Une plateforme ouverte et flexible pour l'exploitation des contenus

http://www.webcontent-project.org

Gaël de Chalendar

CEA LIST / LIC2M

Fontenay-aux-Roses - France

Objectifs du Projet

- Fournir à partir d'une plateforme centrale :
  - Modules Technologiques dédiés au traitement et au stockage des contenus multimédia
  - Outils et tutoriels pour l'intégration et la construction d'applications
- Identifier et intégrer des modules existants, mais « épars »
- Développer et intégrer certains modules allant au-delà de l'Etat de l'Art:
  - TAL Multilingue
  - Gestion de gros volumes de données en P2P
  - Enrichissement Sémantique (automatique ou semi-automatique)
  - Déploiement de services Web
- Proposer la plateforme à d'autres partenaires et d'autres projets Français ou Européens

Problèmes et Besoins

Les données accessibles sont

- Principalement non-structurées (80% de l'information disponible)
- Hétérogènes dans leurs contenus et leurs formats
- Toujours plus nombreuses et volumineuses

Leur utilisation implique des difficultés fonctionnelles pour leur

- Collecte, Acquisition,
- Stockage, Indexation,
- Transformation, Normalisation,
- Description, Annotation,
- Visualisation, Présentation,
- Structuration, Ordonnancement,
- Dissémination, Partage,
- ...

Partenaires

Académiques

- CEA LIST
- INRA Mét@risk
- INRIA - GEMO
- INRIA - Mostrare
- INRIA - InSitu
- INRIA - Exmo
- LIP6
- PriSM
- LIG
- LIMSI-CNRS
- Grimm
- PSY.CO

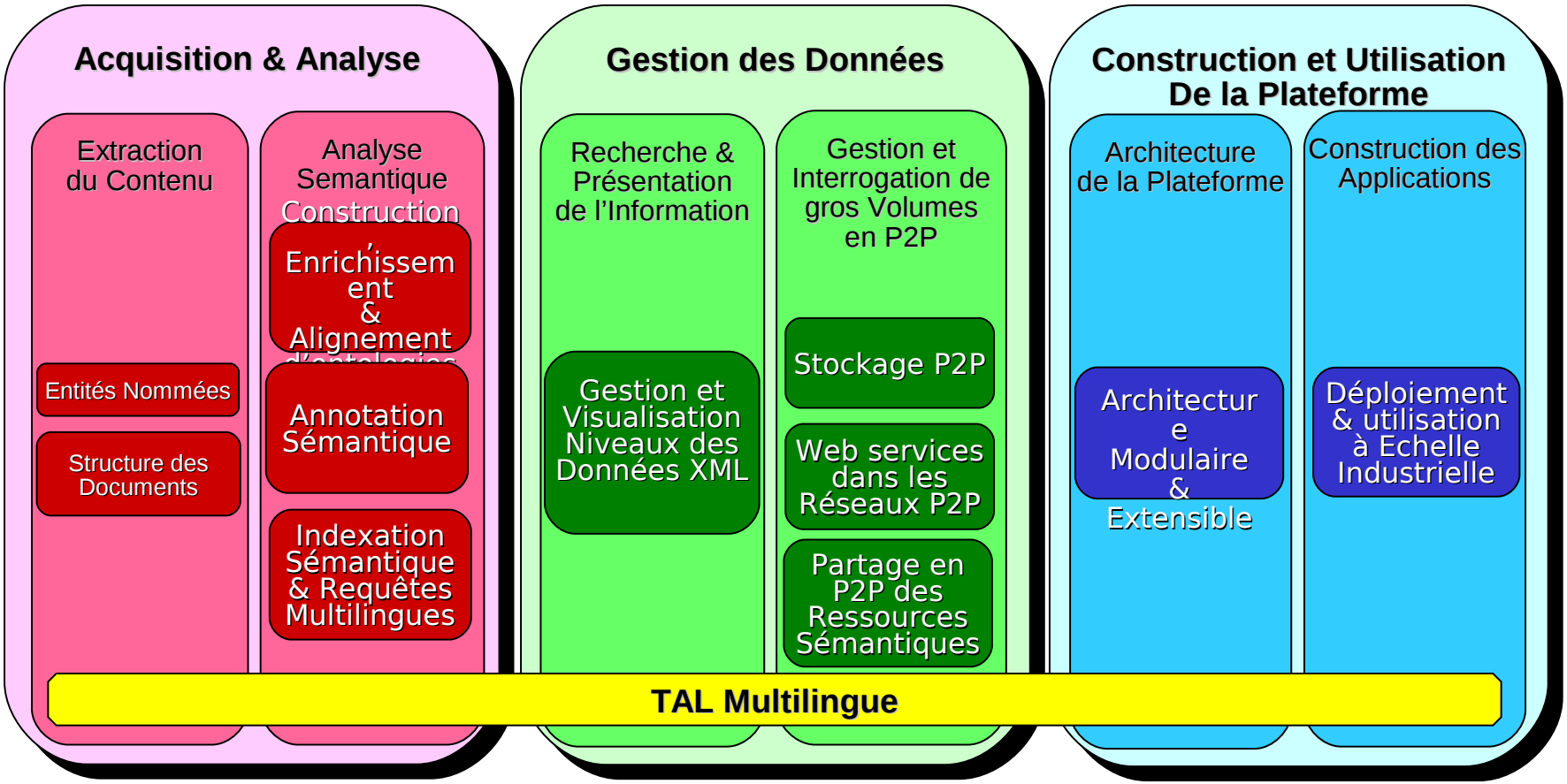
Industriels

- EADS DCS
- Thales R&T
- Exalead

Utilisateurs

- ADRIA
- Soredab (Bongrain)
- CEA DAM

Aspects Abordés



Cas d'Utilisation

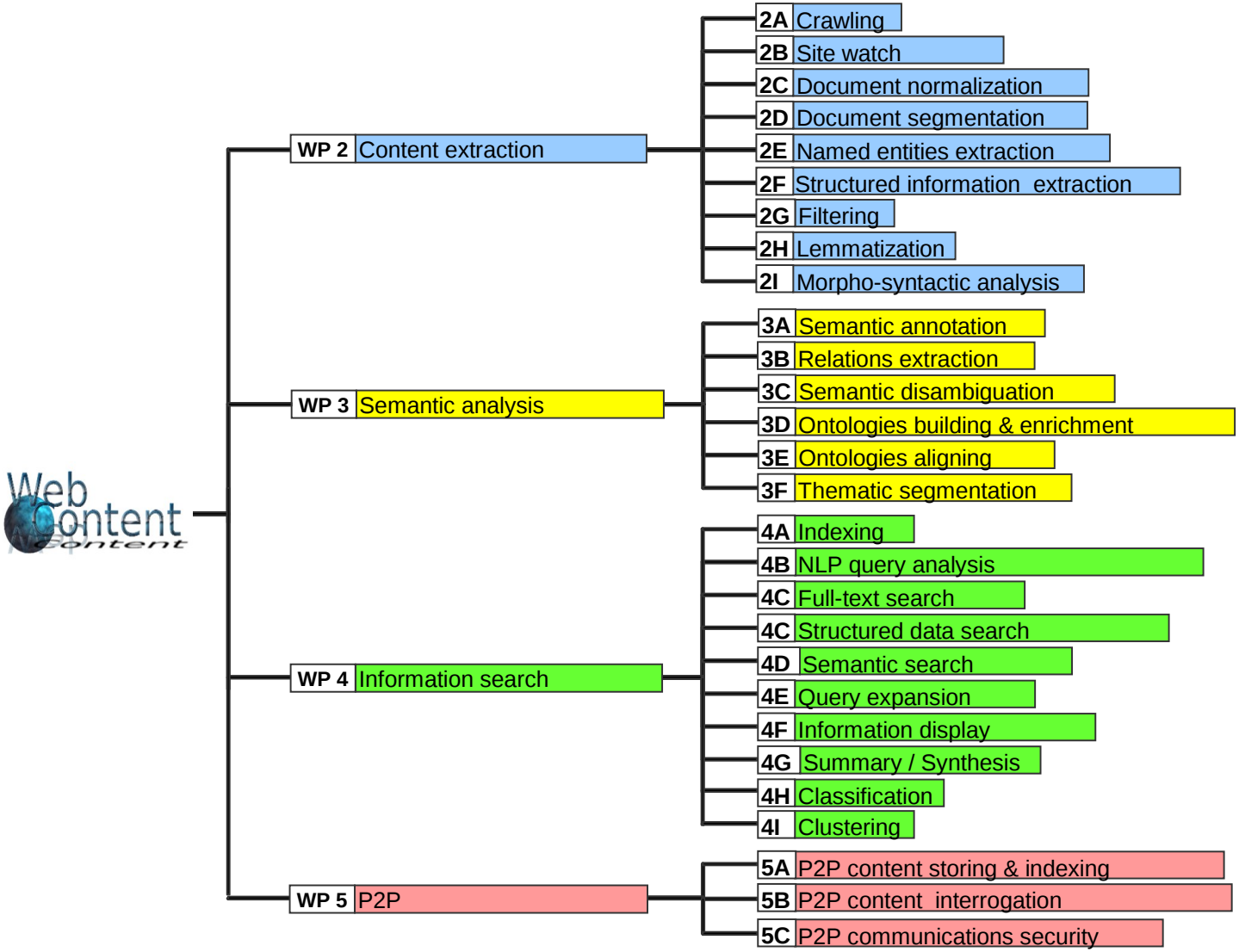
4 Applications de « Test » :

