



Du texte au portail sémantique : cas d'utilisation lié à des données temporelles

Charles Teissèdre, Delphine Battistelli, Jean-Luc Minel

► To cite this version:

Charles Teissèdre, Delphine Battistelli, Jean-Luc Minel. Du texte au portail sémantique : cas d'utilisation lié à des données temporelles. IC 2010, Jun 2010, France. pp.209-220. halshs-00493727

HAL Id: halshs-00493727

<https://shs.hal.science/halshs-00493727>

Submitted on 23 Jun 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Du texte au portail sémantique : cas d'utilisation lié à des données temporelles

Communication appliquée

Charles Teissèdre^{1 2}, Delphine Battistelli³, Jean-Luc Minel¹

¹ MoDyCo UMR 7114, Université de Paris-Ouest Nanterre La Défense - CNRS

{charles, teissedre}@gmail.com

{jean-luc, minel}@u-paris10.fr

² Mondeca - 3, cité Nollez, 75018 Paris

³ Equipe STIH, Université de Paris-Sorbonne,

{delphine, battistelli}@paris-sorbonne.fr

Résumé : L'article décrit une méthode et des ressources pour manipuler des données temporelles, à la fois pour un utilisateur final qui souhaite interroger un portail avec des filtres temporels et, en amont, pour le peuplement d'ontologies. Le système assiste les personnes ayant à charge la saisie d'information dans une ontologie en leur permettant d'exprimer en langage naturel les propriétés temporelles liées à l'accessibilité d'un lieu (dates et horaires d'ouverture). Le système projette leur interprétation sur un calendrier éditable, afin de permettre à l'utilisateur de contrôler l'analyse et si nécessaire d'y modifier des informations. Cette projection du texte sur le calendrier est opérée à l'aide d'un module de raisonnement qui calcule en extension les données temporelles. Cet ensemble d'outils prépare la mise en œuvre, dans un Portail Sémantique, de mécanismes permettant d'interroger un système de Recherche d'Informations en utilisant des critères temporels.

Mots-clés : Web Sémantique, Modélisation et visualisation de connaissances temporelles, Raisonnement temporel, Annotation linguistique.

1 Introduction

Si de nombreux travaux envisagent le recours à l'annotation textuelle afin de faciliter l'alimentation des ontologies, l'objectif dans ce cadre est d'en faciliter l'industrialisation pour un cas d'utilisation lié à la temporalité, en ne faisant porter l'analyse que sur des portions restreintes et bien identifiées de la langue et en couplant l'ensemble avec des outils de normalisation et de raisonnement permettant un contrôle de l'information saisie. L'objectif est ainsi de proposer des solutions d'assistance au peuplement d'ontologies pour les informations concernant les dates et horaires d'ouverture et de fermeture. Ces étapes précèdent la mise en place d'un portail sémantique qui permettra d'interroger les données en utilisant des filtres temporels.

Si le Web Sémantique tel qu'il est envisagé par (Berners-Lee et al., 01) peut à raison s'entendre comme un dispositif qui doit permettre aux machines d'assister plus efficacement les utilisateurs pour l'accès aux ressources sur le Web, dans le mouvement inverse, la puissance d'interrogation qu'il offre pour les utilisateurs finaux s'accompagnent également d'une difficulté croissante pour ceux qui ont à charge de modéliser, de maintenir et d'alimenter les bases de connaissances au cœur de cette infrastructure. En ce sens, l'assistance à l'alimentation manuelle des bases de connaissances répond en partie à cette problématique : en s'appuyant sur des traitements linguistiques, ainsi que sur des processus de raisonnement pour construire des données structurées à partir du texte libre, il devient possible de faciliter la saisie d'informations complexes, telles que certaines propriétés temporelles.

Dans le cas d'étude qui retient notre attention, l'objectif est double. Il s'agit, d'un côté, de fournir des outils pour renseigner de façon simple, dans une base de connaissances, des périodes d'accès et, d'un autre, d'offrir des outils pour les interroger à travers un portail : *quand un site touristique, un musée, un restaurant est-il ouvert ? Quel jour et à quelle heure ont lieu les séances d'une pièce de théâtre ?* Dans ce cadre, la notion d'accessibilité recouvre aussi bien l'accès à un musée, à un restaurant, la programmation d'un festival ou de séances de cinéma. Désormais, nous désignerons par « période d'accessibilité » tout énoncé renvoyant à des propriétés temporelles caractérisant l'accessibilité d'un lieu. Aujourd'hui, la plupart des magazines ou des sites en ligne délivrent ces informations sous une forme textuelle : telles sont les périodes d'ouverture ou de fermeture, tels sont les horaires, tel est le programme de la salle. Cette présentation de l'information suffit lorsqu'un utilisateur sait où il souhaite se rendre (c'est-à-dire lorsqu'il veut un complément d'information sur un sujet déterminé) ; en revanche, elle rend difficile la réponse à une question plus ouverte du type : *je me déplace à tel endroit, à telle date, quels sont les musées ouverts le matin, quels sont les films qui vont être diffusés vers 19h*, car il faut alors consulter les horaires et programmes de chacun des sites, de chacune des salles ou de chacun des films. C'est au demeurant une des difficultés auxquelles sont confrontés les SGBD relationnels et à laquelle le Web Sémantique peut proposer une réponse efficace.

