Génération de documents multimédia adaptatifs dans une perspective analytique

Julie Bourbeillon, Catherine Garbay, Françoise Giroud

Génération de documents multimédia adaptatifs dans une perspective analytique

Application : Conception et évaluation de Tissue Microarrays

Julie Bourbeillon — Catherine Garbay — Françoise Giroud

Laboratoire TIMC-IMAG,
IN3S, Faculté de Médecine,
38706 La Tronche cedex
{Prenom.Nom}@imag.fr

RÉSUMÉ. *Le volume de documents disponible sur l'Internet rend difficile les tâches de recherche d'informations. Parallèlement, des efforts vers la personnalisation de l'Internet permettent l'adaptation des pages web à l'utilisateur. Pour ce faire, les travaux s'orientent vers les concepts des systèmes d'informations adaptatifs et de la génération de résumés de collections de pages. La plupart de ces systèmes cherchent à rendre compte d'une diversité par un synthèse ou à rendre compte d'une cohérence narrative par une organisation spatiale. Nous nous proposons quant à nous d'extraire une collection représentative constituant une cohérence thématique selon une perspective analytique afin de générer un document synthétique qui pourra servir de support à une fouille de données. Le système envisagé sera dans un premier temps consacré au domaine médical et en particulier à la génération de documents autour de la technologie des Tissue Microarrays.*

ABSTRACT. *The amount of documents available on the Internet makes information retrieval a difficult task. In parallel some efforts towards the Internet personalization allow web pages user-adaptivity. In order to fulfil those objectives current works rely on adaptive information systems and documents summarizing concepts. In this context most systems intend to represent a diversity through a synthesis or a narrative coherence through spatial organisation. We intend to extract a representative collective presenting a thematic coherence according several analytic points of view in order to generate a synthetic document which could be used for data mining. The considered system will be at first designed for the medical field and in particular in the Tissue Microarray technology field.*

MOTS-CLÉS : *système d'information adaptatif, document multimédia, document de synthèse, domaine médical, Tissue MicroArrays .*

KEYWORDS: *adaptive information system, multimedia document, synthetic document, medical domain, Tissue MicroArrays.*

1. Introduction

Le volume de documents disponible sur l'Internet rend problématique les tâches de recherche d'informations. Parallèlement, des efforts vers la personnalisation de l'Internet permettent l'adaptation du contenu et de l'apparence des pages web à l'utilisateur. Afin de remplir ces objectifs, la communauté scientifique s'oriente vers les concepts des systèmes d'informations adaptatifs et de la génération de résumés de collections de pages. Dans ce contexte, la plupart des systèmes d'extraction d'informations cherchent à rendre compte de la diversité des documents analysés par une synthèse ou à rendre compte d'une cohérence narrative entre ces documents par une organisation spatiale. Ces processus sont guidés soit par un profil utilisateur, afin de ne prendre en compte que les informations d'intérêt en fonction de ce qui est connu de l'usager, soit par des critères intrinsèques à l'application elle-même.

Les points de vues ainsi générés sur la collection de documents apportent une vision épurée qui limite le nombre de documents à compulser pour l'utilisateur humain. Ces techniques facilitent le travail de recherche de documents ou données intéressants. Mais pour ce faire, des documents sont éliminés ou condensés, parce que moins pertinents par rapport au but recherché, ce qui réduit le champ couvert par les documents. Cette vision résumée sur la collection de documents est réalisée au détriment de la diversité intrinsèque à toute collection. La perte ou du moins l'altération de cette dimension statistique limite les possibilités de fouille de données et le nombre d'axes de lecture sur l'extrait de la collection. Or une analyse statistique peut être elle aussi source d'informations complémentaires.

Nous nous proposons donc de mettre en place un système d'information adaptatif permettant d'extraire une collection représentative constituant une cohérence thématique, selon une perspective analytique, qui pourra ultérieurement servir de support à une fouille de données. Le système devra construire à la volée une représentation de la collection de documents selon une requête utilisateur exprimant le but de l'étude à réaliser. Pour être utile, cette représentation devra consister en un ensemble de données hétérogène et complexe, regroupant textes, images, vidéos, sons, données brutes. Elle résultera d'un échantillonnage raisonné guidé par une requête utilisateur. Cela conduit à considérer cette représentation comme un document multimédia virtuel synthétique.

Le système envisagé sera dans un premier temps consacré au domaine médical et en particulier à la génération de documents autour de la technologie des Tissue Microarrays (TMA) (Kallioniemi *et al.*, 2001)(Hoos *et al.*, 2001). Cette technologie récente est déjà très utilisée en oncologie Elle permet, en complément d'études moléculaires globales, la visualisation rapide des cibles moléculaires (séquences ADN, ARN ou protéines) *in situ* dans des milliers de spécimens de tissu à la fois.

Dans cet article nous présentons une approche préliminaire pour formaliser le processus d'adaptation. Après une revue de l'état de l'art dans la section 2, la section 3 donne un aperçu de notre approche du problème. La section 4 présente les besoins

de représentation des connaissances pour la mise en place du système. La section 5 introduit un début de réflexion sur l'architecture du futur système. La section 6 présente une application au domaine des TMA.

