, o_4, o_5\}$. Dans G, sans o_α en entrée, on retrouve la même structure : v_α dérivant de o_1 et v_β de $\{o_2, o_3, o_4, o_5\}$. Ainsi, la correspondance structurelle est parfaitement respectée et le score $\text{same}(V_G(a), V_{G'}(a))$ vaut 1 pour cette adresse. En revanche, cette analyse ne permet pas de conclure sur la cohérence des temps valides : il est possible que v_α et v'_α n'aient pas exactement le même intervalle temporel, ce qui est abordé dans le paragraphe suivant.

Cohérence temporelle Le score structurel précédent ne garantit pas que les intervalles de validité des versions d'attributs sont cohérents. Pour évaluer la cohérence temporelle, on exploite le fait que l'information sur le temps valide est associée aux événements qui induisent les changements sur les versions.

Soit $C_G(a)$ et $C_{G'}(a)$ les ensembles des changements appliqués à un attribut de type géométrie pour l'adresse a dans les graphes G et G', respectivement. Pour chaque changement $c \in C_G(a)$, on définit :

$$\text{changeTime}(c) = t \quad \text{tel que} \quad \texttt{dependsOn}(c, e) \wedge \texttt{hasTime}(e, t).$$

Ensuite, la correspondance temporelle entre les ensembles de changements $C_G(a)$ et $C_{G'}(a)$ s'exprime avec la fonction sameTimes :

$$\text{sameTimes}(C_G(a), C_{G'}(a)) = \begin{cases} 1 & \text{si } \forall c \in C_G(a),\ \exists c' \in C_{G'}(a), \\ & \text{changeTime}(c) = \text{changeTime}(c'), \\ 0 & \text{sinon.} \end{cases}$$

Pour l'adresse a, $s_{\text{temp}}(a) = \text{sameTimes}(C_G(a), C_{G'}(a))$ définit son score de cohérence temporelle, ce qui permet d'introduire le score global sur l'ensemble des adresses :

$$S_{\text{temp}}(G, G') = \frac{1}{|A|} \sum_{a \in A} s_{\text{temp}}(a).$$

Ainsi, $S_{\text{temp}}(G, G') = 1$ signifie que chaque changement dans G possède au moins un changement correspondant dans G' appliqué au même instant, garantissant la cohérence temporelle des versions d'attributs de type géométrie des adresses. En ajoutant des données fragmentaires décrivant des événements, on s'attend à ce que ces informations complètent et affinent les connaissances temporelles, ce qui conduit à un score $S_{\text{temp}}(G, G')$ relativement faible dans ce cas.

7.5.4 Résultats de l'évaluation

Évaluation 1 - Comparaison avec le jeu de référence. Dans le jeu de référence R, à partir des observations issues des six sources ayant fait l'objet d'un processus de liage, nous obtenons 7874 adresses. Le graphe G, résultat de l'application de la méthode de peuplement, comprend lui aussi 7874 adresses, toutes alignées avec celles du jeu de référence.

Le score $S(R, G)$ obtenu est de $0,9998$, ce qui indique que la méthode de peuplement a échoué à reconstruire correctement l'évolution d'une seule adresse. Cette divergence est due à un effet de seuil lors du calcul de la similarité des géométries de deux observations successives. En effet, nous avons fixé un seuil $d_{\max}$ de 5 mètres pour considérer deux géométries comme identiques. Pour l'adresse concernée, deux de ses observations, o_i et o_j, ont chacune une géométrie distantes de 5 mètres. Alors que la construction du jeu de référence R a calculé une distance $d_{ij} \leq d_{max}$, la méthode de peuplement a évalué une distance légèrement supérieure, $d_{ij} = 5.001 > d_{max}$, ce qui crée une divergence et entraîne la petite différence dans le score global. Malgré cet échec localisé, le score très proche de 1 montre que la méthode est globalement capable de reconstruire correctement l'évolution des adresses tout en préservant l'ordre temporel des entités à travers les *snapshots* successifs.

Évaluation 2 - Robustesse face à l'ajout de données fragmentaires. Pour cette seconde évaluation, nous examinons la robustesse de la méthode lorsqu'elle est confrontée à des données fragmentaires ajoutées artificiellement. Comme dans l'expérience précédente, le nombre d'adresses considérées reste inchangé. Un ensemble de 5821 observations supplémentaires décrivant des états de géométrie a été généré, ainsi que 7874 événements fictifs indiquant des changements de géométrie à différents instants. Ces données synthétiques ne sont pas ajoutées sur l'ensemble des entités mais sur un échantillon d'adresses, de manière à tester l'impact d'informations incomplètes et hétérogènes.

À partir de ces données, nous construisons les graphes G_{st} (ajout des seuls états) et $G_{\text{st+ev}}$ (ajout des états et des événements), que nous comparons ensuite au graphe G selon les deux scores définis précédemment, structurel (S_{stru}) et temporel (S_{temp}). Les résultats correspondants sont présentés dans le tableau 7.3.

En perturbant le graphe G uniquement avec les états des adresses — c'est-à-dire sans introduire de nouveaux événements — les deux scores de cohérence structurelle et temporelle sont égaux à 1. Cela indique que l'ajout de ces états fragmentaires ne modifie pas l'agencement des versions d'attributs géométriques ni les instants associés aux changements.

	$S_{\text{stru}}(G, G')$	$S_{\text{temp}}(G, G')$
$G' = G_{\text{st}}$	1	1
$G' = G_{\text{st+ev}}$	1	0,1330

TABLEAU 7.3 – Scores de cohérence structurelle et temporelle mesurant l'impact des données fragmentaires sur la reconstruction de l'évolution des adresses.

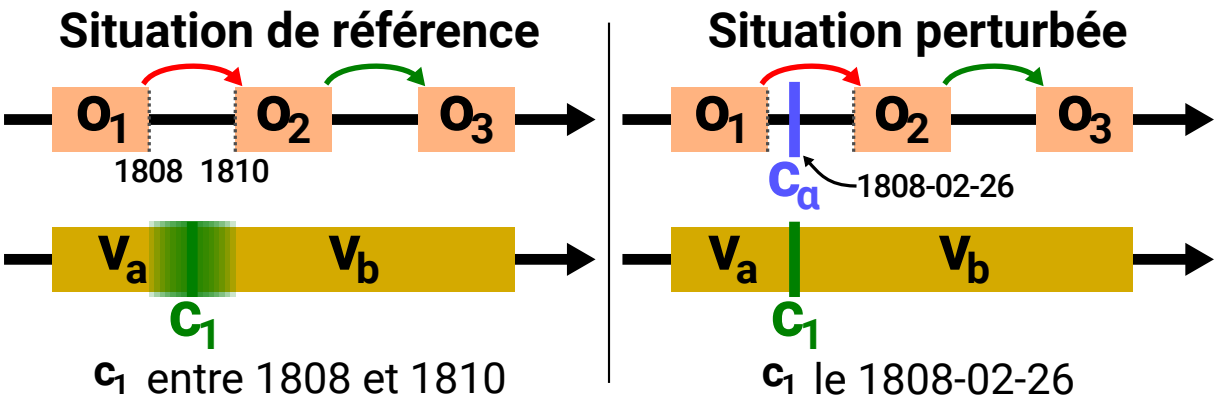


FIGURE 7.9 – Comparaison de deux lignes de vie de l'adresse *2 rue du Pas de la Mule* obtenues avec la méthode de peuplement proposée. La situation de référence est obtenue à partir des observations initiales, tandis que la situation perturbée est générée à partir des mêmes observations, auxquelles est ajouté un changement fictif c_α. $\rightarrow$: *a une géométrie différente de*, $\rightarrow$: *a une géométrie similaire à*.

En revanche, lorsque G est perturbé par l'ajout simultané des états et des événements associés aux adresses, le score structurel reste égal à 1, confirmant que la séquence des versions d'attributs géométriques est préservée. Le score de cohérence temporelle, en revanche, chute à 0,1330, ce qui était attendu. Cela reflète que l'introduction de nouveaux événements modifie les instants des événements impliquant les changements, ce qui est attendu puisque ces événements fournissent une information plus précise sur les temps valides des versions. Reprenons l'exemple de l'adresse « 2 rue du Pas de la Mule » illustré par la figure 7.9. Dans la situation de référence, c'est-à-dire dans le graphe G, deux versions d'attributs de géométries v_a et v_b sont initialement séparées par un changement c_1 associé à un instant flou (entre 1808 et 1810). L'ajout d'un événement datant du 26 février 1808 et impliquant un changement c_α permet désormais d'en déduire les temps valides de manière plus précise.

Ces résultats montrent que la méthode préserve la structure des évolutions des adresses (versions et traces) même en présence de données fragmentaires, tout en ajustant de manière appropriée les temps valides lorsque de nouveaux événements sont introduits. La reconstruction est donc robuste et cohérente, tant sur la dimension structurelle que temporelle.

Outil de visualisation et évaluation qualitative Pour faciliter l'interprétation et l'inspection manuelle, nous avons développé un outil de visualisation permettant d'afficher l'évolution des entités géographiques dans le temps ou l'état du territoire à un instant donné. Cet outil permet des évaluations qualitatives du graphe reconstruit et ouvre la voie à des analyses exploratoires plus approfondies. Les travaux futurs pourraient inclure des évaluations qualitatives basées sur des données fragmentaires synthétiques spécifiquement

conçues pour perturber l'évolution des entités. De telles expériences permettraient de vérifier si la méthode se comporte de manière cohérente et prédictible lorsqu'elle est confrontée à des données temporelles erronées ou contradictoires.

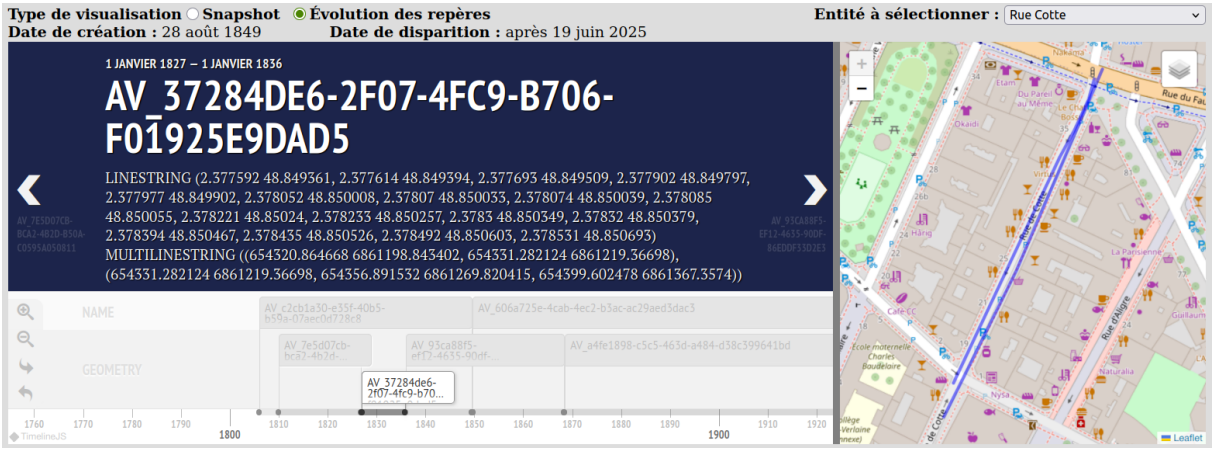


FIGURE 7.10 – Outil de visualisation montrant l'évolution des entités géographiques, ici focalisé sur la *Rue de Cotte*. La géométrie affichée est valide de 1827 à 1836.

La figure 7.10 illustre la fonctionnalité de l'outil de visualisation affichant l'évolution d'une entité. Il fournit une frise chronologique pour chaque attribut d'une entité géographique sélectionnée. En comparant les résultats de la reconstruction automatique aux archives historiques, nous observons une cohérence générale des données. Cependant, certaines incohérences subsistent, provenant soit de conflits entre sources, soit de critères de similarité trop stricts appliqués aux versions successives (voir sous-sous-section 6.4.2.2).

Par exemple, le graphe signale un changement de nom au 28 août 1849, où la rue de Cotte devient rue de Cotte. L'erreur ici ne provient pas de notre méthode, mais de l'interprétation des sources historiques. La nomenclature des voies de Paris indique qu'un arrêté du 28 août 1849 concerne la dénomination de cette voie. En réalité, c'est la *rue Trouvée* qui est absorbée par elle. L'arrêté entraîne donc, dans nos données, l'application d'un changement sur l'attribut de type nom de l'entité, conservé lors de la construction du graphe de faits et créant ainsi une incohérence temporelle. Cette interprétation erronée est confirmée par sa présence sur le plan de Verniquet, antérieur à l'arrêté. À l'inverse, notre méthode infère un changement de géométrie dans les années 1850, alors qu'aucune source ne le mentionne, ce qui résulte d'un défaut de fusion des versions jugées à tort insuffisamment similaires.

Par ailleurs, l'outil de visualisation temporelle a été utilisé pour générer des *snapshots* (des instantanés) du territoire à des dates spécifiques, choisies en fonction de leur correspondance avec les intervalles de validité des sources ayant servi à la construction du graphe. La figure 7.11 en est un exemple. Elle représente l'outil de visualisation affichant le résultat d'une requête sur le graphe, permettant de sélectionner l'état du territoire parisien en 1896. Les géométries en vert correspondent aux entités dont l'existence est attestée avec certitude par le graphe à cette date. En revanche, les entités affichées en rouge présentent une incertitude : le graphe indique que la date la plus proche attestant leur existence en 1896 est inférieure à un seuil fixé ici à 20 ans.

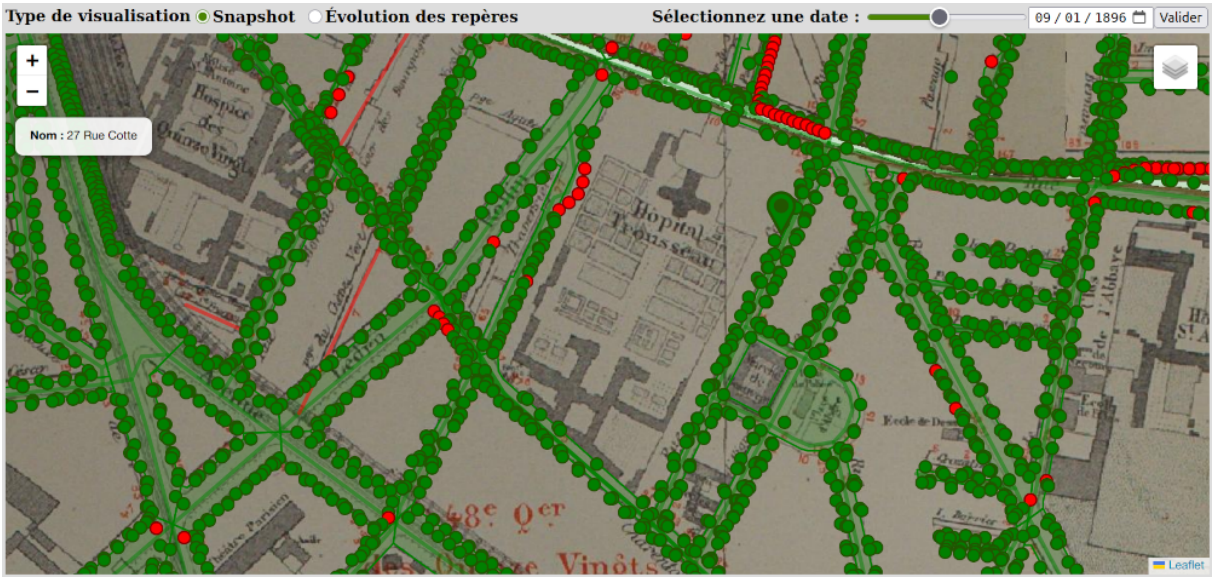


FIGURE 7.11 – Outil de visualisation montrant un instantané du secteur autour de la *Rue de Cotte* en 1896. La carte de fond est le plan municipal de 1888.

Conclusion

Ce chapitre a présenté la mise en œuvre de la méthode de peuplement de l'ontologie *PeGazUs* pour générer d'un graphe de connaissances reconstituant l'évolution des entités géographiques à partir de données hétérogènes et fragmentaires.

La mise en œuvre s'est appliquée au quartier du Faubourg Saint-Antoine en utilisant des sources cartographiques, textuelles et numériques. Cette étape a permis d'illustrer la diversité et le caractère fragmentaire des données disponibles.

La mise en application de la méthode de peuplement a été réalisée à l'aide de scripts Python et de requêtes SPARQL, permettant de transformer les données selon l'ontologie définie et de créer le graphe de faits. La vérification de la cohérence du graphe a ensuite été menée en s'appuyant sur des contraintes définies sur les ressources et les propriétés, garantissant l'intégrité et la cohérence des relations établies entre les entités.

L'évaluation de la méthode a consisté à comparer le graphe reconstitué avec un jeu de données de référence, ainsi qu'à tester l'introduction de données fragmentaires. Les résultats montrent que la méthode permet de restituer fidèlement les évolutions historiques et que l'ajout de nouvelles données cohérentes n'entraîne pas de perturbations significatives dans le graphe. Ces observations confirment la robustesse de la méthode et sa capacité à construire un référentiel géohistorique.

Enfin, le graphe construit offre un potentiel intéressant pour l'analyse des dynamiques urbaines. Des outils comme l'outil de visualisation développé permettent de représenter l'état du territoire à un instant donné et de suivre l'évolution de chaque entité géographique. En exploitant les données du graphe, cet outil ouvre la voie à des analyses des transformations urbaines. Ces résultats montrent que le graphe généré, combiné à des outils permettant de l'exploiter, constitue un moyen prometteur pour enrichir la compréhension des évolutions urbaines à partir de données fragmentaires et hétérogènes.

Chapitre 8

Discussion et perspectives

Dans cette thèse, nous avons introduit une ontologie permettant la représentation des adresses dans leur diversité ainsi que celles des entités auxquelles elles font référence tout en proposant la représentation de leurs évolutions. Nous avons aussi proposé une approche de création et de peuplement automatique d'un graphe de connaissances géohistorique à partir de sources diverses contenant des données hétérogènes et fragmentaires. Ces propositions ont été appliquées à un ensemble de données issues de sources décrivant le territoire parisien depuis la fin du XVIIIe siècle.

Ce chapitre présente les limites des travaux effectués durant cette thèse et les pistes d'amélioration que nous entrevoyons. Il est structuré en plusieurs parties. La section 8.1 s'intéresse à la représentation des incertitudes dans l'ontologie, en abordant notamment la provenance des données, l'enrichissement collaboratif, et la gestion des imperfections géométriques. Par la suite, la section 8.2 discute des limites de la mise en œuvre du peuplement automatique de l'ontologie, tandis que la section 8.3 propose une réflexion critique autour du critère d'identité des entités géographiques et son influence sur le référentiel construit. La section 8.4 suggère d'utiliser davantage les relations entre entités géographiques pour améliorer leur localisation. La section 8.5 propose une exploitation plus fine des événements et enfin la section 8.6 explore les perspectives autour des documents textuels pour enrichir le graphe de connaissances.

8.1 Représentation des incertitudes dans l'ontologie

8.1.1 Incertitudes et provenance des données

Le travail sur l'ontologie présenté au chapitre 4 a permis de représenter les dynamiques spatio-temporelles et certaines incertitudes, notamment par l'intégration des temps flous. Cependant, la représentation des imperfections directement dans le graphe demeure partielle, en particulier pour les géométries : à ce stade, celles-ci sont simplement extraites telles qu'elles apparaissent dans les sources, sans indication sur leur fiabilité ou leur précision.

Il est important de rappeler que les données utilisées ne sont pas parfaites, en grande partie en raison de la nature des sources historiques et des limites des processus d'extraction d'informations. Pour pallier ces limitations, au moins en partie, le graphe explicite systématiquement la provenance des données : chaque assertion est liée à sa source, ce qui permet de contextualiser sa création et de mettre en perspective, voire de remettre en question, les informations qu'elle contient. En parallèle, on sépare les factoïdes des faits, dont les triplets sont stockés dans des graphes nommés distincts. En revanche, aucune in-

formation n'est actuellement fournie sur la qualité ou les imperfections des données, ce qui pourrait avoir un impact important sur le processus de construction du référentiel. Pour améliorer la gestion des conflits entre données, il serait pertinent d'attribuer des poids aux factoïdes selon leur fiabilité, de manière à ce que des assertions issues de sources plus fiables aient un impact plus important en cas de contradiction avec d'autres informations.

Ces contradictions peuvent se manifester de différentes manières. Il arrive que plusieurs sources décrivent une même entité sur des périodes qui se chevauchent, et que les valeurs de certains attributs divergent. Par exemple, l'emprise spatiale d'une même rue peut être représentée de façon très différente selon les sources, entraînant la coexistence simultanée de deux versions contradictoires d'un même attribut. Des divergences peuvent également exister au niveau des événements : un même événement peut se voir attribuer des dates différentes selon les sources consultées. Enfin, il peut exister des contradictions entre descriptions d'état et événements : une source peut indiquer l'apparition d'une voie à une certaine date alors qu'une autre atteste de l'existence effective de cette voie sur une période antérieure. C'est le cas, par exemple, de la rue Gérard : la nomenclature des voies de Paris mentionne sa création en 1851, tandis que le cadastre napoléonien, dont la période valide est 1845-1849, indique que la rue existait déjà.

8.1.2 Enrichissement collaboratif

Afin de corriger les conflits entre les sources ou d'autres incohérences dans le graphe, un enrichissement collaboratif du référentiel pourrait être bénéfique. La construction cumulative par les contributions d'utilisateurs nécessite la mise en place d'un outil intuitif et simple d'utilisation, à l'image de l'éditeur iD [1] pour OpenStreetMap ou des interfaces d'édition de Wikimedia. La qualité de ces apports doit être assurée par des mécanismes de contrôle, permettant de limiter le vandalisme et d'encourager des contributions sourcées et vérifiables.

