Projet TourinFlux : Apport des Technologies du Web Sémantique pour la Gestion des Données du Tourisme

Fayrouz Soualah-Alila, Mickaël Coustaty, Cyril Faucher, Rouaa Wannous

Projet TourinFlux : Apport des Technologies du Web Sémantique pour la Gestion des Données du Tourisme

Fayrouz Soualah-Alila, Mickaël Coustaty, Cyril Faucher, Rouaa Wannous
Laboratoire L3I, Université de la Rochelle
23 avenue Albert Einstein, 17000 La Rochelle, France
{fayrouz.soualah-alila, mickael.coustaty, cyril.faucher, rouaa.wannous}@univ-lr.fr

Résumé : Le tourisme représente aujourd'hui l'un des secteurs d'activité le plus important puisqu'il génère de nombreuses retombées économiques et stimule plusieurs autres secteurs d'activités. Dans un secteur aussi important, la maîtrise des données devient fondamentale : donner une visibilité sur l'activité touristique d'un territoire, connaître les profils des visiteurs, pouvoir anticiper les comportements et adapter l'offre touristique en conséquence, tels sont les défis auxquels sont confrontés les différents acteurs privés et les pouvoirs publics. Or, le secteur du tourisme est caractérisé par un fort afflux de données, souvent complexes, hétérogènes et fortement dynamiques. Dans ce contexte, l'industrie du tourisme doit s'adapter pour mieux gérer, d'une part les données touristiques internes, et d'autre part les données externes disponibles sur le Web. Afin de répondre aux exigences aussi bien d'ordre qualitatives que quantitatives des données, l'exploitation des technologies du Web sémantique permet de répondre aux verrous de l'hétérogénéité des sources d'information et de manipulation de volumes élevés de données. La gestion de diverses données touristiques sémantiquement annotées est facilitée par l'utilisation des ontologies qui fournissent des méthodes et des standards permettant un accès plus intelligent aux données. L'objectif de ce papier est de décrire de façon synthétique comment l'information touristique est gérée dans le cadre du projet TourinFlux. Dans cet article, une architecture pour la gestion des données touristiques est décrite. Cette architecture permet de résoudre plusieurs verrous technologiques: l'hétérogénéité, l'interopérabilité, la réutilisation et la standardisation.

Mots-clés : TourinFlux, TourInFrance, TIFSem, Ontologie, Schema.org, TimeML, SentiML.

1. INTRODUCTION

Le tourisme est considéré comme une industrie à forte production de données et un domaine très dynamique où l'information joue un rôle important pour la prise de décision. En effet, aujourd'hui dans un monde en constante évolution, beaucoup de données liées au tourisme sont produites. Ceci est principalement dû à la forte numérisation du secteur. D'une part, un nombre croissant d'opérations d'achats de produits touristiques (nuits d'hôtels, visites de sites...) se fait par Internet, grâce la démocratisation des services comparatifs dédiés au tourisme, comme par exemple, Booking, TripAdvisor et Yelp. D'autre part, les touristes ont une tendance naturelle à partager leurs opinions et à décrire leurs expériences sur le Web. En même temps, de plus en plus de données sont générées par les capteurs, les téléphones mobiles et les objets connectés.

Ces données qui sont de type Big-Data sont des sources extrêmement riches d'informations. La plupart de ces données pourraient être collectées et utilisées par les décideurs politiques afin d'affecter efficacement les fonds publics pour accroître la fréquentation touristique d'un territoire et augmenter la satisfaction des visiteurs. Seulement ces données sont la plupart du temps non utilisées ou exploitées d'une manière inefficace en raison d'un manque d'outils appropriés. Il devient alors de plus en plus difficile pour les décideurs d'orienter leurs décisions en vue de donner une bonne visibilité à leur territoire, et d'analyser les avis et la

vision des visiteurs. On peut déjà identifier deux principaux problèmes qui émergent lors de la gestion des données touristiques : la massivité des données et leur hétérogénéité.

La massivité se traduit par un accroissement constant et considérable des données à traiter suite à l'émergence de nouveaux usages de ces données principalement induits par le développement du Web. Les données touristiques sont aujourd'hui massivement produites par différents experts (agences de voyages, offices de tourisme...) mais aussi par les visiteurs, constituant ainsi des données sémantiquement hétérogènes, souvent incomplètes et inconsistantes. Ces données peuvent être composées d'informations reliées à des objets touristiques (hôtels, restaurants, évènements...), d'informations temporelles ou encore d'informations sur les avis des visiteurs. L'explosion du nombre de ces informations accessibles via le Web multiplie les besoins en techniques d'intégration de sources de données autonomes et hétérogènes. L'intégration des données est le processus par lequel plusieurs sources de données autonomes, réparties et sous forme hétérogène sont intégrées sous forme de source unique représentée par un schéma global. Il existe déjà différentes taxonomies et catalogues conçus et utilisés par les experts du tourisme pour les aider à gérer en interne des données hétérogènes. Les efforts sont aujourd'hui centrés sur la proposition d'une norme pour faciliter l'intégration et l'échange des données internes, mais aussi des données externes accessibles sur le Web.

