

Aide à la classification de documents hétérogènes

Troisième Revue

27 Février 2004

Proposé par:

Talel Abdessalem

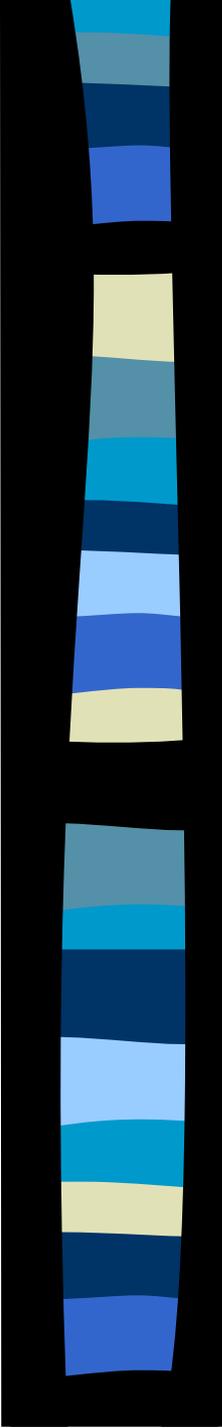
Préparé par:

Adrien Missemmer

Céline Monthéard

Nicolas Boulanger

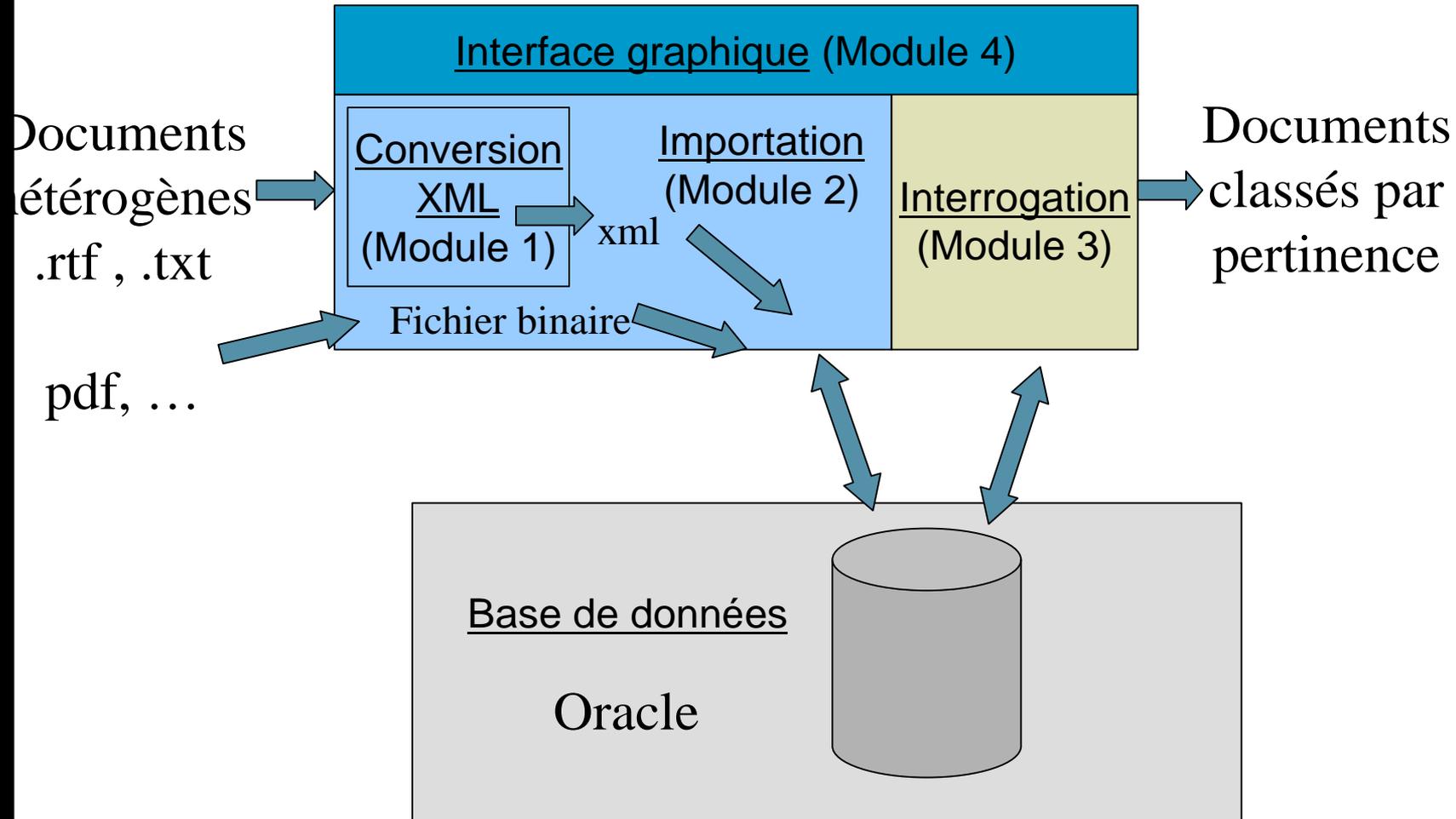
Rana BouJaoudé



Plan de la présentation

- Rappel : Logiciel MyDOX
- Constitution de la base
 - Conversion en XML
 - Analyse
- Requêtes sur la base
- Noyau
- Mise à jour de la planification du projet

Rappel : Logiciel MyDOX



1- Conversion en XML

 Txt2xml
<ul style="list-style-type: none"> fichierin: File fichierout: File s0: String s1: String
<ul style="list-style-type: none"> Txt2xml() getPathin() getPathout() convert()

 Rtf2xml
<ul style="list-style-type: none"> fichierin: File fichierout: File s0: String s1: String
<ul style="list-style-type: none"> Rtf2xml() getPathin() getPathout() convert()