Il est particulièrement important que l'utilisateur n'ait pas besoin de maîtriser le vocabulaire de l'ontologie, qui peut être complexe et facilement mal interprété. Par exemple, l'éditeur iD est assez intuitif, mais certaines opérations restent difficiles pour un débutant en raison des spécificités d'OSM : représenter un polygone peut être fait par le biais d'un chemin fermé ou bien par une collection de chemins où chacun a un rôle défini.

Différentes stratégies pourraient être envisagées pour faciliter l'interaction :

— proposer à l'utilisateur des cas spécifiques et lui demander de décider s'il faut effectuer des corrections (Shah et Zhou, 2016) ;
— laisser l'utilisateur naviguer librement et effectuer des modifications dès qu'il identifie un problème (Haklay, 2010).

Pour la validation des contributions, deux approches sont possibles :

— considérer que la dernière correction réalisée est la bonne, comme c'est le cas dans OSM ou les projets liés à Wikimedia (Bennett, 2010) ;
— utiliser une conciliation des retours de plusieurs utilisateurs pour déterminer la version finale la plus fiable (Karger *et al.*, 2011).

Il serait également pertinent de considérer ces contributions comme des informations à part entière, stockées dans un graphe nommé et facilement sourçable selon l'ontologie PROV-O. Cela permettrait de reconstruire le graphe de faits de manière cumulative, sans nécessiter que les utilisateurs refassent des contributions similaires à chaque étape de reconstruction du graphe, tout en tenant compte des différents niveaux de fiabilité et de provenance des données.

1. https://wiki.openstreetmap.org/wiki/ID

8.1.3 Représentation des imperfections géométriques

La typologie proposée par de Runz (2008) pourrait être intégrée à l'ontologie PeGazUs afin de représenter de manière plus fine les imperfections, notamment pour les géométries. Dans le graphe de faits, une même entité géographique peut être associée à plusieurs géométries issues de sources distinctes. Par exemple, une version v_1 d'une entité peut avoir trois représentations g_1, g_2 et g_3, provenant respectivement des sources s_1, s_2 et s_3. Il serait alors pertinent de définir une géométrie consolidée qui concilie ces différentes valeurs, tout en reflétant l'incertitude inhérente à l'information. Une telle approche pourrait s'inspirer des modèles de géométrie floue proposés par Schneider (1999), qui distinguent trois types de géométries floues : *fpoint*, *fline* et *fregion*. Le *fpoint* correspond à une union de points flous, chaque point étant associé à un degré d'appartenance. Une contrainte spécifique, la ligne floue (non *fline*), peut être appliquée pour que les points se situent le long d'une courbe continue, ce qui est particulièrement utile pour représenter les numéros d'immeuble par interpolation. Les *flines* et *fregions* suivent des principes similaires, mais présentent des différences adaptées à des entités linéaires ou surfaciques. Ce modèle de géométrie floue a été implémenté par Dilo *et al.* (2007), utilisant une région noyau et un support, et le degré d'appartenance de chaque point étant calculé par triangulation. L'intégration de ces concepts dans le graphe permettrait non seulement de mieux formaliser l'incertitude géométrique et de concilier des sources divergentes, mais aussi d'envisager la réalisation d'opérations spatiales directement dans les requêtes SPARQL sur des géométries floues en s'inspirant notamment des travaux de Cadorel *et al.* (2022b) et Car (2024).

L'incertitude doit également être considérée avec prudence : si la présence répétée d'une information dans plusieurs sources peut renforcer la confiance, il faut tenir compte des effets de propagation des erreurs, comme lorsque des plans ultérieurs reprennent des informations erronées issues de documents précédents. Enfin, certaines informations figurant dans les sources consultées décrivent des entités projetées, telles que des rues à percer ou des adresses réservées pour de futurs bâtiments (Pinon *et al.*, 2014). Ces entités peuvent ne jamais avoir existé, avoir été modifiées, c'est-à-dire que la conception initiale de l'entité diffère de sa réalisation sur le terrain (par exemple, une rue planifiée pour 300 mètres qui n'en fait finalement que 200), ou avoir été reportées, c'est-à-dire prévues à une date donnée mais réalisées ultérieurement (par exemple, un percement prévu en 1875 mais exécuté en 1880). Cela souligne les limites dans l'intégration de ces données lors du peuplement de l'ontologie et la nécessité d'une interprétation critique. Il conviendrait alors de distinguer les informations correspondant à des observations réelles de celles relevant de projections ou de plans futurs, afin de décider comment les intégrer dans le graphe de connaissances, par exemple en les annotant comme observations prospectives.

8.2 Limites de la mise en œuvre du peuplement de l'ontologie

La mise en œuvre et l'évaluation du référentiel, présentées au chapitre 7, ont également révélé plusieurs défis. Sur le plan opérationnel, le code associé à la mise œuvre avait d'abord été testé sur un petit quartier. Lorsqu'il a été appliqué à une emprise plus étendue, en intégrant davantage de sources, le temps de traitement a considérablement augmenté : plusieurs optimisations ont donc été nécessaires pour diminuer le temps de traitement. D'une part, la méthodologie de peuplement a été homogénéisée : le code de transformation des données issues des sources a été unifié pour générer des triplets compatibles avec l'ontologie *PeGazUs*, évitant ainsi de créer un script spécifique pour chaque source.

D'autre part, la construction du graphe de faits a été optimisée afin de réduire drastiquement le temps de traitement, notamment par l'amélioration des requêtes SPARQL. Les perspectives d'amélioration portent sur l'application à l'ensemble de Paris et l'intégration de sources supplémentaires. Certaines opérations, en particulier dans les requêtes, restent coûteuses, mais le principal facteur limitant est l'inférence réalisée par GraphDB après chaque requête, générant ou supprimant des triplets implicites. L'optimisation de cette étape est donc un enjeu majeur du passage à l'échelle de l'approche.

Du point de vue théorique, la construction d'un graphe complet à partir des sources reste possible, mais dépend fortement de leur disponibilité et de leur exhaustivité. Par exemple, le peuplement d'un graphe sur le quartier parisien de la Butte-aux-Cailles, appartenant à Gentilly jusqu'en 1860, a montré que même en combinant plusieurs sources, comme le cadastre napoléonien et celles présentées dans la section 7.1, certaines lacunes demeurent inévitables. Toutefois, l'identification de ces lacunes constitue en elle-même un résultat précieux, car elle permet de cibler les zones d'incertitude et d'orienter la recherche de nouvelles sources pertinentes. L'approche proposée favorise ainsi une accumulation progressive des connaissances : par exemple, si l'on dispose d'une centaine d'adresses non localisées en 1840, l'analyse du graphe permet de déterminer quelles sources pourraient combler ces manques, de les intégrer et de mettre à jour automatiquement la représentation des entités correspondantes dans le graphe.

L'absence de vérité terrain accessible conduit à adopter des heuristiques pour construire une référence approximative ; ce graphe de connaissances de référence, construit, peut se rapprocher de la réalité mais reste par nature incomplet et jamais définitif.

Dans le cadre de notre démarche d'évaluation, nous avons exploité un ensemble de *snapshots* représentant l'état du territoire à différentes périodes. Néanmoins, l'application de notre méthode avec cet ensemble présente plusieurs limites. D'une part, les jeux de données disponibles ne se chevauchent pas dans le temps et ne décrivent que des états, à l'exception des données synthétiques, ce qui restreint la diversité des situations testées et ne permet pas idéalement d'évaluer la capacité de la méthode à intégrer des événements. D'autre part, certains paramètres indépendants de la méthode, tels que les critères de similarité entre versions d'attributs, notamment les seuils utilisés pour comparer les noms et les géométries, peuvent influencer les résultats de l'évaluation. Il serait pertinent de faire varier ces paramètres et d'explorer différentes techniques de calcul de similarité afin d'évaluer la robustesse de la méthode et de s'assurer que ces paramètres n'influent pas lors du peuplement. De plus, bien que nous ayons ajouté des données fragmentaires cohérentes avec le jeu de référence, il serait intéressant de générer des données délibérément incohérentes pour mesurer l'impact de la méthode de peuplement sur le graphe final, en comparant ce que l'ajout de ces données devrait produire avec ce qui est effectivement généré. Enfin, pour évaluer la capacité de la méthode à reconstituer l'état du territoire en l'absence de données directes sur une période donnée, reprenons l'ensemble des jeux de données $D = \{d_1, \ldots, d_n\}$. Pour $i \in [1, n]$, nous générons un graphe G_i en appliquant la méthode sur le sous-ensemble $D \setminus \{d_i\}$, puis nous extrayons l'état du territoire dans G_i sur la période correspondant au temps valide de d_i. Cette comparaison permet de déterminer si les états reconstitués par la méthode sont proches de ceux effectivement observés dans d_i, et ainsi d'évaluer dans quelle mesure la méthode peut reconstruire fidèlement l'évolution du territoire malgré l'absence de données couvrant certaines périodes. Ces perspectives offrent des possibilités supplémentaires pour avoir des évaluations plus rigoureuses ou diversifiées, dans le but de mieux caractériser les limites et la fiabilité de la méthode de peuplement.

8.3 Critiques autour du critère d'identité

La méthode de peuplement détaillée au chapitre 6 fait apparaître plusieurs axes d'amélioration. La question de la similarité entre entités pourrait être affinée, afin de mieux gérer les conflits complexes. Par exemple, certains plans historiques montrent des voies en construction dont l'existence sur le terrain n'a jamais été effective, tandis que les noms prévus pour ces voies ont parfois été attribués à d'autres voies situées à proximité, rendant difficile la distinction entre les entités correspondantes dans les graphes de factoïdes et générant des erreurs dans le graphe de faits. De même, la question de l'identité des numéros d'immeuble se pose dans des situations où un numéro identique apparaît à des dates différentes dans des sources distinctes, associé à des coordonnées proches. Si le bâtiment initial a été démoli et remplacé par un autre, il pourrait être pertinent de considérer ces numéros comme appartenant à deux entités géographiques distinctes, et non pas similaires : ceci illustre la complexité de définir l'identité d'une entité géographique au fil du temps.

L'identité des entités géographiques constitue un élément central, difficile à définir de manière uniforme, et l'évolution observée dépend directement de la manière dont cette identité est conceptualisée. Par exemple, la place de la Nation à Paris a connu plusieurs noms au cours de son histoire : *place du Trône*, *place du Trône Renversé* pendant la Révolution française, *place du Trône* de 1815 à 1880, puis *place de la Nation* depuis. Avec le critère d'identité adopté dans notre approche, cette succession d'odonymes témoigne de l'existence de quatre entités distinctes se succédant dans le temps, sur une emprise géographique similaire. En revanche, en utilisant les critères d'identité proposés par Bernard *et al.* (2018) ou Charles et Hernandez (2023), la construction du graphe conduirait à considérer qu'il s'agit d'une seule entité ayant subi trois changements de nom. Cet exemple illustre comment le choix du critère d'identité influence directement la structure du graphe et la manière dont l'évolution des entités est représentée. Enfin, notre approche consistant à ne prendre en compte qu'une seule propriété pour définir l'identité d'une entité a ses limites. En effet, Bernard *et al.* (2025c) utilise l'ontologie *PeGazUs* pour représenter des parcelles et leurs évolutions qui sont décrites dans le cadastre napoléonien. Or, dès lors qu'une parcelle est scindée — comme le montre la figure 8.1 avec la parcelle n°1170 —, les parcelles résultantes héritent de l'identifiant de la parcelle d'origine. Notre critère d'identité n'est pas suffisant dans ce cas : pour permettre le suivi de l'évolution des parcelles, il doit combiner plusieurs paramètres, comme la géométrie et l'identité du propriétaire. De plus, il faut adapter en conséquence l'opération de mise en correspondance des observations issues des différentes sources.

8.4 Utiliser les relations spatiales entre entités géographiques pour améliorer la localisation

En plus des références spatiales contenues dans le graphe de connaissances, les relations spatiales entre entités géographiques constituent une source précieuse d'informations pour affiner ou estimer la localisation d'entités dont la position n'est pas connue avec précision. Dans notre graphe, de nombreuses relations sont décrites — d'appartenance, de voisinage, de contiguïté, d'inclusion ou encore de position relative — mais leur potentiel pour la reconstruction spatiale reste partiellement exploité. Comme présenté dans la sous-sous-section 4.2.1.2, plusieurs travaux antérieurs décrivent la manière de représenter des relations entre entités géographiques. D'autres travaux ont mis en lumière des approches

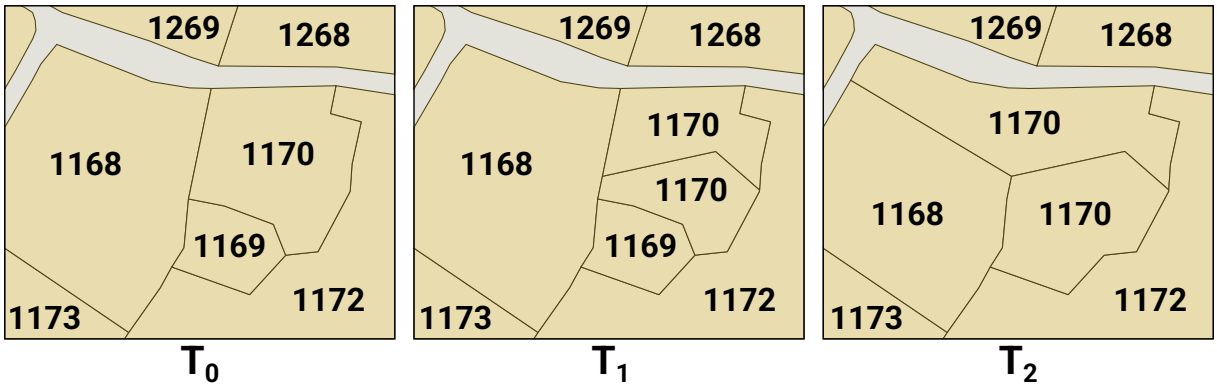
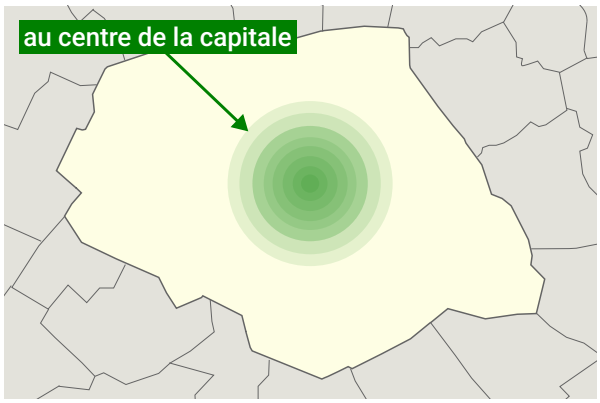


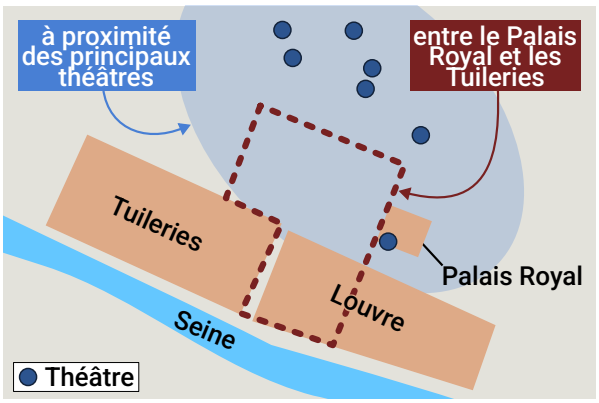
FIGURE 8.1 – Exemple d'évolution cadastrale où la parcelle initiale numérotée 1170 est subdivisée en plusieurs entités qui conservent le même identifiant.

utilisant ces informations afin d'inférer la position probable d'une entité inconnue à partir de celles de ses voisines (Dutta, 1991; Frank, 1992; Syed *et al.*, 2025). Prenons par exemple trois entités A, B et C dont les géométries respectives sont partiellement connues. Si les positions de A et B sont connues, et que l'on sait que A est au nord de C tandis que B est au sud de C, il devient possible d'estimer une position probable pour C. Ce raisonnement pourrait être formalisé afin de déduire une géométrie approximative ou floue pour C.

Cette approche rejoint les perspectives développées précédemment à propos de la représentation des géométries incertaines (voir sous-section 8.1.3). En combinant ainsi les relations spatiales et la représentation de l'incertitude géométrique, il serait envisageable de déduire et représenter la localisation d'entités entièrement dépourvues de références spatiales directes, en s'appuyant sur leurs relations. Elle pourrait notamment s'appliquer aux cas où les sources textuelles ne fournissent que des informations relatives, tels que « la rue Gérard prolonge la rue du Père Guérin » ou « la rue de Lobau est située derrière l'hôtel de ville », permettant aussi de reconstituer la topologie d'un espace urbain partiellement reconstitué.



(a) Représentation floue de la position d'une entité située « *au centre de la capitale* » (Paris).



(b) Représentation des positions d'une entité située « *entre le Palais-Royal et les Tuileries* » et « *à proximité des principaux théâtres* ».

FIGURE 8.2 – Estimation de la position de la cible de l'adresse dont le label est « *au centre de la capitale, entre le Palais-Royal et les Tuileries, à proximité des principaux théâtres* » au début du XIXe siècle.

De même, si l'on reprend l'exemple de l'adresse dont l'énoncé est « *au centre de la ca-*

pitale, entre le Palais-Royal et les Tuileries, à proximité des principaux théâtres » (voir figure 4.4), il serait possible d'estimer la position de sa cible en s'appuyant sur les segments d'adresse existants, comme l'illustre la figure 8.2. En effet, la représentation de cette adresse selon l'ontologie PeGazUs indique que la cible comporte trois relations spatiales avec d'autres entités géographiques. Le premier segment d'adresse, « *au centre de la capitale* », annonce une relation d'inclusion dans Paris, en ajoutant une précision quand à la localisation centrale de la cible dans la ville. La figure 8.2a illustre cette relation en représentant le territoire de Paris tel qu'il existait au début du XIXe siècle, correspondant à la période de validité de l'adresse. Les deux autres segments, « *entre le Palais-Royal et les Tuileries* » et « *à proximité des principaux théâtres* », correspondent à des relations spatiales plus localisées. La figure 8.2b donne une précision sur l'estimation de la position de la cible, en complément de l'indication donnée par le premier segment d'adresse.

8.5 Vers une exploitation plus fine des événements

Actuellement, les événements associés aux entités géographiques sont principalement utilisés pour situer des changements dans le temps, sans typage explicite (par exemple, fusion, scission, renommage). Une exploitation plus fine consisterait à inférer ces types d'événements à partir des données disponibles, afin d'enrichir la compréhension des transformations.

Certaines configurations de changements permettent en effet de déduire la nature de l'événement :

- une **fusion** peut être identifiée lorsqu'un même événement regroupe plusieurs changements correspondant à la disparition de plusieurs entités et à l'apparition d'une nouvelle entité unique ;
- une **séparation** peut être inférée lorsqu'un changement de géométrie sur une entité existante est appliqué, et qu'un autre changement associé à l'événement, décrit l'apparition d'une entité de même type ;
- de manière analogue, des règles similaires peuvent être établies pour l'**annexion**, la **scission** ou la **réallocation**, selon la combinaison des changements observés.

Cependant, du fait de la pluralité des cas et de la complexité de certains événements, il n'existe pas de cadre strict pour leur typage. Par exemple, considérons l'événement résultant de la *loi sur l'extension des limites de Paris* du 16 juin 1859, qui implique que le territoire de la ville s'étend à partir de 1860 jusqu'à l'enceinte de Thiers. Cet événement engendre un grand nombre de changements correspondant aux entités intégrées à Paris — voies, quartiers, numéros d'immeuble, etc. On peut ainsi observer des changements illustrant, par exemple, la disparition de la relation ⟨Rue_Gérard dans Gentilly⟩ et l'apparition de ⟨Rue_Gérard dans Paris⟩. Typer un tel événement complexe n'est pas aisé. Cela souligne qu'avant de proposer un ensemble de règles pour inférer automatiquement le type des événements, il est nécessaire de définir une typologie claire et cohérente pour ces derniers. Cette typologie constituerait la base sur laquelle des règles formalisées pourraient ensuite être appliquées, permettant un raisonnement systématique malgré la diversité et la complexité des cas.

De plus, certains événements n'impliquent pas directement de modifications dans nos données. Il peut s'agir, par exemple, de l'inauguration d'une voie ou d'événements recensés par la nomenclature des voies de Paris, tels que le classement, le déclassement, le nivellement ou l'assainissement. Il existe également des événements projetés, liés notamment à la planification. Concernant la rue de Rivoli, l'arrêté du 17 vendémiaire an X (9 octobre 1801) qui prévoit que « sera percée [cette] rue dans toute la longueur du passage du

Manège jusqu'à celle de Saint-Florentin ». Tous ces événements pourraient être intégrés dans le référentiel, mais cela soulève des questions méthodologiques : comment représenter un événement qui n'est pas constitué d'une combinaison de changements appliqués aux données existantes ? Comment indiquer le caractère hypothétique d'un événement et comment l'utiliser lors de la construction du référentiel ?

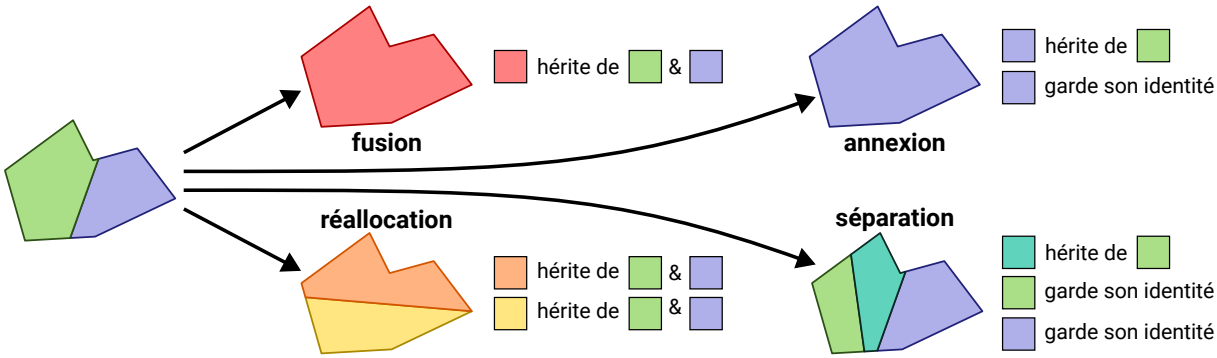


FIGURE 8.3 – Proposition d'une typologie des liens de filiation en fonction des événements.

