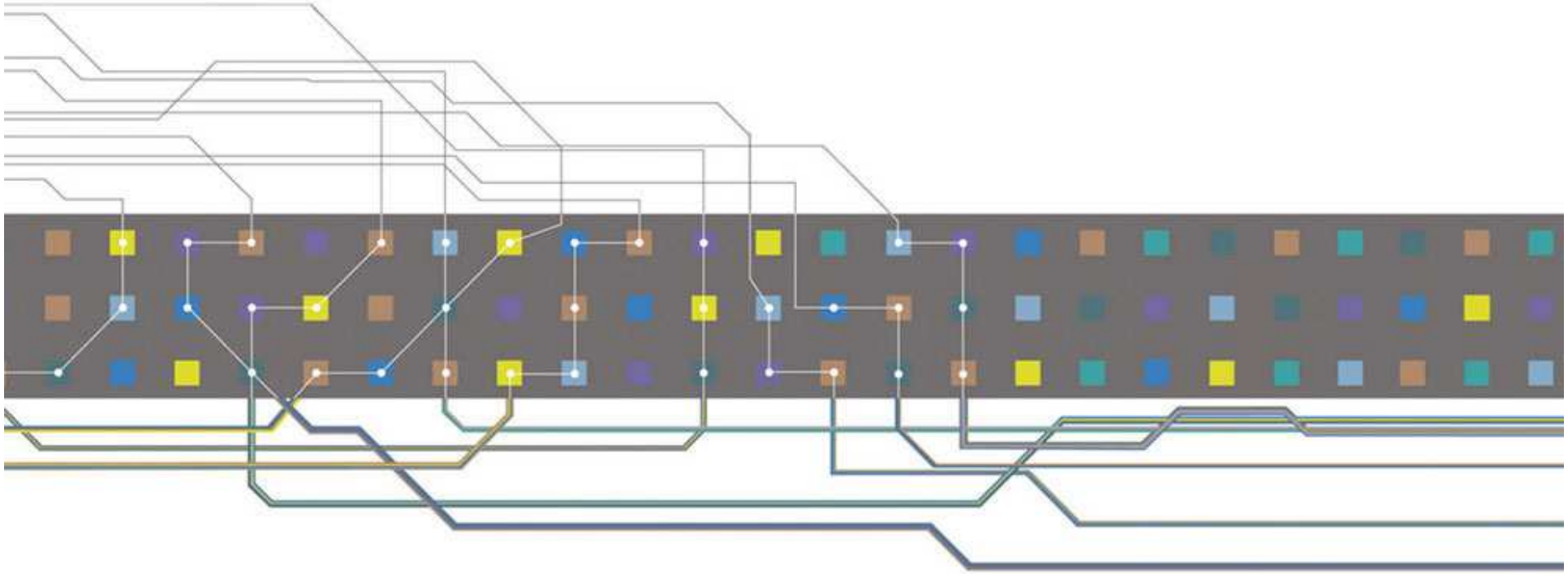
THALES



# Projet OUTILEX
# Rapport d'étude final
# Octobre 2006

date / references

Systèmes Terre et Interarmées

# 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

# Contributions attendues de Thales Communications

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)
- ✓ Dévelopement 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 intialement prévu portant sur l'analyse des incidents dans le domaine de l' automobile n'a pas pu être développé dans le cadre du projet.

# Contributions attendues de Thales Communications



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 plate-forme Outilex dans un contexte industriel, en vue de répondre à des besoins métier non triviaux

# Les besoins

# Besoins en matière d'extraction d'information

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

# Corpus d'expérimentation

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.