Dans un premier mouvement, nous montrons l'intérêt de s'appuyer sur le traitement des expressions en langage naturel, pour faciliter à la fois la saisie et le stockage, au sein d'une base de connaissances, des conditions d'accessibilité à un site. Nous décrivons ensuite la façon dont le système traite ces expressions en langage naturel, pour les interpréter conformément à un modèle qui repose sur le formalisme OWL. Dans la dernière section, nous décrivons les mécanismes de raisonnement à partir desquels les périodes d'accessibilité sont transformées en périodes effectives, définies en extension et transposables sur un calendrier, avant de montrer comment les différents outils décrits préparent la mise en œuvre d'un portail, dans lequel un utilisateur pourra effectuer des recherches avec des filtres temporels.

2 La temporalité linguistique et le temps calendaire

Permettre une saisie d'informations temporelles en langage naturel vise à faciliter le travail des opérateurs qui s'occupent d'alimenter manuellement une base de connaissances. Un des enjeux de la communauté du Web Sémantique est de parvenir à exploiter, de la façon la plus automatisée possible, la connaissance portée par les textes, pour la rendre interopérable et manipulable par des agents logiciels, comme le montrent par exemple (Bontcheva et Cunningham, 03). De nombreux travaux de recherche s'intéressent plus spécifiquement à l'extraction et l'annotation d'informations temporelles dans les textes, notamment les travaux entrepris autour de TimeML (Pustejovsky et al., 03). Pour ce qui regarde les travaux sur les expressions calendaires en particulier, le plus souvent, l'objectif poursuivi est de les ancrer sur un calendrier et de les relier à des événements (Schilder et Habel, 01), (Setzer et Gaizauskas, 00), (Filatova et al., 01), soit afin d'ordonner les événements décrits par les textes, soit afin de mettre en œuvre des systèmes de Question/Réponse (TERQAS, 02). Ces travaux cherchent ainsi à ramener les expressions calendaires présentes dans les textes à un format calendaire normé.

Plutôt que de transposer directement les références temporelles présentes dans les textes sur un calendrier, il nous a semblé important, pour notre démarche, de distinguer, au sein des traitements, les représentations temporelles sous-jacentes au calendrier et celles portées par les expressions calendaires qui définissent des périodes d'accessibilité. La temporalité linguistique présente en effet des spécificités qui la distinguent du *temps social* au sens de (Benveniste, 74), dont l'histoire a cristallisé progressivement la représentation dans la norme du calendrier. La langue - ce en quoi elle est économe par rapport aux représentations temporelles normées du calendrier - permet de ne pas avoir à désigner toujours des périodes à travers des intervalles de temps précisés en extension. Les représentations calendaires qui se conforment par exemple à la norme iCalendar RFC2445¹ (un standard utilisé notamment par Google Calendar ou Microsoft Outlook), définissent essentiellement des événements caractérisés par des intervalles de temps. La langue, pour sa part, permet de désigner des périodes itératives (*les lundis*), imprécises (*aux alentours de la mi-mars*), définies les unes par rapport aux autres (*deux jours après*) ou définies par rapport à l'acte énonciatif (*aujourd'hui, demain, ce week-end*). Les travaux autour d'OWL-Time (Hobbs et Pan, 04) se sont intéressés à la modélisation des expressions calendaires sans les réduire pour autant à une représentation du temps rive à la norme calendaire : ils décrivent ainsi la façon dont on peut construire des agrégats temporels, définir des périodes itératives ou encore lier entre elles la définition de plusieurs périodes. Conformément à cette démarche, mais en posant d'emblée l'écart entre la représentation normée du calendrier et la représentation linguistique des expressions calendaires, les travaux présentés ici s'appuient sur une modélisation linguistique présentée dans (Battistelli et al., 08) que l'on a étendue pour pouvoir rendre compte

¹ <http://tools.ietf.org/pdf/rfc2445.pdf>

de la définition de périodes d'ouverture et de fermeture, dont voici plusieurs exemples récupérés sur différents sites Web :

Ex. 1 : *Ouvert du mardi au samedi de 10h à 19h et le dimanche de 13h à 19h, sauf les jours fériés suivant : 1^{er} jan, dimanche et lundi de Pâques, 1^{er} et 8 mai.*

Ex. 2 : *Ouvert de 10h à 4h et le vendredi et samedi de 10h à 5h.*

Ex. 3 : *Horaires : Du lundi au jeudi : 9h à 21h30. Vendredi et samedi : de 9h à 22h30. Dimanche : de 9h à 18h30. Fermé le 1^{er} janvier.*