Enfin, l'une des limites de l'ontologie proposée est l'absence de liens de filiation. Or, ces liens pourraient s'avérer particulièrement utiles pour traiter certains cas où un changement de nom génère la création d'une nouvelle entité. En effet, un changement de nom officiel ne correspond pas nécessairement à un changement dans l'usage, ce qui fait que des sources continuent d'utiliser l'ancienne appellation. Par exemple, considérons l'adresse « 2 place du Trône » qui se trouverait dans une source datant de 1882. À cet instant, la place n'existe plus dans notre référentiel. Cependant, si l'on indique que la place de la Nation est l'héritière de la place du Trône, cela permettrait d'interpréter que l'adresse fait finalement référence au « 2 place de la Nation ». Une telle représentation, comme proposée par la figure 8.3, permettrait d'aller au-delà de l'identité des entités portant sur le nom.

8.6 Extraction d'information dans les sources textuelles

Dans cette thèse, nous avons essentiellement travaillé sur des données numérisées à partir de plans anciens et des jeux de données structurées publiés par diverses organisations. Pour enrichir le graphe de connaissances, il serait intéressant d'exploiter davantage les sources textuelles. Si certaines informations structurées (par exemple celles issues d'annuaires) peuvent être exploitées après leur extraction, une grande partie des documents historiques demeure non structurée, rendant leur traitement plus complexe. Au-delà du fait qu'elles ne sont généralement pas structurées, ces informations sont souvent dispersées au sein d'une quantité importante de texte, contenant de nombreux éléments qui ne sont pas directement pertinents pour nos besoins. Il est donc nécessaire de disposer de méthodes capables d'identifier spécifiquement les informations utiles au suivi des évolutions des voies et des adresses, puis de les structurer afin de les rendre compatibles avec l'ontologie.

Dans cette perspective, un stage de recherche a été consacré à l'extraction automatisée d'informations textuelles relatives à l'évolution des voies parisiennes. L'objectif était d'évaluer des approches d'apprentissage automatique (en particulier les grands modèles

de langage) dans leur capacité à identifier des événements d'évolution et les structurer de manière exploitable dans notre graphe de connaissances. Trois approches ont été explorées et comparées : une méthode basée sur BERT pour l'extraction conjointe d'entités et de relations, une méthode de génération de triplets RDF guidée par une ontologie et utilisant un grand modèle de langage (LLM), et une approche hybride combinant modèles de type *transformer* et LLM.

Les résultats préliminaires montrent que les LLM sont compatibles avec l'extraction d'informations relative à des événements d'évolution mais soulignent qu'il est nécessaire de mettre en place des mécanismes de contrôle et de post-traitement pour garantir la fiabilité des extractions. Cette piste ouvre la voie à l'intégration de corpus textuels variés mais aussi à l'extraction et la structuration d'adresses historiques. Le détail de ces travaux ainsi que l'analyse complète des résultats sont présentés en annexe A.

8.7 Limites de l'évaluation proposée

L'évaluation proposée montre que la méthode de peuplement permet de restituer fidèlement l'évolution des entités géographiques et qu'elle est robuste à l'ajout de données fragmentaires cohérentes, qu'il s'agisse d'états spatio-temporels ou d'événements.

Cependant, cette évaluation présente plusieurs limites.

8.7.1 Limitation du périmètre expérimental

Pour commencer, l'expérimentation a été menée sur un périmètre restreint, correspondant au quartier du Faubourg Saint-Antoine, en utilisant des données issues de six sources cartographiques pour un total d'environ 8 000 adresses. Bien que ce volume soit satisfaisant, il resterait pertinent d'élargir à la fois le nombre de sources, en intégrant une plus grande diversité de types de documents (et non uniquement des cartes), ainsi que la diversité des emprises spatiales considérées. En particulier, on pourrait appliquer la méthode à d'autres territoires, notamment à l'étranger afin d'évaluer sa généricité dans des contextes où les systèmes d'adressage et la nature des sources diffèrent.

8.7.2 Absence d'évaluation avec des recouvrements temporels

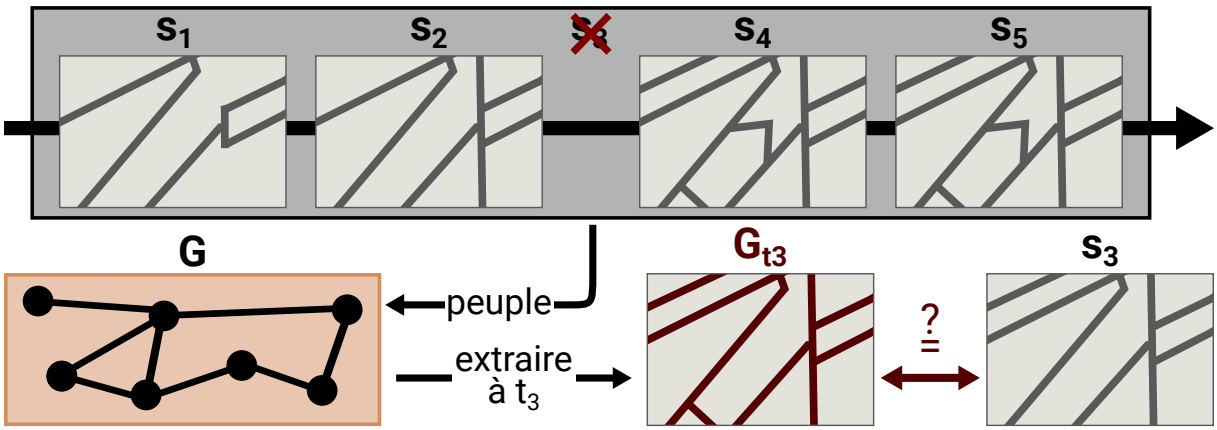


FIGURE 8.4 – Proposition d'une chaîne de traitement pour un test d'ablation.

Ensuite, l'évaluation repose sur la comparaison entre un graphe G reconstruit et un jeu de référence R. Toutefois, les données utilisées correspondent uniquement à des *snapshots* temporellement disjoints, c'est-à-dire dont les temps valides ne se chevauchent pas. Dans ce cadre, la méthode reconstitue correctement les évolutions, de manière comparable aux approches classiques. En revanche, le cas de sources décrivant des états simultanés ou partiellement chevauchants n'a pas été évalué. Une piste d'amélioration consiste à mettre en place un test d'ablation. Le principe, illustré par la figure 8.4, est de construire un graphe G à partir d'un ensemble de données en excluant celles issues d'une source donnée (par exemple s_3, associée à un temps valide t_3). Une fois G construit, un sous-graphe G_{t_3}, représentant l'état du territoire à t_3, peut être extrait. Il devient alors possible de comparer ce sous-graphe avec les données initiales de la source exclue (représentées dans un graphe nommé G_{s_3}), afin d'évaluer la capacité de la méthode à reconstruire correctement un état en présence de recouvrements temporels (s'il existe des données dont le temps valide chevauche celui de la source exclue).

8.7.3 Absence de test de robustesse face à l'ajout de données contradictoires

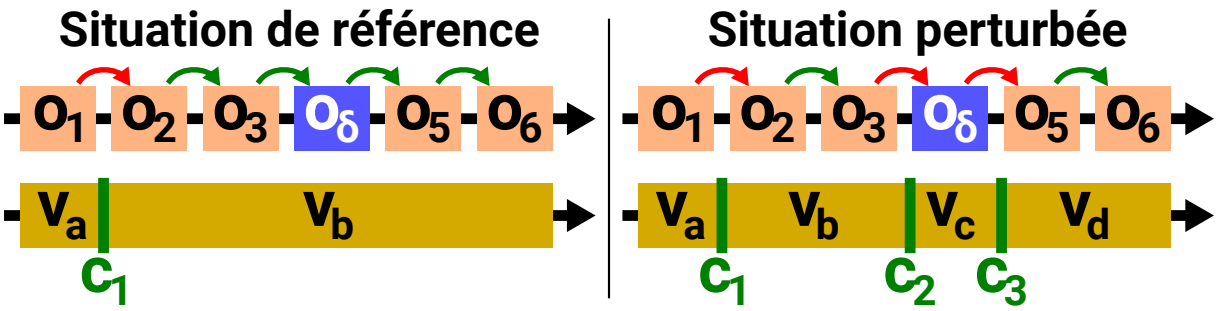


FIGURE 8.5 – Comparaison de deux lignes de vie de l'adresse *2 rue du Pas de la Mule* obtenues avec la méthode de peuplement proposée. La situation de référence est obtenue à partir des observations initiales, tandis que la situation perturbée est générée à partir des mêmes observations, auxquelles est ajouté une observation c_δ décrivant une version fictive contradictoire avec l'évolution attendue. $\rightarrow$: *a une géométrie différente de*, $\rightarrow$: *a une géométrie similaire à*.

Troisièmement, la seconde partie de l'évaluation considère uniquement l'ajout de données fragmentaires cohérentes. Or, il serait également pertinent d'analyser le comportement de la méthode en présence de données contradictoires. Une première approche consiste à vérifier que l'introduction d'une observation contradictoire perturbe effectivement les lignes de vie reconstruites. Par exemple, pour l'adresse *2 rue du Pas de la Mule* (voir la figure 8.5), l'ajout d'une observation o_δ contradictoire avec les évolutions décrites par les sources conduit à une augmentation du nombre de versions d'attributs géométriques (passant de deux à quatre), traduisant une incohérence dans la reconstruction.

8.7.4 Nombre limité d'expérimentations pour l'évaluation de robustesse

Enfin, l'évaluation de robustesse repose sur une unique comparaison entre G et G'. Bien que cette comparaison ait été répétée plusieurs fois avec des données fragmentaires

générées aléatoirement, produisant des scores similaires, il serait pertinent de systématiser cette démarche en réalisant un plus grand nombre d'expérimentations et moyennant les scores obtenus. Cela permettrait de renforcer la validité des conclusions faites en apportant une évaluation plus robuste.

Conclusion

Ce chapitre clôture la thèse en proposant un ensemble de critiques autour des travaux menés et en introduisant plusieurs perspectives d'amélioration. La discussion s'est articulée autour de plusieurs axes notamment autour du besoin de représenter l'incertitude dans l'ontologie. Ce besoin découle, d'une part, des caractéristiques des données issues des sources historiques, notamment en ce qui concerne la représentation des géométries, et d'autre part des conflits possibles entre ces données. Parallèlement, l'enrichissement collaboratif peut constituer une stratégie pour corriger d'éventuelles imperfections. Ensuite, nous avons présenté les limites de la mise en œuvre du peuplement de l'ontologie ainsi que les perspectives d'amélioration de l'évaluation. Les critiques autour du critère d'identité, l'utilisation des relations spatiales pour améliorer la localisation, l'exploitation plus fine des événements ainsi que l'extraction d'information dans les sources textuelles offrent des pistes d'amélioration tant pour la constitution des données servant à construire le référentiel que pour l'exploitation de ce dernier. Enfin, nous avons présenté les limites de l'évaluation proposée tout en y proposant des stratégies d'améliorations.

Ainsi, ce chapitre joue un double rôle : il porte un regard critique sur les travaux réalisés et oriente les développements futurs. Il propose des axes de recherche et des perspectives d'amélioration pour la reconstitution géohistorique des adresses et des entités géographiques. Il souligne aussi la nécessité de compléter les données, de diversifier les sources et de renforcer les protocoles d'évaluation afin de garantir la cohérence et la fiabilité du graphe de connaissances final.

Conclusion

Notre thèse s'intéresse à la construction de référentiels géohistoriques d'adresses à partir de sources de données hétérogènes et fragmentaires. Considérant l'adresse comme un énoncé structuré désignant un lieu à partir d'une suite ordonnée de relations entre entités géographiques, reconstruire son évolution revient à reconstituer celle des entités qui la composent et de leurs relations. Pour atteindre cet objectif, nous proposons l'ontologie PeGazUs, qui permet de représenter les adresses et leurs changements dans le temps au sein d'un graphe de connaissances unifié, alimenté par des informations issues de multiples sources historiques. De plus, cette thèse propose une méthode de peuplement de ce graphe. Elle repose sur l'intégration de multiples "graphes factoïdes" : chacun représentant les connaissances souvent fragmentaires tirées d'une source de données, historique ou non, renseignant sur les changements ou sur les valeurs des propriétés des adresses représentées.

Les adresses sont importantes non seulement par leur volume dans les sources, mais également par leurs applications, puisqu'elles permettent de servir de pivot pour relier les informations provenant de sources diverses. Elles évoluent en fonction des modifications des entités géographiques qui leur sont associées, ce qui souligne l'intérêt de créer un référentiel géohistorique d'adresses. La construction de ce référentiel à partir de plusieurs sources est nécessaire, car aucune source unique ne décrit l'historique complet des adresses ; en revanche, l'accumulation et la combinaison des informations issues des différentes sources permettent de reconstituer partiellement, voire globalement, l'évolution de ces entités.

Les sources utilisées sont très diverses. Les sources cartographiques décrivent généralement des informations spatiales sur les entités géographiques auxquelles sont associés quelques propriétés comme leur nom ou leur nature. L'information temporelle associée est le plus souvent un intervalle correspondant au temps de validité du plan. Chaque plan peut ainsi être considéré comme un ensemble d'états spatio-temporels d'entités géographiques. Les documents textuels présentent un contraste, fournissant rarement des informations spatiales explicites sur l'emprise des entités, mais contenant souvent des descriptions de relations. Parmi ces documents, on distingue les sources structurées, telles que les annuaires, qui accumulent des adresses et témoignent ainsi de l'existence d'une adresse ou d'une entité géographique à un instant donné, éventuellement en décrivant des relations entre entités, et les dictionnaires topographiques, qui intègrent des événements documentant l'évolution des entités géographiques. D'autres documents sont peu structurés, comme certains documents administratifs, rendant l'extraction d'information plus complexe. Néanmoins, les avancées récentes dans le traitement automatique du langage, notamment via les grands modèles de langage (LLM), permettent désormais d'extraire ces informations de manière plus efficace. Enfin, les sources numériques sont généralement structurés, mais la diversité des standards nécessite une transformation des données pour homogénéiser les informations. L'hétérogénéité des sources se manifeste également dans la manière dont sont décrits les phénomènes, alternant entre états spatio-temporels et événements.

L'objectif est de créer un référentiel spatio-temporel à partir de données à la fois fragmentaires et hétérogènes, c'est-à-dire des ensembles d'informations incomplets et discontinus, ne couvrant qu'une partie de la réalité observée, et présentant des modalités de description différentes selon les sources. Un état de l'art a permis d'identifier les approches existantes pour représenter les dynamiques des entités géographiques ainsi que les supports permettant cette représentation. Ces analyses nous ont conduit à adopter une modélisation des entités géographiques qui permet de représenter de manière indépendante les dynamiques de chacune de leurs dimensions, à la fois spatiales, temporelles et thématiques, en utilisant des graphes de connaissances, particulièrement adaptés pour représenter des données fragmentaires et incomplètes.

Contributions principales

Pour répondre aux objectifs de cette thèse, trois contributions majeures ont été développées, portant respectivement sur la conception d'une ontologie, la proposition d'une méthode de peuplement d'un graphe de connaissances géohistoriques, et la définition d'une stratégie d'évaluation adaptée à l'absence de vérité terrain et aux difficultés majeures que poserait sa création.

La première contribution est l'ontologie *PeGazUs*, conçue pour décrire les entités géographiques et les adresses dans toute leur diversité, tout en intégrant la dimension spatio-temporelle de leur évolution. Elle permet une représentation flexible, adaptée au caractère incomplet, fragmentaire et hétérogène des sources historiques : chaque entité peut être décrite uniquement selon les informations effectivement disponibles, sans que l'absence de certaines caractéristiques, comme la géométrie, ne soit compromettante. L'ontologie rend possible l'expression de relations variées, notamment spatiales, et permet de relier des entités géographiques selon ce qu'indiquent explicitement ou pas les sources. L'introduction de la temporalité permet d'étudier les dynamiques d'évolution, tout en tenant compte de l'imperfection de cette dimension lorsque les données sont imprécises ou incertaines. La modélisation spatio-temporelle ne se limite pas au niveau des entités : elle s'étend aux attributs eux-mêmes, traités comme des éléments indépendants dont les versions successives reflètent les changements observés. Cette granularité permet de mieux représenter la nature fragmentaire des données : par exemple, une source dépourvue d'information géométrique sur une entité ne perturbe pas la description de ses autres aspects ni la reconstruction de ses dynamiques spatiales.

La seconde contribution consiste en une méthode de peuplement d'un référentiel géohistorique d'adresses à partir de données hétérogènes et fragmentaires. Elle s'appuie directement sur l'ontologie *PeGazUs* et vise à concilier des informations issues de sources variées, structurées ou non, dans un cadre unifié. Ainsi, pour chaque source, un graphe dit « de factoïdes » est constitué par l'ensemble des triplets obtenus lors de la conversion des données de la source selon l'ontologie. Ces graphes de factoïdes, considérés comme des ensembles d'observations indépendantes, servent ensuite à la création d'un graphe, appelé "« graphe des faits » représentant la conciliation des connaissances. Les observations similaires y sont reliées afin d'agréger l'information redondante, de détecter les divergences et d'inférer de nouvelles relations et des nouveaux faits. Ce processus permet de transformer un ensemble de données initialement fragmentaires, issues de sources multiples, en une représentation cohérente et évolutive des entités géographiques et des adresses.

Enfin, la troisième contribution concerne l'évaluation de la méthode de peuplement et de l'ontologie. L'évaluation de l'ontologie est celle proposée par la méthode SAMOD, qui a guidé sa conception. En revanche, l'évaluation de la méthode de construction s'avère

plus complexe, en raison de l'absence de vérité terrain exhaustive permettant une comparaison directe avec le graphe construit automatiquement. Pour pallier cette difficulté, une stratégie d'évaluation spécifique a été élaborée, incluant notamment un test de robustesse consistant à introduire des données fragmentaires synthétiques dans les sources initiales. Cette approche permet d'observer la stabilité du processus de reconstitution face à des perturbations contrôlées, et d'étudier la capacité de la méthode à produire des résultats fiables à partir d'informations imparfaites.

Perspectives générales

Les travaux présentés dans cette thèse ouvrent la voie à des perspectives de recherche et d'application dépassant le cadre du cas d'étude parisien. La modélisation proposée, ainsi que la méthode de construction et de peuplement du graphe de connaissances géohistoriques, peuvent en effet être généralisées à d'autres contextes territoriaux et à des ensembles de données de plus grande ampleur. On pourrait appliquer, à la condition d'optimiser la mise en œuvre de la méthode, cette approche à l'échelle nationale. Il serait possible de l'appliquer par exemple à l'ensemble des adresses de l'IGN ou d'OpenStreetMap afin de constituer un référentiel géohistorique des adresses couvrant l'ensemble du territoire français, pour les périodes couvertes par ces deux référentiels. Un tel référentiel offrirait une vision unifiée des adresses et des entités géographiques à partir d'une multitude de sources. Ainsi, il pourrait être mobilisé pour enrichir les moteurs de recherche des institutions patrimoniales et culturelles en améliorant les capacités de géocodage et de désambiguïsation des lieux mentionnés dans les corpus documentaires. Une autre perspective intéressante consiste à construire un référentiel d'adresses actuelles pour une ville, une région ou un pays qui n'en disposerait pas, à partir de sources fragmentaires disponibles sur les dernières années. L'ontologie développée dans ce travail est générique et supporte des adresses de tous types, tandis que la méthode proposée permet de générer le graphe de connaissances à partir de l'ensemble des données collectées. À partir de ce graphe, il est alors possible, via des requêtes SPARQL, de produire un *snapshot* correspondant à la dernière année disponible. Ce type de référentiel constituerait ainsi une base géohistorique ouverte, interopérable et réutilisable par la communauté scientifique, servant de support de référence pour les chercheurs en humanités numériques.

La masse considérable des sources d'adresses disponibles — notamment issues des documents textuels historiques ou administratifs — représente toutefois un défi majeur. Ces documents, riches en informations, présentent une forte hétérogénéité de structure, de vocabulaire et de précision. L'un des enjeux futurs réside ainsi dans le développement d'approches automatiques pour extraire, harmoniser et relier ces informations à grande échelle. Les récents progrès des LLM offrent des perspectives intéressantes pour extraire et structurer les adresses ainsi que les entités géographiques qui y sont référencées en facilitant la reconnaissance d'adresses ainsi que la détection des entités et des relations régies entre elles.

Enfin, la réutilisation et l'alignement des données produites constituent une perspective majeure de valorisation. Les connaissances issues du graphe pourraient être alignées et partagées avec d'autres projets collaboratifs tels que *Wikidata*, *OpenStreetMap* (ou sa version historique *OpenHistoricalMap*), favorisant ainsi une meilleure diffusion et un enrichissement des données géohistoriques. Une telle interconnexion pourrait s'étendre aux bases de connaissances des GLAM (Galeries, Bibliothèques, Archives et Musées), afin d'intensifier la mise en relation entre les collections patrimoniales via leurs dimensions spatiales et temporelles. Une telle intégration renforcerait aussi la cohérence et la

traçabilité des informations sur les adresses et les entités géographiques dans le Web des données, tout en ouvrant la voie à de nouvelles exploitations scientifiques et patrimoniales, notamment pour la recherche en humanités numériques et la valorisation du patrimoine géographique.