1- Conversion en XML

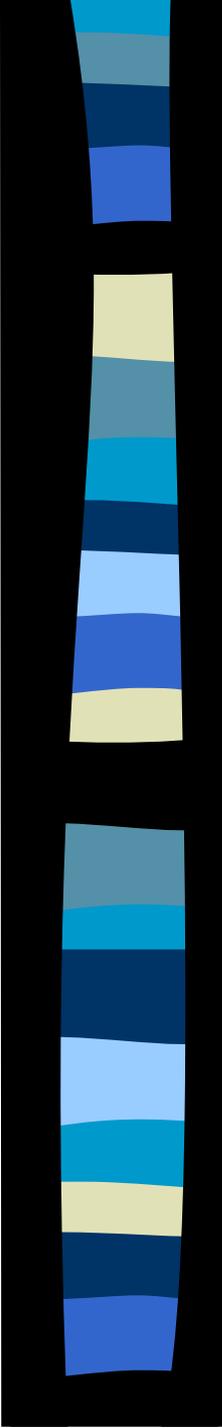
- convert de Rtf2xml

- **Appel Majix**

- Runtime : Accéder à l'environnement d'exécution
 - Process : Exécution de Majix dans un processus séparé ➡ .xml

- **Elimine XSL**

- BufferedReader : Lecture du .xml
 - BufferedWriter : Génération du nouveau .xml



1- Conversion en XML

- convert de Txt2xml

- BufferedReader : Lecture du .txt
- BufferedWriter : Génération du .xml
 - Balises : `<info> <title> ... </title> </info>`
`<p> ... </p>` ou `<p />`
 - Suit la dtd de Majix

1- Conversion en XML

■ Document XML final

```
<?xml version="1.0" encoding="ISO-8859-1" standalone="no"?>
```

```
<!DOCTYPE mydoc (View Source for full doctype...)>
```

```
- <mydoc>
```

```
  - <info>
```

```
    <title> Test </title>
```

```
  </info>
```

```
- <p>
```

```
  <b> Module 1 </b>
```

```
  </p>
```

```
</mydoc>
```