D'un autre côté, les données embarquées dans les pages Web sont à l'origine conçus pour être compréhensibles par l'homme, elles sont pour la plupart conservées dans de grandes collections de documents textuels. Face à la croissance intensive du Web en taille et en complexité, le besoin d'automatiser des tâches telles que la recherche de données pertinentes, l'extraction et l'interprétation, se fait ressentir. Dans ce contexte, de nombreux verrous sont à lever à la fois pour la collecte la plus large possible d'informations sur le Web, la normalisation de ces informations, l'analyse, la manipulation, l'échange entre organismes et les interfaces homme-machine correspondantes.

Les nouvelles technologies du Web sémantique offrent des solutions prometteuses à ces différents verrous, solutions devenues aujourd'hui un atout majeur pour l'industrie du tourisme. Le Web sémantique consiste en l'application d'un ensemble de standards promus par le W3C, permettant d'identifier, modéliser, encoder et interroger de gros volumes de données. Le Web sémantique réclame par définition de construire des ontologies permettant, par leur caractère formel, d'automatiser un certain nombre de tâches liées principalement à la recherche d'informations, à la classification et au partage. Dans le cadre de notre travail de recherche, nous pouvons déjà aborder deux cas d'usages de cette nouvelle technologie:

1. Mettre en place une stratégie de modélisation et de recherche d'informations touristiques adaptées aux besoins des professionnels du tourisme;
2. Publier des informations touristiques d'une façon plus riche, et en améliorer l'indexation par les principaux moteurs de recherche du Web.

Le projet TourinFlux, sélectionnée dans le cadre de l'appel à projets Big-Data du Fonds National pour la Société Numérique et financé dans le programme d'investissements d'avenir, rassemble deux entreprises, une association d'entreprises et le laboratoire L3i de l'Université de la Rochelle, et est réalisé en partenariat avec plusieurs acteurs du tourisme de

France. Ce projet vise à apporter aux acteurs du tourisme (d'abord les institutionnels mais aussi les acteurs privés) un ensemble d'outils leur permettant de gérer à la fois leurs données internes et les informations disponibles sur le Web afin de mieux comprendre comment un territoire est perçu et de mieux agir sur cette perception.
Les outils actuellement à disposition des institutionnels du tourisme sont insuffisants pour répondre à ce besoin du fait des problèmes de collecte, d'analyse, de manipulation et d'échange d'informations réalisés de manière beaucoup trop artisanale.
La plateforme du projet TourinFlux rentre dans le cadre de ce contexte où elle répond à un besoin élémentaire qui est de proposer des tableaux de bord adaptatifs aux institutionnels du tourisme qui permettent d'exploiter:

- Les informations concernant les offres touristiques;
- Les informations à propos des visiteurs (fréquentations, profils des visiteurs, mouvements des visiteurs, avis des visiteurs, préférences des visiteurs...);
- Les informations relatives aux entreprises (activités, services, effectifs, horaires...).

Un tableau de bord touristique est un ensemble d'indicateurs de gestion, construit périodiquement, pour un acteur touristique ou un groupe d'acteurs touristiques. Ces tableaux de bord permettent aux institutionnels du tourisme, quelle que soit leur taille, à travers d'exploitation de données enrichies issues des systèmes d'information touristiques (SIT) et du Web, de visualiser et interpréter l'information disponible par rapport à leur territoire dans l'objectif de prendre des décisions plus efficaces.

