PeGazUs : une méthode de reconstitution de l'évolution des entités géographiques à partir de données hétérogènes et fragmentaires

Charly Bernard[1], Nathalie Abadie[1], Bertrand Duménieu[2], Julien Perret[1]

[1] LASTIG, Université Gustave Eiffel, IGN/ENSG
[2] CRH, EHESS-CNRS

charly.bernard@ensg.eu ; nathalie-f.abadie@ensg.eu ; bertrand.dumenieu@ehess.fr ; julien.perret@ensg.eu

Résumé

Constituer un référentiel géo-historique d'entités spatiales permet de nombreux cas d'application, comme l'étude des dynamiques urbaines. Différentes approches dans la littérature montrent qu'il est possible de construire un tel référentiel, mais imposent d'avoir des jeux de données homogènes et structurés, à différentes dates. La disponibilité croissante des sources d'archives numérisées et les progrès des méthodes d'extraction d'informations dans ces sources permettent désormais de produire de grands volumes de données hétérogènes, fragmentaires et incomplètes, sur les entités géographiques du passé et leurs évolutions. Or, les approches de création de référentiels géo-historiques existantes ne permettent pas encore d'intégrer ces types de données de façon satisfaisante. Dans cet article, nous proposons une méthode de reconstitution de l'évolution spatio-temporelle des entités géographiques à partir de données hétérogènes et fragmentaires provenant de différentes sources. Nous expliquons aussi la manière dont on vérifie la cohérence du graphe de données créé à partir de cette méthode. Enfin, nous mettons cette dernière en application sur le quartier de la Butte aux Cailles à Paris à partir de sources de la fin du XVIII^{e} siècle à nos jours.

Mots-clés

Index géographique urbain historique - Graphe de connaissances - Évolution des entités géographiques.

Abstract

Building a geo-historical repository of spatial entities can be used in a number of applications, such as the study of urban dynamics. Various approaches in the literature show that it is possible to build such a repository, but require homogeneous and structured datasets at different dates. The increasing availability of digitised archive sources and advances in methods for extracting information from these sources now make it possible to produce large volumes of heterogeneous, fragmentary and incomplete data on past geographical entities and their evolution. However, existing approaches to creating geohistorical reference systems are not yet able to integrate these types of data satisfactorily. In this article, we propose a method for reconstructing the spatio-temporal evolution of geographical features using heterogeneous and fragmentary data from different sources. We also explain how to check the consistency of the data graph created using this method. Finally, we apply the method to the Butte aux Cailles district of Paris, using sources from the late 18th century to the present day.

Keywords

Historical urban gazetteer - Knowledge graph - Geographical entities evolution.

1 Introduction

Un index géographique (ou *gazetteer*) est une liste de lieux dans laquelle on représente, pour chaque lieu, un nom, un type et lorsque c'est possible, une localisation la plupart du temps sous la forme de coordonnées géographiques [18]. Représenter ce type d'informations à l'échelle des adresses présente un double intérêt. D'une part, cela permet d'indexer spatialement des documents d'archives numérisés, dont beaucoup regorgent de mentions d'adresses : c'est le cas des annuaires, des plans de ville anciens, des registres administratifs, ou encore des documents notariés par exemple. D'autre part, cela permet éventuellement de dater ces documents en fonction des adresses qu'ils mentionnent et de leurs dates d'existence connues. Au cours de la dernière décennie, un consensus s'est formé autour de l'utilisation de graphes de connaissances pour représenter des *gazetteers* historiques. Ceux-ci s'avèrent en effet particulièrement adaptés pour intégrer des données très hétérogènes et de structure non connue a priori [4].

Représenter des adresses anciennes dans un graphe de connaissances géo-historique nécessite une ontologie adaptée et une approche de peuplement capable d'intégrer des données issues de sources hétérogènes, à différentes dates et fragmentaires. En effet, bien que très représentées dans les sources historiques, les adresses sont des entités géographiques dont la généalogie est peu documentée. Ainsi, on ne dispose généralement que d'attestations d'existence, discontinues dans le temps, présentant tantôt des recouvrements temporels tantôt des lacunes.

Dans cet article, nous proposons une nouvelle approche pour peupler une ontologie représentant des adresses an-

ciennes et leur évolution au cours du temps. À partir de données décrivant des états ou des événements et provenant de différentes sources, avec différents temps valides, nous proposons de reconstituer pour chaque adresse ou élément de la voirie urbaine (rue, place, etc.) l'évolution de ses propriétés avec un enchaînement de versions successives reliées entre elles par des changements. La principale contribution de cette approche est de fournir une représentation continue et cohérente de l'évolution de chaque adresse, inférée à partir d'attestations d'états ou d'événements ponctuelles ou discontinues. L'approche utilise l'ontologie PeGazUs introduite par Bernard *et al.* [7].

Cet article est organisé de la façon suivante : nous présentons tout d'abord un état de l'art sur les méthodes de peuplement d'ontologies visant à reconstituer des évolutions spatio-temporelles du territoire. Ensuite, nous proposons une approche de reconstitution automatique de l'évolution temporelle d'adresses anciennes à partir de données fragmentaires pouvant représenter pour chaque entité géographique, son état sur un intervalle défini ou un événement décrivant son évolution à un instant donné. Puis, nous montrons que notre méthode fournit des données cohérentes et continues dans le temps au sein du graphe. Enfin, nous appliquons notre méthode sur les adresses du quartier de la Butte aux Cailles à Paris, au cours du XIX^{e} siècle.

2 État de l'art et notions préliminaires

2.1 Représentation des dynamiques spatio-temporelles

L'intégration de l'aspect temporel pour représenter les dynamiques territoriales a fait l'objet de travaux très tôt dans le domaine des systèmes d'information géographique [8, 17]. De nombreux travaux se sont concentrés sur la définition de l'identité des entités géographiques [19, 12, 16, 14]. Del Mondo [14] indique que la notion d'identité est cruciale pour conceptualiser et modéliser des phénomènes et qu'elle permet d'en transcrire une représentation dans une base de données. Hallot [19] reprend le modèle de l'*Identity Base Change* [20] et considère que chaque entité a une durée de validité infinie dont l'évolution est marquée par l'alternance d'états spatio-temporels (existence, non existence, etc.). Les travaux visant à décrire les dynamiques territoriales dans des SIG se sont largement fondés sur des modèles de graphes spatio-temporels [14, 15, 13].