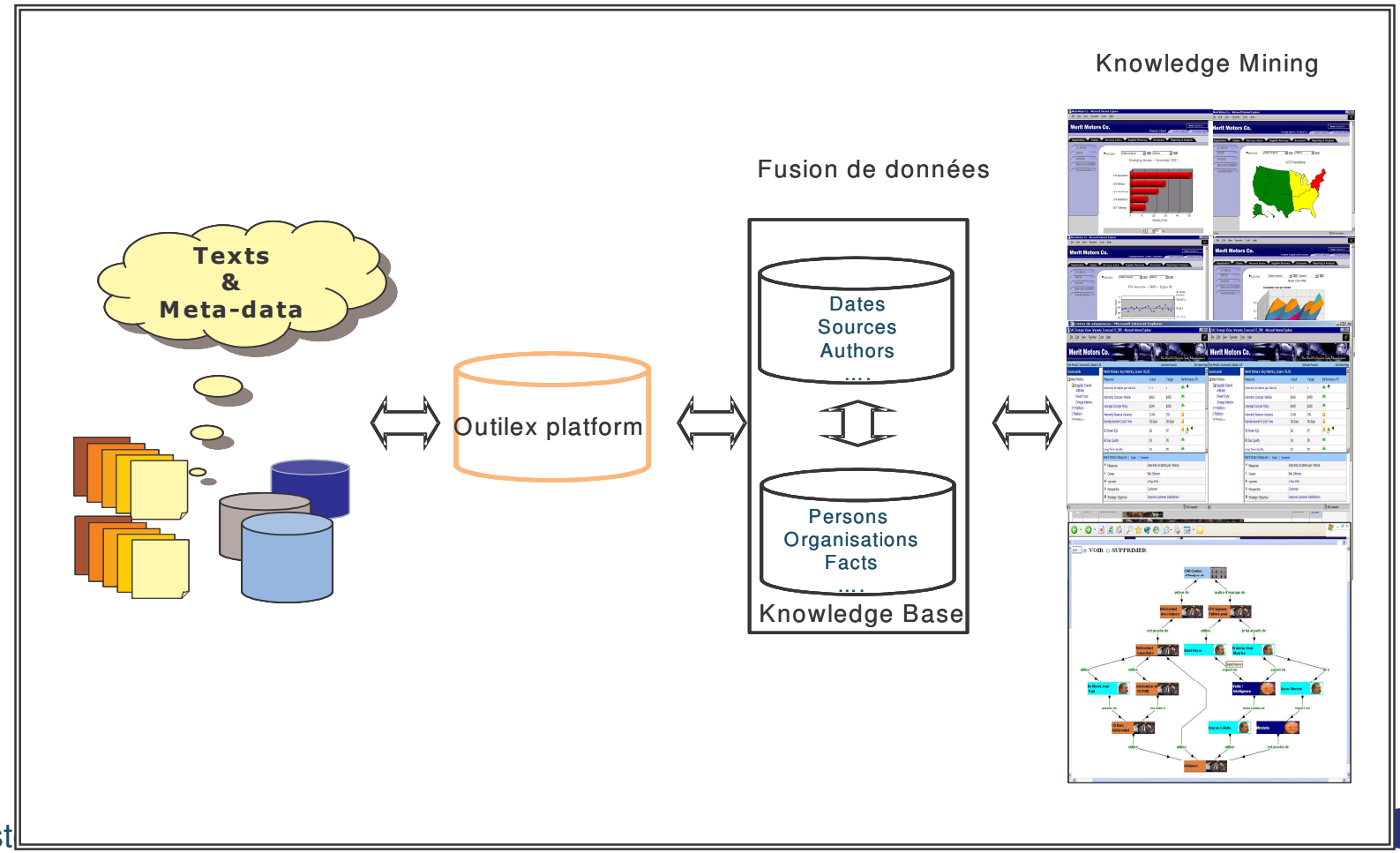
Chaque rapport est composé de données structurées – objet , date , auteur, source - et d'un texte.

