Une approche semi-automatique pour l'extraction d'informations liées aux itinéraires culturels à partir des réseaux sociaux : cas de la Via Francigena

Nathalie Valmond-Leblanc*, Eric Kergosien**
Natalia Grabar* Marta Severo**

*STL, Université de Lille
Domaine Universitaire Point de bois, BP 60149, 59653, Villeneuve d'Ascq, France
prenom.nom@univ-lille3.fr,
**GERiiCO, Université de Lille
Domaine Universitaire Point de bois, BP 60149, 59653, Villeneuve d'Ascq, France
prenom.nom@univ-lille3.fr

Résumé. Dans le cadre de l'analyse des itinéraires culturels, objet complexe du point de vue territorial, symbolique et social, le projet Itinéraire Numérique s'attache à analyser les contenus numériques disponibles en ligne pour la valorisation de l'itinéraire Via Francigena. Dans cet article, nous proposons une méthodologie semi-automatique pour l'extraction d'information spatiale liées à la Via Francigena à partir des données Instagram.

1 Introduction

Le fait que le patrimoine culturel soit disponible au format numérique modifie la manière de le comprendre et de le représenter. En particulier, cette abondance des données numériques a produit une nouvelle effervescence vers les possibilités de représentation du patrimoine. Des projets comme la plateforme de l'Institut Culturel de Google, le projet ORBIS de Stanford (Stanford Geospatial Network Model of the Roman World), ou les différents projets Europeana constituent de bons exemples d'une telle effervescence. L'objectif du projet « Itinéraire Numérique » [1] est de fédérer une approche interdisciplinaire pour étudier les modalités de représentation numérique du patrimoine culturel lié à l'itinéraire Via Francigena. Dans cet article nous décrivons une méthode semi-automatique pour extraire les entités spatiales (ES) à partir de messages Instagram. L'objectif à terme est de pouvoir proposer au grand public une interface dynamique permettant de découvrir la Via Francigena à travers les informations diffusées par des utilisateurs d'instagram (commentaires, photos, etc.). Nous traitons dans ces travaux des ES dites simples, aussi nommées entités spatiales absolues (ESA), à la différence des entités spatiales plus complexes intégrant des relations spatiales appelées entités spatiales relatives (ESR) Lesbegueries et al. (2006). La méthode combine une approche morpho-syntaxique à une approche lexicale s'appuyant sur la liste des étapes de l'itinéraire et sur Wikipedia pour identifier les ES dans les messages Instagram. En section 2, nous présentons un rapide état des

1. http ://www.itinerairesculturels.fr

travaux pour l'extraction des ES à partir de données de réseaux sociaux. En section 3, nous présentons notre méthodologie et les premiers résultats obtenus. Nous concluons en section 4 en précisant les perspectives.

2 Travaux connexes

Le travail de Pospescu et al. Pospecu et al. (2009) a pour objectif l'extraction automatique d'informations touristiques à partir de photos de voyages Flikr. Ils s'intéressent plus particulièrement à ce que les gens visitent, la durée de leur visite et les photos panoramiques. L'extraction des ES contenues dans les titres et les tags s'effectue selon une approche lexicale qui se base sur la construction d'un gazetier à partir de Wikipédia et qui associe à chaque nom de lieu ses coordonnées géographiques et son type. Mohamed et Oussalah (2014) présente une recherche similaire en se focalisant surtout sur la catégorisation des entités nommées (personnes, lieux et organisations) en se basant sur les Infobox de Wikipédia. D'autres méthodes permettant l'extraction d'ES telles que la méthode TEXT2GEO Tahrat et al. (2013) qui présente une approche hybride combinant des patterns linguistiques avec une approche d'apprentissage supervisé. La combinaison des deux méthodes améliore significativement la précision et le rappel des résultats. Cependant, le principal inconvénient de la méthode d'apprentissage supervisée réside dans l'exigence de données annotées manuellement, ce qui est plutôt une tâche coûteuse et complexe.

