Traitements automatisés de conférences

J.C.Moissinac

GET-ENST-CNRS UMR5141

39 rue Dareau, 75014 Paris – France

moissinac@enst.fr

Résumé

'Cours Chemin' est un projet de mise au point d'outils pour la constitution d'archives de conférences. Ce projet veut offrir des méthodes et des outils qui permettent cet archivage en minimisant les coûts, et en autorisant des modes évolués de consultation, grâce d'une part à des traitements linguistiques, d'autre part à des méthodes de traitement et de production de documents ***multimédias*** *largement basées sur* ***XML****.*

Nous présentons ici la démarche, en l'illustrant par le traitement de conférences dans les domaines du droit et des NTIC. Après une brève présentation de l'objectif poursuivi dans le cas du droit, nous présentons les diverses techniques mises en oeuvre pour atteindre cet objectif. Nous insistons sur les techniques XML mises en oeuvre ou explorées pour la représentation des conférences, la présentation des résultats d'indexation pour validation et la présentation finale.

Nous tirons de l'expérience de ce travail d'une part une illustration de la souplesse obtenue en s'appuyant sur un ensemble de grammaires XML, d'autre part une vision critique de l'état de l'art en matière de représentation de scènes multimédia, et des perspectives ouvertes dans ce domaine notamment par ***MPEG-4****.*

Mots clefs

Indexation multimédia, SVG, SMIL, MPEG-4, XML

1 Représentation et traitement des documents

Nous avons entrepris de confronter les méthodes et langages de représentation de séquences multimédia à des cas concrets d'utilisation. Un des cas choisi est celui de la représentation et de l'exploitation de capture vidéo de cours et de conférences. Nous nous plaçons ainsi dans un domaine où d'importants travaux ont déjà été entrepris [1][3][4].

Ce cas d'utilisation nous parait être significativement intéressant :

- d'une part, il s'agit d'une simplification du cas général, proche de l'exploitation audio/vidéo de base, structuré par un déroulement temporel linéaire, et relativement peu interactif,

- d'autre part, il y a un intérêt fort à joindre à une telle séquence audio/vidéo un certain nombre de données associées : métadonnées descriptives, images ('slides' par exemple), documents joints, références externes… Il est nécessaire de maîtriser des moyens de représenter, d'éditer et de présenter conjointement l'ensemble de ces données.

Dans le cas des conférences du domaine des NTIC, nous avons traité des conférences accompagnées de présentations Powerpoint, accompagnées de leur timing de diffusion. Nous avons souhaité obtenir une représentation de l'évènement filmé indépendante d'une part de la technique utilisée pendant la présentation -ici Powerpoint, et d'autre part indépendante de la présentation effectuée ultérieurement de cet évènement.

Nous avons opté pour la grammaire XML définie par le consortium IMS [9]. IMS propose des schémas de représentation pour diverses ressources pédagogiques, dont des cours ou conférences. Un document IMS contient trois portions principales. Une portion permet de décrire le document à l'aide de métadonnées (résumé, mots-clés, durée, auteur, difficulté, langue...). Une deuxième fournit une description séquentielle de la conférence. A chaque partie est associée un titre, d'éventuelles sous parties et un identificateur d'un groupe de ressources associées (voir ci-après). Dans notre cas, nous associons en plus deux marqueurs temporels indiquant le début et la fin de chaque partie. La troisième portion est constituée de groupe de ressources désigné par un identificateur; ces ressources peuvent être des pages HTML, des images Jpeg ou SVG, des séquences sonores; une des ressources désigne l'enregistrement vidéo de la conférence. Le document IMS produit contient donc une description XML de l'enregistrement de la conférence et des références aux

documents externes associés. Cette description est au coeur des différentes exploitations que nous pourrons faire de la conférence.

Les techniques mises en oeuvre pour la génération automatique de la description IMS à partir du fichier Powerpoint et de l'enregistrement s'adaptent bien à d'autres types de présentation. Des adaptations sont en cours pour OpenOffice et PDF.

Nous sommes dans un contexte qui offre des facilités pour obtenir des résultats exploitables en matière de reconnaissance automatique de la parole -en vue de l'indexation-, et, simultanément, qui présente des champs d'investigation où des progrès importants restent à faire. Au rang des facilités, citons la connaissance du domaine linguistique par la connaissance du sujet de la conférence, une bonne qualité de prise de son, souvent, un locuteur connu. Au niveau des difficultés, la principale concerne le traitement de la parole spontanée sur lequel de nombreux progrès restent à faire. Cependant, même à ce niveau, la parole 'spontanée' en conférence n'est souvent pas si spontanée que cela; elle s'appuie sur des notes préparatoires et comporte des formes récurrentes.

