

Fiche de dépôt de projet

Volet A : Administratif

1. Titre du projet

« Exploitation de la sémantique spatiale et temporelle hébergée par les documents comme levier d'interaction »
Valorisation du patrimoine à travers le prisme géographique

2. Axe¹ du projet

Axe 2	« Traitement de l'information et des interactions : architectures, modèles, méthodes orientés utilisateurs »
-------	--

3. Nom du porteur du projet²

Christian Sallaberry

4. Niveau de la demande

Initiale (2ans)

5. Liste des chercheurs du LIUPPA impliqués dans le projet³

Nom	Prénom	Implication en %	Compétence apportée
Bessagnet	Marie-Noëlle	50	Formalis. Domaine / EIAH
Dagorret	Pantxika	50	Formalis. Domaine / interaction
Etcheverry	Patrick	100	Spécification / conception IHM
Gaio	Mauro	100	Modélisation sémantique / IE, IR
Jobard	Bruno	20	Visualisation / IHM
Lacayrelle Le Parc	Annig	100	Modélisation sémantique / IE, IR
Lopistéguy	Philippe	60	Formalis. Domaine / interaction
Luthon	Franck	40	Modélisation sémantique / IE image
Marquesuzaa	Christophe	100	Spécification / conception IHM
Nodenot	Thierry	100	Formalis. / Spécif. / EIAH
Sallaberry	Christian	100	IE, IR / Archi. Log.

Ces différentes compétences s'intégreront dans les domaines ci-dessous. A chacun de ces domaines nous associons un pourcentage représentant la part du volume de travail total estimé pour le projet :

- | | | |
|---|---|-----|
| 1. l'extraction d'informations et représentation de la sémantique | : | 30% |
| 2. l'indexation et la recherche d'information | : | 30% |
| 3. l'interaction dans des contextes d'usages spécifiques | : | 20% |
| 4. l'éducation | : | 20% |

¹ Cocher l'axe correspondant, les deux dans le cas d'un projet trans-axes.

² Un chercheur ne peut être responsable que d'un seul projet ; le projet doit être en adéquation avec le domaine de compétence du responsable.

³ 38 chercheurs minimum par projet sauf avis du CdL ; 2 projets maximum par chercheur.

Notons également que 2 thésards travailleront directement sur ce projet ainsi que 4 stagiaires ingénieur et Master au minimum.

Nom	Prénom	Implication en %	Compétence apportée
Thèses en cours			
Lesbegueries	Julien	temps plein	Modélisation sémantique / IE, IR / Archi. Log.
Loustau	Pierre	temps plein	Modélisation sémantique / IE / Archi. Log.
Stages 2006			
Corbineau	Sandrine	5 mois	Recherche & Développement (M2 Rech.)
De Zambotti	Anthony	5 mois	Développement (M2 Pro.)
Kergozien	Eric	5 mois	Recherche & Développement (M2 Rech.)
Levignat	Matthieu	4 mois	Développement (IUP2 M1)
Mansour	Jihene	5 mois	Développement (Ingé.)
Bouhaouala	Mehdi	5 mois	Développement (Ingé.)

6. Liste des chercheurs n'appartenant pas au LIUPPA

Nom	Prénom	Implication en %	Compétence apportée
Le Nir	Yannick	20	TAL

Volet C : Scientifique

7. Résumé

Le problème des corpus documentaires hérités

Grâce à la maturité des techniques de numérisation et de reconnaissance optique de caractères, des corpus documentaires conséquents deviennent de plus en plus facilement disponibles en format électronique, cependant leur accessibilité reste problématique. Afin d'exploiter au mieux la valeur de ces documents dits hérités, il est important d'identifier et de formaliser leur potentiel informatif. Ce potentiel, une fois cerné, même partiellement, est alors utilisable comme clef d'accès pour dynamiser l'utilisation du corpus. Dans le domaine de STIC de nombreuses actions de recherche ont déjà proposé des techniques et des méthodes autour de tâches bien identifiées dans des thématiques connues sous les acronymes d'IE (extraction d'information) et d'IR (recherche d'information). Actuellement une constatation, assez répandue, pointe les limites des techniques basées exclusivement sur des analyses de type statistique des éléments de base directement « exploitables » par un programme, par exemple les mots qui constituent un texte. En parallèle, se développent des méthodes traitant du « contenu » des documents généralement appréhendés de manière partielle pour des raisons d'efficacité évidente. Le gain attendu est à la fois en termes de rappel (plusieurs mots peuvent tomber sous le même concept objet de la recherche), de richesse de l'interrogation (dépasser la combinaison booléenne d'indicateurs), mais aussi d'appréhension par l'utilisateur des résultats de sa requête. Il s'agit de présenter à l'utilisateur « l'espace documentaire » selon des concepts qui lui sont a priori familiers, qui « font sens » pour lui ; il s'agit donc de cibler des *usages particuliers* suscités par la spécificité de tel ou tel corpus documentaire.

Ainsi, la reconnaissance d'informations qui font « sens » dans le cadre de situations d'usage identifiées et modélisées constitue le challenge de recherche de ce projet. C'est par exemple l'idée qu'une navigation de type « touristique » ou « éducative » dans un ensemble de documents hétérogènes (récits, contes, cartes postales, lithographies, séquences filmées...) liées à un territoire conduit à des stratégies spécifiques d'appropriation, de réutilisation et de « construction du sens ».

Ce projet propose des actions de recherche permettant d'apporter un ensemble de contributions plurielles, obéissant chacune à une cohérence globale, dans cette problématique.