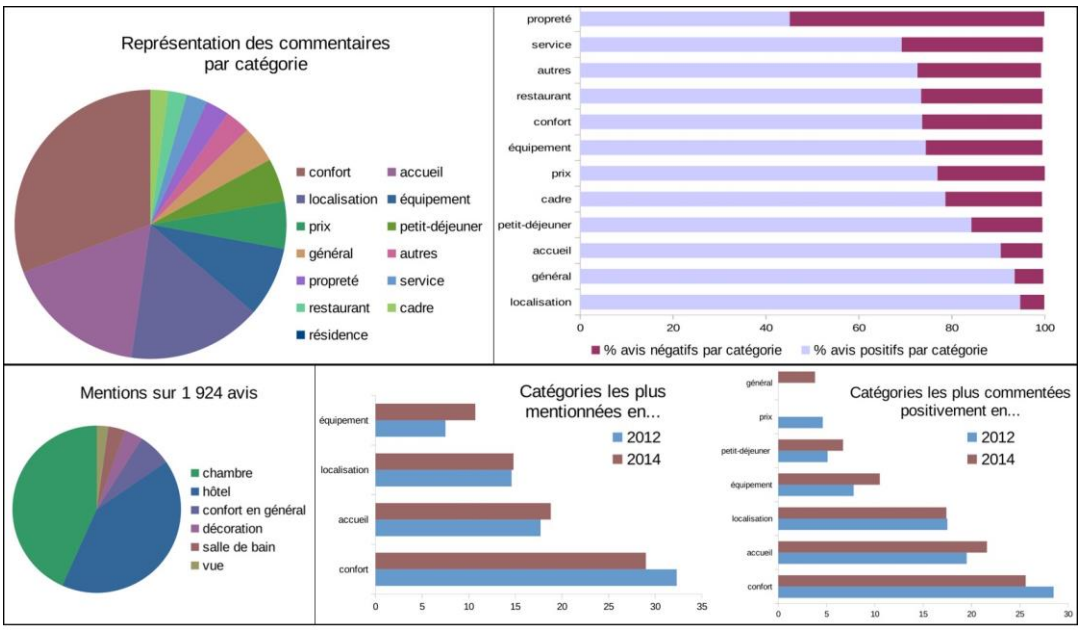

Figure 1. Exemple de tableau de bord

La Figure 1 illustre un exemple de tableau de bord extrait d'un rapport d'information socio-touristique (TourinFlux, 2015).

Aujourd'hui en France, nous pouvons distinguer quatre principaux éditeurs intentionnels de tableaux de bord : les Offices de Tourisme (OT), les Comités Départementales du Tourisme (CDT), la Direction Générale des Entreprises (DGE), l'Institut National de la Statistique et des Etudes Economiques (Insee) et Atout-France. Chacun de ces éditeurs a développé indépendamment diverses méthodes pour évaluer un territoire, et malgré tous les efforts

réalisés jusqu'ici dans l'élaboration de leurs propres tableaux de bord, ceux-ci restent insuffisants pour atteindre les objectifs décrits ci-dessus. En effet (1) ils ne sont pas assez représentatifs de l'activité touristique d'un territoire. Ils sont la plupart du temps focalisés sur les hébergements (offres et fréquentations), mais ne s'attardent pas trop sur les avis et les préférences des visiteurs qu'on peut retrouver sur le Web et qui permettraient d'anticiper une bonne conduite pour le développement des activités touristiques; (2) ils sont limités au niveau du territoire où ils sont développés. En effet il est impossible de générer des tableaux de bord à tous les niveaux hiérarchiques (département, région, national...) ou de faire une comparaison entre les territoires, car les SIT ne sont pas complètements homogènes. Pour l'élaboration des tableaux de bord riches, il est nécessaire d'exploiter le maximum d'informations touristiques afin de refléter une image fidèle du potentiel touristique d'un territoire.

Ce papier décrit les premiers résultats obtenus dans le cadre du projet TourinFlux. Une architecture nommée DataTourism pour la gestion des données touristiques est décrite. Cette architecture est articulée autour d'une ontologie de domaine du tourisme qui permet de résoudre plusieurs verrous technologiques: l'hétérogénéité, la qualité, l'interopérabilité, la réutilisation et la standardisation. Dans la section 2 nous décrivons quels sont les données touristiques et les sources qui nous intéressent et nous décrivons les limites des systèmes existants. La section 3 est dédiée à la description de l'architecture DataTourism pour l'agrégation de données touristiques à partir de différentes sources.

2. E-TOURISME : SOURCES ET TYPES DE DONNEES

2.1. Sources et types de données touristiques

L'industrie du tourisme est par nature fortement basée sur l'échange de données. Dans la dernière décennie, de plus en plus de données sont devenues disponibles pour la recherche et le développement. Ces données proviennent de sources différentes. Les principales sources de données touristiques qui sont à considérer comme une partie de ce projet sont les données disponibles dans les différents SIT, les données disponibles sur le Web et les données Open Data. Ces données pourraient être composées (1) d'informations relatives à des objets touristiques ; (2) d'informations temporelles; et (3) d'opinions.

Notre architecture DataTourism permet de générer des tableaux de bord dynamiques en passant par quatre grandes phases: (1) l'intégration des informations relatives à des objets touristiques dans le système; (2) l'annotation des informations temporelles et d'opinions dans les pages Web; (3) l'enrichissement des objets touristiques par les informations annotées, et enfin; et (4) la génération dynamique de tableaux de bord. Toutes les composantes qui sont utilisés dans chaque phase sont illustrées par la Figure 2.
Les sections suivantes décrivent brièvement comment les objets touristiques sont modélisés et comment les informations temporelles et d'opinions sont annotées dans les pages Web, de manière à compléter la description des objets touristiques.

