AGENCE NATIONALE DE LA RECHERCHE ANR	**CONTENUS ET INTERACTIONS**	Réservé à l'organisme gestionnaire du programme N° de dossier : ANR-08-XXXX-00 Date de révision :
	Document de soumission **B**	**Edition 2008**

Acronyme	**T A N G U Y**
Titre du projet *(en français)*	Des **T**extes aux **A**rguments par **N**avigation **G**raphique de l'**U**tilisateur en **I**nteraction
Titre du projet *(en anglais)*	From **T**ext to **A**rguments through **N**etworks with **G**oals and **U**ser **I**nitiative

Sommaire

1. Programme scientifique et technique

1.1 Problème posé

« *L'utilisateur est relié en permanence et en tous lieux au système d'information planétaire constitué par le web et les réseaux de toutes natures. Il se trouve ainsi confronté à une surabondance d'information qu'il est nécessaire d'organiser, de prioriser, de trier, de résumer, d'interpréter, voire d'oublier. Il est indispensable de concevoir de nouvelles manières d'accéder et de traiter cette information, il s'agit là d'un défi majeur pour la société contemporaine.* » Ces quelques lignes extraites des *Propositions pour la programmation 2008-2010 des activités STIC de l'ANR* posent clairement le cadre des travaux de recherche que nous proposons d'effectuer dans ce projet.

Aujourd'hui, cette surabondance d'informations[1] est traitée principalement en indexant les sources de données dans des moteurs de recherche. Pour améliorer la pertinence des résultats, des outils de traitement automatique de la langue (TAL) et plus particulièrement les outils d'extraction d'informations enrichissent l'indexation. De même, des outils de clustering donnent une vision synthétique des résultats.

Plus récemment, avec les recherches autour du Web Sémantique, des outils de représentation des connaissances sont utilisés pour structurer ces informations dans des ontologies ou des réseaux sémantiques. Quelques travaux se sont intéressés au couplage entre outils d'extraction d'informations d'une part et outils de représentation des connaissances d'autre part[2]. Le projet OSEMINTI (Operational Semantic Intelligence Infrastructure), de l'Agence Européenne de Défense, vise entre autres à coupler réseaux sémantiques et outils de fouille de texte.

On doit aussi mentionner les efforts faits pour développer des interfaces graphiques autour des outils de représentation des connaissances[3] ou pour visualiser les contenus (« text visualization) »[4], par exemple les travaux cherchant à visualiser et fouiller de manière interactive la base WordNet.

Tous ces couplages actuels textes libres / connaissances structurées posent plusieurs problèmes :

1) Il est long et difficile de définir des grammaires d'extraction spécialisées couvrant un large domaine.
2) Il est également complexe de construire l'ontologie modélisant fidèlement un domaine. De plus, les ontologies ont un caractère inévitablement « statique » et peu flexible[5].
3) Les points 1) et 2) supposés résolus, il faut faire un troisième travail pour rapprocher ces deux univers.
4) Enfin, dans les applications réelles, les contours et le contenu d'un domaine varient en permanence, ce qui obligerait en théorie à modifier sans cesse les trois points précédents.

En pratique, cette approche aboutit à des architectures très orientées « ingénierie », avec de nombreux intervenants techniques spécialisés dans les points 1), 2) et 3). Ces architectures sont statiques : toutes les connaissances embarquées sont figées à l'avance par des techniciens éloignés de l'utilisateur final. Elles conduisent à des applications complexes, délicates à spécifier et à mettre au point. Elles doivent amortir leur coût élevé sur une longue période et sur un grand nombre d'utilisateurs supposés partager un même besoin

---

[1] Peter, L. and Varian, H. R. (2003). *How Much Information?* Retrieved March 2006, from http://www.sims.berkeley.edu/how-much-info-2003.
Keim, D. A. (2001). *Visual exploration of large data sets*. Communications of the ACM 44(8): 38-44.

[2] Amardeilh F. (2006). *Web Sémantique et Informatique Linguistique : propositions méthodologiques et réalisation d'une plate-forme logicielle*. Thèse de Doctorat de l'Université de Paris X – Nanterre.

[3] Kalogerakis, E., S. Christodoulakis, et al. (2006). *Coupling Ontologies with Graphics Content for Knowledge Driven Visualization*. IEEE Virtual Reality Conference 2006.
Fluit, C., Sabou, M., et al. (2005). *Ontology-based Information Visualisation*. In : *Visualising the Semantic Web*. V. Geroimenko, Springer Verlag.

[4] Voir par exemple les deux sites Internet suivants : http://www.neoformix.com/2007/ATextExplorer.html, http://www.visualthesaurus.com/. Voir également tous les travaux autour d'outils interactifs pour visualiser et fouiller la base WordnNet.

[5] van Elst, L. and Abecker, A. (2002). *Ontologies for information management: balancing formality, stability, and sharing scope*. Expert Systems with Applications 23(4): 357-366.
Thomas, R. G. (1995). *Toward principles for the design of ontologies used for knowledge sharing*. 43(5-6): 907-928.
Gruber, T. R. (1991). *The Role of Common Ontology in Achieving Sharable Reusable Knowledge Bases*. Second International Conference on Principles of Knowledge Representation and Reasoning, San Mateo, California, Morgan Kaufmann.

arrêté à l'avance. Et l'on constate en pratique que les grammaires et modèles sont rarement modifiés quand bien même le domaine évolue, ce qui dégrade la pertinence des résultats au cours du temps.

Le projet TANGUY se propose d'aborder le problème dans une toute autre direction. Il veut donner à chaque utilisateur le maximum d'informations et d'initiative pour qu'il puisse **traiter sa problématique personnelle du moment en limitant le plus possible les délais et coûts d'ingénierie**.

Plutôt que de « câbler » un pipeline de traitements, l'approche de TANGUY consiste à faire coopérer symétriquement et dynamiquement trois « pôles » :

1) Le pôle d'extraction de « micro-connaissances » présentes dans les textes à l'aide d'outils de traitements de la langue (TAL). Cette extraction aboutit à une certaine structuration des textes en morceaux élémentaires, de manière locale au document et non spécialisée dans un domaine prédéfini.
2) Le pôle de représentation de la connaissance extraite à l'aide de réseaux sémantiques. Il construit une vision globale des relations inter-documents dans le corpus. Ce pôle supporte des opérateurs de navigation, requête et raisonnement pour aider l'utilisateur à converger vers son but. Il permet aussi de créer un maillage de concepts de haut niveau qui matérialisent la compréhension de l'utilisateur.
3) Le pôle de visualisation et d'interaction avec le réseau sémantique. Cette interaction permet d'explorer à différents niveaux d'échelle le continuum entre les informations extraites et les concepts qui les synthétisent. C'est l'instrument qui pilote les actions (incorporer de nouveaux textes, explorer, créer de nouveaux concepts) pour construire et argumenter la solution au problème courant de l'utilisateur.

Un progrès important apporté par l'architecture TANGUY réside autant **dans ce que chaque pôle ne fait pas** que dans ce qu'il fait :

1) L'analyse du texte doit résister à la tentation d'incorporer de la connaissance du domaine dans ses règles linguistiques.
2) La représentation sémantique doit résister à la tentation de décrire la structure fine des phrases du texte.
3) L'interaction graphique ne doit pas imposer des représentations ou opérations purement « géométriques » vides de sens (hors effet de démonstration).

L'enjeu est de faire coopérer harmonieusement l'ensemble des pôles, chacun restant dans son domaine d'excellence.

Le dispositif peut être schématisé dans la Figure 1 ci-dessous :

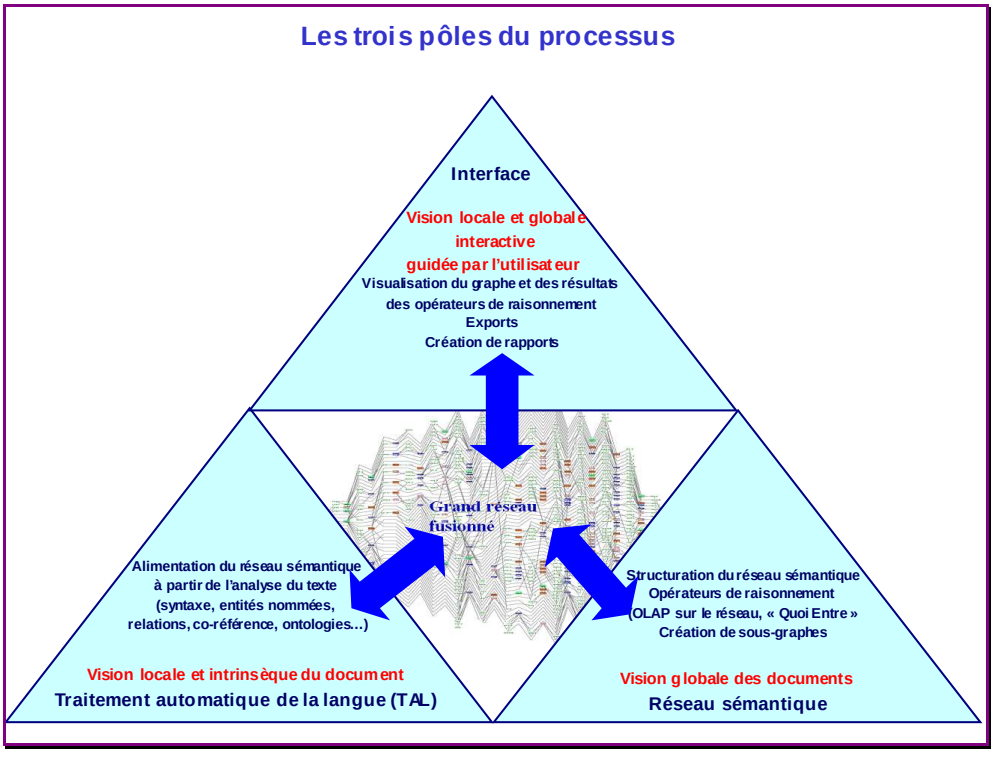


Figure 1 : Les trois pôles du processus TANGUY

A tout moment, l'utilisateur peut prendre l'initiative d'introduire de nouvelles connaissances dans le système, soit « à la main » directement dans le réseau sémantique, soit par import de connaissances préexistantes. Ceci

pourra se faire tout naturellement sous la forme d'introduction de nouveaux textes qui seront analysés « au vol » par le pôle TAL. C'est de cette manière que pourront être introduites des informations structurées, des ontologies sans besoin de recours aux formalismes lourds du Web Sémantique

Dans cette optique, le projet TANGUY est un **projet de recherche industrielle qui vise à aider les utilisateurs à résoudre des problèmes précis à partir d'informations non structurées**, et non pas simplement à rechercher plus facilement de l'information préalablement structurée.

Pour guider et illustrer pleinement les travaux de recherche développés dans ce projet, nous utiliserons un domaine où les textes, la précision et le raisonnement sont fondamentaux : **le domaine juridique**. En effet, que ce soit pour traiter de litiges concernant le droit des affaires, le droit du travail ou les problèmes de cybercriminalité, de terrorisme, les juristes sont amenés à analyser le contenu de données peu ou pas structurées stockées sur ordinateur. L'objectif est d'aider les métiers concernés (analystes, avocats, juges, …) à identifier les informations les plus pertinentes pouvant les aider dans leurs procédures, à construire les réseaux de raisonnement conduisant à l'établissement de preuves.

Cependant, le domaine d'application naturel de TANGUY est plus général : il concerne tous les « cols blancs » ou « travailleurs du savoir » qui ont à raisonner sur beaucoup de documents non structurés. A titre d'exemple , nous projetons d'établir des liens particuliers avec la communauté de l'analyse du contenu de publications scientifiques et de l'aide à leur rédaction.

1.2 Contexte et enjeux du projet

L'actualité la plus récente[6] nous plonge au cœur de l'enjeu général du projet TANGUY : la nature du travail professionnel dans la Société de l'Information, et les manières de rendre ce travail plus satisfaisant à la fois pour les travailleurs et leurs employeurs.
Les NTIC apportent en effet un bouleversement non seulement dans la nature et la quantité de la matière informationnelle transformée par le travail mais aussi dans la nature cognitive de ce travail. Ainsi, dans le numéro 186 d'Octobre 2007 de la revue Sciences Humaines, Emmanuel Sander, responsable de l'équipe Compréhension, Raisonnement et Acquisition des Connaissances de Paris VIII rappelle que « l'être humain est limité dans ses capacités de traitement et guidé par des **principes d'économie cognitive** » et introduit la notion de « coût cognitif » ; pour lui la technologie « devient un outil de pensée ».
Dans le même numéro de la revue, son directeur Jean-François Dortier évoque « de nouvelles formes de pathologies de l'intelligence », des « **labyrinthes documentaires où l'on s'isole, absorbés par un trou noir cognitif** ».
Cette intrusion pathogène du technique dans le cognitif interpelle : l'éditorial de Février 2008 du CXP titre des « Moyens artificiels et naturels de lutter contre l'« infobésité » sous la signature de la journaliste Claire Leroy. Cet article évoque des **pertes de productivité des cols blancs** de 3 à 5 heures par semaine. L'auteur évoque toutes les technologies disponibles (GED, moteurs de recherche, text-mining, ontologies, …), mais considère aussitôt que « trop de logiciels tue la productivité » : le remède serait pire que le mal. Et pour casser ce cercle vicieux remède-mal-remède-mal … il ne faut pas oublier l'essentiel : quelle était la finalité de l'information recherchée ? L'auteur cite une étude de 2005 où les internautes ne passeraient que 1% de leur temps à comprendre ce qu'ils sont venus chercher. L'article conclut que rien ne remplace l'intelligence humaine et l'exercice des méninges, sous-entendu sans aide de la machine.

A partir de ce même diagnostic, le pari des objectifs de recherche de TANGUY est tout autre : précisément rompre cette dichotomie « les outils pour la recherche d'information et les méninges pour le ***vrai travail*** à accomplir ». **Nous mobilisons les technologies de l'information pour aider *aussi* à l'accomplissement du *« vrai travail ».***

Cette approche intéresse plusieurs enjeux sociétaux contemporains :

- la productivité des cadres français et européens, qui augmente moins vite qu'aux USA par exemple (source : service des études et des statistiques industrielles du Ministère de l'Industrie),
- le manque de personnels qualifiés en Europe dans les « nouveaux métiers » qui s'accroît (ibid.),
- la tendance de tous les services publics à offrir plus avec moins d'effectifs. Un exemple qui n'a rien d'anecdotique : les concepteurs des futures Frégates de la Marine Hollandaise espèrent diminuer le nombre de personnels à bord par une meilleure exploitation d'outils de traitement de l'information,
- la judiciarisation de la société qui s'applique aux grandes institutions (Loi Sarbane-Oxley, Norme Bâle 2, Loi Française sur la Sécurité Financière ou loi Mer) : le travail des cols blancs est à la fois plus

[6] Voir en particulier le rapport suivant :
Corman, V. and Ingargiola, E. (2005). *Guide pratique : intelligence économique et PME*. MEDEF. Paris.
Voir également le rapport sur le stress professionnel Nasse-Légeron remis le 13 Mars 2008 au ministre du travail Xavier Bertrand.

surveillé et plus lourd de menaces en cas de non-conformité. Cette tendance s'étend aussi aux droits et obligations des particuliers,

- la réduction du temps de travail, qui ne s'accompagne que rarement d'une réduction de la quantité de travail,
- le besoins d'améliorer l'efficacité de l'Enseignement et de la Formation Continue.

Plus près de notre domaine d'application, il faut noter que la croissance des données non structurées est de 4000% par an, et , parmi celles-ci, la croissance de ***la proportion*** des formats HTML est de 89%, celle de PDF de 10%, et celle des formats classiques de Microsoft de 0%. Alors qu'il y a quelques années on disait couramment que 80% de l'information était non structurée, cette proportion atteindrait aujourd'hui plus de 95%[7]. ***Nous sommes condamnés à rechercher des informations précises dans des formats non structurés***

Sur le plan des retombées économiques à plus long terme, le positionnement des objectifs de recherche de TANGUY pourrait déboucher sur de nouveaux marchés pour l'industrie du logiciel, dans la mesure où il s'adresse à un segment plutôt laissé en friche aujourd'hui par les grands acteurs de l'industrie logicielle, dont la France est assez absente, à l'exception précisément de la CAO, domaine auquel s'apparente l'objectif visé par TANGUY… **L'enjeu considérable est celui de la conception de nouvelles générations d'outils de travail pour les cols blancs**.

1.2.1 Les enjeux autour du TALN

Le traitement automatique de la langue naturelle (TALN), aussi appelé Traitement automatique des langues (TAL) a pour objectif de construire des systèmes capables de comprendre (analyse) ou de produire (génération) des énoncés exprimés en langue naturelle. Dans le cadre de TANGUY c'est la partie analyse du contenu textuel qui nous intéresse dans la mesure où le but est d'extraire des informations structurées, les plus fines possibles, à partir de ressources textuelles non nécessairement structurées comme les e-mails. L'information structurée ainsi extraite servira à peupler une base de connaissance.

Analyse linguistique

L'analyse linguistique constitue l'étape « générique » qui a pour but d'identifier chaque élément d'une phase ainsi que les dépendances entre ces éléments. Elle s'effectue grâce à un moteur d'analyse appliquant des règles propres à une langue particulière. Pour se sortir de la complexité liée au langage naturel, l'analyse linguistique décompose le problème en un ensemble de problèmes plus simples imbriqués les uns dans les autres :

- La segmentation du texte en phrases.
- La segmentation des phrases en mots.
- L'analyse morphologique analyse la structure interne des mots.
- La **syntaxe** organise ces mots dans des groupes de mots et dans des phrases. Elle-même se décompose en différentes étapes :
  - L'identification des groupes syntaxiques (e.g. nominal, verbal, prépositionnel, …)
  - La détection des liens entre ces groupes (Sujet, Verbe, Objet, Direct, Indirect, …)
- La **sémantique** analyse le sens des mots et des phrases.

Chaque niveau de l'analyse linguistique génère un ensemble de résultats, qui peut être exploité par le niveau suivant de l'analyse. Pour la problématique d'extraction d'informations, l'analyse linguistique vise les objectifs suivants :

- La reconnaissance des **entités nommées** : ce sont toutes les formes linguistiques bien identifiées, comme les noms propres (personnes, organisations, lieux) mais également les expressions temporelles (dates, durées, horaires), les quantités (monétaires, unités de mesure, pourcentages), etc.
- Le traitement de la **co-référence** : il s'agit de reconnaître toutes les formes linguistiques qui se réfèrent à une entité nommée. Elle se subdivise en deux sous-tâches :
  - la résolution des **anaphores** : par exemple, dans la phrase « *Les inspecteurs de l'IAEA sont arrivés sur le site de Parchin ; ils sont autorisés à prélever des échantillons.* », le pronom « *Ils* » se réfère aux « inspecteurs de l'IAEA ».
  - L'identification des **variantes de forme des noms propres**, pour trouver toutes les occurrences des mêmes entités orthographiées différemment ou leurs alias, comme pour « *Sarkozy* », « *Le Président de la République* », etc.

[7] Source : Garnier, Alain : *L'Information non structurée dans l'entreprise*. éditions Hermes.

- La reconnaissance des **relations** s'attache à identifier un certain nombre de relations, le plus souvent binaires, entre les entités extraites précédemment. Par exemple, dans l'exemple ci-dessus, la relation de présence dans un lieu peut être identifiée entre les entités personnes « *Inspecteurs de l'IAEA*» et l'entité lieu « *Parchin* ».

Si la plupart des outils arrivent à extraire les entités nommées, il est plus difficile d'en trouver qui traitent efficacement le problème de la co-référence ou qui sachent extraire des relations. Ou bien, les relations extraites sont de très haut niveau comme le fait qu'une personne soit membre d'une organisation, ou qu'une organisation se situe dans un lieu donné. Or, en pratique, il est nécessaire de pouvoir capter toute la sémantique d'un domaine, de manière plus précise que ces relations basiques.

Dans ce projet, il s'agira donc d'aller au-delà de ces résultats (entités nommées et relations en particulier) et d'exploiter au maximum toutes les résultats fournis par l'analyse linguistique.

Analyse syntaxique robuste

L'analyseur syntaxique fournit pour le projet par Xerox (XIP pour Xerox Incremental Parser) accepte en entrée n'importe quel document écrit (ou partie de document) et produit en sortie une représentation grammaticale du contenu textuel de ce document. En d'autres termes, l'analyse syntaxique découpe les phrases selon les différentes unités grammaticales qui les composent.