2. État de l'art

L'adaptation à l'utilisateur et à la tâche sont des problématiques centrales de la personnalisation. Celle-ci fait l'objet de nombreux travaux au sein de la communauté des systèmes d'information :

— adaptation à l'utilisateur : de plus en plus de systèmes d'information visent à adapter leur réponse et leur apparence à un profil utilisateur. Pour ce faire, ces systèmes, et plus spécifiquement l'hypermédia adaptatif et le web adaptatif (Brusilovsky, 2002) (Wu *et al.*, 2000), construisent un profil utilisateur en extrayant le comportement de l'utilisateur à partir des traces d'exécution du système (systèmes adaptatifs) et / ou de préférences saisies par l'utilisateur (systèmes adaptables). L'apparence et le contenu des pages sont alors générés en fonction du profil utilisateur courant. Ces approches se rencontrent entre autres dans le domaine du e-commerce. Par delà cette simple personnalisation selon des préférences, des systèmes comme ceux des journaux en ligne (Iksal *et al.*, 2002), du e-Learing ou des assistants à la réalisation de présentations multimédia (Bes et al., 2002) visent à présenter de surcroît une cohérence narrative par une organisation spatiale des éléments qu'ils manipulent. Par rapport à de telles approches, le système envisagé cherche de plus à atteindre une organisation thématique représentative par rapport à une requête qui par nature est analytique. En conséquence, il semble nécessaire de se focaliser sur la tâche.

— adaptation à la tâche : actuellement, le modèle de tâche n'est souvent pas explicitement distingué du modèle utilisateur. Or la tâche, exprimée par le biais de la requête utilisateur, doit jouer un rôle prééminent dans le système envisagé. En effet, comme dans tout système de Recherche d'Information, il s'agit de retourner des informations pertinentes en fonction d'une requête utilisateur. En outre, comme pour les systèmes de suivi d'actualités (McKeown *et al.*, 2002), de détection du changement ou de résumé de pages web (Google Actualités[1] ou NewsMap[2] par exemple), l'information retournée ne se limite pas à une liste ordonnée par pertinence vis à vis de la requête. Elle doit plutôt être vue comme une collection organisée de documents. Enfin, le système envisagé ne se limite pas à construire une collection organisée : les données collectées doivent pouvoir faire l'objet d'une fouille de données et non d'une simple consultation. En conséquence, les données doivent être sélectionnées et organisées pour présenter un aperçu de la collection complète. Cet aperçu est réalisé par le biais d'un extrait représentatif conservant la variété présente dans la collection originale et permettant des analyses statistiques.

[1] http://news.google.com
[2] http://www.marumushi.com/apps/newsmap/newsmap.cfm.

Afin de mettre en place ces adaptations à l'utilisateur et à la tâche, il est nécessaire d'acquérir et représenter des connaissances sur la tâche d'adaptation et le domaine d'application. L'utilisation d'ontologies (Gruber, 1995) (Guarino, 1998) facilitera aussi l'acquisition, la modélisation et le partage de connaissances par des agents logiciels et humains.

En ce qui concerne l'implantation elle sera basée sur les langages classiquement utilisés dans l'infrastructure du web sémantique (Ding *et al.*, 2004). Ces langages sont utilisés pour extraire ou regrouper des connaissances, ou raisonner sur les informations contenues au sein de ressources accessibles sur l'Internet. On peut citer par exemple XML, XSLT, RDF, OWL.

3. La tâche de synthèse : vue générale

3.1. *La synthèse comme conception d'un document multimédia*

L'objectif du système est de fournir un point de vue statistiquement exploitable sur une collection de documents. Par exemple le système pourrait être consacré à l'étude de la gestion du territoire et permettre une approche analytique sur une collection de documents comprenant des images satellites, des cartes, des éléments cadastraux, des informations concernant les surfaces consacrées à différents types d'exploitations (forêts, cultures, prairies, zones industrielles, zones urbaines...), des textes de directives et réglementations officielles, des rapports économiques...

Étant donnée une requête, la représentation à construire peut être considérée comme un arrangement de documents, références à documents ou extraits de documents, chacun de ces éléments étant caractérisé par un ensemble d'informations issues d'un dépôt de données. En conséquence, cette représentation peut être considérée comme une collection de documents multimédia qui est générée en fonction d'une question utilisateur. Ce document de synthèse inclut:

— la requête utilisateur : celle-ci permet de définir l'étude à réaliser sur la collection de documents. Dans le cas de l'exemple précédent, il peut s'agir de l'évolution au cours du temps de la forêt amazonienne entre 1950 et 2000,

— une grille documentaire : il s'agit de l'assemblage de documents sélectionnés et organisés spatialement selon la requête utilisateur. Dans l'exemple précédent, l'assemblage logique serait de consacrer chaque case de la grille à une année entre 1950 et 2000 et de proposer un classement chronologique. Chaque élément de la grille peut être associé à d'autres éléments:

- document correspondant complet : pour le cas considéré, il s'agirait d'une fiche présentant la forêt pour une année donnée, comportant par exemple une image satellite, une carte, la surface boisée, la surface défrichée, des liens vers les réglementations, les rapports économiques ou scientifiques émis cette année là,

- autres données associées au document : auteur, date de publication, indice de confiance...

