TALN 2003, Batz-sur-Mer, 11–14 juin 2003

MDA-XML : une expérience de rédaction contrôlée multilingue basée sur XML

Guy Lapalme (1), Caroline Brun (2), Marc Dymetman (2),
(1) RALI - Université de Montréal
Montréal (Québec)
lapalme@iro.umontreal.ca
(2) Xerox Research Centre Europe
6, chemin de Maupertuis, 38240 Meylan
{Caroline.Brun,Marc.Dymetman}@xrce.xerox.com

Mots-clefs – Keywords

rédaction contrôlée multilingue, XML
multilingual controlled authoring, XML

Résumé - Abstract

Nous décrivons dans cet article l'implantation d'un système de rédaction contrôlée multilingue dans un environnement XML. Avec ce système, un auteur rédige interactivement un texte se conformant à des règles de bonne formation aux niveaux du contenu sémantique et de la réalisation linguistique décrites par un schéma XML. Nous discutons les avantages de cette approche ainsi que les difficultés rencontrées lors du développement de ce système. Nous concluons avec un exemple d'application à une classe de documents pharmaceutiques.

We describe an XML implementation of a multilingual authoring system. Using this system, an author can interactively write a text conforming to well-formedness content and realization rules described by an XML schema. We discuss the advantages of such an approach as well as the problems we met during the implementation of our system. We show in the end an example of an application for a class of pharmaceutical documents.

La production moderne de documentation fait de plus en plus l'objet d'applications informatisées, en particulier pour des domaines où la forme et le contenu des documents sont des paramètres cruciaux, comme c'est le cas par exemple pour la documentation technique ou les manuels d'instructions. Les besoins en termes de formatage sont maintenant bien identifiés, et la plupart des outils de traitement de texte proposent des fonctionnalités telles que des feuilles de style ou des modèles de documents. On ne peut cependant pas réduire la notion de rédaction de document à celle de formatage. Ainsi, des recherches en cours visent à développer des méthodes pour aider un auteur à composer le contenu du texte, en particulier lorsque les documents appartiennent à des domaines restreints. En effet, certains standards imposent des contraintes strictes sur l'information véhiculée dans les documents, et dans certains cas, la forme même des phrases peut être décrite au moyen d'une langue contrôlée.

Parmi ces différentes recherches, le système "Multilingual Document Authoring" (MDA), (Brun *et al.*, 2000; Dymetman *et al.*, 2000; Dymetman, 2002; Brun & Dymetman, 2002), développé ces dernières années dans la lignée des travaux de (Power & Scott, 1998) et de (Ranta, 2002), permet d'encoder sous forme de grammaires l'information nécessaire à la constitution d'un document valide. Nous montrerons un exemple tiré du domaine pharmaceutique illustrant la capacité de ces grammaires à décrire les connaissances du domaine (le fait qu'un produit pharmaceutique soit commercialisé sous une certaine forme contraint à la fois son conditionnement et son mode d'administration) ainsi que des contraintes sur l'organisation du document (la forme pharmaceutique et le mode d'administration doivent apparaître dans des sections spécifiques d'une notice).

Le système MDA développé à XRCE a été utilisé avec succès pour la production de notices pharmaceutiques et de comptes-rendus d'expériences biologiques (Brun *et al.*, 2002). Ce système possède de fortes connexions avec XML, car dans les deux cas le document est représenté comme une combinaison de structures d'arbres balisées et d'éléments de surface en texte libre.

Dans l'approche proposée par MDA, la rédaction est fondée sur des décisions sémantiques indépendantes de la langue permettant la construction de structures également indépendantes de la langue. Ce sont des mécanismes propres aux différentes langues couvertes par le système qui sont ensuite utilisés pour produire les réalisations textuelles multilingues. La structure sémantique guide l'auteur dans la spécification du document, à la manière d'une DTD ou d'un schéma guidant la structure d'un document XML. Mais les DTDs et les schémas XML sont très limités dans l'expressivité des dépendances sémantiques ou grammaticales qu'ils peuvent établir entre les différents sous-arbres de la structure du document. De plus, les outils de gestion de styles comme XSLT ou CSS, développés pour opérer des transformations sur la présentation des documents, sont mal adaptés aux traitements linguistiques. Cependant, les outils de rédaction basés sur XML sont de plus en plus largement utilisés pour la création de documentation technique, le contrôle de la qualité des documents et l'amélioration de leur réutilisabilité. Du point de vue du traitement automatique des langues, la nouveauté des représentations XML n'est *a priori* pas évidente. Néanmoins, XML possède un large potentiel en tant que *lingua franca*, et draine une large communauté d'utilisateurs vers des pratiques de rédaction où l'explicitation du contenu prend une importance cruciale.