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

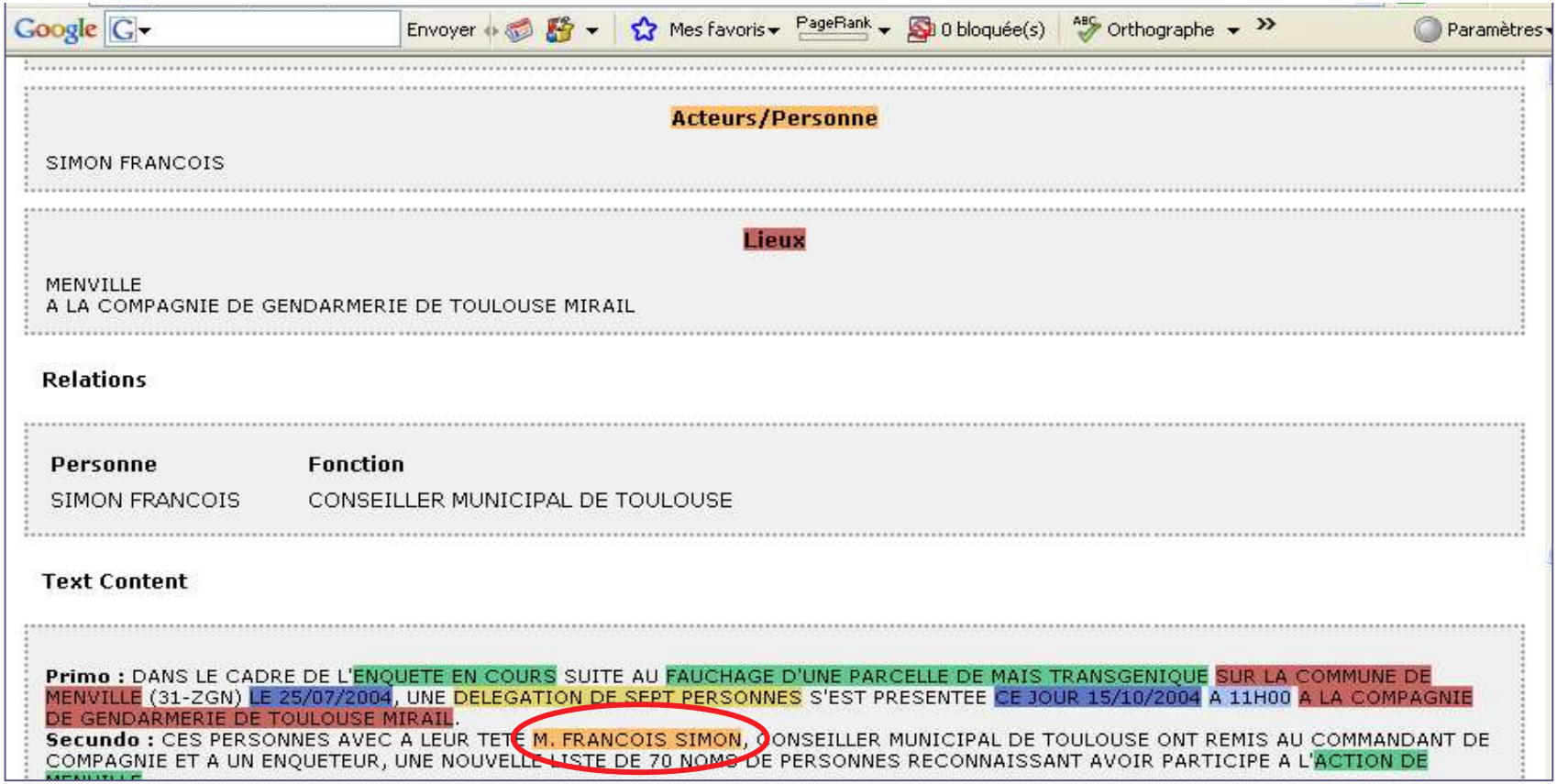
# Schéma du démonstrateur

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.