L'espace et le temps ; des clés d'accès universelles

Notre choix pour ces dimensions est fondé sur leur potentiel de référent. En effet, lorsqu'un document colporte des signes spatiaux (resp. temporels) relativement à un espace (resp. un temps) de référence tels que des objets, des bâtiments ou des régions dénotant des zones « connues » (resp. procès, événements, périodes ou années dénotant des moments « connues ») alors cette référence peut servir de clef d'accès pour son exploitation.

Une hypothèse forte consiste à considérer que dans les corpus *homogènes*, du point de vue référence à un territoire, la régularité des expressions faisant référence à des zones géographiques ou à des moments historiques

doit rendre possible la réalisation de chaînes d'analyses et d'interprétations sémantiques ciblées, raisonnablement générales et portables.

Les situations d'usage ; des clés d'accès complémentaires

De façon concomitante, ce projet de recherche vise à faciliter la sélection de documents en adéquation avec les situations d'utilisation. En effet, il est reconnu que selon les situations, les utilisateurs privilégient des documents au détriment d'autres. La sélection faite par les opérateurs est naturellement induite par le potentiel informatif contenu dans le document mais également par son adéquation à la situation d'exploitation.

Ainsi, ce projet de recherche se focalise sur le potentiel informatif spatial et temporel des documents et sur son exploitation dans des situations d'interaction.

Quatre axes de recherche dérivés d'activités de conception d'applications exploitant des documents hérités

La conception d'applications exploitant des documents hérités suppose la réalisation d'activités incontournables telles que : le cadrage de situations d'utilisation, la stabilisation et la formalisation du domaine informatif recherché dans les documents, la production d'outils d'aide au repérage d'information dans les documents ou la production d'environnements médiatisés adaptés aux situations d'usage.

Les contributions attendues de ce projet de recherche visent à faciliter la réalisation de ces activités⁴, le domaine informatif recherché étant les dimensions spatiales et temporelles. Le choix visant à couvrir un ensemble significatif d'activités de conception permet d'avoir une vue intégrée des résultats de recherche. A ce titre, la démarche scientifique sera expérimentale et l'équipe projet s'entourera de compétences complémentaires : dans le domaine du traitement automatique de la langue naturelle (laboratoire GREYC⁵ et ERSS⁶, équipe SIGNES⁷), et, dans le domaine de la linguistique focalisant le patrimoine culturel local (laboratoire IKER⁸). Notre travail de modélisation aura la particularité de s'appuyer sur des situations réelles. L'expérimentation de ces modèles permettra ensuite de confronter les résultats obtenus à des situations toujours proches de la réalité, elle portera sur des espaces thématiques plus restreints et des tâches plus spécifiques, à savoir l'exploitation de corpus de documents géo-situés liés à un territoire et à son histoire dans le cadre de la production d'environnements adaptés à des situations de découverte et d'éducation. Les compétences couvrent des méthodes et des techniques en Modélisation et Traitements des Contenus (Représentation sémantique de l'information, Extraction et Recherche d'Information) ainsi qu'en Interaction et en Education.

8. Contenu scientifique

Nous découpons la description du contenu scientifique de notre projet en quatre axes étroitement liés. Dans ce contexte documentaire, nous proposons deux premiers axes dont le but est de proposer des modèles décrivant les situations d'utilisation, d'une part, et le domaine informatif recherché, d'autre part. Deux autres axes ont ensuite pour objectif de mettre en œuvre ces modèles dans le cadre de travaux d'extraction et de recherche d'information, d'une part, et de production d'environnements médiatisés adaptés à des situations d'usage spécifiques, d'autre part.

Les principaux verrous scientifiques de ce projet concernent la modélisation, l'extraction puis, l'exploitation de la connaissance orientés par des usages spécifiques.

Le cadre applicatif de ce projet, quant à lui, est fourni par la MIDR⁹ et probablement le Musée de Bayonne qui procèdent à la numérisation et à l'OCRisation de parties importantes de leurs fonds. Les documents concernés ont trait aux Pyrénées, au Pays Basque, au Béarn, etc. De même, un autre élément caractéristique de ces fonds est leur connotation temporelle : la MIDR, par exemple, conserve des collections de quotidiens et hebdomadaires sur plusieurs dizaines d'années, des cartes postales, des récits de voyages, des contes, etc. De ce fait, le dénominateur commun à tous les documents de tels corpus est la notion de localisation spatiale et temporelle. Enfin, une dimension importante que nous devons traiter concerne la notion de nouveaux publics et d'usages qui pourraient être fait de ces documents jusque là oubliés.

⁴ Ces activités couvrent un spectre significatif du processus de conception d'applications valorisant des documents hérités. Les activités telles que la sélection et la numérisation de documents, la mise en exploitation effective, la commercialisation d'outils n'entrent pas dans nos préoccupations de recherche.

5GREYC - UMR 6072 (Groupe de Recherche en Informatique, Image, Automatique) et Instrumentation de Caen)

<http://www.greyc.ensicaen.fr/>

6ERSS - UMR 5610 (Equipe de Recherche en Syntaxe et Sémantique) <http://www.univ-tlse2.fr/erss/>

7SIGNES - INRIA (Signes linguistiques, grammaire et sens : algorithmique logique de la langue)

http://www.inria.fr/recherche/equipes_ur/signes.fr.html

8IKER - UMR 5478 (Centre de Recherche sur la Langue et les Textes Basques) <http://www.iker.cnrs.fr/>

9Médiathèque Intercommunale à Dimension Régionale de Pau

AXE 1 : Cadrage de situations d'utilisation

Actions de recherche : Spécification de situations d'interaction exploitant un corpus documentaire enrichi d'un marquage sémantique spatial et temporel.