2.2. Modélisation des objets touristiques - limites actuelles

L'interopérabilité des SIT est un défi majeur pour le développement du tourisme. Plusieurs initiatives institutionnelles nationales, européennes et internationales ont proposé différentes normes pour répondre aux besoins spécifiques des professionnels du tourisme,

mais aucune norme internationale n'a été définie avec succès jusqu'à aujourd'hui (World Tourism Organization, 2004). En France, les acteurs du tourisme institutionnel ont créé la norme TourInFrance (TIF) en 1999 afin de faciliter l'échange de données touristiques. La plupart des SIT comme Raccourci[1], TourinSoft[2] et Sitra[3], ont adopté TIF au début des années 2000. La norme est aujourd'hui utilisée par plus de 3000 offices de tourisme en France, par les Comités Départementales du Tourisme et par différents tours opérateurs, pour faciliter l'échange de données entre ces acteurs. Depuis 2005, cette norme a cessé d'évoluer. En conséquence, les professionnels du tourisme ont adapté la norme à leurs propres besoins (nouveaux tags ajoutés, différentes syntaxes...) et ont proposé chacun leur propre évolution d'une manière inorganisée. Avec l'évolution des technologies du Web et la démocratisation de l'Open Data, cette norme est devenue obsolète et les TIS ont perdu leur inter-compatibilité et ne peuvent plus se partager directement leurs données. Enfin, en l'absence de normes internationales, les informations touristiques restent piégés dans leur propre territoire, il est donc compliqué d'agréger ces informations (Bittner et al., 2005).

En nous basant sur ce constat, nous présentons dans la section suivante un nouveau type de système pour la gestion des données touristiques. Le défi confronté dans ce papier impose deux restrictions majeures. Premièrement, le système conçu doit être capable de modéliser et de structurer des connaissances du domaine du tourisme et ceux du domaine du langage naturel, tout en étant interconnecté avec les systèmes sémantiques existants. Deuxièmement, ce système doit être capable de traiter de gros volumes de données comme les données d'opinion qui sont produites chaque jour sur le Web. Nous proposons donc d'utiliser une combinaison entre des langues d'annotation dédiés pour être en mesure de traiter rapidement de gros corpus, avec les ontologies comme modèle définie de façon claire, abstraite et accepté par toute la communauté du tourisme pour structurer les données et assurer leurs inter-compatibilité. Enfin, ce système peut être facilement lié aux technologies du Web sémantiques afin de faciliter la production de tableaux de bord et l'échange de données.

3. ORGANISATION GENERALE DE LA PLATEFORME PROPOSEE

3.1. Vue d'ensemble de la plateforme

Afin de surmonter les limites présentées précédemment, nous proposons de faire évoluer la norme de TourInFrance pour lui permettre de partager les connaissances qu'elle renferme et également assurer l'interopérabilité des données. Notre plateforme s'articule autour d'un modèle ontologique afin de représenter la terminologie standard du domaine. Ce modèle, initié et validé sur des données françaises, s'appuie sur des outils et une méthodologie issue de la communauté du Web sémantique et est compatible avec l'ensemble des données internationales. Ainsi, toutes les composantes technologiques restent génériques et peuvent être facilement adaptées à d'autres cas d'utilisation.

Le choix d'un modèle basé sur des ontologies repose sur leur capacité à spécifier explicitement une conceptualisation, ou plus exactement, à spécifier dans un langage formel les concepts d'un domaine et leurs relations de manière concrète tout en assurant une

[1] http://www.raccourci.fr/

[2] http://www.tourinsoft.com/

[3] http://www.sitra-tourisme.com/

expressivité assurée au travers d'un vocabulaire partagé (Hirst, 2004). Ainsi, la définition d'une base sémantique commune facilite les échanges et réduit les goulots d'étranglement liés à l'interopérabilité des systèmes (Fodor et Werthner, 2005), tout en assurant une intégration de données hétérogènes. Enfin, l'évolutivité de ce type de systèmes permettra l'insertion de nouvelles données non structurées.

Plusieurs travaux se sont ainsi intéressés à ce sujet dans le domaine du tourisme. Plusieurs ontologies du tourisme sont disponibles et montrent l'intérêt de ce modèle, par exemple : OTA-Open Travel Alliance (OTA, 2000), l'ontologie Harmonise (Dell'Erba et al., 2002), l'ontologie Hi-Touch (Legrand, 2004), l'ontologie QALL-ME (Ou et al., 2008), le catalogue Tourpedia (Cresci et al., 2014), etc. L'inconvénient majeur de ces modèles repose sur le fait qu'ils se concentrent chacun sur des éléments spécifiques du domaine du tourisme, mais aucun d'eux ne traite de l'ensemble du domaine. Aucune vision d'ensemble des données n'est donc possible, ce qui limite les informations à extraire pour la génération de tableaux de bord complets d'un territoire. En l'état actuel, et à notre connaissance, aucune ontologie unique n'existe pour pallier ce problème.

Comme le montre la Figure 2, l'organisation générale de la norme proposée repose sur une structure modulaire qui est en fait composé de trois éléments principaux:

1. Une évolution de TIF en TIFSem afin de stocker des données touristiques dans un format compatible avec les technologies du Web sémantique afin de faciliter le partage et la recherche de données;
2. Une évolution de la norme TimeML afin de l'adapter aux spécificités des données temporelles;
3. Une évolution de la norme SentiML afin d'être en mesure de traiter les données d'opinion.

