

➔ **Projet OUTILEX**
Rapport d'étude final
Octobre 2006

date / references

Contexte et Objectifs du document



Ce document a été élaboré dans le cadre du projet Outilex, il présente le rapport d'étude final de Thales Communications

date / references

Contributions attendues de Thales Communications



ales

Pour rappel, les annexes technique et financière de Thales Communications ont été revues en mars 2005, suite au transfert du contrat Outilex par la société Kalima vers la société Thales Communications.

Dans le cadre de ces révisions, les contributions de Thales Communications au projet Outilex ont été définies comme suit, à compter de la date du transfert :

- ✓ Développement de composants d'extraction d'information (grammaires locales et ressources associées)
- ✓ Développement d'un démonstrateur métier dans le domaine de l'analyse des incidents

Les composants d'extraction d'information (grammaires locales et ressources associées) ont été développés pour le traitement des données du démonstrateur.

Pour des questions de fourniture des données par le client, le démonstrateur métier initialement prévu portant sur l'analyse des incidents dans le domaine de l'automobile n'a pas pu être développé dans le cadre du projet.



On s'est donc orienté, en cours de projet et avec l'accord du consortium, vers le développement d'un démonstrateur métier portant sur l'extraction d'informations dans des textes de type dépêches et rapports à des fins d'alimentation d'une base de connaissances, laquelle est exploitée par des outils d'analyse de type réseaux sémantiques et data mining.

Le domaine métier retenu est le domaine de la Sécurité Nationale.

Les travaux réalisés ont principalement visé à tester et valider « l'utilisabilité » de la plateforme Outilex dans un contexte industriel, en vue de répondre à des besoins métier non triviaux



Les besoins



Les besoins en matière d'extraction d'information ont été spécifiés par les utilisateurs finaux

Quatre grands types de besoins ont été définis :

- ✓ Extraction d'entités nommées (personnes, organisations, lieux, dates et heures)
- ✓ Extraction de faits
- ✓ Extraction de marqueurs d'ambiance
- ✓ Détection de relations élémentaires entre les entités extraites

En vue d'être stockées dans la base de connaissances puis exploitées par les outils d'analyse et de visualisation, les données extraites ont été normalisées graphiquement, syntaxiquement et sémantiquement, via l'utilisation de grammaires de normalisation, de ressources lexicales et de processus de transformation des informations

date / référence

Les informations contenues dans ce document sont la propriété exclusive du Groupe Thales. Elles ne doivent pas être divulguées sans l'accord écrit de Thales



Le corpus utilisé dans le cadre du démonstrateur est composé d'environ un millier de rapports de divers organismes de renseignement portant sur la thématique des campagnes de fauchage OGM.

Chaque rapport est composé de données structurées – objet , date , auteur, source - et d'un texte.

Pour des questions de confidentialité, les données structurées hors l'objet des rapports, ont été supprimées dans la version finale du démonstrateur.



