Recital 2008, Avignon, 9-13 juin 2008

Informations spatio-temporelles et objets touristiques dans des pages Web : repérage et annotation

Stéphanie Weiser

MoDyCo, UMR7114, CNRS – Université Paris X, 200 av. de la République, 92001 Nanterre

Résumé. Cet article présente un projet de repérage, d'extraction et d'annotation d'informations temporelles, d'informations spatiales et d'objets touristiques dans des pages Web afin d'alimenter la base de connaissance d'un portail touristique. Nous portons une attention particulière aux différences qui distinguent le repérage d'information dans des pages Web du repérage d'informations dans des documents structurés. Après avoir introduit et classifié les différentes informations à extraire, nous nous intéressons à la façon de lier ces informations entre elles (par exemple apparier une information d'ouverture et un restaurant) et de les annoter. Nous présentons également le logiciel que nous avons réalisé afin d'effectuer cette opération d'annotation ainsi que les premiers résultats obtenus. Enfin, nous nous intéressons aux autres types de marques que l'on trouve dans les pages Web, les marques sémiotiques en particulier, dont l'analyse peut être utile à l'interprétation des pages.

Abstract. This paper presents a project for the detection, extraction and annotation of temporal and spatial information and of tourism objects in order to fill the knowledge base of a tourism Web portal. We focus on the differences that exist between extraction from structured documents and extraction from Web pages. First, the different types of information to extract are presented. We then discuss methods for linking these pieces of information together – for example relating the name of a restaurant to its opening hours – and how to annotate the extracted data. The program we have developed to perform the extraction and annotation, as well as an evaluation of this program, are presented here. Finally, we focus on the semiotic marks which appear on the Web and show they also prove useful in interpreting Web pages.

Mots-clés : extraction d'information, annotation, informations spatio-temporelles, tourisme, pages Web.

Keywords: information extraction, annotation, spatial & temporal information, tourism, Web pages.

1 Introduction

Avec les méthodes du Web Sémantique, des portails applicatifs reposant sur des ontologies peuvent être créés. Pour ces applications ainsi que pour de nombreux services proposés sur le Web, les informations temporelles et spatiales sont souvent primordiales. Par exemple, un portail touristique sur Internet nécessite des informations sur différents objets touristiques, leur type et leur localisation spatio-temporelle. Par ailleurs, une base de connaissance, calquée sur l'ontologie utilisée par le portail, devrait être capable de stocker ces informations.

Dans cet article, nous nous focaliserons sur les informations temporelles et spatiales contenues dans des pages Web touristiques. Ces informations doivent être détectées, extraites et ensuite annotées. Le format d'annotation est basé sur des outils XML préexistants (Stern, 2007). Trois grands types de difficultés ont vu le jour pour extraire automatiquement ces informations. Premièrement, des informations temporelles ou spatiales complexes et imprécises doivent pouvoir être traitées. Bien sûr, les dates simples sont faciles à repérer, mais nous avons également besoin d'extraire des expressions plus complexes et de les catégoriser, comme des périodes ou des informations répétitives. Deuxièmement, une fois les informations extraites, elles doivent être liées aux bons objets du domaine. Si la page Web ne mentionne qu'un seul objet touristique, le lien est direct : l'information extraite fait référence à cet objet. Par contre, il arrive fréquemment qu'une page Web concerne plusieurs objets touristiques. Une analyse plus fine est alors nécessaire pour créer des références entre les objets et les informations extraites. Troisièmement, nous cherchons à traiter des pages Web d'un certain type (ce sont toutes des pages touristiques), mais celles-ci sont faites par différents acteurs, dans des formats hétérogènes. Elles varient donc beaucoup. Nous essayerons de montrer que certaines spécificités nécessiteraient l'étude de marques sémiotiques.