- Veille économique et technologique dans l'aéronautique ↔ **EADS / AIRBUS**
- Veille Stratégique ↔ **THALES**
- Risques Microbiologique et chimique dans l'alimentation ↔ **SOREDAB (groupe Bongrain)**
- Veille événementielle (Sismique) ↔ **CEA DAM**

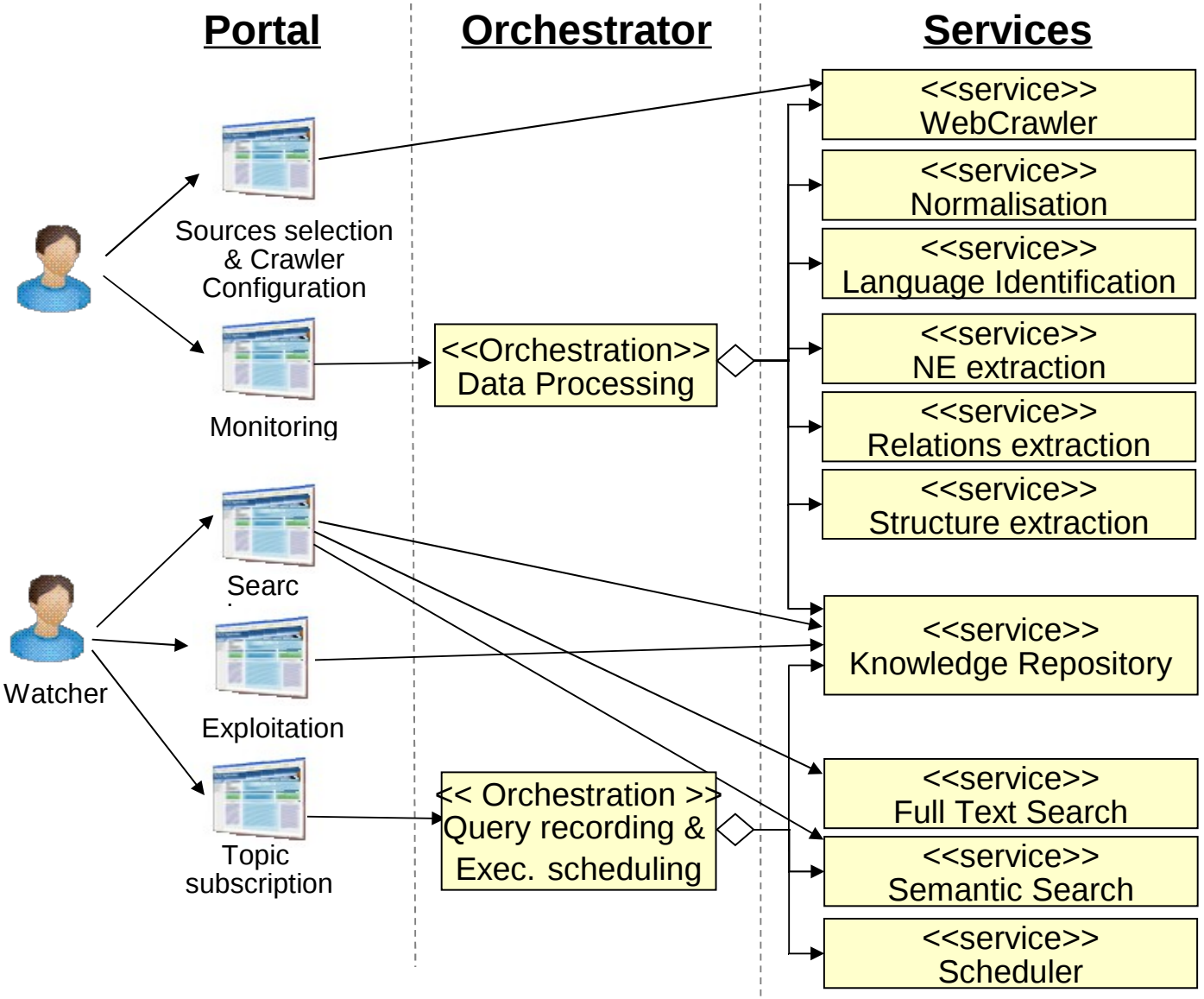
Approche Adoptée

1) Elaboration d'une carte Fonctionnelle
2) Listage des modules logiciels utilisables
3) Evaluation de la couverture fonctionnelle
4) Choix technologique pour intégrer/fédérer les modules
5) Définition d'architecture d'intégration
6) Spécification des divers composants de la plateforme
7) Codage/intégration des composants
8) Construction des applications opérationnelles

Carte Fonctionnelle



Application Airbus



Méthode de Conception

- La plateforme WebContent doit offrir des **services** permettant de construire des **applications**.
- Les services sont offerts par des **composants** amenés par divers fournisseurs
- La plateforme présente une **infrastructure** d'**intégration** de services
- Diverses **chaînes de traitement** peuvent être élabores par composition de services
- Un **portail** donne accès à l'interface graphique des applications et aux ressources disponibles

⇒ ARCHITECTURE ORIENTEE SERVICE

(Service Oriented Architecture, SOA)