Ex. 4 : *Ouvert tous les jours de 10h à 19h jusqu'à dimanche 3 janvier. A partir du lundi 4 janvier : ouvert le mercredi, samedi et dimanche : 10h-19h ; le vendredi : 14h-19h. Fermé le 1^{er} janvier.*

La décomposition de ces informations ne va pas de soi, car elles sont de nature diverses : elles peuvent être composées (a) d'intervalles (*du 2 février au 13 avril*), (b) d'expressions calendaires itératives (*le lundi*), ou (c) d'expressions calendaires absolues (*le 1^{er} mai 2009*). En outre, il peut y avoir (d) des spécifications pour préciser, à une granularité plus fine ou plus large, une période d'accessibilité (*le dimanche de 13h à 19h* ou *fermé le lundi du 15 mars au 30 septembre*) ou encore (e) des exceptions (*ouvert tous les jours sauf le mardi*). La mise en œuvre de notre procédure permet à des opérateurs de saisir ces informations en langage naturel, moyennant parfois quelques corrections dont on précise plus loin la nature. Ce système leur évite ainsi le recours à des formulaires qui découpent l'information en tiroirs multiples, comme le font généralement les formulaires destinés à produire de nouvelles instances dans une base de connaissances. L'application annote les expressions définissant des périodes d'accessibilité et projette leur interprétation sur un calendrier éditable afin que l'utilisateur puisse contrôler la justesse de cette interprétation et éventuellement y modifier, ajouter ou corriger les informations, interagissant ainsi entre le texte et le calendrier. La synchronisation entre les deux représentations, textuelles et iconiques, est maintenue par l'applicatif : une modification sur le calendrier entraîne la modification du texte et inversement.

Du reste, la capacité qu'offre la langue de condenser, dans des formules brèves, des périodes dont la définition en extension peut être coûteuse en termes de stockage (définitions itératives vs. définitions en extension), explique pourquoi la saisie libre est une bonne alternative aux représentations standards et normées du temps calendaire, pour définir les propriétés temporelles relatives à l'accessibilité d'un site. En effet, le système proposé permet de stocker, selon les besoins, ou bien des périodes définies en intension ou bien des intervalles de temps définis en extension. Pour que les moteurs de recherche puisse traiter des requêtes contenant des filtres temporels, telles que « *musées ouverts ce week-end* », les propriétés temporelles gagnent à être stockées sous une forme extensive plus simple à indexer et à interroger : pour ce faire, il faut pouvoir calculer l'extension d'une expression comme « *ouvert le lundi de 9h à 19h* » afin d'obtenir des informations du type « *ouvert lundi 13 mars, de 9h à 19h* », « *ouvert lundi 20 mars, de 9h à 19h* », etc. Ce calcul de l'extension des périodes d'accessibilité, qui consiste à passer du modèle linguistique au modèle

calendaire, ne peut pas être opéré à la volée, car la procédure affecterait les temps de réponse. Dans l'architecture logicielle sur laquelle nous nous appuyons, décrite dans (Noël et Azémard, 08), la base de connaissances contenant les données sur l'accessibilité des sites est couplée à un moteur de recherche : la saturation des connaissances (le calcul de l'extension temporelle) s'effectue au moment de l'export de la base vers les moteurs de recherche. La base de connaissances, pour sa part, ne stocke ainsi l'information sur les périodes d'accessibilité que sous la forme du modèle linguistique, ce qui permet de ne traiter dans la base qu'un réseau de taille réduite.

3 Du texte libre aux données structurées

3.1 Le modèle des périodes d'accessibilité

Nous proposons un premier modèle linguistique des périodes d'accessibilité qui tient compte de leur spécificité linguistique face à la représentation normée du calendrier, tout en étant susceptible de fournir toutes les informations nécessaires pour la mise en œuvre des traitements de raisonnement et de requêtage décrits plus loin.

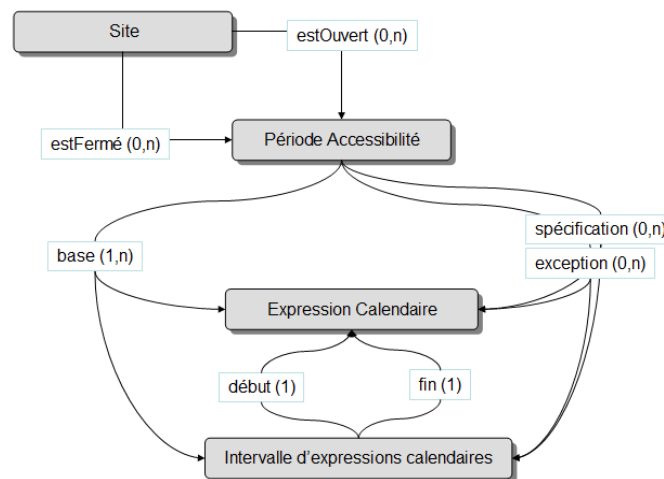


Fig. 1 – Proposition d'un modèle linguistique des périodes d'accessibilité à un site