Kauppinen *et al.* [21] proposent une ontologie pour représenter les municipalité finlandaises et y ntroduisent le concept de Change Bridge qui consiste à décrire les changements qui relient deux états successifs d'une entité géographique.

Bernard *et al.* [6] proposent une ontologie appelée TSN (Territorial Statistical Nomenclature) qui permet de représenter les unités territoriales de la NUTS [1]. Son extension TSN-Change permet de tracer l'évolution des entités géographiques grâce aux différentes versions de la NUTS en réutilisant le concept de Change Bridge. Alors que Kauppinen *et al.* [21] utilisent un changement pour décrire une transition entre deux états successifs, un changement pour [6] constitue un ensemble de modifications élémentaires sur plusieurs données. Par exemple, lors de la fusion deux entités impliquant la création d'une troisième, un changement va modéliser trois modifications élémentaires : deux disparitions et une apparition.

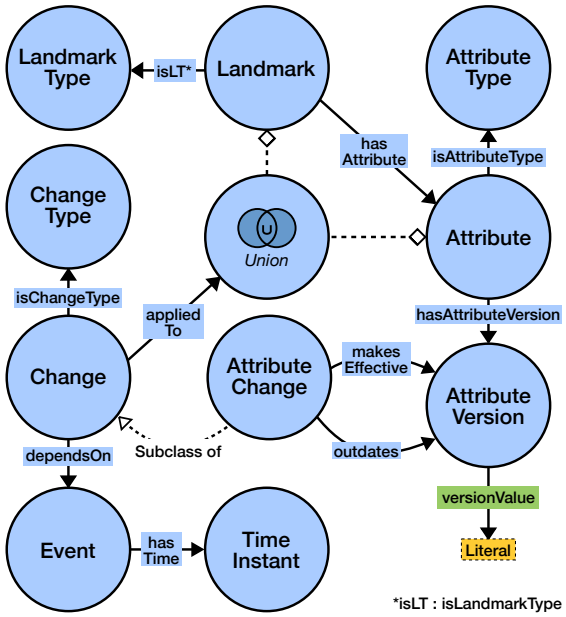

FIGURE 1 – Partie de l'ontologie PeGazUs pour modéliser l'évolution temporelle des entités géographiques.

Charles *et al.* [9] introduisent l'ontologie Hierarchical Historical Territory (HHT) et réutilisent aussi le concept de *Change Bridge* pour représenter l'évolution des unités territoriales de l'Ancien Régime et leurs multiples hiérarchies (religieuse, fiscale, judiciaire, etc.). Comme dans l'ontologie TSN-Change, cette évolution est représentée sous la forme de versions successives de ces unités territoriales, leurs propriétés demeurant constantes pour une même version. Dans le but de représenter de manière plus flexible les évolutions des emprises spatiales des unités territoriales, une extension de l'ontologie HHT a été ajoutée qui vise à la décrire sous la forme de zones élémentaires pouvant être intégrées à l'une ou l'autre des unités territoriales représentées selon les connaissances extraites des sources historiques [10].

L'ontologie PeGazUs [7] permet de décrire les entités géographiques et leurs évolutions (voir figure 1), mais se distingue des précédentes en versionnant non plus les entités géographiques, mais leurs attributs et en représentant les changements à leur niveau. Une entité géographique y est représentée dans la classe `Landmark`. On lui associe sa nature `LandmarkType` via le prédicat `isLandmarkType`. Elle possède des attributs (`Attribut`) typés (*cf.* `AttributeType`). Chaque attri-

1. Nomenclature des unités territoriales statistiques, voir https://ec.europa.eu/eurostat/fr/web/nuts/background

but compte un nombre non limité de versions définis par la classe `AttributeVersion`. Une version comporte une valeur qui est un `Literal`.
Un événement, associé à un instant (`TimeInstant`), décrit une évolution du territoire (fusion, scission, changement de nom, de géométrie, etc) entraînant une ou plusieurs modifications des ressources de type (`Landmark` ou `Attribut`) représentées par un changement. Par exemple, la disparition d'une entité géographique est décrite par un changement de type `LandmarkDisappearance`. `AttributeChange` est une sous-classe de `Change` et permet de décrire une évolution dans les données au niveau des attributs. Elle permet de préciser la façon dont le changement s'applique à l'attribut en décrivant la mise en effectivité (avec `makesEffective`) et/ou l'obsolescence (avec `outdates`) de versions.