D'un point de vue utilisateur final, tout processus d'exploitation de corpus documentaire comporte typiquement des étapes de recherche et de sélection de documents ainsi que des étapes d'utilisation proprement dites des documents sélectionnés, sachant que la qualité des résultats des étapes de recherche et de sélection dépend du degré d'adéquation des documents sélectionnés pour la complétion des situations d'usage. En effet, la sélection faite par les opérateurs est naturellement induite par le potentiel informatif contenu dans les documents mais également par leur adéquation à la situation d'exploitation. Réciproquement, les situations d'exploitation initialement prévues sont également ajustées aux documents réunis.

Les étapes de recherche et de sélection de documents constituent des situations d'interaction avec le corpus, au même titre que les étapes d'utilisation. Toutefois, la modélisation des situations d'utilisation (en fin de chaîne) avec une mise en exergue des propriétés de documents adéquats, facilite la mise en place d'étapes amont efficaces. Pour construire le plan de travail envisagé, nous utiliserons cette distinction entre, d'une part, les situations de recherche et de sélection et, d'autre part, les situations d'utilisation :

- *Action 1.1* : Spécification de situations de recherche et de sélection de documents, prenant en compte les critères spatiaux et temporels. Par exemple, pour un corpus donné : recherche et sélection de documents relatifs à une période et/ou un territoire particuliers.
- *Action 1.2* : Spécification de situations d'utilisation pédagogiques et/ou touristiques de documents. Par exemple situations de type « course d'orientation », respectivement « itinéraire virtuel pyrénéen », ou encore, « rallye découverte » (recherche d'information pour la construction ou la résolution d'énigmes).
- *Action 1.3* : Mise en exergue de propriétés de documents utiles aux situations d'utilisation spécifiées en Action 1.1. Par exemple, propriétés factuelles et/ou propriétés descriptives de l'espace dans lequel la « course d'orientation » est envisagée.

Stabilisation des abstractions utiles pour concevoir la situation d'utilisation : les abstractions décrivant les activités proposées aux utilisateurs doivent être compatibles avec les abstractions spatiales et temporelles « à extraire », « à rechercher » dans les documents choisis. Cette stabilisation passe par un processus de négociation qu'il s'agit de formaliser et de faciliter (outils de prototypage rapide de type *LinguaStream*¹⁰ incluant le concepteur d'activités dans la boucle d'analyse du marquage des documents).

- *Action 1.4* : Spécification de situations de recherche et de sélection prenant en compte les propriétés mises en exergue en Action 1.3. Par exemple, dans le cas « course d'orientation », le concepteur doit pouvoir rechercher, sélectionner et adapter des documents, mais aussi adapter le scénario de la situation aux résultats trouvés.

Ces actions seront menées conjointement et de façon itérative, avec des résultats plus marqués en Année 1 pour les Action 1.1 et Action 1.2 et des résultats plus marqués en Année 2 pour les Action 1.3 et Action 1.4. L'effort de production sera conduit sur un corpus documentaire hébergeant des informations spatiales et temporelles et, les résultats obtenus auront vocation à être discutés et utilisés dans les axes a2, a3 et a4 du projet.

AXE 2 : Stabilisation et formalisation du domaine informatif recherché dans les documents

Actions de recherche : Ingénierie documentaire visant la formalisation de l'information portée par les contenus des documents.

En premier lieu, il est nécessaire de souligner l'importance du croisement de savoirs dans différentes thématiques de l'informatique et l'apport de la pluridisciplinarité.

L'objectif est d'améliorer par différentes approches la modélisation opératoire de contenus à composante spatiale et temporelle de documents d'un corpus « homogène ». Cette modélisation permettra la mise au point d'approches et d'outils facilitant les étapes de recherche et d'identification rapide d'une information dans un document ou ensemble de documents textuels et iconographiques et, éventuellement de recomposer un nouveau document à partir de plusieurs sources ainsi analysées :

- Concernant *la composante spatiale*, le travail doit permettre de confirmer et d'étendre aux autres modes d'expressions ce qu'un premier travail [11] [12] [13] a permis de mettre en évidence sur un corpus de documents textuels, à savoir : **la relation entre l'expression syntaxique de cette composante et sa structure sémantique**. Il faut noter au préalable que la notion importante de coréférence n'est pas pour le moment prise en compte.

¹⁰<http://www.linguastream.org/whitepaper.html>

En simplifiant, à partir d'éléments « noyaux » des noms de lieu (toponymes) dénotant des zones « connues » et repérables grâce à un géoréférencement, la langue procède par application d'un certain nombre d'opérateurs : opérateurs spatiaux (ou « géométriques ») tels que "le nord/sud... de", "le triangle X Y Z", " de X à Y " etc. ; et opérateurs de sélection d'entités au sein d'une zone donnée, selon divers types de critères : sociologiques, administratifs, physiques...

- Concernant la *composante temporelle* la modélisation pourrait appliquer les mêmes principes, ceci devant être confirmé par un travail équivalent à celui mené pour la composante spatiale. Les éléments « noyaux » sont ici les dates ; des opérateurs définissent des intervalles ("de X à Y", "entre X et Y", "les années X") soumis à leur tour à une nouvelle classe d'opérateurs ("le début de X", "aux alentours de X" ...).

La modélisation opératoire préconisée consiste à repérer cette organisation et à la traduire en structures symboliques (selon une représentation formelle réursive) exploitables par des outils d'indexation. Dans les deux cas, des mécanismes compositionnels forts peuvent être formalisés.

Un des principes fondateurs de la démarche consiste à privilégier et même encourager la complémentarité des modèles. Nous faisons en effet l'hypothèse qu'une modélisation intégrant les manières dont l'humain pourrait mener ses recherches d'information compte tenu de ses objectifs doit adopter successivement plusieurs regards sur le même matériau (corpus). A ces différents regards répondront peut être des formalismes distincts. Nous privilégierons des formalismes déclaratifs (grammaires DCG par exemple) afin de rapprocher autant que possible la description des connaissances sémiotiques de la prescription de patrons ou de contraintes destinées à l'outil informatique.

