

Modélisation DAPA

Présentation du schéma XML DAPA

Réalisé par

AJLSM

17 rue Vital Carles

33000 Bordeaux, France

Téléphone : +33 5 57 14 25 25 // Télécopieur : +33 5 56 44 08 47

Personne contact :

Martin Sévigny, sevigny@ajlsm.com

21 juin 2004

Le schéma XML de la DAPA est un schéma qui peut être utile dans de nombreux contextes, pour de nombreuses applications documentaires, généralistes ou spécialisées. Dans ce document, nous allons expliquer la nature de ce schéma, sa structure générale, ainsi que les contextes d'utilisation où il peut être particulièrement approprié.

Nature du schéma

Selon une approche documentaire relativement classique, on pourrait présenter en premier lieu la nature *rédactionnelle* du schéma DAPA. En effet, l'un des objectifs principaux du schéma est de permettre la création de documents rédigés pour être lus de manière séquentielle, comme une monographie, un article, etc. Cette approche rédactionnelle est à opposer à l'approche *base de données* qui consiste à mettre des informations dans des cases précises et à les restituer selon une certaine logique applicative. On peut par contre faire une analogie entre cette approche et les documents produits à l'aide d'un traitement de texte, même si cette analogie est très superficielle.

Il est toutefois important d'aller au-delà de cette première approche classique pour réaliser que le schéma DAPA permet de créer des *documents structurés*, c'est-à-dire des documents qui contiennent de l'information à propos de leur structure, que cette structure soit logique (séparation en parties, sections, paragraphes, etc.) ou sémantique (rôle de certaines parties du texte). La mise en place d'un système d'information basé sur des documents structurés est essentielle si l'on veut associer à la fois les mérites des documents de nature rédactionnelle – clarté, facilité de lecture, souplesse pour exprimer un discours – et les avantages de l'information structurée telle qu'on la retrouve dans des bases de données – possibilité de manipuler l'information de différentes manières, recherches précises. Ainsi, à partir d'un corpus de documents XML respectant le schéma DAPA, il est possible de créer des outils qui vont effectuer des recherches uniquement dans les titres (pour une meilleure précision), qui vont extraire automatiquement des tables des matières pour les inclure dans un système de diffusion sélective d'information, qui vont produire des listes d'illustrations, etc.

Mais le schéma DAPA permet d'aller beaucoup plus loin que la création de documents structurés de nature rédactionnelle. Il permet de créer de véritables *dossiers électroniques*, où l'on retrouvera toutes les informations numériques – textes, images, animations, vidéo, son – à propos d'une entité, que celle-ci soit une œuvre du patrimoine, une personne, un lieu, etc. Cette notion de dossier électronique fait largement appel à la possibilité de relier entre eux de nombreux documents, de définir le rôle de ces relations mais également des documents à l'intérieur des dossiers et d'intégrer de manière cohérente et structurée des éléments multimédia tels que des images.

Pour arriver à représenter de tels dossiers électroniques – et leurs constituantes – le schéma DAPA propose une approche cohérente : l'utilisation d'un même jeu de métadonnées, basé sur les normes *Dublin Core* et *Geographical Markup Language*, pour décrire l'ensemble des constituantes d'un dossier électronique. Ces métadonnées peuvent ainsi décrire non seulement les documents, mais leurs parties et ce jusqu'à un niveau très fin (on peut ainsi préciser l'auteur ou les droits associés à une section), elles peuvent décrire des objets externes comme des images, mais aussi des relations pour préciser leur rôle ou leur fonction.

Par cette généralisation de l'utilisation d'un même format de métadonnées, il est possible de non seulement constituer des corpus de documents structurés en format XML, mais de leur donner une cohérence d'ensemble qui permet de produire de véritables dossiers électroniques.

Structure générale du schéma

Organisation du contenu

La plupart des documents XML créés en respectant le schéma DAPA adopteront une structure rédactionnelle classique, qui consiste à écrire un document composé de différentes parties, ces parties pouvant elles-mêmes contenir d'autres parties.

Cette structure éditoriale peut prendre différentes formes avec le schéma DAPA. Une première consiste à créer un document complexe qui sera représenté par un élément `book` qui aura un titre et sera composé de parties `part` s'il s'agit d'un document volumineux ou très complexe, ou encore directement de chapitres `chapter` ou même de sections `section`.