Le modèle de données que nous proposons possède donc l'avantage de pouvoir intégrer des données hétérogènes tout en assurant leur manipulation et l'extraction de connaissances.

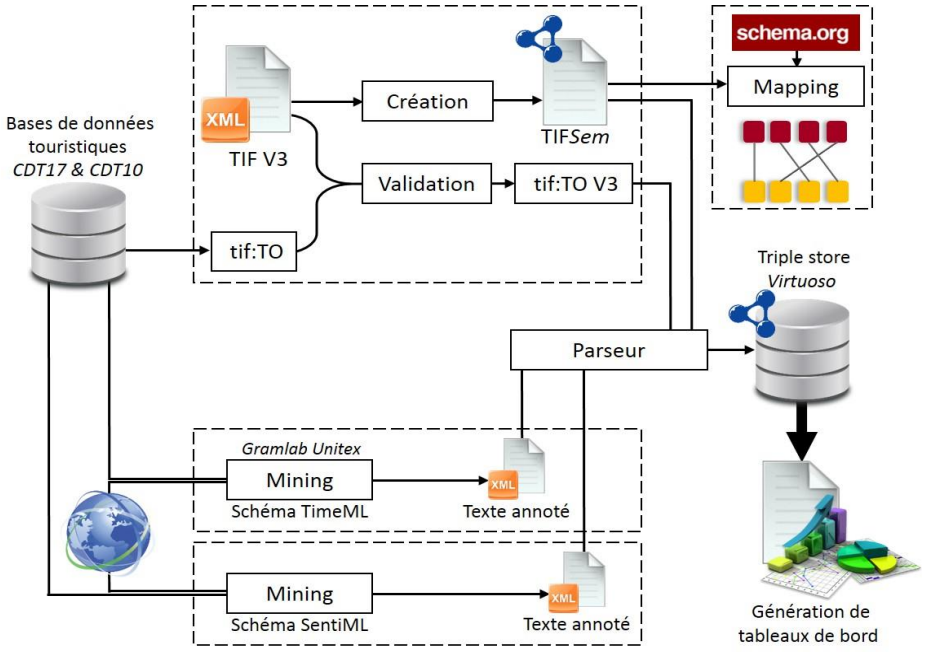

Figure 2. Architecture générale de DataTourism

3.2. Norme pour les données touristiques

Comme mentionné précédemment, aucune norme internationale n'existe aujourd'hui pour échanger des données sur le tourisme et raisonner dessus. La norme que nous proposons doit être en mesure d'offrir deux types de services (l'extraction des connaissances et le raisonnement, et le partage des connaissances), et de traiter des informations hétérogènes (informations textuelles, position GPS, informations sur la météo, informations temporelles...). Une solution est proposée, combinant une ontologie pour la structuration et le raisonnement, avec une formalisation suivant Schema.org[4] reconnu pour sa capacité à partager et à diffuser des connaissances sur le Web.

La première partie du modèle, une ontologie nommée TIFSem (TIF Sémantique) est proposée pour décrire d'une manière globale les objets touristiques en mixant des sources hétérogènes (Soualah-Alila et al., 2015). Nous avons choisi de reconcevoir la norme TIF en TIFSem à des fins de raisonnement. Une ontologie de domaine du tourisme permet d'implémenter des mécanismes de raisonnement déductifs et permet d'assurer l'interopérabilité entre les SIT. Ce modèle permet aussi d'enrichir l'information touristique pour d'un côté proposer des parcours touristiques personnalisés et adaptées aux préférences des visiteurs, et d'un autre côté, du point de vue des experts en tourisme, d'analyser et de mieux gérer les données en ligne sur leur territoire.

Afin d'élaborer l'ontologie TIFSem, nous avons échangé avec différents acteurs du tourisme afin de collecter l'ensemble des concepts spécifiques au domaine. Nous avons travaillé avec des sources provenant du Comité Départemental du Tourisme de la Charente Maritime[5] (CDT17) et du Comité Départemental du Tourisme de l'Aube[6] (CDT10). Nous sommes également en train d'étendre l'ontologie TIFSem en rassemblant des ressources provenant de plusieurs autres fournisseurs de services touristiques en France et à l'International.

Notre deuxième objectif est de veiller à ce que TIFSem soit compatible avec les nouveaux formats de publication des données sur le Web. Vu que TIF est incapable de partager facilement et interagir avec les normes mondiales du Web, nous proposons d'enrichir l'ontologie TIFSem avec le modèle Schema.org. Schema.org est un effort conjoint par Google, et un certain nombre d'autres moteurs de recherche comme Bing et Yahoo. Ce Schéma a été initié il y a quelques années pour essayer d'obtenir plus d'informations sémantiques sur les pages Web. Il préconise essentiellement aux propriétaires de contenu de penser comme les programmeurs en créant des propriétés héritées qu'une machine peut comprendre et se rapporter dessus si c'est indiqué sur votre site Web correctement. En d'autres termes, Schema.org fournit un vocabulaire pour ajouter de l'information au contenu HTML avec un format de micro-données, favorisant le référencement par les grands moteurs de recherche. Lorsque ces micro-données sont utilisées dans une page Web, les moteurs de recherche peuvent mieux interpréter le sens de ses ressources.