Dans la spécification de ce modèle sous sa forme OWL (fig. 1), un Site est *ouvert* ou *fermé* sur une ou plusieurs Périodes d'Accessibilité. Les relations de *base*, de *spécification* et d'*exception* (des Object Properties en OWL) lient les Périodes d'Accessibilité aux Expressions Calendaires et Intervalles d'Expressions Calendaires qui les composent. Un Intervalle d'Expressions Calendaires est lié à deux Expressions Calendaires, dont l'une est le début, l'autre la fin. Une Expression Calendaire (l'objet au cœur du modèle) est composée de plusieurs attributs (ou DataType Properties en OWL), dont les principaux sont les suivants : (i) les grains qui sont susceptibles d'entrer dans la composition d'une Expression Calendaire (*minute*, *heure*, *jour de la semaine*, *jour du mois*, *semaine du mois*, *semaine de l'année*, *mois*, *année*), (ii) les

parties du jour (*{matin, après-midi, soir, nuit}*), les parties du mois et de l'année (*{début, mi, fin}*), ainsi qu'un (iii) label d'affichage. Pour une description plus détaillée de la sémantique des expressions calendaires et des ressources pour les annoter, nous renvoyons aux travaux décrits dans (Battistelli et al., 08) et (Teissèdre et al., 10).

3.2 Instanciation du modèle

Les expressions définissant des périodes d'accessibilité sont analysées à l'aide d'un ensemble de grammaires locales implémentées¹ à l'aide du logiciel Unitex². La procédure d'annotation conduit à une première interprétation de ces expressions, préparant l'instanciation du modèle. Les transducteurs décrivent la composition linguistique des périodes d'accessibilité et les annotent pour produire, en sortie, un premier niveau d'interprétation. À l'issue du processus d'annotation, l'interprétation des périodes d'accessibilité est encore incomplète : la portée des exceptions ou des spécifications doit encore être déterminée et certaines tournures elliptiques demandent à être dépliées. Ainsi, pour que l'interprétation d'une expression telle que « *les 7 et 8 janvier* » soit complète, elle doit être analysée comme « *les 7 janvier et 8 janvier* ». Cette difficulté, liée à la coordination, se pose également au niveau de l'interprétation du statut d'accessibilité (ouvert ou fermé) : l'algorithme mis en œuvre, qui succède à l'annotation, interprète ainsi l'expression « *fermé en août et le dimanche* » comme « *fermé en août et fermé le dimanche* ». Cette procédure de complétion a également à charge de fixer la portée des spécifications et exceptions. Pour l'expression « *ouvert tous les jours sauf le lundi, de 9h à 19h* », il faut ainsi que l'outil interprète les horaires (*de 9h à 19h*) comme une spécification qui n'est pas liée à l'exception (*sauf le lundi*), mais à la période d'accessibilité de référence (*tous les jours*), puisqu'il s'agit des horaires d'ouverture standard. Pour ce faire, l'outil compare la granularité et la nature des expressions calendaires qui suivent un marqueur d'exception : tant qu'elles sont de même granularité et de même nature (« *sauf lundi et mardi* »), on ajoute à la définition de la période d'accessibilité de nouvelles exceptions. Si la granularité diffère (« *de 9h à 19h* »), l'outil interprète ce changement comme la fin de la portée de l'exception. Cette seconde étape de structuration vient ainsi compléter et achever la structuration partielle produite par le processus d'annotation. Il faut noter, qu'à ce stade, l'analyse s'en tient délibérément à la granularité temporelle de l'expression saisie ; elle ne cherche pas à descendre à une granularité plus fine (jour, mois, année, heure, par exemple). À la fin de ces deux processus, une interprétation des périodes d'accessibilité conforme au modèle linguistique est délivrée³.

¹ Les ressources linguistiques exploitées dans ces grammaires sont le résultat de la capitalisation de plusieurs travaux réalisés dans les projets ANR Eiffel et Conique.

² Unitex: <http://www-igm.univ-mlv.fr/unitex>

³ Il convient de noter que cette interprétation peut s'avérer erronée dans un contexte socio-culturel donné. Par exemple « ouvert tous les jours » doit être interprété comme fermé le samedi et le dimanche s'il s'agit des horaires d'ouverture d'une mairie en France (cf. Weiser, 2010).

3.3 Evaluation des ressources pour l'analyse des expressions saisies en langage naturel

Les expressions saisies par les personnes alimentant la base de connaissances sont soumises à un service Web qui intègre les ressources décrites et permet d'établir, sur une période donnée, l'accessibilité du site concerné. Le modèle linguistique est ainsi instancié après l'annotation et la phase de complétion ; l'extension des périodes d'accessibilité est alors calculée sur un intervalle de temps donné, selon une granularité qui dépend des besoins, à l'aide de mécanismes de raisonnement décrits dans la section suivante. L'ensemble des traitements se répartit ainsi en trois tâches distinctes : (1) l'annotation, (2) la complétion des informations pour instancier le modèle et (3) la transposition des périodes d'accessibilité dans un format calendaire normé, selon une granularité paramétrable et dans un intervalle de temps donné.