Fonctionnalités des Services

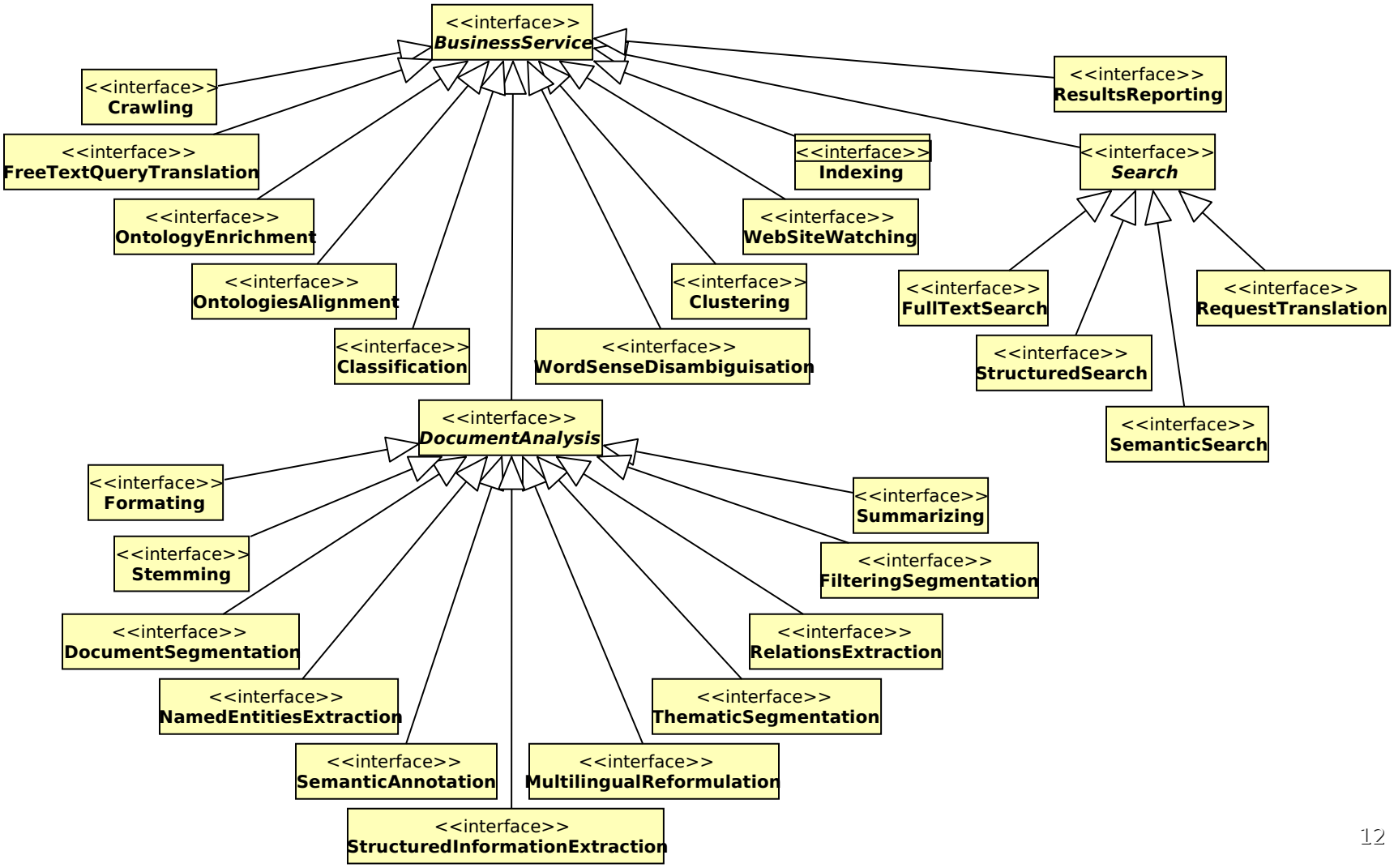
Chaque service est

- Indépendant (du point de vue de l'utilisateur)
- Défini par un contrat (interface + conditions d'utilisation)
- Normalisé (spécification de l'interface)
- Possiblement implémenté de plusieurs manières

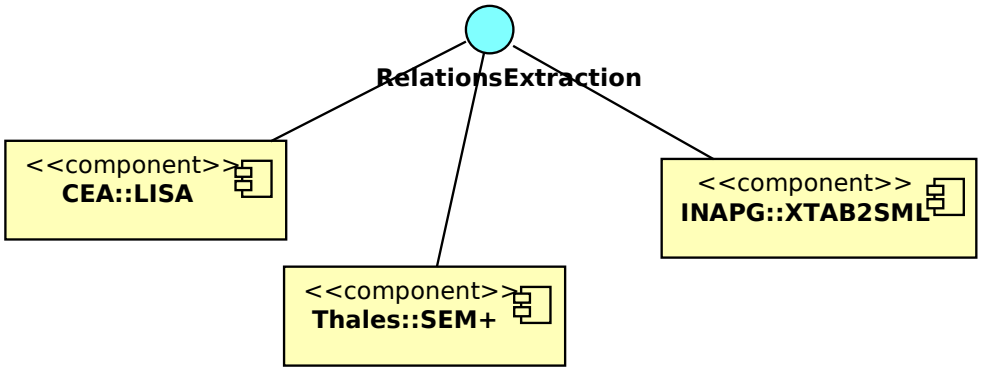
Services « Métier »: réalisent les fonctions identifiées dans la carte fonctionnelle

Services Techniques: sont une base commune pour le développement d'applications

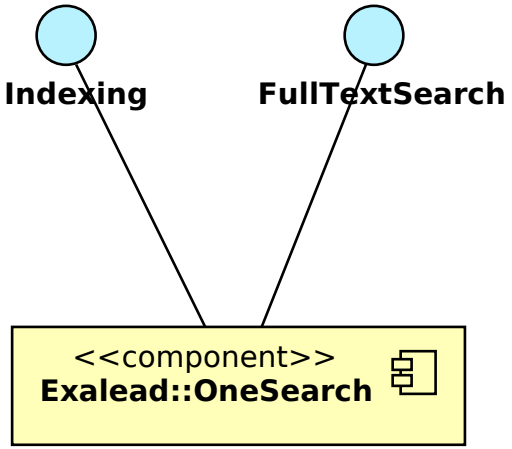
Les services « Métier » de WebContent



Implémentation des Services



Plusieurs implémentations d'un même service



Plusieurs services remplis par un même composant

Bus de Services (ESB)

- **Intégration et réduction du couplage**
  - Exposition et appel des services (médiateur entre le consommateur et le fournisseur)
  - Connecteurs, agrégation de services, etc.
- **Distribution et routage**
  - Appels de services routés vers le bon fournisseur
  - Routage par tables ou dépendant du contenu XML (CBR)
  - Choix du protocole de transport (http, JMS, SMTP, etc.)
  - Quality of Service (garantie de transport des messages)
- **Transformation**
  - Transformation des formats de données d'une application à une autre
- **Orchestration de services**
  - Assemblage de services pour créer des services plus puissants
- **Gestion technique des activités**
  - Traçabilité des échanges
  - Gestion du cycle de vie des services
  - Traitement d'erreurs