Le travail présenté dans cet article se situe dans le cadre du projet Eiffel (ANR / RNTL) dont l'objectif est de créer, sur le Web, un portail touristique comprenant de nombreuses fonctionnalités. Ce portail inclut un moteur de recherche spécialisé et a pour but de permettre à ses utilisateurs de trouver et stocker des informations essentielles et précises en contexte. Par ailleurs, les régions françaises pourront, grâce à ce portail, promouvoir leurs services. Il s'agit d'un vaste projet (Noël & al., 2008), basé sur les technologies du Web Sémantique, sur la représentation de la connaissance et sur des méthodes et expertises linguistiques. Il inclut la détection, l'extraction et l'annotation automatique de différents types d'informations sur le Web ; tâches pour lesquelles il s'appuie sur une ontologie du territoire, créée pour cette application, par des experts.

Un corpus de plus de 5000 pages Web francophones a été constitué automatiquement par l'un de nos partenaires pour les besoins de notre étude. Ces pages sont toutes liées au domaine du tourisme : sites d'hôtels, de restaurants, d'événements ponctuels, spectacles, concerts, etc. Ces pages ont été transformées en documents XML, format plus adéquat pour un traitement automatique, et « nettoyées ». C'est-à-dire que seules les informations utiles à notre analyse ont été conservées. De nombreuses balises ont donc été supprimées. Certaines, au contraire, ont été précieusement conservées, notamment les balises donnant des indications sur la mise en forme des pages : couleur, disposition, tableaux, etc. Dans le corpus, chaque page est indépendante : il se peut que plusieurs pages soient issues du même site mais, une fois l'aspiration effectuée, elles ne sont plus liées.

Cet article est organisé de la manière suivante : la partie deux présente les différences qui distinguent le repérage et l'extraction d'information dans une page Web ou dans un document structuré ainsi que les différentes informations que l'on cherche à extraire et annoter. On y

trouve ensuite une classification des différentes informations que l'on cherche à extraire, ainsi que des précisions sur la façon dont on peut les lier et les annoter. La partie trois présente le système que l'on a mis au point pour effectuer cette tâche de repérage et d'annotation et son évaluation.

2 L'extraction d'information dans des pages Web

Le repérage et l'extraction d'information forment un des domaines du traitement automatique des langues qui a déjà été largement étudié. Les travaux d'extraction d'informations temporelles ou calendaires (Battistelli & al., 2006) et d'informations de localisation spatiale (Piton & Maurel, 2004) sont donc déjà répandus. Mais ils ne concernent pas des textes dont l'écriture et la structuration découlent en partie de la forme des pages Web. Nous cherchons à montrer que l'extraction d'un même type d'informations est différente selon qu'elle est effectuée dans une page Web ou dans un document structuré.

Dans les guides touristiques classiques (papier), l'information est beaucoup plus structurée et standardisée que dans des pages Web touristiques, qui sont très variées et ne respectent pas un modèle de page commun. Dans un guide, toutes les entrées correspondant au même type d'objet touristique sont structurées de la même manière et contiennent le même type d'informations. Des marqueurs linguistiques peuvent alors permettre de repérer chaque information à extraire. De plus, la ponctuation est très rigoureuse, ce qui facilite grandement l'interprétation, surtout si celle-ci est automatisée. Les points et points-virgules sont bien utilisés : les points séparent les informations qui sont de natures différentes tandis que les points-virgules séparent les différentes informations de même type (les différents prix par exemple). Dans les pages Web, au contraire, la ponctuation – s'il y en a – est rarement utilisée à bon escient. La plupart du temps, ce sont des espaces blancs ou des « lignes sautées » qui servent de séparateur.