Dans le futur on peut envisager d'inclure des références à des requêtes similaires conduites avec l'outil ou de la bibliographie pertinente.

Cette description sommaire des éléments d'un document de synthèse peut conduire à la réalisation d'un modèle de document de synthèse spécifique d'un domaine d'application. Ce modèle de document pourrait être paramétrable au sein de la requête ou par des préférences stockées dans un profil utilisateur.

Les documents générés peuvent servir de support pour des publications dans les revues pertinentes du domaine et permettre d'enrichir la base de connaissances.

3.2. *La synthèse comme tâche complexe*

À partir des données et des connaissances disponibles, l'outil à concevoir doit générer à la volée un document multimédia de synthèse, soit construire une vue sur la collection de documents selon une requête. Il s'agit d'un problème d'**adaptation** complexe qui peut être décomposé en trois sous-problèmes:

— problème de **sélection** : construire un document de synthèse est assimilable à la recherche, au sein d'une liste d'éléments (l'ensemble originel de documents), d'une collection (la liste des documents pertinents) qui correspondent à la demande (la requête utilisateur) et respectant certaines règles générales,

— problème d'**organisation spatiale** : il s'agit ensuite de placer des objets (documents, références à documents ou portions de documents) sur une grille,

— problème de **présentation** : l'ensemble précédemment généré doit être présenté à l'utilisateur sous la forme la plus lisible et conviviale possible, en respectant d'éventuelles préférences.

La complexité du problème est liée à plusieurs facteurs:

— données à manipuler : cette génération implique l'agrégation d'un ensemble d'objets de nature complexe. Ces données peuvent être hétérogènes dans leur type : textes, images, vidéos, données brutes.. Certaines peuvent être pérennes et d'autres à validité limitée dans le temps ou l'espace. Certaines peuvent être quantitatives et d'autres qualitatives. Ces données peuvent être appréhendées à diverses échelles. De plus, les points de vue sur ces données sont multiples, dépendants de l'utilisateur ou de l'objectif qu'il veut atteindre. Enfin, la difficulté peut être accrue par la combinatoire importante associée, si la collection originelle présente un volume important de documents, et par la diversité des documents à construire.

— but de la requête : la requête n'est pas limitée à des critères d'inclusion / exclusion et sa finalité est différente du simple retour d'informations pertinentes, puisqu'elle vise à proposer une collection de données et documents qui serviront de support pour une analyse statistique.

4. Ontologies pour la synthèse

4.1. *Ontologie de domaine*

Afin de proposer un document de synthèse adapté au mieux à la question posée par l'utilisateur il faut tout d'abord formaliser et représenter les connaissances concernant le domaine d'application : cette formalisation passe par la représentation des objets et concepts du domaine ainsi que des relations existant entre eux, soit une **ontologie de domaine**. Dans l'exemple utilisé précédemment, il faudrait constituer une ontologie géographique, économique, politique, agronomique...

4.2. *Ontologie de tâche*

4.2.1. *Ontologie de la requête*

Il faut représenter comment manipuler les objets et concepts inclus dans l'ontologie de domaine afin de permettre l'adaptation du document de synthèse à la question de l'utilisateur. Il s'agit d'une **ontologie de tâche** qui comprend en premier lieu une ontologie de la requête. Considérons l'exemple évoqué précédemment : « évolution au cours du temps de la forêt amazonienne entre 1950 et 2000 ». La représentation de la requête doit inclure :

— objectif : liste ordonnée d'éléments qui guideront l'organisation spatiale des documents sur la grille de la synthèse. Dans l'exemple du Tableau 1 il s'agit de l'évolution au cours du temps représentée par l'élément [Temps] [Année],

— critères d'inclusion : série de contraintes sur des valeurs d'éléments associées aux documents qui permettront de sélectionner les documents pertinents à intégrer. Dans l'exemple, ces critères portent sur la région (l'Amazonie), le type d'exploitation (la forêt), la période de temps à considérer (entre 1950 et 2000) et sont représentés par les éléments: [Région] = [Amazonie], [Type exploitation] = [Forêt], [Année] => [1950] et [Année] =< [2000],

— modèle logique de document : squelette de document de synthèse qui permettra de définir quelles sont les informations à présenter dans le document de synthèse. Pour le même exemple, il s'agira d'un modèle générique pour une étude territoriale spécifié par le couple [Modèle de Document] = [Analyse territoriale],

— préférences : ensemble de valeurs pour des paramètres associés au modèle de document choisi, permettant d'affiner ce modèle, par exemple nombre d'éléments par page ou par case de la grille. Aucune préférence n'étant spécifiée dans la requête d'exemple, une taille de grille par défaut sera utilisée.

Cette décomposition devrait permettre de construire une ontologie de la requête spécifique à chaque domaine d'application, comme présenté dans le Tableau 1.