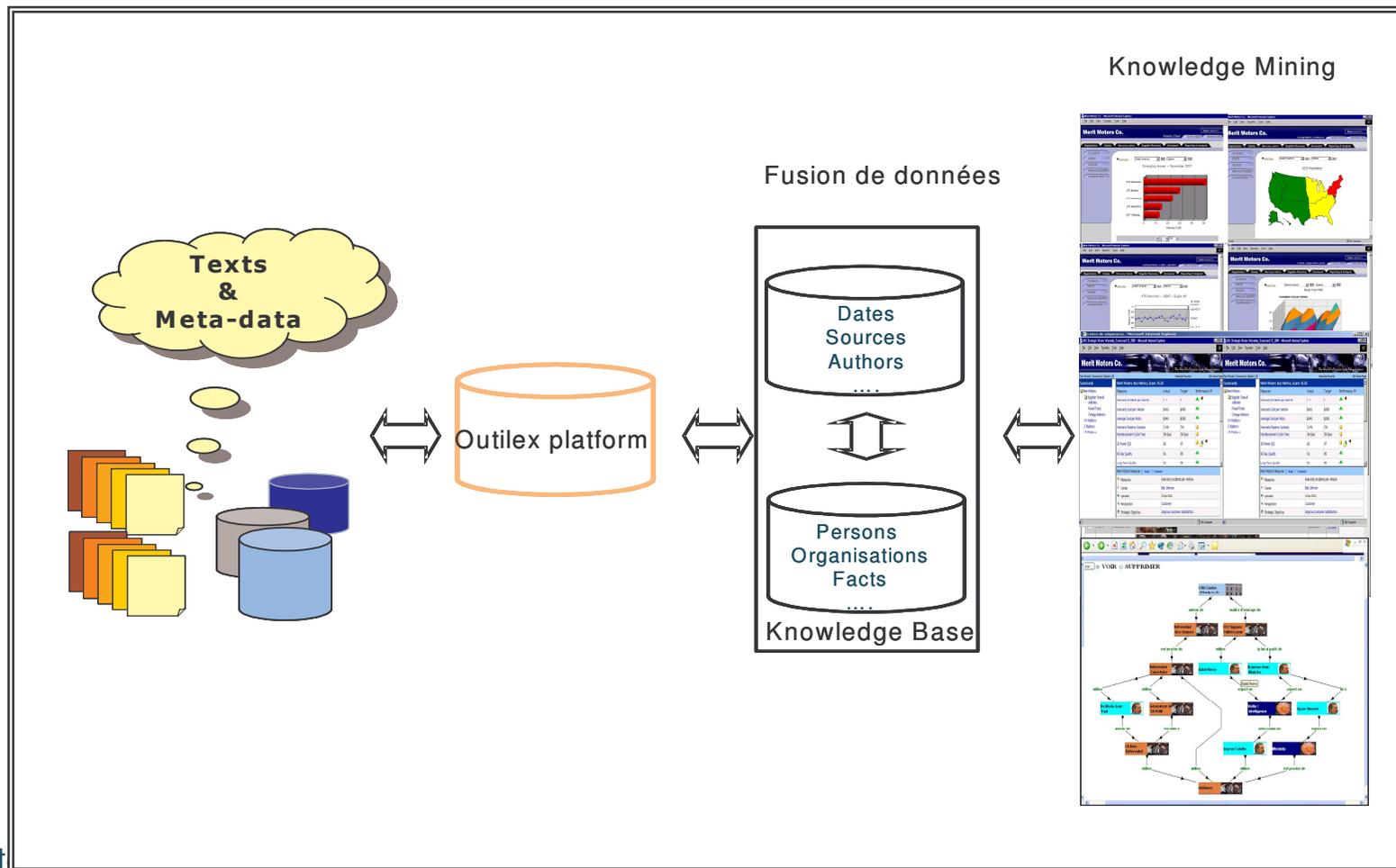
Aperçu du démonstrateur implémenté

date / references

Schéma du démonstrateur



Les données d'entrée du démonstrateur sont constituées de données non structurées et de données structurées - données signalétiques associées aux textes (date, source, auteur, ...) – Ces données sont couplées aux données structurées issues des processus d'extraction d'information et exploitées par les outils d'analyse.



date / references



Etape 1 : Extraction d'information

date / references

Exemple Entités Nommées/Personnes



Acteurs/Personne

SIMON FRANCOIS

Lieux

MENVILLE
A LA COMPAGNIE DE GENDARMERIE DE TOULOUSE MIRAIL

Relations

Personne	Fonction
SIMON FRANCOIS	CONSEILLER MUNICIPAL DE TOULOUSE

Text Content

Primo : DANS LE CADRE DE L'ENQUETE EN COURS SUITE AU FAUCHAGE D'UNE PARCELLE DE MAIS TRANSGENIQUE SUR LA COMMUNE DE MENVILLE (31-ZGN) LE 25/07/2004, UNE DELEGATION DE SEPT PERSONNES S'EST PRESENTEE CE JOUR 15/10/2004 A 11H00 A LA COMPAGNIE DE GENDARMERIE DE TOULOUSE MIRAIL.

Secundo : CES PERSONNES AVEC A LEUR TETE M. FRANCOIS SIMON, CONSEILLER MUNICIPAL DE TOULOUSE ONT REMIS AU COMMANDANT DE COMPAGNIE ET A UN ENQUETEUR, UNE NOUVELLE LISTE DE 70 NOMS DE PERSONNES RECONNAISSANT AVOIR PARTICIPE A L'ACTION DE

date / references

Les informations contenues dans ce document sont la propriété exclusive du Groupe Thales. Elles ne doivent pas être divulguées sans l'accord écrit de Thales

Exemple Entités Nommées/Organisation



Google G Envoyer Mes favoris PageRank 0 bloquée(s) Orthographe Paramètres

Acteurs/Organisation

[REDACTED]