Principalement à l'aide de Sirocco [7], nous obtenons une transcription automatique de chaque conférence au format Trans-13. Nous exploitons un ensemble de ressources linguistiques basées sur un large vocabulaire défini avec sa prononciation [17]. Trans-13 est la grammaire XML définie comme format de sortie de l'outil Transcriber [2] de transcription par un opérateur humain ; nous avons réalisé des transcriptions avec Transcriber pour les données de référence servant à la validation de nos outils. Dans les grandes lignes, ce fichier donne un découpage en phrases, avec un instant de début attaché au début de chaque phrase.

Le fichier Trans-13 nous donne donc une suite de phrase avec leurs timings de début et de fin. Ces timings peuvent être comparés à ceux disponibles dans le fichier IMS pour obtenir un alignement de la transcription sur les données audio, vidéo et images de la conférence. Outre l'utilité de cette transcription en matière d'indexation que nous développons au paragraphe suivant, elle permet également de réaliser certaines opérations d'édition en respectant des contraintes porteuses de sens ; par exemple, grâce à l'alignement obtenu, on peut obtenir la suppression d'un transparent d'une présentation en coupant la vidéo correspondante depuis un début de phrase jusqu'à une fin de phrase.

Les deux fichiers XML, IMS et Trans-13, nous permettent de produire diverses présentations multimédia de l'enregistrement à l'aide de simples transformations XSLT ou de programmes Java exploitant l'API JavaDOM. Nous verrons comment plus loin.

2 L'indexation et la validation

Nous allons illustrer notre démarche dans le cas du droit. Nous travaillons avec l'entreprise Droit In-Situ qui produit des Cd-Roms de consultation de séries conférences sur un thème juridique, tournées en vidéo. Le travail d'indexation est réalisé par des juristes spécialement formés. Notre objectif n'est pas de remplacer ces juristes par une indexation totalement automatique, mais d'aider ces juristes à gagner du temps et de la qualité dans leur travail. Nous sommes donc amené à leur présenter le contenu multimédia de la conférence sous une forme différente de sa présentation finale. Tous les éléments ne sont pas présentés à chaque étape et les modes d'interaction sont très différents de ceux de l'utilisateur final. XML nous donne la souplesse nécessaire pour obtenir ces différentes présentations.

Des traitements linguistiques sur la transcription au format Trans-13 nous permettent de trouver des points d'intérêt dans l'enregistrement: références juridiques, mots-clés, items du plan accompagnés du timing correspondant. Le résultat de ce traitement est un fichier trans-13 dans lequel ont été ajoutés des balises typées -référence, mot-clé...- au niveau de chaque point d'intérêt [13].

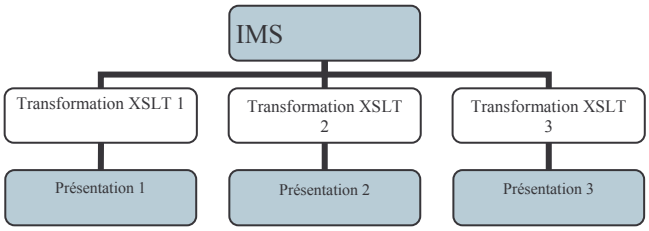

Figure 1 Une représentation (IMS), des présentations

Des transformations XSLT nous permettent de générer des documents HTML+TIME pour présenter les résultats obtenus à un juriste qui va les valider en pouvant facilement écouter l'extrait de conférence correspondant à une point d'intérêt marqué. HTML+TIME a été proposé par Microsoft au W3C en 1998. HTML+TIME est un système simple de description de scènes multimédia exploitant la syntaxe HTML pour la mise en page et la syntaxe SMIL pour la synchronisation de médias. L'ensemble respecte donc une syntaxe XML et peut donc être aisément produit par la transformation du fichier Trans-13. Les dernières versions d'Internet Explorer permettent de jouer des documents HTML+TIME [10].