AXE 3 : Extraction et recherche d'information ciblée dans les documents

Actions de recherche : Modélisation et développement de techniques et d'outils d'extraction et de recherche d'informations spatiales et temporelles dans des documents (TAL, Extraction/Recherche d'Information)

La première hypothèse sous-tendant ce projet considère que « la régularité des expressions spatiales et temporelles dans les corpus *homogènes du point de vue référence à un territoire* doit rendre possible la réalisation de chaînes d'analyses et d'interprétations sémantiques ciblées, raisonnablement générales et portables ». Afin de donner une première assise à cette hypothèse, une chaîne a donc déjà été réalisée pour ce qui concerne des documents textuels et le marquage d'expressions spatiales ; elle permet de construire une représentation symbolique du sens colporté par les syntagmes nominaux identifiés. Une première expérimentation réalisée avec cette chaîne semble effectivement confirmer nos hypothèses de faisabilité. Ce travail a été initié dans le cadre des projets PIV¹¹ et Geosem²¹² en collaboration avec la MIDR (Médiathèque) et le laboratoire Greyc de l'Université de Caen [11] [12] [13].

Il s'agit désormais d'étendre cette chaîne de traitements minimaliste à la détection d'expressions spatiales plus élaborées ainsi qu'à la détection d'expressions temporelles. Il en ressort un certain nombre de questions concernant :

- d'une part, la portabilité de la méthode permettant de constituer des chaînes de traitement pour les documents textuels, celle-ci devant sans doute se préciser, par l'amélioration de la formalisation de la structure des expressions porteuses de sens, mais également par des techniques d'acquisition de lexiques d'expressions complexes voire d'ontologies spécifiques ;
- d'autre part, le « transfert » de la méthode vers des documents iconographiques. Le projet comporte un volet visant l'interprétation de document de type « images » (cartes postales, lithographies, séquences filmées...) il s'appuiera sur l'environnement d'analyse d'image et de séquences vidéo M.A.I.S (*Motion Analyses for Image Sequences*) du LIUPPA. Il s'agit de repérer des formes significatives : concentration, polarisation, contraste, position relative, taux d'occupation... en correspondance avec des entités géographiques répertoriées et des éléments thématiques des textes d'accompagnement ;
- enfin, le problème du « matching sémantique » entre la représentation sémantique d'une expression spatiale extraite et formalisée grâce aux chaînes de traitement et un besoin formulé par l'utilisateur : question complexe et délicate dès lors qu'il est laissé à l'utilisateur une grande liberté dans les moyens de formulation de son besoin.

L'approche que nous avons envisagée reprend globalement les principes des systèmes d'extraction et de recherche d'information traditionnels : un traitement « off-line » permet d'analyser les collections de documents et de produire des index. Puis, pour chaque besoin d'information formulé par un utilisateur, un moteur se charge

¹¹Projet « Pyrénées Itinéraires Virtuels » mené en collaboration par la MIDR et le LIUPPA (contrat Recherche Communauté d'Agglomération de Pau – LIUPPA, mai 2005)

¹²Traitements sémantiques pour l'Information Géographique. Expérimentation, Valorisation de la plateforme de TAL et Prolongements européens - <http://infodoc.unicaen.fr/geosem/>

de chercher dans les index les entrées témoignant des plus grands appariements. Ce sont les mêmes chaînes qui sont utilisés pour traiter les documents et les besoins formulés par l'utilisateur.

Notre moteur présente néanmoins un certain nombre de spécificités quant aux méthodes qui seront utilisées pour l'indexation et la recherche de l'information :

- Indexation sémantique : une première spécificité concerne l'indexation qui ne s'appuiera pas uniquement sur des mots clefs mais sur les représentations symboliques des expressions identifiées. Il en résulte un pouvoir expressif beaucoup plus riche. A titre d'illustration et concernant l'expression textuelle « toutes les communes du sud du Béarn » on pourra obtenir automatiquement une représentation symbolique sous la forme d'une structure de traits : {[determinant: (type: exhaustif)] [type: (administratif: commune)] [zone: [entité_géo: (nom : Béarn, type_zone: pays)] [localisation : interne] [relation: nord]]}. Sur le plan technique, cela entraîne l'écriture de comparateurs spécifiques permettant l'appariement entre besoin formulé par l'utilisateur et cette forme d'indexation.
- Indexation temporelle et spatiale : cette indexation s'appuiera sur le principe qu'à un domaine donné correspond un ensemble de descripteurs que l'on estime pertinent pour la recherche documentaire dans ce domaine. Dans le cas présent, il s'agira donc d'informations spatiales et temporelles s'appliquant à des phénomènes dans le cadre de situations d'usage au préalable identifiées.

Ces méthodes visent essentiellement à étendre les outils de recherche plein texte afin d'affiner des recherches d'information portant notamment sur des critères spatiaux ou temporels. Par exemple, si nous recherchons des documents évoquant le *sud de Pau*, un résultat d'une démarche classique serait composé de documents mentionnant *Pau* et de documents mentionnant *sud* (ou les deux), mais les documents évoquant des communes au sud de Pau, telles que *Lons* ou *Gan*, ne seraient pas proposés. Enfin, nous nous intéressons également aux utilisateurs concernés par la mise en ligne de telles ressources et aux usages qu'ils pourraient envisager.

AXE 4 : Production d'environnements médiatisés adaptés aux situations d'usage

Actions de recherche : La problématique consiste à intégrer à la fois la dimension spatiale et temporelle et la dimension usage (touristique/éducatif) dans des environnements dédiés à l'interrogation de collection de documents, d'une part, et à l'exploitation (navigation / visualisation) de collections de documents (résultant d'une recherche), d'autre part. Ainsi, nos travaux visent la conception et la mise en oeuvre de services logiciels tenant compte du contexte d'utilisation et favorisant l'interaction avec un corpus documentaire enrichi d'un marquage sémantique spatial et temporel.