CIVAM
CENTRE D'INITIATIVE POUR FAVORISER L'AGRICULTURE ET LE MILIEU RURAL
ASSOCIATION TERRES D'ARREE-BRO AN ARE ET PRODUITS DU CAP SIZUN
ASSOCIATION GREENPEACE
ASSOCIATION JARDINS DU MONDE
ASSOCIATION PEUPLES ET FORETS PRIMAIRES
ASSOCIATION SORTIR DU NUCLEAIRE

Lieux

LIEU-DIT TREMATOUARN
MAHALON
EXPLOITATION AGRICOLE BIO
DANS LES PAYS DU SUD

Text Content

SERVIS PAR UN AUTRE MOYEN TO/PREFET FINISTERE INFO/ [REDACTED]
OCTOBRE 2004, A COMPTER DE 10 HEURES, AU LIEU DIT TREMATOUARN A MAHALON (29-ZGM), LE CIVAM 29 (CENTRE D INITIATIVE POUR FAVORISER L'AGRICULTURE ET LE MILIEU RURAL), LES ASSOCIATIONS PAGE 2 RFFDCC 1602 NON PROTEGE TERRES D'ARREE-BRO AN ARE ET PRODUITS DU CAP SIZUN ORGANISENT LA JOURNEE CAMPAGNES VIVANTES. LA MANIFESTATION SE PRODUIRA DANS UNE EXPLOITATION AGRICOLE BIO OU IL SE DEROUlera DEUX CONFERENCES/DEBATS: - A 10 HEURES 30 : SUR LE LE THEME L'IMPACT DES OGM SUR LA PRESERVATION DE LA BIODIVERSITE DANS LES PAYS DU SUD - A 14 HEURES 00 : SUR LA QUESTION DES RISQUES ENVIRONNEMENTAUX, SOCIAUX ET ECONOMIQUES ENGENDREES PAR UNE PROLIFERATION NON MAITRISEE DES OGM. AU COURS DE CETTE JOURNEE IL SERA

date / references

Les informations contenues dans ce document sont la propriété exclusive du Groupe Thales. Elles ne doivent pas être divulguées sans l'accord écrit de Thales

Exemple Entités Nommées/Dates et Heures



Google G

Envoyer

Mes Favoris

PageRank

0 bloquée(s)

Orthographe

Paramètres

Dates

28/01/04
28/01/05
01/06/65
15/05/70
30/01/05

Heures

13:45
12:15
13:30
14:00
20:00

Text Content

LE 28/01/04 A 13H45

Primo : VERS 12 HEURES 15, DEUX MILITANTS GREENPEACE SONT MONTES A BORD DU "GOLDEN LION" AFIN D'Y INSTALLER UNE BANDEROLE. LES FORCES DE L'ORDRE ONT INTERPELLE CES INDIVIDUS QUI ONT FAIT L'OBJET D'UN CONTROLE D'IDENTITE AU COMMISSARIAT DE POLICE. ILS ONT RAPIDEMENT ETE REMIS EN LIBERTE. VERS 13 HEURES 30, DEBUT DE DISLOCATION DE LA MANIFESTATION. JOSE BOVE ET ARNAUD APOTEKER, RESPONSABLE CAMPAGNE EUROPEENNE ANTI-OGM FRANCE SONT A BORD DE L'ESPERANZA. LE DECHARGEMENT DE SOJA DU GOLDEN LION N'A TOUJOURS PAS DEBUTE. -