S'il s'agit plutôt d'un document plus simple qui fait partie d'un ensemble tel un dossier électronique, il est plutôt intéressant d'utiliser l'élément `article` pour représenter ce document, celui-ci pouvant être composé de sections `section`.

Mais une structure éditoriale de cette nature n'est pas la seule possibilité offerte par le schéma DAPA. En effet, si l'on veut représenter uniquement des données très structurées, voire des métadonnées, l'élément `record` le permet aisément. Cela peut être utile pour, par exemple, récupérer des contenus provenant de bases de données et pour les représenter en XML respectant ce schéma. Par ailleurs, si l'on veut créer un seul document pour plusieurs jeux de métadonnées, telle une base de données complète, l'élément `set` sera utilisé pour contenir un ensemble de jeux de métadonnées `record`.

Métadonnées

Les métadonnées associées à un document, voire une partie d'un document, sont toujours structurées à l'intérieur d'un élément nommé `info`. Cet élément peut se retrouver à l'intérieur de la plupart des éléments sauf ceux qui sont plutôt destinés à venir enrichir le texte. Ainsi, on peut attribuer des métadonnées à un document, une section, un paragraphe, un tableau, une liste, une citation, une procédure, etc. Ce mécanisme est très souple et surtout très puissant, car il permet, pour ne citer que ces deux exemples, d'attribuer des droits d'utilisation spécifiques pour les différentes entités dans un document ou pour leur attribuer une mention de responsabilité telle que l'auteur.

Ces métadonnées sont elles-mêmes des métadonnées qui respectent les conventions et règles du *Dublin Core*¹. Ainsi, partout où l'on peut associer des métadonnées, on peut préciser n'importe quelle propriété Dublin Core, et on peut même ajouter d'autres propriétés libres ou définies par l'utilisateur, tout comme on peut associer des coordonnées géographiques complexes en respectant la norme GML².

Enrichissement du texte

Partout où il y a du texte structuré en schéma DAPA, on peut enrichir ce texte de différentes manières. Cet enrichissement peut bien sûr être de nature typographique, mais il peut surtout être sémantique ou hypertextuel. L'enrichissement sémantique permet de baliser dans le texte des concepts ou le rôle des différentes parties du texte. On peut ainsi baliser le texte *monument religieux* de manière à ce qu'il soit explicitement identifié comme un concept issu d'un thésaurus et que ce concept décrit le propos du document. Ainsi, un outil pourra automatiquement ajouter de la valeur en permettant une recherche sur ces concepts ou encore des liens hypertextes automatiques pour voir d'autres documents indexés par ce concept, etc.

L'enrichissement hypertextuel permet de pousser encore plus loin l'idée du dossier électronique. En effet, en ajoutant dans les documents des liens vers d'autres documents ou parties de documents, pas uniquement pour structurer un contenu général mais aussi pour établir des relations très fines de type *voir aussi* ou *a été créé par le même architecte*, un corpus peut devenir un véritable environnement hypermédia pertinent et efficace pour les utilisateurs.

Aller au-delà du schéma

Dans un schéma XML, ou toute autre forme de grammaire telle une DTD, on peut définir un certain nombre de contraintes que les documents devront respecter pour être *valides* selon cette grammaire. Ces contraintes sont toutefois fort limitées, et se résument à :

- Le nom des éléments, autrement dit les éléments qui peuvent être utilisés dans le document, voire dans certains cas l'espace de noms auquel ils doivent appartenir.
- Les attributs que l'on peut associer à un élément, et le caractère obligatoire ou non de ces attributs.

¹ Voir <http://dublincore.org/> .

² *Geographical Markup Language*, voir <http://www.opengis.org/docs/02-023r4.pdf> .

- Les valeurs que peuvent prendre des attributs ou des éléments, en fonction de règles simples, de patrons de valeurs définies par des expressions régulières, de listes de valeurs ouvertes ou fermées, etc.
- Les éléments que l'on peut inclure à l'intérieur d'un élément, de même que la possibilité d'y ajouter du texte ou non.
- Le caractère obligatoire ou répétable des éléments.