# Etape 1 : Extraction d'information

# Exemple Entités Nommées/Personnes



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

**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

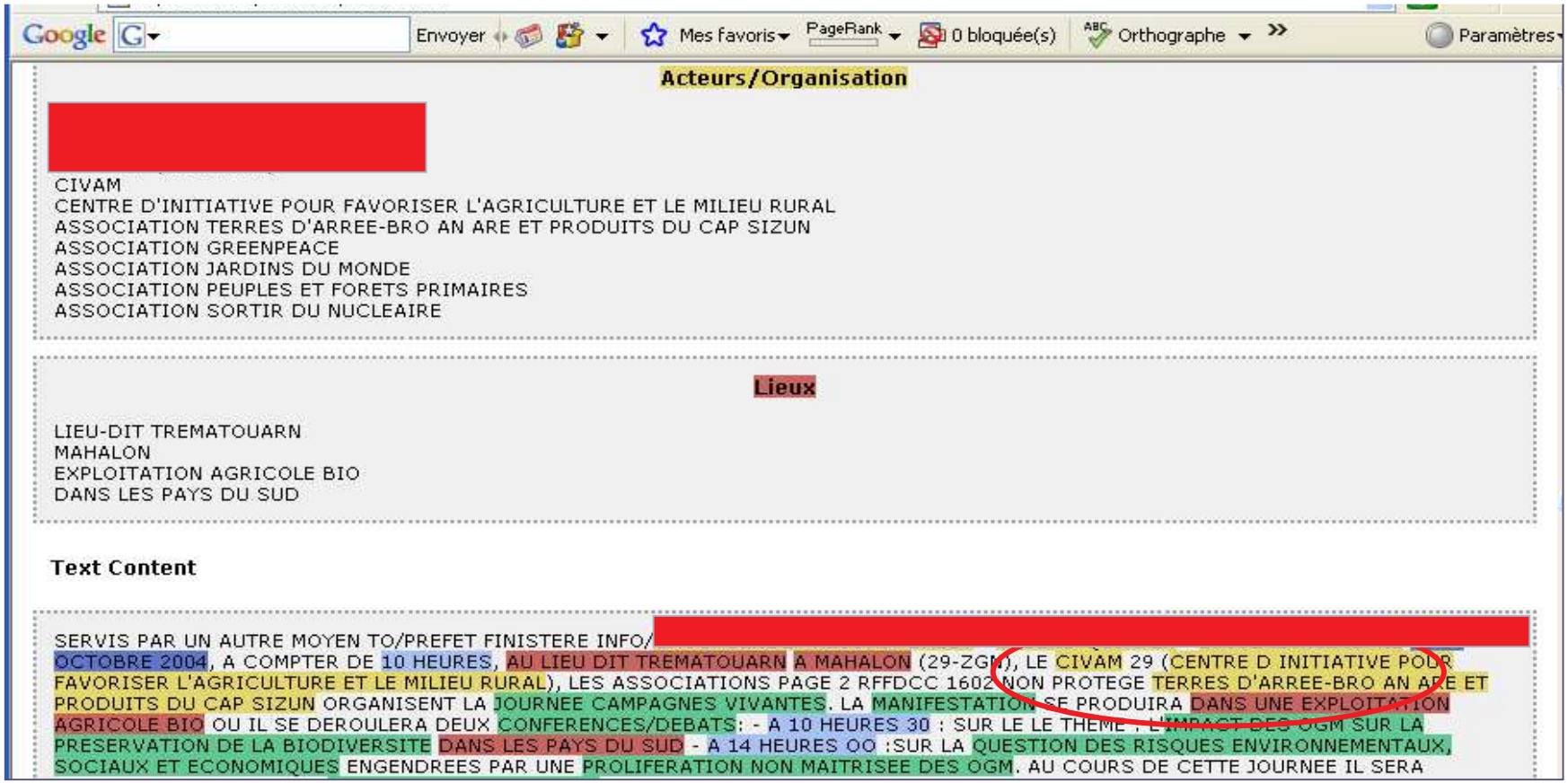
date / references





THALES

# Exemple Entités Nommées/Organisation



Google | 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 OO :SUR LA QUESTION DES RISQUES ENVIRONNEMENTAUX, SOCIAUX ET ECONOMIQUES ENGENDREES PAR UNE PROLIFERATION NON MAITRISEE DES OGM. AU COURS DE CETTE JOURNEE IL SERA