Nous avons donc jugé intéressant d'approfondir les connexions existant entre MDA et XML en développant le prototype MDA-XML, qui implante les principes du système MDA dans un environnement basé sur XML. Nous décrivons dans les sections suivantes l'état d'avancement de notre prototype de rédaction multilingue interactive, et discutons certaines des difficultés rencontrées dans la poursuite de son développement.

Intégration dans XML

Les *feuilles de styles XML*, (eXtensible Stylesheet languages), proposent une notation déclarative pour les transformations de structures XML, elles-mêmes décrites en XML. (Cawsey, 2000) et (Wilcock, 2001) ont décrit quelques expériences de génération automatique en utilisant des feuilles de style XSLT initialement conçues pour la présentation de documents. Comme les transformations d'arbres XSLT ne s'appliquent que sur l'arbre d'entrée sans accès à la sortie partielle d'un traitement intermédiaire, ceci induit une vision "pipeline" du processus de génération avec des séparations de traitement assez artificielles. Dans un contexte interactif (non envisagé par ces auteurs) comme celui de MDA, ceci impliquerait de générer une séquence de transformations après chaque choix de l'utilisateur, les appliquer, récupérer l'arbre transformé et l'afficher. C'est pourquoi ils suggéraient d'autres implantations où les transformations sont réalisées à l'aide de programmes standard. Deux caractéristiques importantes de XML sont la vérification de types *a priori* ainsi que la validation du document accomplie avant le traitement informatique. L'information décrivant les types d'un document XML peut être donnée par des schémas, qui offrent un système de typage flexible et puissant. Les éditeurs XML actuels ne vont cependant pas bien au delà du niveau de la présentation. De nombreuses recherches sont menées concernant le *Semantic Web* ou des extensions de XML pour des applications spécifiques, mais elles se concentrent sur les méthodes de structuration et d'annotation de données, qui seront par la suite présentées de diverses façons ou bien encore traitées informatiquement. À notre connaissance, seul (Boardman, 1999) a aidé des auteurs à produire un texte grammatical et conforme à des schémas XML.

Dans le système MDA, les interactions sont réalisées par le biais du texte de sortie : bien que les choix soient de nature sémantique, l'auteur n'interagit jamais directement avec la structure sémantique mais seulement avec une représentation textuelle de cette structure, les valeurs sémantiques (équivalentes à des balises) étant cachées. Le but d'un système comme MDA n'est donc pas uniquement de produire un document valide comme c'est le cas pour XML, mais également des réalisations textuelles grammaticales. La structure interne du document est constamment mise à jour et les dépendances sémantiques sont propagées entre les différentes parties du texte, et ceci dans l'ensemble des langues couvertes par le système. La notion de présentation (*rendering*) XML est ainsi considérablement étendue dans le système MDA. Nous décrivons dans la section suivante un expérience visant à obtenir les fonctionnalités du système MDA dans un environnement reposant sur les schémas XML.

Description du prototype MDA-XML

Les principes du système de rédaction assistée décrits dans cet article sont implantés au sein d'une interface utilisateur conviviale qui permet une interaction en temps réel avec des grammaires multilingues. Actuellement, le prototype MDA-XML dispose de deux grammaires parallèles en français et en anglais décrivant une classe de notices pharmaceutiques.

L'interface utilisateur (cf. figure 1) comprend une fenêtre de texte par langue ainsi qu'une fenêtre présentant l'arbre sémantique construit par le biais d'interactions dans les fenêtres de texte. Cette fenêtre est présentée ici à des fins d'illustration, mais n'est pas nécessaire dans une phase de production. L'auteur interagit dans les fenêtres de texte en cliquant sur un mot : s'il s'agit d'un non-terminal (cf. le mot *form* dans le menu de la figure 1), un menu apparaît

dans lequel les choix valides sont présentés à l'utilisateur qui sélectionne un élément parmi ces possibilités ou effectue un retour-arrière. Lorsque un choix est effectué pour un langage donné (*cream* dans la figure 1), l'arbre sémantique est mis à jour pour refléter le choix: *pharmForm* prend la valeur *creme* dans la fenêtre sémantique de la figure 2 et le mot correspondant *crème* est produit dans le texte français. Dans ce cas précis, le mode "automatique" du système est actif : toutes les mises à jour obligatoires qui dépendent directement du choix que vient de faire l'auteur sont effectuées. Sur la figure 2, on peut observer que *conditionnement* est devenu *tube*, puisqu'il s'agissait de la seule possibilité compatible avec la forme pharmaceutique *creme*. De la même façon, le mode d'administration *modeEmploi* a été mis à jour par sa réalisation textuelle (à nouveau la seule compatible avec *creme*). Les actions correspondantes sont également réalisées en français.