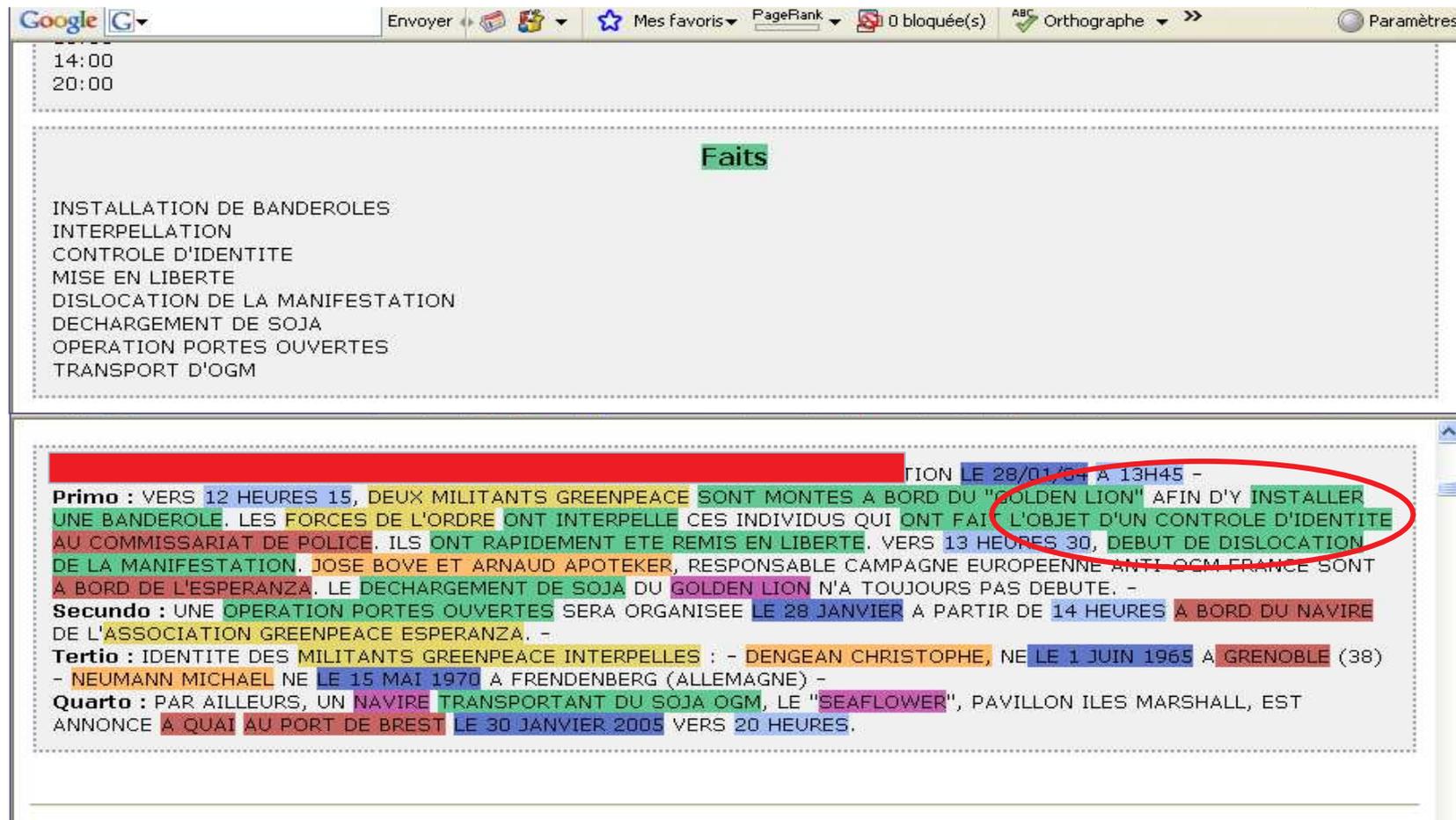
Secundo : UNE OPERATION PORTES OUVERTES SERA ORGANISEE LE 28 JANVIER A PARTIR DE 14 HEURES A BORD DU NAVIRE DE L'ASSOCIATION GREENPEACE ESPERANZA. -

Tertio : IDENTITE DES MILITANTS GREENPEACE INTERPELLES : - DENGAN CHRISTOPHE, NE LE 1 JUIN 1965 A GRENOBLE (38) - NEUMANN MICHAEL NE LE 10 MAI 1970 A FRENENBERG (ALLEMAGNE) -

Quarto : PAR ALLEURS, UN NAVIRE TRANSPORTANT DU SOJA OGM, LE "SEAFLOWER", PAVILLON ILES MARSHALL, EST ANNONCE A QUAI AU PORT DE BREST LE 30 JANVIER 2005 VERS 20 HEURES.

date / references

Les informations contenues dans ce document sont la propriété exclusive du Groupe Thales. Elles ne doivent pas être divulguées sans l'accord écrit de Thales



14:00
20:00

Faits

- INSTALLATION DE BANDEROLES
- INTERPELLATION
- CONTROLE D'IDENTITE
- MISE EN LIBERTE
- DISLOCATION DE LA MANIFESTATION
- DECHARGEMENT DE SOJA
- OPERATION PORTES OUVERTES
- TRANSPORT D'OGM