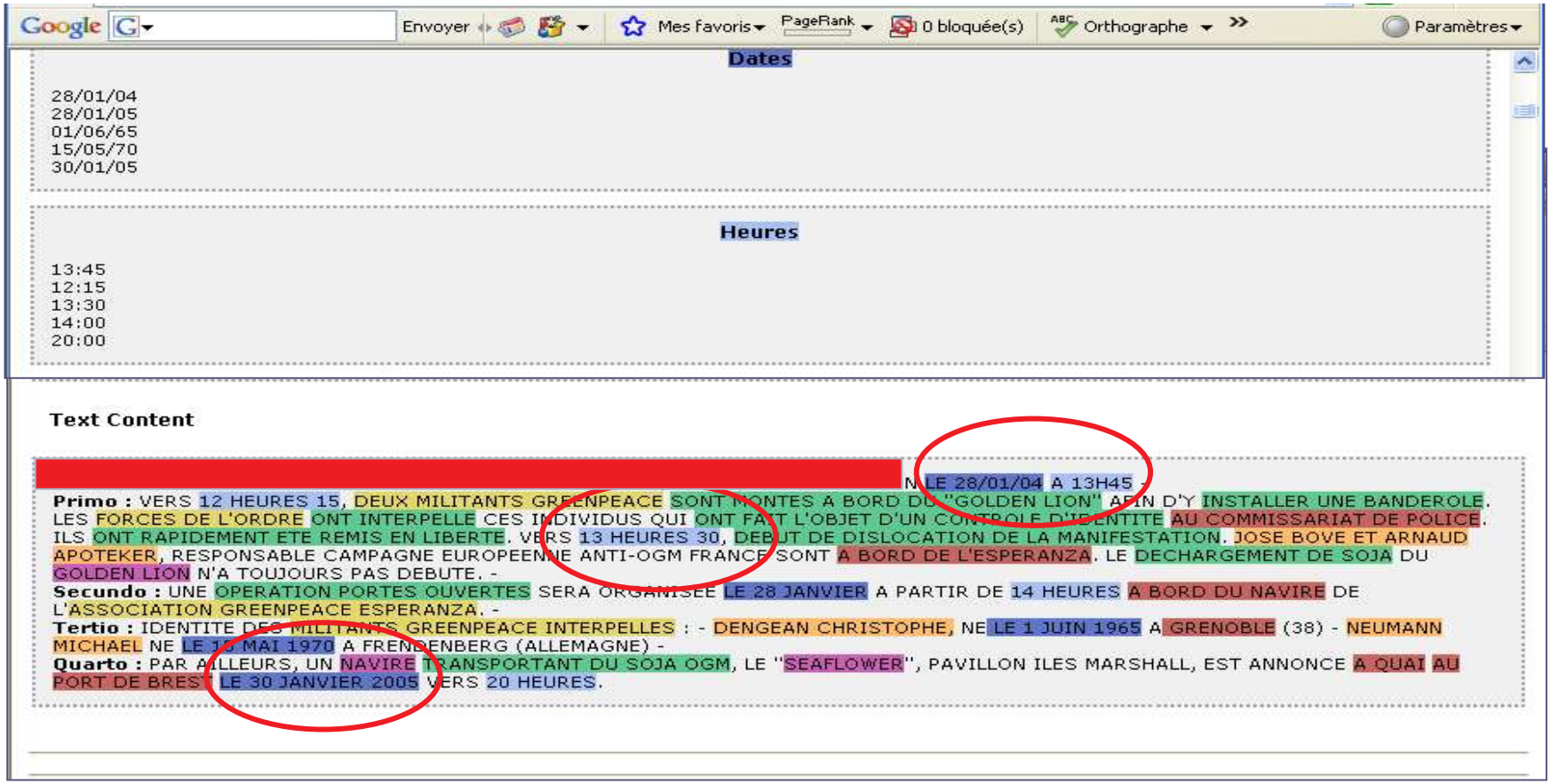
date / references





THALES

# Exemple Entités Nommées/Dates et Heures



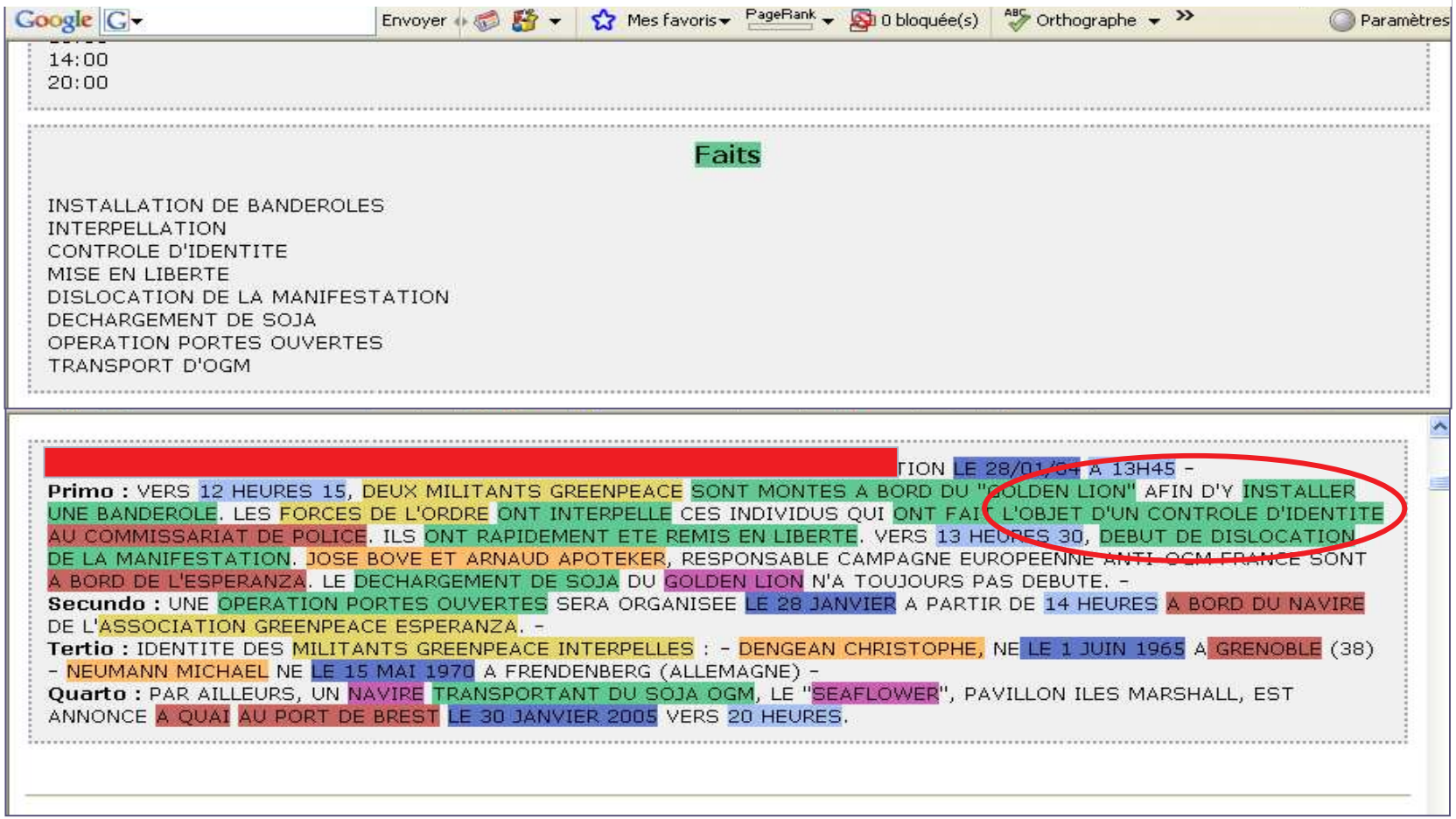

date / references





THALES

# Exemple Faits




date / references

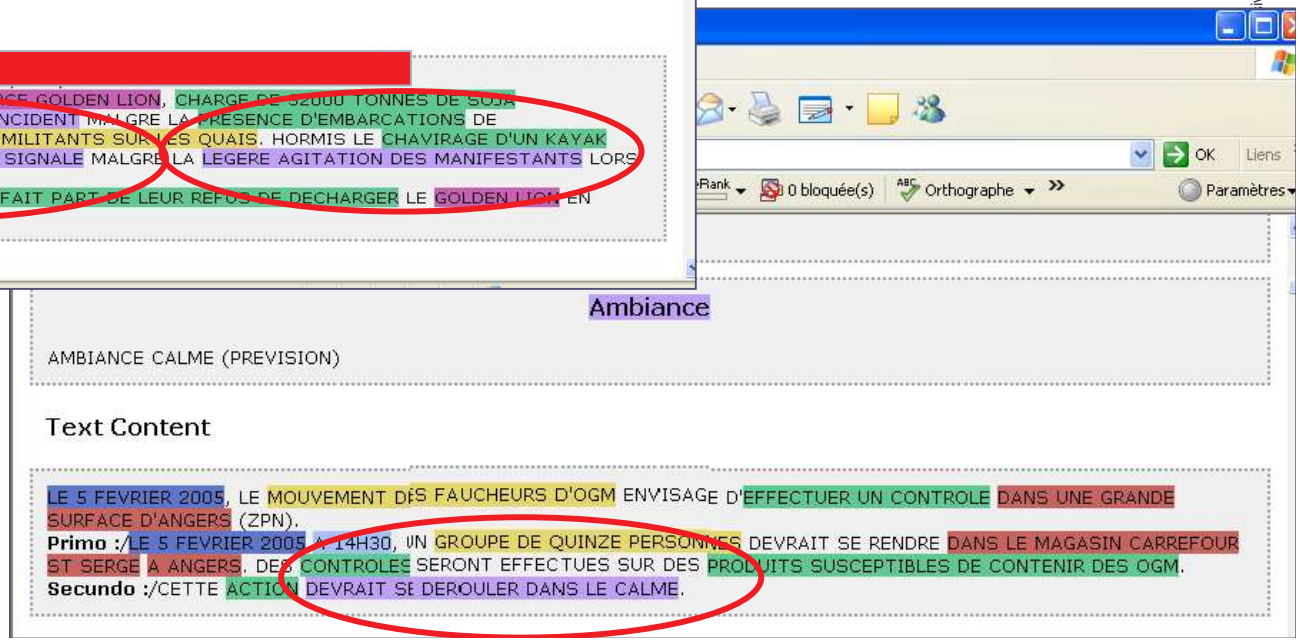




THALES

# 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

# 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 ….

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

| 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 MARIA LOURDES | 25/02/62 | 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 ZENARRUZABEITIA | VICE-LEHENDAKARI |