3 Travaux réalisés

Le corpus de départ est un ensemble de 7034 posts multilingues (italien, anglais et français), publiés entre septembre 2011 et janvier 2016, extraits du réseau social Instagram car mentionnant le tag « Via Francigena ». La méthodologie que nous appliquons dans nos travaux se décompose en 2 étapes : Extraction des étapes de la Via Francigena et des ES à partir de la ressource Wikipedia ; Désambiguisation des ES.

Pour **l'extraction des ES**, nous avons combiné une méthode morpho-syntaxique et une approche lexicale s'appuyant sur la liste des étapes de la Via Francigena et la liste d'ES générée à partir d'articles Wikipédia. Pour cela, nous avons sélectionné dans Wikipedia les articles dont les lieux sont géocalisés dans les régions que traverse la Via Francigena. Un article contenant différentes métadonnées (commune, région, province, coordonnées, description, lien de la page) est attaché à chaque ES de la liste obtenue. Le marquage du corpus Instagram avec la liste des étapes et les données Wikipédia permet d'obtenir une liste d'ES candidates à décrire la Via Francigena. Pour chaque ES candidate, une (voir deux) étiquette est associée pour préciser son type en nous appuyant sur la description et sur le nom de l'article provenant de Wikipedia. Nous avons défini pour cela plusieurs motifs s'appuyant sur des lexiques, afin d'identifier cinq types distincts (étapes, monuments, lieux religieux, lieux naturels, autres lieux).

La **désambiguïsation** s'applique aux ES identifiées à partir de la liste Wikipédia qui ne correspondent pas à une étape de l'itinéraire. Nous retenons une ES si elle est localisée dans une région, province ou commune par laquelle passe la Via Francigena. Les coordonnées géographiques extraites des données Wikipédia nous permettent de calculer la distance entre un lieu et une étape. Au regard des expérimentations menées sur le corpus visant à identifier la

distance adéquate entre les étapes de la Via Francigena et les ES candidates (5, 10, 15, 20, 30, 50, 100km), nous retenons les ES qui se trouvent à moins de 15 km d'une étape. Par exemple, l'ES « passo della cisa », identifiée dans 3 posts et caractérisée dans Wikipedia comme appartenant à une région par laquelle passe la Via Francigena, est identifiée comme pertinente car localisée à 6km de l'étape Berceto.

Des **analyses statistiques** résultent de ces premières expérimentations. Tout d'abord, à partir des 7034 posts de départ, 2598 contiennent une ou plusieurs ES pertinente(s). Aussi, 6186 occurrences d'ES sont identifiées, à savoir 3874 pour les étapes et 2312 pour les autres ES pertinentes. Les figures 1 et 2 mettent en avant les ES les plus fréquentes, respectivement les étapes et les ES provenant de Wikipedia.

Un **démonstrateur Web**, développé dans le cadre du projet [2], permet de (1) visualiser l'itinéraire culturel Via Francigena de façon dynamique ; et (2) afficher sur la carte les informations pertinentes relatives aux ES présentes dans Wikipedia. Les liens vers la description plus complète des lieux provenant de Wikipedia sont en cours d'intégration.

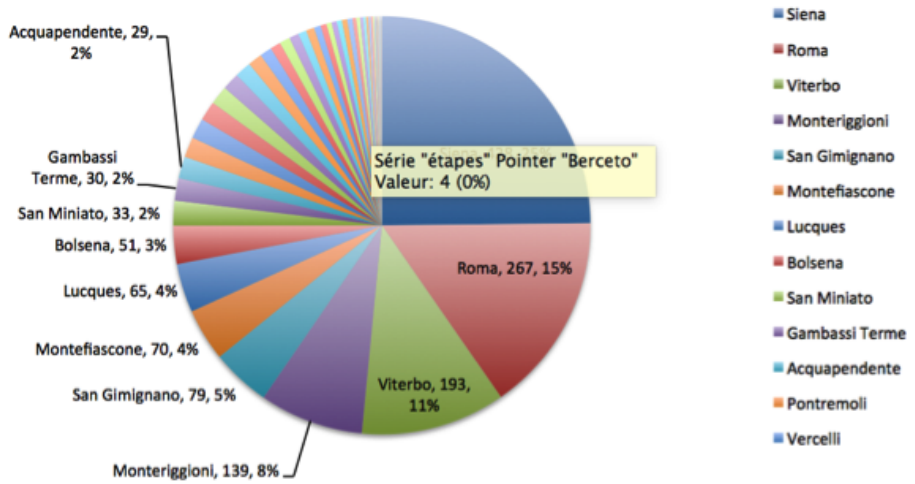

FIG. 1 – *Répartition nb occurrences étapes ViaFrancigena dans les posts Instagram.*