Les ressources développées pour l'annotation et l'instanciation du modèle (tâches (1) et (2)) couvrent déjà un nombre important d'expressions définissant des périodes d'ouverture et de fermeture. Une évaluation a été menée sur un corpus fourni par un utilisateur contenant 400 expressions qui définissent des périodes d'accessibilité. Le corpus provient de différents sites Web de musées, de restaurants, de cafés ou d'autres sites dédiés au tourisme et au loisir. Saisies telles quelles dans le système, sur les 400 expressions, 71 ne sont pas complètement ou correctement interprétées par le système, qui présente ainsi un taux de précision de 82,25%. Dans ce contexte, la mesure du taux de rappel (qui permet d'apprécier la couverture d'un système et sa capacité à repérer les énoncés qu'il vise) ne fait pas vraiment sens, puisque l'utilisateur ne soumet au système que des périodes d'accessibilité : le système n'analyse donc pas des textes complets dont il lui faudrait extraire les périodes d'accessibilité. Les difficultés auxquelles se heurtent les ressources pour annoter et instancier le modèle, en leur état actuel, relèvent essentiellement de trois catégories :

- (a) le cas de l'ambiguïté de certaines expressions, comme « *visite les mardis, jeudis et samedis après-midi* », pour lesquelles les ressources ne livrent qu'une seule interprétation : en l'occurrence, l'interprétation est la suivante « *ouvert les mardis et jeudis toute la journée et l'après-midi le samedi* », alors qu'une autre interprétation était possible (*ouvert l'après-midi, les mardis, jeudis et samedis*).
- (b) le cas où des connaissances externes aux textes sont nécessaires pour l'interpréter correctement (*fermé pendant les vacances scolaires et jours fériés*).
- (c) le cas d'expressions dont l'interprétation s'appuie sur des inférences simples pour un utilisateur mais complexe à formaliser et systématiser. Ainsi des expressions telles que « *lundi et mercredi, ouvert dès 8h* » ou « *les 24 et 31 décembre : ouverture jusqu'à 20h.* » sont définies par rapport à d'autres périodes. Leur interprétation exige ainsi d'établir un lien avec des horaires standards d'ouverture.

Ces limites du système ne sont pas pour autant des limites incontournables pour l'utilisateur, dans la mesure où l'outil lui donne à voir ce qu'il est parvenu à interpréter : cette interprétation lui est restituée sous une forme textuelle normalisée et sous la forme visuelle d'un calendrier éditable. Dans le cas où l'expression saisie

n'est pas correctement interprétée, l'utilisateur peut éventuellement modifier cette expression (remplacer, par exemple, « *les 24 et 31 décembre : ouverture jusqu'à 20h* » par « *ouvert les 24 et 31 décembre, de 8h à 22h* »). L'expressivité possible est ainsi directement fonction de la couverture que sont en mesure de proposer les processus d'analyse linguistique : plus grande est cette couverture, plus la saisie peut être « naturelle ». Pour tenir compte de ces limites, une notice invite les opérateurs à se conformer à quelques règles d'écritures simples : en particulier, éviter de saisir des informations qui font intervenir des connaissances que ne possède pas le système, comme les dates des vacances scolaires et des jours fériés ou éviter de définir des périodes par rapport à d'autres qui précèdent (« *ouvert dès 8h le lundi* »). Ce qu'il importera surtout d'évaluer par la suite, c'est le retour des utilisateurs et leur capacité à saisir, dans la base de connaissances, les informations qu'ils souhaitent, de façon simple, rapide et contrôlable. En ce sens, une campagne d'évaluation mesurant les retours des opérateurs qui ont à charge la saisie dans la base de connaissances a été entamée et livrera ses résultats d'ici quelques mois : quel type d'énoncé est mal analysé ? ces problèmes ont-ils pu être contournés en transformant l'expression saisie ? les recommandations d'usage sont-elles faciles à suivre ? y a-t-il des cas où elles ne suffisent pas pour exprimer une période d'accessibilité ? la procédure de contrôle de l'interprétation est-elle simple ?, etc.

4 Du raisonnement temporel à l'interrogation dans un portail sémantique

Les traitements décrits visent à enrichir un portail dit « sémantique », au sens où les utilisateurs finaux peuvent interroger les données en croisant différentes facettes (type d'objets cherchés dans une taxonomie) et différents filtres (temporel, géographique). L'objectif en l'occurrence est d'offrir à l'utilisateur final la possibilité d'effectuer des recherches avec des filtres temporels : « *musées ouverts à Venise le week-end du 1^{er} mai* », « *restaurants ouverts le dimanche après 22h à Toulouse* ». Dans le même ordre d'idée, une application mobile pourrait par exemple fournir tous les sites ouverts à proximité du lieu où se trouve l'utilisateur. Pour cela, des mécanismes de raisonnement et d'interrogation sont mis en œuvre, afin de passer du modèle linguistique au modèle calendaire et de s'appuyer sur le modèle calendaire pour construire des index.