Bien sûr, l'extraction d'information dans des documents structurés n'est pas triviale, mais ce que nous voulons souligner, c'est que pour ce type de texte, il existe déjà de nombreux outils syntaxiques permettant de faciliter l'extraction automatique : par exemple, des outils d'analyse syntaxique partielle, de chunking ou de parsing sont disponibles. Il subsiste toutefois des difficultés, comme la résolution d'anaphores, qui rendent l'analyse automatique de textes complexe. Ce qui nous intéresse, ce sont les difficultés que l'on peut rencontrer lors de l'analyse automatique de pages Web et qui sont très différentes des difficultés que présente une simple analyse automatique de texte. Par exemple, dans L*e guide du routard*[1], on trouve pour chaque restaurant son nom suivi de son adresse, numéro de téléphone et station de métro puis les indications d'ouverture comme *ouvert tous les jours.* Chaque entrée est facile à analyser automatiquement. Le problème est que, sur une page Web, le même type d'information prendrait plutôt la forme présentée sur la page Web de la figure 1.

Pour l'instant, aucun outil n'est capable d'interpréter que *Chez Paul et Bernadette Colas* est le nom d'une auberge ou que *Ouvert du 15 février au 30 novembre* correspond aux informations d'ouverture de cette auberge. C'est pour cela que les méthodes d'extraction d'information dans des documents structurés ne peuvent pas être directement appliquées à des pages Web dans

[1] *Le guide du routard – Paris balades* 2007 p.231 : « Le Saint-Amour 2, av.Gambetta, 75020. 01-47-97-20-15. Métro Père-Lachaise. Ouvert tous les jours »

lesquelles la structure n'est pas standardisée et où, en plus d'une ponctuation presque inexistante, la syntaxe n'est souvent pas conforme à la norme.

Figure 1 : extrait d'une page Web

Notre travail est proche de (Tenier & al., 2006) dont l'objectif est d'extraire, dans des pages Web, les informations concernant des équipes de recherche. Notre approche est plus générale dans le sens où les pages que ces auteurs analysent respectent des règles de disposition spatiale. Nous nous rapprochons également de (Bry & al., 2003) qui ont travaillé sur des modules de raisonnement temporel et de localisation pour des pages Web. Mais la différence entre leur travail et le nôtre est que ces auteurs travaillent sur des documents XML dans lesquelles l'information est déjà sémantiquement structurée ; les expressions temporelles sont déjà marquées et n'ont donc pas à être repérées.

Par ailleurs, TimeML[2] est un langage de marquage des expressions temporelles et d'événements. Nous avons choisi de ne pas nous appuyer sur ce langage car, pour nos besoins, il est trop riche et pas assez spécifique. En effet, il permet d'annoter en détail des expressions temporelles, de les ordonner et de faire certains raisonnements mais ces expressions sont trop nombreuses et variées pour nous. Les informations que nous cherchons à annoter sont spécifiques au domaine du tourisme et ont donc des caractéristiques propres qu'il serait dommage d'ignorer.

2.1 Informations à extraire

Lorsqu'un utilisateur effectue une recherche sur le portail en vue de planifier un voyage, il porte un intérêt tout particulier aux informations spatio-temporelles : que peut-il trouver aux alentours de son lieu de vacances, le musée de la ville voisine est-il ouvert à cette période de l'année, y a-t-il un événement particulier à ne pas manquer, etc.? C'est pour cela qu'il est important de rechercher et d'extraire ces informations des pages Web touristiques. Par ailleurs, au niveau de la chaîne de traitement, ces informations tiennent une place dans la base de connaissance du projet. Cette base est calquée sur une ontologie du territoire qui modélise à la fois l'offre (objets touristiques) et l'approche marketing du territoire. Pour les objets touristiques, un travail de représentation des informations temporelles et spatiales a été intégré

[2] http://www.timeml.org

dans l'ontologie. Ainsi, on y trouve les concepts d'heure d'ouverture et de fermeture, de date mais aussi de périodicité. Le repérage d'informations dans les pages Web est donc guidé par les besoins de cette ontologie en vue de pouvoir ensuite effectuer des requêtes pertinentes sur ces données.