L'omniprésence de la dimension spatiale au sein des corpus considérés nous incite à utiliser les critères spatiaux comme clés d'accès aux documents. Du fait de la complexité et de la richesse des critères et relations permettant de décrire un contexte spatial nos travaux iront au delà des outils « classiques » permettant d'exprimer une requête à partir d'un formulaire de saisie (recherche par auteur, par thème, etc.) [14]. Des approches telles que [15], bien que pauvres d'un point de vue des interactions, permettent un accès au corpus à partir d'un point d'entrée géographique et nous semblent plus appropriées étant donné la spécificité territoriale des corpus que nous considérons. Nous envisageons ainsi de coupler des méthodes d'expressions graphiques (à partir de carte géographique par exemple) avec des méthodes d'expressions plus ouvertes comme le langage naturel, un peu à la manière de [16] où la requête est exprimée à partir de mots clés et où les résultats sont restreints à une zone géographique spécifiée préalablement sur une carte.

Vis-à-vis de la présentation des résultats et de la navigation, nous retenons l'approche qui vise à organiser les résultats de recherche par thème comme dans [14] ou [17]. Du fait de l'ancrage territorial du corpus considéré, les thèmes que nous considérerons seront relatifs à l'espace. Chaque document retourné sera associé aux lieux qu'il évoque invitant ainsi l'utilisateur à naviguer parmi les documents comme s'il naviguait d'un lieu à un autre.

L'interaction avec un corpus documentaire concerne donc l'ensemble des échanges réalisés entre un utilisateur et une application logicielle incorporant un système de gestion électronique de documents (géositués dans notre cas). Parmi ces interactions, nous considérons :

- celles qui permettent à un utilisateur d'exprimer une attente ou un besoin en terme de contenu informatif. L'intérêt porte sur :
 - les modes d'expression du besoin (textuels, graphiques, ...)
 - les modes de représentation de l'interprétation de ces besoins par le système ;
 - les processus de raffinement possibles permettant de rapprocher les attentes réelles de l'utilisateur avec les attentes de l'utilisateur telles qu'elles sont interprétées par le système.
- celles qui permettent d'exploiter le contenu informatif retourné par le système en vue de satisfaire les attentes initiales de l'utilisateur. L'intérêt porte alors sur :
 - la manière de véhiculer ce contenu vers l'utilisateur afin de favoriser son appropriation (comment présenter les résultats) ;
 - les possibilités offertes à l'utilisateur pour naviguer/explore ce contenu informatif ;

- la manière de favoriser le repérage de l'utilisateur lors de sa navigation ;
- les différentes façons de lire/consulter un même contenu informatif ;
- les possibilités offertes à l'utilisateur pour traiter/travailler un contenu donné.

Par conséquent, la conception de chacune de ces interactions et leur intégration au sein d'environnements médiatisés prendra en compte :

- o la situation d'usage du système ;
- o la dimension spatiale et temporelle omniprésente dans le corpus considéré.

Ces deux critères serviront de cadre pour " typer " chacune des interactions citées précédemment et les intégrer au sein de prototypes. Une situation d'usage définit une manière d'exploiter le corpus pour un public et un objectif donnés. Les situations d'usage considérées concerneront le domaine touristique, dans un premier temps, puis, le domaine éducatif, dans un second temps. Pour chaque situation d'usage envisagée, il s'agira d'intégrer au sein d'un prototype informatique, les modèles d'interaction conçus dans l'axe 1.

9. Objectifs

Après la description du contenu scientifique de notre projet en quatre axes, nous allons détailler les résultats attendus en respectant le même découpage.

AXE 1 : Cadrage de situations d'utilisation

Dans le cadre de cet axe, nous allons décrire les situations suivantes :

- o Dans le cas du domaine touristique, la situation d'usage considérée sera la découverte d'un territoire au travers d'un *itinéraire virtuel*. Les situations de recherche et sélection seront guidées par la présence et l'évocation dans les documents, des lieux composant et/ou avoisinant l'itinéraire.
- o Dans le cas du domaine éducatif, les situations d'utilisation viseront principalement la compréhension de textes par l'apprenant via des méthodes actives : actions sur le texte (extraction et organisation par l'apprenant des concepts significatifs), production de texte ou de croquis (représentations de l'individu versus représentations partagées et partageables), actions (prise de décision) se basant sur la compréhension intra-texte et inter-texte ...

Simultanément à ces travaux de modélisation de situations, trois types de contributions sont envisagés pour cet axe :

- o Langage et/ou modèle de spécification de situations d'utilisation : il s'agit d'outils (modèles ou langages) facilitant l'articulation de spécifications de situations d'interaction avec des propriétés documentaires adaptées aux situations. Ces contributions seront construites sur la base de langages de spécification de scénarios d'interaction ou de cas d'utilisation UML [8] et de langages de spécifications de tâches, tels que CTT [9], combinés avec des langages de modélisation de domaine.
- o Approche conceptuelle de situations d'utilisation : il s'agit de recommandations et/ou d'une méthode à destination de concepteurs de situations d'utilisation. Nous prévoyons d'utiliser les patrons de conception [10] pour formaliser des schémas de situations récurrents et nous orienterons nos recommandations de conception vers des approches itératives à base de prototype.
- o Spécification d'outillages informatiques : il s'agit de la spécification d'un ensemble de moyens logiciels facilitant de façon intégrée, la recherche et sélection de documents, et la conception et mise en œuvre de situations d'utilisation. Conformément aux langages de spécification de situations et aux approches itératives, les outillages spécifiés (produits dans l'axe 4) permettront aux utilisateurs manipulant le corpus :
 - A. d'exprimer une attente ou un besoin en terme de contenu informatif et,
 - B. d'exploiter un contenu informatif retourné par le système en vue de satisfaire leurs attentes initiales.