Par exemple, si la phrase en entrée est la suivante : « *Marie achète une nouvelle imprimante dans un grand magasin* », XIP analyse cette phrase, pour découvrir que :

- Les Groupes nominaux sont : *Marie*, *une nouvelle imprimante, un grand magasin*
- le groupe prépositionnel est : *dans un grand magasin*
- Le verbe principal est *acheter*. Ce verbe est au présent de l'indicatif, à la troisième personne
- du singulier
- *Marie* est le sujet du verbe *acheter*
- *Une nouvelle imprimante* est l'objet du verbe *acheter*

XIP peut analyser des phrases très complexes du point de vue de la structure et traite également divers styles d'écriture (e-mail, « chat », forum…). C'est la raison pour laquelle on dit qu'il est « robuste ». La Figure 2 ci-dessous donne un exemple (tiré du corpus de documents de l'affaire Enron[8]) de l'analyse syntaxique d'un e-mail :

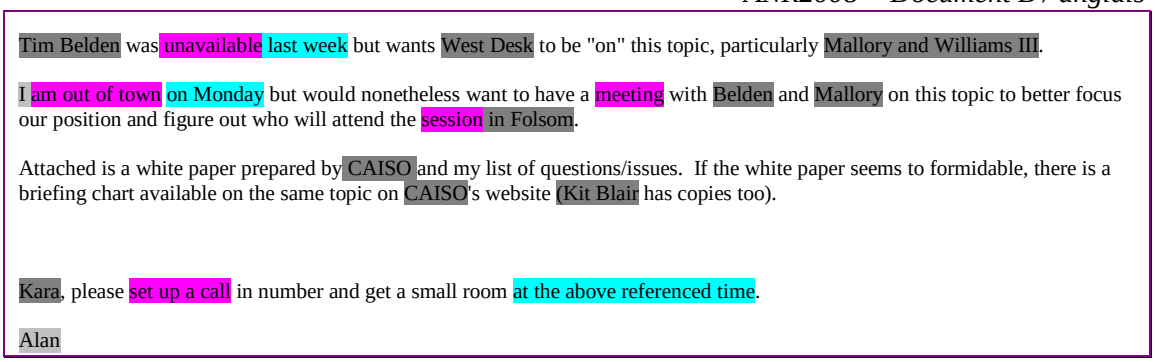
Auteur
Destinataire
Sujet
Time
Event
Named entity

Message-ID: <6561434.1075839926050.JavaMail.evans@thyme>
Date: Sat, 24 Nov 2001 19:41:12 -0800 (PST)
From: alan.comnes@enron.com
To: alan.comnes@enron.com, tim.belden@enron.com, chris.mallory@enron.com,
kit.blair@enron.com, kara.ausenhus@enron.com
Subject: Uninstructed Deviations Proposal of CAISO
Cc: bill.williams@enron.com
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
Bcc: bill.williams@enron.com
X-From: Comnes, Alan </O=ENRON/OU=NA/CN=RECIPIENTS/CN=ACOMNES>
X-To: Comnes, Alan </O=ENRON/OU=NA/CN=RECIPIENTS/CN=Acomnes>, Belden, Tim
</O=ENRON/OU=NA/CN=RECIPIENTS/CN=Tbelden>, Mallory, Chris
</O=ENRON/OU=NA/CN=RECIPIENTS/CN=Notesaddr/cn=90dbe75-fe4c8f62-8625676e-5e14b6>, Blair, Kit
</O=ENRON/OU=NA/CN=RECIPIENTS/CN=Kblair>, Ausenhus, Kara
</O=ENRON/OU=NA/CN=RECIPIENTS/CN=Notesaddr/cn=9832d04d-13ccfce3-88256999-648308>
X-cc: Williams III, Bill </O=ENRON/OU=NA/CN=RECIPIENTS/CN=Bwillia5>
X-bcc:
X-Folder: \ExMerge - Williams III, Bill\Calendar
X-Origin: WILLIAMS-W3
X-FileName:

On Tues and Wed 27-28 Nov 01 CAISO will be having "focus" group meetings on its proposal to begin assessing penalties on uninstructed deviations.

I have already briefed Foster and Williams III on this topic.

[8] http://www.cs.cmu.edu/~enron/.


Tim Belden was unavailable last week but wants West Desk to be "on" this topic, particularly Mallory and Williams III.

I am out of town on Monday but would nonetheless want to have a meeting with Belden and Mallory on this topic to better focus our position and figure out who will attend the session in Folsom.

Attached is a white paper prepared by CAISO and my list of questions/issues. If the white paper seems to formidable, there is a briefing chart available on the same topic on CAISO's website (Kit Blair has copies too).

Kara, please set up a call in number and get a small room at the above referenced time.

Alan

Figure 2 : Exemple d'analyse syntaxique effectuée par XIP sur un e-mail.

L'analyseur syntaxique est le module de base de toute analyse « intelligente » des textes. En d'autres termes, avoir un analyseur syntaxique c'est être en mesure de retrouver des informations de façon précise, de faire du filtrage de questions (par exemple de filtrer, à l'aide des dépendances syntaxiques des documents trouvés à l'aide de mots clés), d'extraire des termes ou des expressions (par exemple, *imprimante couleur*, *le 5Février 2005, Johnny Hallyday*).

L'analyseur syntaxique robuste XIP nous fournira en premier lieu une sortie générique décrivant les différentes relations de dépendance reliant les différents constituants ou entités nécessaires à l'analyse. C'est cette sortie qui servira de base à des couches de traitement ultérieures.

1.2.2 Les enjeux en représentation des connaissances

Position de notre problème.

Traditionnellement, le domaine de la représentation des connaissances fait fortement référence à des notations formelles : ainsi les propositions récentes du Web Sémantique (RDF, OWL, SPARQL, SWRL …) introduisent une syntaxe très codifiée en XML, et les nombreux outils fondamentaux de représentation des connaissances de l'Intelligence Artificielle font usage de langages comme LISP et PROLOG, à la syntaxe elle aussi bien définie.
Ceci n'entre pas réellement dans notre paysage : un utilisateur TANGUY ne voit, ne manipule que des documents textuels et des visualisations de graphes. Les questions pour nous sont donc :

- quelle capacité de représentation veut-on présenter à l'utilisateur à travers nos graphes ?
- quelles capacité d'action (raisonnement, calcul de nouveaux graphes) veut-on lui fournir ?

dans le but de l'aider à traiter ses problèmes, d'une manière acceptable par des professionnels de différents secteurs d'activité et de différents profils intellectuels (par exemple des juristes).

La manière dont ces connaissances et ces opérations sont traitées par la machine, les capacités d'importer / exporter des connaissances avec le monde extérieur – y-compris le monde plus formel – sont d'autres questions, indépendantes de ce que voit l'utilisateur.

Représentation des connaissances et Web Sémantique

Sir Tim Berners-Lee a expliqué récemment[9] que le terme « web sémantique » était depuis le départ une erreur … sémantique : ce qu'il veut faire, et tout le W3C avec lui, c'est un « web de données », et c'est de fait la définition affichée sur le site du W3C. Nous sommes donc dans le monde de la programmation, de la modélisation, du génie logiciel, des bases de données : sa vision est de traduire toutes les bases de données existantes en une base – potentiellement unique – de triplets RDF. Le Web Sémantique est fait pour écrire des applications sur cette base universelle. Une autre question est de savoir comment fabriquer cette base universelle, c'est à dire comment faire s'entendre tout le monde sur le sens des mots (d'où des domaines comme celui de l'alignement des ontologies).

Nos objectifs sont tout autres :

- à la fois plus simples car nous ne nous intéressons pas aux connaissances universelles, mais à celles relatives à la résolution de problèmes très particuliers (une enquête sur une entorse au droit international par exemple)
- à la fois plus complexes, car les raisonnements à mener, les concepts à manipuler, sont plus subtils et imprévisibles que ceux que pourraient raisonnablement décrire et traiter la logique formelle du monde automatique du Web Sémantique, à supposer que l'on dispose de temps et des personnels nécessaires pour décrire exhaustivement un cas en logique formelle.

[9] La Recherche, Septembre 2007

Remarquons que ce côté formel, universel et automatique du Web Sémantique a entraîné en réaction certaines initiatives. D'une part les outils et pratiques de tagging ou folksonomies du Web2.0, qui sont en quelque sorte des anti-ontologies, où la connaissance émerge d'annotations libres d'une communauté. D'autre part le domaine naissant du « Web Pragmatique » qui considère qu'à la notion d'*ontologie* – *ce qui existe* – il faut adjoindre celle de *praxéologie* – *ce qui se passe*.

Représentation « naturelle » des connaissances

L'utilisateur – et non le programmeur – étant aux commandes dans TANGUY, de même qu'il travaille sur du langage naturel et non du langage formel, de même nous devons le pourvoir d'outils de **représentation *naturelle*** et non formelle des connaissances.
Cette problématique est peu abordée pas les communautés académiques, mais, parce qu'elle correspond à un besoin évident, elle fait l'objet d'usages et de réalisation significatifs depuis des années.

Un des objectifs du projet est précisément de porter un regard scientifique sur un domaine qui concerne la vie quotidienne de la majorité des « cols blancs ».
Nous citerons en particulier les outils de « MindMapping » (appelées aussi « cartes heuristiques » en français) et les outils de gestion de réseaux sémantiques, utilisés en particulier dans le monde des enquêtes policières et dans celui du Renseignement.

Le MindMapping

Dans son acception la plus simple, le MindMapping est une technique de prise de notes, qui peut se pratiquer avec du papier et un crayon. Elle consiste à dessiner autour d'un sujet central un écheveau d'informations sous forme arborescente qui vont s'étaler dans toute la page. Il existe depuis plusieurs années beaucoup de logiciels libres ou commerciaux de grande qualité visuelle pour instrumenter cette méthode (MindManager, TheBrain, FreeMind, …). Elle est souvent utilisée en gestion de projets, en visualisation de connaissances encyclopédiques. Elle se révèle être un excellent outil de communication et d'aide à l'élaboration de conceptualisations partagées.

De nombreux ouvrages[10] et communautés s'intéressent à cette pratique, maintenant assez répandue (MindManager a vendu plus de 1 million de licences dans le monde).

Les Outils de réseaux sémantiques

Les outils de MindMapping ont pour eux leur grande simplicité, leur « naturel » évident, et, à ce titre nous les considérerons comme une sérieuse source d'inspiration dans TANGUY : imaginons une « carte heuristique » remplie ou complétée automatiquement à partir de textes…

Ils présentent cependant des limitations certaines : la seule structure acceptée est l'arbre, ce qui signifie qu'un objet ou concept ne peut être utilisé à deux endroits différents dans une description, et encore moins partagé entre deux descriptions. Dans un arbre il n'existe qu'un chemin entre deux objets, ce qui ne permet pas de représenter une grande complexité. En d'autres termes, c'est un plan de classement d'items, mais en aucun cas un réseau maillé.

Dans une autre direction, sont apparues au début des années 1990 des initiatives pour simplifier les outils de modélisation complexes, et les mettre entre les mains de n'importe quel utilisateur, sans aucun recours à des techniciens informatiques. De manière plus ou moins implicite, tous ces outils reposent sur la notion de réseau sémantique. Cette notion est elle-même issue directement des outils de représentation des connaissances en Intelligence Artificielle, et plus auparavant , des travaux sur la logique formelle, pour ne pas remonter à Aristote et ses catégories[11].

Par réseau sémantique, ces outils, susceptibles, comme le MindMapping, d'être pratiquées par des cols blancs de n'importe quel métier, entendent des notions très simples :

- des objets
- des relations
- des énoncés de la forme Objet1 / relation / Objet2
- des catégories auxquelles peuvent appartenir les objets

Le passage au réseau maillé apporte une puissance et une souplesse d'expression très grande, qui, pour faire court, permet de représenter à peu près n'importe quel modèle d'information connu. Les outils de gestion de réseaux sémantiques offrent de plus grandes possibilités de visualiser eu manipuler l'information sous forme graphique.

[10] Le Bihan, F., Deladrière, J.-L., Mongin, P., Rebaud, D.. *Organisez vos idées avec le Mind Mapping*. Editions Dunod
[11] Une histoire complète des réseaux sémantiques est donnée par Sowa (1992) : http://www.jfsowa.com/pubs/semnet.htm.

Parmi les outils de cette famille, on peut citer Analyst Notebook et IDELIANCE
Analyst Notebook (www.i2inc.com) est un logiciel très utilisé par les polices des pays anglo-saxons dans le domaine des investigations sur la fraude et les réseaux criminels. IDELIANCE[12] est un logiciel français développé depuis 1993, qui a été utilisé par de grandes entreprises dans le domaine de l'Intelligence Economique, et qui est maintenant la propriété de Thales Communications, qui l'applique au Renseignement Militaire.

Une caractéristique commune à ces types de logiciels est qu'ils proposent aux utilisateurs des formalismes très simples et très souples (pas d'héritage, pas de typage, grande évolutivité du modèle), tout en leur permettant de structurer leurs connaissances. Ainsi, Idéliance fait un grand usage de la notion d'émergence : plutôt que de suivre un cadre prédéfini, le système examine en permanence l'état courant des informations et propose à l'utilisateur de les enrichir par analogie avec les structures existantes.

A l'expérience, les utilisateurs de tels outils justifient une approche comme celle de TANGUY, dans la mesure où ils expriment :

- le souhait de faciliter le passage le plus automatisé possible de documents textuels au réseau sémantique, c'est dire du « langage naturel » à la « représentation naturelle des connaissances »
- le souhait de disposer d'outils de visualisation, d'interaction et de raisonnement beaucoup plus puissants
- mais en parallèle la volonté de garder une facilité de mise en œuvre qui exclue tout recours à de l'ingénierie informatique

Un des objectifs de recherche du projet sera de voir comment tirer le meilleur profit de ces trois filières aux origines très différentes (MindMapping, réseaux sémantiques, sans négliger le contexte du Web Sémantique dans la mesure où l'on pourra s'abstraire de certains aspects formels) pour satisfaire nos besoins en représentation des connaissances naturelles et interactives.

1.2.3 Les enjeux autour de la visualisation de graphes

Les deux étapes précédentes auront conduit, au départ, à la construction d'un réseau sémantique conséquent. L'exploitation de ce réseau sémantique exige de proposer une interface pour y accéder, le consulter et l'enrichir. Il ne suffit cependant pas de voir ce réseau comme une simple base de données relationnelle stockant l'ensemble des données et pour laquelle il faudrait développer une interface de requêtes. TANGUY se propose d'élaborer une démarche se plaçant dans le courant de recherche en visualisation d'information, et mise sur la visualisation interactive du réseau, de sa forme – dans tous les sens que l'on pourra être amené à donner à ce terme, et des informations qu'il stocke. Selon Ware[13], 40% de notre cortex est dévolu à la perception et à l'analyse de signaux visuels. L'humain, mieux que toute procédure implémentable, est capable de repérer et d'interpréter des motifs graphiques et structuraux dans l'image. Ainsi, le champ de recherche en visualisation d'information s'inspire d'idées ancrées dans des traditions d'origines diverses, dont la statistique graphique, la cartographie, le graphisme par ordinateur, l'interaction homme-machine, la psychologie cognitive, la sémiotique, le design graphique et l'art graphique[14].

Le défi posé par TANGUY exige de trouver les approches pertinentes de visualisation de graphes pour venir en appui aux différentes tâches de l'utilisateur, depuis la construction du réseau sémantique en intégrant les informations obtenues partant de l'analyse linguistique, jusqu'à sa navigation. C'est bien là que se trouve la valeur ajoutée de la visualisation : **l'utilisateur devient un élément essentiel de la construction de la représentation du réseau et de la découverte du sens capturé dans cette structure symbolique**. Le monde informatique n'a eu de cesse de proposer des interfaces graphiques de plus en plus évoluées comme support d'interaction à l'utilisateur. Si les mondes virtuels et les interfaces exploitant les capacités du graphisme 3D présentent un attrait en approchant au mieux le monde réel, la visualisation de représentations 2D d'informations abstraites présentent des avantages évidents, sur les plans cognitifs et opérationnels : ces

[12] Sur les expériences avec IDELIANCE , voir Rohmer, J. (2005) :
http://www.semanticdesktop.org/xwiki/bin/view/Wiki/IDELIANCE
et Perchet (2006) : http://www.cdef.terre.defense.gouv.fr/publications/doctrine/doctrine09/fr/retex/art05.pdf

[13] Ware, C. (2000). *Information Visualization: Perception for Design*. Orlando, FL, Morgan Kaufmann Publishers.

[14] Voir les références suivantes :
Herman, I., Marshall, M. S. et al. (2000). *Graph Visualisation and Navigation in Information Visualisation: A Survey*. IEEE Transactions on Visualization and Computer Graphics 6(1): 24-43.
Card, S. K., Mackinlay, J. D. et al. (1999). *Readings in Information Visualization*. San Francisco, Morgan Kaufmann Publishers.
Spence, R. (2001). *Information Visualization*. Harlow, England, ACM Press/Addison-Wesley.
Tufte, E. R. (1983). *The Visual Display of Quantitative Information*. Cheshire, CT, USA, Graphics Press.
Tufte, E. R. (1990). *Envisioning Information*. Cheshire, CT, USA, Graphics Press (8th printing, June 2001).
Tufte, E. R. (1997). *Visual Explanations*. Cheshire, CT, USA, Graphics Press.

représentations n'imposent pas une charge cognitive importante à l'utilisateur habitué aux conventions de la statistique graphique et reste intuitive ; elles ne reposent pas sur des dispositifs de navigation avancés requérant un apprentissage de la part de l'utilisateur et exigeant un investissement lourd tant en terme de développement que de déploiement.

Voilà une réponse que peut apporter TANGUY aux enjeux de société soulignés à la section 1.2.1, en mobilisant la recherche en visualisation d'information et en développant un interface de fouille interactive et de visualisation 2D contribuant à **définir un environnement archétype pour les travailleurs du savoir, et s'intégrant ultimement à son environnement de travail quotidien**.

La taille du réseau, mais aussi la complexité inhérente aux informations qu'il capture, à leur hétérogénéité, à leur caractère changeant voire instable, posent un défi central à relever. Les techniques à développer placeront l'utilisateur au cœur du processus dès le départ, assemblant des tâches de fouille et d'organisation du réseau en début de cycle ; l'affinage et la formulation d'hypothèses, elle-même accompagnées d'ajustement de l'espace d'information suivront ; à terme, la validation de ces hypothèses et leur diffusion arriveront, secondées de représentations synthétiques (graphiques) pertinentes.

C'est ce processus itératif que Thomas et Cook[15] appelle le « sense making loop » (voir Figure 3 ci-dessous). Ces métaphores, comme la variante proposée par van Wijk[16] (Figure 4 ci-dessous) soulignent le caractère essentiel de l'interaction, permettant à l'utilisateur d'affiner son analyse et la vue qu'il se donne sur l'espace d'information. C'est là un aspect fondamental que TANGUY compte apporter, et qui le distingue de solutions comme celles mentionnées à la section 1.2.2.

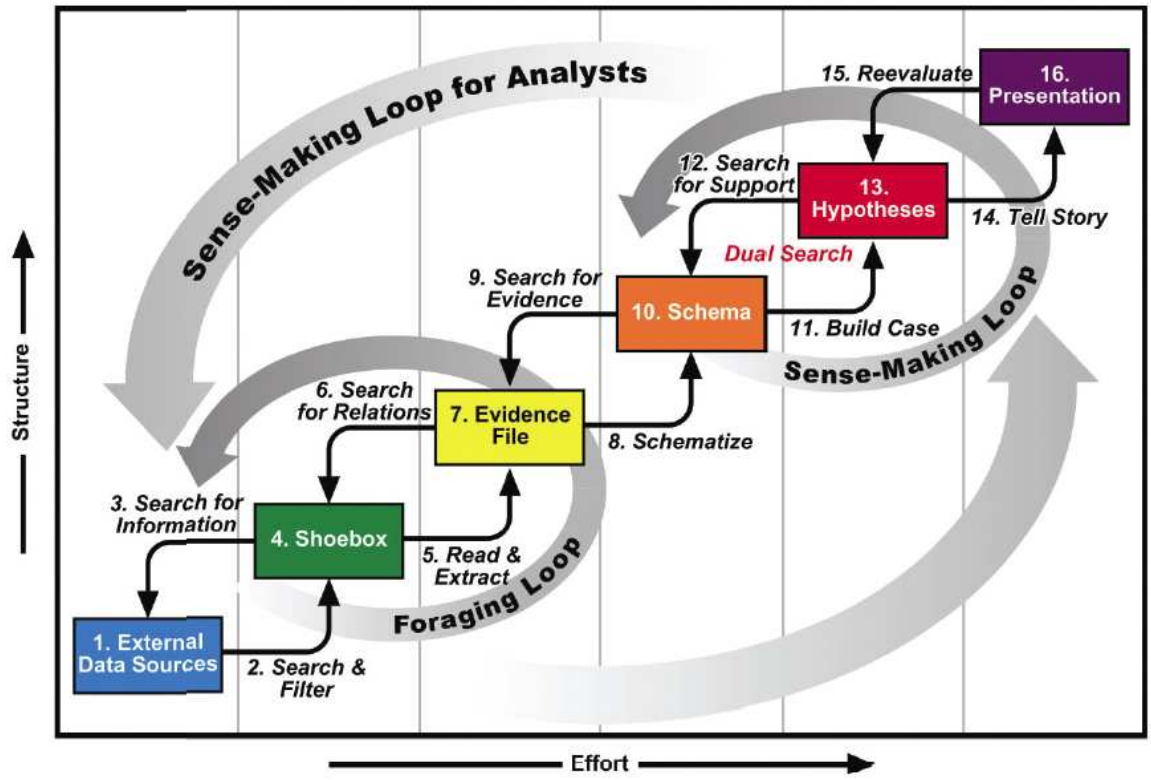


Figure 3. Le "sense making loop" selon (Thomas et Cook 2006).