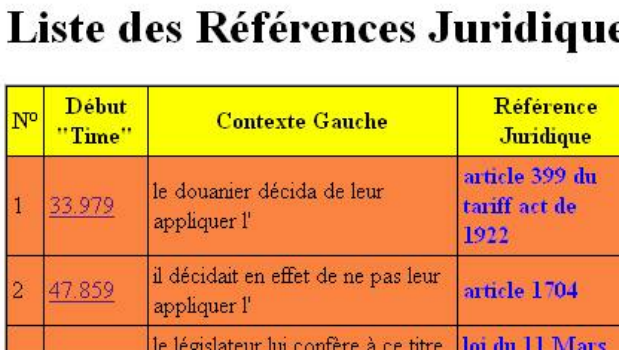

Figure 2 – *Détail d'une iinterface de validation*

3 Les documents finaux

A l'aide de diverses transformations XSLT, nous avons produit des scènes multimédia permettant la présentation des enregistrements de conférences à des utilisateurs finaux. Ce type de production est facilité par l'existence de grammaires XML qui permettent de représenter des documents multimédia.

Une première version est constituée d'une page HTML incluant des scripts pour la synchronisation de médias, des images JPEG pour les transparents, un objet vidéo RealVideo. Il apparaît clairement que ce type de représentation pose d'importants problèmes de portabilité entre navigateurs et systèmes d'exploitation. Cependant, notre solution actuelle fonctionne bien avec le player RealOne sur Macintosh et sur Pc Windows et avec Internet Explorer sur Pc.

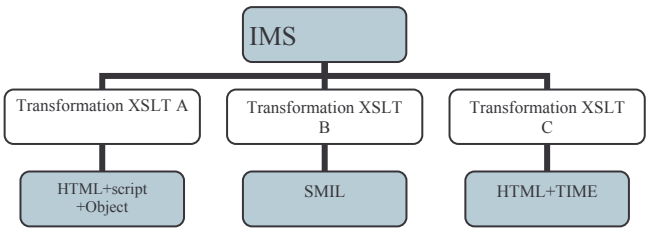

Figure 3 Une représentation (IMS), des techniques de présentation (HTML, SMIL...)

Une variante de ce mode de présentation a été effectué avec des images SVG –Scalable Vector Graphics [18]- au lieu des images JPEG. La nature vectorielle du SVG permet une meilleure adaptation automatique de l'affichage des transparents à la taille de la fenêtre du client qui visionne la séquence. Cette solution pose cependant des problèmes liés à des défauts d'implémentation des players disponibles.

Une autre transformation XSLT permet de produire des documents SMIL. A cette occasion, nous avons observé des différences importantes entre les players capables de jouer des séquences SMIL (nous avons principalement testé Quicktime Player et RealOne Player). En dehors de problèmes temporaires d'implémentation, certains problèmes nous paraissent liés à la spécification SMIL elle-même. SMIL ne spécifie pas comment le player SMIL interagit avec les players de médias élémentaires auquel il fait appel. Il en résulte que rien n'assure une bonne interaction entre ces éléments. SMIL apparaît donc comme un bon langage de spécification de la présentation et de la synchronisation d'une scène multimédia, mais apparaît insuffisant pour en assurer la diffusion dans des conditions prévisibles d'un player à l'autre.

Nous avons entrepris d'exploiter les possibilités de MPEG-4. MPEG-4 intègre des possibilités qui couvrent nos besoins et recouvrent bien les possibilités que nous utilisions précédemment avec HTML+javascript, SVG, SMIL ou HTML+TIME : affichage graphique bitmap ou vectoriel, diffusion audio/vidéo, interactivité. De plus, MPEG-4 permet de se placer dans un environnement de diffusion totalement intégré, garantissant une meilleure portabilité théorique que SMIL, pourvu qu'on utilise de players respectant la norme (profil Core 2D). De plus, cet environnement couvre des considérations de gestion des droits d'accès, de streaming... L'obtention des documents MPEG-4 est un peu plus complexe que pour les formats précédents. Nous utilisons XMT [14], une grammaire qui permet une représentation XML d'une séquence multimédia MPEG-4 BIFS [14] (binaire compressé). Une transformation XSLT nous permet d'obtenir un fichier XMT qui doit ensuite être 'compilé' en scène binaire MPEG-4 BIFS, prête à être jouée dans un player MPEG-4. La compilation peut être effectuée par exemple à l'aide de l'outil mp4box du projet sourceforge GPAC initié par l'ENST [8].

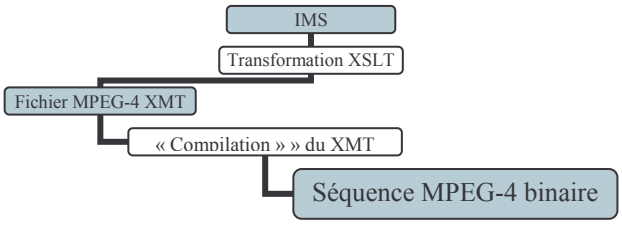

Figure 4 Production de MPEG-4 à partir de XML

Il faut pourtant noter que, pour l'instant, MPEG-4 est mal intégré aux navigateurs et que la diffusion des players MPEG-4 intégrant BIFS est encore relativement

confidentielle, malgré l'existence de la solution OpenSource Osmo4 [8][15].