Au-delà du cas d'étude parisien, limité à la période allant de la fin du XVIIIe siècle à aujourd'hui, notre approche ouvre également vers la généralisation de la construction de référentiels d'adresses pour d'autres villes, régions ou pays disposant de sources diverses. L'ontologie proposée, ainsi que la méthode de peuplement associée, ont été conçues de manière générique afin de s'adapter à différents territoires et types de sources.

Annexe A

Exploitation approfondie des documents textuels

Aujourd'hui, les technologies d'OCR et d'HTR sont suffisamment robustes pour extraire efficacement des informations à partir de sources textuelles historiques. Lorsque ces informations sont déjà structurées, comme dans le cas des annuaires ou des registres, leur intégration dans notre graphe de connaissances est relativement aisée. En revanche, la plupart des sources sont non structurées, et les informations pertinentes sont souvent noyées dans de larges volumes de texte, ce qui rend leur exploitation plus complexe. Cependant, avec les récents progrès des grands modèles de langage (LLM), il devient envisageable d'extraire et de structurer ces informations pour les intégrer dans notre graphe.

A.1 Méthodes pour l'extraction d'informations textuelles

Face à ces défis, un stage a été réalisé portant sur l'extraction et la structuration des informations relatives à l'évolution des rues à partir de corpus textuels, en utilisant des grands modèles de langage afin de faciliter l'alignement des données avec l'ontologie *PeGazUs*. Cette recherche s'est concentrée sur l'exploitation des données issues de la nomenclature des voies de Paris, une source numérique tabulaire contenant certaines informations en langage naturel, notamment sur l'historique des rues. Afin d'évaluer la pertinence d'utiliser des LLM pour extraire de telles informations, trois méthodes ont été utilisées et comparées :

— la méthode PURE (Zhong et Chen, 2021), basée sur BERT, est une approche permettant l'extraction d'entités et de relations et constitue notre méthode de référence ;
— l'approche Text2KGBench (Mihindukulasooriya *et al.*, 2023) qui permet de générer des triplets RDF à partir de LLM en fournissant en entrée l'ontologie, un prompt et des données à partir desquels les triplets doivent être générés ;
— une approche hybride utilisant BERT et des LLM (Ding *et al.*, 2024).

A.1.1 Méthode PURE (BERT)

L'approche PURE (*Positional Unsupervised Relation Extraction*), basée sur le modèle de langage BERT, permet d'extraire simultanément des entités et les relations qui les lient, contrairement aux méthodes classiques reposant sur BERT qui procèdent en deux

étapes distinctes. Dans PURE, les entités sont extraits sans annotation manuelle explicite, et les relations binaires sont prédites en fonction des couples d'entités identifiés dans le texte. Cette extraction simultanée contribue à réduire les erreurs. Cependant, cette méthode reste limitée à des jeux de données standards et est adaptée à un ensemble de relations défini. Rawsthorne *et al.* (2024) proposent une extension de PURE adaptée aux informations géographiques. La figure A.1 illustre un exemple d'une telle extraction. Il est important de noter que les extractions ne se limitent pas au simple nom des entités géographiques, mais peuvent inclure des informations plus détaillées. Pour ce faire, l'approche s'appuie sur les étiquettes introduites par Moncla (2015), qui distingue les *caractéristiques géographiques* (noms communs désignant un type d'entité spatiale, comme la « route »), les *noms* (noms propres, ex. « Fontainebleau ») et les *noms géographiques* (noms complets d'entités spatiales comme « route de Fontainebleau »). Néanmoins, cette approche reste restreinte. Par exemple, elle a une capacité limitée à généraliser dans des contextes complexes ou ouverts. D'une part, il n'est pas possible de détecter des relations implicites ou raisonnées. D'autre part, les relations pouvant être extraites sont binaires, il est donc impossible de générer des relations correctes pour dans des phrases du type « La rue Gérard est située entre les rues Bobillot et du Moulin des Prés ».

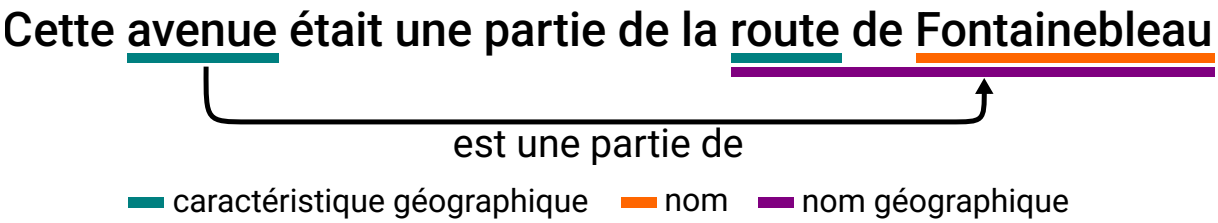


FIGURE A.1 – Extraction d'entités et de relations au sein de la phrase « Cette avenue était une partie de la route de Fontainebleau », issue du dictionnaire d'Hillairet traitant de l'avenue d'Italie.

A.1.2 Méthode Text2KGBench

En second lieu, Mihindukulasooriya *et al.* (2023) proposent Text2KGBench, un benchmark orienté vers la génération de triplets RDF à partir d'un corpus de texte guidée pour une ontologie en utilisant des LLM. La chaîne de traitement permettant la génération est illustrée par la figure A.2. Le processus commence par l'entrée de l'ontologie et des corpus textuels, qui servent à générer automatiquement un prompt. Ce prompt est ensuite utilisé pour l'extraction de l'information à l'aide de modèles de fondation, c'est-à-dire de grands modèles de langage pré-entraînés capables de traiter une variété de tâches linguistiques sans entraînement spécifique, tels que GPT-4 ou Mistral 7B. Enfin, une étape de post-traitement et de validation est réalisée afin de produire le graphe de connaissances final. L'objectif de leur travail est de vérifier si de tels triplets sont conformes à l'ontologie fournie.

A.1.3 Méthode hybride BERT + LLM

Comme l'affirment Wei *et al.* (2024), les LLM sont capables de générer des triplets simples sans apprentissage grâce à des prompts bien définis. Néanmoins, utiliser uniquement ces grands modèles de langage est limité car le corpus d'entraînement des LLM est généralement constitué de phrases simples impliquant d'avoir un biais marqué vers

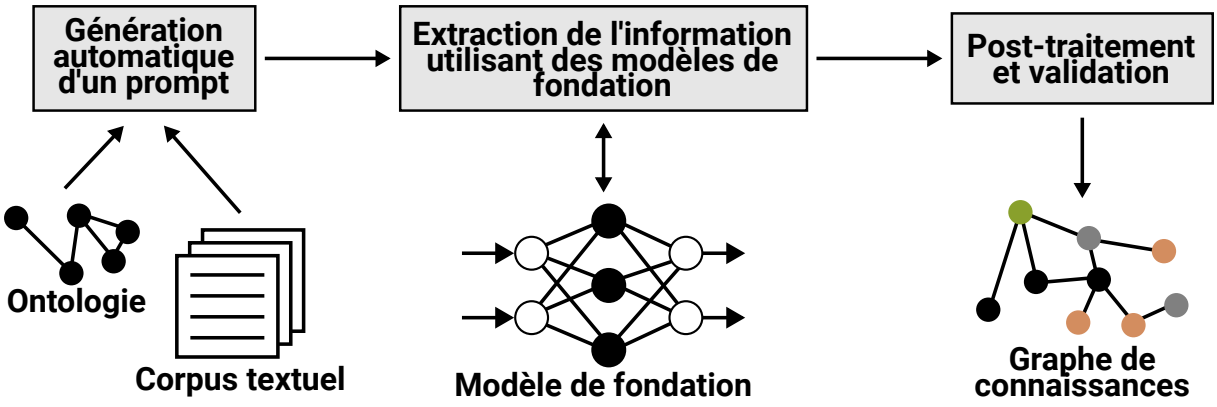


FIGURE A.2 – Schéma expliquant les étapes de la génération des triplets RDF avec des LLM à partir d'une ontologie et d'un corpus de textes.

des structures contenant peu de triplets. Par conséquent, lorsqu'il s'agit d'extraire plusieurs entités et relations dans des phrases longues ou complexes, les LLM n'arrivent pas à extraire la totalité des informations nécessaires. Au contraire, les approches avec les modèles traditionnels comme BERT tendent à extraire davantage d'informations que nécessaires (Chu *et al.*, 2020). C'est pour relever ce défi majeur que Ding *et al.* (2024) proposent une méthode hybride combinant un petit modèle traditionnel et un LLM. Leur approche repose sur un modèle d'évaluation-filtrage fondé sur une architecture Transformer, destiné à générer des paires d'entités candidates et à les fournir comme partie intégrante du prompt pour guider le LLM à attribuer la relation correcte.

A.2 Données et vérité terrain

Pour évaluer nos trois méthodes, nous avons comparé leurs performances afin de déterminer si l'utilisation d'un modèle coûteux comme un LLM vaut l'investissement par rapport aux modèles traditionnels. Pour ce faire, nous avons construit une vérité terrain à partir de la nomenclature des voies de Paris, qui décrit les 6 500 voies de la ville. Nous avons sélectionné un échantillon représentatif de 300 rues en respectant deux critères : la répartition spatiale et la répartition thématique. La répartition spatiale assure que la distribution des rues selon les arrondissements dans l'échantillon reflète celle de la réalité. La répartition thématique garantit la représentativité des descriptions des évolutions des voies, en supposant que plus une voie est longue, plus elle a subi de modifications. Ainsi, l'échantillon comprend à la fois des voies courtes, comme des impasses ayant subi peu de changements, et des rues longues ayant connu de nombreuses évolutions.

Pour chaque phrase décrivant une évolution d'une rue, nous avons associé le triplet attendu. Les phrases proviennent des colonnes *Historique*, *Classement*, *Numérotage* et *Ouverture* de notre base de données, combinées dans un tableau simplifié (voir tableau A.1).

En raison de la complexité de l'ontologie, nous avons constaté, après quelques tests, que les modèles de langage de grande taille (LLM) se montraient peu fiables pour générer directement des triplets contenant des nœuds blancs. Afin de contourner cette difficulté, il a été choisi de produire des triplets plus simples, qui pourront ensuite être convertis en structures conformes à l'ontologie *PeGazUs*.

Par exemple, pour la phrase « La partie comprise entre les rues Bobillot et du Moulin des Prés a été dénommée rue du Père Guérin par arrêté municipal du 30 août 1978. »,

ID	Dénomination	Historique
11028	rue Gérard	La partie comprise entre les rues Bobillot et du Moulin des Prés a été dénommée rue du Père Guérin par arrêté municipal du 30 août 1978.
12744	rue du Père Guérin	Précédemment partie de la rue Gérard.
9360	rue Bobillot	
11369	avenue d'Italie	Précédemment partie de la grande route ou route de Fontainebleau (partie de la RN n° 7). Elle figure à l'état de chemin sur le plan de Jouvin de Rochefort (1672).

Tableau A.1 – Exemple simplifié du jeu de données avec les phrases historiques associées aux rues.

dont la représentation conforme à l'ontologie complète est présentée dans la figure 4.8, nous générons les triplets simplifiés suivants (en complément de ceux décrivant les entités géographiques) :

```
<RuePereGuerin   appearsOn   "1978-08-30">
<RueGerard   hasGeometryChangeOn   "1978-08-30">
```

Une autre raison de cette simplification est la nécessité de comparer les méthodes d'extraction, certaines — comme *PURE* — ne produisant que des relations binaires. Une fois la vérité terrain constituée (environ 850 événements), le jeu de données a été réparti en trois sous-ensembles : 70% pour l'entraînement, 20% pour le test et 10% pour la validation. Ces ensembles ont ensuite été utilisés pour évaluer les trois méthodes retenues.

A.3 Résultats

Méthode	Précision	Rappel	F1-score
PURE (CamemBERT)	0.55	**0.82**	0.66
Hybride (BERT + LLM)	0.78	0.73	0.75
LLM	**0.90**	0.89	**0.89**

Tableau A.2 – Comparaison des résultats obtenus par les trois méthodes selon les métriques utilisées par Mihindukulasooriya *et al.* (2023). Les meilleurs scores pour chaque métrique sont indiqués en gras.

Les résultats permettent d'obtenir des premières tendances. La méthode *PURE*, basée sur le modèle *CamemBERT* — une variante francophone de *BERT* — présente une précision relativement faible (0,55), ce qui traduit un nombre important de faux positifs. Toutefois, elle obtient le meilleur rappel (0,82) indiquant une bonne couverture des triplets attendus. Cette approche s'avère donc utile pour maximiser la couverture, mais au prix d'une perte de précision. La méthode hybride offre un compromis équilibré entre précision et rappel. Avec une précision de 0,78 et un rappel de 0,73, elle atteint un F1-score de 0,75, supérieur à celui de l'approche précédent. Elle combine la rigueur des modèles de type BERT pour la détection de paires candidates et la capacité interprétative d'un LLM pour la validation et la complétion, assurant ainsi un bon équilibre entre qualité et couverture. Enfin, la méthode fondée uniquement sur un LLM surpasse les deux autres approches sur

la plupart des indicateurs. Elle atteint la meilleure précision (0,90) et le meilleur F1-score (0,89), tout en maintenant un rappel élevé (0,89). Ces performances montrent la capacité des LLM à extraire de manière fiable des triplets complets, avec un alignement satisfaisant sur l'ontologie. Toutefois, ce type de modèle reste plus sujet aux hallucinations d'objets, ce qui peut limiter son emploi dans des contextes exigeant une fiabilité totale des données. Le fait que les LLM surpassent la méthode hybride peut s'expliquer par plusieurs facteurs. D'une part, leur pré-entraînement massif sur des corpus multilingues et multidomaines leur confère une capacité de généralisation supérieure. D'autre part, la méthode hybride souffre d'un effet de goulot d'étranglement : la génération de paires candidates par BERT limite le rappel global, puisqu'un triplet manqué à ce stade ne peut plus être récupéré par le LLM. Enfin, les LLM disposent d'une aptitude implicite à la complétion et à la correction d'informations manquantes, que l'étape de filtrage de la méthode hybride tend à restreindre. En résumé, l'approche *PURE* maximise la couverture, la méthode hybride assure un bon compromis entre précision et rappel, tandis que le LLM offre la meilleure performance globale. Le choix de la méthode dépendra ainsi des priorités de l'application : maximiser la couverture, garantir un équilibre entre qualité et quantité, ou viser la performance optimale en termes de qualité d'extraction.

A.4 Discussion

Il convient toutefois de mettre ces résultats en perspective. Le jeu de données utilisé contient des phrases structurées de manière spécifique, et ne reflète pas nécessairement la diversité ou la complexité syntaxique que l'on pourrait trouver dans d'autres documents historiques ou corpus hétérogènes. De plus, les résultats obtenus dépendent fortement de la répartition des événements entre les ensembles d'entraînement, de test et de validation. Pour obtenir une estimation plus robuste de la performance des différentes méthodes, il serait nécessaire de réaliser plusieurs répartitions aléatoires du jeu de données, d'appliquer les méthodes sur chaque répartition, puis de calculer la moyenne des métriques obtenues. Cette approche permettrait de limiter l'effet d'une division particulière et de fournir une évaluation plus stable et généralisable.

Conclusion

Les LLM constituent des outils intéressants pour structurer des informations issues de sources textuelles, mais ces solutions doivent être utilisées avec précaution. Dans notre cas, ils peuvent être employés pour générer des triplets simples, qui pourront ensuite être transformés en triplets plus complexes conformes à l'ontologie *PeGazUs* à l'aide d'un programme Python, par exemple. Il convient de rappeler qu'avant ce travail, nous avions déjà généré manuellement des triplets à partir de sources textuelles afin de disposer de ce type d'informations pour alimenter notre graphe de connaissances. Enfin, le travail présenté ici a été appliqué pour extraire des événements d'évolution d'entités géographiques mais il serait aussi pertinent de faire de même avec les adresses qui sont très diverses dans les sources. De plus, les LLM pourraient être appropriés pour les structurer selon l'ontologie.

Annexe B

Questions de compétence

B.1 Questions du modelet des adresses

Cette partie présente des requêtes SPARQL traduisant les questions de compétence liées au *modelet* des adresses, accompagnées de leur résultat.

B.1.1 Liste d'adresses situées le long d'une voie donnée

Quelles sont les adresses situées le long de la rue du Pas de la Mule ?

```
SELECT DISTINCT ?addr ?addrLabel ?geom WHERE {
    VALUES (?lLabel ?lLang) {("rue du Pas de la Mule", "fr")}

    ?l a :Landmark ; :isLandmarkType ltype:Thoroughfare ; rdfs:label ?lName .
    FILTER(LCASE(STR(?lName)) = LCASE(?lLabel) && LANG(?lName) = ?lLang)
    ?addr a :Address ; rdfs:label ?addrLabel ; :hasStep [ :relatum ?l ] .

    OPTIONAL {
        ?addr :targets [ :hasAttribute [ :isAttributeType atype:Geometry ;
        ↪  :hasAttributeVersion [:versionValue ?geom] ] ] .
    }
}
```

Listing B.1 – Requête SPARQL pour identifier les adresses le long d'une voie donnée.

```
address,addrLabel,geom
ex:ADDR_0436fb16ccfe4f8aabda4330915fd22b,"6 Rue du Pas de la Mule, 75003 Paris
↪  3e Arrondissement",POINT (2.367113 48.856027)
ex:ADDR_15ca185bfd95479dbfa57d3defde3e8f,"4b Rue du Pas de la Mule, 75003
↪  Paris 3e Arrondissement",POINT (2.367544 48.855953)
...
ex:ADDR_c8a23ed9623d41ef98be40c7deafb02a,"15 Rue du Pas de la Mule, 75004
↪  Paris 4e Arrondissement",POINT (2.366797 48.85597)
ex:ADDR_e493ef98f9094428bac794c143702de6,"13 Rue du Pas de la Mule, 75004
↪  Paris 4e Arrondissement",POINT (2.366972 48.855943)
```

Listing B.2 – Extrait du résultat en CSV de la requête SPARQL du Listing B.1 identifiant les adresses le long de la rue du Pas de la Mule. Le résultat renvoie 13 lignes.

B.1.2 Coordonnées d'une adresse à partir de son libellé

Quelles sont les coordonnées de l'adresse « 14 Place des Vosges, 75004 Paris 4e Arrondissement » ?

```
SELECT ?addr ?addrLabel ?geom
WHERE {
    BIND("14 Place des Vosges, 75004 Paris 4e Arrondissement" AS ?addrLabel)
    BIND("fr" AS ?addrLang)

    ?addr a :Address ; rdfs:label ?addrLabel ; :targets ?target .
    ?target :hasAttribute [ :isAttributeType atype:Geometry ;
    ↪  :hasAttributeVersion [ :versionValue ?geom ] ] .

    FILTER(LCASE(STR(?addrLabel)) = LCASE(?addrLabel) && LANG(?addrLabel) =
    ↪  ?addrLang)
}
```

Listing B.3 – Requête SPARQL pour obtenir les coordonnées d'une adresse à partir de son libellé.

```
addr,addrLabel,geom
ex:ADDR_69ca1f621a3844d6bb1117d2047e2b7c;"14 Place des Vosges, 75004 Paris 4e
↪  Arrondissement","POINT (2.366516 48.855382)
```

Listing B.4 – Résultat en CSV de la requête SPARQL du Listing B.3 identifiant les coordonnées associées à l'adresse dont le libellé est « 14 Place des Vosges, 75004 Paris 4e Arrondissement ».

B.1.3 Liste d'adresses situées dans une zone géographique

Quelles sont les adresses situées dans la zone définie par le polygone WKT exprimé dans le système de coordonnées WGS84 : `POLYGON ((2.367404 48.856463, 2.364829 48.854656, 2.368562 48.854247, 2.367404 48.856463))` ?

```
SELECT DISTINCT ?addr ?addrLabel ?geom
WHERE {
    BIND("<http://www.opengis.net/def/crs/OGC/1.3/CRS84> POLYGON ((2.367404
    ↪  48.856463, 2.364829 48.854656, 2.368562 48.854247, 2.367404
    ↪  48.856463))"^^geo:wktLiteral AS ?area)

    ?addr a :Address ; rdfs:label ?addrLabel ; :targets ?target .
    ?target :hasAttribute [ :isAttributeType atype:Geometry ;
    ↪  :hasAttributeVersion [ :versionValue ?geom ] ] .

    FILTER (geof:sfWithin(?geom, ?area))
}
```

Listing B.5 – Requête SPARQL pour lister les adresses situées dans une zone géographique donnée.

```
addr,addrLabel,geom
ex:ADDR_0436fb16ccfe4f8aabda4330915fd22b,"6 Rue du Pas de la Mule, 75003 Paris
↪  3e Arrondissement",POINT (2.367113 48.856027)
ex:ADDR_04e11e3b7aba401f9e561930f50cbab4,"10 Rue Jean Beausire, 75004 Paris 4e
↪  Arrondissement",POINT (2.368213 48.854401)
ex:ADDR_0e571b3cd0554c429120e551bc0f3993,"18 Place des Vosges, 75004 Paris 4e
↪  Arrondissement",POINT (2.366622 48.855663)
```

Listing B.6 – Extrait du résultat (3 sur 62 lignes) de la requête SPARQL du Listing B.5 retournant les adresses dont les coordonnées associées sont situées dans la zone définie.

B.2 Questions du modelet des entités géographiques et de leurs évolutions

Cette section regroupe des questions de compétence et les requêtes SPARQL associées pour d'interroger la dimension temporelle du graphe de connaissances. Les requêtes concernent l'existence d'entités géographiques à une date donnée, la validité d'adresses dans le temps et l'historique des changements affectant les géométries de voies.