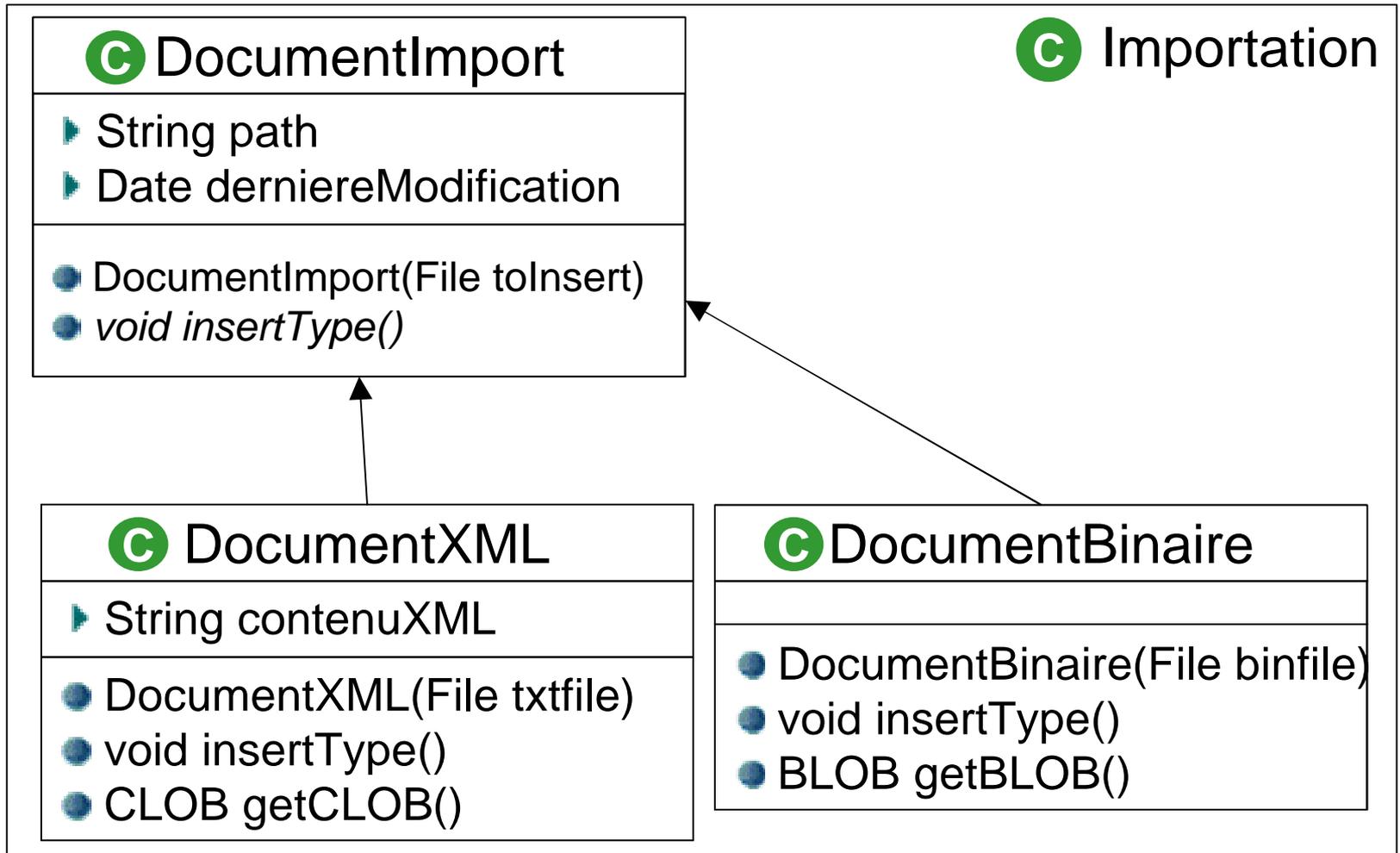
2 – Importation des documents

Importation

▶ private Connection connexionOracle

- public Importation()
- public void close()
- public void inserer(String pathname)
- public void mettreAJour(String pathname)
- private DocumentImport loadFile(String pathname)