...TION LE 28/01/04 A 13H45 -

Primo : VERS 12 HEURES 15, DEUX MILITANTS GREENPEACE SONT MONTES A BORD DU "GOLDEN LION" AFIN D'Y INSTALLER UNE BANDEROLE. LES FORCES DE L'ORDRE ONT INTERPELLE CES INDIVIDUS QUI ONT FAIT L'OBJET D'UN CONTROLE D'IDENTITE AU COMMISSARIAT DE POLICE. ILS ONT RAPIDEMENT ETE REMIS EN LIBERTE. VERS 13 HEURES 30, DEBUT DE DISLOCATION DE LA MANIFESTATION. JOSE BOVE ET ARNAUD APOTEKER, RESPONSABLE CAMPAGNE EUROPEENNE ANTI OGM FRANCE SONT A BORD DE L'ESPERANZA. LE DECHARGEMENT DE SOJA DU GOLDEN LION N'A TOUJOURS PAS DEBUTE. -

Secundo : UNE OPERATION PORTES OUVERTES SERA ORGANISEE LE 28 JANVIER A PARTIR DE 14 HEURES A BORD DU NAVIRE DE L'ASSOCIATION GREENPEACE ESPERANZA. -

Tertio : IDENTITE DES MILITANTS GREENPEACE INTERPELLES : - DENGAN CHRISTOPHE, NE LE 1 JUIN 1965 A GRENOBLE (38) - NEUMANN MICHAEL NE LE 15 MAI 1970 A FRENDBERG (ALLEMAGNE) -

Quarto : PAR AILLEURS, UN NAVIRE TRANSPORTANT DU SOJA OGM, LE "SEAFLOWER", PAVILLON ILES MARSHALL, EST ANNONCE A QUAI AU PORT DE BREST LE 30 JANVIER 2005 VERS 20 HEURES.

date / references

Les informations contenues dans ce document sont la propriété exclusive du Groupe Thales. Elles ne doivent pas être divulguées sans l'accord écrit de Thales

Exemple Marqueurs d'ambiance



La plupart des rapports analysés portent sur les démonstrations collectives, publiques et organisées des acteurs anti-ogm. Les marqueurs d'ambiance détectés au moyen de grammaires locales dédiées permettent de qualifier l'atmosphère matérielle et morale de ces démonstrations

ive du Groupe Thales. Elles ne doivent pas être divulguées sans l'accord écrit de Thales

NAVIRE GOLDEN LION

Lieux

PORT DE LORIENT

Ambiance

AMBIANCE CALME
MANIFESTANTS LEGEREMENT AGITES

Text Content

Primo : LE 28 JANVIER A 6 HEURES LE NAVIRE DE COMMERCE GOLDEN LION, CHARGE DE 32000 TONNES DE SOJA TRANSGENIQUE, A ACCOSTE AU PORT DE LORIENT SANS INCIDENT MALGRE LA PRESENCE D'EMBARCATIONS DE L'ASSOCIATION GREENPEACE ET D'UNE CINQUANTAINE DE MILITANTS SUR LES QUAIS. HORMIS LE CHAVIRAGE D'UN KAYAK N'AYANT PAS FAIT DE VICTIME, AUCUN INCIDENT N'A ETE SIGNALÉ MALGRÉ LA LEGERE AGITATION DES MANIFESTANTS LORS DES MANOEUVRES D'ACCOSTAGE.

Secundo : LES DOCKERS DU PORT DE LORIENT AURAIENT FAIT PART DE LEUR REFUS DE DECHARGER LE GOLDEN LION EN RAISON DE LA "FORTE PRESENCE POLICIERE" (UNE CRS).

Ambiance

AMBIANCE CALME (PREVISION)

Text Content

LE 5 FEVRIER 2005, LE MOUVEMENT DES FAUCHEURS D'OGM ENVISAGE D'EFFECTUER UN CONTROLE DANS UNE GRANDE SURFACE D'ANGERS (ZPN).

Primo : LE 5 FEVRIER 2005 A 14H30, UN GROUPE DE QUINZE PERSONNES DEVRAIT SE RENDRE DANS LE MAGASIN CARREFOUR ST SERGE A ANGERS. DES CONTROLES SERONT EFFECTUES SUR DES PRODUITS SUSCEPTIBLES DE CONTENIR DES OGM.