B.2.1 Entités géographiques existent à une date donnée

Quelles voies existaient à Paris en 1860 dans le calendrier grégorien (`wd:Q1985727`) ?

```
SELECT DISTINCT ?lm ?lmLabel ?lmType WHERE {
    VALUES (?tStamp ?tPrecision ?tCal) {
        ("1860-01-01T00:00:00"^^xsd:dateTime time:unitDay wd:Q1985727)
    }

    ?lm a :Landmark ; :isLandmarkType ?lmType ; rdfs:label ?lmLabel ;
        :changedBy [:isChangeType ctype:LandmarkAppearance ; :dependsOn
        ↪  ?evApp], [:isChangeType ctype:LandmarkDisappearance ; :dependsOn
        ↪  ?evDis] .

    ?evApp a :Event ; ?pApp ?timeApp .
    ?evDis a :Event ; ?pDis ?timeDis .
    ?timeApp :timeStamp ?tsApp ; :timePrecision ?tpApp ; :timeCalendar ?tCal .
    ?timeDis :timeStamp ?tsDis ; :timePrecision ?tpDis ; :timeCalendar ?tCal .

    FILTER(?pApp IN (:hasTime, :hasTimeBefore, :hasTimeAfter))
    FILTER(?pDis IN (:hasTime, :hasTimeBefore, :hasTimeAfter))
    FILTER(
        ((?pApp = :hasTime       && ?tStamp >= ?tsApp) ||
         (?pApp = :hasTimeBefore && ?tStamp >= ?tsApp))
        &&
        ((?pDis = :hasTime       && ?tStamp <= ?tsDis) ||
         (?pDis = :hasTimeAfter  && ?tStamp <= ?tsDis))
    )
}
```

Listing B.7 – Requête SPARQL pour identifier les voies existantes à une date donnée.

```
lm,lmLabel,lmType
ex:LM_9ccc60d5-7f39-438b-aeb2,Châtenay,ltype:Municipality
ex:LM_a150063e-964a-4901-a9dd,17,ltype:StreetNumber
ex:LM_bed8d11c-abaf-4438-9d9a,Rue Richard Lenoir,ltype:Thoroughfare
ex:LM_12996c02-4475-43ea-be49,155,ltype:StreetNumber
```

Listing B.8 – Extrait du résultat de la requête SPARQL du Listing B.7 identifiant les entités géographiques existant en 1860. L'exécution renvoie 1355 résultats au total.

B.2.2 Temps valide d'une adresse

Pendant quelles années trouve-t-on l'adresse « 14 Place des Vosges, 75004 Paris » ?

```
SELECT DISTINCT ?addr ?geomValue ?timeStampME ?timeStamp0
WHERE {
    BIND("14 Place des Vosges, 75004 Paris"@fr AS ?addrLabel)
    ?addr a :Address ; rdfs:label ?addrLabel ;
             :targets [:hasAttribute [:isAttributeType atype:Geometry ;
             ↪  :hasAttributeVersion ?geomVersion]].
    ?geomVersion :versionValue ?geomValue ; :isMadeEffectiveBy [:dependsOn
    ↪  ?eventME] ;
             :isOutdatedBy [:dependsOn ?event0].
    OPTIONAL {?eventME :hasTime [:timeStamp ?timeStampME]}
    OPTIONAL {?event0 :hasTime [:timeStamp ?timeStamp0]}
}
```

Listing B.9 – Requête SPARQL pour déterminer la période de validité d'une adresse.

B.2.3 Changements de géométrie affectant une voie

Quels événements liés à un changement de géométrie ont concerné la rue de Cotte ?

```
SELECT DISTINCT ?lm ?change ?tStamp ?tStampAfter ?tStampBefore ?odVers ?meVers
WHERE {
    BIND("rue de Cotte"@fr AS ?lmLabel)

    ?lm a :Landmark ; :isLandmarkType ltype:Thoroughfare ;
        rdfs:label ?lmLabel ; :hasAttribute ?attrGeom .
    ?attrGeom a :Attribute ; :isAttributeType atype:Geometry .
    ?change :appliedTo ?attrGeom ; :dependsOn ?event .
    ?event a :Event .

    OPTIONAL { ?event :hasTime [ :timeStamp ?tStamp ] }
    OPTIONAL { ?event :hasTimeAfter [ :timeStamp ?tStampAfter ] }
    OPTIONAL { ?event :hasTimeBefore [ :timeStamp ?tStampBefore ] }
    OPTIONAL { ?change :outdates [ :versionValue ?odVers ] }
    OPTIONAL { ?change :makesEffective [ :versionValue ?meVers ] }
}
```

Listing B.10 – Requête extrayant les événements de changement de géométrie d'une voie.

```
street,change,timeStamp,timeStampAfter,timeStampBefore,outdatedVersion,madeE
    ffectiveVersion
ex:LM_94cc4644-dca9-430d-936c-579eee02182b,ex:CG_33bb5b46-07c8-4e6b-9415-4df
    74ebb2ae8,,1889-01-01T00:00:00Z,2025-04-01T00:00:00Z,"MULTILINESTRING(..
    .)", "POLYGON(...)"
ex:LM_94cc4644-dca9-430d-936c-579eee02182b,ex:CG_ffda6b4e-154f-43a9-ad7d-e81
    f49181ec8,,1850-01-01T00:00:00Z,1887-01-01T00:00:00Z,"MULTILINESTRING(..
    .)","MULTILINESTRING(...)"
ex:LM_94cc4644-dca9-430d-936c-579eee02182b,ex:CG_78296b15-8c83-459e-9a8e-364
    892f4c6d2,,,1810-01-01T00:00:00Z,,"LINESTRING(...)"
ex:LM_94cc4644-dca9-430d-936c-579eee02182b,ex:CG_bfad96c1-d2c4-44d9-8ceb-351
    43969cba2,,2025-06-19T18:03:20.227837Z,,"POLYGON(...)"
```

Listing B.11 – Extrait du résultat CSV de la requête SPARQL du Listing B.10. Les géométries (`LINESTRING`, `MULTILINESTRING`, `POLYGON`) sont abrégées pour la lisibilité.

B.2.4 Évolution de la géométrie d'une voie

Quelle sont les différentes géométries de la rue Cotte (ensemble des géométries successives avec leur intervalle de validité) ?

```
SELECT DISTINCT
  ?street ?geomValue ?change
  ?timeStampME ?timeStampAfterME ?timeStampBeforeME
  ?timeStampO ?timeStampAfterO ?timeStampBeforeO
WHERE {
  BIND("rue de Cotte"@fr AS ?streetLabel)

  ?street a :Landmark ; :isLandmarkType ltype:Thoroughfare ;
          :hasAttribute ?attrGeom ; rdfs:label ?streetLabel .

  ?attrGeom a :Attribute ; :isAttributeType atype:Geometry ;
            :hasAttributeVersion ?geomVersion .

  ?geomVersion :versionValue ?geomValue ;
               :isMadeEffectiveBy [:dependsOn ?eventME] ;
               :isOutdatedBy [:dependsOn ?eventO] .

  OPTIONAL { ?eventME :hasTime [:timeStamp ?timeStampME] }
  OPTIONAL { ?eventME :hasTimeAfter [:timeStamp ?timeStampAfterME] }
  OPTIONAL { ?eventME :hasTimeBefore [:timeStamp ?timeStampBeforeME] }

  OPTIONAL { ?eventO :hasTime [:timeStamp ?timeStampO] }
  OPTIONAL { ?eventO :hasTimeAfter [:timeStamp ?timeStampAfterO] }
  OPTIONAL { ?eventO :hasTimeBefore [:timeStamp ?timeStampBeforeO] }
}
```

Listing B.12 – Requête SPARQL pour retracer l'histoire des géométries successives d'une voie donnée.

Annexe C

Algorithmes pour la construction de graphes et la mise en œuvre

C.1 Algorithmes de la méthode de construction du graphe

L'ensemble des algorithmes illustre la méthode décrite dans le chapitre 6. Les algorithmes correspondent à la création du graphe de faits, la déduction des dates d'apparition et de disparition des entités et des relation ainsi que la reconstitution des évolutions d'attributs.

C.1.1 Initialisation des entités et propagation des liens

Ici, les algorithmes initialisent les entités du graphe des faits et propagent les liens de provenance vers les relations, attributs, changements et événements.

Algorithme C.1 : Initialisation des entités du graphe des faits

Input : O_L : ensemble des observations d'entités,
α : seuil de similarité pour rattacher à une observation à une entité réelle
Output : L : ensemble des entités du graphe,
C : ensemble des changements créés (apparition / disparition),
Triplets `hasTrace` et `appliedTo` associés aux entités et changements
$L \leftarrow \emptyset$
$C \leftarrow \emptyset$
for $o \in O_L$ **do**
  $l^* \leftarrow \arg\max_{l \in L} \text{sim}'(o, l)$; `// entité la plus similaire existante`
  if $l^* \neq \emptyset$ ***and*** $sim'(o, l^*) \geq \alpha$ **then**
    Générer les triplets `hasTrace`(l^*, o)
  else
    Générer une nouvelle entité l;
    $L \leftarrow L \cup \{l\}$;
    Générer les triplets `hasTrace`(l, o);
    Générer c_l^{app} de type `ctype:LandmarkAppearance`;
    Générer c_l^{dis} de type `ctype:LandmarkDisappearance`;
    $C \leftarrow C \cup \{c_l^{app}, c_l^{dis}\}$;
    Générer les triplets `appliedTo`(c_l^{app}, l) et `appliedTo`(c_l^{dis}, l)

Algorithme C.2 : Propagation des liens de provenance vers relations, attributs, changements et événements

Input : O_R : observations de relations,
O_A : observations d'attributs,
O_C : observations de changements,
O_E : observations d'événements,
L : ensemble des entités,
R : ensemble des relations
Output : Triplets `hasTrace` pour toutes les observations et `owl:sameAs` pour des événéments similaires

$C_R, E_R \leftarrow \emptyset, \emptyset$
for $o \in O_R$ **do**
  $\mathcal{O}_o \leftarrow \{o_l \mid \texttt{locatum}(o, o_l) \vee \texttt{relatum}(o, o_l)\}$
  if $\forall o_l \in \mathcal{O}_o, \exists l \in L, \textit{hasTrace}(o_l, l)$ **then**
    $r^* \leftarrow \arg\max_{r \in R} \text{sim_relation}(o, r)$; `// relation la plus similaire`
    if $r^* \neq \emptyset$ **then**
      Générer le triplet $\texttt{hasTrace}(o, r^*)$
    else
      Générer une nouvelle relation r;
      $R \leftarrow R \cup \{r\}$;
      Générer le triplet $\texttt{hasTrace}(o, r)$;
      Générer c_r^{app} de type `ctype:LandmarkRelationAppearance`;
      Générer c_r^{dis} de type `ctype:LandmarkRelationDisappearance`;
      $C_R \leftarrow C_R \cup \{c_r^{app}, c_r^{dis}\}$;
      Générer les triplets $\texttt{appliedTo}(c_r^{app}, r)$ et $\texttt{appliedTo}(c_r^{dis}, r)$

for $o \in O_A$ **do**
  Trouver $o_l \in O_L$ associé à o;
  Trouver $l \in L$ tel que $\texttt{hasTrace}(o_l, l)$;
  if *un attribut a de l du même type que o existe* **then**
    Générer $\texttt{hasTrace}(o, a)$
  else
    Générer un nouvel attribut a pour l;
    Générer $\texttt{hasTrace}(o, a)$

for $o \in O_C$ **do**
  Identifier l'entité ou la relation $x \in (L \cup R)$ appliquée;
  Trouver le changement c associé à x;
  Générer le triplet $\texttt{hasTrace}(o, c)$

for $o_e \in O_E$ **do**
  for $o_c \in O_C$ *dépendant de* o_e **do**
    Trouver $c \in C$ tel que $\texttt{hasTrace}(o_c, c)$;
    Trouver $e \in E$ tel que $\texttt{dependsOn}(c, e)$;
    Générer le triplet $\texttt{hasTrace}(o_e, e)$
  if $\exists e_1, e_2 \in E$ *avec* $\textit{hasTrace}(o_e, e_1)$ *et* $\textit{hasTrace}(o_e, e_2)$ **then**
    Générer le triplet $\texttt{owl:sameAs}(e_1, e_2)$

C.1.2 Inférence des dates d'apparition et de disparition

Les algorithmes présentés dans cette sous-section permettent d'inférer les dates des événements associés aux changements d'apparition et de disparition des entités, ainsi que des relations entre entités. Ils traitent les observations disponibles pour générer des contraintes temporelles, estimer des intervalles flous si nécessaire, et produire les triplets temporels correspondant aux événements d'apparition et de disparition dans le graphe.

Algorithme C.3 : Inférence des dates d'apparition et de disparition d'une entité ou relation. La fonction *estimation_contrainte_temporelle()* est introduite dans l'algorithme C.4.

Input : x une entité ou une relation
Output : Contraintes temporelles pour les événements e_{app} et e_{dis}
$O \leftarrow obs(x)$;
$C(e_{app}) \leftarrow \emptyset$;
$C(e_{dis}) \leftarrow \emptyset$;
foreach $o \in O$ **do**
  if *o décrit un état spatio-temporel* **then**
    récupérer l'intervalle associé $T = [t_{deb}, t_{fin}]$;
    ajouter $(\texttt{hasTimeBefore}, t_{deb})$ à $C(e_{app})$;
    ajouter $(\texttt{hasTimeAfter}, t_{fin})$ à $C(e_{dis})$;
  else if *o décrit une évolution* **then**
    récupérer l'instant t associé à l'événement observé $e \in \{e_{app}, e_{dis}\}$;
    ajouter $(\texttt{hasTime}, t)$ à $C(e)$;

foreach $e \in \{e_{app}, e_{dis}\}$ **do**
  estimation_contrainte_temporelle(e, C);
return triplets temporels estimés pour e_{app} et e_{dis};

Algorithme C.4 : Fonction *estimation_contrainte_temporelle()* : calcule et génère les triplets temporels pour un événement à partir des contraintes observées.

Input : e un événement,
C l'ensemble des contraintes temporelles associées
Output : Triplets temporels générés pour e
if *une unique contrainte* $(\texttt{hasTime}, t)$ *existe dans* $C(e)$ **then**
  générer le triplet $\texttt{hasTime}(e, t)$;
else if *plusieurs contraintes* $(\texttt{hasTime}, t_i)$ *existent* **then**
  définir un instant flou t conciliant les valeurs t_i;
  générer le triplet $\texttt{hasTime}(e, t)$;
else
  if *il existe des couples* $(\texttt{hasTimeAfter}, t_i)$ *dans* $C(e)$ **then**
    $t_{max} \leftarrow \max\{t_i \mid (\texttt{hasTimeAfter}, t_i) \in C(e)\}$;
    générer le triplet $\texttt{hasTimeAfter}(e, t_{max})$;
  if *il existe des couples* $(\texttt{hasTimeBefore}, t_i)$ *dans* $C(e)$ **then**
    $t_{min} \leftarrow \min\{t_i \mid (\texttt{hasTimeBefore}, t_i) \in C(e)\}$;
    générer le triplet $\texttt{hasTimeBefore}(e, t_{min})$;

C.1.3 Reconstitution de l'évolution des attributs

Les algorithmes présentés dans cette sous-section permettent de reconstituer l'évolution des attributs élémentaires des entités dans le graphe. Ils couvrent l'initialisation des versions d'attributs élémentaires ainsi que leur exploitation et fusion pour produire une représentation cohérente de l'évolution temporelle des attributs.

C.1.3.1 Initialisation des versions d'attributs élémentaires

Cette partie inclut des algorithmes créant et organisant les versions élémentaires pour chaque attribut. Ils permettent d'initialiser les changements et versions élémentaires, d'associer chaque version à ses changements et de préparer les données pour la reconstitution.

Algorithme C.5 : Initialisation des changements et des versions élémentaires pour un attribut a

Input : Versions $V = \{v_1, \ldots, v_m\}$, changements $C = \{c_1, \ldots, c_n\}$
Output : Ensemble de changements élémentaires Γ, ensemble de versions élémentaires B
$C_V \leftarrow \emptyset$; `// changements fictifs pour les versions avec temps valide`
for $v_i \in V$ **do**
  Générer γ_i^{α} et γ_i^{β} ; `// changement d'apparition et de disparition`
  $C_V \leftarrow C_V \cup \{\gamma_i^{\alpha}, \gamma_i^{\beta}\}$
$\Omega \leftarrow C \cup C_V$;
Associer un instant t_ω à chaque $\omega \in \Omega$;
for *tous les sous-ensembles* $S \subseteq \Omega$ *avec* t_ω *identique pour tout* $\omega \in S$ **do**
  Générer $\gamma_S \in \Gamma$; `// changement élémentaire`
Ajouter $\gamma_{-\infty}, \gamma_{+\infty}$ à Γ;
Trier Γ temporellement et générer des triplets `hasNextChange` entre deux changements successifs;
for *chaque paire de changements successifs* (γ_i, γ_j) **do**
  Générer une version élémentaire β avec `makesEffective`(γ_i, β) et `outdates`(γ_j, β)

Algorithme C.6 : Affiliation des changements et versions aux factoïdes pour un attribut a

Input : Changements élémentaires Γ, versions élémentaires B, changements C, versions V
Output : Triplets `hasTrace` et `derives`
for $c \in C$ **do**
  Trouver $\gamma \in \Gamma$ tel que γ dérive de c;
  Générer le triplet `hasTrace`(γ, c);
for $v \in V$ **do**
  for $\beta \in B$ **do**
    if β *et* v *partagent un changement* γ *dérivé de* c **then**
      Générer le triplet `hasTrace`(β, v);
for *chaque paire de versions* (β_i, β_j) *avec* ***hasTrace***(β_i, v) *et* ***hasTrace***(β_j, v) **do**
  for β *entre* β_i *et* β_j *via* ***hasNextVersion***$^+$ **do**
    Générer le triplet `hasTrace`(β, v)

C.1.3.2 Reconstitution de l'évolution des attributs avec leurs versions élémentaires

Cette partie présente les algorithmes qui utilisent les versions élémentaires pour reconstruire l'évolution complète des attributs. Ils gèrent d'une part la suppression des versions lacunaires, et d'autre part la fusion des versions similaires successives et l'intégration des versions dans le graphe afin de produire une représentation cohérente de l'évolution temporelle des attributs.

Algorithme C.7 : Suppression des versions lacunaires

Input : Versions élémentaires B, changements élémentaires Γ, triplets `hasTrace`
Output : Versions élémentaires B après suppression et fusion de versions non tracées

for $\beta \in B$ **do**
  if $\nexists v \in V, \mathit{hasTrace}(\beta, v)$ **then**
    Identifier $\gamma_{app}, \gamma_{dis}$: changements d'apparition et de disparition de β;
    if *aucun des deux changements n'est tracé* **then**
      Fusionner γ_{app} et γ_{dis};
      Supprimer β;
    else
      if *un seul changement tracé* **then**
        Identifier γ_{nt} : le changement non tracé parmi $\{\gamma_{app}, \gamma_{dis}\}$;
        Identifier γ_t : l'autre changement (tracé);
        Identifier β' et $r \in \{\texttt{makesEffective}, \texttt{outdates}\}$ tels que $r(\gamma_{nt}, \beta')$;
        Supprimer le triplet $r(\gamma_{nt}, \beta')$;
        Générer le triplet $r(\gamma_t, \beta')$;
        Supprimer γ_{nt};
      else
        skip (ne rien faire)

Algorithme C.8 : Fusion des versions élémentaires similaires successives

Input : Versions élémentaires B, triplets `sameVersionValueAs` et `differentVersionValueFrom`, changements Γ
Output : Versions élémentaires B après fusion des versions similaires

for *toutes les paires de versions* (β_i, β_j) *consécutives* **do**
  if *il existe* $\gamma \in \Gamma$ *non tracé tel que* $\mathit{makesEffective}(\gamma, \beta_i)$ *et* $\mathit{outdates}(\gamma, \beta_j)$ **then**
    if *les versions* β_i *et* β_j *dérivent de versions similaires* **then**
      `// Indiquer qu'il faut fusionner`
      Générer le triplet `toBeMergedWith`(β_i, β_j);

for *chaque ensemble de versions* $\mathcal{B} \subseteq B$ *connecté par* `toBeMergedWith` *transitif* **do**
  `// Créer la nouvelle version fusionnée`
  Générer nouvelle version β avec c_{app} = changement de la version la plus ancienne de $\mathcal{B}$, et c_{dis} = changement de la version la plus récente de $\mathcal{B}$;
  Supprimer toutes les versions de $\mathcal{B}$;

C.2 Mise en œuvre de la méthode

Cette section présente l'algorithme principal (algorithme C.9) de construction du graphe de faits et de reconstitution des attributs. Il s'appuie sur des fonctions utilitaires permettant de comparer les versions des attributs, que ce soit pour les noms (voir algorithme C.10) ou pour les géométries (voir algorithme C.11).

