UFR Sciences et Techniques

Master 1, Technologie de l'Internet
— Interopérabilité des Données et des Connaissances —
Sujet de Devoir (comptant pour le Contrôle Continu)

Contexte

DSL (Domain Specific Language)

Nous utilisons depuis longtemps et dans diverses activités des DSL parfois sans même nous en apercevoir. Peuvent être considérés comme des DSL le langage SQL, le langage de requête d'un moteur de recherche, un ensemble d'expressions régulières.

Un DSL (Domain Specific Language), par opposition aux langages de programmation généralistes tels que Python, Java, C++... et aux langages de modélisation généralistes tels que UML, se différencie pour les trois raisons suivantes :

Il est spécifique : Un DSL est un langage, de transformation, de modélisation, de programmation ou d'interrogation, selon ce pour quoi il a été conçu. Il est spécifique à un domaine métier et souvent peut ne pas répondre à toutes les problématiques liées à ce domaine.

Il est doté d'un vocabulaire précis et concis : Un DSL est simple d'utilisation et non ambigu. Son expressivité est basée sur un vocabulaire ciblé, propre au domaine métier.

Il aura été créé ou adapté pas le ou les concepteurs : La syntaxe du DSL est simple à personnaliser, elle se base sur un vocabulaire et des règles syntaxiques entièrement définies lors de sa conception.

Grâce à sa syntaxe personnalisable, exprimant des concepts communs à son domaine d'application, un DSL est paramétrable par rapport au métier, à une communauté ou à un projet. Facile à interpréter, il est un outil de communication entre les experts du domaine et le développeur, il permet aux experts de participer à la conception fonctionnelle de l'application.

Text Encoding Inititative (TEI)

Les Recommandations de la TEI – Text Encoding Inititative (TEI) Guidelines [1] – s'adressent à tous ceux qui souhaitent échanger des informations stockées sous forme électronique. Elles mettent l'accent sur l'échange des données textuelles mais d'autres types de données comme les images et les sons sont également pris en compte. Les Recommandations peuvent être appliquées aussi bien pour créer de nouvelles informations que pour échanger des informations existantes.

Les Recommandations fournissent le moyen de rendre explicites certaines caractéristiques d'un texte, de façon à faciliter le traitement de ce texte par des programmes informatiques pouvant s'exécuter sur des plates-formes différentes.

Les objectifs sont :

— de spécifier à partir d'une **spécialisation de la TEI**, un ou plusieurs DSL d'annotation permettant de marquer finement « au fil de l'eau » une ou des classes particulières d'informations contenuent dans les documents mis à votre disposition. Chaque langage ayant sa propre finalité et donc sa propre logique.
Par exemple : l'annotation syntaxique des mots ou combinaison de mots des textes, l'annotation par catégorisation sémantique d'expressions particulières (comme par des expressions textuelles exprimant le temps, ou l'espace, ou encore les actions ...)

1. http://www.tei-c.org/release/doc/tei-p5-doc/fr/html/index.html

— d'intégralement concevoir et spécifier un DSL permettant cette fois-ci d'organiser dans une structure régulière de type Base de Données (BD), les éléments et atrributs (obligatoires et optionnelles) pour telle ou telle catégorie d'information.
— Implémenter grâce à XSLT un premier ensemble de transformations permettant de passer du ou des langages DSL « au fil de l'eau » vers le DSL permettant l'organisation en une structure régulière.
— Implémenter (toujours en XSLT) un second ensemble de transformations permettant de passer, selon le besoin, des langages DSL spécifiés dans ce projet vers des langages pour des usages précis, qu'ils soient conçus selon les principes XML (GPX, SVG, KML, OWL...) ou pas (JSON, GeoJSON, PDF, DOT(Graphviz), LaTeX, PostScript, ...).

Le travail à réaliser consiste :

1. à spécifier le lexique et la syntaxe de chaque langage DSL pour répondre aux besoins que vous aurez mis en évidence par l'étude préalable de l'annotation des documents pré-annotés donnés en exemple et des objectifs de représentation de l'information fixés. Cette étude sera réalisée lors des séances en présentiel prévues à cet effet, elle constituera une sorte de mini cahier de charge (CC) pour la spécification des langages ;

   la spécification des propriétés de chaque langage sera ensuite obligatoirement réalisée grâce au langage `XML-Schema` (une première tentative pouvant être réalisée en `DTD`).
2. à tester, sur les exemples de documents mis à votre disposition, les capacités de contrôle et d'expressivité des DSL spécifiés pour un marquage « au fil de l'eau ».
3. de réaliser des transformations contrôlées grâce au langage `XSLT` et au langage de description `XML-Schema`
4. de réaliser grâce au langage `XSLT` les autres transformations nécessaires aux besoins exprimés dans le CC.

Quelques précisions :

— Les langages spécifiés pour des annotations « au fil de l'eau », étant à minima deux, doivent pouvoir coexister grâce aux espaces de nom qu'il sera nécessaire de définir. À titre d'exemple :
   1. si le premier permet d'annoter les catégories grammaticales des mots du texte ;
   2. et le second permet de marquer les expressions dans le texte qui font référence à des entités selon différentes catégories sémantiques. Catégories qui devront être définies préalablement ;

   alors il doit être envisageables d'avoir dans un même document les annotations provenant de ces deux langages.
— Le DSL modélisant les *entités* doit par exemple, quand il s'agit d'un lieu, permettre d'intégrer des informations supplémentaires comme la nature (catégorie sémantique) del'entité ou encore des coordonnées permettant de la « géo-localisation » sur une carte.
— Le DSL modélisant *les catégories grammaticales ou POS en anglais* doit prévoir d'intégrer, par exemple, les lemmes des mots.
— Une transformation vers un langage du monde XML pourrait être la traduction de toutes les « entités nommées de lieu » et informations associées en élément adéquat à des fins de cartographie grâce à l'api Leaflet (http ://leafletjs.com/) et aux données d'OpenStreetMap (http ://www.openstreetmap.org/).