Le projet ici planifié sur deux années n'a pas pour ambition de couvrir l'ensemble de ces contributions. Conformément aux Action 1.1 à Action 1.4 présentées au point 10 (Contenu Scientifique), les résultats prioritairement attendus seront :

1. La formalisation de situations d'utilisation ; par exemple du type *itinéraire virtuel* ou *course d'orientation*,
2. La description de propriétés documentaires utiles à la mise en œuvre efficace de situations ; par exemple du type *course d'orientation*,
3. La spécification, sur la base de prototypage, de moyens informatiques facilitant la recherche et la sélection de documents géo-situés par affinements successifs et itératifs ; par exemple exploitation d'un corpus pour un *itinéraire virtuel*.

AXE 2 : Stabilisation et formalisation du domaine informatif recherché dans les documents

Nous allons produire des modèles formalisant les concepts du domaine et leurs relations de dépendance. Pour les axes 3 et 4 suivants nous aurons besoin de :

1. *modèles* permettant la description d'informations spatiales issues indifféremment de documents textuels ou iconographiques ;
2. *modèles* permettant la description d'informations temporelles issues de documents textuels ;
3. d'ontologie géographique utile notamment en visualisation de collection de documents.

AXE 3 : Extraction et recherche d'information ciblée dans les documents

Nous envisageons dans un premier temps de reprendre l'expérimentation de la chaîne de traitement initiale sur un échantillon de documents conséquent et représentatif.

Les résultats obtenus nous permettront notamment de proposer des règles et des grammaires permettant de détecter des informations spatiales et temporelles dans des documents afin de les marquer conformément aux modèles (cf. axe 2).

Nous pourrions, dès lors, concevoir et mettre oeuvre une chaîne de traitements complète, intégrant ces nouvelles règles et grammaires. L'objectif principal est de construire des index génériques (indépendants de la forme d'expression de l'information indexée) dédiés aux composantes spatiales et temporelles des contenus des documents. Nous travaillerons enfin sur les techniques de recherche d'informations via ces index.

Notons que nous envisageons de développer une version de la chaîne de traitement des documents (partie Extraction d'Information) et de la procédure générique d'interrogation de ces documents (partie Recherche d'Information) selon une architecture basée sur des services web. Ceci, afin de rendre nos modules interchangeables et accessibles à distance dans le cadre de nos différentes collaborations.

Axe 4 : Production d'environnements médiatisés adaptés aux situations d'usage de type touristique et/ou éducatif

La contribution attendue au sein de cet axe correspond à la production d'environnements médiatisés adaptés aux situations d'usage de type touristique et éducation. Ces productions logicielles devront intégrer et expérimenter les modèles élaborés dans l'axe 1 mais aussi alimenter les réflexions sur les modèles en cours de développement.

Notre premier objectif consistera à développer un prototype permettant à un touriste de découvrir un territoire par le biais d'un corpus documentaire territorial. Le territoire considéré sera le Béarn et le fonds documentaire utilisé sera mis à disposition par la MIDR de Pau. A partir d'une requête spatiale exprimée par l'utilisateur (" je souhaite visiter le sud de Pau "), ce prototype aura pour objectif de proposer un parcours découverte de la zone géographique concernée. La découverte sera réalisée en laissant l'utilisateur naviguer au sein de sa zone géographique d'intérêt (représentée par une carte) et en lui proposant au fur et à mesure des documents évoquant les lieux qu'il traverse. Nous expérimenterons la possibilité de représenter le territoire exploré par une carte en trois dimensions dans laquelle l'utilisateur pourra se déplacer et consulter les documents rattachés aux lieux visualisés.

Parallèlement au développement de ce prototype nous étudierons l'opportunité d'exploiter ce même corpus documentaire à des fins d'apprentissage assisté par ordinateur. Nous nous intéresserons aux outils logiciels permettant à un pédagogue d'exploiter le corpus pour concevoir un scénario pédagogique mais également aux outils permettant d'exploiter ce corpus en situation d'apprentissage. Dans le premier cas, il s'agira de considérer le corpus documentaire comme un vivier potentiel de documents pédagogiques. Nous proposerons alors des outils logiciels pour assister le pédagogue dans la recherche de documents adaptés à une situation d'apprentissage donnée mais aussi sur la manière d'intégrer ces documents dans la situation. Dans le second cas, il s'agira d'exploiter le corpus documentaire dans une situation d'apprentissage intégrant un dialogue système-apprenant sur un exercice de compréhension de texte.

10. Planning sur les deux ans

Ventilation des actions sur les 2 années à venir :