2.1.1 Les informations temporelles

Tout d'abord, voici quelques exemples des expressions que l'on cherche à repérer :

- *Le 21 avril*
- *Du 10 au 21 mars*
- *Du lundi au vendredi, 9h – 11h*
- *Inauguration du musée le 7 juillet 2006*
- *En Juillet et Août ouvert tous les jours. Hors cette période, fermeture le mercredi soir, le jeudi toute la journée et le dimanche soir. Horaires d'ouverture: de 12h00 à 14h00 de 19h00 à 22h00.*

Comme ces exemples le montrent, les entités temporelles à repérer sont de différents types. Deux catégories principales peuvent être identifiées : les informations temporelles qui concernent un événement particulier et les informations temporelles répétitives. La première comprend des dates (*concert le 1er octobre*), des périodes (*festival de mai à juin*), des heures (*le concert commence à 8h*). La seconde comprend des horaires (*le musée ouvre à 10h*), des périodes (*le restaurant est ouvert du lundi au samedi*) et des exceptions (*le camping est ouvert toute l'année sauf en janvier*). Des exemples d'une complexité plus grande peuvent également prendre place dans cette classification comme *de mai à juin, ouvert tous les jours sauf le mardi*.

Certaines informations temporelles ne nous intéressent pas car elles ne concernent pas un objet touristique. Elles ne doivent pas être repérées. Par exemple, une date seule, non introduite par une marque lexicale du type *date d'ouverture* ne concerne pas toujours un objet touristique. Les dates isolées concernent souvent la page Web elle-même comme dans : *Dernière modification : 25/10/2006*. Rappelons que les informations temporelles que l'on cherche à identifier sont celles qui concernent directement un objet touristique en particulier.

De plus, certaines expressions peuvent être ambiguës. Par exemple dans *vendredi et samedi soir*, doit-on comprendre qu'il s'agit des deux soirées ou de la journée de vendredi et de la soirée de samedi ? Le contexte est nécessaire pour lever ce type d'ambiguïté. Pour cet exemple, s'il s'agit d'un concert, la première interprétation sera probablement la bonne.

2.1.2 Les informations spatiales

Voici quelques exemples des expressions de localisation que l'on cherche à repérer :

- *8, rue de l'église, 58000 Nevers*
- *Office de tourisme 58230 Brisson*

- *Le champ Thierry 58170 Luzy*
- *13, rue Henri Renaud - BP 5 58360 Saint-Honoré-les-Bains*
- *Au centre du village de Tamnay en Bazois (D978 direction Nevers)*
- *Quai Jules Moineau B.P. 123 58206 Cosne Cours Sur Loire CEDEX*

Le repérage des informations spatiales s'est pour l'instant focalisé sur les adresses prototypiques comme *8, rue de l'église, 58000 Nevers* et quelques unes de leurs variantes (pas de numéro, une information de boîte postale etc.). Le schéma de ce type d'adresse correspond à un numéro, un marqueur lexical indiquant le type de voie, une suite de mots indiquant le nom de la voie, puis un numéro pour le code postal et enfin un nom de ville. Nous disposons d'un lexique de noms de ville qui sert à déclencher des analyses contextuelles.

Un travail plus sémantique qui concernera des informations de localisation est nécessaire et prévu, pour repérer les informations du type *Au centre du village de Tamnay en Bazois (D978 direction Nevers)*. Pour ce type d'expression, il faudra alors déterminer plus précisément ce que l'on cherche à extraire et de quelle manière cela peut prendre place dans l'ontologie.

2.1.3 Les objets touristiques

Le repérage des informations temporelles et spatiales est donc primordial dans le cadre d'une application touristique, mais ces informations n'ont de sens que si l'on sait à quel objet touristique chacune se rapporte. Il est alors nécessaire de nous intéresser au repérage de ces objets touristiques.

Tout d'abord, voici ce que nous entendons par « objet touristique » : un objet touristique peut correspondre à une activité (concert) ou à une infrastructure (hôtel). Cette terminologie est calquée sur ce que l'ontologie du projet modélise, afin de faciliter ensuite la communication avec la base de connaissance. Tout ce qui peut faire l'objet d'une page dans notre corpus et avoir des informations temporelles et/ou spatiales associées peut être considéré comme un objet touristique : cela englobe aussi bien les lieux de type hôtel, restaurant, auberge, maison d'hôte, etc. que les événements de type concert, festival, fête, marché, etc.