2.2 Inférences des dynamiques spatio-temporelles à partir de descriptions d'états

Plusieurs approches ont été proposées dans la littérature [21, 6, 11], pour peupler des gazetiers historiques concernant des unités administratives ou statistiques. Les deux premières utilisent des jeux de données géographiques structurées versionnés par année de validité des données. La dernière permet d'intégrer des données associées à des dates de validité différentes au sein du même jeu de données, mais décrivant l'évolution des entités géographiques sans recouvrements ni lacunes. Seule l'approche proposée par [7] permet d'intégrer des données sur d'autres types d'entités géographiques, dont la généalogie est moins bien documentée dans les sources historiques et pour lesquelles on ne dispose généralement que d'attestations d'existence, discontinues dans le temps, présentant tantôt des recouvrements temporels tantôt des lacunes. Cependant, cette approche comporte encore des limites : elle ne permet pas de prendre en compte des données décrivant les événements qui affectent les entités géographiques ; les changements inférés peuvent présenter des incertitudes temporelles ; ils peuvent également présenter des incohérences en cas d'attestations contradictoires concomitantes.
Qu'elles utilisent les systèmes d'information géographique [15], ou les graphes de connaissances [21, 5], les premières approches proposées dans la littérature pour peupler un *gazetteer* historique utilisent des jeux de données géographiques structurées, versionnés par année de validité des données. Sur la base d'un critère d'identité donné, elles comparent les états successifs des entités géographiques deux à deux afin de détecter des changements potentiels entre eux et d'en déduire les types d'événements du monde réel qui ont pu conduire à de tels changement au niveau des données.
L'approche proposée par [11] s'en distingue en permettant de traiter des données géographiques associées à des dates de validité différentes au sein du même jeu de données, mais décrivant les états des entités géographiques concernées sans recouvrements ni lacunes temporels entre ces états. Cette approche a été récemment étendue pour détecter de qualifier des changements des emprises spatiales des unités territoriales représentées même lorsque les données de géométrie destinées à les représenter sont absentes ou imprécises. Utilisant le principe de blocs élémentaires pour constituer des géométries, cette approche est particulièrement adaptée dans le cas où les entités géographiques sont hiérarchisées par des relations d'inclusion spatiale comme c'est le cas pour les unités administratives. Cette approche se limite cependant à l'étude de l'évolution de géométries : elle ne peut être appliquée pour des attributs comme le nom, le code INSEE, etc.
La limite de toutes ces approches est qu'elles ne s'appliquent pas aux entités géographiques, dont la généalogie est moins bien documentée dans les sources historiques que dans le cas des unités administratives, et pour lesquelles on ne dispose que d'attestations d'existence, discontinues dans le temps, présentant tantôt des recouvrements temporels tantôt des lacunes. De plus, l'identité des entités géographiques y est soit connue a priori, soit triviale à retrouver. Ainsi, les faits qui décrivent la même entité sont bien associés à cette dernière dès leur intégration dans le graphe.
L'approche de peuplement d'un graphe de connaissances géo-historique à partir de données hétérogènes, fragmentaires et incomplètes provenant de sources multiples proposée par Bernard *et al.* [7] vise à dépasser ces limites. Chaque source historique contient des informations à intégrer qu'on appelle factoïdes [23]. Ces dernières sont représentées selon l'ontologie Pegazus et leurs triplets sont stockés dans un graphe nommé regroupant l'ensemble des factoïdes liés à la source. Le résultat final est un graphe de faits qui est la réconciliation de l'ensemble des graphes de factoïdes où chaque fait est sourcé par un ou plusieurs factoïdes. La première étape consiste à détecter les faits au sein des différentes sources qui décrivent les mêmes entités géographiques sur la base d'un ou plusieurs critères. Puis, les faits sont ordonnés temporellement afin d'en déduire les événements qui décrivent l'apparition et la disparition des entités géographiques qu'ils décrivent. Ce tri temporel ne repose pas sur l'algèbre des intervalles d'Allen [1] du fait de la présence d'instants dans la phase de tri et des possibles ambiguïtés dans la combinaison de relations, notamment en procédant à un raisonnement par transitivité. L'évolution des attributs de chaque entité géographique est ensuite inférée en détectant les changements entre deux versions d'attributs successives.
Si elle présente l'avantage de permettre d'intégrer des données fragmentaires extraites de sources historiques offrant des représentations partielles, discontinues et potentiellement redondantes ou complémentaires des entités géographiques au cours du temps, cette approche comprend plusieurs limites. Premièrement, les changements inférés sont temporellement flous. Si deux versions successives sont espacées d'un siècle, alors l'incertitude sur la date du changement sera de l'ordre du siècle. Au lieu d'utiliser uniquement des versions pour en déduire des changements comme le font toutes les méthodes, il serait donc utile d'intégrer des changements dans les données de départ qui permettraient d'inférer des versions. Enfin, la méthode a des problèmes

d'inférence de changements dans le cas où des versions différentes se chevauchent temporellement. Elle en déduit des événements dont la période temporelle associée est située dans un intervalle $[t_i, t_j]$ où $t_i > t_j$, ce qui est constitue une incohérence.

3 Reconstruction de l'évolution des entités

Dans cette section, nous introduisons une extension de PeGazUs [2] qui, en plus d'introduire une ontologie, propose une méthode de reconstitution de l'évolution des entités géographiques à partir de données hétérogènes et fragmentaires [7]. La méthode se décompose en trois grandes étapes. En premier lieu, il s'agit de représenter les données sur les entités géographiques conformément à l'ontologie PeGazUs puis de lier les données décrivant les mêmes entités en fixant un critère de similarité. Pour les quartiers et les voies de communication, ce critère repose sur la similarité du nom tandis pour les numéros d'habitation, deux entités sont équivalentes si elles possèdent le même numéro en plus d'être liés à la même voie par le biais d'une relation (`LandmarkRelation`) de même nature.
Puis, nous créons une succession de versions d'attributs élémentaires indivisibles qui ne se chevauchent pas temporellement et où il n'existe pas de période temporelle sans version d'attribut (voir figure 4). Enfin, nous fusionnons les versions successives initialisées lors de l'étape précédente selon plusieurs critères. Par exemple, deux versions successives de "Nom" dotées de labels identiques pour la même rue seront fusionnées.

3.1 Initialisation multi-source des entités

Comme le fait Bernard *et al.* [7] (voir section 2.2), il faut débuter par la description des données issues des sources selon notre ontologie. Les triplets générés sont dans des graphes nommés associés chacun à une source. Ils permettent la création du graphe de faits résultant de la construction multi-source du référentiel. Chaque fait est lié à un ou plusieurs factoïdes comme le montre la figure 2. Pour une ressource r du graphe des faits et une de ses attestations a dans une source donnée, il existe un triplet tel que $hasTrace(r, a)$. On dit que r est tracé par a.
Ce liage permet d'inférer les changements d'apparition et de disparition ainsi que les événements dont ils dépendent. Si, pour une entité géographique, il existe dans les sources des attestations de sa création (respectivement de sa disparition), alors on peut générer une instance de la classe `Change` de type `LandmarkAppearance` (respectivement `LandmarkDisappearance`) et déduire quand cette dernière est apparue (respectivement a disparu) à partir du temps valide associé à cette attestation dans la source. Dans le cas contraire, on ne peut pas avoir une valeur temporelle précise mais une estimation. En extrayant

2. *PErpetual GAZeteer of approach-address UtteranceS*. Sa documentation, les données et le script permettant de construire le graphe sont disponibles sur le dépôt https://github.com/charlybernard/pegazus-extension