Axe 1	Année 1	Année 2
Action 1.1 : Spécification de situations de recherche et de sélection de documents, prenant en compte les critères spatiaux et/ou temporels.	X	
Action 1.2 : Spécification de situations d'utilisation pédagogiques et/ou touristiques de documents.	X	
Action 1.3 : Mise en exergue de propriétés de documents utiles aux situations d'utilisation spécifiées		X
Action 1.4 : Spécification de situations de recherche et de sélection prenant en compte les propriétés mises en exergue en Action 1.3.		X
Axe 2	Année 1	Année 2
Action 2.1 : <i>modèles</i> permettant la description d'informations spatiales issues de documents textuels	X	
Action 2.2 : <i>modèles</i> permettant la description d'informations spatiales issues de documents iconographiques		X
Action 2.3 : <i>modèles</i> permettant la description d'informations temporelles issues de documents textuels	X	
Action 2.4 : <i>modèles</i> d'ontologie géographique utile notamment en visualisation de collection de documents ?		X
Axe 3	Année 1	Année 2
Action 3.1 : expérimentation de la chaîne de traitement sémantique initiale	X	
Action 3.2 : règles et grammaires permettant de détecter des informations spatiales dans doc. textuels	X	
Action 3.3 : règles et grammaires permettant de détecter des informations spatiales dans doc. iconographiques		X
Action 3.4 : <i>règles et des grammaires</i> permettant de détecter des informations temporelles dans des documents textuels		X
Action 3.5 : transformation de la chaîne de traitements sémantique en services web	X	
Axe 4	Année 1	Année 2
Action 4.1 : expression de la requête spatiale	X	
Action 4.2 : interprétation de la requête et représentation spatiale (sur une carte géographique) de cette interprétation pour l'utilisateur	X	
Action 4.3 : raffinement de la requête utilisateur	X	
Action 4.4 : présentation des documents résultants de la requête	X	
Action 4.5 : navigation / repérage parmi les documents résultants		X
Action 4.6 : lecture / consultation d'un document		X
Action 4.7 : visualisation de terrain, positionnement d'objets et navigation 3D en vue de géolocalisation de collections de documents sur une représentation d'une région		X
Action 4.8 : Exploitation de documents géositués pour la conception de scénarios pédagogiques		X

11. Valorisation et rayonnement scientifique

La revitalisation de fonds patrimoniaux culturels est au centre des préoccupations de ce projet. Aujourd'hui, de nombreuses archives, bibliothèques, médiathèques, ... optent pour la numérisation et l'OCRisation de documents. En effet, cette conversion modifie grandement les modes d'utilisation traditionnels. Dès lors, ces documents numérisés deviennent utilisables avec toutes les commodités que procure le format électronique : consultation à distance, téléchargement, manipulation, annotations, recherche d'information, etc. Ces fonds ont plusieurs particularités : il sont fortement territorialisés et les usages envisagés sont très variés. De ce fait, le dénominateur commun à tous les documents de tels corpus est la notion de localisation spatiale et temporelle, d'une part, et la notion d'usage, d'autre part. Ainsi, les développements que nous envisageons dans les axes 3 (Extraction et Recherche d'Information ciblées dans les documents) et 4 (Production d'environnements médiatisés adaptés aux situations d'usages) de notre projet sont autant de solutions tenant compte de ces

particularités. Ils ont notamment pour but d'expérimenter puis, d'améliorer et, enfin, de valider les modèles proposés dans les axes 1 (Cadrage de situations d'usage) et 2 (Stabilisation et formalisation du domaine informatif recherché dans les documents). Durant les deux années du projet nous allons progressivement intégrer ces développements dans un prototype. Ce dernier exploitera un échantillon du fonds documentaire de la MIDR. L'objectif de ce travail, dont la livraison est prévue en début 2008, est triple :

- expérimenter nos propositions sur une situation proche de la réalité ; les résultats obtenus nous permettront de donner de nouvelles orientations et perspectives à nos travaux ;
- asseoir notre partenariat avec la MIDR autour d'une production qui sera disponible dans leurs locaux ;
- communiquer à partir d'un prototype et initier une série de démonstrations ayant pour but d'intéresser de nouveaux partenaires.

A l'évidence, la pluridisciplinarité induite des travaux engagés ne nous permettra pas de nous focaliser sur un seul domaine scientifique. Ainsi, nous allons publier et organiser des manifestations scientifiques dans les domaines du document électronique et plus généralement, des systèmes d'information, de l'interaction et de l'éducation. Nous serons notamment co-organisateurs de deux colloques en 2006 :

- ISDD 06 - Colloque International : Discours et Document, Caen (France) 15-17 Juin 2006, <http://discours2006.info.unicaen.fr/>
- colloque « SI et bibliothèques - l'organisation des savoirs », novembre 2006, Pau, donnera suite au colloque sur « L'avenir des systèmes d'information des bibliothèques », MIDR, novembre 2005, dans le cadre duquel la présentation de nos travaux et de leur cadre applicatif a reçu un très bon écho national, http://www.adbgv.asso.fr/index.php?page=2005_01_pau

Durant l'année 2006, nous développerons les échanges et travaux avec :

- le laboratoire IKER¹³ pour ce qui est de la valorisation des fonds documentaires du pays basque ;
- les chercheurs impliqués dans le projet SIGNES¹⁴.

Ces échanges entamés depuis plusieurs mois devraient nous amener à formaliser des accords de coopération scientifique assurant une meilleure visibilité du laboratoire LIUPPA.

D'autre part, nous sommes partenaires du laboratoire GREYC¹⁵ dans le cadre du projet GéoSem 2¹⁶. Nous envisageons de poursuivre cette collaboration à travers de nouveaux projets communs.

12. Divers : Opportunité scientifique de ce projet

Le projet proposé cadre avec les préoccupations actuelles de la communauté scientifique. Ainsi, le dernier appel à projets de la National Software Foundation (janvier 2006) dans le domaine de la valorisation des fonds documentaires (appel à projets NSDL sur le thème « Technology, Engineering and Mathematics Education Digital Library »¹⁷ identifie des actions prioritaires de recherche à mener pour les trois axes suivants :