Dans le contexte de notre travail, nous proposons de correspondre les termes de TIFSem avec les termes de Schema.org en utilisant des relations sémantiques (Figure 3). De plus, en

[4] https://schema.org

[5] http://www.charente-maritime.org/

[6] http://www.aube-champagne.com/

travaillant avec la communauté Schema.org, nous avons l'intention d'étendre leur modèle, soit formellement en proposant d'ajouter de nouveaux termes Schema.org, ou informellement en définissant comment Schema.org peut être combiné avec quelques termes supplémentaires.

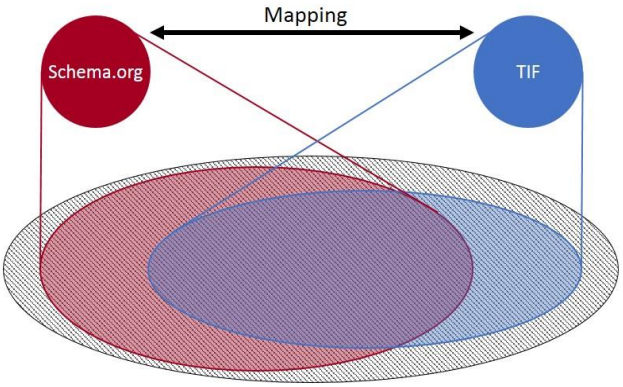

Figure 3. Mapping Schema.org/TIF

Pour alimenter ce modèle, les sections suivantes présentent comment nous avons complété TIFSem avec des techniques de traitement automatique pour l'extraction d'informations basiques, sur le temps et les opinions, liées au domaine du tourisme. Ces techniques permettent d'extraire à partir de ressources des mots clés et des annotations, les insérer dans notre modèle pour finalement déduire de nouvelles connaissances. Dans les deux sections suivantes, nous décrivons comment nous procédons pour compléter la description de TIFSem respectivement par des informations temporelles et d'opinion.

3.3. Données temporelles du tourisme

La première extension du modèle TIFSem repose sur l'intégration de données temporelles dans le corpus. Les données textuelles du tourisme constituent un corpus riche de phénomènes pour l'analyse linguistique. Il suffit de parcourir les sites Web, les bases de données Open Data ou les annuaires pour être frappé par le foisonnement d'entités nommées temporelles. La plupart des objets touristiques sont associés à des évènements et sont caractérisés par des entités temporelles: dates, durées, horaires d'ouverture, horaires de fermeture, conditions d'ouverture, etc. Ces types de données sont parfois toutes présentes dans le même document, tel qu'un calendrier de manifestation. Nous souhaitons annoter les données temporelles présentes dans les documents afin d'enrichir la description des objets pédagogiques dans TIFSem. La reconnaissance automatique des expressions temporelles et d'événements dans un texte en langage naturel est récemment devenue un domaine de recherche actif en informatique linguistique et la sémantique.

Plusieurs travaux de recherche se sont focalisés sur l'annotation temporelle, comme dans les travaux sur TIDES (Translingual Information Detection, Extraction, and Summarization) (Ferro et al., 2001), STAG (Sheffield Temporal Annotation Guidelines) (Setzer, 2001) et encore TimeML (Pustejovsky et al., 2005). Tous proposent visent à fournir un langage de balisage pour l'information temporelle. Dans notre cas, nous optons pour TimeML. Un état de l'art complet sur le sujet est proposé par (Drat, 2014) afin de justifier l'utilisation de ce langage. Dans le domaine du traitement automatique du langage naturel, la norme ISO TimeML s'est imposée comme un standard de facto pour l'annotation de la temporalité. Des

corpus annotés suivant le format TimeML existent ainsi pour différentes langues. Actuellement, afin de couvrir l'essentiel des informations temporelles que l'on peut associer à un texte, le standard prévoit les fonctionnalités suivantes : l'annotation des événements, l'étiquetage des expressions temporelles et la normalisation de leurs valeurs, ainsi que la mise en évidence des relations aspectuelles, temporelles ou de subordination qui peuvent exister entre ces deux types d'entités temporelles. Le standard propose trois types d'entités annotables : (1) les adverbiaux, par exemple date, temps, durée… (TIMEX3); (2) les évènements (EVENT); et (3) les signaux, typiquement des mots pour décrire des relations entre objets temporels, par exemple, quand, pendant, avant, après… (SIGNAL). Comme la présentation de TimeML n'est pas le but de ce papier, nous vous proposons de vous référer au manuel de (Sauri et al., 2009) où une description complète du langage est donnée.