[15] Thomas, J. J. and Cook, K. A., Eds. (2006). *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. IEEE Computer Society.
[16] van Wijk, J. J. (2005). *The Value of Visualization*. IEEE Visualization, IEEE Computer Society.

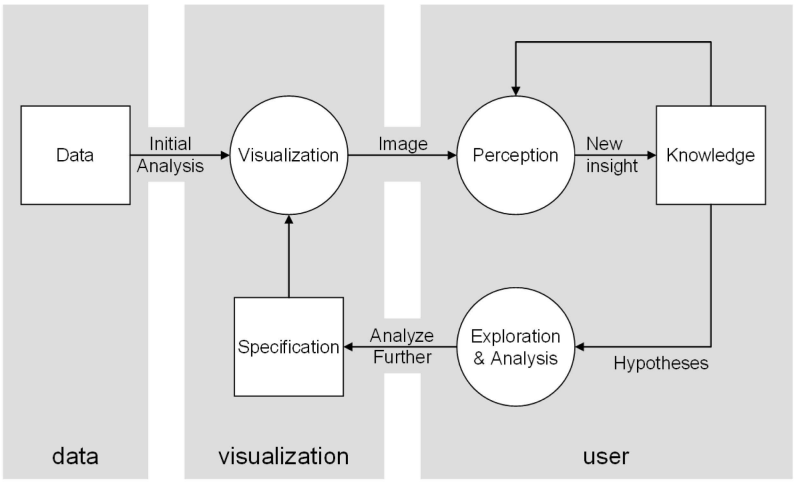


Figure 4. La construction de connaissance vue comme un système dynamique dont le moteur est (en partie) alimenté par la visualisation et la fouille interactive (van Wijk 2005).

Comme le confirme l'expérience des partenaires et les résultats de la littérature, on peut s'attendre à devoir proposer à l'utilisateur des modes de représentations lui permettant de passer à volonté d'une vue globale à une vue locale (et vice-versa) sur les données.

On peut d'ores et déjà identifier des pistes relevant de la fouille interactive et de la visualisation de graphes pour répondre aux enjeux posés par TANGUY. La communauté de recherche en représentations de connaissances pose *a priori* un point de vue qualitatif sur le réseau sémantique, mettant surtout l'accent sur les types de relations qui peuvent lier deux objets et sur la catégorisation des objets. Si cette dimension qualitative peut être « embarquée » dans la visualisation au travers d'artifices visuels, il reste néanmoins indispensable de prendre la mesure du graphe sous-jacent et de sa *topologie* (la structure induite par la distance entre les sommets dans le graphe). **Les propriétés topologiques d'un graphe** donnent en effet accès à certains modes de représentations mettant en valeur certains critères esthétiques : les arbres peuvent être visualisés sous forme de diagramme hiérarchique, les graphes sans circuit sont souvent disposés par niveaux, les graphes planaires admettent des dessins sans croisement d'arêtes, etc.[17]

La notion de distance devient dès lors un ingrédient incontournable pour la construction d'une vue locale dans le réseau, l'enjeu étant de trouver comment harmoniser proximité sémantique et proximité au sens de la topologie et comment identifier le mode de représentations le plus pertinent pour la tâche de l'utilisateur[18].

Nous envisageons évidemment d'étudier d'autres paradigmes de représentations, comme les TreeMaps, qui mettent l'accent sur les attributs des éléments du graphe plutôt que sur sa topologie. Les allers-retours entre vue globale et vue locale reposent aussi sur l'exploitation de la distance sémantique et/ou topologique, mettant en jeu des mécanismes comme ceux introduits au départ par Furnas[19] et repris par van Ham, van Wijk et Abello van Ham[20], ou exploitant des décompositions de graphes[21]. Le calcul de hiérarchies de graphes (sous-graphes imbriqués, Figure 5) offre une piste prometteuse pour proposer des mécanismes de requête et de navigation des réseaux sémantiques.

[17] di Battista, G., Eades, P. et al. (1998). *Graph Drawing: Algorithms for the Visualisation of Graphs*, Prentice Hall.
Kaufmann.
Wagner, M. and D. Eds. (2001). *Drawing Graphs, Methods and Models*. Lecture Notes in Computer Science, Springer.

[18] Purchase, H. C. (1998). *Which Aesthetic has the Greatest Effect on Human Understanding?* Symposium on Graph Drawing GD '97, Berlin, Springer–Verlag.
Ware, C., Purchase, H. et al. (2002). *Cognitive Measurements of Graph Aesthetics*. Information Visualization 1(2): 103-110.

[19] Furnas, G. W. (1986). *Generalized Fisheye Views*. Human Factors in Computing Systems CHI '86, ACM Press.

[20] Abello, J., van Ham, F. et al. (2006). *ASK-GraphView: A Large Scale Graph Visualization System*. IEEE Transactions on Visualization and Computer Graphics 12(5): 669-676.
van Wijk, J. J. and van Ham F. (2004). *Interactive Visualization of Small World Graphs*. IEEE Symposium on Information Visualisation, Austin, TX, USA, IEEE Computer Science press.

[21] Archambault, D., Munzner, T. et al. (2007). *Grouse: Feature-Based, Steerable Graph Hierarchy Exploration*. Eurographics/ IEEE-VGTC Symposium on Visualization, Eurographics Association.

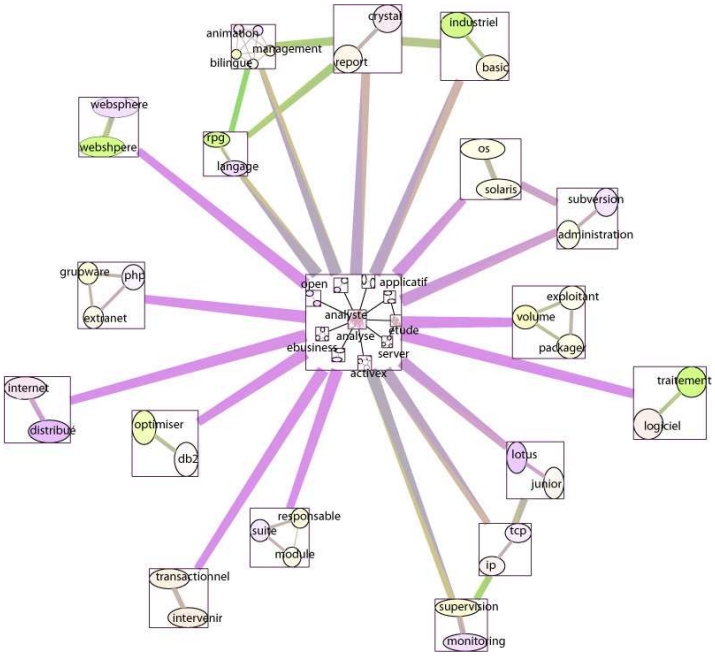


Figure 5. Visualisation de sous-graphes imbriqués (vue égocentrique).

Des mécanismes de vues coordonnées[22] seront aussi vraisemblablement envisagées, un peu à l'image du travail de Kang *et al*.[23] faisant appel à des visualisations de données multi-dimensionnelles comme les coordonnées parallèles[24] en complément des visualisations de graphes. L'enjeu consiste à trouver le bon agencement entre visualisation et tâches de l'analyste au travers du cycle de développement décrit dans le « sense making loop » (Figure 3 et Figure 4 ci-dessus).

1.3 Objectifs et caractère ambitieux / novateur du projet

L'idée innovante du projet est de dépasser les limites connues des trois technologies mises en oeuvre (TAL , représentation de connaissances, visualisation graphique) en les faisant progresser par une mise en synergie étroite et originale.

On peut en effet isoler des verrous pour chacune de ces technologies prises isolément:

- le TAL n'arrive pas à lui seul à analyser et comprendre utilement des textes car il ne mobilise pas assez, au sein de ses mécanismes proches de la langue, des connaissances sémantiques et pragmatiques indispensable à la restitution du sens ;
- la représentation des connaissances propose certes des formalismes artificiels intéressants pour représenter et raisonner, mais ne « décolle pas » en pratique car il est trop coûteux, fastidieux de passer du monde réel, naturel, à un univers codifié et formalisé ;
- les visualisations graphiques progressent dans leur compréhension « géométrique » des réseaux, mais restent difficiles à interpréter, car elles travaillent sur des informations trop pauvres en sens.

On est en présence d'un véritable cercle vicieux, d'un interblocage (« deadlock ») chaque technologie isolée manquant de ressources qui pourraient venir des deux autres.

En rapprochant les trois pôles et en les faisant communiquer, nous cherchons à **enclencher une réaction en chaîne positive** : la représentation des connaissances incorpore avec plus de facilité des informations contenues dans les textes, l'analyse du langage dispose de plus de données structurées pour reconnaître des éléments textuels, enfin l'utilisateur perçoit mieux, en temps réel, ce qui est « compris » par le système ; il est à même de le « doper » en injectant ses propres connaissances humaines pertinentes (car un système ne possèdera jamais ***à l'avance*** les connaissances nécessaires ***plus tard*** dans chaque contexte précis : c'est en refusant de reconnaître cette évidence que beaucoup de systèmes coûteux conduisent à des résultats décevants)

[22] Baldonado, M. Q. W., Woodruff, A. et al. (2000). *Guidelines for using multiple views in information visualization*. Advanced Visual Interfaces AVI 2000, Palermo, Italy, ACM.

[23] Kang, H., Plaisant, C. et al. (2006). *NetLens: Iterative Exploration of Content-Actor Network Data*. IEEE Symposium on Visual Analytics Science and Technology 2006 (VAST2006).

[24] Fua, Y.-H., Ward, M. O. et al. (1999). *Hierarchical Parallel Coordinates for Exploration of Large Datasets*. IEEE Visualization '99, IEEE CS Press.

1.3.1 Travaux sur l'architecture globale

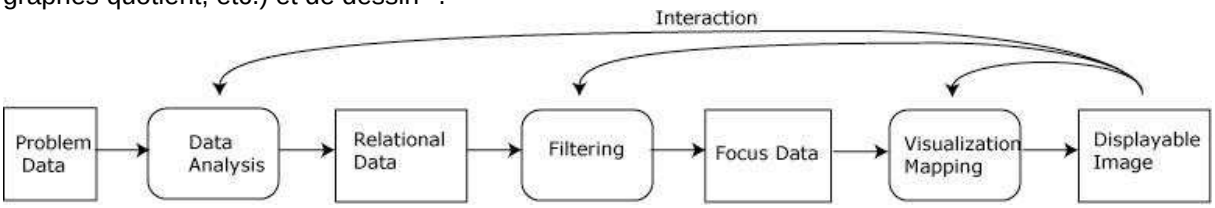
Un travail majeur de ce projet consistera à faire communiquer de façon la plus harmonieuse et la plus optimale possible les résultats des trois pôles :

- Le TALN,
- La représentation des connaissances,
- Les interfaces graphiques et la visualisation de grands graphes.

Cette communication sera définie par un méta-modèle que le projet se propose de construire et qui représente à la fois les « micro-connaissances » extraites par les outils du TALN, la structure du réseau sémantique initial et le support sur lequel les graphes seront construits.

L'architecture tripolaire de TANGUY est donc en elle-même une innovation qui vise à faire sauter des verrous technologiques rencontrés sur chacun des pôles individuellement, qui, dans leur domaine, ne progressent pas autant qu'on pourrait le souhaiter.

Nous envisageons d'utiliser la plate-forme de visualisation Tulip comme pilier technologique du projet. L'architecture logicielle de Tulip suit du « pipeline » de visualisation allant de l'extraction des données dans la phase la plus amont, jusqu'à la production d'un rendu à l'écran, en passant par des phases d'organisation (encodage dans une structure de graphe), de filtrage (seuillage de valeurs, clustering de graphes, calcul de graphes quotient, etc.) et de dessin[25].



L'interaction est vue comme une **boucle inversée** dans le pipeline qui permet à l'utilisateur de relancer les phases amont. Tulip est une plate-forme OpenSource disponible sur SourceForge et développé par l'équipe MABioVis du LaBRI (partenaire de l'équipe INRIA GRAVITÉ)[26]. Tulip connaît un succès grandissant avec près de 1000 téléchargements par mois. La plate-forme est au cœur des recherches menées par l'équipe GRAVITÉ ; plusieurs logiciels dédiés ont déjà été déclinés depuis la plate-forme[27].

Son modèle de données a été conçu de manière à pouvoir manipuler et visualiser de très grands graphes. De plus, un mécanisme d'héritage original à Tulip facilite la construction de graphes imbriqués encodant les représentations multi-niveaux et la manipulation de graphes quotients (des graphes dont les sommets sont eux-mêmes des graphes).

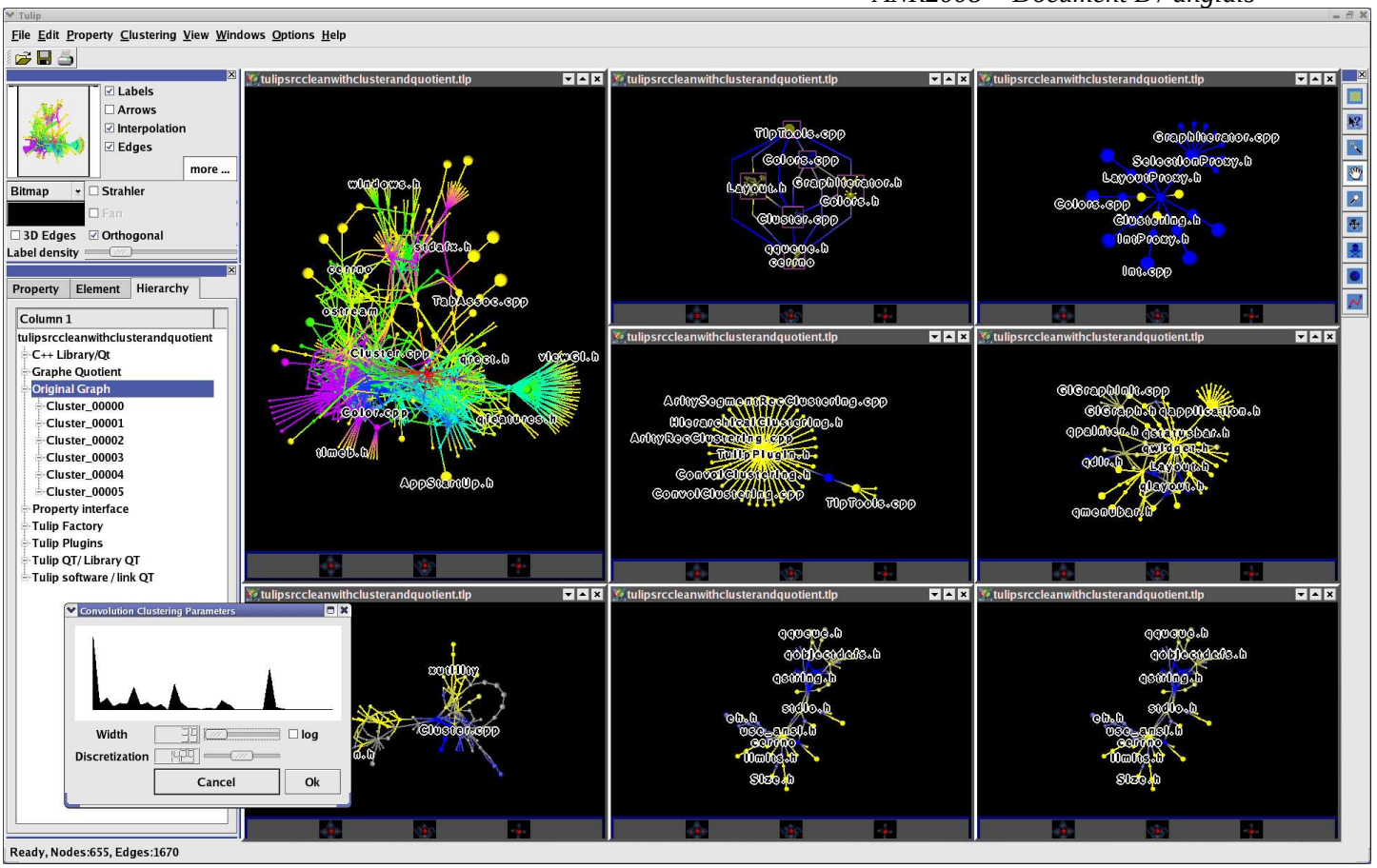
L'architecture de la plate-forme convient parfaitement à une adoption en projet multi-partenaires. L'ajout de nouvelles fonctionnalités répond au mécanisme de plug-in, facilitant l'enrichissement de la plate-forme sans avoir à recompiler l'ensemble. (L'équipe expérimente actuellement la mise en place de services web permettant de recueillir les nouveaux plug-ins et la mise à jour depuis la plate-forme elle même, comme c'est le cas pour l'IDE Eclipse développé par AlphaWorks d'IB, par exemple).

---

[25] Wright, H., K. Brodlie, et al. (1996). The dataflow visualization pipeline as a problem solving environment. Eurographics workshop on Virtual environments and scientific visualization '96, Monte Carlo, Monaco, Springer-Verlag. dos Santos, S. and K. Brodlie (2004). "Gaining understanding of multivariate and multidimensional data through visualization." Computers & Graphics 28(3): 311-325.

[26] Voir le site web http://tulip.labri.fr

[27] Pour plus de détails voir le rapport d'activité de l'équipe INRIA GRAVITÉ
www.inria.fr/recherche/equipes/gravite.fr.html
Auber, D. (2003). Tulip - A huge graph visualization framework. Graph Drawing Software. P. Mutzel and M. Jünger, Springer Verlag.
Delest, M., T. Munzner, et al. (2004). Exploring InfoVis Publication History with Tulip (2nd place - InfoVis Contest). IEEE Symposium on Information Visualization, IEEE Computer Society.
Auber, D., Y. Chiricota, et al. (2007). Visualisation de graphes avec Tulip: exploration interactive de grandes masses de données en appui à la fouille de données et à l'extraction de connaissances. Revue des Nouvelles Technologies de l'Information (actes EGC 2007). M. Noirhomme-Fraiture and G. Venturini, Cépaduès. 1: 147-156.



Ce mécanisme de plug-ins réduit l'effort d'intégration de manière conséquente, en le reléguant à la construction de passerelle d'import ou d'export selon qu'il s'agit des phases amont ou aval du pipeline. Nombre de plug-ins de calcul et de dessin sont déjà disponibles.

Cela dit, le travail au niveau interaction – qui constitue la réelle valeur ajoutée de TANGUY – reste à faire. Le projet requiert aussi la mise au point d'un méta-modèle et de son implémentation dans la plate-forme de manière à tirer profit des mécanismes disponibles et du modèle d'héritage de sous-graphes (héritage de la structure mais aussi des propriétés).

Les bénéfices de l'utilisation de Tulip sont multiples, même sans compter le fait que l'outil émane directement de l'un des partenaires. L'adoption de Tulip par chacun des partenaires apporte au consortium à la fois une base pour échanger sur les avancées du projet et un cadre défini qui contiendra les résultats de nos travaux.

Aussi, Tulip sera utilisé dès le départ pour prendre en main les données issues de l'analyse linguistique, puis pour prendre en charge le méta-modèle, offrant dès les premières phases un retour visuel sur le travail amont. Ainsi, même si Tulip doit *in fine* proposer un outil de travail à l'utilisateur final, son utilisation précoce donnera aux acteurs du projet l'occasion de poser un regard critique sur leur méthodologie de travail, lançant ainsi les premières itérations du « sense-making loop ».