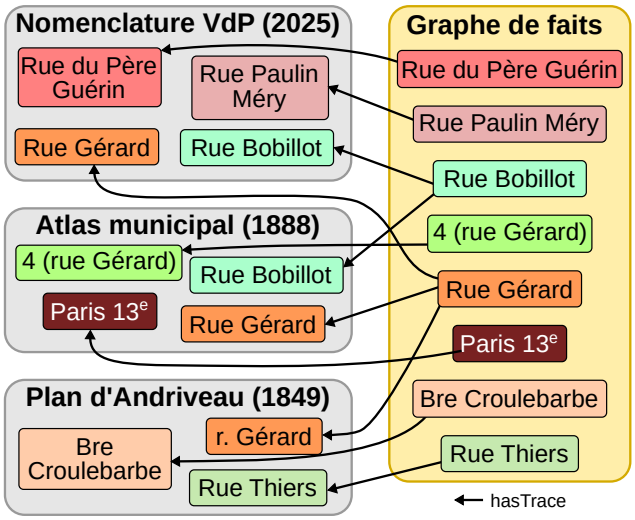

FIGURE 2 – Initialisation du graphe de faits à partir des graphes de factoïdes décrivant chacun une source.

le temps valide associé à la plus ancienne attestation d'une entité dans une source, on en déduit que son apparition se déroule avant la date extraite. On fait de même avec la disparition en sélectionnant la date de l'attestation la plus récente, dans la source considérée. En prenant l'exemple de la figure 2, la rue Bobillot est mentionnée dans une source datant de 1888 ainsi que dans une autre datant 2025. Ainsi, on peut déduire que la voie apparaît avant 1888 et disparaît après 2025.
Toutefois, l'apparition et la disparition ne sont pas les seuls changements qui s'appliquent aux entités géographiques. Les évolutions de noms ou de géométrie constituent d'autres changements à représenter. L'ontologie PeGazUs modélise les attributs de manière indépendante, ce qui permet de représenter leur évolution avec un ensemble de versions associées chacune à une période de validité délimitée par des changements. Néanmoins, ce processus demande une méthode spécifique dont la première partie demande d'effectuer une représentation élémentaire de leur évolution.

3.2 Représentation élémentaire de l'évolution des attributs

Après l'étape initiale, 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 des changements et $V = \{v_1, ..., v_m\}$ pour celui des versions.
Pour commencer, il faut procéder à un découpage élémentaire comme décrit par la figure 4. L'objectif est de générer une succession de versions élémentaires et indivisibles en fonction de l'ordre relatif des attestations. Pour cela, deux phases de traitement sont nécessaires : l'initialisation de

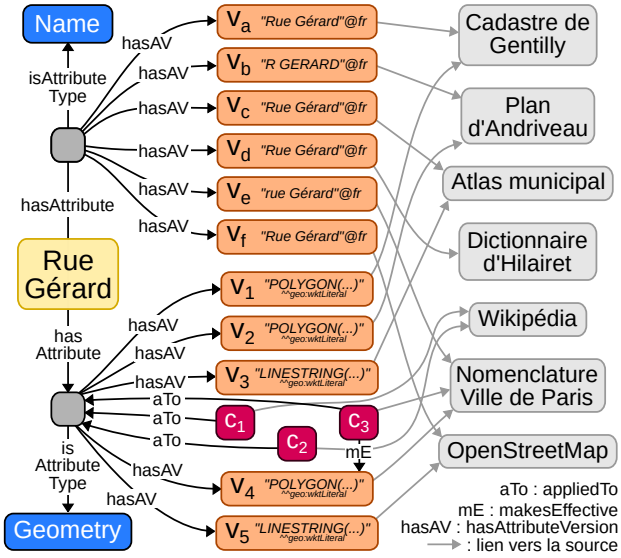

FIGURE 3 – Regroupement des attestations des différentes sources sur les attributs de la rue Gérard.

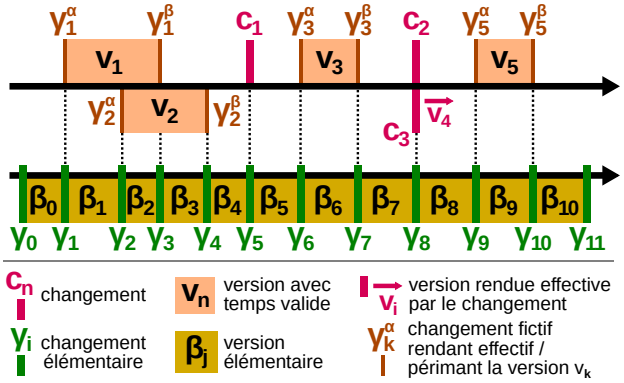

FIGURE 4 – Création de l'ensemble de changements Γ et de versions B (frise du bas) à partir de versions et de changements. La frise du haut montre l'agencement temporel des éléments de C et V pour l'attribut "géométrie" de la rue Gérard.

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 3, 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 ;
- 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 temporel des temps valides de v_1 et v_2 que, d'autre part, 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.

3.2.1 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 séparés par des changements comme le montre la frise du bas de la figure 4. 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.

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 $\forall i \in [\![1; m]\!], makesEffective(\gamma_i^\alpha, v_i) \wedge outdates(\gamma_i^\beta, v_i)$. 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 associés chacun à deux changements fictifs indiquant sa mise en effectivité et sa 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 $\Omega = \left\{c_1, c_2, c_3, \gamma_1^\alpha, \gamma_1^\beta, \gamma_2^\alpha, \gamma_2^\beta, \gamma_3^\alpha, \gamma_3^\beta, \gamma_5^\alpha, \gamma_5^\beta\right\}$. Les changements c_2 et c_3 ont lieu en 1979 donc leur simultanéité permet d'en déduire qu'ils sont similaires. En agrégeant les changements simultanés de Ω, on génère l'ensemble Γ valant $\{\gamma_1, \ldots, \gamma_{10}\}$.

On y ajoute 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 4). Enfin, on trie temporellement ces éléments en reliant deux changements successifs avec le triplet $hasNextChange(c_i, c_j)$. Pour $\gamma \in \Gamma$, si $\exists \delta = \underset{x \in \Omega \setminus \{\gamma\}, t(x)-t(\gamma)>0}{\arg\min} t(x) - t(\gamma)$ où $t(x)$ est la valeur temporelle associé au changement x alors $hasNextChange(\gamma, \delta)$.