Comme son nom le suggère, MDA-XML est une reconstruction systématique des idées développées dans le projet MDA, avec cependant certaines variantes dans les choix d'implantation : les grammaires sont décrites dans un format XML contrôlé par un schéma, au lieu de l'être dans la notation propriétaire inspirée de Prolog; toutes les interactions et interprétations sont faites en Java, au sein duquel une version simplifiée (unification de termes atomiques seulement) du mécanisme d'unification de Prolog est réimplantée : l'utilisation d'un interpréteur Prolog comme dans la version actuelle du système MDA n'est plus nécessaire.

Idéalement nous aurions souhaité implanter les règles sémantiques et grammaticales dans des schémas XML. Bien que des expressions régulières ou des contraintes d'occurrence puissent être définies pour des éléments, le standard XML actuel ne dispose que de contraintes limitées entre différents éléments : seules quelques contraintes relativement simples concernant l'unicité d'identificateurs et de références peuvent être imposées. Il existe des propositions plus étendues de modèles de contraintes, mais elles reposent pour la plupart sur des processus externes (par exemple XSLT) et ne sont pas adaptées à un environnement interactif comme MDA-XML. Nous avons donc défini notre propre format de règle : il s'agit d'une adaptation plus ou moins directe du format des grammaires à clauses définies se conformant aux conventions XML et inspirée de la syntaxe des définitions et des appels de "modèles" (templates) des feuilles de style XSLT. Ces règles peuvent être validées syntaxiquement par un schéma, mais il serait intéressant d'en permettre une validation d'ordre sémantique en cours d'édition. Un autre avantage important est que ce schéma peut être utilisé dans le cadre d'un éditeur XML qui peut ainsi suggérer les expansions possibles des règles à chaque étape de leur définition.

Une grammaire "MDA-XML" bilingue français/anglais (environ 2500 lignes de XML) a été développée pour la rédaction assistée de notices pharmaceutiques. Un des aspects intéressants de cette expérience réside dans l'utilisation de standards industriels XML; par exemple les vues "Authentic" et tabulaires de XMLSpy (XML Spy, 2002), qui permettent d'utiliser des suggestions déduites du schéma, simplifient la saisie des règles. Les balises XML sont cachées au développeur dans ce mode. Contrairement au système MDA initial, MDA-XML est entièrement développé en Java et ne dépend donc pas d'une communication avec un interpréteur Prolog externe pour les calculs et la mise à jour du texte, ce qui permet de conjuger portabilité et rapidité. Il présente cependant certains inconvénients, car le pouvoir expressif de l'interpréteur Prolog permet de simplifier les grammaires par l'utilisation du non-déterminisme. Actuellement MDA-XML ne dispose pas d'une implantation complète du mécanisme d'unification, ce qui rend difficile la description de listes récursives corrélées. Comme dans le système MDA, l'interaction se fait par le biais du texte, ce qui nous semble une approche assez naturelle. Cependant, il serait intéressant de voir si l'arbre sémantique peut aussi être utilisé à cette fin. Même si aucune difficulté technique ne se pose pour proposer des choix sémantiques dans la fenêtre de vue