Élément de la requête	**Élément du problème**	**Formalisation** **([Élément père]...) [Élément] (= [Value])**
Objectif	Organisation	[Objet de l'étude] [Temps] [Année]
Critères d'inclusion	Sélection	[Critère d'inclusion] [Région] = [Amazonie] [Critère d'inclusion] [Type exploitation] = [Forêt] [Critère d'inclusion] [Année] => [1950] [Critère d'inclusion] [Année] =< [2000]
Modèle logique de document	Sélection / Organisation / Présentation	[Modèle de Document] = [Analyse territoriale]
Préférences	Sélection / Organisation / Présentation	[Taille Grille] = [défaut]

Tableau 1. *Exemple de requête formalisée où les éléments sont référencés selon l'ontologie de la requête*

4.2.2. *Collection de critères*

Comme exposé précédemment, la génération d'un document de synthèse est un processus de sélection et d'organisation de documents sur une grille, puis de présentation de cette grille à l'utilisateur. C'est un processus d'adaptation du document final à une requête qui est dirigé par un ensemble de critères. Une partie de ces critères sont propres au Niveau Domaine d'application, d'autres sont spécifiés au Niveau Requête. Des préférences utilisateur, exprimées en accompagnement de la requête ou au sein d'un profil peuvent influer sur ces ensembles de critères.

Une vue organisée de l'espace des critères est présentée dans le Tableau 2. Cette hiérarchie doit s'affiner et se spécialiser en fonction du domaine d'application spécifique pour constituer une collection complète de critères.

5. Procédure d'adaptation

5.1. *Modèles d'adaptation*

Étant donnée cette collection de critères, générer le document de synthèse va consister à sélectionner et ordonner un groupe de critères pertinents puis de les spécialiser en fonction de la requête et des préférences utilisateur afin de proposer un **plan d'adaptation**. Mais la manipulation d'une telle collection de critères est un problème complexe, du fait de sa taille et de possibles contradictions. Il s'agit alors de trouver un moyen pour faciliter cette manipulation.

Une analyse de différents plans d'adaptation suggère que pour un même domaine d'application, il existe des familles de requêtes et qu'il est possible de construire des

Type de critères	*Critères*	
Sélection	*Spécifiques au domaine*	
		Critères de représentativité : ils correspondent à une approche statistique d'échantillonnage proche de celle utilisée pour les sondages d'opinion
		Critères de validité : les documents inclus doivent être valides dans le contexte spatio-temporel de la requête (par exemple des documents trop anciens ne seront pas toujours pertinents)
		Critères d'économies : il faut chercher à réutiliser des listes de documents sélectionnées pour d'autres requêtes
	Spécifiques à la requête	
		Critères d'inclusion : voir Tableau 1
Organisation spatiale	*Spécifiques au domaine*	
		Critères physiques : la taille de la grille est limitée et la grille a une géométrie particulière
		Critères d'économies : il faut chercher au maximum à réutiliser des portions de grilles existantes
	Spécifiques à la requête	
		Objectif de l'étude : voir Tableau 1
Présentation	*Spécifiques au domaine*	
		Préférences utilisateur : il faut prendre en compte le profil utilisateur
	Spécifiques à la requête	
		Préférences : voir Tableau 1
		Modèle de document : voir Tableau 1

Tableau 2. *Vue hiérarchisée de l'espace des critères*

modèles d'adaptation pour résoudre le problème. Un modèle d'adaptation consiste en un ensemble de critères plus spécialisé extrait de la collection complète. Dans le cadre de la gestion du territoire, des familles de requêtes pourraient être:

— « Évolution au cours du temps » : requête du paragraphe 4.2.1 par exemple,

— « Comparaison d'évolutions » : un des axes de la grille est alors utilisé comme axe temporel et l'autre comme axe géographique,

— « Bilan régional » : à une date donnée, les documents concernant les aires géographiques de la région étudiée sont organisés sur la grille en fonction de leurs positions géographiques réelles relatives.

Afin de réaliser l'adaptation du document de synthèse à la requête utilisateur, ces modèles d'adaptation doivent être spécialisés avec des données extraites de la requête. Par exemple, la spécialisation du modèle « Évolution au cours du temps » pour la requête du paragraphe 4.2.1 consiste à y ajouter entre autres les critères d'inclusion tels que la région ([Critère d'inclusion] [Région] = [Amazonie]).

En conséquence les modèles d'adaptation constituent un niveau de représentation dépendant du but de l'étude. Ce Niveau Objectif est intermédiaire entre les Niveaux Domaine et Requête. Son utilisation facilite ainsi la génération du plan d'adaptation et sa spécialisation en fonction d'une requête spécifique.

5.2. *Décomposition du processus d'adaptation*

Partant de la section précédente, trois niveaux de spécialisation, dont la dépendance vis à la vis de la requête est croissante, peuvent être définis:

— Niveau Domaine : ce niveau contient la collection complète des critères pour le domaine d'application considéré,

— Niveau Objectif : ce niveau contient un extrait de la collection précédente, sélectionné, organisé et paramétré en fonction de la famille de requêtes correspondante, appelé modèle d'adaptation,

— Niveau Requête : ce niveau contient un modèle d'adaptation spécialisé en fonction de la requête courante et d'éventuelles préférences utilisateur.

Parallèlement, trois niveaux d'adaptation, correspondant aux trois sous-problèmes à résoudre peuvent être proposés:

— Niveau Factuel : il correspond à l'étape de sélection où les données ou faits sont analysés afin de proposer une liste d'éléments pertinents,

— Niveau Logique : il correspond à l'étape d'organisation spatiale où un arrangement thématique de la liste précédente est réalisé,

— Niveau de Présentation: il correspond à l'étape de présentation où le document précédent est préparé pour l'affichage en prenant en compte d'éventuelles préférences utilisateur.