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 : $\forall(\gamma_i, \gamma_j) \in \Gamma^2, hasNextChange(\gamma_i, \gamma_j) \implies \exists \beta \in B, AttributeVersion(\beta) \wedge makesEffective(\gamma_i, \beta) \wedge outdates(\gamma_j, \beta)$.

3.2.2 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 : $\forall c \in C, \exists \gamma \in \Gamma, hasTrace(\gamma, c)$ (cela vaut aussi

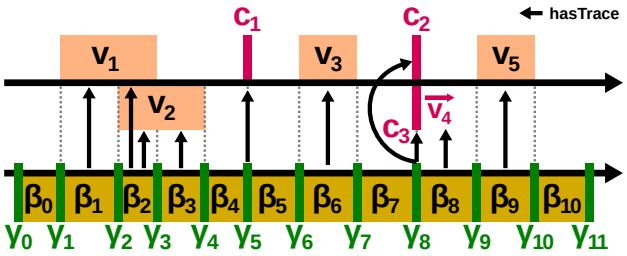

FIGURE 5 – Affiliation des changements et des versions pour l'attribut géométrie de la rue Gérard.

respectivement pour V et B). L'intérêt de l'affiliation est de lier les données que l'on vient d'initialiser aux factoïdes.

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 $hasTrace(\gamma, c)$. Ainsi, pour la rue Gérard, les triplets générés sont $hasTrace(\gamma_5, c_1)$, $hasTrace(\gamma_8, c_2)$ et $hasTrace(\gamma_8, c_3)$.

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 5, β_1 et β_2 sont tracées par v_1 et β_2 et β_3 sont tracées par v_2. Pour $(v, \beta) \in (V, B)$, la relation $hasTrace(\beta, v)$ est satisfaite si l'une des deux conditions suivantes est remplie : soit si l'intersection de leurs temps valides est non vide, soit s'ils dépendent chacun d'un changement, et l'un de ces changements est la trace de l'autre. Autrement dit, $\exists(c, \gamma) \in (C, \Gamma)$ tel que $makesEffective(c, v) \wedge makesEffective(\gamma, \beta) \wedge hasTrace(\gamma, c)$. Cette condition reste valable si l'on remplace $makesEffective$ par $outdates$. 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 $hasTrace(\beta_i, v_j)$ 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).

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

3.3.1 Suppression de versions lacunaires

Le découpage élémentaire présenté en section 3.2 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 prédécesseure et/ou leur successeure) permettrait de combler les lacunes sur ces intervalles temporels. Différents critères doivent être pris en compte

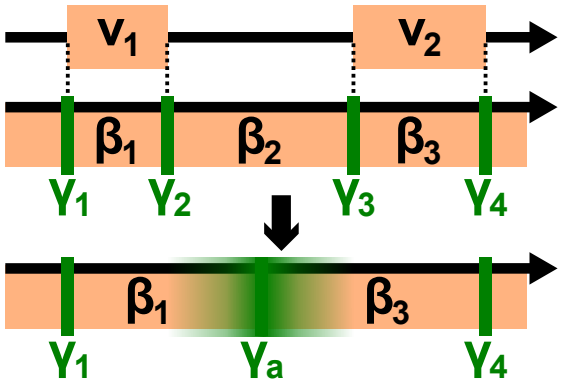

FIGURE 6 – Cas d'une version non tracée (ici β_2) délimitée par deux changements non tracés.

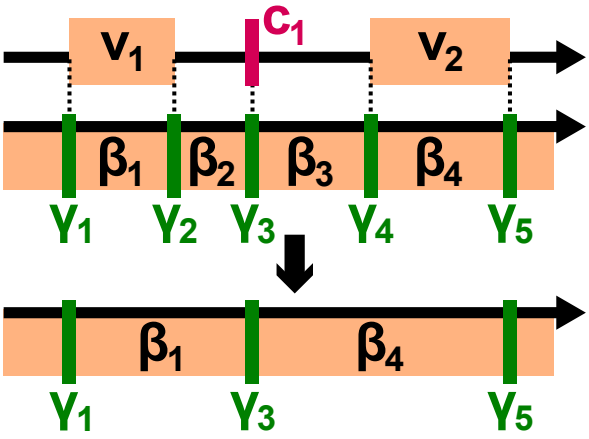

FIGURE 7 – Cas d'une version non tracée (ici β_2 et β_3) délimitée par un seul changement non tracé.

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 résulte 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 5 ne pourront être fusionnées.

Pour une version β non tracée, trois cas existent :

- ses deux changements ne sont pas tracés ;
- un seul des deux changements est tracé ;
- ses deux changements sont tracés.

Le premier cas est illustré par la figure 6. 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 qui rend obsolète β_1 (et rend effectif β_2) soit le même que celui qui rend effectif β_3 (et périme β_2). 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 $t_1 \leq t \leq t_2$.

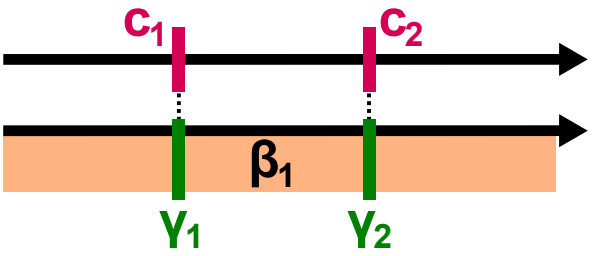

FIGURE 8 – Version non tracée (ici β_1) délimitée par deux changements tracés.

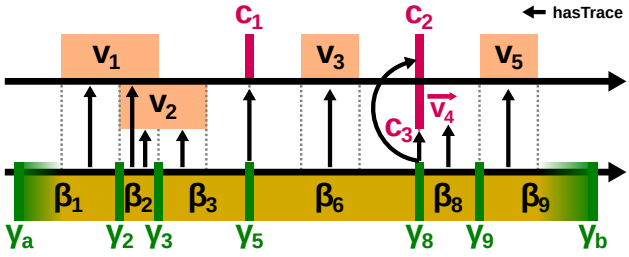

FIGURE 9 – Reconstitution de l'évolution temporelle après suppression de versions non tracées pour l'attribut géométrie de la rue Gérard.