1.3.2 Travaux de recherche du pôle Analyse

Enrichissement des sorties de XIP

Les systèmes de traitement de la langue seront adaptés et enrichis pour analyser et structurer un texte afin de récupérer le maximum de connaissances qui seront intégrées dans le réseau sémantique. Par exemple, excepté peut-être les articles, quasiment tout ce que contient le texte doit être représenté d'une manière ou d'une autre dans le réseau sémantique. Il s'agit donc d'aller bien au-delà de l'extraction d'entités nommées et de relations entre ces entités. Il faudra travailler le plus possible indépendamment du contexte, par définition inconnu à ce stade. Il faudra exploiter au maximum toutes les résultats produits par ces outils de traitement automatique de la langue, en particulier les arbres syntaxiques qui seront enrichis pour être intégrés tels quels dans le réseau sémantique. Les travaux porteront donc sur tout ce qui est intrinsèque à la langue : tout ce qui est lié au dictionnaire, à la représentation et au traitement de la synonymie et de l'antonymie dans le réseau

sémantique, au traitement de la co-référence… **Le résultat de ces travaux sera une des entrées du réseau sémantique**.

La nouveauté de ces travaux consistera donc à aller au delà de l'extraction des éléments du contenu. L'analyse syntaxique et les ressources sémantiques mentionnées ci-dessus permettront l'élaboration de **ressources conceptuelles**. L'analyse de ces ressources permettra de formuler des requêtes véritablement conceptuelles qui permettront à l'utilisateur final du système de faire abstraction de la réalisation linguistique des textes.

Analyse Sémantique
La méthode d'analyse que nous utiliserons est le "concept matching" ; cette méthode a été développée par Xerox. Elle permet de détecter une large variété d'expressions partageant un contenu conceptuel commun (par exemple expression d'un événement, d'un lieu, d'un moment) tout en se dégageant de la réalisation purement lexicale.

L'idée directrice du « concept matching » s'appuie sur le constat qu'un type de contenu particulier est exprimé par les mêmes concepts constituants y compris si leurs réalisations linguistiques de surface sont très différentes. Le « Concept-matching » est une méthodologie à mi-chemin entre les paquets de mots (et leurs synonymes) et les méthodes prenant en compte les dépendances syntaxiques.

Dans des systèmes utilisant une approche de type « concept matching » on examinera la cooccurrence de constituants correspondants à des concepts qui sont liés par une relation de dépendance syntaxique au sein d'une phrase. On écrira ensuite un ensemble de règles de cooccurrence qui régissent les combinaisons possibles de ces concepts constituants, décrivant ainsi les réalisations possibles du concept cible. Cette méthodologie a été implantée sous forme de règles XIP pour rendre compte de différents concepts cible : nouveauté scientifique, danger médical, et risque pays dans le cadre par exemple du projet d'Infom@gic.

Une fois les concepts identifiés, la sémantique des textes devra alors déterminer la présence ou non de relations spécifiques entre ces concepts de manière à identifier des événements, des personnes (acteurs ou autres participants à ces événements), le temps des événements et les lieux. Ces événements peuvent être vus de façon simpliste comme des patrons prédéfinis décrivant une information spécifique. Une même information/fait pouvant être décrit de diverses façons, c'est toute la puissance de l'analyse linguistique qui aura pour but de reconnaître et d'unifier ces différentes formes dans une représentation normalisée et labellisée. Une fois de plus, la définition des faits ainsi que l'identification des différentes façons de les exprimer s'effectuera en étroite collaboration entre experts métiers et experts du traitement automatique du langage. Les règles ainsi créées seront intégrées au moteur d'analyse linguistique[28].

Étude de genre e-mail
Notre travail d'analyse devra prendre en compte que les e-mails constituent un genre particulier de texte écrit.[29] L'e-mail est moins structuré que d'autres genres écrits comme les articles de presse ou des écritures scientifiques. Le style est plus familier et l'orthographe moins soignée. Cela implique des ajustements dans l'analyse automatique qui peuvent être exécutés après une étude détaillée.

1.3.3 Travaux de recherche du pôle Réseau sémantique

Idée de modélisation naturelle
Nous travaillerons à concevoir et expérimenter plusieurs ***modélisations naturelles*** (de même que nous parlons plus haut de ***représentations naturelles***), pour optimiser le nécessaire compromis entre acceptabilité (donc simplicité) et expressivité (donc complexité)

Une modélisation naturelle doit être une manière d'organiser l'information :
- plus précise que le langage naturel
- sans recours au langage formel du logicien ou de l'informaticien
- capable de représenter des problèmes complexes
- susceptible d'opérations automatiques (fusion, éclatement, comparaison, …)
- praticable sans long apprentissage par une grande majorité de « cols blancs »

Au-delà des méta-modèles classiques de représentation sémantiques (Graphes Conceptuels [http://conceptualgraphs.org/], RDF [http://www.w3.org/RDF/], Topic Maps [http://www.topicmaps.org/]), nous étudierons systématiquement les propriétés des modélisations naturelles implicites quotidiennes, sous-jacentes

[28] Sandor, A., Kaplan, A., Rondeau, G. Discourse and citation analysis with concept matching. ISDD 2006.
[29] Rice, R.P. The rhetoric of e-mail: an analysis of style. Professional Communication Conference, 1995. IPCC '95 Proceedings.

aux outils bureautiques (répertoires de fichiers, tableaux Excel, présentations PowerPoint, e-mails), qui, bien qu'universelles, sont négligées par la plupart des travaux de recherche.

Un autre terrain d'investigation sera la structuration de l'information – souvent tout aussi implicite – contenue dans les documents textuels : sémantique des découpages et des intitulés des titres, sous-titres, notions de passages et de paragraphes. Il ne s'agit pas ici de découpage physique, mais de la découverte et du nommage des concepts complexes incarnés dans un passage. (Exemple : « preuve de la collusion d'intérêts entre la société S et Mme M »). En particulier, dans le cas privilégié par le projet des courriels, ces structures, passages, concepts, sont encore plus implicites, et leur découverte est un enjeu important.

Plus ces structures sont complexes, moins une approche automatique peut à elle seule les discerner, et plus le principe d'interaction de TANGUY se justifie. Cela inclura l'étude de stratégies d'annotation, dans l'esprit du « WEB 2.0 » (tagging, folksonomies), mais adaptées au surlignage explicite de concepts de haut niveau. Ces concepts pouvant à leur tour aider, par apprentissage et raisonnement analogique, à en suggérer automatiquement d'autres. Nous nous inspirerons dans ces domaines, des travaux naissants dans les communautés s'intéressant au « Web pragmatique » [http://www.pragmaticweb.info/] et à la théorie de l'Enaction[30].

Idée d'aide au raisonnement naturel

Le but n'est pas seulement de structurer l'information, mais de s'en servir pour résoudre un problème. Il s'agit d'aider à la réflexion, au raisonnement. Si c'est au travers de l'interaction graphique que l'utilisateur a accès à ces outils, c'est au niveau du réseau sémantique qu'ils doivent être représentés. Notre liste des cinq points caractérisant plus haut une modélisation naturelle s'applique aussi à la notion de raisonnement naturel.

Par contre, les points de départ sont moins répandus que ne l'était la modélisation implicite des documents bureautiques. Nous en trouverons quelques-uns dans des outils comme IDELIANCE , à travers des notions comme :

- le « quoi-entre » : pour calculer, visualiser et interpréter les chemins entre deux (-ensembles de-) points
- le concept de « signal faible » ou « chaînon manquant »[31] : calculer des conditions nécessaires pour qu'une nouvelle information contribue à la réponse à un problème explicitement représenté dans la base de connaissances
- des calculs de nature statistique d'agrégation entre informations, de type tableaux de contingence OLAP, similarité entre objets, clustering d'objets
- des mécanismes de raisonnement par analogie, – naturels s'il en est par opposition aux règles de logique formelle – que ce soit à partir de connaissances réelles du corpus, ou de cas d'école, s'inspirant des travaux de « raisonnement par cas »[32].

La recherche consistera à approfondir et harmoniser ces idées de départ en un ensemble de mécanismes non disparates, à les implémenter efficacement et à expérimenter leur efficience sur des cas significatifs.

L'ensemble des dispositifs de raisonnement naturel visent à construire des « objets de raisonnement » comme des arguments, des preuves, des contradictions, des réfutations. Nous nous rapprocherons dans ce domaine de la communauté de recherche sur l'argumentation, en particulier l'ISSA : International Society for the Study of Argumentation.

Une contrainte importante du travail sur la représentation sera de ne pas progresser isolément, mais d'étudier et d'optimiser en permanence le couplage des modélisations et raisonnements naturels avec les deux autres pôles : la structure des résultats issus du TAL, et les capacités de visualisation et d'interaction de l'interface graphique

1.3.4 Travaux de recherche du pôle Visualisation de graphes

Outre la nécessité d'envisager de manipuler et de visualiser des graphes de grandes tailles, le défi consiste à pouvoir visualiser et interagir avec des graphes dynamiques évoluant dans le temps ou par le biais d'interactions de la part de l'utilisateur.

---

[30] Varela, F., Thompson, E. and Rosch, E. (1991). *The Embodied Mind : Cognitive Science and Human Experience*. MIT Press.

[31] Rohmer, J. (2006). *Signaux Faibles et Chaînons Manquants.* CESS, Paris. http://www.afscet.asso.fr/resSystemica/Paris05/rohmer.pdf

[32] Kolodner J. (1992). *An Introduction to Case-Based Reasoning*, Artificial Intelligence Review 6, 3-34.

Stratégies d'exploration et analyse de la topologie des graphes

Il apparaît déjà nécessaire d'étudier les propriétés topologiques des graphes issus du processus d'acquisition des réseaux sémantiques et mis en œuvre par Xerox. Lors de l'élaboration du réseau sémantique global, on peut s'attendre à ce que les thèmes forts émergeants de la collection de documents constituent un graphe invariant d'échelle (scale-free) organisé autour de thèmes ou concepts jouant le rôle de « hubs »[33]. Le modèle « **petit monde** », pertinent pour la donnée linguistique, est à regarder de près tant du point de vue de l'analyse[34] que de la visualisation[35]. Ainsi, il peut être intéressant d'examiner de près comment ces réseaux se forment au fur et à mesure de l'ajout de documents, agrégeant les nouvelles entités ou nouveaux concepts autour de concepts déjà présents. Le modèle « scale-free », par nature, se présente sous la forme d'une hiérarchie de graphes ordonnant les sommets en fonction de leur degré – qui pourrait bien être corrélé à leur portée sémantique. D'autres modèles, comme le modèle ERMG (Erdös-Rényi Mixed Graphs) offrent de alternatives à étudier à des fins d'analyse[36].

L'objectif est d'identifier, voire de développer, des modèles évolutifs pertinents capables de **rendre compte de l'évolution de la topologie du réseau sémantique**, à mesure de sa construction lors de l'analyse des documents mais aussi au gré des interactions utilisateur. Les résultats qui suivront de cette analyse détermineront nombre d'éléments de la visualisation : stratégie de filtrage, algorithmes de dessin, construction de représentations multi-niveaux, etc.

Identification de motifs, requêtes et interaction

Le graphe est donc vu comme dépositaire de la connaissance extraite des documents. C'est sur cet objet que s'effectuera l'ensemble des recherches menées par l'utilisateur. Ainsi, une requête effectuée sur cette base de connaissance doit pouvoir restituer une portion du graphe global. En effet, même si la réponse à la requête consiste en un ensemble d'éléments y répondant, on peut fouiller le graphe, tenir compte de la « distance sémantique » et construire un contexte dans lequel se place ces éléments de réponses au vu de ce que la base de connaissance contient.

Ce défi se situe en réalité au carrefour des trois pôles : non seulement faut-il pouvoir calculer ce voisinage dans le graphe, mais il faut déjà avoir calculé la distance sémantique à la fois en terme de structure linguistique et/ou documentaire, et en tenant compte des enrichissements du réseau par apport de nouvelles données et relations, par exemple. Nous comptons de surcroît concevoir et réaliser un cadre dans lequel l'usage du réseau par l'utilisateur a un impact sur le calcul de la distance sémantique : en gardant une trace de l'usage, **le système se ferait gardien de la mémoire de l'analyste**, et pourrait être à même de compléter la requête par une information contextuelle tenant compte de l'historique de navigation (dans le graphe).

On peut s'attendre à ce que l'analyste développe, au gré de la navigation, une connaissance du réseau, de son interprétation. L'ensemble des chemins – des causalités – allant d'un objet à un autre, donne lieu à la visualisation d'une portion connexe du graphe global et offre un point de vue centré sur les objets de la requête. Dans le même ordre d'idée, l'analyste pourrait être intéressé à repérer un motif précis dans le graphe, autant en terme de topologie (structure du voisinage) que des attributs portés par les liens, voire de l'historique de navigation. Ainsi, nous envisageons qu'une requête puisse être formulée en terme de motifs à rechercher (dans le graphe). Cette question – difficile parce qu'elle sous-tend d'une certaine manière le problème du calcul d'homomorphismes de graphes – rejoint une problématique très actuelle en fouille interactive des graphes, notamment dans le domaine de la bio-informatique[37].

Dessin, representations multi-niveaux et interactions

Le problème de la représentation par dessin d'un graphe (le « Graph Drawing ») occupe une place bien définie dans la littérature en informatique (cf. références rappelées dans la note 17 ci-dessus). S'ajoutent à ces représentations par sommets et arêtes (segments de droites), d'autres modes mettant moins l'accent sur la topologie du graphe que sur les attributs de ses éléments.

---

[33] Dorogovtsev, S. N. and Mendes, J. F. F. (2003). *Evolution of Networks : From Biological Nets to the Internet and WWW*, Oxford University Press.

[34] Gaume, B., Venant, F. et al. (2006). *Hierarchy in lexical organization of natural language*. In *Hierarchy in natural and social sciences*. D. Pumain, Springer: 121-143.

[35] Auber, D., Chiricota, Y. et al. (2003). *Multiscale navigation of Small World Networks*. IEEE Symposium on Information Visualisation, 75-81, IEEE Computer Society.

[36] Daudin, J.-J., Picard, F. et al. (2006). *A mixture model for random graphs*, INRIA Futurs, report no RR-5840 (to appear in Stat. Comput. 2008).

[37] Lacroix, V., Fernandes, C. G. et al. (2006). *Motif Search in Graphs: Application to Metabolic Networks*. IEEE/ACM Transactions on Computational Biology and Bioinformatics 3(4): 360-368.

Jin, R., McCallen, S. et al. (2007). *Trend Motif: A Graph Mining Approach for Analysis of Dynamic Complex Networks*. Seventh IEEE International Conference on Data Mining, 541-546, IEEE Computer Society.

A l'évidence, TANGUY exige de revoir cet attirail d'approches afin de concevoir des représentations se pliant aux exigences des graphes dynamiques. Soulignons que dans notre cas, la dynamique peut venir tout autant du caractère changeant des données que de l'interaction utilisateur. De plus, les approches décidant de dessin a posteriori, faisant en quelque sorte le bilan de toutes les évolutions que le graphe a connu[38], n'apportent qu'une solution partielle à la situation que nous souhaitons étudier. Il s'agit bien de pouvoir ajuster la représentation au fur et à mesure de l'évolution du graphe, impossible à prédire tant dans son résultat que dans son origine. Des approches par animation de dessins successifs ont déjà été proposées et pourraient être envisagées[39]. Cela exige de bien prendre en compte l'effet de ces bouleversements sur la « carte mentale » de l'utilisateur[40].

Le dessin lui-même – les positions calculées pour chacun des sommets, ne suffira vraisemblablement pas à rendre compte de l'évolution du réseau. Par ailleurs, le dessin à lui seul ne suffit pas pour mettre en exergue les éléments saillants du graphe. Il faut en tout état de cause calculer pour les éléments du graphe des attributs graphiques (gradient de couleur, taille, etc.) associés à leur importance relative.

Nombre de statistiques sur les graphes (centralités dans les réseaux sociaux, ranking dans les hypertextes, indices de cohésion, etc.) peuvent être mobilisées à cette fin. Cela dit, il reste à trouver comment elles peuvent contribuer à suivre l'évolution, à repérer le passage d'un élément d'un rôle marginal à un rôle plus central (l'identification des signaux faibles). Vient ensuite la nécessité de décider d'artifices graphiques adaptés à la dynamique : nombre de caractéristiques graphiques ont un effet préemptif (perception visuelle immédiate) et sollicite de manière très efficace l'utilisateur ; on peut penser à des variations du gradient de couleur, à des effets oscillatoires (gradient de couleur / position / taille) ; ou encore à des effets radar – un peu à l'image des écrans des contrôleurs aériens, pour indiquer la région de l'écran qui subit l'impact des changements récents[41].

Représentations multi-niveaux et interactions

Le scénario de navigation typique est celui où l'utilisateur alterne entre vues générales et vues détaillées de l'espace d'information pour localiser un voisinage d'intérêt, avant que de se déplacer dans les zones périphériques à son centre d'intérêt. Partant de la vue générale, on peut imaginer dévoiler plus de détails dès lors que l'utilisateur effectue un zoom sur la représentation : de ce fait, le zoom géométrique s'accompagne d'un *zoom sémantique*.

Le mécanisme de vues « focus+contexte », parfois aussi appelé « Overview + detail », conjugue vues locales et vues globales en offrant une vue détaillée centrée sur un élément du graphe placée dans un contexte calculé à partir de la globalité du graphe. En d'autres mots, la requête de l'utilisateur aura marqué son intérêt pour certains éléments dont il veut connaître le voisinage (au sens de la distance sémantique et de la topologie du graphe). Cela dit, on peut tout de même maintenir une vue globale du graphe pour indiquer comment le reste de l'espace d'information s'organise par rapport à la requête de départ. Des éléments, même s'ils ne sont pas vus comme voisins de la requête, se trouvent peut-être assez proches ou mériteraient que la navigation s'y attarde.

Tout comme le zoom sémantique, le mécanisme « focus + contexte » exige de hiérarchiser l'espace, propose de fait une représentation multi-niveaux de l'espace d'information. Notons que cette hiérarchie de niveaux apparaît naturellement dans nombre de domaines[42]. Il y a ici à trouver une approche montrant la pertinence représentations multi-niveaux dans le cadre de la construction et de l'exploration des réseaux sémantiques.

---

[38] Voir les references suivantes :
Frishman, Y. and Tal, A. (2004). *Dynamic Drawing of Clustered Graphs*. IEEE Symposium on Information Visualization (INFOVIS'04).
Gaertler, M. and Wagner, D. (2006). *A Hybrid Model for Drawing Dynamic and Evolving Graphs*. Graph Drawing, 13th International Symposium, GD 2005, Limerick, Ireland, Springer.
Frishman, Y. and Tal, A. (2007). *Online Dynamic Graph Drawing*. Eurographics / IEEE VGTC Symposium on Visualization (EuroVis).

[39] Garton, L., Haythornthwaite, C. et al. (1997). *Studying Online Social Networks*. Journal of Computer-Mediated Communication 3(1).
Huang, M. L. and Eades, P. (1998). *Online Animated Graph Drawing using a Modified Spring Algorithm*. Journal of Visual Languages and Computing 9(6): 623-645.

[40] Misue, K., Eades, P. et al. (1995). *Layout Adjustment and the Mental Map*. Journal of Visual Languages and Computing 6: 183–210.

[41] Saulnier, A. (2005). *La perception du mouvement dans les systèmes de visualisation d'informations*. 17ème Conférence Francophone sur l'Interaction Homme-Machine, Toulouse, France, ACM Press.

[42] Pumain, D., Ed. (2006). *Hierarchy in Natural and Social Sciences*. Methodos Series, Springer.

Une technique commune consiste à calculer un clustering du graphe global encodé dans un arbre ; l'arbre décrit alors comment les clusters s'emboîtent les uns dans les autres. Classiquement, on donne accès à l'utilisateur à la structure d'emboîtements de manière à lui permettre de choisir le niveau de détail souhaité (continent -> pays -> région -> ville, par exemple), un peu comme l'explorateur de fichiers usuel.

Partant de cet arbre et d'une mesure de distance entre les clusters (généralisant la distance entre les sommets du graphe), la mécanique calcule une coupe de l'arbre fonction d'un point d'intérêt. La coupe est alors plus profonde autour du point d'intérêt et remonte au niveau les plus haut (les pus abstrait) lorsqu'on s'en éloigne. C'est là l'essentiel de la proposition de Furnas (1986)[43] qui reste à la base des interfaces « zoomables »[44] et autres vues « focus + context »[45].

Le défi encore une fois consiste à revoir ces mécanismes de navigation et prendre en compte la dynamique du graphe. Partant d'une vue « focus+contexte » (locale/globale) du graphe basé sur une hiérarchisation de l'espace : comment tenir compte de modifications apportées au graphe ? Comment tenir compte de l'historique de navigation ? Les approches de clustering connues ne proposent pas de mécanismes adaptatifs d'ajustement local du clustering / plutôt que de relancer le calcul sur l'ensemble du graphe (et sans effet mémoire). Comment adapter le mécanisme de Furnas à la navigation de graphes dynamiques ? Comment faire évoluer la coupe au gré de l'évolution de l'espace d'information ?