Algorithme C.9 : Construction du graphe de faits et reconstitution des attributs

Input : Graphes nommés de factoïdes $G_1, \ldots, G_n$,
$d_{\max}$: distance maximale pour considérer deux géométries ponctuelles similaires,
$\alpha_{\min}$: seuil minimal de recouvrement pour des géométries de type polygone
Output : Graphe de faits G_F avec triplets `hasTrace`, `sameVersionValueAs` et `differentVersionValueFrom`

$G_F \leftarrow \emptyset$
Créer graphe temporaire $G_T \leftarrow \emptyset$
Créer graphe des comparaisons $G_C \leftarrow \emptyset$
Créer graphe de mapping factoïdes / faits $G_M \leftarrow \emptyset$
for $G_i \in \{G_1, \ldots, G_n\}$ **do**
  for *chaque ressource* $R_S \in G_i$ **do**
    Créer ou récupérer l'entité réelle $R_F \in G_F$ décrite par R_S;
    Générer le triplet $\texttt{hasTrace}(R_F, R_S)$ dans G_M;
for *entité ou relation* $E \in G_F$ **do**
  Déduire dates d'apparition et de disparition à partir des graphes nommés
for *paires de versions* (v_i, v_j) *d'un attribut* A **do**
  if A *est de type identifiant* **then**
    $sim \leftarrow (v_i = v_j)$
  else if A *est de type nom* **then**
    $sim \leftarrow versions_noms_similaires(v_i, v_j)$
  else if A *est de type géométrie* **then**
    $sim \leftarrow versions_geometries_similaires(v_i, v_j, d_{max}, \alpha_{min})$
  if *sim* **then**
    Générer le triplet $\texttt{sameVersionValueAs}(v_i, v_j)$ dans G_C
  else
    Générer le triplet $\texttt{differentVersionValueFrom}(v_i, v_j)$ dans G_C
for *attribut* A *d'une entité* $E \in G_F$ **do**
  $C \leftarrow \{c_i \mid \texttt{Attribut}(A) \wedge \texttt{hasTrace}(A, A_s) \wedge \texttt{appliedTo}(c_i, A_s)\}$;
  $V \leftarrow \{v_i \mid \texttt{Attribut}(A) \wedge \texttt{hasTrace}(A, A_s) \wedge \texttt{hasAttributeVersion}(A_s, v_i)\}$;
  Avec C et V, générer l'évolution de l'attribut A sous forme de :
    — Γ, l'ensemble des changements reconstitués pour A ;
    — B, l'ensemble des versions reconstituées pour A.
for *triplet* $t \in G_T$ **do**
  if t *concerne un élément de* B *ou* Γ **then**
    if t *est un triplet* `hasTrace` **then**
      Insérer t dans G_M
    else
      Insérer t dans G_F
Supprimer graphe temporaire G_T

Algorithme C.10 : Fonction *versions_noms_similaires()* : test de similarité entre versions de type nom

Input : v_i, v_j : versions d'un attribut de type nom
Output : *sim* : booléen indiquant si les noms sont similaires
Récupérer n_i tel que $\texttt{versionValue}(v_i, n_i)$;
Récupérer n_j tel que $\texttt{versionValue}(v_j, n_j)$;
Récupérer l_type_i tel que
$\texttt{Landmark}(l) \wedge \texttt{isLandmarkType}(l, l_type_i) \wedge \texttt{hasAttribute}(l, a) \wedge$
$\texttt{isAttributeType}(a, \texttt{atype:Name}) \wedge \texttt{hasAttributeVersion}(a, v_i)$;
Récupérer l_type_j de la même manière pour v_j;
Récupérer les langues $lang_i$ et $lang_j$ associées aux labels;
$s_i \leftarrow \texttt{simplifier_nom}(n_i, lang_i, l_type_i)$;
$s_j \leftarrow \texttt{simplifier_nom}(n_j, lang_j, l_type_j)$;
$sim \leftarrow (s_i = s_j)$;
return *sim*;

Algorithme C.11 : Fonction *versions_geometries_similaires()* : test de similarité entre versions de type géométrie

Input : v_i, v_j : versions d'un attribut de type géométrie,
$d_{\max}$: distance maximale pour les entités ponctuelles,
$\alpha_{\min}$: seuil minimal de recouvrement
Output : *sim* : booléen indiquant si les versions ont des géométries
$sim \leftarrow$ **false**
Récupérer g_i tel que $\texttt{versionValue}(v_i, g_i)$ et g_j tel que $\texttt{versionValue}(v_j, g_j)$;
if $type(g_i) = point$ ***and*** $type(g_j) = point$ **then**
  $d \leftarrow distance(g_i, g_j)$
  if $d \leq d_{\max}$ **then**
    $sim \leftarrow$ **true**
else
  if $type(g_i) = polyligne$ **then**
    $g_i \leftarrow buffer(g_i)$
  if $type(g_j) = polyligne$ **then**
    $g_j \leftarrow buffer(g_j)$
  $\alpha \leftarrow \frac{surface(g_i \cap g_j)}{surface(g_i \cup g_j)}$,
  if $\alpha \geq \alpha_{\min}$ **then**
    $sim \leftarrow$ **true**
return *sim*

Bibliographie

Nathalie Abadie : *Formalisation, acquisition et mise en œuvre de connaissances pour l'intégration virtuelle de bases de données géographiques : les spécifications au cœur du processus d'intégration*. phdthesis, Université Paris-Est, novembre 2012. URL https://theses.hal.science/tel-00794395.

Jacky Akoka, Isabelle Comyn-Wattiau et Cédric du Mouza : Évaluer la crédibilité des sources historiques. *In Atelier EGC DAHLIA*, Bruxelles, Belgique, janvier 2020. URL https://cnam.hal.science/hal-02464556.

James F. Allen : Maintaining Knowledge about Temporal Intervals. *In* Daniel S. Weld et Johan de Kleer, éditeurs : *Readings in Qualitative Reasoning About Physical Systems*, pages 361–372. Morgan Kaufmann, janvier 1990. ISBN 978-1-4832-1447-4. URL https://www.sciencedirect.com/science/article/pii/B978148321447450033X.

Adolphe Alphand et Louis-François Sébastien Fauve : Atlas municipal des vingt arrondissements de la Ville de Paris dressé sous l'Administration de M. Ferdinand Duval, Préfet, sous la Direction de M. Alphand ; par les soins de M. L. Fauve, géomètre en chef, avec le concours des Géomètres du Plan de Paris, 1888. URL https://catalogue.bnf.fr/ark:/12148/cb40583717c.

American Planning Association : Street Naming and House Numbering Systems. PAS Report 13, American Planning Association, Chigago, USA, avril 1950. URL https://www.planning.org/pas/reports/report13.htm.

J Andriveau-Goujon : Plan de Paris fortifié et des communes environnantes : 1849 / Le plan et la lettre gravés par P. Rousset, 1849. URL https://bibliotheques-specialisees.paris.fr/ark:/73873/pf0002192000.

Marc P Armstrong : Temporality in spatial databases. *GIS/LIS 88 Proceedings : Accessing the world*, pages 880–889, 1988.

Luigi Asprino, Enrico Daga, Aldo Gangemi et Paul Mulholland : Knowledge Graph Construction with a *Façade* : A Unified Method to Access Heterogeneous Data Sources on the Web. *ACM Transactions on Internet Technology*, 23(1):1–31, février 2023. ISSN 1533-5399, 1557-6051. URL https://dl.acm.org/doi/10.1145/3555312.

Robert Audi, éditeur. *The Cambridge dictionary of philosophy*. Cambridge University Press, Cambridge, cambridge university press édition, 1999. ISBN 978-1-107-01505-0 978-1-107-64379-6.

Srividya K. Bansal et Sebastian Kagemann : Integrating Big Data : A Semantic Extract-Transform-Load Framework. *Computer*, 48(3):42–50, mars 2015. ISSN 1558-0814. URL https://ieeexplore.ieee.org/abstract/document/7063172.

Senén Barro, Roque Marín, José Mira et Alfonso R. Patón : A model and a language for the fuzzy representation and handling of time. *Fuzzy Sets and Systems*, 61(2):153–175, janvier 1994. ISSN 0165-0114. URL https://www.sciencedirect.com/science/article/pii/0165011494902313.

Jacques Barzun et Henry F. Graff : *The modern researcher : the classic work on research and writing*. Harcourt Brace Jovanovich, San Diego, 4. ed édition, 1985. ISBN 978-0-15-161479-0 978-0-15-562512-9.

Hannah Bast, Patrick Brosi, Johannes Kalmbach et Axel Lehmann : An Efficient RDF Converter and SPARQL Endpoint for the Complete OpenStreetMap Data. *In Proceedings of the 29th International Conference on Advances in Geographic Information Systems*, pages 536–539, Beijing China, novembre 2021. ACM. ISBN 978-1-4503-8664-7. URL https://dl.acm.org/doi/10.1145/3474717.3484256.

Jonathan Bennett : *OpenStreetMap : be your own cartographer*. Packt Publishing, Birmingham, 1st ed édition, 2010. ISBN 978-1-84719-751-1 978-1-84719-750-4.

Camille Bernard, Marlène Villanova-Oliver, Jérôme Gensel et Hy Dao : TSN et TSN-change : ontologies pour représenter l'évolution des découpages territoriaux statistiques. *In SAGEO 2017 - Spatial Analysis and GEOmatics*, Rouen, France, novembre 2017. GdR Magis CNRS. URL https://hal.science/hal-01634424.

Camille Bernard, Marlène Villanova-Oliver, Jérôme Gensel et Hy Dao : Ontologies pour représenter l'évolution des découpages territoriaux statistiques. *Revue Internationale de Géomatique*, 28(4):409–437, décembre 2018. ISSN 1260-5875, 2116-7060. URL https://rig.revuesonline.com/10.3166/rig.2019.00069.

Charly Bernard, Nathalie Abadie, Bertrand Duménieu et Julien Perret : PeGazUs : une méthode de reconstitution de l'évolution des entités géographiques à partir de données hétérogènes et fragmentaires. *In* Fleur Mougin, éditeur : *36es Journées francophones d'Ingénierie des Connaissances (IC 2025) @ Plate-Forme Intelligence Artificielle (PFIA 2025)*, Dijon, France, juillet 2025a. URL https://hal.science/hal-05131104.

Charly Bernard, Nathalie Abadie, Bertrand Duménieu et Julien Perret : Reconstructing the Temporal Evolution of Geographic Entities from Fragmentary Knowledge. *In International Knowledge Graph and Semantic Web Conference*, Leipzig, Germany, novembre 2025b. URL https://hal.science/hal-05308831.

Charly Bernard, Nathalie Abadie, Julien Perret et Bertrand Duménieu : Création d'un référentiel géo-historique d'adresses à partir de sources multiples. *In* Nida MEDDOURI, Loïc SALMON et Aurélie LEBORGNE, éditeurs : *GAST -Gestion et l'Analyse de données Spatiales et Temporelles*, Dijon, France, janvier 2024. URL https://hal.science/hal-04490732.

Charly Bernard, Solenn Tual, Nathalie Abadie, Bertrand Duménieu, Joseph Chazalon et Julien Perret : PeGazUs : A Knowledge Graph Based Approach to Build Urban Perpetual Gazetteers. *In* Mehwish Alam, Marco Rospocher, Marieke Van Erp, Laura Hollink et Genet Asefa Gesese, éditeurs : *Knowledge Engineering and Knowledge Management*, volume 15370, pages 364–381. Springer Nature Switzerland, Cham, 2025c. ISBN 978-3-031-77791-2 978-3-031-77792-9. URL https://link.springer.com/10.1007/978-3-031-77792-9_22.

Adolphe Berty, Henri Legrand, Lazare-Maurice Tisserand, Théodore Vacquer et Camille Platon : *Topographie historique du vieux Paris. Région du Louvre et des Tuileries*, volume 1. Imprimerie impériale, Paris, 1866. URL https://gallica.bnf.fr/ark:/12148/bpt6k6517947c.

Vinayak Borkar, Kaustubh Deshmukh et Sunita Sarawagi : Automatically Extracting Structure from Free Text Addresses. *IEEE Data Eng. Bull.*, 23:27–32, janvier 2000.

Lucas Bourel, William Charles, Nathalie Jane Hernandez, Nathalie Aussenac-Gilles, Victor Gay et Sébastien Poublanc : Graphes de connaissances pour représenter et analyser l'évolution des territoires en Histoire. *In* Nicolas Lasolle, Olivier Bruneau et Jean Lieber, éditeurs : *Journées humanités numériques et Web sémantique 2022*, pages 23–37, Nancy, France, juin 2022. Zenodo. URL https://ut3-toulouseinp.hal.science/hal-03760353.

Suzanne Briet : *Qu'est-ce que la documentation ?* Éditions documentaires, industrielles et techniques, Paris, 2024. URL https://hal.science/hal-04816907.

André Buisson : Jean-Luc Arnaud, Analyse spatiale, cartographie et histoire urbaine. *Géocarrefour*, 84(4):228–228, 2009. ISSN 1627-4873, 1960-601X. URL https://journals.openedition.org/geocarrefour/7332.

Mattia Bunel : Un état de l'art sur l'imprécision spatiale et sa modélisation. *Cybergeo : European Journal of Geography*, février 2021. ISSN 1278-3366. URL https://journals.openedition.org/cybergeo/36126.

Lucie Cadorel, Denis Overal et Andrea G. B. Tettamanzi : Fuzzy representation of vague spatial descriptions in real estate advertisements. *In Proceedings of the 6th ACM SIGSPATIAL International Workshop on Location-based Recommendations, Geosocial Networks and Geoadvertising*, pages 1–4, Seattle Washington, novembre 2022a. ACM. ISBN 978-1-4503-9540-3. URL https://dl.acm.org/doi/10.1145/3557992.3565994.

Lucie Cadorel, Andrea G. B. Tettamanzi et Fabien Gandon : Towards a representation of uncertain geospatial information in knowledge graphs. *In Proceedings of the 1st ACM SIGSPATIAL International Workshop on Geospatial Knowledge Graphs*, pages 1–2, Seattle Washington, novembre 2022b. ACM. ISBN 978-1-4503-9538-0. URL https://dl.acm.org/doi/10.1145/3557990.3567588.

Wentao Cai, Shengrui Wang et Qingshan Jiang : Address Extraction : Extraction of Location-Based Information from the Web. *In* Yanchun Zhang, Katsumi Tanaka, Jeffrey Xu Yu, Shan Wang et Minglu Li, éditeurs : *Web Technologies Research and Development - APWeb 2005*, Lecture Notes in Computer Science, pages 925–937, Berlin, Heidelberg, 2005. Springer. ISBN 978-3-540-31849-1.

Nicholas J Car : Representing spatial uncertainty and allowing for probabilistic topological functions with SUFF, an extension to GeoSPARQL. *In* Timo Homburg, Beyza Yaman, Mohamed Ahmed Sherif et Axel-Cyrille Ngonga Ngomo, éditeurs : *GeoLD 2024 : 6th International Workshop on Geospatial Linked Data at ESWC 2024*, Hersonissos, Greece, mai 2024. URL https://ceur-ws.org/Vol-3743/paper4.pdf.

Nicholas J. Car, Timo Homburg, Matthew Perry, Frans Knibbe, Simon J.D. Cox, Joseph Abhayaratna, Mathias Bonduel, Paul J. Cripps et Krzysztof Janowicz : GeoSPARQL – A Geographic Query Language for RDF Data. OGC Standard 22-047r1, Open Geospatial Consortium, 2012. URL https://www.ogc.org/standards/geosparql/.

Edgardo D. Carosella, Bertrand Saint-Sernin, Philippe Capelle et S.E. Marcelo Sánchez Sorondo : *L'identité changeante de l'individu : La constante construction du soi*. Éditions L'Harmattan, 2008. ISBN 978-2-296-19883-8. URL https://books.google.fr/books?id=mRT77fqLIjUC.

Simona Cerutti et Isabelle Grangaud : Sources et mises en contexte. Quelques réflexions autour des conditions de la comparaison. *In* Florent Brayard, éditeur : *Des contextes en histoire*, Actes du Forum du Centre de recherches historiques, pages 91–102. Publications du Centre de recherches historiques, Paris, 2014. URL https://crh.ehess.fr/docannexe/file/3845/05bcrh13cerutti.pdf.

William Charles : *From document and data interpretation to historical knowledge : an ontological approach applied to the case of territories*. Theses, Université de Toulouse, juillet 2025. URL https://theses.hal.science/tel-05262507. Issue : 2025TLSEP056.

William Charles et Nathalie Jane Hernandez : HHT : An ontology to represent territorial dynamics for digital humanities. *In* Nathalie Aussenac-Gilles, Antony Glaton, Torsten Hamman et Maria M. Hedblom, éditeurs : *Formal Ontology in Information Systems - Proceedings of the Thirteenth International Conference, FOIS 2023.Frontiers in Artificial Intelligence and Applications*, volume 377 de *Frontiers in Artificial Intelligence and Applications*, pages 257 – 271, Sherbrooke, Canada, juillet 2023. IOS Press. URL https://hal.science/hal-04310997.

Yizi Chen, Joseph Chazalon, Edwin Carlinet, Minh Ôn Vũ Ngoc, Clément Mallet et Julien Perret : Automatic vectorization of historical maps : A benchmark. *PLOS ONE*, 19(2):e0298217, février 2024. ISSN 1932-6203. URL https://dx.plos.org/10.1371/journal.pone.0298217.

Peter Christen : A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication. *IEEE Transactions on Knowledge and Data Engineering*, 24(9):1537–1555, septembre 2012. ISSN 1558-2191. URL https://ieeexplore.ieee.org/abstract/document/5887335.

Zhendong Chu, Haiyun Jiang, Yanghua Xiao et Wei Wang : InSRL : A Multi-view Learning Framework Fusing Multiple Information Sources for Distantly-supervised Relation Extraction, décembre 2020. URL http://arxiv.org/abs/2012.09370. Issue : arXiv :2012.09370 arXiv : 2012.09370 [cs].

Christophe Claramunt et Marius Thériault : Managing Time in GIS An Event-Oriented Approach. *In* James Clifford et Alexander Tuzhilin, éditeurs : *Recent Advances in Temporal Databases*, Workshops in Computing, pages 23–42, London, 1995. Springer. ISBN 978-1-4471-3033-8.

Christophe Claramunt, Marius Thériault et Christine Parent : A qualitative representation of evolving spatial entities in two-dimensional topological spaces. *In Innovations In GIS 5*, pages 128–142. CRC Press, mars 1998. ISBN 978-0-7484-0810-8 978-1-4822-6311-4. URL http://www.crcnetbase.com/doi/10.1201/b16831-16.

Eliseo Clementini, Paolino Di Felice et Peter Van Oosterom : A small set of formal topological relationships suitable for end-user interaction. *In* David Abel et Beng Chin Ooi, éditeurs : *Advances in Spatial Databases*, volume 692, pages 277–295. Springer Berlin Heidelberg, Berlin, Heidelberg, 1993. ISBN 978-3-540-56869-8 978-3-540-47765-5. URL http://link.springer.com/10.1007/3-540-56869-7_16.

Serena Coetzee et Antony K. Cooper : What is an address in South Africa ? *South African Journal of Science*, 103(11):449–458, décembre 2007. ISSN 0038-2353. URL http://www.scielo.org.za/scielo.php?script=sci_abstract&pid=S0038-23532007000600006&lng=en&nrm=iso&tlng=en.

Serena Coetzee, Antony K Cooper et Jeofrey Ditsela : Towards good principles for the design of a national addressing scheme. *In 25th International Cartographic Conference (ICC 2011)*. French Committee of Cartography Paris, France, 2011.

Claude Collet : Analyse spatiale, géomatique et systèmes d'information géographique. *Revue internationale de géomatique*, 15(4):393–414, décembre 2005. ISSN 12605875. URL http://geo.revuesonline.com/article.jsp?articleId=7272.

Laurent Costa : La construction de référentiels géohistoriques : un enjeu pour l'interdisciplinarité dans les sciences historiques. *L'Espace géographique*, 41(4):340–351, 2012. ISSN 0046-2497. URL https://shs.cairn.info/revue-espace-geographique-2012-4-page-340.

Benoît Costes : *Vers la construction d'un référentiel géographique ancien : un modèle de graphe agrégé pour intégrer, qualifier et analyser des réseaux géohistoriques*. phdthesis, Université Paris-Est, novembre 2016. URL https://theses.hal.science/tel-01565850.

Simon Cox : RDF representation of 2016 edition of International Chronostratigraphic Chart (Geologic Timescale), 2017. URL https://data.csiro.au/collections/#collection/CIcsiro:22654v1.

Karen Coyle et Tom Baker : Guidelines for Dublin Core™ Application Profiles, mai 2009. URL https://www.dublincore.org/specifications/dublin-core/profile-guidelines/.

Rémi Cura, Bertrand Dumenieu, Nathalie Abadie, Benoit Costes, Julien Perret et Maurizio Gribaudi : Historical Collaborative Geocoding. *ISPRS International Journal of Geo-Information*, 7(7):262, juillet 2018. ISSN 2220-9964. URL https://www.mdpi.com/2220-9964/7/7/262.

Souripriya Das, Seema Sundara et Richard Cyganiak : R2RML : RDB to RDF Mapping Language. Rapport technique, World Wide Web Consortium, 2012. URL https://www.w3.org/TR/r2rml/.

Clodoveu A. Davis et Frederico T. Fonseca : Assessing the Certainty of Locations Produced by an Address Geocoding System. *GeoInformatica*, 11(1):103–129, mars 2007. ISSN 1573-7624. URL https://doi.org/10.1007/s10707-006-0015-7.

Cyril de Runz : *Imperfection, temps et espace : modélisation, analyse et visualisation dans un SIG archéologique*. phdthesis, Université de Reims - Champagne Ardenne, novembre 2008. URL https://theses.hal.science/tel-00560668.

Géraldine Del Mondo : *Un modèle de graphe spatio-temporel pour représenter l'évolution d'entités géographiques*. phdthesis, Université de Bretagne occidentale, Brest, octobre 2011. URL https://theses.hal.science/tel-00651682.

Jean Denègre et François Salgé : *Les systèmes d'information géographique*. Numéro 3122 *in* Que sais-je ? Presses universitaires de France, Paris, 1996. ISBN 978-2-13-047932-1.

Thierry Depaulis : Vaugeois, tabletier, connu et inconnu (1re partie). *Le Vieux Papier*, 41(421):106–113, juillet 2016. URL https://www.academia.edu/42192495/_Vaugeois_tabletier_connu_et_inconnu_I_.

Thomas Devogele : *Processus d'intégration et d'appariement de Bases de Données Géographiques : Application à une base de données routières multi-échelles*. phdthesis, Université de Versailles-Saint Quentin en Yvelines, décembre 1997. URL https://theses.hal.science/tel-00085113.

Thomas Devogele, Christine Parent et Stefano Spaccapietra : On spatial database integration. *International Journal of Geographical Information Science*, 12(4):335–352, juin 1998. ISSN 1365-8816. URL https://doi.org/10.1080/136588198241824.