Ces objets touristiques sont parfois clairement énoncés dans le contexte direct d'une information temporelle (*le concert commence à 20h*) et sont donc assez simples à extraire. Mais ce n'est pas toujours le cas et il faut alors procéder à une recherche plus approfondie dans le reste de la page, notamment dans le titre.

2.1.4 Appariement des informations repérées

Comme on vient de le voir, une information temporelle ou spatiale isolée est difficilement exploitable. En effet, l'expression ne prend son sens que lorsqu'elle correspond à un objet touristique.

On distingue principalement deux cas de figure : soit la page Web ne concerne qu'un seul objet touristique, soit elle fait référence à plusieurs objets. Dans le premier cas, l'opération d'appariement est donc relativement simple : on peut faire l'hypothèse que toutes les informations spatiales ou temporelles de la page concernent l'unique objet touristique mentionné. Dans le second, une analyse plus fine est nécessaire pour établir les relations entre

les différentes informations repérées. Différents critères sont exploitables pour lier les entités repérées les unes aux autres : critère de proximité, d'ordre, critères sémiotiques prenant en compte la disposition de la page, titres, couleurs, etc.

2.2 Annotation des expressions extraites

Afin d'être exploitables, les informations extraites doivent être intégrées dans la base de connaissance du projet. Elles doivent alors correspondre aux besoins de cette base, elle-même calquée sur l'ontologie du territoire présentée plus haut. Elles doivent donc être annotées en fonction de ce que l'ontologie permet de modéliser. Le format XML a été choisi pour effectuer cette tâche.

2.2.1 Une annotation en deux étapes

L'annotation se fait en deux temps : dans un premier temps, l'expression est repérée et délimitée et dans un second temps, les composants des expressions repérées sont identifiés, notamment la granularité pour les expressions temporelles. Ainsi, seules quelques balises générales sont d'abord insérées : il s'agit d'une balise <UT> qui encadre toute expression repérée et de balises plus spécifiques comme <HeureOuverture>, <dateOuverture>, <periodeOuverture>, <Adresse> ou <Infrastructure>. La seconde étape de l'annotation permettra d'insérer des balises plus précises et par exemple de transformer une période en <DateDebut> et <DateFin> ou d'ajouter les balises <Ville> ou <CodePostal> pour la spatialisation.

Voici quelques exemples d'expressions annotées lors de la première étape :

- <UT> <dateOuverture>10 février </dateOuverture>au<dateFermeture>21 mars </dateFermeture> </UT>

- <UT> <periodeOuverture> du lundi au vendredi <heureOuverture> 9h-11h </heureOuverture> </periodeOuverture> </UT>

- <Localisation><Adresse>34 Rue Saint Gildard 58000 NEVERS</Adresse></Localisation>

Les composants des expressions seront ensuite identifiés. Pour les expressions temporelles, il s'agit des détails de jour, mois et année. Pour les localisations, les villes et codes postaux seront identifiés. Lors de l'extraction d'information, on a choisi de s'en tenir à la tâche d'annotation, sans effectuer de raisonnement. Par exemple l'expression *ouvert tous les jours sauf le mardi* ne sera pas convertie en *ouvert le lundi, le mercredi, le jeudi, etc*. Ce raisonnement sera traité au niveau de la base de connaissance ou au niveau des requêtes.

2.2.2 Nomenclature des balises d'annotation

Six balises permettent d'annoter les informations temporelles : <UT> délimite l'expression repérée, les autres balises sont plus précises. <dateOuverture> et <dateFermeture> permettent d'indiquer le début et la fin d'un événement, <periodeOuverture> et <periodeFermeture> servent à annoter les informations périodiques comme dans *ouvert tous les jours*. Les balises <dateOuverture> et <dateFermeture> servent donc à annoter les informations ponctuelles, même si, dans la langue, on pourrait considérer qu'il s'agit d'une période (pour *du 1er au 30*

décembre par exemple) tandis que les balises <periodeOuverture> et <periodeFermeture> indiquent réellement un phénomène de périodicité. Les exceptions, comme *ouvert toute l'année sauf en juillet* sont marquées par la balise <exception>.