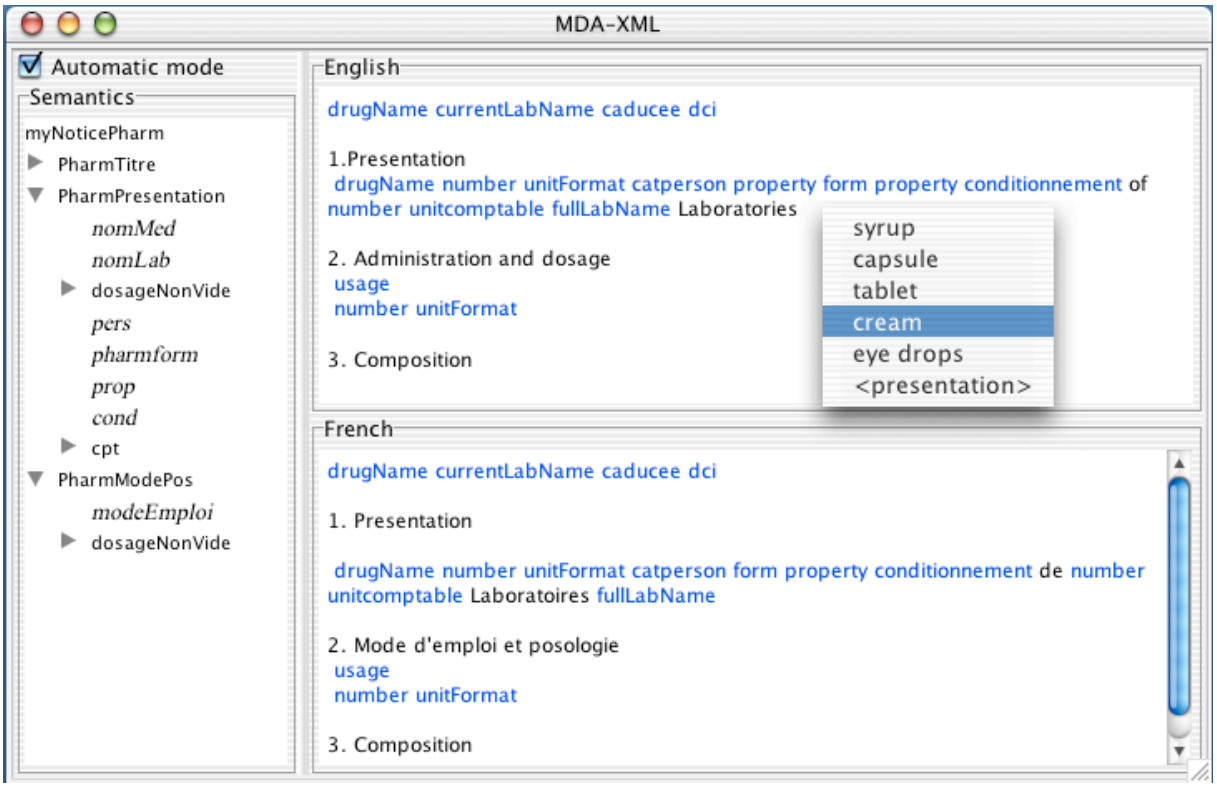

Figure 1: Après que l'utilisateur a sélectionné *form*, un menu des choix possibles lui est présenté.