Cette hiérarchisation de l'espace d'information se trouve aussi à l'interface du pôle « Réseau sémantique » dans la mesure où elle pourrait avoir un impact sur la recherche d'information effectuée pour répondre aux requêtes utilisateur. Plus en amont, l'analyse linguistique pourrait bien guider le calcul de la hiérarchie.

1.4 Positionnement du projet

Nous résumons dans le tableau 1 ci-dessous les axes de recherche différentiants du projet TANGUY vis-à-vis de l'état de l'art des outils alliant analyse du texte, représentation des connaissances et visualisation interactive.

	Etat de l'Art	**Objectif TANGUY**	**Travaux de Recherche**
Philosophie générale	Service centralisé, mutualisé, pour accéder à de grosses masses d'information. Gros travail d'ingénierie au préalable. Reste dans une logique documentaire	Usage individuel personnalisé pour résoudre des problèmes précis. C'est l'utilisateur qui fait émerger les éléments de résolution de son problème par interaction avec des « petits morceaux » de textes et de connaissances	Concevoir une architecture d'usage où l'utilisateur a l'initiative de la résolution, après avoir formulé son problème, en étant guidé par un « retour » visuel
Architecture	Etapes séquentielles figées et indépendantes, définies par des spécialistes de grammaires, modèles de données, ontologies. Utilisateur en position de consommateur éloigné des producteurs	Couplage fort et symétrique entre linguistique, repré-sentation des connaissances et visualisation. Utilisateur aux commandes en position d'acteur. Il peut injecter à tout moment de nouvelles informations ou connaissances sous forme de textes	Rendre chaque module dynamique, coopérant, interactif, rapide. . Grande isotropie des informations, tout en permettant à l'utilisateur de s'y retrouver à sa manière
Pôle Linguistique	Spécialisation, verticalisation complexe et coûteuse des grammaires pour augmenter la compréhension par des connaissances non linguistiques. Système longuement « câblé » pour un corpus et un métier donnés	Etendre horizontalement les capacités linguistiques générales pour capter plus de sens, indépendamment de toute connaissance métier . Système immédiatement opérationnel pour analyser un corpus	Etudier des phénomènes nouveaux intrinsèques à capter localement dans les textes : temporalité, enchaînement, rupture, modalités d'expression

[43] Furnas, G. W. (1986). *Generalized Fisheye Views*. Human Factors in Computing Systems CHI '86, ACM Press.
[44] Bederson, B. B., J. Meyer, et al. (2000). *Jazz: An Extensible Zoomable User Interface Graphics Toolkit in Java. User Interface and Software Technology* (UIST 2000), ACM Press.
[45] van Wijk, J. J. and van Ham F. (2004). *Interactive Visualization of Small World Graphs*. IEEE Symposium on Information Visualisation, Austin, TX, USA, IEEE Computer Science press.

Pôle Représentation	Représente un univers simplifié : entités nommées, catégories, attributs, relations binaires, évènements locaux à une phrase	Faire émerger et représenter des objets complexes non modélisables à priori : un conflit, une conjonction d'intérêt, un mensonge, une relation de cause effet	Etudier des métamodèles variés de représentation, mettre en place des opérateurs de calcul et de raisonnement sur les structures (« calcul littéraire »)
Pôle Visualisation	Partie visible (par l'utilisateur) de l'ensemble du système, et souvent perçue comme la valeur ajoutée mais non essentielle. Souvent placée en aval de la chaîne et/ou en lecture seule. Représentation d'objets « artificiels » : graphes de réseaux sémantiques.	Point d'entrée et poste de pilotage permanent du dispositif. Moyen d'action sur les autres pôles. Représenter visuellement des phénomènes qui ont du sens : analogie, cause à effet, …	Interagir avec un continuum d'informations entre des phrases réelles et des concepts issus de l'interprétation. Introduire plusieurs échelles / niveaux de représentation.

Tableau 1 : Vision schématique du projet TANGUY

En résumé, l'hypothèse d'innovation qui sous-tend TANGUY est que – in fine – seul l'individu donne du sens à l'information , qu'elle soit textuelle ou graphique, et que la clé est d'utiliser les technologies pour amplifier la zone de contact entre l'information et l'individu

Le projet est déposé dans le cadre de l'appel Contenus et Interactions. Tanguy traite explicitement de « **Contenus** » et « **Interactions** » : nous travaillons sur des contenus sous forme textuelle, qui sont en train de devenir le matériau de travail de très loin le plus utilisé des cols blancs, et nous nous intéressons à de nouvelles manières d'interagir avec ces contenus. Tanguy assiste le travailleur du savoir dans sa tâche essentielle : créer de nouveaux contenus à forte valeur ajoutée à partir de contenus existants.

Plus précisément, nous répondons aux **thématique 2** (*Production, Assemblage et mise à disposition de contenus et de connaissances*) **thématique 3** (*Accès et échange de contenus*) de cet appel. Nous avons souligné dans les extraits des sous-thèmes les éléments clés que nous souhaitons appliquer dans TANGUY :

- Axe thématique 2 - sous-thème Agrégation de contenus et de connaissances, édition, Web2.0
  « *Il s'agit donc de faciliter la **mise en forme de contenus riches dans un mode prêt à la consommation**. Ces programmes devront permettre un **accès linéaire et non-linéaire aux contenus**. Il faut s'appliquer à fournir des solutions pour l'**agrégation de contenus hétérogènes de sources diverses** (professionnelles, privées, Web2.0,…) et de formats variés. Ces agrégations pourront se faire en temps réel ou différé et avec des outils semi-automatiques ou automatiques d'enrichissement* »

- Axe thématique 2 - sous-thème Passage du contenu à la connaissance
  « *Dans le foisonnement des contenus multimodaux de sources diverses, il est problématique d'identifier ce qui relève de la connaissance. De plus, la forme de mise à disposition des médias audiovisuels n'est pas forcément adaptée au transfert de connaissances. **Le passage du contenu à la connaissance nécessitera de nouveaux processus de présentation, de mise à disposition et d'enrichissement interactif des informations multimodales.** On peut citer l'élaboration d'ontologies adaptées au contenu multimédia. […] L**es résultats attendus seront principalement liés à des sujets développant des outils de mise en forme de contenus qui intègrent des points intermédiaires interactifs permettant de mesurer la manière dont la connaissance est perçue*** ».

- Axe thématique 2 - sous-thème Gestion du patrimoine numérique et indexation
  *« L'enjeu ici consiste à pouvoir traiter et gérer de façon simultanée de **grandes masses d'informations** (en flux ou en stock), **complexes, hétérogènes** et multimodales afin de **faciliter leur accès, leur visualisation**, leurs modifications et leur stockage. De plus, pour être viable, la création de contenus, de connaissances et programmes nécessite d'implémenter la possibilité de réutilisation simplifiée des informations par des utilisateurs d'origines diverses : professionnels, publics ou privés. Il convient donc de développer des technologies d'indexation enrichissant les contenus et facilitant l'accès aux contenus et à la connaissance* ».

- Axe thématique 3 – sous-thème Les outils de recherche et de navigation multimédia, multilingues, multimodaux, interactifs, coopératifs et adaptatifs
  « *Ce thème concerne le développement d'**outils de recherche et de navigation**, qui permettent d'**améliorer l'accès aux contenus sur des bases sémantiques**, et sémiologique pour des contenus multimédia et multilingues. Ces outils doivent fournir des **moyens nouveaux de navigation multimodale et des outils d'assistance pour la fusion et le traitement de données, adaptables ou personnalisables pour les utilisateurs et permettant un enrichissement des contenus**. Ce thème*

concerne aussi les outils coopératifs pour la recherche et l'annotation de contenus par les utilisateurs ».

1.5 Description des travaux : programme scientifique et technique

Le projet vise à mettre en place un système pour stocker des « micro-connaissances », extraites des textes, dans un réseau sémantique. Des outils de visualisation de graphes et des opérateurs d'aide au raisonnement seront développés sur ce réseau pour en exploiter les données contenues.

Nous proposons de découper le projet par les sous-tâches suivantes :

- T1 : Architecture générale du système
- T2 : Traitement automatique de la langue
- T3 : Réseau sémantique et opérateurs de raisonnement
- T4 : Visualisation de graphes
- T5 : Plate-forme
- T6 : Expérimentations

La tâche T1 du projet consiste, après un état de l'art des différentes méthodes (TAL, réseau sémantique et visualisation/interaction) à construire un méta-modèle, commun aux trois méthodes, qui sera utilisé pour faire communiquer les différents composants entre eux.

La tâche T2 porte sur le pôle Traitement Automatique de la Langue. Outre le développement des ressources et processus automatiques pour le traitement des documents (dictionnaires, analyses syntaxique et sémantique - identification des entités nommées, des relations entre entités, détection d'événements, traitement de la co-référence, de la spécificité des e-mails pour le traitement automatique…), cette tâche se focalisera sur l'élaboration de méthodes d'extraction de concepts et de leur représentation dans le réseau sémantique.

La tâche T3 porte sur le pôle Réseau sémantique. Une nouvelle implémentation de gestionnaire de réseaux sémantiques sera construite à partir de l'expérience du développement du produit Ideliance de Thales et des limitations ressenties par ses utilisateurs. Ce serveur permettra des enrichissements manuels du réseau sémantique pour permettre l'expression directe de l'interprétation de l'utilisateur. Des opérateurs de calcul sur le réseau sémantique seront développés pour exploiter les résultats et aider au raisonnement. Des rapports d'analyse publiables seront construits pour tracer et résumer l'activité de l'utilisateur.

La tâche T4 porte sur la visualisation et l'interaction. Un module de visualisation et de navigation dans le graphe sera construit. Des visualisations spécifiques seront développées sur les résultats des opérateurs construits dans la tâche T3. Un module de production de rapports synthétiques sur les résultats trouvés sera également fourni.

La tâche T5 porte sur l'intégration du méta-modèle (tâche T1) et des composants développés précédemment (tâches T2, T3 et T4) sur une même plate-forme, facilitant d'autant les échanges entre les partenaires du projet. La plate-forme sera développée au-dessus de la plate-forme TULIP développée par l'INRIA. Cette plate-forme contiendra la base de données recueillant l'ensemble des données. Des modules d'import de données et d'export de données seront intégrés à la plate-forme.

La tâche T6 est une tâche d'expérimentations. Nous avons choisi le domaine juridique car c'est un domaine où les textes, la précision et le raisonnement sont fondamentaux. Des cas réels de données de procédures seront utilisés. De plus, un cas avec des données volumétriques sera utilisé avec les données ENRON (disponibles sur Internet). Enfin, des données d'autres utilisateurs seront utilisées.

L'architecture d'un tel système peut être représentée dans le schéma ci-dessous (Figure 6) :

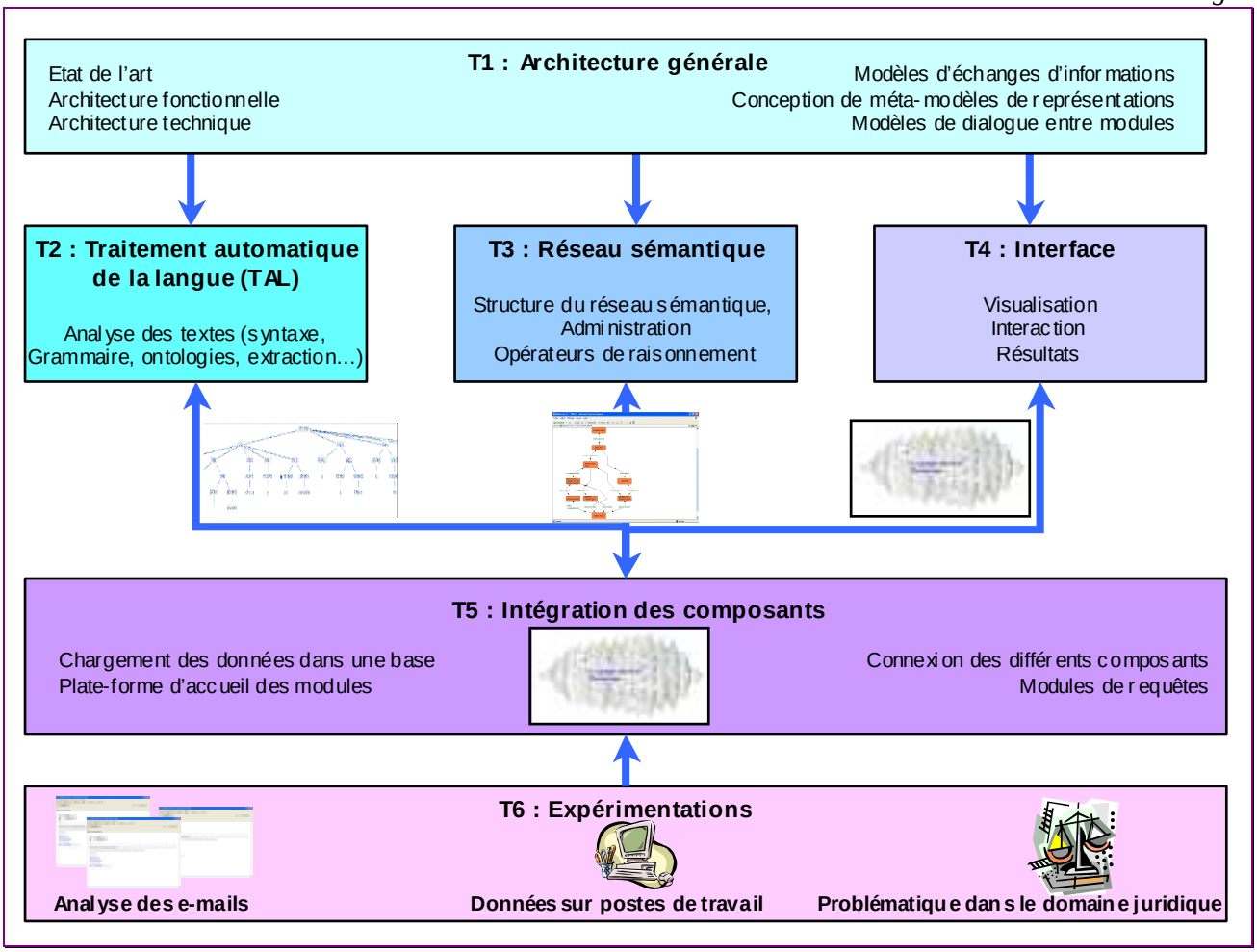


Figure 6 : Architecture de TANGUY.

1.5.1 Architecture générale

Tâche	T.1	Architecture générale
Responsable	THALES	

Partenaires	XEROX	INRIA			

Objectifs : Le système mis en place doit stocker les « micro-connaissances », extraites des textes, dans un réseau sémantique. Ce réseau sera ensuite exploité par des outils de visualisation de graphes et par des outils de requêtes. A partir d'un état de l'art effectué pour les trois pôles de recherche, un méta-modèle de représentation des connaissances et des modèles d'échange d'informations et de dialogue seront définis dans cette tâche pour permettre la communication entre les trois pôles : extraction (des « micro-connaissances »), représentation des connaissances (dans le réseau sémantique) et visualisation des données (visualisation de graphes). La philosophie nouvelle de l'interaction entre l'utilisateur et l'information sera également définie dans cette tâche.

Démarche détaillée :

ST1.1 : Etat de l'art des différentes approches

L'état de l'art réalisé pour chacun des trois pôles (TAL, réseau sémantique et visualisation/interaction) sera réalisé et devra faire apparaître des moyens de faire communiquer les trois pôles entre eux. Cet état de l'art servira à la définition du méta-modèle décrit en ST1.2.

ST1.2 : Définition d'un méta-modèle de représentation des informations

La sortie de l'extraction d'informations (résultats du TALN) sera l'entrée du réseau sémantique et le réseau sémantique sera l'entrée de la visualisation de graphes. Par conséquent, un méta-modèle doit être défini avec tous les partenaires impliqués dans le projet pour partager une même vision des données. Ce méta-modèle sera construit en partant de l'étude des sorties du produit XIP de Xerox, des résultats de la définition

d'un nouveau modèle de « représentation naturelle », et de la structure des données du produit Tulip de l'INRIA. Il se rapprochera le plus possible d'un « modèle naturel » au sens défini plus haut .

ST1.3 : Définition des modèles d'échange d'informations et de dialogue
Des modèles d'échange d'informations et de dialogue seront définis pour passer d'un pôle à un autre : visualisation des opérations de raisonnement sur le réseau sémantique, enrichissement manuel du réseau et prise en compte par les outils du TALN (lexique, ontologies utilisateur…).

ST1.4 : Définition de l'architecture fonctionnelle et technique
Une architecture fonctionnelle et technique seront définies pour construire les briques d'un possible futur prototype. S'agissant d'un projet de recherche, nous privilégierons les capacités de l'architecture à permettre l'expérimentation, l'implémentation rapide de variantes d'interaction, tout en garantissant la performance. Nous envisageons de travailler avec la plate-forme libre Tulip comme base de l'architecture, ce qui revient à mettre en avant l'interaction plutôt que le stockage de l'information comme cela est fait traditionnellement.

Fournitures :
- **L1.1** : Rapport sur l'état de l'art
- **L1.2** : Structure du méta-modèle de représentation des données
- **L1.3** : Modèles d'échange d'informations et de dialogue
- **L1.4** : Architecture fonctionnelle et technique

1.5.2 Traitement automatique de la langue (TAL)

Tâche	**T.2**	**Traitement automatique de la langue (TAL)**
Responsable	**XEROX**	

Partenaires	**THALES**	**INRIA**			

Objectifs : Développement des ressources et processus automatiques pour le traitement des documents : dictionnaires, synonymes, antonymes… ; analyse syntaxique ; analyse sémantique : Identification des entités nommées, des relations entre entités, la détection d'événements, le traitement de la co-référence, de la spécificité des e-mails pour le traitement automatique… Elaboration des méthodes d'extraction de concepts et de leur représentation dans le réseau sémantique.

Démarche détaillée :
ST2.1 : La prise en compte des spécificités du genre e-mail pour l'analyse automatique
- Prétraitement des erreurs d'orthographe
- Style épistolaire plus proche du langage parlé que les textes écrits pour un public
- Extraction et utilisation des champs de format e-mail : auteur, destinataire(s), sujet, date

ST2.2 : Analyse automatique des documents
- Extraction des :
  - entités nommées (personnes, organismes, lieux)
  - relations entre entités
  - descriptions des événements pertinents. A chaque événement nous assignons la source d'information ainsi qu'un indicateur de factualité
  - Expressions temporelles
- Traitement de la coréférence,
- Analyse conceptuelle pour aider la recherche dans les collection de documents

ST2.3 : Identification des éléments des analyses effectuées en ST1.2 repris dans les réseaux sémantiques
- Normalisation des entités nommées
- Normalisation des événements
- Choix pour la prise en compte des structures syntaxiques
- Choix pour la prise en compte des concepts

Fournitures :

- **L2.1** : Prétraitement des e-mails
- **L2.2** : Implémentation des règles pour le traitement sémantique
- **L2.3** : Représentation des résultats de l'analyse textuelle sous la forme de réseau sémantique

1.5.3 Réseau sémantique

Tâche	**T.3**	**Réseau sémantique**
Responsable	**THALES**	

Partenaires	**XEROX**	**INRIA**			

Objectifs : Une nouvelle implémentation de gestionnaire de réseaux sémantiques sera construite à partir de l'expérience du développement du produit Ideliance de Thales et des limitations ressenties par ses utilisateurs, en remettant complètement en cause les principes actuels. En particulier, les mécanismes d'émergence seront généralisés, et des mécanismes seront proposés pour représenter des concepts plus complexes que les seules entités, relations et catégories actuelles. Cet outil assurera les fonctions d'administration et de gestion du réseau sémantique. Ce serveur permettra des enrichissements manuels du réseau sémantique pour permettre l'expression directe de l'interprétation de l'utilisateur. Des opérateurs de calcul sur le réseau sémantique seront développés pour exploiter les résultats et aider au raisonnement. Des rapports d'analyse publiables seront construits pour tracer et résumer l'activité de l'utilisateur.

Démarche détaillée :

ST3.1 : Définition des objets représentés et de sémantique d'interaction et de visualisation

Choix pour représenter le continuum d'information depuis la structure des documents de départ jusqu'aux argumentations / raisonnement construits par l'utilisateur

Détermination précise des paradigmes naturels d'interaction (édition, navigation, requêtes)

ST3.2 : Conception et réalisation de l'implémentation

Choix des structures de données de bas niveau pour favoriser l'interactivité et le dynamisme, au-dessus des structures de graphe de Tulip.