Dans le cadre du projet TourinFlux, afin de faciliter l'extraction de données temporelles, un corpus de pages Web liées au tourisme à est créé. Ce corpus est composé de:

- Un corpus en texte libre décrivant des fêtes et des manifestations, fourni par le Groupe d'Action Local Othe Armance[7]. Ce corpus est disponible sous licence LGPL/LR (Lesser General Public License for Linguistic Resources);
- Un corpus de données touristiques fourni par le Comité Départemental de Tourisme de l'Aube (CD10). Ce corpus contient des descriptions d'objets touristiques: hôtels, restaurants, musées, contenant notamment des horaires et des dates d'ouverture et de fermeture. Ce corpus est multilingue et structuré selon la norme TourInFrance;
- Des données ouvertes, notamment les données concernant les musées nationaux.

La reconnaissance et l'annotation des expressions temporelles constituent naturellement la première étape de tout traitement de la temporalité. L'objectif poursuivi est de baliser le plus complètement et précisément possible un ensemble d'expressions temporelles diverses. Cette étape est réalisée dans notre cas à l'aide d'un ensemble de transducteurs à états finis, développé par Gramlab Unitex[8]. Unitex est un système de traitement et d'analyse de corpus textuels en langage naturel en utilisant des dictionnaires et des grammaires. La détection et l'annotation d'éléments touristiques temporels par la grammaire, dans notre cas TimeML, s'effectue en deux phases principales : La première phase consiste en la préparation et le prétraitement des données (normalisation des caractères spéciaux, déclaration du début et de fin de document, segmentation du texte, déclaration des noms propres...) (Paumier, 2008). Une fois le texte nettoyé, la deuxième phase consiste à utiliser Unitex pour détecter et annoter le texte puis calculer la valeur des attributs pour chacune des étiquettes comme spécifié par TimeML. Pour cela, nous avons utilisé le logiciel GramLab IDEling qui est un environnement ajouté à Unitex dans le cadre du projet GramLab. Les résultats de ce travail sur l'annotation avec TimeML sont présentés dans (Drat, 2014).

3.4. Données d'opinion

L'analyse d'opinions est un domaine de recherche qui se concentre sur l'identification et la classification des opinions dans les données textuelles. Les opinions aident à analyser une situation sur différents aspects et prendre une décision appropriée. L'opinion d'un individu peut influencer l'opinion d'un autre individu, et donc le concept de l'opinion publique est généré. L'opinion publique est très importante dans le domaine du tourisme.

[7] http://www.tourisme-othe-armance.com/

[8] http://www-igm.univ-mlv.fr/~unitex/

La quantité des données d'opinion sur les sites de tourisme a augmenté de façon exponentielle, surtout après l'apparition et la croissance rapide des réseaux sociaux en ligne. Avec la disponibilité et la popularité de riches ressources d'opinion, nous avons besoin de mécanismes fiables pour identifier tous les aspects de l'opinion dans un texte et extraire des informations connexes utiles. Ainsi, nous introduisons le concept d'opinion mining. L'opinion mining, aussi appelé sentiment analysis ou sentiment classification ou encore subjectivity analysis (Cambria et al., 2013), est l'analyse des sentiments à partir de sources textuelles dématérialisées sur de grandes quantités de données (Liu, 2012).

Il existe beaucoup sous-tâches connexes à l'opinion mining, comme l'annotation sémantique des opinions. Cette tâche a fait l'objet de plusieurs études dans le but de déterminer et de représenter les descripteurs pertinents dans ce cadre d'extraction de connaissances à partir des expressions d'opinion. Certains modèles d'annotation ont été proposés par la communauté de recherche. Nous distinguons trois principaux modèles d'annotation d'opinions: SentiML (Di Bari et al., 2013), OpinionMining-ML (Robaldo and Caro, 2013) et EmotionML (Schröder et al., 2011). Dans le cadre du projet TourinFlux, une étude comparative de ces schémas d'annotation a été proposée par (Malik et al., 2014).

Dans notre cas nous nous sommes intéressés à SentiML pour l'annotation des données d'opinion. Dans SentiML nous parlons de sentiments plutôt que d'opinions. L'analyse des sentiments consiste à analyser une grande quantité de données pour déterminer les opinions ou les sentiments exprimés dans les textes. Développée dans les années 2000, l'analyse des sentiments est utilisée pour détecter les opinions des utilisateurs sur des sujets divers. Le but est alors d'attribuer une polarité (positive, négative ou neutre) à des opinions présentes dans les documents (commentaires des internautes, des forums ou plus récemment sur les réseaux sociaux). SentiML est basé sur Appraisal Framework (AF) qui est une forte théorie linguistique. Les résultats de ce travail sur l'annotation avec SentiML sont présentés dans (Malik et al., 2014).