L'ESB PEtALS

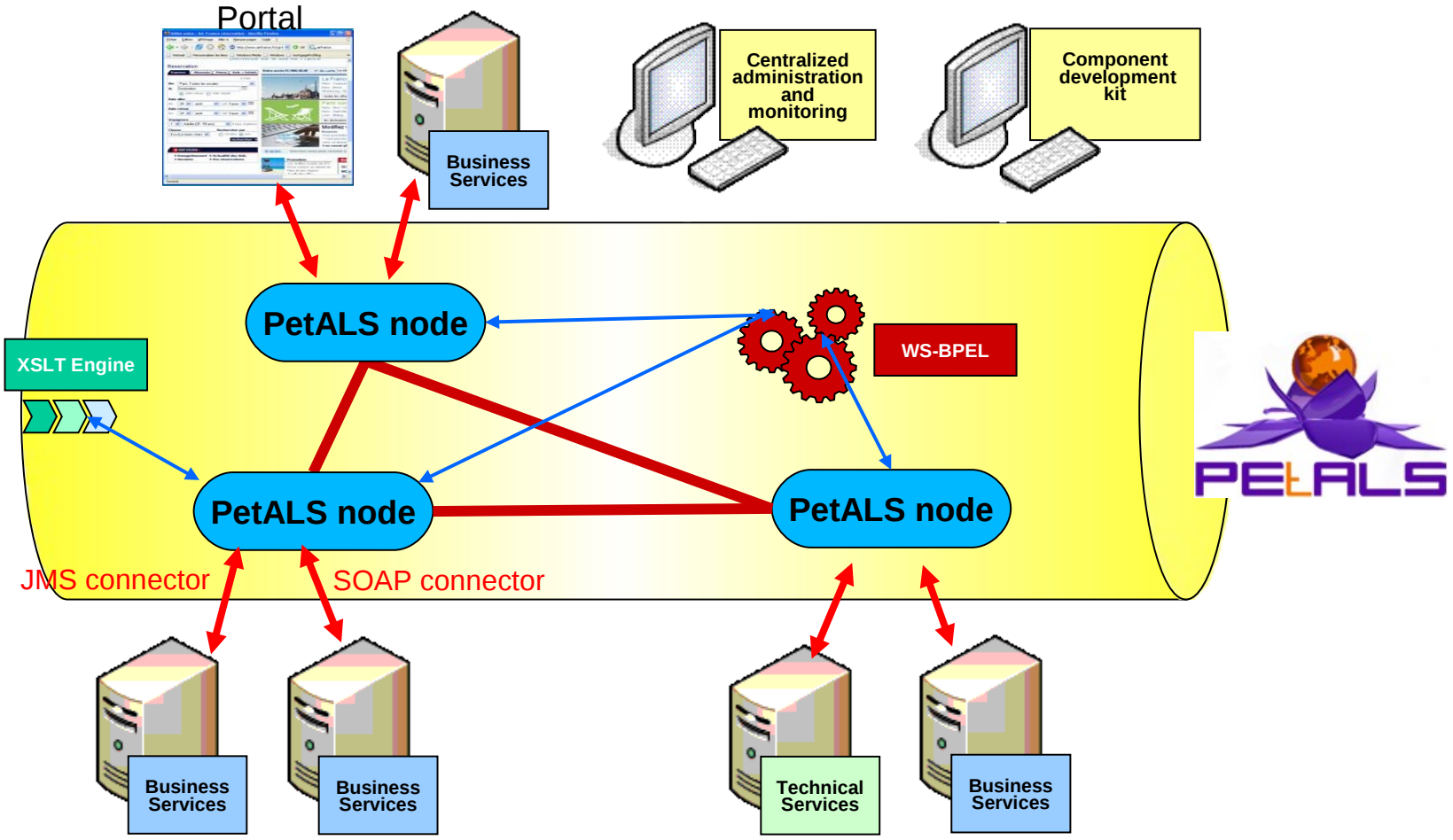
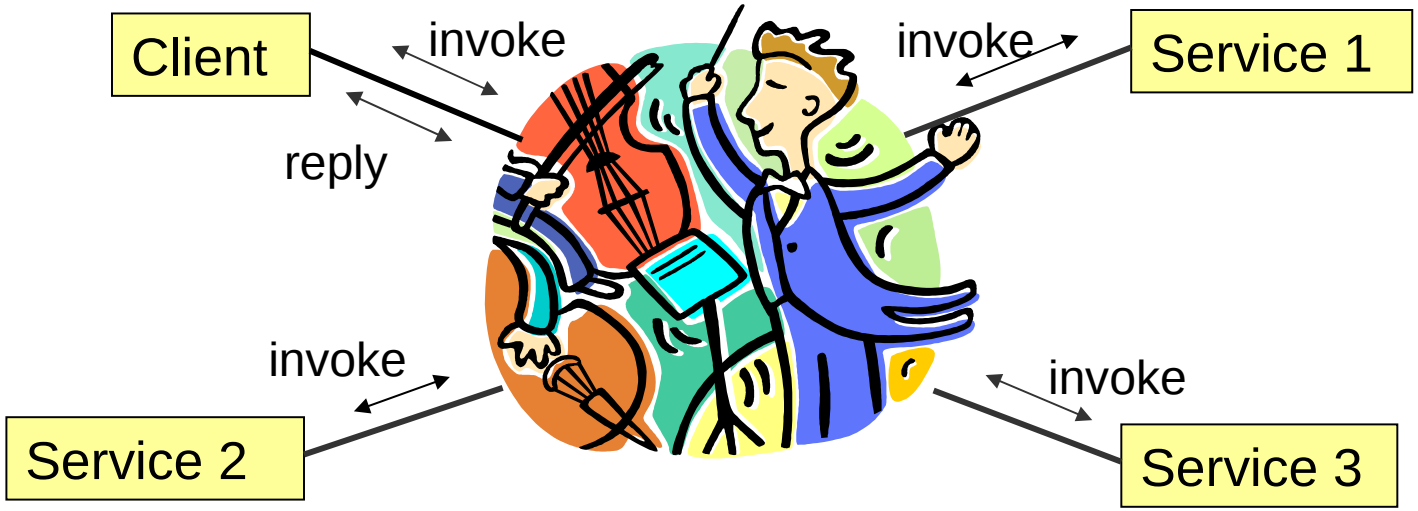


Schéma EBM Websourcing

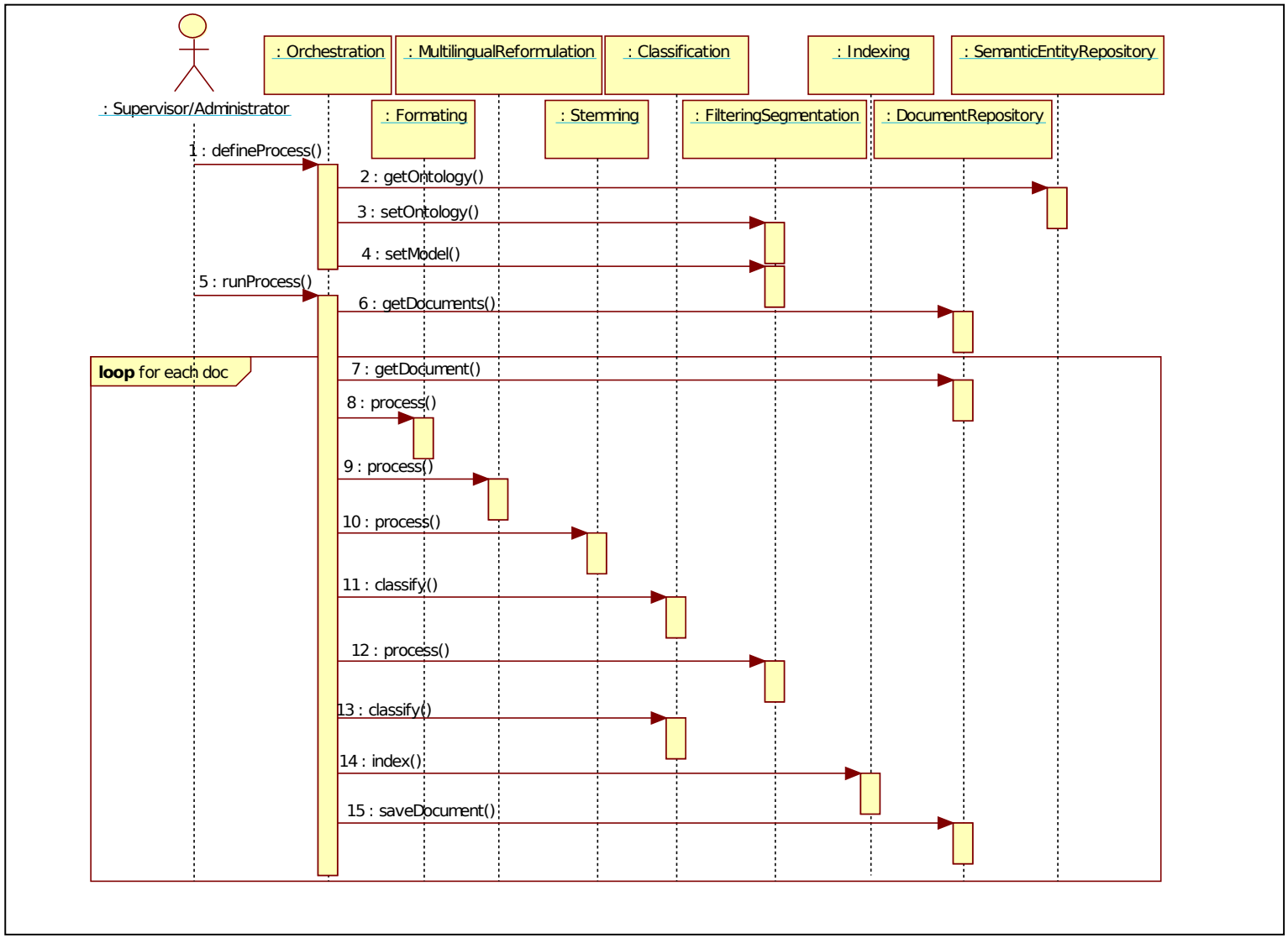
Orchestration de Services

- Pour réaliser un cas d'utilisation, divers services doivent en général être appelés
- Il faut définir le processus en spécifiant l'ordre exact d'invocation des services, les séquences, les alternatives, les itérations, etc.
- Langage BPEL
- Orchestra (BULL/OW2)
- Approche alternative : P2P, Chorégraphie de services (INRIA Gemo)