Deeptendu Bikash Dhar et Bhabatosh Chanda : Extraction and recognition of geographical features from paper maps. *International Journal of Document Analysis and Recognition (IJDAR)*, 8(4):232–245, septembre 2006. ISSN 1433-2833, 1433-2825. URL http://link.springer.com/10.1007/s10032-005-0010-9.

E. W. Dijkstra : A note on two problems in connexion with graphs. *Numerische Mathematik*, 1(1):269–271, décembre 1959. ISSN 0945-3245. URL https://doi.org/10.1007/BF01386390.

Arta Dilo, Rolf A. De By et Alfred Stein : A system of types and operators for handling vague spatial objects. *International Journal of Geographical Information Science*, 21 (4):397–426, avril 2007. ISSN 1365-8816, 1362-3087. URL http://www.tandfonline.com/doi/abs/10.1080/13658810601037096.

Zepeng Ding, Wenhao Huang, Jiaqing Liang, Deqing Yang et Yanghua Xiao : Improving Recall of Large Language Models : A Model Collaboration Approach for Relational Triple Extraction, avril 2024. URL http://arxiv.org/abs/2404.09593. Issue : arXiv :2404.09593 arXiv : 2404.09593 [cs].

AnHai Doan, Alon Halevy et Zachary Ives : *Principles of Data Integration*. Elsevier, juin 2012. ISBN 978-0-12-391479-8.

Filipe Drapeau Contim : *Qu'est-ce que l'identité ?*, volume 5 de *Chemins philosophiques*. Vrin, Paris, 2010. ISBN 978-2-7116-2256-6.

Gift Dumedah : Address points of landmarks and paratransit services as a credible reference database for geocoding. *Transactions in GIS*, 25(2):1027–1048, 2021. ISSN 1467-9671. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/tgis.12716.

Bertrand Dumenieu : *Un système d'information géographique pour le suivi d'objets historiques urbains à travers l'espace et le temps*. phdthesis, École des Hautes Études en Sciences Sociales, Paris, décembre 2015. URL https://theses.hal.science/tel-03262391.

Jean-Luc Dupouey, Jean Bachacou, Régine Cosserat-Mangeot, Serge Aberdam, Daniel Vallauri, Gérard Chappart et Marie-Anne Corvisier de Villèle : Vers la réalisation d'une carte géoréférencée des forêts anciennes de France. *Le Monde des Cartes*, 191:85, 2007. URL https://hal.inrae.fr/hal-02664500.

Soumitra Dutta : Approximate spatial reasoning : Integrating qualitative and quantitative constraints. *International Journal of Approximate Reasoning*, 5(3):307–330, mai 1991. ISSN 0888-613X. URL https://www.sciencedirect.com/science/article/pii/0888613X9190015E.

Duverneuil et Jacques de La Tynna : *Almanach du commerce de Paris, pour l'an IX de la République française*. Bureau de l'Almanach, Paris, 1801. URL https://gallica.bnf.fr/ark:/12148/bpt6k1175057p.

Jérôme Euzenat et Pavel Shvaiko : *Ontology Matching*. Springer, Berlin, Heidelberg, 2013. ISBN 978-3-642-38720-3 978-3-642-38721-0. URL https://link.springer.com/10.1007/978-3-642-38721-0.

Expert Group on Archival Description (EGAD) : Records in Contexts Ontology (RiC-O), version 1.1. Rapport technique, International Council on Archives, mai 2025. URL https://www.ica.org/standards/RiC/ontology.

Abdelfettah Feliachi et Nathalie Abadie : Évaluation de la précision planimétrique des données géoréférencées sur le Web de données. *In Atelier Qualité des Données du Web (QLOD), conférence Extraction et Gestion des Connaissances (EGC)*, Grenoble, France, 2017. URL https://hal.science/hal-02399178.

Abdelfettah Feliachi, Nathalie Abadie et Fayçal Hamdi : An Adaptive Approach for Interlinking Georeferenced Data. *In Proceedings of the Knowledge Capture Conference, K-CAP 2017*, pages 1–8, Austin, United States, décembre 2017. ACM Press. URL https://hal.science/hal-02388078.

Ambroise Firmin-Didot et Hyacinthe Firmin-Didot : *Annuaire-almanach du commerce, de l'industrie, de la magistrature et de l'administration : ou almanach des 500.000 adresses de Paris, des départements et des pays étrangers : Firmin Didot et Bottin réunis*. Firmin-Didot frères, Paris, 1894. URL ark:/12148/bpt6k9732740w.

Peter Fisher, Alexis Comber et Richard Wadsworth : Nature de l'incertitude pour les données spatiales. *In* Rodolphe Devillers et Robert Jeansoulin, éditeurs : *Qualité de l'information géographique*, pages 49–64. Hermes, 2005.

Peter F. Fisher : Models of uncertainty in spatial data. *Geographical information systems*, 1:191–205, 1999. URL https://www.geos.ed.ac.uk/~gisteac/gis_book_abridged/files/ch13.pdf.

Jacques Fize, Mathieu Roche et Maguelonne Teisseire : Spatial Textual Representation (STR) ou comment représenter la spatialité des donnnées textuelles. *In* Université de Rouen Normandie, éditeur : *Actes de Spatial Analysis and GEOmatics 2017*, page 14, novembre 2017. URL https://hal.science/hal-01643368.

Trine Fjordback Søndergaard, Jack Andersen et Birger Hjørland : Documents and the communication of scientific and scholarly information : Revising and updating the UNISIST model. *Journal of Documentation*, 59(3):278–320, juin 2003. ISSN 0022-0418. URL https://doi.org/10.1108/00220410310472509.

M. Fleury : *Concordance entre la numérotation sectionnaire et la numérotation de type actuel de 2466 maisons de Paris*. Commission du Vieux, Rotonde de La Villette, Institut d'histoire de Paris, 1995. ISBN 978-2-85738-000-9. URL https://books.google.fr/books?id=Oq_NAQAACAAJ.

Andrew U Frank : Qualitative spatial reasoning about distances and directions in geographic space. *Journal of Visual Languages & Computing*, 3(4):343–371, décembre 1992. ISSN 1045926X. URL https://linkinghub.elsevier.com/retrieve/pii/1045926X92900079.

Victor Gay, Sébastien Poublanc et Jean-Luc Emmanuel Demonsant : Plan de gestion de données du projet ANR ObARDI. Technical Report, MSHS Toulouse ; FRAMESPA ; TSE - Toulouse School of Economics ; IAST - Institute for Advanced Study in Toulouse, juin 2021. URL https://hal.science/hal-03461923.

Lise Getoor et Ashwin Machanavajjhala : Entity resolution : theory, practice & open challenges. *Proc. VLDB Endow.*, 5(12):2018–2019, août 2012. ISSN 2150-8097. URL https://dl.acm.org/doi/10.14778/2367502.2367564.

Sylvain Giroux et Ginette Tremblay : *Méthodologie des sciences humaines : la recherche en action*, volume 324. Éd. du Renouveau pédagogique ERPI, 2009.

Daniel W Goldberg, John P Wilson et Craig A Knoblock : From Text to Geographic Coordinates : The Current State of Geocoding. *URISA journal*, 19(1):33–46, 2007.

Michael F Goodchild et Anne Kelly Knowles : Combining space and time : New potential for temporal GIS. *Placing history : How maps, spatial data, and GIS are changing historical scholarship*, pages 179–198, 2008. Edition : ESRI Press Place : Redlands, CA.

Mike Graves, Adam Constabaris et Dan Brickley : FOAF : Connecting People on the Semantic Web. *In Knitting the Semantic Web*, pages 191–202. Routledge, 2007.

Antonio Grieco, Francesco Nucci et Alfredo Anglani : Representation of fuzzy time variables in discrete event simulation. *Integrated Computer-Aided Engineering*, 10(4): 305–318, septembre 2003. ISSN 18758835, 10692509. URL https://journals.sagepub.com/doi/full/10.3233/ICA-2003-10402.

Ralf Hartmut Güting : An introduction to spatial database systems. *The VLDB Journal*, 3(4):357–399, octobre 1994. ISSN 0949-877X. URL https://doi.org/10.1007/BF01231602.

Udo Hahn, Katrin Tomanek, Elena Beisswanger et Erik Faessler : A Proposal for a Configurable Silver Standard. *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 235–242, juillet 2010. URL https://aclanthology.org/W10-1838/.

Mordechai Haklay : How Good is Volunteered Geographical Information ? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets. *Environment and Planning B : Planning and Design*, 37(4):682–703, août 2010. ISSN 0265-8135, 1472-3417. URL https://journals.sagepub.com/doi/10.1068/b35097.

Pierre Hallot : *L'identité à travers l'espace et le temps. Vers une définition de l'identité et des relations spatio-temporelles entre objets géographiques.* phdthesis, ULiège - Université de Liège, mars 2012.

Harry Halpin, Patrick J. Hayes, Jamie P. McCusker, Deborah L. McGuinness et Henry S. Thompson : When owl :sameAs Isn't the Same : An Analysis of Identity in Linked Data. *In* Peter F. Patel-Schneider, Yue Pan, Pascal Hitzler, Peter Mika, Lei Zhang, Jeff Z. Pan, Ian Horrocks et Birte Glimm, éditeurs : *The Semantic Web – ISWC 2010*, pages 305–320, Berlin, Heidelberg, 2010. Springer. ISBN 978-3-642-17746-0.

Benjamin Harbelot, Helbert Arenas et Christophe Cruz : LC3 : A spatio-temporal and semantic model for knowledge discovery from geospatial datasets. *Journal of Web Semantics*, 35:3–24, décembre 2015. ISSN 1570-8268. URL https://www.sciencedirect.com/science/article/pii/S1570826815000840.

Olaf Hartig : Provenance Information in the Web of Data. *LDOW*, 538, 2009. URL https://ceur-ws.org/Vol-538/ldow2009_paper18.pdf.

Nicholas William Joseph Hazelton : *Integrating time, dynamic modelling and geographical information systems : Development of four-dimensional GIS.* phdthesis, University of Melbourne, 1991.

Nicholas William Joseph Hazelton, Francis J. Leahy et Williamson : Integrating dynamic modeling with geographic information systems. *Journal of the Urban and Regional Information Systems Association*, pages 47–58, 1992. Edition : 4.

Jacques Hillairet : *Dictionnaire historique des rues de Paris.* Les Éditions de Minuit, Paris, 1963. ISBN 978-2-7073-1054-5.

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia D'Amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab et Antoine Zimmermann : Knowledge Graphs. *ACM Comput. Surv.*, 54(4):71 :1–71 :37, juillet 2021. ISSN 0360-0300. URL https://dl.acm.org/doi/10.1145/3447772.

Kathleen Hornsby et Max J. Egenhofer : Qualitative representation of change. *In* Stephen C. Hirtle et Andrew U. Frank, éditeurs : *Spatial Information Theory A Theoretical Basis for GIS*, pages 15–33, Berlin, Heidelberg, 1997. Springer. ISBN 978-3-540-69616-2.

Kathleen Hornsby et Max J. Egenhofer : Identity-based change : a foundation for spatio-temporal knowledge representation. *International Journal of Geographical Information Science*, 14(3):207–224, avril 2000. ISSN 1365-8816, 1362-3087. URL http://www.tandfonline.com/doi/abs/10.1080/136588100240813.

Gary J. Hunter et Ian P. Williamson : The development of a historical digital cadastral database†. *International journal of geographical information systems*, 4(2):169–179, avril 1990. ISSN 0269-3798. URL http://www.tandfonline.com/doi/abs/10.1080/02693799008941538.

International Organization for Standardization : ISO 19133 :2005 - Information géographique — Services basés sur la localisation — Suivi et navigation. Rapport technique ISO 19133 :2005, Organisation internationale de normalisation (ISO), 2005. URL https://www.iso.org/fr/standard/32551.html.

International Organization for Standardization : ISO - ISO 8601 — Représentation de la date et de l'heure, février 2017. URL https://www.iso.org/fr/iso-8601-date-and-time-format.html.

International Organization for Standardization : ISO 19109 :2025 Information géographique — Modèle général des entités et règles relatives au schéma d'application. Rapport technique ISO 19109 :2025, Organisation internationale de normalisation (ISO), juillet 2025. URL https://www.iso.org/fr/standard/84700.html.

Théodore Jacoubet : Atlas général de la Ville, des faubourgs et des monuments de Paris : 1836 / par Th. Jacoubet, architecte... ; gravé par Bonnet ; dédié et présenté à M. le Comte de Chabrol de Volvic, Conseiller d'Etat, Préfet du Département de la Seine ; écrit par Hacq. graveur du Dépot de la Guerre, 1836. URL https://bibliotheques-specialisees.paris.fr/ark:/73873/pf0000858369.

Mustafa Jarrar et Robert Meersman : Ontology Engineering – The DOGMA Approach. *Advances in Web Semantics I*, pages 7–34, 2009. ISSN 0302-9743, 1611-3349. URL http://link.springer.com/10.1007/978-3-540-89784-2_2. Edition : Springer Place : Berlin, Heidelberg.

B. Jiang et C. Claramunt : A Structural Approach to the Model Generalization of an Urban Street Network*. *GeoInformatica*, 8(2):157–171, juin 2004. ISSN 1573-7624. URL https://doi.org/10.1023/B:GEIN.0000017746.44824.70.

Bin Jiang, Christophe Claramunt et Björn Klarqvist : Integration of space syntax into GIS for modelling urban spaces. *International Journal of Applied Earth Observation and Geoinformation*, 2(3):161–171, janvier 2000. ISSN 1569-8432. URL https://www.sciencedirect.com/science/article/pii/S0303243400850102.

Philip Kahle, Sebastian Colutto, Gunter Hackl et Gunter Muhlberger : Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents. *In 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, pages 19–24, Kyoto, novembre 2017. IEEE. ISBN 978-1-5386-3586-5. URL http://ieeexplore.ieee.org/document/8270253/.

David R. Karger, Sewoong Oh et Devavrat Shah : Iterative learning for reliable crowdsourcing systems. *In Proceedings of the 25th International Conference on Neural Information Processing Systems*, NIPS'11, pages 1953–1961, Red Hook, NY, USA, décembre 2011. Curran Associates Inc. ISBN 978-1-61839-599-3. URL https://dl.acm.org/doi/10.5555/2986459.2986677.

Tomi Kauppinen, Riikka Henriksson, Reetta Sinkkila, Robin Lindroos, Jari Vaatainen et Eero Hyvonen : Ontology-based Disambiguation of Spatiotemporal Locations. *IRSW*, 2008.

Tomi Kauppinen et Eero Hyvönen : Bridging the Semantic Gap between Ontology Versions. *Web Intelligence Symposium, Finnish AI Conference*, 20:2–3, 2004. URL https://seco.cs.aalto.fi/events/2004/2004-09-02-web-intelligence/papers/changebridges-wis04.pdf.

Carsten Kessler, Johannes Trame et Tomi Kauppinen : Tracking editing processes in volunteered geographic information : The case of OpenStreetMap. *In Identifying objects, processes and events in spatio-temporally distributed data (IOPE), workshop at conference on spatial information theory*, volume 12, pages 6–8, 2011.

George J. Klir et Bo Yuan : *Fuzzy sets and fuzzy logic : theory and applications*. Prentice Hall, Upper Saddle River, 1995. ISBN 978-0-13-101171-7.

Holger Knublauch et Dimitris Kontokostas : Shapes Constraint Language (SHACL), juillet 2017. URL https://www.w3.org/TR/shacl/.

Konstantinos Kotis et George A. Vouros : Human-centered ontology engineering : The HCOME methodology. *Knowledge and Information Systems*, 10(1):109–131, juillet 2006. ISSN 0219-3116. URL https://doi.org/10.1007/s10115-005-0227-4.

Marin Kreenfelt et François-Claude Mangeot, éditeurs. *Almanach de Paris, contenant la demeure, les noms & qualités des personnes de condition dans la ville & fauxbourgs de Paris*. Lesclapart, Paris, 1789. URL https://gallica.bnf.fr/ark:/12148/bpt6k1265821s.

Kostis Kyzirakos, Ioannis Vlachopoulos, Dimitrianos Savva, Stefan Manegold et Manolis Koubarakis : GeoTriples : a Tool for Publishing Geospatial Data as RDF Graphs Using R2RML Mappings. *In Terra Cognita 2014*, pages 33–44, Riva del Garda, Trentino, Italy, octobre 2014. URL https://cgi.di.uoa.gr/~koubarak/publications/2014/terra2014_3.pdf.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami et Chris Dyer : Neural Architectures for Named Entity Recognition. *In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 260–270, San Diego, California, 2016. Association for Computational Linguistics. URL http://aclweb.org/anthology/N16-1030.

Gail Langran : *Time in Geographic Information Systems*. CRC Press, 1 édition, 1992. ISBN 978-1-003-06259-2. URL https://www.taylorfrancis.com/books/9781000123968.

Gail Langran et Nicholas R Chrisman : A Framework For Temporal Geographic Information. *Cartographica*, 25(3):1–14, décembre 1988. ISSN 0317-7173, 1911-9925. URL https://utppublishing.com/doi/10.3138/K877-7273-2238-5Q6V.

Éditions Larousse : Définitions du mot "source", 2025. URL https://www.larousse.fr/dictionnaires/francais/source/73725.

Aurélie Leborgne, Adrien Meyer, Henri Giraud, Florence Le Ber et Stella Marc-Zwecker : Un graphe spatio-temporel pour modéliser l'évolution de parcelles agricoles. *In SAGEO*, Clermont-Ferrand, France, novembre 2019. URL https://hal.science/hal-02375094.

Prosper Lemoine : *Almanach des marchands et fabricans en cuirs, de la ville de Paris*. Lemoine, Paris, 1813 édition, 1813. URL https://gallica.bnf.fr/ark:/12148/bpt6k6411555f.

Maurizio Lenzerini : Data integration : a theoretical perspective. *In Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, PODS '02, pages 233–246, New York, NY, USA, juin 2002. Association for Computing Machinery. ISBN 978-1-58113-507-7. URL https://doi.org/10.1145/543613.543644.

Stephen C. Levinson : Frames of reference and Molyneux's question : Crosslinguistic evidence. *In Language and space*, Language, speech, and communication, pages 109–169. The MIT Press, Cambridge, MA, US, 1996. ISBN 978-0-262-02403-7.

Linna Li et Michael F. Goodchild : An optimisation model for linear feature matching in geographical data conflation. *International Journal of Image and Data Fusion*, 2 (4):309–328, décembre 2011. ISSN 1947-9832. URL https://doi.org/10.1080/19479832.2011.577458.

Qi Li, Yaliang Li, Jing Gao, Bo Zhao, Wei Fan et Jiawei Han : Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. *In Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, SIGMOD '14, pages 1187–1198, New York, NY, USA, juin 2014. Association for Computing Machinery. ISBN 978-1-4503-2376-5. URL https://dl.acm.org/doi/10.1145/2588555.2610509.

Em. Longeard : Nettoyons. *L'Aurore : contre tous les abus et toutes les iniquités*, 6:4, avril 1922. URL http://catalogue.bnf.fr/ark:/12148/cb32706831s.

Edmond Léry : Les anciens numérotages de Paris et de Versailles. *In* L. Bernard et H. Champion, éditeurs : *Revue de l'histoire de Versailles et de Seine-et-Oise*, volume 19, pages 185–198. Imprimerie J. Aubert et Cie, Versailles, 1917. ISBN 1158-2677. URL https://gallica.bnf.fr/ark:/12148/bpt6k67116j.

Hana Mallek, Faiza Ghozzi et Faiez Gargouri : Towards Extract-Transform-Load Operations in a Big Data context. *International Journal of Sociotechnology and Knowledge Development (IJSKD)*, 12(2):77–95, avril 2020. ISSN 1941-6253. URL https://www.igi-global.com/gateway/article/249233.

Éric Mermet et Anne Varet-Vitu : Paris et son réseau de rues pour une spatialisation des adresses historiques, fin XVIIIe siècle-première moitié du XIXe siècle. *Actes de la conférence francophone ESRI*, 2015. URL https://psigehess.hypotheses.org/paris-et-son-reseau-de-rues-pour-une-spatialisation-des-adresses-historiques-fin-xviiie-siecle-premiere-moitie-du-xixe-siecle.

Nandana Mihindukulasooriya, Sanju Tiwari, Carlos F. Enguix et Kusum Lata : Text2KGBench : A Benchmark for Ontology-Driven Knowledge Graph Generation from

Text, août 2023. URL http://arxiv.org/abs/2308.02357. Issue : arXiv :2308.02357 arXiv : 2308.02357 [cs].

Alistair Miles, Brian Matthews, Michael Wilson et Dan Brickley : SKOS Core : Simple knowledge organisation for the Web. *In Proceedings of the International Conference on Dublin Core and Metadata Applications*. Dublin Core Metadata Initiative, septembre 2005. URL https://dcpapers.dublincore.org/article/952107985.

Marco Minghini et Francesco Frassinelli : OpenStreetMap history for intrinsic quality assessment : Is OSM up-to-date ? *Open Geospatial Data, Software and Standards*, 4(1):9, septembre 2019. ISSN 2363-7501. URL https://doi.org/10.1186/s40965-019-0067-x.

Ludovic Moncla : *Automatic reconstruction of itineraries from descriptive texts*. phdthesis, Université de Pau et des Pays de l'Adour ; Universidad de Zaragoza, décembre 2015. URL https://theses.hal.science/tel-01249999.

Luc Moreau et Paul Groth : The prov Ontology. *In* Luc Moreau et Paul Groth, éditeurs : *Provenance : An Introduction to PROV*, pages 21–38. Springer International Publishing, Cham, 2013. ISBN 978-3-031-79450-6. URL https://doi.org/10.1007/978-3-031-79450-6_3.

Takashi Morita : Cheminement et adresse. *Les Annales de la Recherche Urbaine*, 7(1): 27–61, 1980. URL https://www.persee.fr/doc/aru_0180-930x_1980_num_7_1_992.