La génération d'un document de synthèse adapté à une requête utilisateur implique la manipulation de l'ensemble de ces niveaux par un moteur d'adaptation.

5.3. *Architecture du moteur d'adaptation*

La génération du document de synthèse est un processus en deux étapes, réalisé par le moteur d'adaptation présenté Figure 1. Il faut tout d'abord générer le plan d'adaptation : le parcours des trois niveaux de spécialisation permet de créer un plan d'adaptation spécifique à la requête en cours. Ce parcours passe par :

— Le choix de la Collection de Critères, au Niveau Domaine, pour l'application considérée,

— Le choix d'un Modèle d'Adaptation pour la famille de requêtes au Niveau Objectif,

— La spécialisation du Modèle d'Adaptation en fonction de la requête, qui est construite en suivant un Modèle de Requête spécifique du domaine, au Niveau Requête.

Il s'agit ensuite d'adapter le document TMA à la requête : l'application du plan d'adaptation précédemment généré aux trois niveaux de composition ou étapes de l'adaptation permet de créer le document proposé à l'utilisateur :

— Étape de Sélection : au Niveau Factuel, un processus de composition factuel est utilisé pour sélectionner un ensemble de documents pertinents regroupés au sein d'un Document Virtuel Orienté Collection,

— Étape d'Organisation Spatiale : au Niveau Logique, un processus de composition logique permet d'agencer spatialement la liste précédente en un Document Virtuel Orienté Tâche, en suivant un Modèle Logique de Document,

— Étape de Présentation : au Niveau de Présentation, un processus de composition de présentation sert à préparer l'affichage du Document de Synthèse final, en suivant un Modèle de Présentation de Document.

Cette étude préliminaire a conduit à quelques perspectives d'implémentation. Modèles de Requêtes, Modèles d'Adaptation et Plans d'Adaptation vont consister en

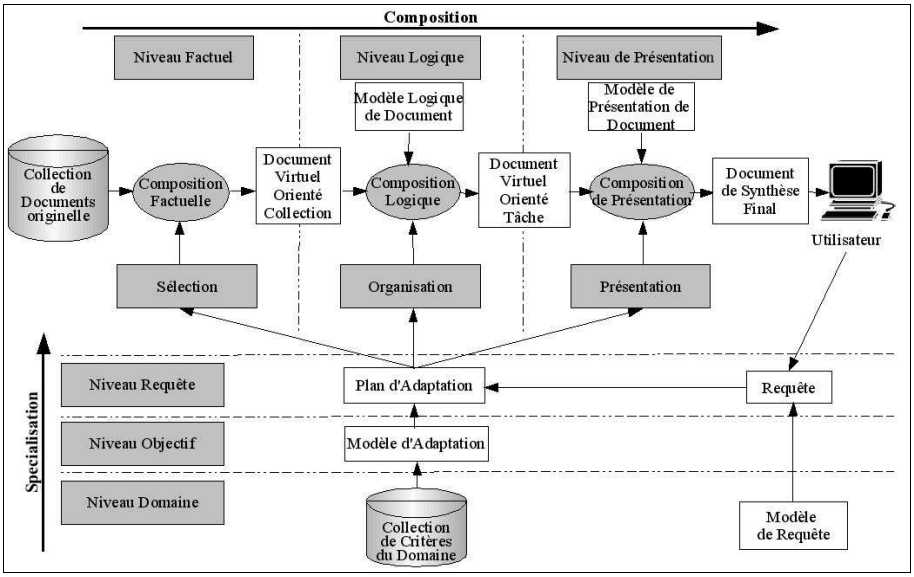

Figure 1. Architecture du moteur d'adaptation

documents XML. Le Modèle Logique de Document sera une DTD XML. Le Modèle de Présentation de Document comportera entre autres des règles XSLT. Les DTDs et des exemples de documents XML sont en cours de réalisation.

6. Application au domaine des Tissue MicroArrays

6.1 *Présentation de la technologie*

La technique des « Tissue MicroArrays » (TMA) est une technologie récente, déjà très utilisée en oncologie pour l'aide au pronostic et au suivi thérapeutique (Kallioniemi *et al.*, 2001)(Hoos *et al.*, 2001). Elle permet, en complément d'études moléculaires globales, la détection rapide des cibles moléculaires (séquences ADN, ARN ou protéines) dans des milliers de spécimens de tissu à la fois. Dans cette technique, on sélectionne, en fonction de l'étude à réaliser, des patients dont des biopsies sont disponibles en archive. Un pathologiste analyse une lame histologique de chacune de ces biopsies et détermine les régions d'intérêt de l'échantillon à étudier. Dans le bloc de paraffine de la biopsie (bloc donneur), des carottes de tissu sont prélevées en correspondance avec les zones pertinentes prédéfinies. Ces carottes sont insérées dans un bloc receveur vierge (bloc TMA) à partir duquel des lames sont réalisées et traitées comme le seraient des lames histologiques conventionnelles. Ces lames TMA font alors l'objet d'une acquisition d'image à différents grossissements. Les images sont ensuite partitionnées en images individuelles de spots (correspondant à la coupe de chaque carotte du bloc), qui font l'objet d'une annotation anatomopathologique et d'une analyse d'image pour quantification de marquage...