Le deuxième cas, illustré par la figure 7, 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 7, β_2 partage un changement non tracé avec β_1 tout comme β_3 avec β_4.

Lorsque les deux changements d'une version sont tracés (illustré par la figure 8), alors il ne faut rien faire.

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 9). 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.

3.3.2 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 9 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.

Il est donc nécessaire, au préalable, de comparer les fac-

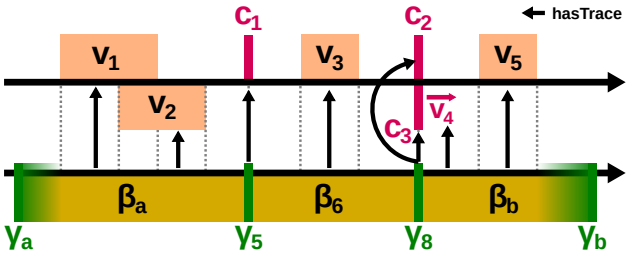

FIGURE 10 – Reconstitution finale de l'évolution temporelle après fusion des versions similaires pour l'attribut géométrie de la rue Gérard.

toïdes décrivant des versions entre eux en fonction de leur valeur via deux prédicats : `sameVersionValueAs` et `differentVersionValueFrom`. Les critères de similarité sont à personnaliser en fonction du type d'attribut considéré. Par exemple, nous avons choisi de comparer les versions d'attributs de nom par stricte égalité de leur nom simplifié. Ce nom simplifié est le nom de la voie auquel on fait les traitements suivants : suppression des signes diacritiques ; remplacement des caractères non alphanumériques par des espaces ; suppression des articles, prépositions, conjonctions et adverbes ; mise en bas de casse des caractères ; tri alphanumérique des caractères (avec la suppression des espaces).

Ainsi, pour une version dont la valeur est « rue du Père-Guérin », sa valeur simplifiée avant le tri alphanumérique sera « rue pere guerin » soit « eeeeginprrruu » après.

Du fait de l'hétérogénéité importante des géométries venant des sources, déterminer la similarité de deux versions v_1 et v_2 d'un attribut de type géométrie ayant des valeurs g_1 et g_2 est moins aisé. Nous proposons ici d'adopter la méthode suivante : en notant $S(g)$ la surface d'une géométrie, on peut considérer que v_1 et v_2 sont similaires si $\frac{S(g_1 \cap g_2)}{S(g_1 \cup g_2)} \geq \alpha_{min}$ où $\alpha_{min} \in [0;1]$ est un coefficient minimal de similarité.

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_1 et β_2 sont les suivantes :

- avoir un changement en commun qui ne soit pas tracé : $\exists \gamma, \nexists hasTrace(\gamma, c) \wedge makesEffective(\gamma, \beta_1) \wedge outdates(\gamma, \beta_2)$;
- les versions dont elles dérivent sont similaires : $hasTrace(\beta_1, v_1) \wedge hasTrace(\beta_2, v_2) \wedge (sameVersionValueAs(v_1, v_2) \vee v_1 = v_2)$.

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 10). 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 $sameVersionValueAs(v_4, v_5)$.

Grâce à cette approche, nous sommes désormais capables de reconstituer l'évolution détaillée de chaque attribut d'une entité géographique, permettant de constituer l'historique de cette dernière. Cependant, pour garantir la fiabilité de ce processus, il est nécessaire d'évaluer la cohérence des données générées par la méthode présentée.

4 Vérification de la cohérence

Une façon d'évaluer la méthode de peuplement que nous proposons consiste à vérifier la cohérence du graphe des faits produit. Plusieurs éléments sont à vérifier :

- les factoïdes modélisés selon l'ontologie ;
- le graphe de faits construit selon l'ontologie ;
- l'alternance de versions et de changements ;
- l'absence de versions lacunaires, supprimées conformément lors de l'étape du traitement décrite en section 3.3.1 ;
- l'absence de versions successives similaires conformément à l'approche décrite dans section 3.3.2.

Une première étape pour s'assurer de la cohérence du graphe de faits consiste à s'assurer que les données à partir desquelles il est construit sont elles-mêmes cohérentes. Pour ce faire, nous testons la cohérence des triplets des graphes de factoïdes représentés avec l'ontologie PeGazUs à l'aide de règles SHACL [22] et de requêtes SPARQL. Ceci nous assure que la forme des données initiales ne peut pas être la raison d'incohérences dans le graphe final. Le même traitement est ensuite appliqué sur le graphe de faits pour vérifier que la méthode de peuplement que nous proposons produit bien un graphe compatible avec l'ontologie PeGazUs. Puis, nous vérifions l'alternance de versions et de changements pour les attributs dans le graphe de faits par l'intermédiaire d'une requête SPARQL. Elle permet aussi de vérifier qu'il n'existe pas de période temporelle non couverte par une version d'attribut conformément aux objectifs présentés en section 1. Pour s'assurer de l'absence de telles lacunes, on vérifie que toutes les versions sont tracées par au moins une affirmation venant des graphes de factoïdes. Si des versions non tracées existent, il convient de vérifier si les changements qui leur sont associés sont tracés comme décrit en figure 8. Dans ce cas, elles restent bien cohérentes. Enfin, l'évaluation de l'étape de fusion de versions successives similaires présentée en section 3.3.2 doit être réalisée manuellement. Pour simplifier cette vérification et éviter d'avoir à naviguer dans le graphe, un outil de visualisation cartographique. D'une part, il permet de visualiser l'évolution d'une entité géographique et celle de l'état du territoire à un instant donné. L'interface présente l'agencement des versions avec leur date de validité et leur(s) valeur(s) sur une frise chronologique. Pour les attributs de type géométrie, une carte interactive affiche les valeurs des versions. D'autre part, l'outil propose d'afficher l'état du territoire à un instant donné.