Définition d'une machine virtuelle de calcul sur le réseau pour implémenter de manière générale les opérations de raisonnement (à priori en logique du premier ordre)

ST3.3 : Procédures d'alimentation du réseau sémantique

Des procédures seront développées pour alimenter le réseau sémantique. Les moyens suivants sont prévus :

- Intégration automatique des « micro-connaissances » faites dans la tâche T2
- prise en compte de données structurées (des e-mails en particulier : émetteur, destinataires, dates),
- enrichissement du réseau par l'utilisateur (données individuelles)
- intégration d'ontologies utilisateur

ST3.4 : Opérateurs d'aide au raisonnement

Des opérateurs seront construits – au dessus de la machine virtuelle logique définie en ST3.2 – pour interroger le réseau sémantique et aider l'utilisateur à exploiter les résultats :

- Tableaux : tableaux simples ou croisés,
- OLAP : tableaux dynamiques permettant de naviguer dans différentes dimensions,
- « Quoi Entre » : chemins entre deux objets du réseau sémantique,
- Signaux faibles : identifier des conditions nécessaires pour qu'une nouvelle information contribue à la réponse à un problème explicitement représenté dans la base de connaissances,
- Clustering : trouver des groupes d'objets du réseau sémantique qui sont les plus homogènes possibles, leur homogénéité étant mesurée par l'ensemble des relations qu'ils ont en commun,
- Inférence : trouver de nouvelles relations dans le réseau, déduites de relations existantes (transitivité…),
- Outils de publication de rapports d'analyse.

Ces opérateurs seront développés étroitement dans la perspective de leur exploitation par les outils de visualisation de graphes.

Fournitures :
- **L3.1** : Définition des objets représentés
- **L3.2** : Rapport de conception et de réalisation de l'implémentation
- **L3.3** : Procédures d'enrichissement du réseau sémantique,
- **L3.4** : Constructions d'opérateurs d'aide au raisonnement et de publication de rapports d'analyse

1.5.4 Visualisation et interaction

Tâche	**T.4**	**Visualisation et interaction**
Responsable	**INRIA**	

Partenaires	**THALES**	**XEROX**			

Objectifs : Module de visualisation et de navigation dans le graphe. Visualisation des résultats des opérateurs construits sur le graphe pour l'aide au raisonnement. Production de rapports synthétiques sur les résultats trouvés.

Démarche détaillée :

ST4.1 : Analyse du réseau sémantique
- Revue (état de l'art) des différentes approches d'analyse de graphes petits mondes et invariant d'échelle (scale-free).
- Extension aux réseaux sémantiques (prises en compte des catégories d'objets, des types de relations).
- Extension aux graphes dynamiques – modifications topologiques et historique de navigation (usage).
- Conception de statistiques adaptatives, étude combinatoire, algorithmique, complexité.
- Calcul du voisinage rapproché (« distance sémantique »).

ST4.2 : Représentations des réseaux sémantiques
- Revue (état de l'art) des différentes approches de dessin des graphes dynamiques.
- Identification des paradigmes de représentations pertinents pour la visualisation des réseaux sémantiques (en liens avec les types d'objets et de relations, voire d'une taxonomie des tâches utilisateur et/ou des requêtes formulées sur les graphes).
- Prise en compte des statistiques adaptatives pour le calcul des représentations.
- Prise en compte des statistiques adaptatives structurelles et sémantiques pour le calcul de voisinage rapproché répondant à une requête ciblée
- Conception d'artifices visuels pertinents (adaptatifs et dynamiques).

ST4.3 : Représentations multi-niveaux
- Revue (état de l'art) des différentes approches de hiérarchisation de graphes (clustering multi-niveaux).
- Extension aux réseaux sémantiques : prise en compte des catégories d'objets et types de relations (combinaison d'indices topologiques et de valeurs nominales et/ou ordinales).
- Extension aux graphes dynamiques (prise en compte des statistiques adaptatives).
- Extension du mécanisme « focus+contexte » aux graphes dynamiques, prise en compte de la sémantique, de l'évolutivité du graphe, de l'historique de navigation.

ST4.4 : Construction de rapports d'analyse
- L'utilisateur pourra suivre les vues construites et annotées au fil de ses analyses, et les exporter dans divers rapports d'analyse.

Fourniture
- **L4.1** : Modules d'analyse (statistiques) du graphe.
- **L4.2** : Modules de représentations (dessins) du graphe.
- **L4.3** : Modules de navigation dans le graphe (représentations multi-niveaux).
- **L4.4** : Modules de construction de rapports de synthèse

1.5.5 Intégration des différents modules

Tâche	**T.5**	**Plate-forme d'intégration**
Responsable	**INRIA**	

Partenaires	**THALES**	**XEROX**			

Objectifs : Les briques développées dans les différentes tâches décrites ci-dessus (T2, T3, et T4) seront intégrées sur une même plate-forme, facilitant d'autant les échanges entre les partenaires du projet. Cette plate-forme contiendra la base de données recueillant l'ensemble des données. Des modules d'import de données et d'export de données seront intégrés à la plate-forme.

Démarche détaillée : le consortium travaillera à l'élaboration de modules spécifiques au projet venant s'ajouter à la plate-forme Tulip. Ainsi, chaque partenaire agira sur un segment du pipeline (section 1.3.1) propre à la nature de son expertise et de ses contributions. L'architecture de la plate-forme profite du mécanisme de plug-in, et facilite l'ajout de fonctionnalités par chaque partenaire. La conception, la réalisation et l'affinage des modules pourra ainsi se faire au travers de rondes d'intégration annuelles ou bi-annuelles.

ST5.1 : Prise en charge du méta-modèle
Le méta-modèle devra être matérialisé de manière cohérente avec le format de description propre à Tulip – qui a la particularité d'être simpliste et très ouvert en vertu de son mécanisme de propriétés très souple. L'import des données de l'analyse linguistique devra aussi se conformer au méta-modèle. Il faudra certainement à cet effet prévoir une passerelle et un module d'import en direction de la plate-forme de manière à articuler Tulip avec les outils d'analyse existants.

ST5.2 : Modules de requêtes
L'interrogation du modèle de données est l'un des éléments moteurs de la navigation. Le dispositif permettant de formuler les requêtes viendra lui aussi se greffer à la plate-forme. Du point de vue de l'implémentation, il s'agit de construire des plug-ins Tulip spécifiques à al plate-forme et au méta-modèle. Des plug-ins consacrés aux calculs déclenchés par les requêtes apparaîtront en sous-tâches.

ST5.3 : Interacteurs
Les modes de navigation propres à l'exploration et à l'interrogation des réseaux sémantiques devront être identifiés et spécifiés. Ils devraient donner lieu au développement de modes d'interaction spécifiques. Dans l'état actuel des choses, ces interacteurs requièrent un travail au niveau du cœur de la plate-forme, impliquant un retour vers le système de fenêtrage et de widgets sous-jacent (QT de TrollTech). A terme, les interacteurs répondront aussi au mécanisme de plug-ins.

ST5.4 : Modules de production de rapports de synthèse
Les rapports de synthèse interviendront typiquement au terme d'une analyse lorsque des vues construites par l'utilisateur final viennent corroborer ses hypothèses. Il s'agit aussi potentiellement de proposer un cadre pour l'échange de points de vue entre plusieurs utilisateurs, intégré à même la plate-forme.

Fournitures
Les fournitures sont listées ici de manière séquentielle, mais seront en réalité réalisées et affinées au travers d'itération annuelle ou bi-annuelle. Ainsi, des livrables de chaque type (1 à 6) pourront être produits pendant toutes les phases du projet et venir enrichir l'environnement de manière incrémentale.

- **L5.1 :** Livraison d'un langage de description prenant en charge le méta-modèle et conforme au langage de description de Tulip.
- **L5.2 :** Livraison de plug-in d'import Tulip permettant de prendre en charge les données extraites par les modules d'analyse linguistiques.
- **L5.3 :** Livraison de plug-in d'interrogation Tulip permettant de formuler des requêtes sur le graphe.
- **L5.4 :** Conception et réalisation d'interacteurs Tulip adaptés aux tâches (sous-jacentes aux types de requêtes envisagées).
- **L5.5 :** Livraison de plug-in d'export Tulip contribuant à la production de rapports de synthèse.

1.5.6 Expérimentations (domaine juridique)

Tâche	**T.6**	**Expérimentations (domaine juridique)**
Responsable	**FIDAL**	

Partenaires	**THALES**	**XEROX**	**INRIA**		

Objectifs : Analyse du besoin métier et étude détaillée de démarches d'analyse et de raisonnement (autour du domaine juridique). Application sur des cas réels.

Démarche détaillée :

ST6.1 : Expression de besoins (autour du domaine juridique)

L'associée du cabinet FIDAL qui coordonne cette tâche est spécialiste des litiges informatiques (problèmes de livrables, retards,…). Elle apportera son expertise dans l'analyse des e-mails sur ses affaires : passer des e-mails aux faits, puis des faits à leur analyse, et enfin de leur analyse à leur qualification juridique.

ST6.2 : Application sur des cas réels

Le cabinet FIDAL fournira des données portant sur des e-mails échangés pour des affaires réelles sur des litiges informatiques (entre 100 et 200 mails échangés par affaire). Ces données seront préalablement anonymisées avant d'être fournies aux différents partenaires du projet. Le cabinet FIDAL apportera son expertise métier sur les résultats qui auront été trouvés par les outils développés.

ST6.3 : Application sur l'affaire ENRON (base de données disponible sur Internet)

L'idée du projet est d'attaquer également de grosses bases d'informations. A cet effet, les données sur l'affaire ENRON (plusieurs centaines de milliers d'e-mails échangés) seront récupérées et analysées dans le cadre du projet. Le cabinet FIDAL apportera son expertise méthodologique pour traiter cette masse d'informations.

ST6.4 : Application sur d'autres données

Au travers des différents contacts des partenaires, d'autres jeux de données seront récupérés pour traiter d'autres problématiques juridiques : droit des grandes catastrophes (par exemple incendie du tunnel du Mont-Blanc), sécurité… Nous sommes également en contact avec le Ministère de l'Intérieur qui est prêt à apporter des expressions de besoins et des études de cas.

Fournitures

- **L6.1** : Expression de besoins pour le domaine juridique
- **L6.2** : Rapports d'étude sur des cas réels (cabinet FIDAL)
- **L6.3** : Rapport d'étude sur l'affaire ENRON
- **L6.4** : Rapports d'étude sur d'autres cas (sécurité, catastrophes, ministère de l'Intérieur)

1.6 Résultats escomptés et Retombées attendues

Tanguy est un projet de recherche industrielle. Ses résultats seront d'abord scientifiques. Le but est de proposer une nouvelle approche multidisciplinaire du traitement de l'information non structurée (linguistique, représentation des connaissances, interaction visuelle). L'activité de publication sera donc un facteur important de succès. Par ailleurs, les développements effectués autour de la plate-forme Tulip seront disponibles pour la communauté scientifique.

Développés dans un cadre d'application précis – le juridique – pour des raisons d'efficacité, les résultats de Tanguy pourront être transposables à beaucoup de situations :

- dans des applications métier spécialisées manipulant des informations non structurées complexes (exemples : dossier médical personnel, aide à la réponse à des appels d'offre, outil pour mener des audits, …)
- mais aussi comme poste de travail généraliste (exemples : aide à la recherche d'informations sur Internet, aide à l'organisation de ses informations personnelles, partage de connaissances,…)

A plus long terme, on peut voir Tanguy comme la préfiguration d'un nouveau type de « poste de travail » au sens « desktop » qui présenterait à chaque utilisateur une vision unifiée, interactive et orientée vers ses tâches de l'ensemble de son espace informationnel.

Critères de réussite et d'évaluation :
Une des manières d'évaluer les potentialiés de Tanguy sera de mesurer le confort apporté aux utilisateurs face au « stress informationnel » et au « trou noir cognitif » évoqués en tout début de cette proposition. Pour reprendre les propos directs d'un analyste professionnel habitué aux outils classiques de traitement de l'information : « j'attends d'un outil intelligent qu'il me permettre de rentrer une heure plus tôt à la maison tout en m'aidant à produire des résultats qui augmentent la satisfaction de mon chef »

Plus objectivement, des mesures de l'apport de Tanguy seront relativement faciles à établir :

- évaluer la part de l'information contenue dans des documents qui devient modélisée explicitement sous forme de représentation structurée
- évaluer le nombre d'items de connaissances qui deviennent visualisables à l'écran au lieu de rester implicitement cachés dans les textes
- mesurer la capacité à découvrir par analyse graphique globale des concepts qui ne peuvent pas être détectés localement par l'analyse linguistique
- sur l'exemple des données de l'affaire Enron, comparer la vitesse d'un utilisateur pour retrouver un concept précis avec Tanguy et avec des outils documentaires traditionnels

Chacun des partenaires a par ailleurs sa vision particulière des retombées du projet TANGUY dans son contexte :

THALES : Thales Communications développe de grands systèmes complexes allant de la fusion de multiples sources d'informations à la prise de décision (systèmes de commandement, systèmes de surveillance). Concevoir des postes de travail adaptés à la charge cognitive d'accès à l'information et au raisonnement est un élément déterminant dans l'architecture de tels projets. De plus, même dans les systèmes techniques, la part de l'information non structurée devient prépondérante. Tanguy apportera aux chefs de projet de Thales Communications de nouveaux paradigmes de conception ainsi que des solutions techniques pour augmenter la valeur ajoutée du service rendu aux utilisateurs, tout en diminuant les coûts et délais d'ingénierie. L'ensemble de ces facteurs accroîtra la compétitivité de Thales Communications sur l'ensemble de ses marchés, en particulier à l'International.

XEROX : Xerox se dirige vers un nouveau champ de recherche autour de la représentation des connaissances à partir des résultats de l'extraction d'informations. En effet, c'est une tendance générale des outils d'extraction (texte, image, vidéo…) à vouloir structurer, représenter, résumer et synthétiser les « micro-connaissances » extraites.

INRIA : l'environnement visé doit permettre à l'utilisateur de faire face aux données dynamiques, hétérogènes et au caractère changeant. Si la visualisation offre la partie visible d'un tel environnement, sa conception requiert une synergie avec toutes les phases amont, présentes ici dans le consortium du projet. Le projet TANGUY cadre parfaitement avec les objectifs que s'est donnée l'équipe INRIA GRAVITÉ nouvellement créée. La représentation graphique de données sémantiques est également un point important pour l'INRIA, qui découvre là un nouveau champ de recherche.

FIDAL : FIDAL traite de nombreuses affaires à partir de l'analyse du contenu des e-mails. Le cabinet espère trouver dans ce projet des moyens pour accroître l'exploitation des données qu'il a à traiter et gagner ainsi en productivité.

1.7 Organisation du projet

La coordination globale du projet se trouve sous la responsabilité de THALES qui cumule une expérience importante en montage et gestion de projets. L'équipe de THALES chargée de la coordination a assuré également le management, le design et la direction scientifique du projet Infom@gic (projet du pôle Cap Digital) qui regroupe 25 partenaires et qui correspond à un enjeu de travail de 40M€.

La coordination se fait à plusieurs niveaux :

- Le **responsable de projet** :
  - Coordonne l'exécution du projet,
  - S'assure de la qualité globale des fournitures,
  - Transmet les rapports et les fournitures au financeur,
  - Diffuse les fournitures documentaires au sein du consortium,
  - Anime le Comité de Pilotage,
  - Informe le Comité de Pilotage des éventuels risques, difficultés ou défaillances de partenaires,

- Saisit le Comité de Pilotage en cas de nécessité d'arbitrage.

- Les **coordinateurs scientifique / technique, un par pôle de recherche (TAL, réseau sémantique, visualition/interaction)** :
  - Coordonnent techniquement les travaux,
  - Consolident les plans de travail et s'assure de leur cohérence,
  - Suivent l'avancement et les fournitures.

- Les **responsables de tâches** :
  - Coordonnent techniquement les travaux et s'assure de leur bonne exécution,
  - Participent à la réunion d'avancement mensuel,
  - Mettent à jour régulièrement le planning de ses travaux et informe le coordinateur technique de l'avancement,
  - Valident les fournitures produites au titre de la tâche,
  - Informent le coordinateur technique en cas de risque, retard, défaillance, difficulté à accéder à des résultats issus d'autres tâches, ou tout autre difficulté.

La responsabilité scientifique/technique du projet sera confiée aux représentants des trois pôles de recherche du projet :

- Frédérique Segond de Xerox, pour le pôle Traitement Automatique du Langage,
- Jean Rohmer de Thales, pour le pôle Réseaux Sémantiques,
- Guy Mélançon, de l'INRIA Bordeaux, pour le pôle Visualisation et Interaction.

La coordination se déroulera de la manière suivante :

- Un comité de pilotage sera organisé deux fois par an. Il sera composé du responsable de projet et des coordinateurs scientifiques/techniques. Il sera animé par le responsable du projet, sa tâche consiste à :
  - S'assurer de la bonne exécution du projet , dont le contrôle et la coordination au quotidien sont délégués au responsable du projet,
  - S'assurer du respect de l'accord juridique,
  - Valider les projets de publications et communications relatives au projet,
  - Réaliser les arbitrages éventuels en cas de défaillance, ou de modification dans la répartition des parts.
- Des réunions d'avancement et de gestion seront organisées deux fois par an et serviront de base pour la rédaction des rapports d'avancement.
- Des réunions techniques entre les différents partenaires seront organisées en moyenne tous les trois mois. Certaines pourront prendre la forme de séminaires de travail de plusieurs jours lors des étapes charnières du projet.
- Un système de travail coopératif à distance sera mis en place pour la collaboration au quotidien.
- Un site du projet sera ouvert afin de mettre à disposition de la communauté scientifique les éléments intéressants obtenus au cours projet (dont les publications).

Le management de projet est détaillé ci-dessous.

Tâche	**T.0**	**Management de projet**
Responsable	**THALES**	

Partenaires	**XEROX**	**INRIA**			

Objectifs : Conduire les activités de gestion de projet et d'assurance qualité. Assurer également la représentation externe du projet vis-à-vis de l'ANR.

Démarche détaillée :

ST0.1 : Plan de management et d'assurance qualité

Un Plan de management et d'assurance qualité du projet sera défini. Il comportera la définition des buts et responsabilités, des documents de référence (normes et standards, etc), de la méthode et des outils de conduite de projet employés, de la documentation et de la gestion des modifications.

ST.0.2 : Coordination des partenaires

Cette tâche se déroulera sur la durée totale du projet et consistera à organiser toutes les réunions définies

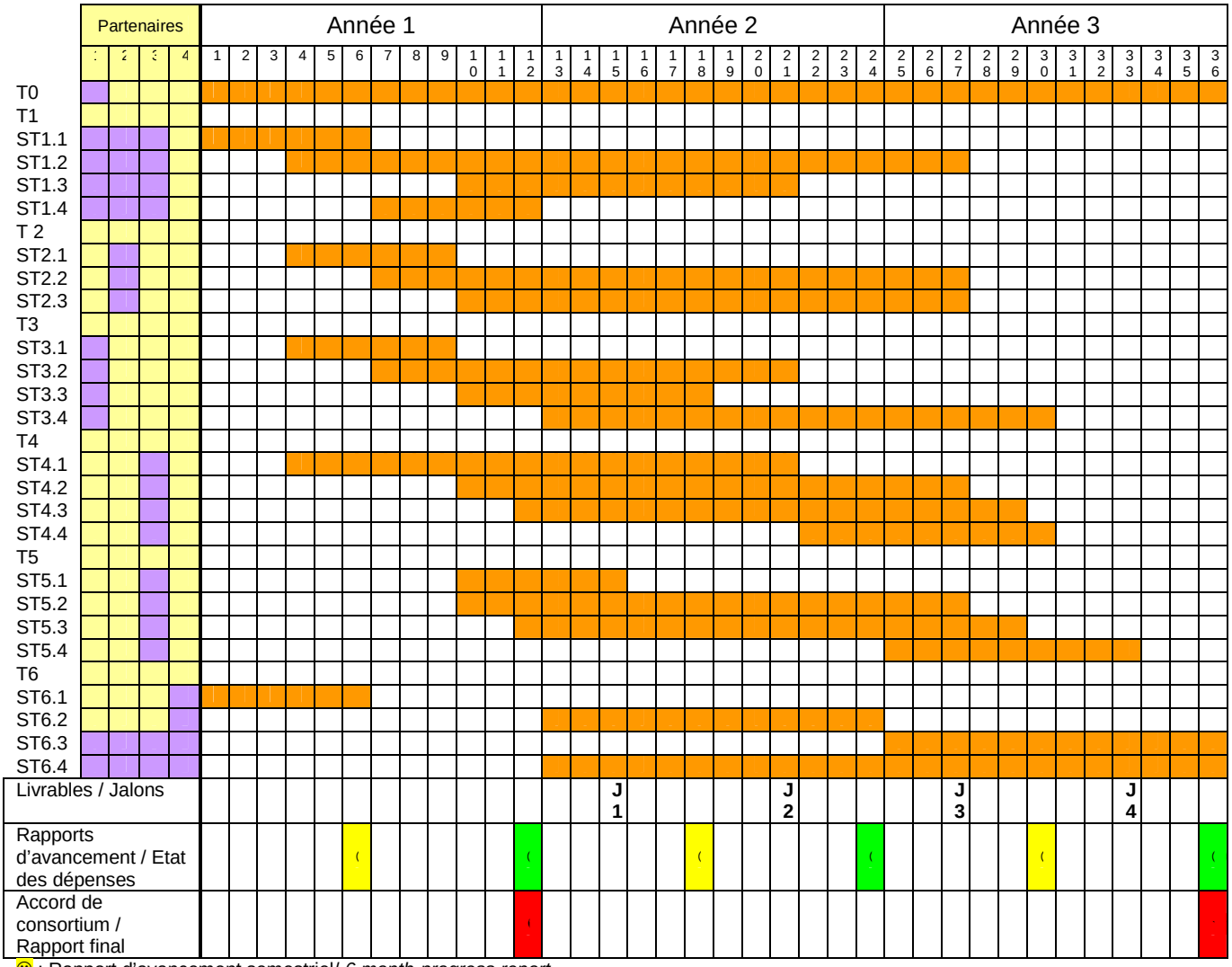
dans le plan de management (comité de pilotage, réunions d'avancement).