4.1 Raisonnement temporel

Les mécanismes de raisonnement implémentés permettent de calculer l'extension temporelle des périodes d'accessibilité. Le raisonneur transforme ainsi les instances du modèle linguistique en un ensemble d'intervalles de temps auxquels un statut d'accessibilité est attribué. Par exemple, si l'on souhaite connaître l'accessibilité d'un site à une granularité du niveau des horaires sur la période du 19 avril 2010 au 19 mai 2010, pour une expression telle que « *Ouvert tous les jours, sauf le mardi, de 8h à 17h. Nocturne le jeudi jusqu'à 20h.* », le raisonneur produit en sortie une liste

d'intervalles de temps : « ouvert le lundi 19 avril 2010 de 8h à 17h », « fermé le 20 avril 2010 », ..., « ouvert le lundi 3 mai 2010 de 8h à 17h », etc. La discrétisation des périodes d'accessibilité - ce qui correspond au passage du modèle linguistique au modèle calendaire - est paramétrable et peut s'opérer, selon les besoins, à une granularité plus ou moins fine. L'affichage sur un calendrier exige que le calcul de l'extension temporelle soit fait à une granularité de l'ordre du jour ou des horaires¹.

Au fil de l'analyse, l'algorithme attribue des statuts d'ouverture aux périodes résolues en extension. Les statuts attribués sont soit définitifs (« ouvert », « fermé », « conflit »), soit provisoires (« présumé ouvert », « présumé fermé », « non renseigné »). Le statut « conflit » est attribué lorsque des informations contradictoires apparaissent. Si le statut d'accessibilité, sur une période donnée, est « présumé fermé » à un stade du traitement, il pourra, par la suite, être modifié en « ouvert » si un traitement ultérieur parvient à cette conclusion ; en revanche, si le statut « fermé » est attribué, les traitements ne pourront plus le modifier ultérieurement en « ouvert » : dans ce cas, le statut « conflit » est attribué. À l'issue des traitements, si aucune information n'a modifié les statuts provisoires « présumé ouvert » ou « présumé fermé », ces derniers sont transformés en « ouvert » et « fermé ». Les mécanismes de raisonnement qui permettent de passer des instances du modèle linguistique à leur transposition en un format calendaire normé se heurtent essentiellement à deux difficultés :

- (a) comme pour l'instanciation du modèle linguistique à partir des annotations, l'interprétation de certaines expressions s'appuie sur des inférences qu'un utilisateur opère sans difficulté, mais qu'il est en revanche difficile de formaliser. Le raisonneur peut ainsi considérer à tort certaines informations comme étant contradictoires. Ainsi, pour une expression telle que « *Ouvert du lundi au samedi, de 8h à 19h. Fermé le mardi.* », le système attribuera le statut « conflit » pour l'ensemble des mardis considérés, car selon la première partie de l'expression, les mardis semblent ouverts, alors que la seconde partie de l'expression précise qu'ils sont fermés. Un utilisateur comprend très bien que la dernière information prévaut sur les précédentes. De même, pour être transposée sur un calendrier, l'expression « *ouvert le dimanche de 19h à 2h du matin* » demande à ce que l'outil infère que le site est ouvert le dimanche de 19h à minuit et le lundi de minuit à 2h. Différentes inférences de ce type ont été implémentées (le dernier exemple est traité), mais il est difficile de les couvrir toutes, la langue permettant un grand nombre de variations dans l'expression des périodes d'accessibilité.

- (b) le raisonneur doit interpréter des expressions contenant des données symboliques qui n'ont pas d'équivalent fixe dans la norme calendaire. Des expressions comme « *fin novembre* » et « *dimanche soir* » renvoient ainsi au problème de la transposition de données symboliques vers des données normalisées. Comme l'ont souligné (Fortin & al., 09), l'interprétation de ces données symboliques varient selon le type d'objet : l'ouverture le soir, par exemple, pour un supermarché et

¹ Notre démarche permettrait, dans un autre contexte d'utilisation, de connaître par exemple tous les sites qui sont ouverts (ou partiellement ouverts) pendant le mois d'août : dans ce cas, le calcul de l'extension temporelle n'aurait pas besoin de descendre à une granularité plus fine que le mois.

un bar, ne s'interprète pas nécessairement de la même manière. Le système permet de paramétrer les règles de transformation d'un modèle à l'autre pour les données symboliques. Ainsi, les expressions annotées comme *Partie De Jour (matin, après-midi, soir, nuit)*, sont traduites en horaires effectifs, mais l'imprécision est alors marquée par un drapeau, qui signale qu'il s'agit d'horaires indicatifs.