5 Mise en œuvre de la méthodologie

5.1 Application sur la Butte aux Cailles

Notre approche a été testée sur un ensemble de données provenant de sources décrivant les voies et numéros d'immeuble du quartier de la Butte aux Cailles situé dans le 13e arrondissement de Paris sur une période allant de la fin du XVIIIe siècle jusqu'à aujourd'hui. Ce quartier était situé dans la commune de Gentilly avant 1860, date à laquelle le territoire parisien s'est étendu jusqu'à l'enceinte de Thiers. Cet événement a permis la transformation urbaine de ce quartier agricole en quartier résidentiel assez dense, ce qui implique de nombreux changements sur les entités géographiques de cette zone.

5.2 Présentation des sources utilisées

Pour reconstituer l'évolution des adresses et des rues de la Butte aux Cailles, nous avons utilisé des données contemporaines comme les *Dénominations des emprises des voies actuelles* [3] et les *Dénominations caduques des voies* [4] de la ville de Paris qui décrivent des voies de communication. La *Base Adresse Nationale* [5] (BAN) décrit des numéros d'immeubles. Des données d'OpenStreetMap, de Wikidata et de Wikipédia sont aussi prises en compte.
Nous avons également intégré des données vectorisées manuellement à partir de plans anciens décrivant le territoire parisien : le cadastre napoléonien de Gentilly (1847), le plan Andriveau de 1849 [3], le plan parcellaire municipal (1871) et l'atlas municipal de 1888 [2].

5.3 Évaluation du graphe final

Ne disposant pas de vérité terrain sur laquelle s'appuyer, nous ne sommes pas en mesure de produire une évaluation quantitative et systématique du graphe final. Nous pouvons toutefois faire une étude qualitative sur le graphe obtenu en le comparant avec l'historique des voies fournie par la nomenclature des *Dénominations des emprises des voies actuelles* de la ville de Paris, historique que nous n'utilisons pas en entrée du processus de peuplement du graphe. L'outil de visualisation mentionné en section 4 et illustré par la figure 11, fournit une frise chronologique pour chaque attribut de l'entité sélectionnée. En comparant les résultats de la reconstitution automatique de l'évolution des rues, on remarque que les données sont globalement en accord avec celles fournies dans cet historique. Néanmoins, il demeure quelques incohérences, causées par des conflits entre sources ou bien par des critères de similarité entre versions successives trop stricts (voir section 3.3.2). Par exemple, le graphe nous indique qu'un changement de nom aurait eu lieu le 9 août 1888 où la rue Bobillot devient la rue Bobillot : l'erreur ici est qu'avant cette date, il n'existait aucun nom pour la voie. Cette erreur s'explique par le fait que le plan parcellaire municipal de la ville de Paris a été établi

3. https://opendata.paris.fr/explore/dataset/denominations-emprises-voies-actuelles
4. https://opendata.paris.fr/explore/dataset/denominations-des-voies-caduques
5. https://adresse.data.gouv.fr/

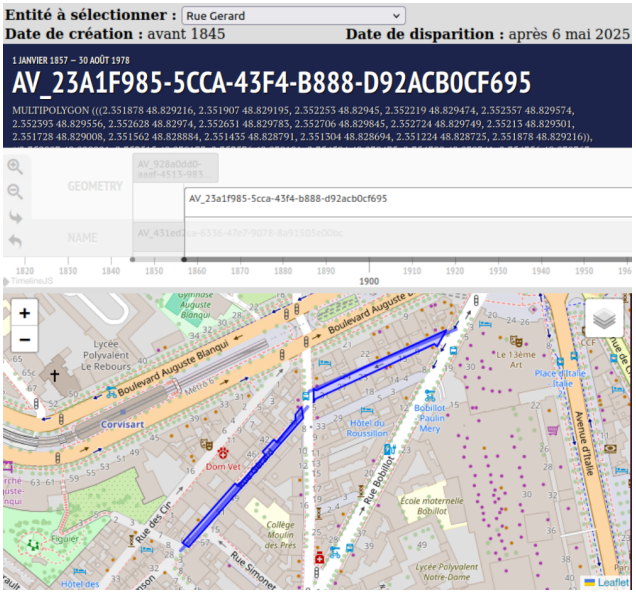

FIGURE 11 – Outil de visualisation de l'évolution des entités géographiques affichant la rue Gérard. La géométrie affichée est valable entre 1857 et 1978.

FIGURE 12 – Outil de visualisation de l'évolution des entités géographiques affichant un snapshot pour l'année 1888 autour de la rue Gérard. Le fond de plan est le plan d'Andriveau.

grâce à des relevés réalisés entre 1871 et 1896. Le temps valide des entités a été fixé ici autour de 1871 alors que le relevé pour cette rue est plus récent. Ici, c'est donc la source utilisée qui cause une incohérence dans les données, pas notre méthode de peuplement. À l'inverse, pour la rue Gérard, notre méthode déduit qu'il y a eu un changement de géométrie dans les années 1850 alors qu'aucune source ne le mentionne. Cette erreur est due à un problème de non fusion de versions similaires, estimées à tort comme non suffisamment similaires.

Parallèlement, l'outil de visualisation temporelle a été utilisé pour générer des snapshots du territoire à des dates spécifiques, sélectionnées en fonction de leur inclusion dans les intervalles de validité des sources utilisées pour la construction du graphe. Par exemple, l'année 1888 a été retenue afin d'évaluer la conformité des entités reconstruites avec le plan Andriveau de cette période.

6 Conclusion

Dans cet article, nous avons présenté une approche pour peupler un graphe de connaissances géo-historique d'adresses à partir de données hétérogènes et fragmentaires. La contribution de ce travail est la méthodologie de construction de l'évolution spatio-temporelle des entités géographiques avec des données décrivant des états ou des événements sans que les sources dont elles sont issues ne couvrent temporellement la totalité de la période d'existence des entités géographiques. À l'inverse, certaines sources utilisées peuvent présenter des chevauchements temporels et proposer des attestations contradictoires. Pour vérifier la cohérence des données en sortie, nous fournissons un ensemble de préconisations avant de mettre en œuvre la méthode sur un jeu de données décrivant le quartier de la Butte aux Cailles à Paris depuis la Révolution française.

Par la suite, nous comptons évaluer cette méthode de peuplement de façon quantitative et systématique, en l'appliquant sur des jeux de données d'adresses récents, exhaustifs, et à différents temps valides, dont on retire tour à tour l'un ou l'autre des millésimes, destiné à servir de vérité terrain pour son année de validité.