Exemple de Chaine

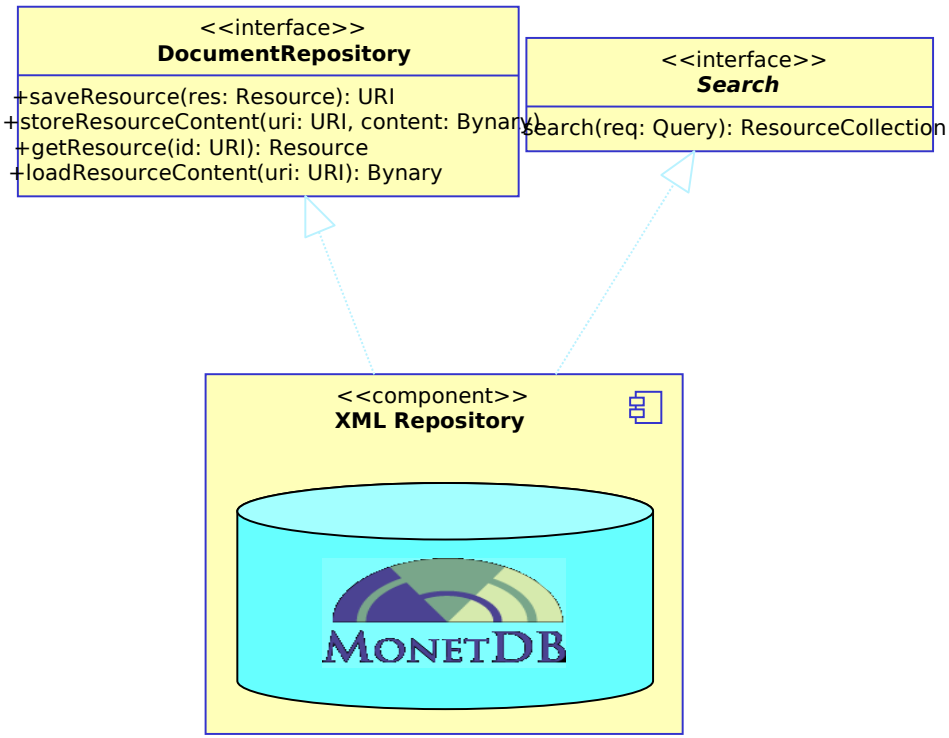
(Prétraitement et indexation)



Service d'Accès aux Données

Implémentation d'un entrepôt **XML**

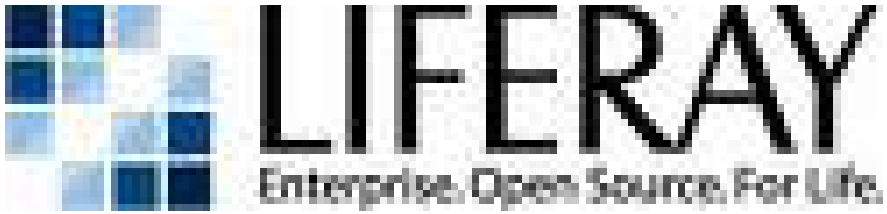
- **Stockage des ressources**
  - Descriptions des documents (XML)
  - Descriptions des services (WSDL)
  - Annotations (RDF/XML), stockables aussi dans un entrepôt RDF dédié
  - Ontologies (RDFS-OWL/XML)
  - Documents dans leurs formats natifs
- **Interrogation en XQuery**
- **Interface Web Service**
- **Implémentations Standard et P2P (INRIA Gemo)**



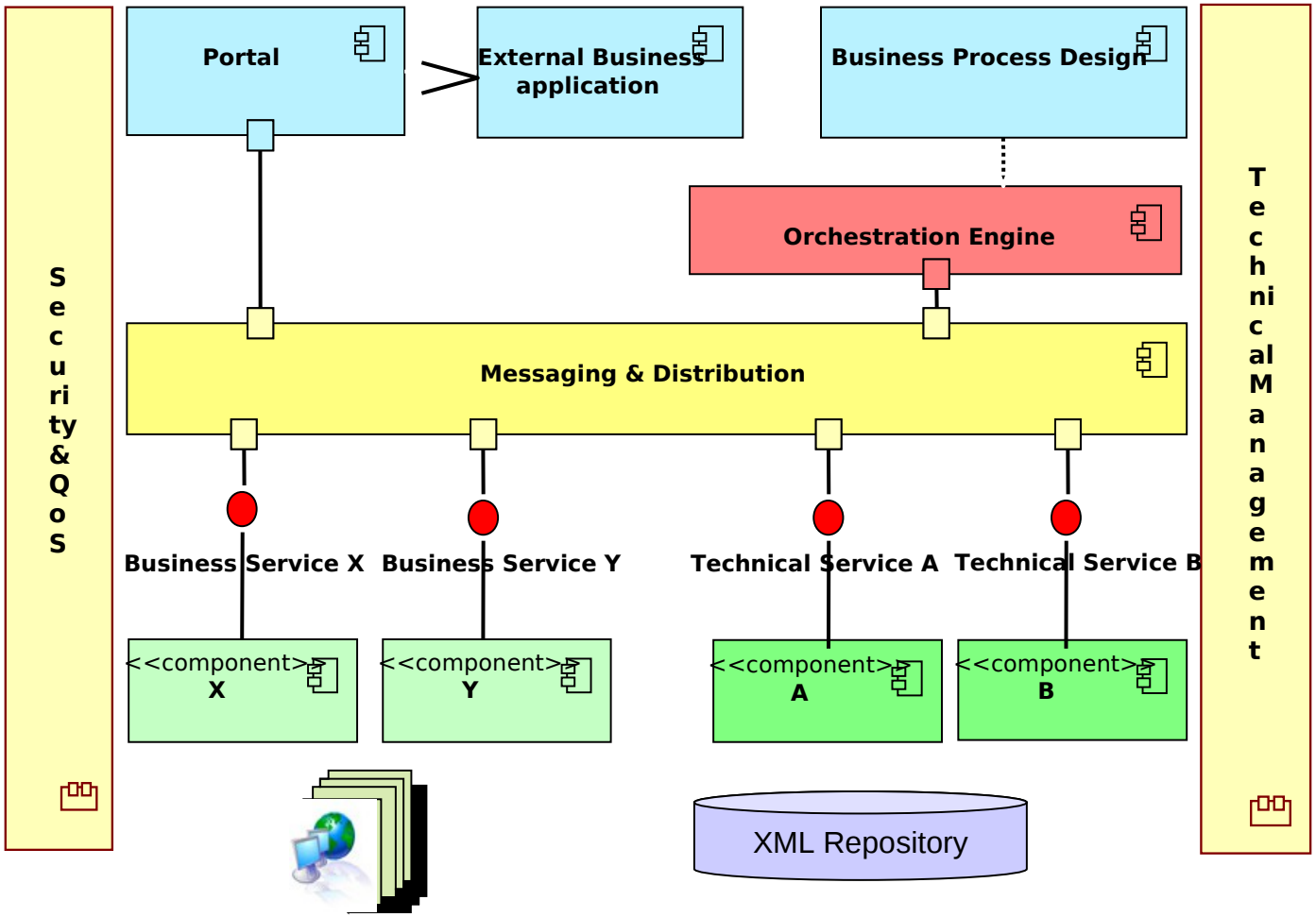
Accès aux services

Les interfaces graphiques des applications WebContent sont préférentiellement réalisées à l'aide d'un portail Web

- Organise et contrôle les composants IHM (portlets)
- Présente les données d'une manière organisée et unifiée
- Offre un accès unique à toutes les ressources disponibles
- Contrôle les accès et gère les utilisateurs
- Permet de personnaliser les espaces de travails selon
  - Les applications
  - Les profils utilisateurs



Technical architecture



Interopérabilité des Services

Pour être utilisés facilement dans une orchestration, les services doivent être **interopérables**:

- Techniquement
  - Normalisation des protocoles
  - Normalisation des formats de données
  - Normalisation des interfaces
- Sémantiquement
  - Normalisation des concepts et des relations caractérisant les données échangées
  - Normalisation des références sémantiques utilisées

Description d'un Contrat de Service

- Description: Service mission & responsibilities
- Catégorie: Reference(s) to the fonctional map
- Interface offerte:

Description de chaque opération ...

Exemple:

process

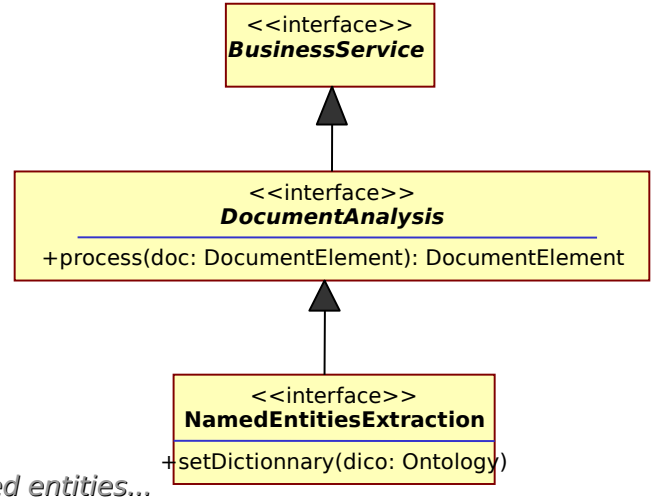
description: Detects the named entities in the input media unit....

input: media unit in the pivot format.

output: media unit in which are defined text fragments for the extracted entities...

pre-condition: the input media unit is in the exchange format.