ST.0.3 : Représentation vis-à-vis de l'ANR
Cette tâche se déroulera sur la durée totale du projet et consistera à produire tous les rapports pour l'ANR.

Fournitures :

- **L0.1** : Plan de management et d'assurance qualité
- **L0.2** : Comptes rendus de réunion (comité de pilotage, réunions d'avancement)
- **L0.3** : Rapports d'étape pour l'ANR

Le planning général du projet et le chemin critique sont décrits dans la figure ci-dessous. Des versions itératives de la plate-forme intégrée commenceront en T0+15 (ce sera le premier jalon du projet) et des versions progressives de la plate-forme seront proposées tous les six mois (aux jalons T0+21, T0+27, T0+33).
Chacun de ces cycles de 6 mois concernera en fait l'ensemble des acteurs du projet : il sera l'occasion de la fixation de nouveaux objectifs à explorer et implémenter, tant en ce qui concerne les besoins exprimés, les concepts nécessaires à leur prise en compte, leur intégration et leur expérimentation.
Ainsi, toute la seconde partie du projet sera rythmée par ces itérations qui permettront d'orienter les recherches avec pertinence.



☺ : Rapport d'avancement semestriel/ *6 month-progress report*
☺ : Rapport d'avancement semestriel + état des dépenses/ *Progress report + expenses*
✪ : Accord de consortium/ *Consortium agreement*
☆ : Rapport de synthèse + récapitulatif des dépenses/ *Final report + expenses summary*

Les risques et les actions de réduction associées sont identifiées à chaque niveau de management (tâche, sous-projet, projet) et reportés au moins mensuellement aux responsables de SP et responsable de projet.

- Risques initiaux / Actions de réduction
  - Défaut de cohérence de certains travaux (redondances, incomplétude, incompatibilité technique …) :
    - Etablissement de plans de travail par sous-projet,
    - Mise en cohérence,
    - Ajustement au premier jalon d'avancement.
  - Limitation de la coopération due à des contraintes commerciales
    - Définition claire des connaissances existantes au début du projet et des résultats attendus au titre du projet,
    - Définition claire des droits de propriété et d'utilisation dès le début du projet,
    - Visibilité entière sur les engagements, les moyens mis en œuvre et les résultats produits entre partenaires.
  - Défaillance d'un partenaire
    - Réunion du comité de pilotage du projet en vue d'envisager son remplacement et/ou la redéfinition du périmètre du projet.

TABLEAU des LIVRABLES et des JALONS (le cas échéant)/ *Deliverables and milestones*			
Tâche	Intitulé et nature des livrables et des jalons	Date de fourniture nombre de mois à compter de T0	Partenaire responsable du livrable/jalon
0. Management de projet			
	L0.1 Plan de management et d'assurance qualité	T0+6	Thales
	L0.2 Comptes rendus de réunion	T0+6, 12, 18, 24, 36	Thales
	L0.3 Rapports d'étape pour l'ANR	T0+6, 12, 18, 24, 36	Thales
1. Architecture générale			
	L1.1 Rapport sur l'état de l'art	T0+6	Thales
	L1.2 Structure du méta-modèle de représentation des données	T0+12, T0+24	Thales
	L1.3 Modèle d'échange d'informations et de dialogue	T0+12, T0+21	Thales
	L1.4 Architecture fonctionnelle et technique	T0+12	Thales
2. Traitement automatique de la langue			
	L2.1 Prétraitement des e-mails	T0+9	Xerox
	L2.2 Implémentation des règles pour le traitement sémantique	T0+12, T0+18, T0+27	Xerox
	L2.3 Représentation des résultats de l'analyse textuelle sous la forme de réseau sémantique	T0+12, T0+24	Xerox
3. Réseaux sémantiques			
	L3.1 Définition des objets représentés	T0+9	Thales
	L3.2 Rapport de conception et de réalisation de l'implémentation·	T0+15, T0+21	Thales
	L3.3 Procédures d'enrichissement du réseau sémantique	T0+ 18	Thales
	L3.4 Constructions d'opérateurs d'aide au raisonnement et de publication de rapports d'analyse	T0+21, T0+30	Thales
4. Visualisation et interaction			
	L4.1 Modules d'analyse (statistiques) du graphe	T0+9	INRIA
	L4.2 Modules de représentations (dessins) du graphe	T0+15	INRIA
	L4.3 Modules de navigation dans le graphe (représentations multi-niveaux).	T0+21	INRIA
	L4.4 Modules de construction de rapports de synthèse	T0+27	INRIA
5. Intégration des différents modules			
	L5.1 Livraison d'un langage de description prenant en charge le méta-modèle et conforme au langage de description de Tulip	T0+15	INRIA
	L5.2 Livraison de plug-in d'import Tulip permettant de prendre en charge les données extraites par les modules d'analyse linguistiques	T0+15	INRIA
	L5.3 Livraison de plug-in d'interrogation Tulip permettant de formuler des requêtes sur le graphe	T0+21	INRIA
	L5.4 Conception et réalisation d'interacteurs Tulip adaptés aux tâches (sous-jacentes aux types de requêtes envisagées)	T0+27	INRIA
	L5.5 Livraison de plug-in d'export Tulip contribuant à la production de rapports de synthèse	T0+33	INRIA
6. Expérimentations (domaine juridique)			
	L6.1 Expression de besoins	T0+6	FIDAL
	L6.2 Rapport sur un cas réel	T0+24	FIDAL
	L6.3 Rapport d'étude sur l'affaire ENRON	T0+36	FIDAL
	L6.4 Rapports d'étude sur d'autres cas (sécurité, catastrophes, ministère de l'Intérieur)	T0+36	FIDAL

1.8 Organisation du partenariat

1.8.1 Pertinence des partenaires

Chacun des partenaires possède un noyau de compétences qui le situe parmi les mieux placés en Europe vis à vis des objectifs du projet Tanguy.

- **XEROX**, avec XIP, est un des seuls acteurs mondiaux à poursuivre depuis des années (400 h/ ans cumulés) le développement et l'amélioration de grammaires générales dans plusieurs langues**.** Le moteur XIP (pour Xerox Incremental Parsing) accepte en entrée n'importe quel document écrit (ou partie de document) et produit en sortie une représentation grammaticale du contenu textuel de ce document. XIP peut analyser des phrases très complexes du point de vue de la structure. C'est la raison pour laquelle on dit qu'il est « robuste ». L'analyseur syntaxique est le module de base de toute analyse "intelligente" des textes. Cet analyseur syntaxique robuste XIP fournira en premier lieu une sortie générique décrivant les différentes relations de dépendance reliant les différents constituants. C'est cette sortie qui servira de base à des couches de traitement ultérieures. C'est une des seules technologies disponibles sur le marché qui fournit des résultats de compréhension de texte indépendants des connaissances du domaine, qui sont précisément représentées ailleurs dans Tanguy.

- **Thales** Thales possède avec Idéliance une expérience rare dans le développement et surtout l'utilisation de réseaux sémantiques comme outil quotidien de représentation de connaissances, depuis plus de dix ans. Thales a formé à l'usage des réseaux sémantiques des utilisateurs très exigeants dans le domaine du renseignement militaire, et amène de ce fait une expérience et une expression de besoin uniques. Thales maîtrise également le savoir-faire d'implémentation « au niveau du bit » des réseaux sémantiques, avec une efficacité bien supérieure aux moteurs génériques issus des travaux sur le Web Sémantique, tout en garantissant des passerelles avec ce dernier. Cette maîtrise d'implémentation va être fusionnée avec celle possédée par l'INRIA pour faire évoluer la plate–forme Tulip.

- L'équipe **GRAVITÉ** de l'**INRIA Bordeaux – Sud-Ouest** travaille à la visualisation et l'exploration interactive de graphes. http://www.inria.fr/recherche/equipes/gravite.fr.html. L'équipe développe des formalismes, des méthodes et des outils venant en appui à la fouille et à l'analyse de données massives, en mettant l'accent sur les données modélisées par des graphes, favorisant la découverte et la compréhension de phénomènes sous-jacents aux grands corpus de données. Au-delà des verrous posés par le volume massif des données à traiter, l'équipe développe des approches permettant de traiter des données dynamiques et leur caractère incertain et changeant. L'équipe développe la plate-forme de visualisation de graphes Tulip (www.tulip-software.org) destinée à la visualisation interactive de grands graphes. Tulip est écrit en C++ et repose sur QT (TrollTech) et OpenGL ; la plate-forme est disponible depuis SourceForge et connaît un succès grandissant. Les évolutions futures de la plate-forme incluent la mise-à-jour et l'ajout de fonctionnalités au travers de web services, évoluant ultimement en un atelier comparable à Eclipse (IDE développé par IBM AlphaWorks). Avec TULIP, L'INRIA dispose d'une plate-forme de développement d'environnements de visualisation de très grands graphes, qui dépasse de très loin les habituels outils de visualisation en environnement Java. En particulier, la combinaison de la maîtrise des structures de données de bas niveau optimisées et d'une importante bibliothèque de calcul sur les graphes offre des fonctionnalités que seules deux ou trois autres acteurs peuvent offrir dans le monde.

- Le cabinet **FIDAL** a été choisi comme end-user pour illustrer l'utilisation des travaux sur la plate-forme. Nous rappelons que le domaine juridique a été choisi car c'est un domaine où les textes, la précision et le raisonnement sont fondamentaux. Le cabinet FIDAL, avec son expertise dans le domaine des nouvelles technologies de l'information et de communication et des exemples qu'il apportera, permettra de valider d'un point de vue métier les résultats trouvés dans ce projet.

1.8.2 Complémentarité des partenaires

Xerox et Thales ont déjà collaboré ensemble sur des projets d'extraction d'informations et de représentation des connaissances, en particulier sur le projet Infom@gic. Xerox fournit actuellement un des meilleurs outils du

marché qui extrait un grand nombre d'entités nommées, de relations binaires entre ces entités et de détection d'événements, de concept matching.

Thales, avec les travaux effectués autour d'Ideliance, a acquis une expérience importante sur la définition d'une bonne architecture pour optimiser l'utilisation et l'exploitation des réseaux sémantiques par des end-users. Les réseaux construits seront à partir des données fournies par Xerox seront d'une taille considérable et il est nécessaire d'avoir des outils de visualisation très performants pour travailler sur de tels réseaux.

Concernant l'INRIA, la plate-forme Tulip permet de visualiser de très grands graphes (dans le domaine de la biologie, de l'analyse de l'information, …). C'est donc un partenaire idéal pour pouvoir traiter de graphes sémantiques. Les structures internes de Tulip étant particulièrement optimisées en vitesse et occupation mémoire peuvent être directement réutilisées par Thales pour implémenter une nouvelle génération de réseau sémantique. Cala garantit d'une part une totale harmonie entre les aspects représentation et visualisation, et évite d'autre part de dépenser des ressources dans des travaux de programmation sans grande valeur ajoutée.

Le cœur de compétence de chacun des partenaires n'est pas aujourd'hui maîtrisé par les autres, et le projet Tanguy permet de les rassembler avec une forte synergie.

1.8.3 Qualification du coordinateur du projet

Stéphane Lorin travaille au sein du service CeNTAI/DT (Centre des Nouvelles Technologies d'Analyse de l'Information à la Direction Technique de Thales Land&Joint Systems). Stéphane Lorin est spécialiste en statistiques et Data Mining. Stéphane Lorin a travaillé pendant plus de 10 ans chez IBM où il a participé au développement de la solution DB2 Intelligent Miner, progiciel de Data-mining. Il a été membre de l'ECAM (European Centre for Applied Mathematics) d'IBM Europe et consultant sur de nombreux projets de data-mining et de data-warehouse (auprès des banques principalement). Il a été chef de nombreux projets de tailles plus importantes que le Projet CARATS : (par exemple chez DEXIA, chef de Projet sur une grosse application « Bâle 2 », chez BNP sur un Projet très important autour de la création d'un Datawarehouse pour l'application « Titres » etc..). Il travaille actuellement au CeNTAI de Thales Communications en tant qu'architecte sur les processus d'analyse de l'information. Il est par ailleurs coordinateur du Sous-projet : Analyse et fusion de l'information (SP3), du projet INFOM@GIC du Pôle Cap Digital.

1.9 Stratégie de valorisation et de protection des résultats

Propriété Intellectuelle

Un accord de Consortium sera établi entre les partenaires au démarrage du projet. Il fixera pour chacun les droits de propriété industrielle et intellectuelle, ainsi que les droits d'exploitation des résultats du projet.
De prime abord, il est convenu que chacun des partenaires est propriétaire des modules qu'il aura lui-même développés. En cas de volonté de commercialiser un logiciel faisant appel à plusieurs modules, un accord commercial de « royalties back » sera établi entre les partenaires concernés.
Une façon simple de répondre à cette question est la suivante : Thales étant Porteur et Chef de ce projet auprès de l'ANR, Thales prévoit de faire appliquer entre les membres du Consortium, le même type d'accord que celui que Thales avait fait élaborer dans le cadre de Cap Digital pour le projet INFOM@GIC (un projet qui a compté jusqu'à 29 partenaires). Cet accord ayant été jugé, excellent, d'une part par les partenaires membres du milieu académique (9 au total), d'autre part par ceux provenant du milieu industriel (PME et Grandes Entreprises). Il a été reconnu , du fait de sa couverture, comme un modèle tout à fait adéquat de coopération et de propriété intellectuelle. Ce modèle a été accepté ensuite comme modèle « standard » par et pour le Pôle Cap Digital dans son ensemble.
C'est précisément ce modèle « Pôle de Compétitivité » Cap Digital qui est prévu d'être signé entre les différents partenaires du projet.

2. Programme scientifique et technique

2.1 Partenaire 1 : THALES

2.1.1 Equipement

Dans le cadre de ce projet, un serveur de développement et de partage d'informations, ainsi que des logiciels pour le développement, sont prévus pour un coût total de 5000 Euros.

2.1.2 Personnel

Thales mettra à disposition du projet un coordinateur de projet, un spécialiste des réseaux sémantiques et un ingénieur de recherche. La répartition se fera comme indiqué dans le tableau ci-dessous :

Nom, prénom	Emploi	% du temps dédié au projet	Rôle dans le projet
ROHMER Jean	Conseiller scientifique	50% 18 mois	Chercheur spécialisé en réseau sémantique
LORIN Stéphane	Chef de projet	40% 14,4 mois	Coordination globale du projet Travaux sur les opérateurs d'aide au raisonnement (en particulier opérateurs de clsutering)
Ingénieur	Ingénieur de recherche	70% 25,2 mois	Conception et expérimentations de nouvelles implémentations d'un réseau sémantique

2.1.3 Prestation de service externe

Néant.

2.1.4 Missions

Concernant les missions, le projet se déroulant entre Bordeaux, Grenoble et Paris, deux à trois réunions de projet par an sont prévues : le calcul est basé sur un déplacement de deux jours (300 € par personne), et pour trois personnes. Le montant total est donc de 8100 € (trois personnes, trois réunions, trois ans).

Trois conférences nationales et internationales seront également prévues pour une personne. Avec une moyenne de 1500 Euros par conférence, cela fait donc un total de 4500 Euros.

2.1.5 Dépenses justifiées sur une procédure de facturation interne

Néant.

2.1.6 Autres dépenses de fonctionnement

Néant.

2.2 Partenaire 2 : Xerox

2.2.1 Equipement

Néant.

2.2.2 Personnel

Xerox mettra à disposition un coordinateur de projet, un linguiste pour la création des règles d'analyse nécessaires et un ingénieur pour les réalisations techniques mineures requises par le projet. La répartition se fera comme suit :

Nom, prénom	Emploi	% du temps dédié au projet	Rôle dans le projet
SEGOND Frédérique	Chef de projet	5% 2 mois	Gestion et coordination des développements liés au projet
SANDOR Agnès	Chercheur en linguistique	61% 22 mois	Mise au point des règles nécessaires à l'analyse syntaxique
Profil recherche	Ingénieur de développement	33% 12 mois	Réalisation des tâches techniques en collaboration avec le linguiste

2.2.3 Prestation de service externe

Néant.

2.2.4 Missions

Néant.

2.2.5 Dépenses justifiées sur une procédure de facturation interne

Néant.

2.2.6 Autres dépenses de fonctionnement. *Other expenses.*

Néant.

2.3 Partenaire 3 : INRIA BORDEAUX

2.3.1 Equipement

La demande de financement concerne :

- l'équipement des personnels non permanents (pour la durée du projet) : 5000€ pour deux postes pour les 36 mois du projet ;
- l'équipement du personnel permanent en cours de recrutement (campagne du printemps 2008), correspondant à peu-près à l'amortissement du matériel mis à disposition par l'INRIA : 2500€.

2.3.2 Personnel

Pour le coût global du projet, nous indiquons la charge du personnel permanent. Trois personnes interviendront sur le projet : un professeur et deux maîtres de conférence. En tant qu'enseignants chercheurs, le potentiel de recherche est de six hommes.mois par an. Le personnel permanent intervenant à 33% sur le projet, chaque intervenant sera donc affecté deux mois par an sur le projet.

La demande en personnel consistera principalement en du personnel non permanent pour remplir les engagements de l'INRIA sur le projet :

- une thèse (36 mois) pour travailler sur le fond dès le premier jour,
- un post-doctorant (sur les 12 premiers mois du projet) pour doper le projet le plus tôt possible,
- un ingénieur pour travailler sur l'architecture, le prototypage/développement et le travail de coordination autour de la solution Tulip (à partir T0+12 pour une période de 24 mois une fois que les orientations et les choix architecturaux auront été définis).

Les profils des différents intervenants sont indiqués dans le tableau ci-dessous :

Nom, prénom	Emploi	% du temps dédié au projet	Rôle dans le projet
MELANCON Guy	Professeur des universités	33%* 2 mois	Gestion et coordination du projet. Recherche et encardrement
AUBER David	Maître de conférence	33%* 2 mois	Recherche et encadrement
Maître de conférence	Maître de conférence	33%* 2 mois	Recherche et encadrement

Thésard	Thèse	100% 36 mois	Voir le sujet ci-dessous. Le travail concernera principalement les tâches ST4.2 et ST5.3
Post-doctorant	Post-doctorant	33% 12 mois	Voir le sujet ci-dessous. Le travail concernera principalement les tâches ST4.3 et ST5.3
Ingénieur	Ingénieur	66% 24 mois	Architecture, prototypage, coordination

(*) Pour les chercheurs, les pourcentages sont calculés uniquement sur leur temps de recherche, c'est-à-dire sur 50% de leur temps.

La thèse se déroulera sur la durée du projet, contribuera au travail d'état de l'art et étudiera des questions de fonds sans négliger d'apporter des solutions concrètes au projet.

Le travail prévu pour le stagiaire post-doctoral exige de bonnes connaissances du domaine, que l'on peut difficilement attendre d'un étudiant de master. On s'attend à ce que le post-doctorant puisse avancer et construire rapidement des solutions pour le projet (12 mois). Le travail d'ingénierie desservira l'ensemble du projet, en assurant l'interface entre l'architecture technique Tulip et l'architecture TANGUY. Nous prévoyons lancer ce travail peu avant la fin du stage post-doctoral, et le faire suivre sur les deux dernières années du projet.

Sujet de thèse : Graphes dynamiques : statistiques adaptatives, indices visuels et dessin. Applications à la fouille interactive et la visualisation de réseaux sémantiques.

Peu de travaux se sont encore réellement penchés sur la visualisation interactive des graphes dynamiques, les travaux les plus marquants faisant usage de la connaissance a posteriori de l'évolution des graphes. La plupart des algorithmes de dessin sont alors en mesure de faire le bilan des évolutions et de décider d'un plongement du graphe sur la base d'un consensus calculé à partir de l'ensemble des évolutions qu'a connu le graphe. (Voir la section 1.3.4 de ce document).