4 Conclusion et perspectives

Nous avons présentés les premiers résultats des travaux menés dans le cadre du projet Itinéraire Numérique visant à valoriser l'itinéraire Via Francigena. Nous remarquons, en analysant les informations relatives à l'itinéraire présentes dans les réseaux sociaux et notamment dans Instagram, que les internautes s'expriment sur le sujet (posts et photos associées). Le démonstrateur est en cours d'amélioration pour ajouter l'accès aux articles Wikipedia, et pour le rendre plus attractif. En perspectives, nous souhaitons étendre l'analyse à Tweeter puis Facebook.

2. Prochainement disponible à l'adresse suivante : http ://geriico-demo.univ-lille3.fr/ItineraireNumerique

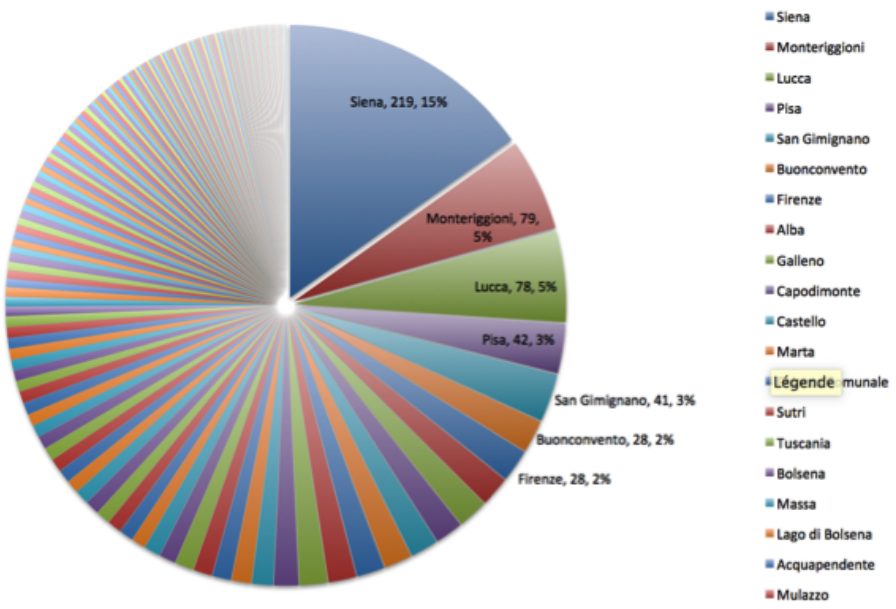

FIG. 2 – *Répartition nombre d'occurrences ES Wikipedia dans les posts Instagram.*

Références

Lesbegueries, J., C. Sallaberry, et M. Gaio (2006). Associating spatial patterns to text-units for summarizing geographic information. *ACM SIGIR 2006. GIR, Geographic Information Retrieval, Workshop, LIUPPA*, 40–43.

Mohamed, A. et M. Oussalah (2014). Identifying and extract named entities from wikipedia database using entity infoboxes. *International Journal of Advanced Computer Science and Applications 5, n7*, 1713–1716.

Pospecu, M., G. Grefenstette, et P.-A. Moellic (2009). Mining tourist information from user-supplied collections. *CIKM '09 Proceedings of the 18th ACM conference on Information and knowledge management*, 1713–1716.

Tahrat, S., E. Kergosien, S. Bringay, M. Roche, et M. Teiseire (2013). Text2geo : from textual data to geospatial information. *WIMS: International Conference on Web Intelligence, Mining and Semantics*, 4.

Summary

As part of the analysis of cultural routes, complex object of the symbolic, social and territorial point of view, the project Digital Route focuses on the analysis of digital content available online for the development of the Via Francigena route. In this article, we propose a semi-

automatic methodology for the extraction of spatial information related to Via Francigena from Instagram data.