- o Axe « Pathway tracks » (page 6 et 7) :
 - Définir et maintenir des critères pour identifier, sélectionner, annoter et générer des méta-données à forte valeur ajoutée sur des corpus documentaires en phase d'exploitation ;
 - Proposer des systèmes de stockage préservant l'utilisabilité des documents archivés ;
 - Anticiper et développer des services à valeur ajoutée permettant d'exploiter les documents archivés par la communauté éducative.
- o Axe « Selection Services » (page 7 et 8) :
 - Sélectionner, baliser et mettre à disposition d'une large communauté des contenus (documentaires) ;
 - Développer des services d'annotation à haute valeur ajoutée et étudier l'utilisabilité de ces annotations pour des publics ciblés ;
 - Concevoir des outils pour permettre aux développeurs de contenus de combiner différentes ressources et de mettre à disposition de publics ciblés ces ressources agrégées.
- o Axe « Other Services » (page 8)
 - Proposer des services permettant d'utiliser des critères spécifiques de recherche de contenus ;
 - Annoter de manière automatique les images, les vidéos et les ressources audio ;

¹³IKER - UMR 5478 (Centre de Recherche sur la Langue et les Textes Basques) <http://www.iker.cnrs.fr/>

¹⁴SIGNES - INRIA (Signes linguistiques, grammaire et sens: algorithmique logique de la langue) http://www.inria.fr/recherche/equipes_ur/signes.fr.html

¹⁵GREYC - UMR 6072 (Groupe de Recherche en Informatique, Image, Automatique) et Instrumentation de Caen) <http://www.greyc.ensicaen.fr/>

¹⁶Traitements sémantiques pour l'Information Géographique. Expérimentation, Valorisation de la plateforme de TAL et Prolongements européens - <http://infodoc.unicaen.fr/geosem/>

¹⁷National Science Foundation : <http://www.nsf.gov/pubs/2006/nsf06533/nsf06533.htm>

- Proposer des systèmes hybrides capables de s'appuyer à la fois sur les capacités humaines d'analyse de corpus et sur les capacités d'analyse automatique de contenus ;
- Proposer des métriques pour l'évaluation des services et outils d'exploitation de contenus.

Dans le cadre de notre projet, ces actions sont déclinées sur un domaine particulier : l'information géographique dans des corpus composés de textes, d'images et de lithographies. De même, nous nous intéressons à des situations d'usage particulières : les situations de recherche d'informations touristiques et culturelles et, les situations d'apprentissage humain (éducation). D'un point de vue technologique, notre projet cadre également avec ce même appel à projets de la NSF qui cite explicitement les technologies suivantes : XML, OAI¹⁸, Dublin Core¹⁹ et Web Services.

13. Références

1. Draheim D., Weber G., (2005) « Modelling form-based interfaces with bipartite state machines », *Interacting with Computers* 17 p.207–228
- 1b. Winckler M.A., Palanque P., Freitas C., (2004) « Task analysis and diagrams for task models: Tasks and scenario-based evaluation of information visualization techniques », in *proc. of the 3rd annual conference on task models diagrams Tamodia'04*, ACM Press.
2. Tchounikine P., (2002), « Pour une ingénierie des Environnements Informatiques pour l'Apprentissage Humain » in *Revue I3*, Vol. 2, n°1, p. 59-95
3. Schneider D. (2004), « Conception and implementation of rich pedagogical scenarios through collaborative portal sites, in Mario Tokoro and Luc Steels (eds.) *The Future of Learning II, Sharing representations and Flow in Collaborative Learning Environment* », IOS Press.
4. Paquette G., (2004) « Instructional Engineering in Networked Environments », ISBN: 0-7879-6466-2, 304 pages, Pfeiffer
5. Enjalbert P. (éd.) (2005), « Sémantique et traitement automatique du langage naturel », 410 pages, Hermès.
6. Dukham M., Goodchild M.F., Worboys M.F., (2003), « Foundations of Geographic Information Science », 224 pages, Travel.
7. Cunningham, H., (2005), « Information Extraction, Automatic », in *Encyclopedia of Language and Linguistics*, 2nd Edition, Elsevier
8. G. Booch, J. Rumbaugh, I. Jacobson, "The Unified Modeling Language Reference Manua"¹, ISBN 0-321-24562-8, [Addison-Wesley](http://www.addison-wesley.com), 2005
9. F.Paternò, "ConcurTaskTrees: An Engineered Notation for Task Models", Chapter 24, in Diaper, D., Stanton, N. (Eds.), *The Handbook of Task Analysis for Human-Computer Interaction*, pp.483-503, Lawrence Erlbaum Associates, Mahwah, 2003
10. Erich Gamma, Richard Helm, Ralph Johnson, John Vlissides, « Design Patterns - Catalogue de modèles de conceptions réutilisables », ISBN 2-7117-8644-7, 1999
11. Jean Casenave, Christophe Marquesuzaà, Pantxika Dagorret, Mauro Gaiò « La revitalisation numérique du patrimoine littéraire territorialisé », EBSI-ENSSIB 2004, Montréal, Canada, <http://www.ebsi.umontreal.ca/ebsi-enssib-colloque.html>
12. Christophe Marquesuzaà, Patrick Etcheverry, Julien Lesbegueries, "Exploiting Geospatial Markers to Explore and Resocialize Localized Documents", *First International Conference on GeoSpatial Semantics GeoS*, Mexico, 2005, <http://www.geosco.org/home.htm>
13. Enjalbert, P., Gaiò, M., « Projet GéoSem : Traitements sémantiques pour l'Information Géographique, textes et cartes », *Rapport de synthèse, Actes du colloque de bilan du programme pluridisciplinaire Société de l'Information*, Ed. CNRS, p. 93-96, 19-21 mai 2005.
14. Visual Catalog : <http://visualcatalog.univ-paris8.fr/vc2/>
15. Maison de l'Orient : <http://www.mom.fr/bibliotheque/bibnum/>
16. Mirago : <http://www.mirago.fr/>
17. Ujiko : http://www.ujiko.com/fr_index.htm

¹⁸<http://www.openarchives.org/OAI/openarchivesprotocol.html>

¹⁹<http://dublincore.org>