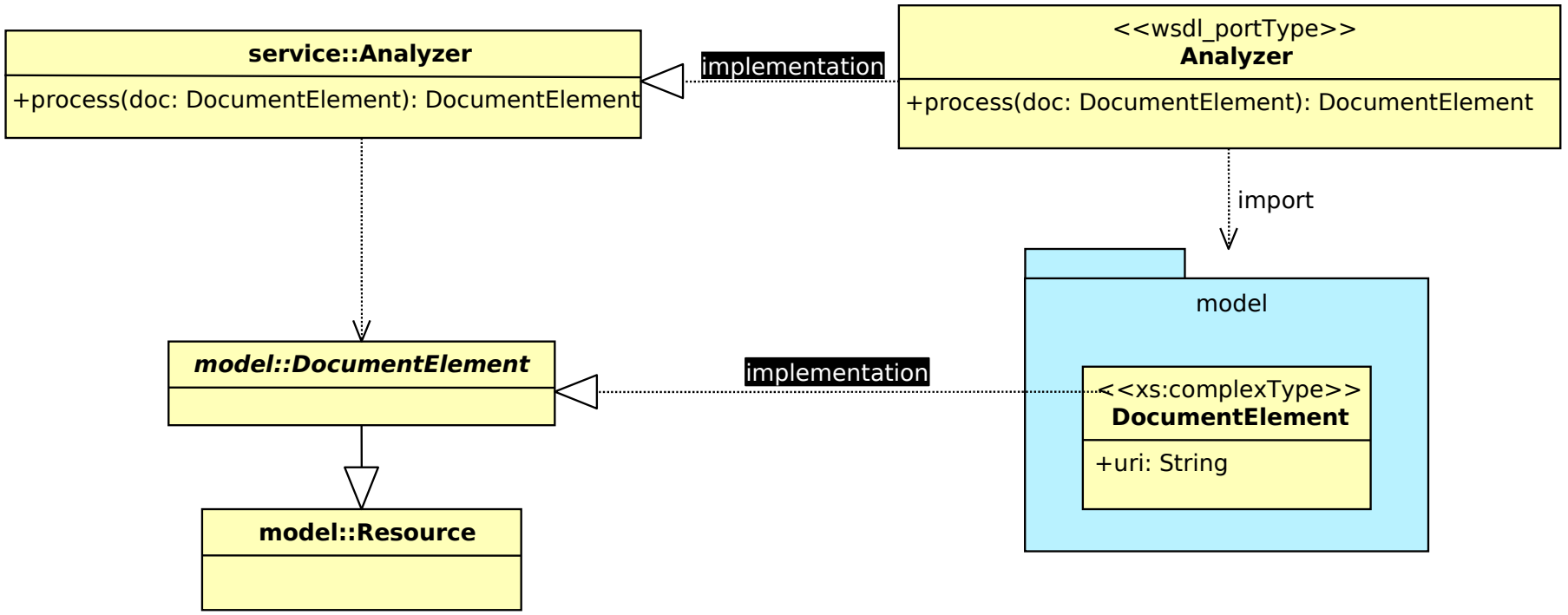
post-condition: none



- Ressources utilisées: données de configuration, paramètres, etc.
- Autres besoins: performances, limites de qualité, etc.
- Implémentation: identification des composants offrant le service

Traduction des Spécifications UML

Produit du WSDL et du XSD



Interopérabilité Sémantique

Pour assurer l'interopérabilité des services, il n'est pas suffisant de normaliser les protocoles et les interfaces

Il faut aussi définir l'organisation et la sémantique des données échangées

Modèle d'échange de données

- Chaque document WebContent peut être décrit par ce modèle
- Chaque document source est « placé en cache permanent » et reste accessible
- Le modèle est formalisé en UML (diagramme de classes) pour aider la transcription des concepts vers un langage de programmation

Fonctionnalités Nécessaires pour le Modèle

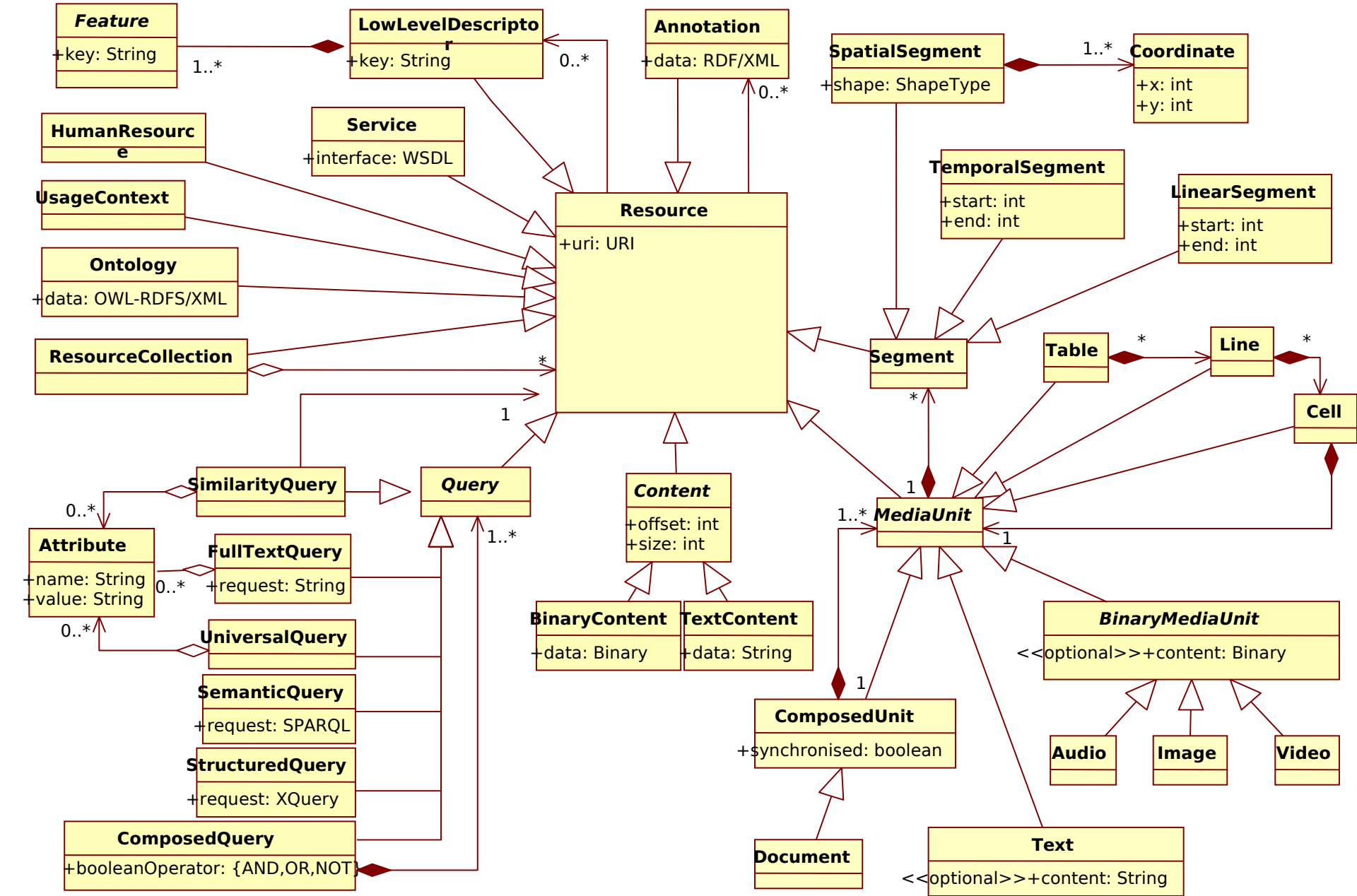
Le modèle doit permettre de définir

- Une référence commune pour tous les utilisateur
- Le contenu des messages échangés entre les services
- Un format pivot fondé sur XML