Mark A. Musen : The Protégé Project : A Look Back and a Look Forward. *AI matters*, 1(4):4–12, juin 2015. ISSN 2372-3483. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4883684/.

Sébastien Mustière : *Intégration de données géographiques*. phdthesis, Université Paris-Est, 2014.

Elisabeth Métais, Fatma Ghorbel, Fayçal Hamdi, Nebrasse Ellouze, Noura Herradi et Assia Soukane : Representing Imprecise Time Intervals in OWL 2. *Enterprise Modelling and Information Systems Architectures*, pages 120–132 Pages, février 2018. URL https://www.emisa-journal.org/emisa/article/view/191.

Felix Nauman et Melanie Herschel : *An Introduction to Duplicate Detection*. Synthesis Lectures on Data Management. Springer International Publishing, Cham, 1st ed. 2010 édition, 2010. ISBN 978-3-031-01835-0.

Hélène Noizet : Les plans d'îlots Vasserot, support d'un système de l'information géographique historique de Paris. *EAV (Ecole nationale supérieure d'architecture de Versailles)*, 14:86–95, 2009. URL https://shs.hal.science/halshs-00419684.

Natalya F. Noy et Michel Klein : Ontology Evolution : Not the Same as Schema Evolution. *Knowledge and Information Systems*, 6(4):428–440, juillet 2004. ISSN 0219-3116. URL https://doi.org/10.1007/s10115-003-0137-2.

Natalya F Noy et Deborah L McGuinness : Ontology Development 101 : A Guide to Creating Your First Ontology. *Knowledge Systems Laboratory*, 32, janvier 2001.

H.J. Ohlbach : Relations between fuzzy time intervals. *In Proceedings. 11th International Symposium on Temporal Representation and Reasoning, 2004. TIME 2004.*, pages 44–51, juillet 2004. URL https://ieeexplore.ieee.org/abstract/document/1314418.

J. O'Keefe : The Spatial Prepositions in English, Vector Grammar and the Cognitive Map Theory. *In* P. Bloom, M. A. Peterson, L. Nadel et M. F. Garrett, éditeurs : *In : Bloom, P and Peterson, MA and Nadel, L and Garrett, MF, (eds.) Language and Space. (pp. 277-316). The MIT Press : Cambridge, MA, USA. (1996)*, pages 277–316. The MIT Press, Cambridge, MA, USA, 1996. URL https://mitpress.mit.edu/books/language-and-space.

Ana-Maria Olteanu : *Fusion de connaissances imparfaites pour l'appariement de données géographiques : proposition d'une approche s'appuyant sur la théorie des fonctions de croyance.* phdthesis, Université Paris-Est, octobre 2008. URL https://theses.hal.science/tel-00469407.

Open Geospatial Consortium : OpenGIS® Implementation Standard for Geographic information - Simple feature access - Part 1 : Common architecture. OGC Standard, Open Geospatial Consortium (OGC), mai 2011. URL http://www.opengis.net/doc/is/sfa/1.2.1.

Krzysztof Ostafin, Dominik Kaim, Tadeusz Siwek et Anna Miklar : Historical dataset of administrative units with social-economic attributes for Austrian Silesia 1837–1910. *Scientific Data*, 7(1):208, juin 2020. ISSN 2052-4463. URL https://www.nature.com/articles/s41597-020-0546-z.

George Papadakis, Ekaterini Ioannou, Emanouil Thanos et Themis Palpanas : Entity Resolution : Past, Present, and Yet-to-Come. *In The Four Generations of Entity Resolution*, pages 1–3. Springer International Publishing, Cham, 2021. ISBN 978-3-031-00750-7 978-3-031-01878-7. URL https://link.springer.com/10.1007/978-3-031-01878-7_1.

Michele Pasin et John Bradley : Factoid-based prosopography and computer ontologies : towards an integrated approach. *Digital Scholarship in the Humanities*, 30 (1):86–97, avril 2015. ISSN 2055-7671. URL https://doi.org/10.1093/llc/fqt037.

George Percivall : OGC Reference Model. Rapport technique, Open Geospatial Consortium Inc., septembre 2003.

Bernardo Pereira Nunes, Alexander Mera, Marco Antônio Casanova, Besnik Fetahu, Luiz André P. Paes Leme et Stefan Dietze : Complex Matching of RDF Datatype Properties. *In* Hendrik Decker, Lenka Lhotská, Sebastian Link, Josef Basl et A Min Tjoa, éditeurs : *Database and Expert Systems Applications*, volume 8055, pages 195–208. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. ISBN 978-3-642-40284-5 978-3-642-40285-2. URL http://link.springer.com/10.1007/978-3-642-40285-2_18.

Silvio Peroni : A Simplified Agile Methodology for Ontology Development. *In* Mauro Dragoni, María Poveda-Villalón et Ernesto Jimenez-Ruiz, éditeurs : *OWL : Experiences and Directions – Reasoner Evaluation*, Lecture Notes in Computer Science, pages 55–69, Cham, 2017. Springer International Publishing. ISBN 978-3-319-54627-8.

Donna J. Peuquet et Niu Duan : An event-based spatiotemporal data model (ESTDM) for temporal analysis of geographical data. *International journal of geographical information systems*, 9(1):7–24, janvier 1995. ISSN 0269-3798. URL https://doi.org/10.1080/02693799508902022.

Donna J. Peuquet et Elizabeth Wentz : An approach for time-based analysis of spatiotemporal data. *In* T.C. Waugh et R.G. Healey, éditeurs : *Advances in GIS research. Proc. 6th symposium.*, volume 1, pages 489–504. Taylor & Francis, London, Edinburgh, 1994. URL https://www.scopus.com/pages/publications/0028564506.

Pierre Pinon, Bertrand Le Boudec et Dominique Carré : *Les plans de Paris : histoire d'une capitale*. Le Passage Atelier parisien d'urbanisme Bibliothèque nationale de France, Paris, 2014. ISBN 978-2-84742-243-6.

Brandon Plewe : The Nature of Uncertainty in Historical Geographic Information. *Transactions in GIS*, 6(4):431–456, 2002. ISSN 1467-9671. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-9671.00121.

Christine Plumejeaud, Hélène Mathian, Jérôme Gensel et Claude Grasland : Spatio-temporal analysis of territorial changes from a multi-scale perspective. *International Journal of Geographical Information Science*, 25(10):1597–1612, octobre 2011. ISSN 1365-8816. URL https://doi.org/10.1080/13658816.2010.534658.

Nicolas Poirier : Du point à l'espace (rural) : localisation de mentions textuelles et mise à l'épreuve de normes socio-spatiales. *Bulletin du centre d'études médiévales d'Auxerre \textbar BUCEMA*, Hors-série n° 9, février 2016. ISSN 1623-5770. URL https://journals.openedition.org/cem/13819.

María Poveda-Villalón, Asunción Gómez-Pérez et Mari Carmen Suárez-Figueroa : OOPS! (OntOlogy Pitfall Scanner!) : An On-line Tool for Ontology Evaluation. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 10(2):7–34, 2014.

Jeanne Pronteau : *Les Numérotages des maisons de Paris, du XVe siècle à nos jours ...* Commission des travaux historiques. Préfecture de la Seine, Service des travaux historiques, 1966.

E. M. Prétot : *Annuaire de la typographie parisienne et départementale*, volume 1. Krabbe, Paris, 1844. URL https://gallica.bnf.fr/ark:/12148/bpt6k63680039.

Erhard Rahm et Philip A. Bernstein : A survey of approaches to automatic schema matching. *The VLDB Journal*, 10(4):334–350, décembre 2001. ISSN 0949-877X. URL https://doi.org/10.1007/s007780100057.

David A. Randell, Zhan Cui et Anthony G. Cohn : A spatial logic based on regions and connection. *In KR'92 : Proceedings of the Third International Conference on Principles of Knowledge Representation and Reasoning*, KR'92, pages 165–176. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1992. ISBN 1-55860-262-3.

Helen Mair Rawsthorne : *Creation of geospatial knowledge graphs from heterogeneous sources*. phdthesis, Université Gustave Eiffel, janvier 2024. URL https://theses.hal.science/tel-04599846.

Helen Mair Rawsthorne, Nathalie Abadie, Eric Kergosien, Cécile Duchêne et Éric Saux : Extraction automatique d'entités spatiales imbriquées et de relations spatiales à partir de texte pour la création de graphes de connaissances : Une approche et deux jeux de données. *In TextMine '24 : Atelier sur la Fouille de Textes*, pages 75–86, Dijon, France, janvier 2024. URL https://hal.science/hal-04444358.

Gilles Sahut : « Citez vos sources » : archéologie d'une règle au cœur du savoir wikipédien (2002-2008). *Études de communication. langages, information, médiations*, pages 97–110, juin 2014. ISSN 1270-6841. URL https://journals.openedition.org/edc/5721.

Ashok Samal, Sharad Seth et Kevin Cueto1 : A feature-based approach to conflation of geospatial sources. *International Journal of Geographical Information Science*, 18 (5):459–489, juillet 2004. ISSN 1365-8816. URL https://doi.org/10.1080/13658810410001658076.

Ettore Sarzotti, Nicola Gabellieri, Paolo Zatelli et Angelo Besana : From Paper to GIS : An Example of Historical Maps Vectorization. *e-Perimetron*, 20(1):11, 2025. URL https://iris.unitn.it/handle/11572/464130.

Fatiha Saïs : *Knowledge Graph Refinement : Link Detection, Link Invalidation, Key Discovery and Data Enrichment*. Habilitation à diriger des recherches, Université Paris Sud, juin 2019. URL https://hal.science/tel-02441156.

Markus Schneider : Uncertainty Management for Spatial Datain Databases : Fuzzy Spatial Data Types. *In* Ralf Hartmut Güting, Dimitris Papadias et Fred Lochovsky, éditeurs : *Advances in Spatial Databases*, pages 330–351, Berlin, Heidelberg, 1999. Springer. ISBN 978-3-540-48482-0.

Wei Liang Seow, Iti Chaturvedi, Amber Hogarth, Rui Mao et Erik Cambria : A review of named entity recognition : from learning methods to modelling paradigms and tasks. *Artificial Intelligence Review*, 58(10):315, juillet 2025. ISSN 1573-7462. URL https://doi.org/10.1007/s10462-025-11321-8.

Service du plan de Paris et Service des eaux et de l'assainissement : Atlas administratif des 20 arrondissements de la ville de Paris publié d'après les ordres de M. le baron G. E. Haussmann, sénateur préfet de la Seine. 1870. Ce plan est une reproduction du plan général à l'échelle de 1/5000 dressé par les géomètres du Service municipal du plan de Paris, 1870. URL https://bibliotheques-specialisees.paris.fr/ark:/73873/pf0000935043.

Nihar B. Shah et Dengyong Zhou : No oops, you won't do it again : mechanisms for self-correction in crowdsourcing. *In Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pages 1–10, New York, NY, USA, juin 2016. JMLR.org. URL https://dl.acm.org/doi/10.5555/3045390.3045392.

Longxiang Shi, Shijian Li, Xiaoran Yang, Jiaheng Qi, Gang Pan et Binbin Zhou : Semantic Health Knowledge Graph : Semantic Integration of Heterogeneous Medical Knowledge and Services. *BioMed Research International*, 2017(1):2858423, 2017. ISSN 2314-6141. URL https://onlinelibrary.wiley.com/doi/abs/10.1155/2017/2858423.

Cogan Shimizu, Karl Hammar et Pascal Hitzler : Modular ontology modeling. *Semantic Web*, 14(3):459–489, avril 2023. ISSN 1570-0844. URL https://journals.sagepub.com/action/showAbstract.

Barry Smith : On drawing lines on a map. *In* Andrew U. Frank et Werner Kuhn, éditeurs : *Spatial Information Theory A Theoretical Basis for GIS*, pages 475–484, Berlin, Heidelberg, 1995. Springer. ISBN 978-3-540-45519-6.

Barry Smith et Achille C. Varzi : Fiat and Bona Fide Boundaries. *Philosophy and Phenomenological Research*, 60(2):401–420, 2000. ISSN 0031-8205. URL https://www.jstor.org/stable/2653492.

R. Smith : An Overview of the Tesseract OCR Engine. *In Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pages 629–633, septembre 2007. URL https://ieeexplore.ieee.org/abstract/document/4376991.

Laurent Spéry, Christophe Claramunt et Thérèse Libourel : A Spatio-Temporal Model for the Manipulation of Lineage Metadata. *GeoInformatica*, 5(1):51–70, mars 2001. ISSN 1573-7624. URL https://doi.org/10.1023/A:1011459921552.

Kostas Stefanidis, Ioannis Chrysakis et Giorgos Flouris : On Designing Archiving Policies for Evolving RDF Datasets on the Web. *In* Eric Yu, Gillian Dobbie, Matthias Jarke et Sandeep Purao, éditeurs : *Conceptual Modeling*, volume 8824, pages 43–56. Springer International Publishing, Cham, 2014. ISBN 978-3-319-12205-2 978-3-319-12206-9. URL http://link.springer.com/10.1007/978-3-319-12206-9_4.

York Sure : A Tool-Supported Methodology for Ontology-Based Knowledge Management. *In The Ontology and Modelling of Real Estate Transactions*. Routledge, 2003. ISBN 978-1-315-23797-8.

Mehtab Alam Syed, Elena Arsevska, Mathieu Roche et Maguelonne Teisseire : GeospatRE : extraction and geocoding of spatial relation entities in textual documents. *Cartography and Geographic Information Science*, 52(3):221–236, mai 2025. ISSN 1523-0406, 1545-0465. URL https://www.tandfonline.com/doi/full/10.1080/15230406.2023.2264753.

Sabiha Tahrat, Eric Kergosien, Sandra Bringay, Mathieu Roche et Maguelonne Teisseire : Text2Geo : des données textuelles aux informations géospatiales. *In Actes de la conférence EGC 2023*, volume RNTI-E-24, pages 407–412. RNTI, janvier 2013. URL https://hal-lirmm.ccsd.cnrs.fr/lirmm-00816294.

Temporary MIWP 2021–2024 sub-group 2.3.1 : D2.8.I.5 INSPIRE Data Specification on Addresses – Technical Guidelines. Rapport technique, INSPIRE Maintenance and Implementation Group (MIG), juillet 2024. URL https://knowledge-base.inspire.ec.europa.eu/publications/inspire-data-specification-addresses-technical-guidelines_en.

Thora Tenbrink et Werner Kuhn : A Model of Spatial Reference Frames in Language. *In* Max Egenhofer, Nicholas Giudice, Reinhard Moratz et Michael Worboys, éditeurs : *Spatial Information Theory*, volume 6899, pages 371–390. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. ISBN 978-3-642-23195-7 978-3-642-23196-4. URL http://link.springer.com/10.1007/978-3-642-23196-4_20.

Melissa Terras : Opening Access to collections : the making and using of open digitised cultural content. *Online Information Review*, 39(5):733–752, septembre 2015. ISSN 1468-4527. URL https://doi.org/10.1108/OIR-06-2015-0193.

Laurent-Étienne Testu : *Almanach national de France, année commune M. DCC. XCIII. L'An IIe de la République.* Testu, Paris, 1793. URL https://gallica.bnf.fr/ark:/12148/bpt6k1041389c.

Hui Tian, Wenwen Sheng, Hong Shen et Can Wang : Truth finding by reliability estimation on inconsistent entities for heterogeneous data sets. *Knowledge-Based Systems*, 187:104828, janvier 2020. ISSN 0950-7051. URL https://www.sciencedirect.com/science/article/pii/S095070511930303X.

Guillaume Touya, Sandrine Balley, Cécile Duchêne, Kusay Jaara, Nicolas Regnauld et Nicholas Gould : Towards an Ontology of Spatial Relations and Relational Constraints. *In 15th ICA workshop on generalisation and multiple representation*, Istanbul, Turkey, septembre 2012. ICA. URL https://hal.science/hal-02290906.

Solenn Tual, Nathalie Abadie, Bertrand Duménieu, Joseph Chazalon et Edwin Carlinet : Création d'un graphe de connaissances géohistorique à partir d'annuaires du commerce parisien du 19 ème siècle : application aux métiers de la photographie. *In Actes de la 34es Journées francophones d'Ingénierie des Connaissances (IC 2023)*, IC2023 : 34es Journées francophones d'Ingénierie des Connaissances, Strasbourg, France, juillet 2023. URL https://hal.science/hal-04121643.

Université du Québec à Montréal : Catégories des sources d'information, 2025. URL https://infosphere.uqam.ca/preparer-sa-recherche/identifier-type-travail/categories-sources-dinformation/.

Philibert Vasserot : *Grand plan cadastral de la ville de Paris, par quartiers et arrondissements, dressé de 1807 à 1825.* Philibert Vasserot, Paris, 1825.

Edme Verniquet : *Atlas du plan général de la ville de Paris, levé géométriquement par le cen Verniquet rapporté sur une échelle d'une demie ligne pour toise, divisée en 72 planches, compris les cartouches et plan des opérations trigonométriques. Dessiné et gravé par les cens Bartholomé et Mathieu...* Edme Verniquet, Paris, 1795. URL https://gallica.bnf.fr/ark:/12148/bpt6k3167995.

Ville de Paris : Agrément donnée pour la dénomination « villa des Enfants Terribles » de la voie privée identifiée par l'indicatif Z/20, commençant 157 rue Pelleport et finissant en impasse à Paris (20e), août 2025. URL https://bovp.apps.paris.fr/doc_num_data.php?explnum_id=42285.

Alexander von Lünen : Tracking in a New Territory : Re-imaging GIS for History. *In* Alexander von Lünen et Charles Travis, éditeurs : *History and GIS : Epistemologies, Considerations and Reflections*, pages 211–239. Springer Netherlands, Dordrecht, 2013. ISBN 978-94-007-5009-8. URL https://doi.org/10.1007/978-94-007-5009-8_14.

Volker Walter et Dieter Fritsch : Matching spatial data sets : a statistical approach. *International Journal of Geographical Information Science*, 13(5):445–473, juillet 1999. ISSN 1365-8816. URL https://doi.org/10.1080/136588199241157.

Dominique Waquet : Almanachs, cadastre, terriers clés du décodage des numéros sectionnaires d'immeubles à Paris, 1791 - 1805. *Annales historiques de la Révolution française*, 392(2):173–183, 2018. ISSN 0003-4436. URL https://www.cairn.info/revue-annales-historiques-de-la-revolution-francaise-2018-2-page-173.htm.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, Yong Jiang et Wenjuan Han : ChatIE : Zero-Shot Information Extraction via Chatting with ChatGPT, mai 2024. URL http://arxiv.org/abs/2302.10205. Issue : arXiv :2302.10205 arXiv : 2302.10205 [cs].

Gerhard Weikum, Luna Dong, Simon Razniewski et Fabian Suchanek : Machine Knowledge : Creation and Curation of Comprehensive Knowledge Bases, mars 2021. URL http://arxiv.org/abs/2009.11564. Issue : arXiv :2009.11564 arXiv : 2009.11564 [cs].

Chris Welty et Richard Fikes : A Reusable Ontology for Fluents in OWL. *In Proceedings of the 2006 conference on Formal Ontology in Information Systems : Proceedings of the Fourth International Conference (FOIS 2006)*, pages 226–236, NLD, mai 2006. IOS Press. ISBN 978-1-58603-685-0.

Bernhard Westfechtel et Reidar Conradi : Software configuration management and engineering data management : Differences and similarities. *In* Boris Magnusson, éditeur : *System Configuration Management*, pages 95–106, Berlin, Heidelberg, 1998. Springer. ISBN 978-3-540-69086-3.

Xander Wilcke, Peter Bloem et Victor de Boer : The knowledge graph as the default data model for learning on heterogeneous knowledge. *Data Science*, 1(1):39–57, décembre 2017. ISSN 2451-8484. URL https://doi.org/10.3233/DS-170007.

Michael Worboys : A model for spatio-temporal information. *Proceedings the 5th International Symposium on Spatial Data Handling, 1992*, 2:602–611, 1992. URL https://cir.nii.ac.jp/crid/1573387450469304192.

May Yuan : Wildfire conceptual modeling for building GIS space-time models. *In Proceedings of GIS/LIS*, volume 94, page 860-869, octobre 1994.

L. A. Zadeh : Fuzzy sets. *Information and Control*, 8(3):338–353, juin 1965. ISSN 0019-9958. URL https://www.sciencedirect.com/science/article/pii/S0019958659024 1X.

Paul A. Zandbergen : A comparison of address point, parcel and street geocoding techniques. *Computers, Environment and Urban Systems*, 32(3):214–232, mai 2008. ISSN 0198-9715. URL https://www.sciencedirect.com/science/article/pii/S0198971507000890.

Christine Zanin et Marie-Laure Trémélo : *Savoir faire une carte*. Belin, 2003. URL https://shs.hal.science/halshs-00175132.

Meng Zhang, Wei Shi et Liqiu Meng : A generic matching algorithm for line networks of different resolutions. *Workshop of ICA commission on generalization and multiple representation computering faculty of a Coruña University-Campus de Elviña, Spain*, 9:101–110, juillet 2005.

Zexuan Zhong et Danqi Chen : A Frustratingly Easy Approach for Entity and Relation Extraction. *In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 50–61, Online, 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.naacl-main.5.

Chenghu Zhou, Hua Wang, Chengshan Wang, Zengqian Hou, Zhiming Zheng, Shuzhong Shen, Qiuming Cheng, Zhiqiang Feng, Xinbing Wang, Hairong Lv, Junxuan Fan, Xiumian Hu, Mingcai Hou et Yunqiang Zhu : Geoscience knowledge graph in the big data era. *Science China Earth Sciences*, 64(7):1105–1114, juillet 2021. ISSN 1869-1897. URL https://doi.org/10.1007/s11430-020-9750-4.

Xiaohan Zou : A Survey on Application of Knowledge Graph. *Journal of Physics : Conference Series*, 1487(1):012016, mars 2020. ISSN 1742-6588, 1742-6596. URL https://iopscience.iop.org/article/10.1088/1742-6596/1487/1/012016.