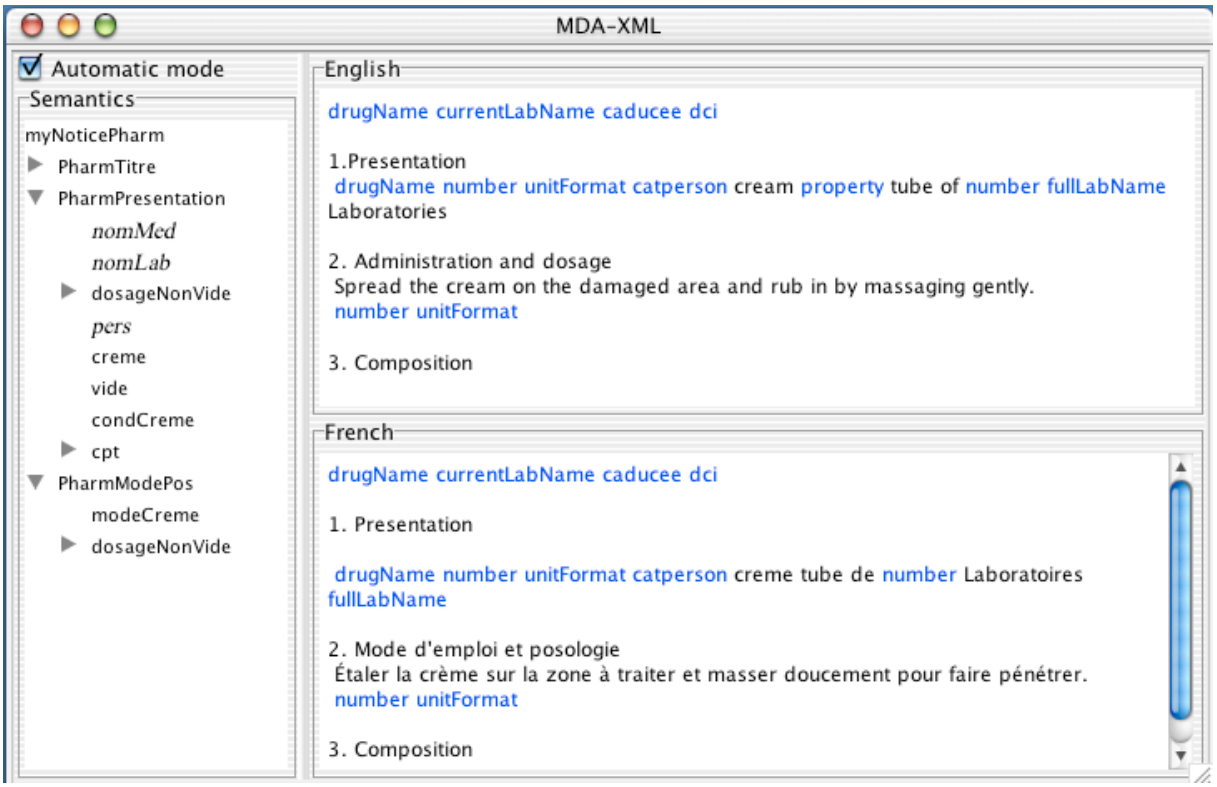

Figure 2: État de MDA-XML après que l'utilisateur a choisi *cream* dans 1. Les choix devenus obligatoires sont mis à jour : *conditionnement* prend la valeur *tube* et *usage* est remplacé par une phrase appropriée. L'information sémantique a aussi été mise à jour dans la fenêtre *Semantics* : *pharmform* prend la valeur *creme*, *cond* prend la valeur *condCreme* et *modeEmploi* prend la valeur *modeCreme*.

de l'arbre sémantique, il faudrait pousser plus avant les expérimentations sur l'interface utilisateur, pour évaluer la convivialité d'une interaction purement sémantique. Actuellement, les grammaires subissent une vérification de cohérence sémantique et de complétude lors de leur analyse par MDA-XML, le schéma actuel ne pouvant exprimer que des vérifications syntaxiques relativement simples. Il serait également intéressant d'approfondir notre réflexion pour intégrer ces vérifications sémantiques dans le schéma lui-même.

Conclusion

Nous avons décrit MDA-XML, un prototype de rédaction contrôlée multilingue développé dans un environnement XML. Les règles de grammaires, écrites en XML, peuvent être vues comme des variantes linguistiquement expressives de schémas XML. Le système, entièrement implémenté en Java (l'analyseur, l'interpréteur de grammaire et l'interface graphique), a été appliqué à une classe de documents pharmaceutiques. Ce travail est un exemple d'utilisation des outils XML dans un contexte de rédaction contrôlée s'étendant au-delà du formatage vers des extensions intégrant des aspects sémantiques et dynamiques dans le processus de rédaction.

Références

BOARDMAN R. (1999). An XML/XSL architecture for language-neutral document authoring. Master's thesis, Centre for Cognitive Science, Edinburgh University.

BRUN C. & DYMETMAN M. (2002). *Rédaction multilingue assistée dans le modèle MDA*, Dans *Multilinguisme et traitement de l'information*, p. 129–152. Traité des sciences et techniques de l'information. Hermès.

BRUN C., DYMETMAN M., FANCHON E. & LHOMME S. (2002). Controlled authoring of biological experiment reports. *à paraître dans EACL'03 demo session*, p. 4 pages.

BRUN C., DYMETMAN M. & LUX V. (2000). Document structure and multilingual authoring. Dans *Proceedings of the First International Natural Language Generation Conference (INLG'2000)*, p. 24–31, Mitzpe Ramon, Israel.

CAWSEY A. (2000). Presenting tailored resource descriptions: Will XSLT do the job? Dans *9th International World Wide Web Conference*.

DYMETMAN M. (2002). Text authoring, knowledge acquisition and description logics. Dans *Proceedings of Coling 2002*, Taiwan.

DYMETMAN M., LUX V. & RANTA A. (2000). XML and multilingual document authoring: converging trend. Dans *Proceedings of the The 18th International Conference on Computational Linguistics (COLING 2000)*, p. 243–249, Saarbruecken: COLING.

POWER R. & SCOTT D. (1998). Multilingual authoring using feedback texts. Dans *Coling-ACL*, p. 1053–1059, Montréal.

RANTA A. (2002). Grammatical framework. a type-theoretical grammar formalism. *to appear in Journal of Functional Programming*.

WILCOCK G. (2001). Pipelines, templates and transformations: XML and natural language generation. Dans *Proceedings of the first XML and NLP workshop*, p. 1–8.

XML Spy (2002). *XML Spy 5 Entreprise Edition Manual*. Altova Corp.