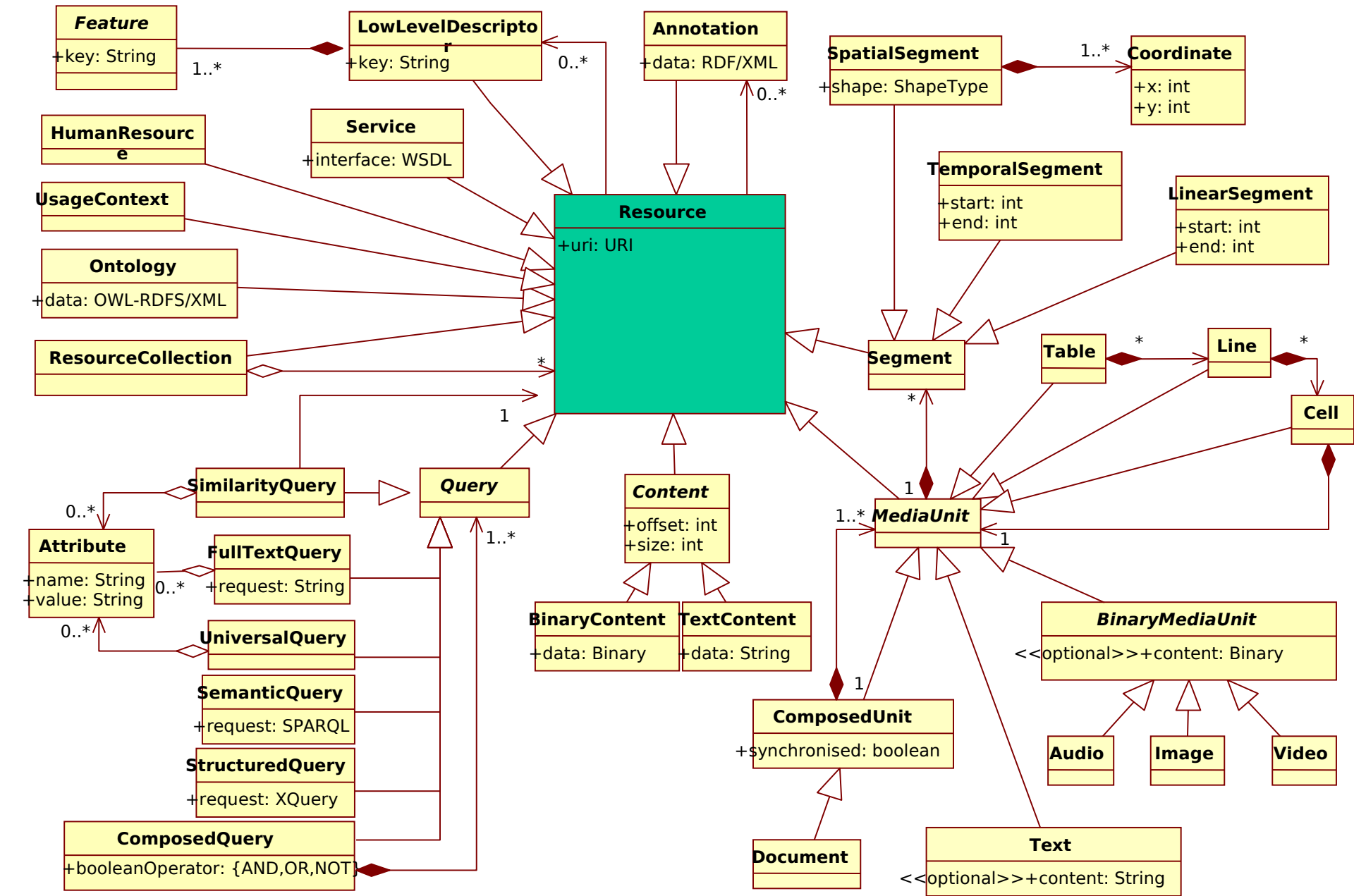
Le modèle doit fournir des mécanismes pour

- identifier chaque document
- annoter globalement un document non-structuré
- identifier les éléments extrais des contenus
- annoter chaque élément extrait
- annoter les autres ressources dans la plateforme