Des balises plus fines seront ajoutées pour effectuer la deuxième étape d'annotation. Il s'agira très probablement de <jour>, <mois> et <année> pour décomposer les dates ; de <jourOuverture>, <jourFermeture>, <moment> (moment de la journée comme *mardi midi*), <heure>, <minute> pour décomposer les périodes.

Les balises <activité> et <infrastructure> sont les deux premières qui permettent d'annoter les objets touristiques. D'autres comme <musée> ou <hôtel> viendront compléter ce jeu.

En ce qui concerne les lieux, la balise <localisation> encadre l'expression repérée. La balise <adresse> indique qu'il s'agit d'une adresse prototypique. Les balises <ville> et <codePostal> viendront compléter les annotations.

3 Système de repérage, d'extraction et d'annotation

C'est sur une approche symbolique, reposant sur des patrons linguistiques, que notre système est basé. L'architecture générale de ce système est présentée dans cette partie, tout comme une première évaluation du module de repérage des informations temporelles.

3.1 Architecture du système

Le système que l'on a créé pour effectuer le repérage et l'annotation des pages Web en XML est composé de plusieurs modules. Un module dédié pour chaque type d'information et un module qui coordonne le tout. Celui-ci est développé en JAVA, les modules dédiés avec l'outil Unitex[3]. Cet outil permet de traiter des corpus en utilisant des dictionnaires (basés sur les tables du LADL[4]) ; et ce au niveau du lexique, de la syntaxe ou de la morphologie. Il permet de repérer et de baliser des structures correspondant à des expressions régulières, représentées par des graphes à états finis. La sortie d'Unitex est stockée dans un fichier texte dans lequel des balises d'annotations, au format XML, ont été ajoutées pour marquer les données identifiées.

Il existe ainsi trois modules dédiés. Le premier sert à repérer et annoter les informations temporelles et il se compose donc de transducteurs créés à l'aide d'Unitex. Le repérage et l'annotation se font à l'aide du même jeu de transducteurs. Le deuxième et le troisième qui concernent le repérage des informations spatiales et des objets touristiques sont en cours de développement.

---

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

[4] Laboratoire d'Automatique Documentaire et Linguistique – les tables ont été créées au LADL par Maurice Gross et contiennent des unités lexicales classées selon des propriétés syntaxiques et distributionnelles

3.2 Evaluation

Une première évaluation (rappel et précision) de l'opération de repérage des informations temporelles (expressions simples et complexes) est présentée ici. Cette évaluation est biaisée car elle a été effectuée sur les cent pages qui ont permis de constituer les transducteurs de repérage. La prochaine étape consistera à appliquer ces transducteurs sur cent nouvelles pages mais cette démarche est coûteuse en temps car, pour calculer les taux de rappel et précision, il faut parallèlement analyser les cent pages manuellement.

	Ens. 1	**Ens. 2**	**Ens. 3**	**Ens. 4**	**Ens. 5**	**Total**
Rappel	7/7	9/9	6/7	18/20	7/8	47/51
	100%	100%	85,7%	90%	87,5%	92,1%
Précision	7/8	9/14	6/6	18/23	7/13	47/64
	87,5%	64,3%	100%	78,2%	53,8%	73,4%

Tableau 1 : Taux de rappel et précision pour les cinq ensembles

Le tableau 1 contient les taux de rappel et précision pour cinq ensembles de vingt pages qui constituent le corpus de test (les cent pages ont aléatoirement été réparties en cinq ensembles) ainsi qu'une moyenne sur les cent pages. Mais pour que ces chiffres prennent toute leur valeur, d'autres données chiffrées sont à prendre en compte. En effet, ce corpus a été constitué de manière tout à fait aléatoire et seules 26 pages contiennent des informations temporelles. Le nombre d'expressions temporelles à extraire est de 51. Par ailleurs, 16 pages ne sont pas en français et 12 sont vides ou inexploitables (cryptées, tout en majuscule, etc.).