4. CONCLUSION

Le travail présenté dans cet article concerne les premiers résultats du projet TourinFlux, pour identifier un nouveau modèle pour la gestion des données du tourisme. Celui-ci possède l'avantage de pouvoir intégrer des données hétérogènes tout en assurant leur manipulation et l'extraction de connaissances nouvelles. L'architecture proposée est formée de trois principales composantes: l'évolution de la norme TourInFrance vers un format compatible avec les technologies du Web sémantique; l'extraction et l'annotation de données temporelles à partir de corpus avec le standard TimeML; et l'extraction et l'annotation des informations d'opinions à l'aide du standard SentiML.

BIBLIOGRAPHIE

TourinFlux (2015). Ville de Rouen: Rapport d'informations socio-touristiques.

Bittner, T., Donnelly, M. & Winter, S. (2005). Ontology and Semantic Interoperability. Zlatanova, S & Prosperi, D. (Eds), *Large-Scale 3D Data Integration: Challenges and Opportunities*, 139-160.

Cambria, E., Schuller, B., Xia, Y. & Havasi, C. (2013). New Avenues in Opinion Mining and Sentiment Analysis. *IEEE Intelligent Systems*, 28(2), 15-21.

Cresci, S., D'Errico, A., Gazze, D., Duca, A. L., Marchetti, A. & Tesconi, M. (2014). Towards a dbpedia of Tourism: the Case of tourpedia. *International Semantic Web Conference*.
Dell'Erba, M., Fodor, O., Ricci, F. & Werthner, H. (2002). Harmonise: A Solution for Data Interoperability. *Towards the Knowledge Society: eCommerce, eBusiness, and eGovernment, the Second IFIP Conference on E-Commerce, E-Business, E-Government*, 433-445.
Di Bari, M., Sharoff, S., Thomas, M. (2013). Sentiml: Functional Annotation for Multilingual Sentiment Analysis. *1st International Workshop on Collaborative Annotations in Shared Environment: Metadata, Vocabularies and Techniques in the Digital Humanities*, DH-CASE'13, ACM, New York, USA, 15:1-15:7.
Drat, L (2014). Projet TourInFlux. Annotation des Expressions Temporelles.
Ferro, L., Mani, I., Sundheim, B., & Wilson, G. (2001). TIDES Temporal Annotation Guidelines, Version 1.0.2. *The MITRE Corporation, McLean, Virginia. Report MTR 01W0000041.*
Fodor, O. & Werthner, H. (2005). Harmonise - a Step Towards an Interoperable e-Tourism Marketplace. *International Journal of Electronic Commerce*.
Hirst, G. (2004). Ontology and the lexicon. Staab, S. & Studer, S. (Eds), *Handbook on Ontologies*: Springer-Verlag, 209-229.
Legrand, B. (2004). Semantic Web Methodologies and Tools for Intra-European Sustainable Tourism. White paper, Paris, Mondeca.
Liu, B. (2012). Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*, Morgan & Claypool Publishers.
Malik, M., Missen, S., Attik, M., Coustaty, M., Doucet, A. & Faucher, C. (2014). SentiML ++: An Extension of the SentiML Sentiment Annotation Scheme. *The 12th Extended Semantic Web Conference (ESWC2015)*.
Ou, S., Pekar, V., Orasan, C., Spurk, C. & Negri, M. (2008). Development and Alignment of a Domain-Specic Ontology for Question Answering. Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odjik, J., Piperidis, S. & Tapias D. (Eds), *the Sixth International Language Resources and Evaluation Conference*, 2221-2228.
OTA (2000). Opentravel Alliance. Opentravel Alliance message specications. Specications Document.
Paumier, S. (2008). UNITEX 2.0: User Manual.
Pustejovsky, J., Ingria, B., Sauri, R., Castano, J., Littman, J., Gaizauskas, R., Setzer, A., Katz, G. & Mani I. (2005). The Specification Language TimeML. *The Language of Time: a Reader,* 545-557.
Robaldo, L. & Caro, L. D. (2013). OpinionMining-ML. *Computer Standards & Interfaces*, 35 (5), 454-469.
Sauri, R., Goldberg, L., Verhagen, M. & Pustejovsky, J. (2009). Annotating Events in English, TimeML Annotation Guidelines, Version TempEval-2010.
Schröder, M., Baggia, P., Burkhardt, F., Pelachaud, C., Peter, C. & Zovato, E. (2011). Emotionml - an Upcoming Standard for Representing Emotions and Related States. *Affective Computing and Intelligent Interaction*, Springer.
Setzer, A. (2001). Temporal Information in Newswire Articles: an Annotation Scheme and Corpus Study. Ph.D. thesis, University of Sheffield, Sheffield, UK.
Soualah-Alila, F., Faucher, C., Bertrand, F., Coustaty, M. & Doucet A. (2015). Applying Semantic Web Technologies for Improving the Visibility of Tourism Data. *CIKM'15 Workshop on Exploiting Semantic Annotations in Information Retrieval ESAIR'15*.

World Tourism Organization (2004). Information and Documentation Resource Centres for Tourism: Guidelines for Establishment and Maintenance. ISBN: 9284407176, 9789284407170, 132.