2 – Importation des documents



2- Constitution de la base

■ La Table des documents

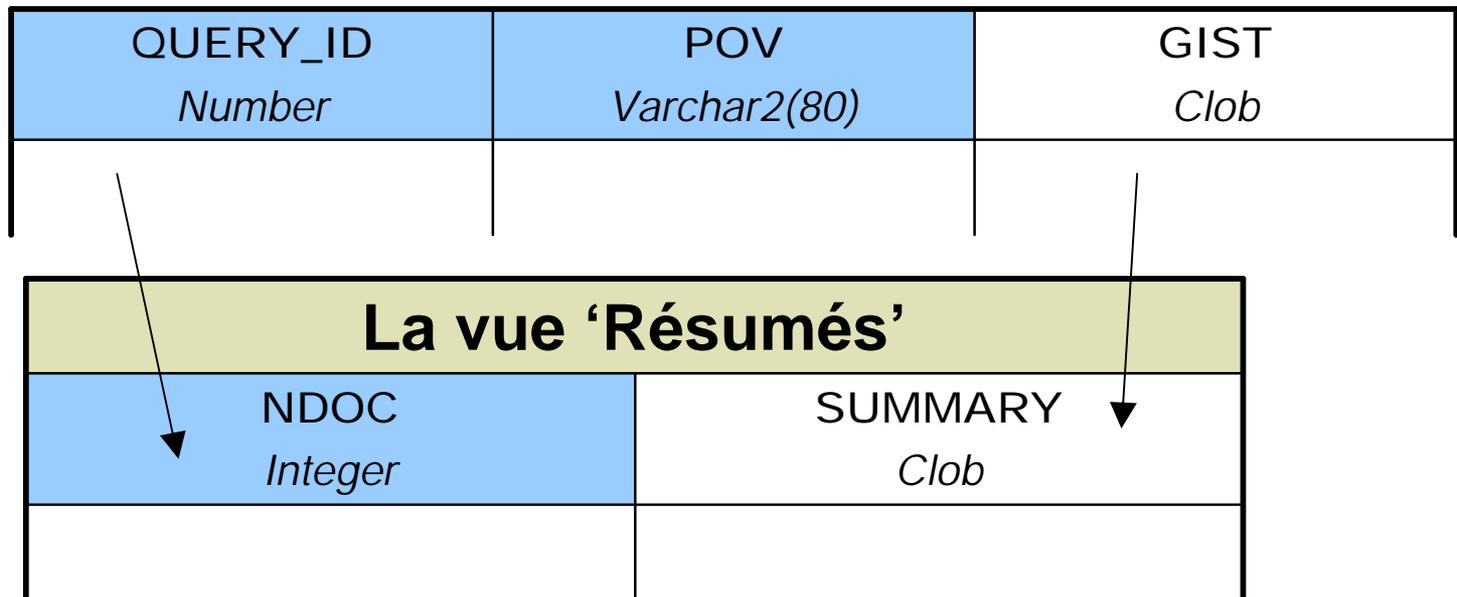
NDOC <i>Integer</i>	URL <i>Varchar(256)</i> <i>unique</i>	CONTENT <i>Clob</i>	MODIFIED <i>Date</i>	TREE <i>Varchar(256)</i>
 Clés				 Arborescence virtuelle

Remarque : NDOC est déterminé automatiquement par Oracle au moment de l'insertion (compteur incrémenté à chaque insertion)

2- Constitution de la base

■ La Table des résumés

- La procédure Oracle `ctx_doc.gist` détermine un résumé d'un document et le stocke dans une table :



2- Constitution de la base

■ Rappel

- Indexation gérée automatiquement par Oracle
- Basée sur 3 tables
 - Table des mots (mot + n° mot + type_index)
 - Table de localisation (n° mot + pointeur document)
 - Table annexe (n°mot + occurrences + localisation fine)
- Et sur des algorithmes spécifiques (STOP Mode, PASS Mode, Column Specific)

3- Requêtes sur la base

C Requete	
△	stmt: Statement
△	conn: Connection
● ^C	Requete()
●	resume(in String)
●	keyword(in String)
●	update(in String)
●	extract(in String)
●	fermeture()

3- Requêtes sur la base

- keyword (String mot)
 - Lit la chaîne de caractères
 - Écrit la requête : `Select ndoc from doc...`
 - Exécute la requête : `stmt.executeQuery`
 - Traduit le résultat en String (sa valeur de retour)
- resume (String resu)
 - Similaire à keyword (String mot)

3- Requêtes sur la base

- update (String urldufichier)
 - Recherche date de stockage du fichier dans la base grâce à une requête SQL
 - Recherche date de la dernière modification du fichier se trouvant sur le disque (File.lastModified())
 - Comparaison des deux dates
 - Retourne un boolean : True si date disque > date base

4- Noyau

- Un noyau stable est déjà en place
 - Convertit fichier txt et rtf en xml
 - Se connecte à la base oracle distante
 - Place le contenu xml dans la base
- Evolution du développement
 - Intégration du module de requêtes
 - Gestion de l'indexation
 - Importation de fichiers non XML
 - Interface graphique

5- Planification

- Planification (heures / groupe)

Etape	Temps estimé	Temps passé
Planification	25	25
Analyse	90	80
Conception	155	90
Codage & Tests	135	
TOTAL	405	