Nous proposons de mettre au point des algorithmes de plongement dynamiques traitant le cas de graphes dont les évolutions ne sont pas connues. Nous souhaitons considérer les évolutions du graphe au niveau de sa topologie (ajout/suppression de sommets et d'arêtes), mais aussi au niveau des attributs de ses éléments (valeurs associées, appartenance à une catégorie, etc.).

Au-delà des seules procédures de dessin, il faudra concevoir les indices visuels dynamiques facilitant la lecture de l'évolution du réseau. Le travail des géographes sur la sémiologie et les cartographies animées sont un premier point de départ[46].

Une stratégie souvent utilisée pour visualiser un graphe de taille conséquente est d'en filtrer les éléments. Une statistique calculée sur les sommets et/ou les arêtes permet alors de filtrer les éléments pour ne garder que ceux qui en forme le « squelette ». Des travaux de l'équipe sur le repérage de changement structuraux et d'éléments saillants (dans des graphes décrivant la structure de séquences vidéos) sont aussi de nature à pouvoir guider les procédures de dessin et de visualisation[47].

Des statistiques pouvant tenir compte de l'évolution du graphe et se soumettant aux exigences de l'interactivité (tant du point de vue de leur pertinence que de considérations algorithmiques) sont à trouver. Nombre de statistiques classiques peuvent vraisemblablement être adaptées ou généralisées au cas des graphes dynamiques. Le domaine de l'analyse des données temporelles doit être aussi exploré. Partant de ces statistiques, on devrait ensuite pouvoir effectuer les calculs de voisinage rapproché, de plongement, et produire des artifices et des indices visuels pertinents, dont nous pourrions envisager de démontrer l'efficacité[48].

[46] Voir par exemple, les travaux de l'UMR Espace, http://www.umrespace.org/ActGCart.htm. L'équipe GRAVITÉ collabore activement avec des membres de cette UMR au travers du projet ANR Massses de données SPANGEO, http://s4.parisgeo.cnrs.fr/spangeo/spangeo.htm.
Schlienger, C., P. Dragicevic, et al. (2006). Les transitions visuelles différenciées : principes et applications 18ème Conférence Francophone sur l'Interaction Homme-Machine (IHM 2006) ACM Press.

[47] Chevalier F., Domenger J.P., Benois-Pineau J., Delest M.. Retrieval of objects in video by similarity based on graph matching,, in "Pattern Recognition Letter", vol. 28, no 8, jun 2007, p. 939–949.

[48] Andrienko, N. and G. Andrienko (2005). Exploratory Analysis of Spatial and Temporal Data - A Systematic Approach, Springer.

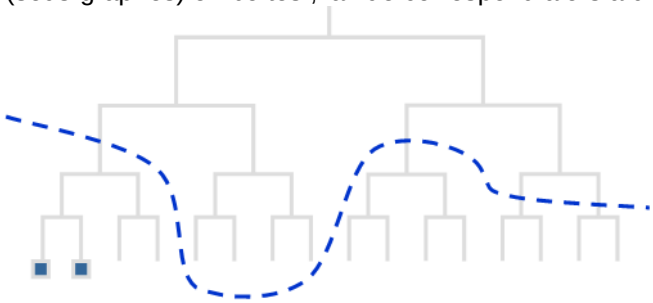
Sujet de stage post-doctoral : Représentations multi-niveaux, approches focus+context et interfaces zoomables pour la visualisation de graphes dynamiques. Applications à la fouille interactive et la visualisation de réseaux sémantiques.

Le sujet de ce stage exige une bonne connaissance des résultats existants sur les techniques de visualisation « focus+contexte » et les interfaces zoomables, qui forment tout un pan de la littérature en visualisation d'information. C'est en cela qu'il est pertinent de rechercher un candidat de niveau post-doctorat. Nous attendons aussi du stagiaire qu'il puisse fournir un effort de prototypage des mécanismes de vues locales/globales s'intégrant à la plate-forme Tulip, et de manière plus générale à toute l'architecture TANGUY.

Nous précisions à la section 1.3.4 les liens entre représentations multi-niveaux et navigation « focus+contexte » permettant à l'utilisateur de passer aisément d'une vue locale à une vue globale de l'espace exploré. Cette représentation multi-niveaux repose elle-même sur une hiérarchisation du graphe sous-jacent à l'espace d'information.

L'objectif du stage est de revoir ces techniques de navigation afin de les faire passer à la fouille interactive visuelle de graphes dynamiques. Plusieurs questions se posent :

- comment tenir compte des évolutions du graphe dans la hiérarchisation ? y a-t-il la possibilité de retarder le calcul de la hiérarchie de l'espace « visible » au moment voulu ? y a-t-il possibilité de faire reposer cette hiérarchisation sur des statistiques tenant compte de la dynamique ?
- rappelons ici le principe des vues « focus+contexte » : la hiérarchisation de l'espace est encodé dans un arbre de clusters (sous-graphes) emboîtés ; la vue correspond alors à une coupe de l'arbre.



- comment tenir compte de la dynamique du graphe dans le calcul de la coupe ?

Le premier point intersecte nettement les problèmes de clustering de graphes étendus aux cas des graphes dynamiques. L'idée de retarder le calcul des clusters au moment nécessaire, et donc de ne calculer que ceux qu'il est utile de calculer, rejoint le calcul de voisinage rapproché. Le travail de hiérarchisation, et le calcul de la coupe, devrait pouvoir s'étendre au cas où les clusters s'intersecte, et donc lorsque la hiérarchisation donne lieu à u graphe orienté sans circuit plutôt qu'un arbre.

Finalement, le travail de conception et sa réalisation devra prendre en compte le méta-modèle développé dans le cadre du projet.

Travail d'ingénierie : Environnement de fouille interactive visuelle de réseaux sémantiques. Intégration des composants développés par les partenaires.

Nous comptons recruter un ingénieur travaillant auprès de l'équipe GRAVITÉ et en relation avec les partenaires. La mission de cette personne comportera deux volets :

- venir en appui à l'équipe GRAVITÉ et contribuer à l'effort de prototypage et de développement
- veiller à une bonne implémentation du méta-modèle afin d'anticiper sur l'intégration des composants ;
- s'assurer de la bonne intégration des composants conçus et réalisés par les partenaires, veiller à ce qu'ils se conforment à la fois à l'architecture Tulip (voir la section 1.5.1) et à l'architecture générale de TANGUY.

2.3.3 Prestation de service externe

Néant.

2.3.4 Missions

Concernant les missions, le projet se déroulant entre Bordeaux, Grenoble et Paris, deux à trois réunions de projet par an sont prévues : le calcul est basé sur un déplacement de deux jours (300 € par personne), et pour trois personnes. Le montant total est donc de 8100 € (trois personnes, trois réunions, trois ans).

De plus, des participations à deux conférences nationales et deux conférences internationales sont prévues pour les deux dernières années du projet pour deux personnes : le calcul est basé sur un coût de 800 € pour une conférence nationale et de 2000 € pour une conférence internationale (incluant le séjour et les frais d'inscription). Le montant total est donc de 3200€ (national) + 8000 € (international) – deux personnes, deux conférences nationales et deux conférences internationales.

Soit un total de 19 300 € pour l'ensemble des frais de missions INRIA sur la durée du projet.

2.3.5 Dépenses justifiées sur une procédure de facturation interne

Néant.

2.3.6 Autres dépenses de fonctionnement. *Other expenses.*

Néant.

2.4 Partenaire 4 : FIDAL

2.4.1 Equipement

Néant.

2.4.2 Personnel

Fidal mettra à disposition un expert métier spécialisé dans le traitement des affaires juridiques sur l'analyse et un stagiaire. La répartition se fera comme suit :

Nom, prénom	Emploi	% du temps dédié au projet	Rôle dans le projet
GAVANON Isabelle	Associée	8% 3 mois	Expertise métier Expertise méthodologique
Stagiaire	Stage	41% 15 mois	Coordination de la tâche T6 Etude de cas Rapports d'étude

2.4.3 Prestation de service externe

Néant.

2.4.4 Missions

Néant.

2.4.5 Dépenses justifiées sur une procédure de facturation interne

Néant.

2.4.6 Autres dépenses de fonctionnement. *Other expenses.*

Néant.

Annexe : Description des partenaires

Partenaire 1 : Thales

Thales est un leader mondial de l'électronique et des systèmes. Partout dans le monde, le Groupe sert les marchés de l'Aéronautique et de l'Espace, de la Défense et de la Sécurité, appuyé par une offre globale de technologies et de services de pointe. Le Groupe optimise le développement parallèle des activités civiles et militaires, et partage une base commune de technologies au service d'un seul objectif : la sécurité des personnes, des biens et des Etats. Avec plus de 22 000 ingénieurs et chercheurs de haut niveau, Thales constitue une capacité unique en Europe pour créer et déployer des systèmes d'information critiques éprouvés. Le Groupe assoit sa croissance sur une stratégie multi-domestique sans équivalent, basée sur des partenariats privilégiés avec les clients nationaux et les acteurs clés des marchés concernés, en s'appuyant sur son expertise globale pour soutenir les technologies et le développement industriel au niveau local. Fort de 68 000 personnes dans cinquante pays, Thales a réalisé en 2007 un chiffre d'affaires de 12,3 milliards d'euros.

Thales CeNTAI (Centre d'Expertise des Nouvelles Technologies d'Analyse de l'Information) est une structure propre à Thales Land&Joint Systems, Division de Thales Groupe, forte de 15000 collaborateurs (ex Thales Communications) . Cette structure de R&D est dédiée à l'Etude des méthodologies et algorithmes relevant de l'Analyse et du Management de l'Information Multimodale (Analyse des Données Textuelles, Analyse des données Numériques, Représentation des Connaissances).

CV des participants

Jean Rohmer

Docteur Ingénieur et Docteur-ès-Sciences de l'Institut National Polytechnique de Grenoble, Jean Rohmer a été chercheur à l'Université de Grenoble, à l'Inria et à Bull dans le domaine des bases de données. Il a ensuite créé et dirigé le CEDIAG, centre de recherche et Business Unit en Intelligence Artificielle du Groupe Bull, qui a développé de grands systèmes industriels à base de règles et de programmation par contraintes. Dans les années 1990, il a créé la société Idéliance, qui a conçu et mis sur le marché l'outil éponyme de gestion de réseaux sémantiques. C'est dans cette voie qu'il poursuit maintenant son activité comme Conseiller Scientifique au sein du Centre des technologies d'Analyse de l'Information de Thales Communications.. Jean Rohmer a été, avec l'invention de la "Méthode d'Alexandre" un des fondateurs de la discipline des base de données déductives (datalog), aujourd'hui à la base des moteurs du Web Sémantique. Il a publié une cinquantaine d'articles scientifiques. Il a récemment présidé le groupe de travail de l'OTAN sur la fusion d'Informations.

Stéphane Lorin

Spécialiste en statistiques et Data-mining, Stéphane Lorin a travaillé pendant plus de 10 ans chez IBM où il a participé au développement de la solution DB2 Intelligent Miner, progiciel de Data-mining. Il a été consultant et chef de projet sur de nombreux projets de Data-warehouse et de Data-mining (auprès des banques principalement). Il travaille actuellement au CeNTAI de Thales Communications en tant qu'architecte et chef de projet sur les processus d'analyse de l'information.

Partenaire 2 : Xerox

Les documents jouent un rôle très important dans les entreprises. La façon dont ils sont crées, utilisés peut avoir un impact très important sur la productivité. Le travail du Centre de Recherche Européen de Xerox (XRCE) se concentre sur l'amélioration de cette productivité grâce au développement de nombreuses technologies visant à rendre ces documents intelligents. XRCE s'attache à travailler à la fois avec des organisations de recherche, mais aussi avec des groupes de productions et leurs clients externes pour s'assurer que ce qui est développé dans son centre n'est pas seulement innovant mais aussi bien en phase avec des besoins réels de l'industrie et constitueront des technologies de rupture permettant à la fois de développer les marchés existants mais aussi d'en créer de nouveaux. Le Centre a pour objectif de concevoir des technologies centrées autour des documents et qui correspondent à un besoin à une utilisation réelle de la part des utilisateurs.

Dans le cadre de ce projet les compétences et technologies apportées par le Centre Européen de Recherche de Xerox (XRCE) seront essentiellement axées sur l'analyse textuelle de documents grâce à des segmenteurs de textes, des analyseurs morphologiques, des extracteurs de dépendances syntaxiques, et des moteurs d'analyse sémantique à base de règles. Toutes ces technologies font parties des logiciels cœurs développés au Centre depuis de nombreuses années. Ces technologies ont déjà été employées dans le cadre de nombreux projets de recherche et même d'applications professionnelles.

CV des participants

Frédérique Segond

Frédérique Segond dirige depuis 2003, le groupe de Recherche « Parsing and Semantics » au sein du laboratoire Européen de Xerox (Xerox Research Centre Europe (XRCE)) de Grenoble. Frédérique Segond a rejoint Xerox en 1993, où elle a été responsable du développement et de l'implémentation d'une grammaire du français, puis chef de projet pour la désambiguïsation sémantique. Avant de rejoindre Xerox, Frédérique Segond a effectué une thèse en Mathématiques Appliquées à l'Ecole des Hautes Etudes en Sciences Sociales de Paris, conjointement avec un sejour de 4 ans au Centre de Recherche d'IBM-France (ECAM), puis à IBMResearch Center Yorktown pour travailler sur les liens entre syntaxe et sémantique. Frédérique Segond, Maître de Conférence associée, est co-auteur de cinq livres et de plus de 50 articles scientifiques. Elle est membre de nombreux comités de programme ainsi que de la Commission technique du Pole de Compétitivité Cap Digital. Elle fournit également son expertise scientifique auprès de la Commission Européenne.

Agnès Sandor

Ágnes Sándor est ingénieur de recherche dans groupe de traitement des langues naturelles du laboratoire Européen de Xerox (Xerox Research Centre Europe (XRCE)) de Grenoble depuis 1996. Ses activités précédentes comprennent la création des analyseurs morphologiques et des désambigüisateurs de parties du discours dans plusieures langues. Depuis quelques années elle se concentre sur l'extraction d'informations et l'analyse du discours. Elle a publié plusieurs articles scientifiques sur ces sujets et participe dans des projets européens.

Partenaire 3 : INRIA Bordeaux

Le centre de recherche INRIA Bordeaux - Sud Ouest créé récemment est basée sur les équipes-projets et services créés à Bordeaux et à Pau dans le cadre de l'unité de recherche Futurs, avec le soutien fort du Conseil Régional d'Aquitaine.
Les équipes de recherche y ont été construites en s'appuyant sur des partenariats forts avec les universités et grandes écoles des sites concernés et en collaboration avec leurs laboratoires réputés.
Grâce au dynamisme de ces collaborations, à l'apport de personnels ayant effectué une mobilité depuis d'autres sites de l'INRIA et en s'appuyant sur une politique de recrutement de chercheurs et de personnels de soutien à la recherche de haute qualité, le centre de recherche emploie directement, début 2008, plus de 100 personnes. Il compte 8 équipes-projets de recherche, et 6 équipes dont l'évaluation est en cours et qui devraient devenir prochainement équipes-projets. Le centre rassemble globalement plus de 250 personnes.
Dans le cadre de son plan stratégique, l'INRIA entend définir son développement à Bordeaux - Sud Ouest autour de thématique prioritaire : systèmes complexes ; simulation, visualisation et interaction ; modélisation du vivant. Des partenaires industriels et académiques seront également étroitement associés. On peut citer par exemple Total, Safran/Turbomeca, Thales, Rhodia, le CEA, le CNRS, France Telecom, EDF, Airbus, SNCF, etc. directement, mais aussi au travers du pôle Aerospace Valley.

CV des participants

Guy Mélançon

Guy Melançon est professeur d'informatique à l'Université Bordeaux I, responsable de l'équipe « Modèles et algorithmes pour la bio-inforamtique et la visualisation » (MABioVis) du LaBRI et responsable de l'équipe INRIA GRAVITÉ (Graph Visualization and Interactive exploration). Diplômé de l'Université du Québec à Montréal où il a obtenu un PhD en mathématique combinatoire (1991), il a d'abord été Maître de Conférences à Bordeaux de 1993 à 1998 avant de rejoindre le CWI à Amsterdam en tant que chercheur pendant près de trois ans. De retour en France en septembre 2000, il a intégré le LIRMM (CNRS UMR 5800). Délégué au CNRS en 2005-2006, puis détaché à l'INRIA en 2006-2007, il est de retour à Bordeaux depuis septembre 2008.

Spécialiste de la visualisation de graphes, il est auteur d'un survey cité plus de 400 fois. Ses travaux portent sur les méthodes de visualisation et de décomposition des graphes mettant à profit la combinatoire, les structures petites mondes et invariant d'échelle observées dans les cas réels.

David Auber

David Auber est Maître de conférences à l'Université Bordeaux I depuis 2003, après avoir bénéficié de la bourse d'excellence Peter Wall de l'Université de Colombie-Britannique que lui ont mérité ses travaux de thèse. Il est membre de l'équipe MABioVis du LaBRI et de l'équipe INRIA GRAVITÉ dont il est un pilier. Il est le concepteur et l'architecte de la plate-forme de visualisation de graphes Tulip, qui fait autorité dans le domaine de la visualisation de graphes. Spécialiste du dessin de graphes et de la visualisation interactive, il mène une collaboration très fructueuse avec nombre de collaborateurs européen et nord-américains.

MCF en cours de recrutement

L'équipe MABioVis envisage de recruter un maître de conférences dont le profil conjuguera expertise en visualisation de graphes et extraction et gestion de la connaissance. Des contacts ont déjà été pris et le recrutement devrait se faire sur un bon vivier de candidatures ; le concours se déroulera à la fin du printemps 2008.

Partenaire 4 : FIDAL

Avec 1175 avocats en France et des partenaires dans 150 pays, FIDAL est le premier cabinet d'avocats d'affaires en France par la taille et le chiffre d'affaires* et le seul cabinet français à figurer au top 100 mondial.
*Source : radiographie des cabinets d'avocats d'affaires en France, Juristes Associés.

Le cabinet offre à ses clients une triple compétence :

- Nationale, avec une forte implantation à Paris (près de 334 avocats) et en régions.
- Européenne avec l'appui de son bureau de Bruxelles spécialisé dans les problématiques communautaires.
- Internationale, en accompagnant nos clients dans leurs opérations transfrontalières avec des équipes dédiées.

Les avocats de FIDAL conseillent 40 000 entreprises de toutes tailles et leurs dirigeants, des groupes internationaux aux entreprises du middle-market, avec la même exigence de qualité et de connaissance du marché de leurs clients.

Huit départements spécialisés couvrent les grands domaines du droit des affaires : Droit fiscal, Droit des sociétés, Droit et gestion sociale, Droit de la concurrence et de la distribution, Droit de la propriété intellectuelle et des technologies de l'information, Droit du patrimoine, règlement des contentieux et Droit public.

Pour répondre aux besoins de marchés spécifiques, le cabinet a parallèlement développé des pôles d'expertise (fusions-acquisitions, Droit boursier, capital investissement, Droit bancaire, Droit des assurances, Droit immobilier, associations et économie sociale, agroalimentaire, santé, Droit du sport, Droit de l'environnement, retraites et prévoyance, sécurité et prévention, …)

CV des participants

Isabelle Gavanon

Avec près de 20 années d'expérience dans le droit des technologies de l'information et de la communication, Isabelle Gavanon a rejoint FIDAL le 1er décembre 2006 en qualité de directeur associé.

Associé depuis 2003 chez Ginestié Magellan Paley-Vincent où elle a exercé pendant près de 8 ans, Isabelle Gavanon dispose d'une expertise pointue notamment dans les techniques contractuelles appliquées aux projets informatiques et de télécommunications, ainsi que pour la gestion des problématiques relatives à Internet et au statut de l'information (bases de données, preuve électronique, droit d'auteur). Une activité contentieuse complète cette pratique de conseil. Cette dernière activité la conduit généralement à devoir analyser, directement ou pas l'intermédiaire de ses clients, un grand volume de données souvent stockées sur support informatique. Ce travail d'analyse pourrait être optimisé grâce à de nouveaux outils de traitement automatisé de l'information.

Une première expérience en tant que juriste chez Apple Computer Inc. (1990-92) a imprimé d'emblée la dominante professionnelle qu'Isabelle Gavanon a ensuite approfondi au sein des cabinets Gide, Loyrette, Nouel (1993-95) et Jeantet & Associés (1996-99).

Isabelle Gavanon est également chargé d'enseignement en droit des télécommunications dans le DESS « Droit du multimédia et de l'informatique » de Paris II, anime des conférences et est auteur de nombreux articles relatifs au droit des technologies de l'information.