Par rapport à des études menées avec des techniques classiques, celles utilisant la technologie des TMA permettent des économies de réactifs et de matériel biologique. De plus, le traitement en masse d'une collection d'échantillons de tissus apporte une dimension statistique au travail du pathologiste. Ces deux avantages peuvent être encore accentués par le recours au concept de lame TMA virtuelle : des images de spots existantes peuvent être sélectionnées et réagencées en fonction d'une nouvelle étude sans nécessiter la construction d'un nouveau bloc.

Même si cette technologie semble prometteuse elle souffre d'un manque de connaissances formalisées et d'automatisation de la préparation du plan d'expérience et de la fouille de données. Étant donné la complexité de ces tâches et considérant la grande quantité de données à manipuler il apparaît nécessaire de concevoir un système informatique pour assister à la réalisation de ces tâches.

Or, même si un grand pas a été franchi avec la définition de la « TMA Data exchange specification » (Berman *et al.*, 2003), les outils développés autour de la technique se consacrent surtout à de la gestion de données (Henshall, 2003) (Shergill *et al.*, 2004). Il paraît donc nécessaire de proposer un outil d'assistance à l'utilisation de cette technologie intervenant à deux étapes du cycle présenté ci-dessus :

— aide à la conception de blocs TMA réels, par génération de représentations virtuelles de blocs TMA à fabriquer en fonction de l'étude à réaliser,

— accompagnement de la fouille de données par génération de lames TMA virtuelles associées à des informations pertinentes pour l'étude en cours.

6.2. *La conception de TMA comme tâche de synthèse*

Étant donné une requête, il s'agit de proposer une représentation de lame ou bloc TMA, consistant en un arrangement de spots ou carottes sur une grille, chaque spot ou carotte étant caractérisé par un ensemble d'informations extraites d'un entrepôt de données. Ces deux types de représentations peuvent être considérés comme des collections de documents multimédia comportant :

— la requête utilisateur, qui permet la définition de l'étude à réaliser en utilisant la technologie (par exemple la comparaison entre deux groupes de patients...),

— une grille TMA, constituée d'un assemblage d'images de spots ou de références de carottes sélectionnés et organisés spatialement en fonction de la requête utilisateur; à chacune peuvent être associées des informations concernant:

– le patient associé : dossier clinique...

– l'analyse et l'annotation d'image : quantification de marquage, description de structures tissulaires...

À l'avenir, on peut envisager d'intégrer des références à des études similaires, de la littérature pertinente (PubMed...), ou des informations concernant les molécules étudiées, tirées de bases de données généralistes telles que SwissProt, GenBank...

Les caractéristiques de la problématique et du document à construire impliquent une bonne adéquation entre le problème de conception de TMA et l'outil de génération de documents de synthèse tel qu'il est envisagé. Les paragraphes suivants présentent donc les ontologies spécifiques à la conception de TMA qui ont été réalisées et les spécificités apparues au sein du processus d'adaptation.

6.3. *Ontologies pour la conception de TMA*

6.3.1. *Ontologie de domaine*

Construire une ontologie de la conception de TMA implique tout d'abord la modélisation du champ de la pathologie étudiée, ici le cancer du côlon. Les ontologies médicales actuellement disponibles sont souvent trop générales (Bodenreider, 2004). Une ontologie du cancer du côlon, répertoriant une centaine de termes du niveau organe au niveau cellule, a été réalisée par un pathologiste (Dr. Simony-Lafontaine, du Centre Régional de Lutte Contre le Cancer de Montpellier).

Il est aussi nécessaire de représenter les autres objets et concepts concernant la technologie (Figure 2). Toutes ces notions ont été intégrées dans une ontologie réalisée avec l'outil Protégé2000 de l'université de Stanford. Cette représentation a guidé la construction d'une base de données relationnelle contenant actuellement les dossiers cliniques et données concernant le matériel biologique pour une centaine de patients.