Secundo : CETTE ACTION DEVRAIT SE DEROULER DANS LE CALME.

date / references

Exemple Détection de Relations



Certaines relations élémentaires ont été identifiées via les grammaires locales. Il s'agit principalement de relations telles que personne/fonction, personne/date de naissance, personne /lieu de naissance, personne/adresse, etc

Personne	Date Naissance	Lieu Naissance
SERBIELLE PIERRE	30/10/55	MAULEON
OXARANGO XAVIER	28/07/70	BAYONNE
OXARANGO MARTIN	13/04/28	MACAYE
DITHURBIDE EPOUSE OXARANGO JEANNE	15/02/26	AYHERRE
OXARANGO GRACIE	05/10/62	HASPARREN
ARRAMBIDE ROBERT	27/06/42	HENDAYE
LICHTAS EPOUSE ARRAMBIDE DINA	25/09/47	KIBOUTZ-DALYA
ARRAMBIDE ADAM	04/05/84	SAINT-JEAN-DE-LUZ
ARRICAU CASSIAU DIDIER	31/03/65	SALIES-DE-BEARN
LAVIE MARYSE	17/08/61	SALIES-DE-BEARN
ANTZA MIKEL	07/06/61	SAINT-SEBASTIEN
AMBOTO	25/04/61	ESCORIAGA
URDAMPILLETA ITUBURU EPOUSE ALCANTARILLA MOZOTA	25/02/62	SAINT-SEBASTIEN
MARIA LOURDES		SAINT-SEBASTIEN
ARANO URBIOLA JOSE RAMON	10/05/59	VIALAVA
INCABY EPOUSE ARANO URBIOLA MYRIAM	23/04/60	BAYONNE
ALCANTARILLA MOZOTA PEDRO MARIA	28/06/60	SAINT-SEBASTIEN
Personne	Fonction	
IDOIA ZENARRIZABETTIA	VICE-LEHENDAKART	

date / references

Les informations contenues dans ce document sont la propriété exclusive du Groupe Thales. Elles ne doivent pas être divulguées sans l'accord écrit de Thales



Etape 2 : Gestion des Connaissances

date / references



Le processus d'extraction permet d'alimenter des formulaires dans lesquels sont renseignées les entités et les relations entre entités

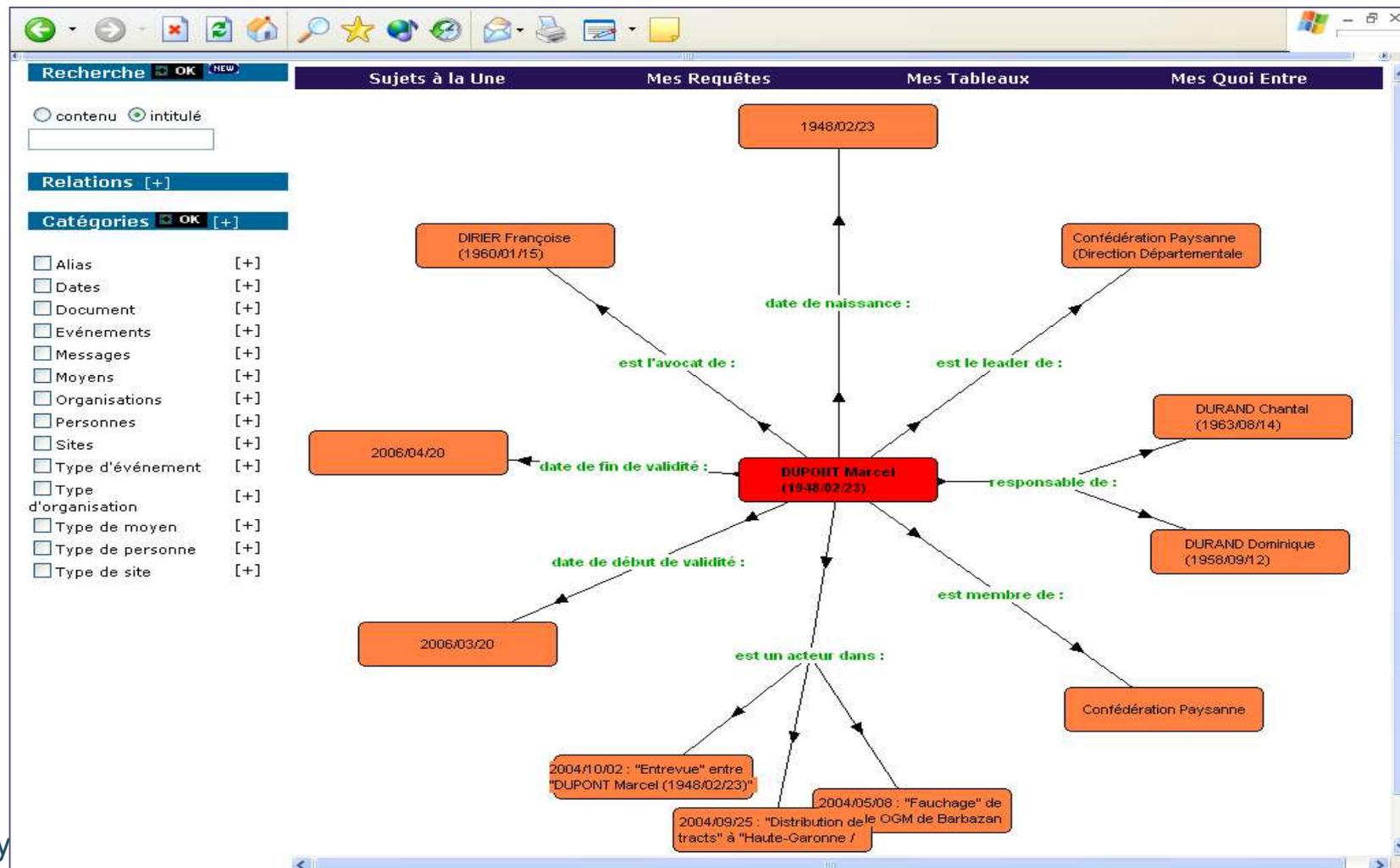
The screenshot shows a web application interface with a search bar and navigation tabs. The search results for 'DOC_0001_PAR_04' are displayed, showing a list of relationships. The relationships are as follows:

- est classé dans ▶ C_paragraphe
- est un(e) : ▶ c_paragraphe
- contient : ▶ relationDateFait /date /28/01/05 /fait /APPEL A MANIFESTATION
- texte : ▶ L'ASSOCIATION GREENPEACE; LA CONFEDERATION PACSANNNE ET LES FAUCHEURS VOLONTAIRES APPELLENT A UNE MANIFESTATION LE 28 JANVIER A 10 HEURES AU PORT DE LORIENT POUR EMPECHER LE DECHARGEMENT DU CARGO GOLDEN LION.
- cite heure : ▶ 10:00
- cite fait : ▶ APPEL A MANIFESTATION
▶ BLOCAGE DU DECHARGEMENT DU NAVIRE GOLDEN LION
- cite organisation : ▶ ASSOCIATION GREENPEACE
▶ CONFEDERATION PAYSANNE
▶ FAUCHEURS VOLONTAIRES
- est contenu dans : ▶ DOC_0001
- est membre de la classe ▶ G1

Gestion des connaissances



Les informations extraites permettent d'alimenter automatiquement une base de connaissances, sur laquelle les utilisateurs effectuent des requêtes. Dans l'exemple ci-dessous, on visualise les informations biographiques associées à une Personne en exploitant notamment les relations entre entités



date / references





Etape 3 : Mining



Des processus d'analyse statistiques sont utilisées en vue de faire l'étude quantitative et qualitative des données analysées