Ces contraintes sont en général suffisantes pour permettre de mettre au point un système d'information complet et robuste au sein d'une institution, d'un groupe, d'un projet, etc. Toutefois, lorsque le schéma a été défini dans un autre contexte ou dans un contexte plus large, plus générique, il est fort probable que les contraintes qu'on y trouve seront trop génériques ou pas totalement appropriées. C'est pourquoi, dans un tel contexte, il est nécessaire de réfléchir à d'autres façons de mettre en place les contraintes nécessaires.

Ce commentaire s'applique parfaitement au schéma DAPA. En effet, même s'il fut réalisé dans le cadre d'un projet précis initié par la direction de l'architecture et du patrimoine, il reste relativement générique, justement dans le but de permettre aux utilisateurs – gestionnaires de projets, auteurs, contributeurs – de créer de l'information appropriée en contenu et en structure. On peut donc qualifier le schéma DAPA de *générique*, et c'est pourquoi il est utile de réfléchir à d'autres moyens d'apporter des contraintes dans un système d'information XML.

Une première approche serait de définir son propre schéma institutionnel ou projet, en s'assurant que les documents qui respectent ce schéma puissent être convertis automatiquement dans un format qui respecte le schéma initial, et ce dans le but de rester compatible avec ce schéma. Cette approche a l'avantage de pouvoir être très précise et très conforme aux souhaits de l'institution ou du projet, mais en contrepartie elle demande des efforts importants pour la mettre en place et elle ne facilite pas les mises à jour si jamais le schéma initial évolue.

C'est pourquoi nous proposons d'autres approches qui consistent à apporter de nouvelles couches de contraintes au-delà du schéma XML, couches qui pourraient s'appliquer à différents endroits dans un système d'information gérant des documents XML.

L'outil de production

Au-delà de la grammaire elle-même, l'outil de production peut offrir des contraintes additionnelles. C'est le cas, par exemple, lorsque l'outil nous permettra de choisir des termes d'indexation dans une liste pour les insérer dans un élément XML, sans que cette liste ne soit définie dans le schéma. Ou encore, si l'on programme l'outil de production pour rendre un élément obligatoire en avertissant l'utilisateur, au moment de la sauvegarde du document, si cet élément n'est pas présent.

La vérification des documents

On peut également imaginer une étape de vérification des documents, externe à l'outil de production, qui permettrait d'effectuer un certain nombre de vérifications plus sophistiquées ou moins précises. Par exemple, on pourrait implémenter une vérification de l'ensemble du document pour voir si la structure générale (chapitres et sections par exemple) respecte les consignes de rédaction propres à un groupe ou un environnement. On pourrait également vérifier si les termes d'indexation saisis dans le document correspondent bien à des entrées d'un vocabulaire contrôlé qui serait accessible dans un autre outil.

Cette couche de contraintes peut être préférable à celle offerte par l'outil de production dans le cas où les contraintes ne sont pas obligatoires. En effet, le retour de cette étape de validation pourrait être à la fois des suggestions et des éléments d'information, pas seulement des obligations de modifications. Par ailleurs, lorsque des contraintes ne peuvent être vérifiées directement dans l'outil de production (par exemple si l'accès à des ressources n'est pas possible ou si l'environnement de production ne peut pas être programmé pour le faire), cette approche peut être très utile et très performante.

La vérification des corpus

Les contraintes par vérification des corpus sont en quelque sorte différentes des autres couches introduites ici car elles s'effectuent pas uniquement sur un document XML, mais sur l'ensemble d'un corpus auquel appartient le document. Par exemple, un document XML pourrait être vérifié avec les autres documents faisant partie d'un dossier électronique, ou parmi l'ensemble des documents d'un corpus de dossiers électroniques.

Cette approche est utile pour vérifier l'intégrité du corpus, notamment les liens qui peuvent exister entre les différents documents. Ou encore pour vérifier l'unicité des identifiants.

En choisissant une approche qui consiste à définir un schéma XML générique et des guides d'utilisation de plus en plus spécifiques, nous avons fait le pari que de telles couches de contraintes allaient être mises en place afin de s'assurer que dans un contexte précis d'utilisation du schéma, une cohérence plus grande que celle contrainte par ce dernier allait pouvoir être mise en place. Toutefois, il convient de souligner que les bénéfices d'un schéma générique dépassent largement cet inconvénient.