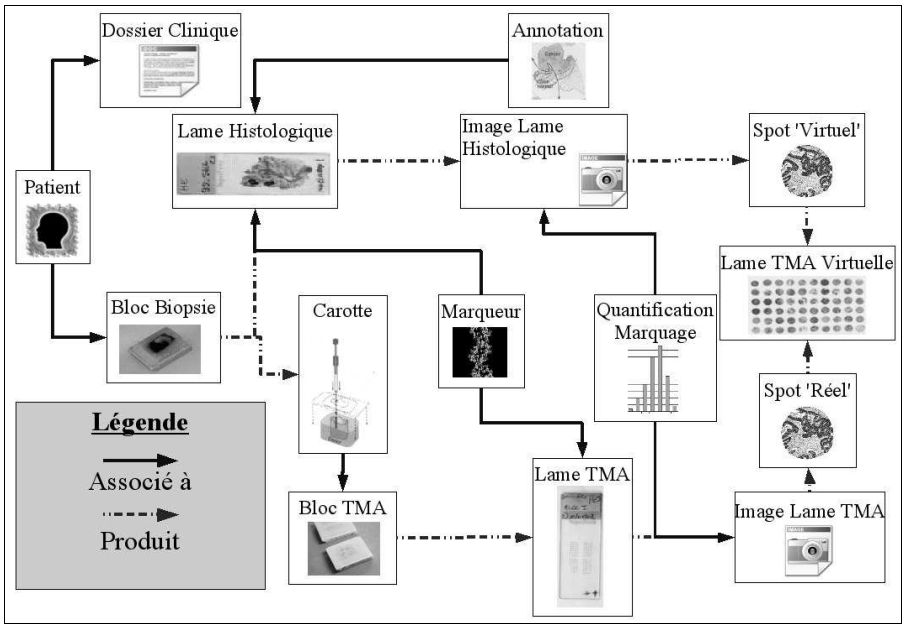

Figure 2. Ontologie simplifiée de la technologie des TMA

6.3.2. *Ontologie de requête*

La requête à formaliser porte sur des éléments du domaine, inclus dans les ontologies précédentes. Ainsi, dans le cas de la conception de TMA, un exemple de question posée par l'anatomopathologiste, exprimée en langage naturel, pourrait être : « Étude de l'évolution du cancer du côlon chez les hommes par une lame TMA virtuelle en utilisant le marqueur Ki67 ». Une telle requête se formalise suivant le modèle général présenté précédemment sous la forme (Tableau 3) :

— un objectif : ici, l'évolution du cancer, représentée par le « stade pTNM »,

— des critères d'inclusion : dans cet exemple, les hommes atteints d'un cancer du côlon, quelle que soit la partie du côlon touchée,

— un modèle de document : ici, lame TMA virtuelle,

— des préférences : utilisées pour spécifier le matériel de laboratoire, dans ce cas le marqueur Ki67.

Une ontologie de la requête spécifique au domaine a donc été réalisée.

Élément de la requête	**Élément du problème**	**Formalisation** **([Élément père]...) [Élément] (= [Value])**
Objectif	Organisation	[Objectif][Patient][Diagnostic] [Stade pTNM]
Critères d'inclusion	Sélection	[Critère d'inclusion][Patient] [État Civil] [Sexe] = [homme] [Critère d'inclusion][Région] = [côlon][tout]
Modèle logique de document	Sélection / Organisation / Présentation	[Modele de Document] = [virtuel]
Préférences	Sélection / Organisation / Présentation	[Matériel] [Taille Grille] = [défaut] [Matériel][Diamètre Aiguille] = [défaut] [Matériel] [Marqueur] = [Ki67]

Tableau 3. *Formalisation d'un exemple de requête pour de la conception de TMA*

6.3.3. *Processus d'adaptation*

Le processus d'adaptation repose sur une spécialisation à trois niveaux de dépendance croissante vis à vis de la requête et une composition entre trois étapes

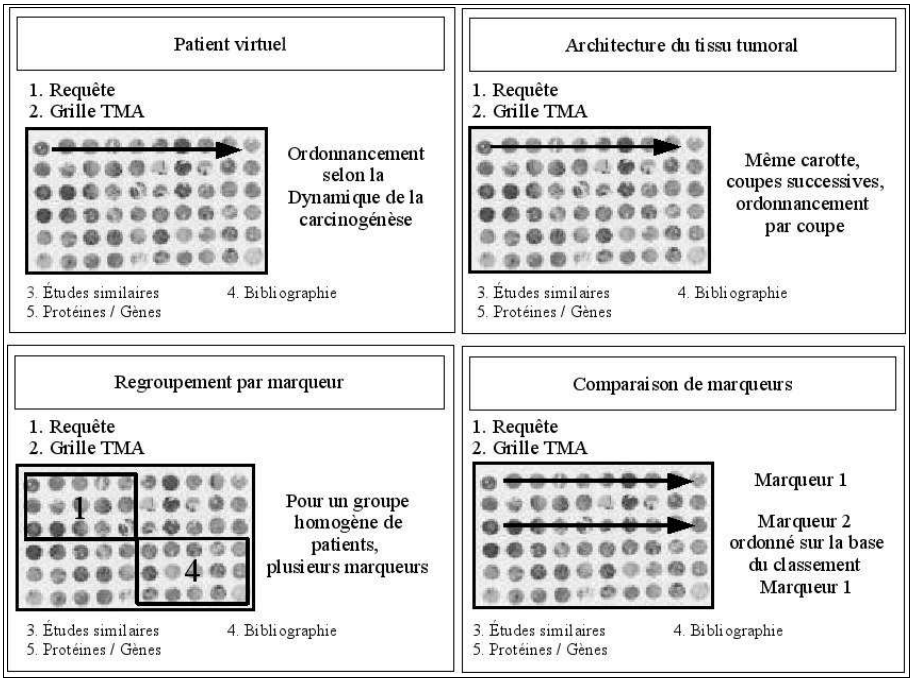

Figure 3. Exemples de modèles d'adaptation pour les documents TMA

d'adaptation. Les niveaux de composition sont génériques, contrairement aux niveaux de spécialisation qui sont spécifiques du domaine :

— Niveau Domaine : aux catégories de critères présentées précédemment, peuvent être ajoutés par exemple des critères techniques : quand il s'agit d'assister la fabrication de blocs TMA réels, le travail du technicien de laboratoire peut être facilité en limitant le nombre de manipulations de blocs par exemple,

— Niveau Objectif : des exemples de modèles d'adaptation spécifiques du domaine des TMA sont présentés Figure 3,

— Niveau Requête : les critères sont ceux définis dans l'ontologie de la requête.