Une fois les informations temporelles calculées en extension pour une période donnée, il devient possible de les présenter sur un calendrier ou un agenda. La fig. 2 montre ainsi les jours d'ouverture et de fermeture d'un site pour la semaine du 26 avril au 2 mai si l'on soumet la précédente expression donnée en exemple.

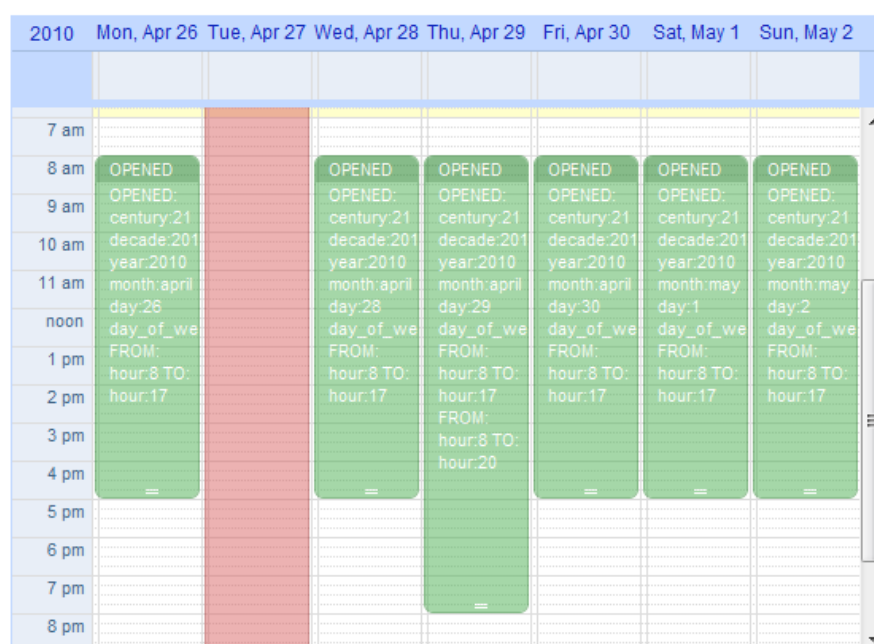


Fig. 2 – Visualisation de l'accessibilité d'un site sur un calendrier pour l'expression « Ouvert tous les jours, sauf le mardi, de 8h à 17h. Nocturne le jeudi jusqu'à 20h. »

Le calendrier où sont projetés les résultats de l'analyse est éditable : un utilisateur peut ainsi y modifier un horaire ou y ajouter un jour d'ouverture ou de fermeture. Le texte saisi est alors automatiquement mis à jour. Un utilisateur peut ainsi, à travers le calendrier, préciser que le samedi 1^{er} mai est un jour de fermeture. La phrase suivante « *Fermé samedi 1 mai 2010.* » est ainsi générée et ajoutée au texte. Pour que le texte et les informations éditées sur le calendrier soient synchronisés, un module permet de retrouver, dans les instances du modèle linguistique, celle éventuellement impactée par la modification. Sa modification entraîne alors en retour la génération d'un nouveau texte. L'utilisateur peut ainsi jongler entre deux manières de définir des périodes d'accessibilité, qui chacune renvoie à un modèle qui lui est propre : le texte et le calendrier. Les informations ajoutées directement sur le calendrier sont toujours

des informations extensives : on ne peut pas préciser via le calendrier que le 1^{er} mai est fermé chaque année ; pour cela, il faut modifier directement le texte (« *Fermé le 1^{er} mai* »). Après la saisie, les données sont stockées sous la forme du modèle linguistique présenté plus haut. Lors de l'export vers les moteurs de recherche pour indexation, le module de raisonnement calcule leur extension sur l'année à venir.

4.2 Interrogation dans un portail sémantique et filtres temporels

Au niveau du requêtage des données, deux grandes classes de recherches liées à des critères temporels sont envisagées : (1) une requête portant sur un objet ciblé (par exemple, le musée du Louvre) dont on souhaite voir les périodes où il est accessible (il s'agit donc d'un complément d'information sur l'accessibilité d'un objet qui intéresse l'utilisateur) et (2) une requête ouverte avec filtres temporels portant sur un objet ciblé ou un type d'objet (par exemple les musées ou les restaurants).

Dans le premier cas, le plus simple, lorsque l'utilisateur est centré sur un objet de la base (un restaurant R, un musée M), le portail peut alors lui proposer de visualiser les périodes d'accessibilité sous forme textuelle (ce qui est fait classiquement) ou bien sous la forme graphique d'un calendrier ou d'un agenda, permettant la visualisation des horaires et jour d'ouverture pour une semaine donnée. L'idée est surtout de permettre aux prestataires de modifier en ligne, quand nécessaire, leurs périodes d'ouverture et de fermeture. Pour le second type de recherche, la requête porte sur un objet ou un type d'objet et est croisée avec un filtre temporel (ou la conjonction de filtres temporels et spatiaux) : « *festival de jazz dans l'Indre en mai* », « *supermarchés ouverts le dimanche matin à Strasbourg* ». Les filtres temporels ont pour but de ne renvoyer, dans la liste des résultats du moteur de recherche, que les objets « accessibles » à l'intérieur de la période d'interrogation. Selon la nature des objets visés par la requête de l'utilisateur, il faut renvoyer tous les résultats dont la période d'accessibilité recoupe celle du filtre (pour un festival ou une exposition, par exemple) ou ne renvoyer que les résultats qui sont accessibles tout au long de la période visée (on souhaite en effet d'abord voir les hôtels disponibles sur toute la période de notre déplacement).