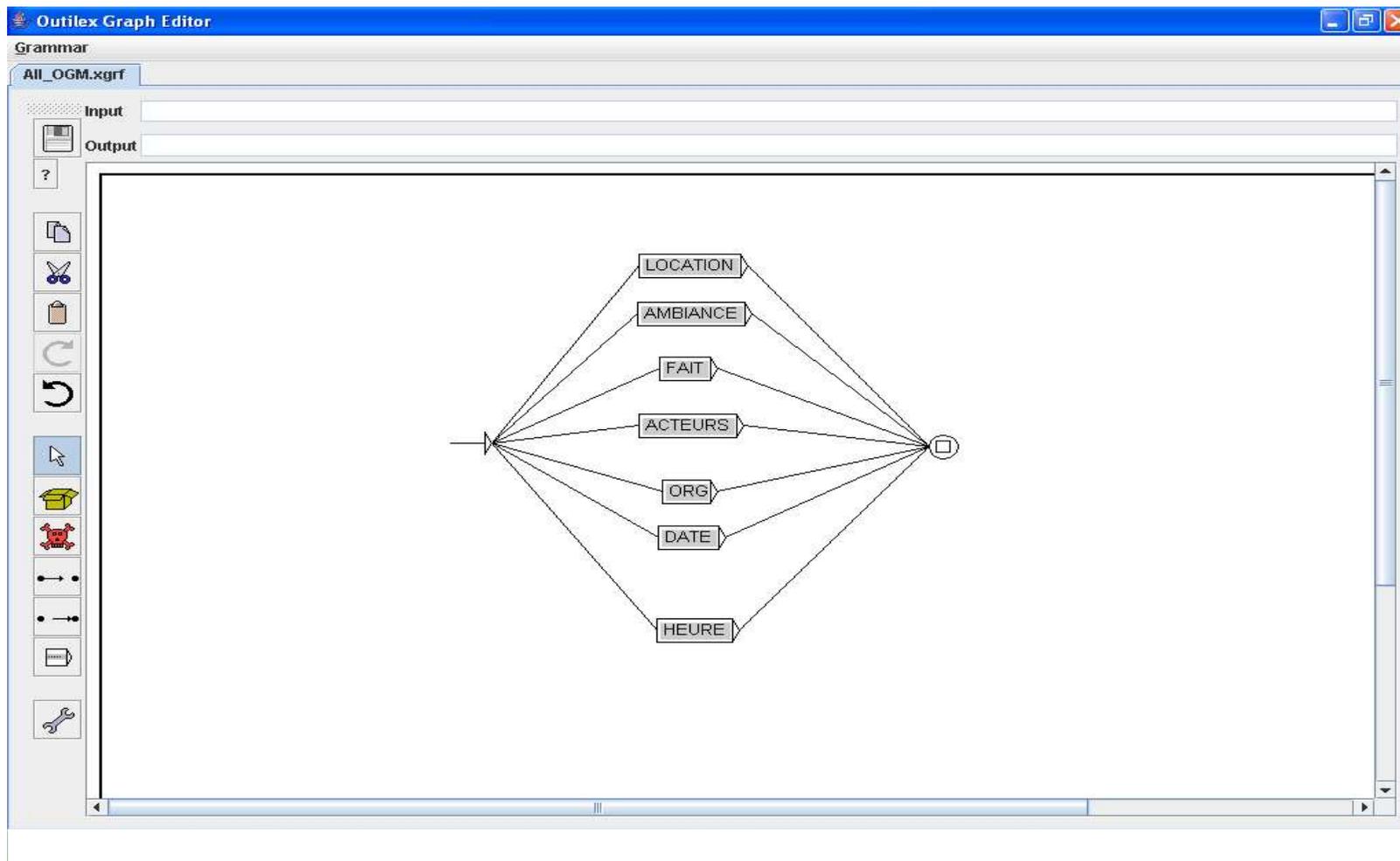
date / references



Aperçu des composants Outilex implémentés

date / references

Les ressources grammaticales ont été implémentées au format Outilex.
La grammaire ci-dessous est la grammaire d'extraction des entités nommées et des faits. En sortie d'analyse, les informations extraites sont étiquetées par des balises indiquant leur type.





```
<?xml version="1.0" ?>
- <infos>
  - <fonction>
    <who>Jose Bové</who>
    <position>Responsable</position>
    <organization>Faucheurs Volontaires</organization>
    <segment>José Bové, responsable des Faucheurs Volontaires</segment>
  </fonction>
</infos>
```



Les ressources lexicales Métier ont été implémentées au format Unitex, puis converties au format Outilex. Les regroupements appliqués aux informations extraites sont effectués à partir de processus de normalisation élémentaires (2 janvier 2005/02/01/2005), de dictionnaires et de règles de grammaires

Niveau morpho-syntaxique



confédération paysanne, .NP+ActeurOrg



Regroupement sémantique



<ActeursOrg> confédération paysanne (PREP DPT)* (PREP REGION)* </ActeursOrg>
GROUPE DE MILITANTS ANTI-OGM



Conclusions

date / references



L'expérimentation réalisée a permis de valider l'intérêt de la plate-forme sur les points suivants :

- ✓ Reprise de l'existant en termes de ressources grammaticales et lexicales (Format Unitex notamment)
- ✓ Développement rapide de nouveaux composants
- ✓ Intérêt de la normalisation des formats des ressources linguistiques
- ✓ Logiciel libre et communauté d'utilisateurs

La pondération sur les grammaires n'a pas été testée, il est clair néanmoins qu'il s'agit d'un point fort.

La taille du corpus d'expérimentation ne permet pas de se prononcer sur les performances en termes de temps de traitement de la plate-forme.