7. Conclusion

Dans cet article, nous avons proposé une architecture pour un moteur d'adaptation permettant la génération de documents de synthèse. L'architecture proposée repose sur une hiérarchie construite selon deux axes:

— Spécialisation : l'adaptation à plusieurs niveaux de spécialisation consiste en un raffinement progressif de la procédure d'adaptation vers un cas particulier. Du point de vue de l'ingénieur, cela facilite la procédure d'acquisition et de représentation des connaissances. Du point de vue de l'utilisateur, cela permet une formulation et reformulation flexibles de la requête,

— Composition : l'adaptation à plusieurs niveaux de composition consiste en une construction progressive du document de synthèse final en étapes successives. Du point de vue de l'ingénieur, cela permet une décomposition de la tâche, ce qui facilite sa manipulation. Du point de vue de l'utilisateur, cela apporte une possibilité de visualiser des résultats intermédiaires et facilite la formalisation de l'expertise.

Le recours à des modèles tels que les Modèles de Requêtes, Modèles d'Adaptation, ou Modèles Logiques de Documents permet une grande souplesse de l'ensemble du processus d'adaptation.

Nous nous intéressons à une classe particulière de documents, que nous appelons de synthèse, encore peu considérés dans la littérature. Nous avons caractérisé le type de requêtes correspondant, les requêtes analytiques, où la représentativité statistique sera un critère important.

La suite des travaux impliquera une caractérisation plus approfondie des critères d'adaptation et la définition d'un système d'évaluation qualité du document de synthèse proposé : adéquation à la requête ou qualité d'échantillonnage entre autres. Il s'agira ensuite d'aller plus loin dans la conception en s'attachant à conserver la généricité de l'outil. La mise en oeuvre d'un prototype devrait permettre la validation du modèle proposé.

8. Bibliographie

Baldonado M., Chang C.-C.K., Gravano L., Paepcke A. « The Stanford Digital Library Metadata Architecture », *Int. J. Digit. Libr*, vol 1, 1997, p. 108-121

Berman JJ., Edgerton ME., Friedman BA., « The tissue microarray data exchange specification: a community-based, open source tool for sharing tissue microarray data », *BMC Med Inform Decis Mak*, vol 23, n° 3, 2003, p. 5

Bes F., Roisin C., « A presentation language for controlling the formatting process in multimedia presentations », *Proceedings of the 2002 ACM symposium on Document engineering,* McLean, Virginia, USA. 8-9 Novembre 2002, New York, USA, ACM Press, pp 2-9

Bodenreider O., « The Unified Medical Language System (UMLS): integrating biomedical terminology », *Nucleic Acids Res. , Database issue,* vol 1, n° 32, 2004, p. 267-270

Brusilovsky P., « From Adaptive Hypermedia to the Adaptive Web », *Communications of the ACM*, vol 45, n°2, 2002, p. 31-33

Ding Y., Fensel D., Klein M., Omelayenko B., « The Semantic Web: Yet Another Hip? », *Data & Knowledge Engineering archive*, vol 41, n°2-3, 2002, p. 205-227

Gruber T., « Toward principles for the design of ontologies used for knowledge sharing », *International Journal of Human-Computer Studies, Special issue: the role of formal ontology in the information technology*, vol 43, n° 5-6, 1995, p: 907-928

Guarino N., « Formal Ontology and Information Systems », *Proceedings of the 1st International Conference on Formal Ontologies in Information Systems, FOIS'98*, Trento, Italy, pp 3-15

Henshall S., « Tissue microarrays », *J Mammary Gland Biol Neoplasia*, vol 8, n° 3, 2003, p. 347-358

Hoos A., Cordon-Cardo C., « Tissue microarray profiling of cancer specimens and cell lines: opportunities and limitations », *Lab Invest*, vol 81, n° 10, 2001, p.1331-8

Iksal S., Garlatti S., « Adaptive Special Reports for On-line NewsPapers »,*Workshop Electronic Publishing,Adaptive Hypermedia, AH 2002*, Malaga, Espagne, 28 Mai 2002.

Kallioniemi OP., Wagner U., Kononen J., Sauter G., « Tissue microarray technology for high-throughput molecular profiling of cancer », *Hum Mol Genet,* vol. 10, n° 7, 2001, p. 657-662

McKeown K., Barzilay R., Evan D., Hatzivassiloglou V., Klavans J., Sable C., Schiffman B., Sigelman S., « Tracking and Summarizing News on a Daily Basis with Columbia's Newsblaster », *Proceedings of HLT 2002: Human Language Technology Conference,* San Diego, USA, 24-27 Mars 2002

Shergill IS., Shergill NK., Arya M, Patel HR., « Tissue microarrays: a current medical research tool », *Curr Med Res Opin,* vol 20, n° 5, 2004, p.707-712

Wu H., De Bra P., Aerts A., Houben G-J., « Adaptation Control in Adaptive Hypermedia Systems ». *Lecture Notes in Computer Science*. vol 1892, 2000, p. 250