Les données représentant des événements utilisées dans cet article proviennent de données textuelles et leur extraction et structuration ont été réalisées manuellement. Une piste d'enrichissement serait de permettre leur reconnaissances et leur structuration automatiques à l'aide de grands modèles de langage (LLM). Les données et les scripts utilisés pour cet article sont disponible sur le dépôt https://github.com/charlybernard/pegazus-extension.

Références

[1] James F. Allen. Maintaining Knowledge about Temporal Intervals. In Daniel S. Weld and Johan de Kleer, editors, *Readings in Qualitative Reasoning About Physical Systems*, pages 361–372. Morgan Kaufmann, January 1990.

[2] Adolphe Alphand and 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.

[3] J Andriveau-Goujon. Plan de Paris fortifié et des communes environnantes : 1849 / Le plan et la lettre gravés par P. Rousset, 1849.

[4] Merrick Lex Berman, Ruth Mostern, and Humphrey Southall, editors. *Placing Names : Enriching and In-*

tegrating Gazetteers. Indiana University Press, August 2016.

[5] Camille Bernard, Christine Plumejeaud, Marlène Villanova-Oliver, Jerome Gensel, and Hy Dao. *An Ontology-based Algorithm for Managing the Evolution of Multi-Level Territorial Partitions*. November 2018.

[6] Camille Bernard, Marlène Villanova-Oliver, Jérôme Gensel, and Hy Dao. Ontologies pour représenter l'évolution des découpages territoriaux statistiques. *Revue Internationale de Géomatique*, 28(4) :409–437, December 2018.

[7] Charly Bernard, Solenn Tual, Nathalie Abadie, Bertrand Duménieu, Joseph Chazalon, and Julien Perret. PeGazUs : A Knowledge Graph Based Approach to Build Urban Perpetual Gazetteers. In Mehwish Alam, Marco Rospocher, Marieke Van Erp, Laura Hollink, and Genet Asefa Gesese, editors, *Knowledge Engineering and Knowledge Management*, volume 15370, pages 364–381. Springer Nature Switzerland, Cham, 2025. Series Title : Lecture Notes in Computer Science.

[8] Peter K Bol. The China Historical Geographic Information System (CHGIS) Choices Faced, Lessons Learned. *Conference on Historical Maps and GIS*, 23, August 2007.

[9] William Charles, Nathalie Aussenac-Gilles, and Nathalie Hernandez. HHT : An Approach for Representing Temporally-Evolving Historical Territories. In Catia Pesquita, Ernesto Jimenez-Ruiz, Jamie McCusker, Daniel Faria, Mauro Dragoni, Anastasia Dimou, Raphael Troncy, and Sven Hertling, editors, *The Semantic Web*, volume 13870, pages 419–435. Springer Nature Switzerland, Cham, 2023. Series Title : Lecture Notes in Computer Science.

[10] William Charles, Nathalie Aussenac-Gilles, and Nathalie Jane Hernandez. Diachronical geometry without polygons : the extended HHT ontology for heterogeneous geometrical representations. volume 15233, page 80. Springer Nature Switzerland ; Springer, November 2024.

[11] Jean Charlet, Bruno Bachimont, and Raphaël Troncy. Ontologies pour le Web sémantique. Technical report, CNRS, réseau thématique pluridisciplinaire Documents publié dans le numéro spécial Web sémantique de la revue I3 (https ://www.irit.fr/journal-i3/hors_serie/annee2004/index_fr.php), 2004.

[12] Christophe Claramunt, Marius Thériault, and Christine Parent. A qualitative representation of evolving spatial entities in two-dimensional topological spaces. In *Innovations In GIS 5*, pages 128–142. CRC Press, March 1998.

[13] Benoît Costes. *Vers la construction d'un référentiel géographique ancien : un modèle de graphe agrégé pour intégrer, qualifier et analyser des réseaux géo-historiques*. PhD Thesis, Université Paris-Est, November 2016.

[14] Géraldine Del Mondo. *Un modèle de graphe spatio-temporel pour représenter l'évolution d'entités géographiques*. phdthesis, Université de Bretagne occidentale, Brest, October 2011.

[15] Bertrand Dumenieu. *Un système d'information géographique pour le suivi d'objets historiques urbains à travers l'espace et le temps*. PhD Thesis, École des Hautes Études en Sciences Sociales, December 2015.

[16] Y. T. Fan, J. Y. Yang, D. H. Zhu, and K. L. Wei. A time-based integration method of spatio-temporal data at spatial database level. *Mathematical and Computer Modelling*, 51(11) :1286–1292, June 2010.

[17] Ian N. Gregory, Chris Bennett, Vicki L. Gilham, and Humphrey R. Southall. The Great Britain Historical GIS Project : From Maps to Changing Human Geography. *The Cartographic Journal*, 39(1) :37–49, June 2002.

[18] Karl Grossner, Krzysztof Janowicz, and Carsten Kessler. Place, Period, and Setting for Linked Data Gazetteers. In Merrick Lex Berman, Ruth Mostern, and Humphrey Southall, editors, *Placing Names*, The Spatial Humanities, pages 80–96. Indiana University Press, Bloomington, IN, 2016.

[19] 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.* PhD Thesis, ULiège - Université de Liège, March 2012.

[20] Kathleen Hornsby and Max J. Egenhofer. Identity-based change : a foundation for spatio-temporal knowledge representation. *International Journal of Geographical Information Science*, 14(3) :207–224, April 2000.

[21] Tomi Kauppinen, Jari Väätäinen, and Eero Hyvönen. Creating and Using Geospatial Ontology Time Series in a Semantic Cultural Heritage Portal. In Sean Bechhofer, Manfred Hauswirth, Jörg Hoffmann, and Manolis Koubarakis, editors, *The Semantic Web : Research and Applications*, volume 5021, pages 110–123. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008. Series Title : Lecture Notes in Computer Science.

[22] Holger Knublauch and Dimitris Kontokostas. Shapes Constraint Language (SHACL), July 2017.

[23] Michele Pasin and John Bradley. Factoid-based prosopography and computer ontologies : towards an integrated approach. *Digital Scholarship in the Humanities*, 30(1) :86–97, April 2015.