date / references





THALES

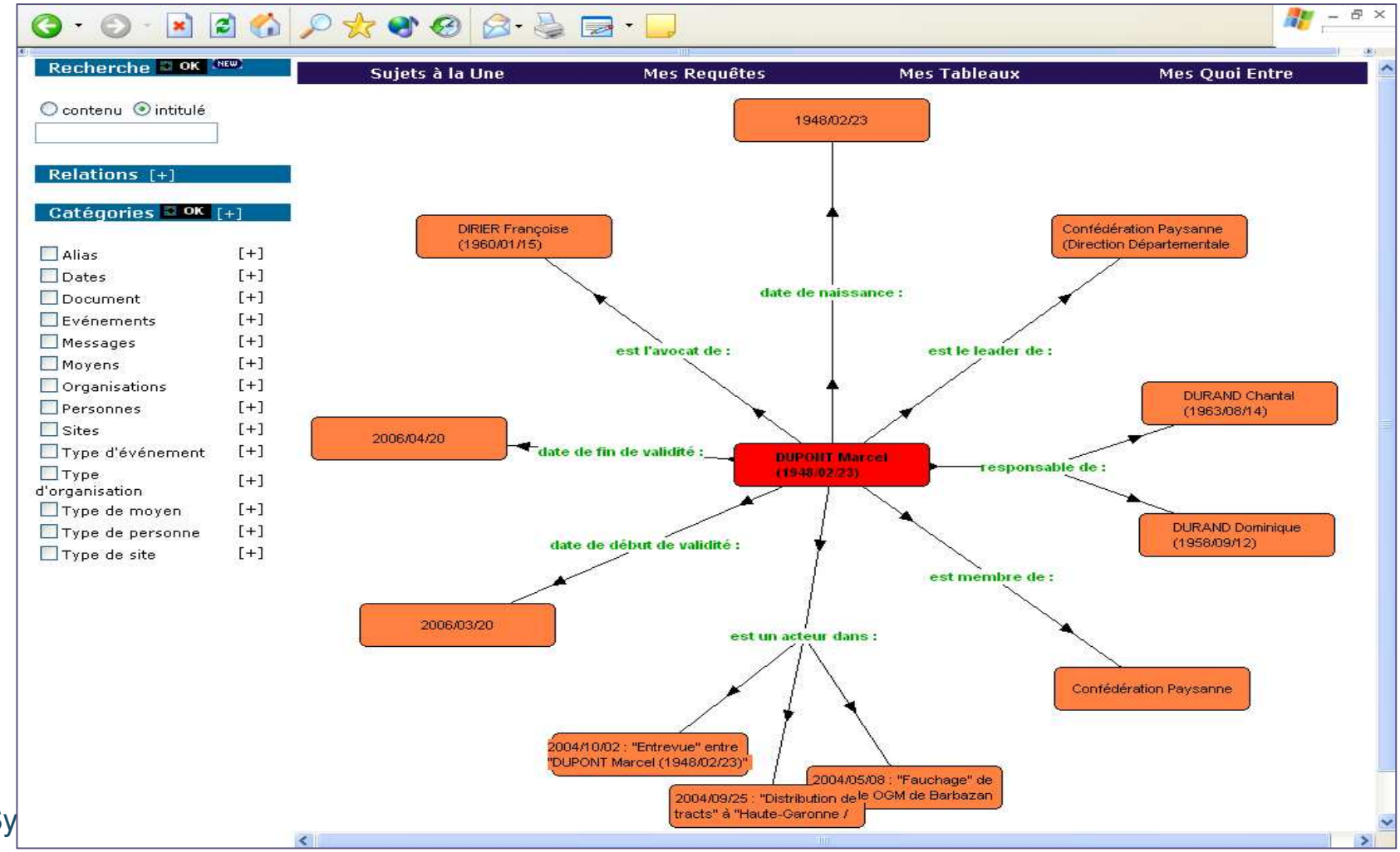
# Etape 2 : Gestion des Connaissances

# Gestion des connaissances

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

# 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



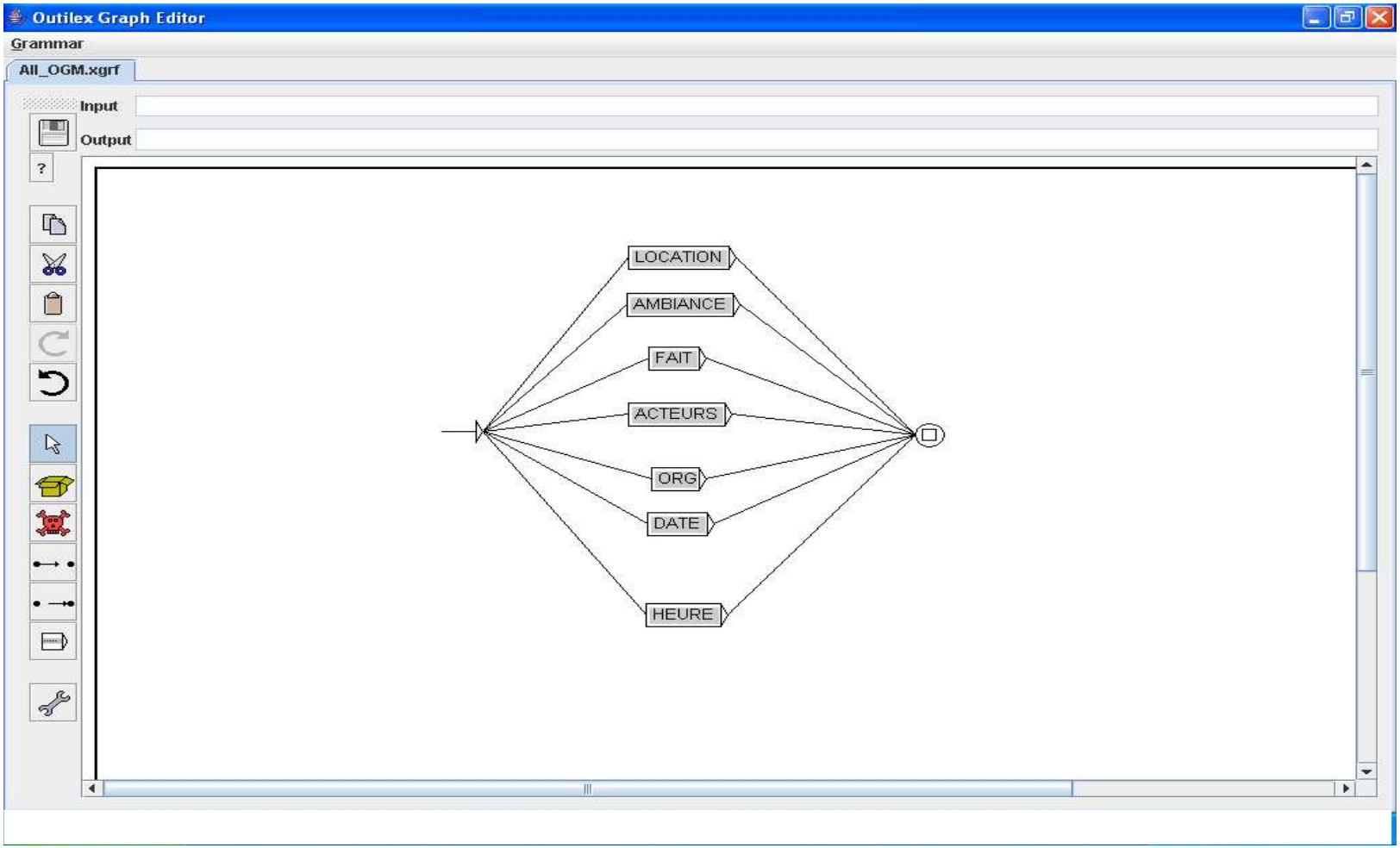
# Etape 3 : Mining

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

# Aperçu des composants Outilex implémentés

# Grammaires locales

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.



# Identification des relations

```
<?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>
```

# Ressources lexicales

Les ressources lexicales Métier ont été implémentées au format Unitex, puis converties au format Outilex. Les regroupements appliquées 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

conféderation paysanne,.NP+ActeurOrg

Regroupement sémantique

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

# Conclusions

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.