4.3 Perspectives

Les différents traitements décrits serviront de brique dans un portail sémantique pour pouvoir proposer l'ajout de filtres temporels dans les requêtes. Dans l'architecture retenue et précitée (Noël et Azémard, 08), qui couple la base de connaissances avec un moteur de recherche, les informations temporelles calculées en extension seront exportées en direction du moteur pour être indexées. La problématique à résoudre devient alors surtout, pour les moteurs de recherche, de repérer dans les requêtes les filtres temporels et d'arrêter un comportement cohérent dans l'ordonnancement des résultats proposés. Le comportement « normal » attendu dans le portail veut qu'on ne renvoie à l'utilisateur que les sites accessibles sur la période de la requête. Mais ce comportement doit être modulé, lorsque les

informations temporelles sont issues de la transposition de données symboliques (*matin, fin novembre*) vers des données calendaires. On peut ainsi envisager de pondérer l'ordonnancement des résultats. Le moteur de recherche renverrait en tête de liste les résultats assurés, et en fin de liste, les résultats dont le statut d'accessibilité est moins certain, en suggérant éventuellement à l'utilisateur cette incertitude (horaires indicatifs, informations à vérifier auprès du prestataire, etc.).

Références

- BATTISTELLI D., COUTO J., MINEL J-L. & SCHWER S. (2008). Representing and Visualizing calendar expressions in texts. In ACTES STEP'08. Venise.
- BENVENISTE E. (1974). Problèmes de linguistique générale (Tome 2). In GALLIMARD. Paris.
- BERNERS-LEE T., HENDLER J. & LASILLA O. (2001). The Semantic Web. In SCIENTIFIC AMERICAN, May 2001.
- BONTCHEVA K. & CUNNINGHAM H. (2003). The Semantic Web: A New Opportunity and Challenge for Human Language Technology. In PROCEEDINGS OF THE SECOND INTERNATIONAL SEMANTIC WEB CONFERENCE, *Workshop on Human Language Technology for The Semantic Web and Web Services*. 20-23 October 2003. p. 89-96. Florida.
- FILATOVA E., HOVY E. (2001). Assigning Time-Stamps to Event-Clauses. In ACTES DE WORKSHOP ON TEMPORAL AND SPATIAL INFORMATION PROCESSING, *ACL'2001*, p. 88-95.
- FORTIN J., CARLONI O., LECLERE M. & WEISER S. (2009). Extraction et exploitation de données temporelles pour un portail d'e-tourisme. In EGC'09. *Fouille de Données Temporelles – Analyse de Flux de Données*. Strasbourg.
- HOBBS J. R. & PAN F. (2004). An Ontology of Time for the Semantic Web. In ACM TRANSACTIONS ON ASIAN LANGUAGE PROCESSING (TALIP). *Special issue on Temporal Information Processing*, Vol. 3, No. 1, March 2004, p. 66-85.
- NOËL L. & AZÉMARD G. (2008). From semantic web data to inform-action: a means to an end. In PROCEEDINGS OF THE SEMANTIC WEB USER INTERACTION WORKSHOP, at *CHI'08 - Exploring HCI Challenges*. Florence.
- PUSTEJOVSKY J., CASTANO J., INGRIA R. , SAURI R., GAIZAUSKAS R., SETZER A. & KATZ G. (2003). TimeML: Robust Specification of Event and Temporal Expressions in Text. In ACTES DE IWCS-5, *Fifth International Workshop on Computational Semantics*.
- SCHILDER F. & HABEL C. (2001). From Temporal Expressions to Temporal Information: Semantic Tagging of News Messages. In ACTES DE ACL'01, *Workshop on temporal and spatial information processing*, p. 65 -72.
- SETZER A., GAIZAUSKAS R. (2000). Annotating Events and Temporal Information in Newswire Texts. In ACTES DE 2E LREC. p. 64-66.
- TEISSEDRE, C., BATTISTELLI, D., MINEL, J-L. (2010). Resources for Calendar Expressions Semantic Tagging and Temporal Navigation through Texts, In 7th international conference on Language Resources and Evaluation (LREC), 19-21 May 2010, Valletta, Malta, accepté.
- WEISER, S. (à paraître). Repérage et typage d'unités temporelles pour la construction d'une plate-forme d'annotation sémantique automatique de pages Web. Thèse de doctorat, Université Paris Ouest Nanterre La Défense.