4 Conclusion

Sur la base de notre expérience de représentation de conférences sous forme de scènes multimédia, nous proposons deux conclusions:

- la première est la grande souplesse obtenue grâce aux représentations XML de scènes multimédia et de données graphiques,
- la deuxième est que SMIL nous apparaît comme un très bon langage de spécification de scènes multimédias, mais que les limites que ses concepteurs se sont données provoque des insuffisances en matière d'intégration de médias et laisse la place à MPEG-4 comme véritable format d'intégration de médias.

Remerciements

Nous remercions la Fondation Louis Leprince-Ringuet qui a soutenu le lancement du projet Cours Chemin, la société Droit In-situ avec qui nous collaborons pour l'exploitation de telles méthodes dans le domaine du droit. François Yvon (ENST) est le principal contributeur pour les questions liées au traitement du langage, notamment le travail sur la transcription automatique des conférences. Slim Ben Hazez a fournit une aide précieuse pour les traitements spécifiques au droit.

Références

[1] G.D. Abowd, L.D. Harvel and J.A. Brotherton "Building a Digital Library of Captured Educational Experiences" Invited paper for the 2000 Int. Conf.on Digital Libraries, Kyoto, Japan, 2000.

[2] C.Barras, E.Geoffrois, Zhibiao Wu and M.Liberman "Transcriber: a Free Tool for Segmenting, Labeling and Transcribing Speech", in the proc. of the First International Conference on Language, 1998.

[3] J.A. Brotherton, Enriching Everyday Experiences through the Automated Capture and Access of Live Experiences: eClass: Building, Observing and Understanding the Impact of Capture and Access in an Educational Domain, Georgia Tech, College of Computing Ph.D. Thesis, December 2001.

[4] J.A. Brotherton, J.R. Bhalodia, and G.D. Abowd (1998). "Automated Capture, Integration, and Visualization of Multiple Media Streams", In the Proceedings of IEEE Multimedia '98.

[5] Cheyrou-Lagreze P., « *Editing SVG with other XML languages*», *SVG Open 2002*, Juillet 2002, Zurich

[6] Concolato C., Dufourd J.-C., « Comparisons of MPEG-4 Bifs and Some Other Multimedia Description Languages», *Workshop and exhibition on MPEG-4*, June 25-27, 2002, San José, California

[7] G.Gravier, F.Yvon, B.Jacob et F.Bimbot (2000). « Sirocco, un système ouvert de reconnaissance de la parole ». Actes des XXIVe Journées d'études sur la parole, Nancy, France, 2000.

[8] GPAC, http://gpac.sourceforge.net/

[9] « IMS Simple Sequencing Information and Behavior Model », Mars 2003, site www.imsproject.org

[10] "Introduction to HTML+TIME", 2004, site http://msdn.microsoft.com/workshop/author/behaviors/time.asp

[11] Francis Kubala & al. (2000). "Integrated technologies for indexing spoken language ». Communications of the ACM, 43:2 pages 48 - 56.

[12] J. Lienhard, T. Lauer (2002). "Multi-Layer Recording as a New Concept of Combining Lecture Recording and Students' Handwritten Notes", in the Proceedings of the 10th ACM International Conference on Multimedia, Juan-les-Pins (France), December 2002.

[13] J.C.Moissinac, F.Yvon, S.Benhazez "Automating Indexing of Classes and Conferences", in the Proceedings of the RIAO Conference, Avignon (France), Avril 2004.

[14] MPEG-4 : « Coding of audio-visual objects – Part 1: Systems», ISO/IEC 14496-1,2000.

[15] Osmo4, site http://www.comelec.enst.fr/osmo4

[16] L.Rabiner, B.-H.Juang (1993), "Fundamentals of Speech Recognition. Prentice Hall, Inc.1993

[17] F.Yvon, C.d'Alessandro, V.Aubergé, P.Boula de mareuïl et J.Vaissière (2000). « Ressource standard pour le français : un large lexique orthographique-phonétique", Actes des JEP, Autrans, France.

[18] W3C, « Scalable Vector Graphics, XML Graphics for the Web », recommendations du W3C, site http://www.w3.org/Graphics/SVG/