Ces premiers résultats sont encourageants, mais seule une évaluation rigoureuse d'un nouvel ensemble de pages permettra de confirmer la pertinence de nos transducteurs. Une évaluation du repérage et de l'annotation des informations spatiales et des objets touristiques reste également à effectuer.

4 Conclusion et perspectives

Nous avons entrepris de repérer, extraire et annoter les informations temporelles, les informations spatiales et les objets touristiques dans des pages Web touristiques afin d'alimenter la base de connaissance d'un portail touristique. Nous avons souligné la différence entre l'extraction d'information dans des documents structurés et l'extraction sur des pages Web.

Après avoir introduit et classifié les différentes informations à extraire, nous avons fait le point sur la façon de lier ces informations entre elles et de les annoter. Nous avons également présenté le logiciel que nous avons réalisé afin d'effectuer cette opération d'annotation et les premiers résultats obtenus. Notre travail étant plus avancé dans le repérage des expressions temporelles, c'est la seule opération pour laquelle nous avons procédé à une évaluation. L'évaluation de l'annotation des expressions spatiales reste à venir et la réalisation des transducteurs pour le repérage des objets touristiques est en cours. De plus, il serait très intéressant de pouvoir exploiter les différentes marques non textuelles que l'on trouve dans les pages Web : la disposition, la structure et la mise en forme (couleurs, polices, etc.).

Par ailleurs, nous travaillons à l'élaboration d'un langage formel de description des connaissances mobilisées pour extraire des informations à partir d'un arbre. Ce langage s'inspirera de LangText (Crispino, 2003) et de (Amardeilh, 2007). L'un des apports de ce langage est de formaliser de manière déclarative la notion d'espace de recherche, d'indicateur et d'annotation d'une unité textuelle. Nous devons néanmoins adapter ce langage au parcours d'un arbre qui représente la structure d'une page Web.

Références

AMARDEILH F., (2007). *Web Sémantique et Informatique Linguistique : propositions méthodologiques et réalisation d'une plateforme logicielle*. Thèse de doctorat. Paris : Université Paris-Sorbonne.

BATTISTELLI D., MINEL J.-L., SCHWER S. (2006). Représentation des expressions calendaires dans les textes : une application à la lecture assistée de biographies. *Traitement Automatique des Langues* 47, 3, 1-26.

BRY F., LORENZ B., OHLBACH H. J., SPRANGER S. (2003). On Reasoning on Time and Location on the Web. In *Workshop on Principles and Practice of Semantic Web Reasoning*, LNCS 2901, Springer-Verlag.

CRISPINO G. (2003). *Une plate-forme informatique de l'Exploration Contextuelle : modélisation, architecture et réalisation (ContextO) – Application au filtrage sémantique de textes*. Thèse de doctorat. Paris : Université Paris-Sorbonne.

NOËL L., CARLONI O., MOREAU N., WEISER S. (2008). Designing a knowledge-based tourism information system. *Int. J. of Digital Culture and Electronic Tourism*, Special Issue on National Tourism Organisations and Exploitation of Information Technologies, en cours de publication.

PITON O., MAUREL D. (2004). Les Noms Propres Géographiques et le Dictionnaire Prolintex. *Cahiers de la MSH Ledoux*, Série Archive, Bases, Corpus 1, 53-76.

STERN R.-D. (2007). *Expression linguistique du temps et représentation ontologique : OWL-Time et étude des adverbiaux temporels.* Mémoire de master. Paris : Université de Paris-Sorbonne.

TENIER S., TOUSSAINT Y., NAPOLI A, POLANCO X. (2006). Instantiation of relations for semantic annotation. Actes de *the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, IEEE, Computer Society* Press, 463-472.