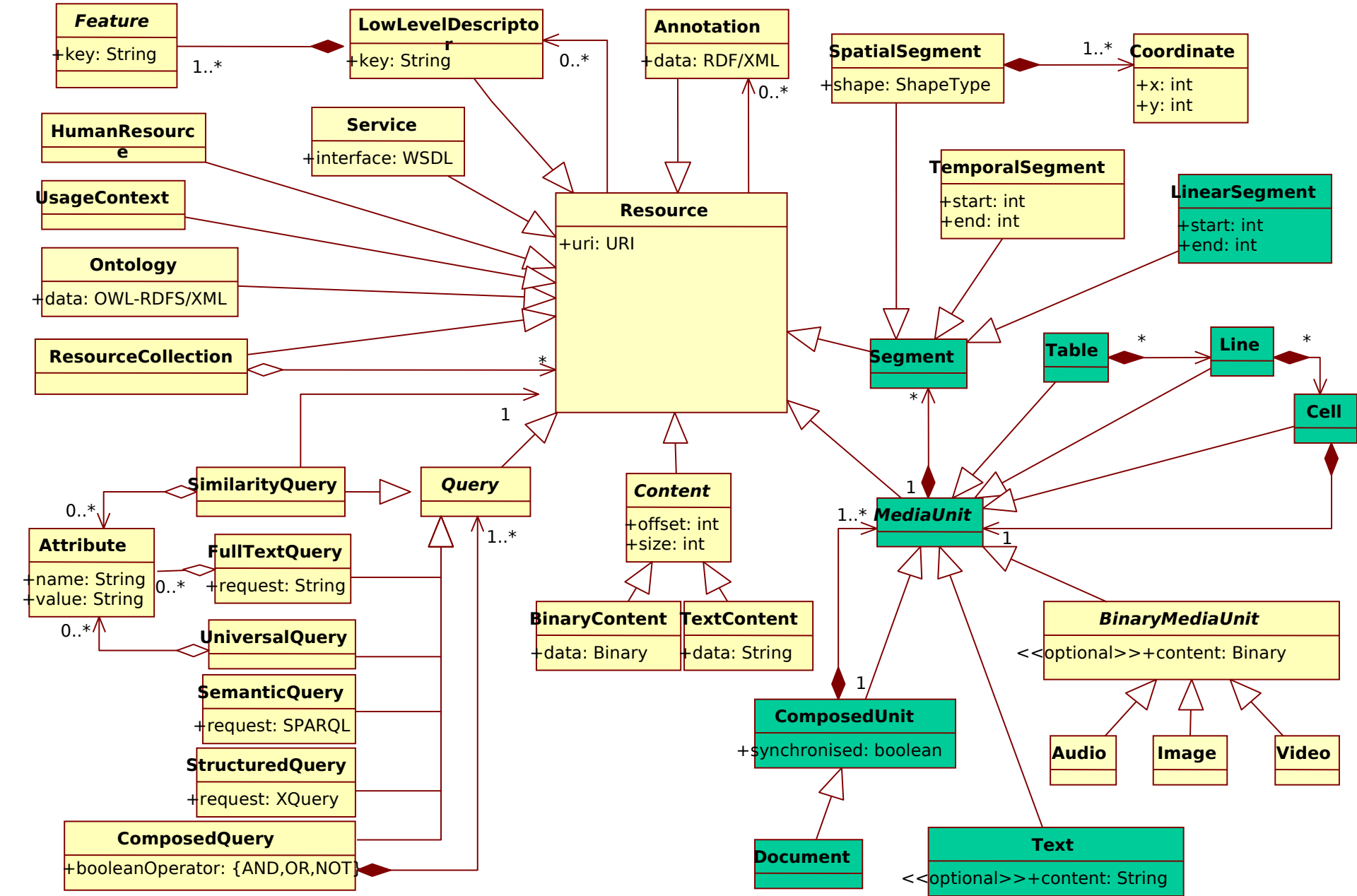
Le Modèle d'Echange WebContent



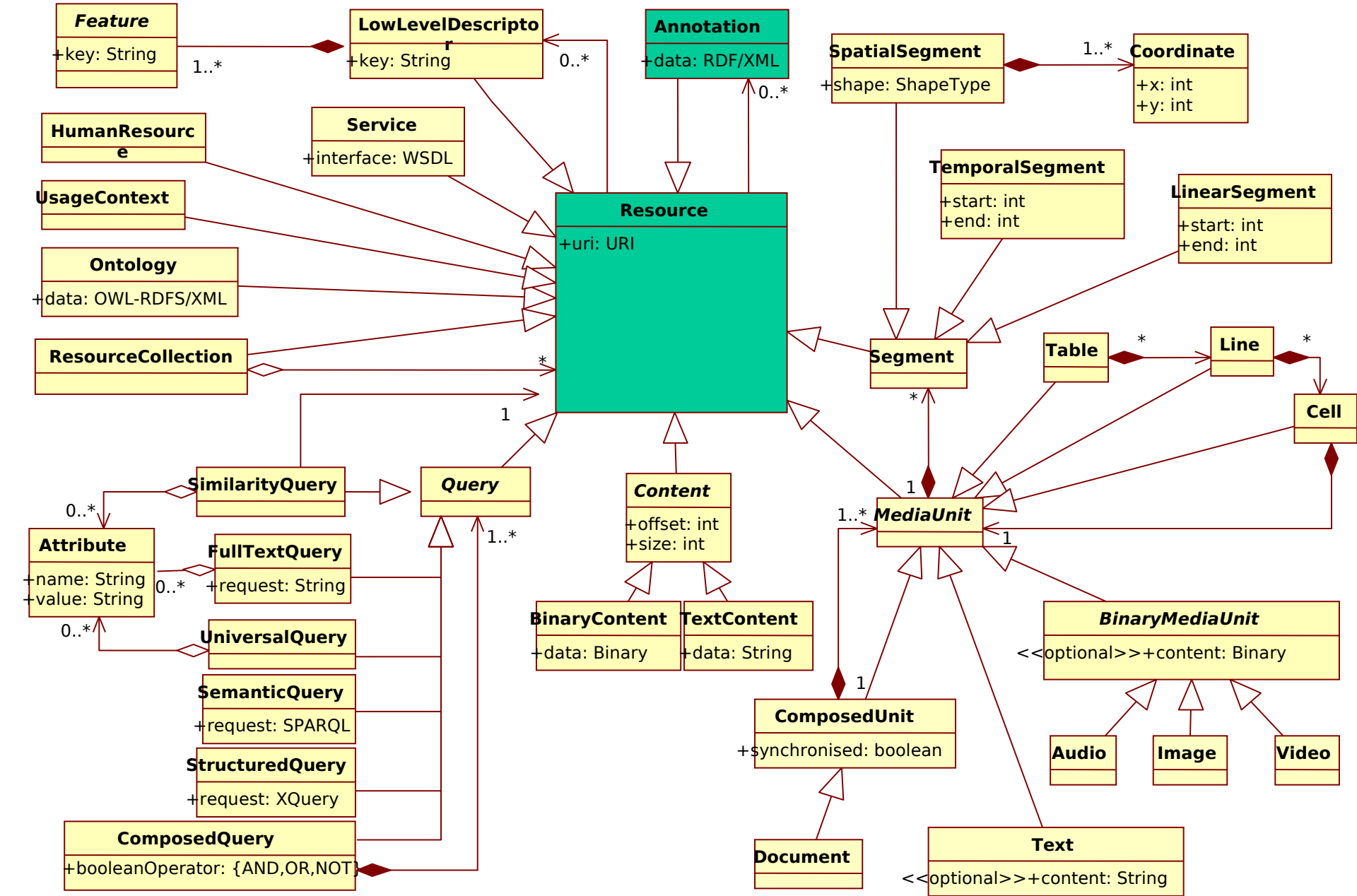
La Ressource



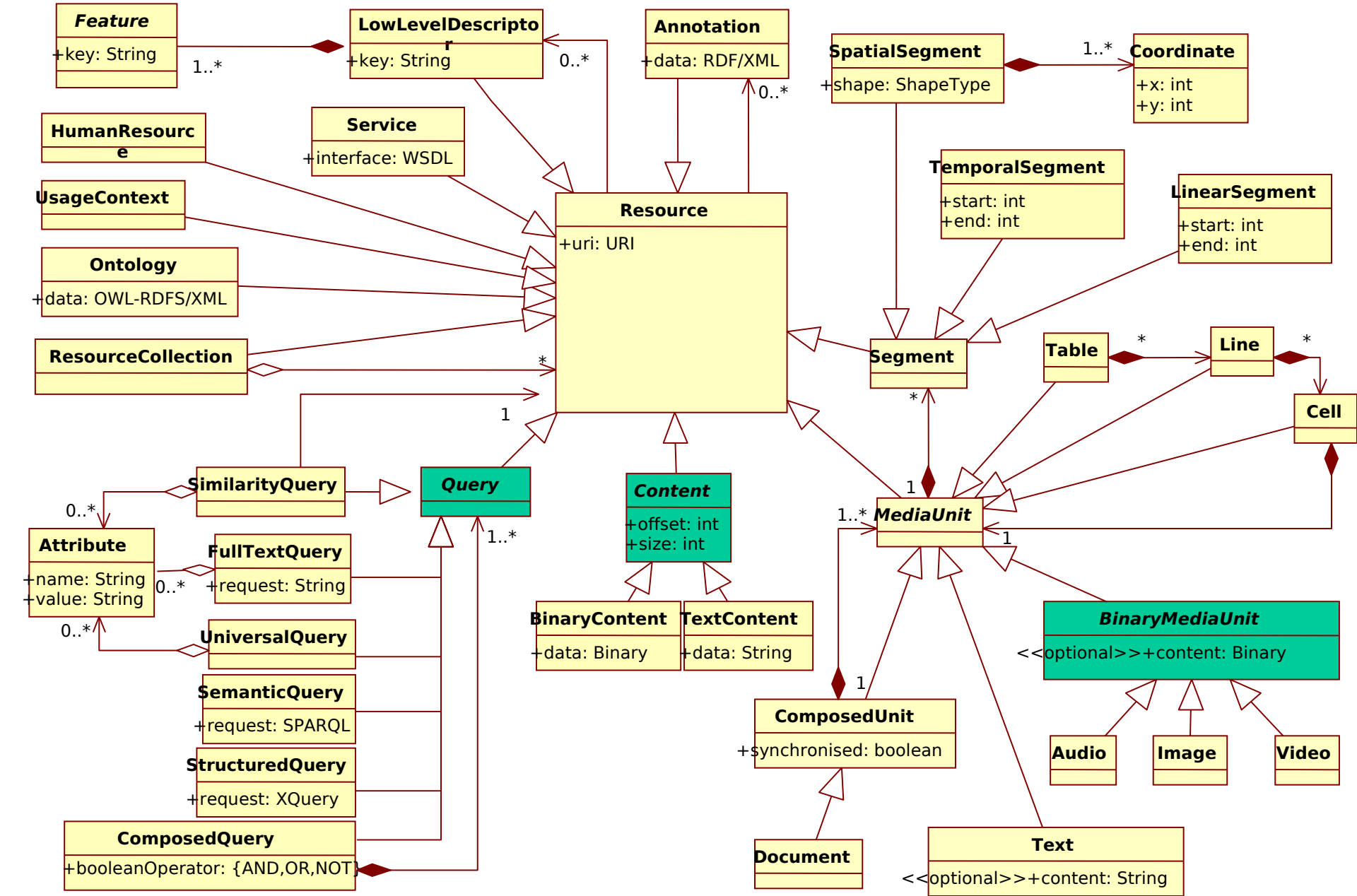
Structure de Document



Annotations



Autres Aspects



Exemple d'Instance de Document

myWS/myDocument

Ceci est l'introduction.

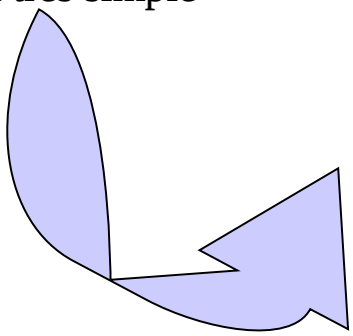
Chapitre 1

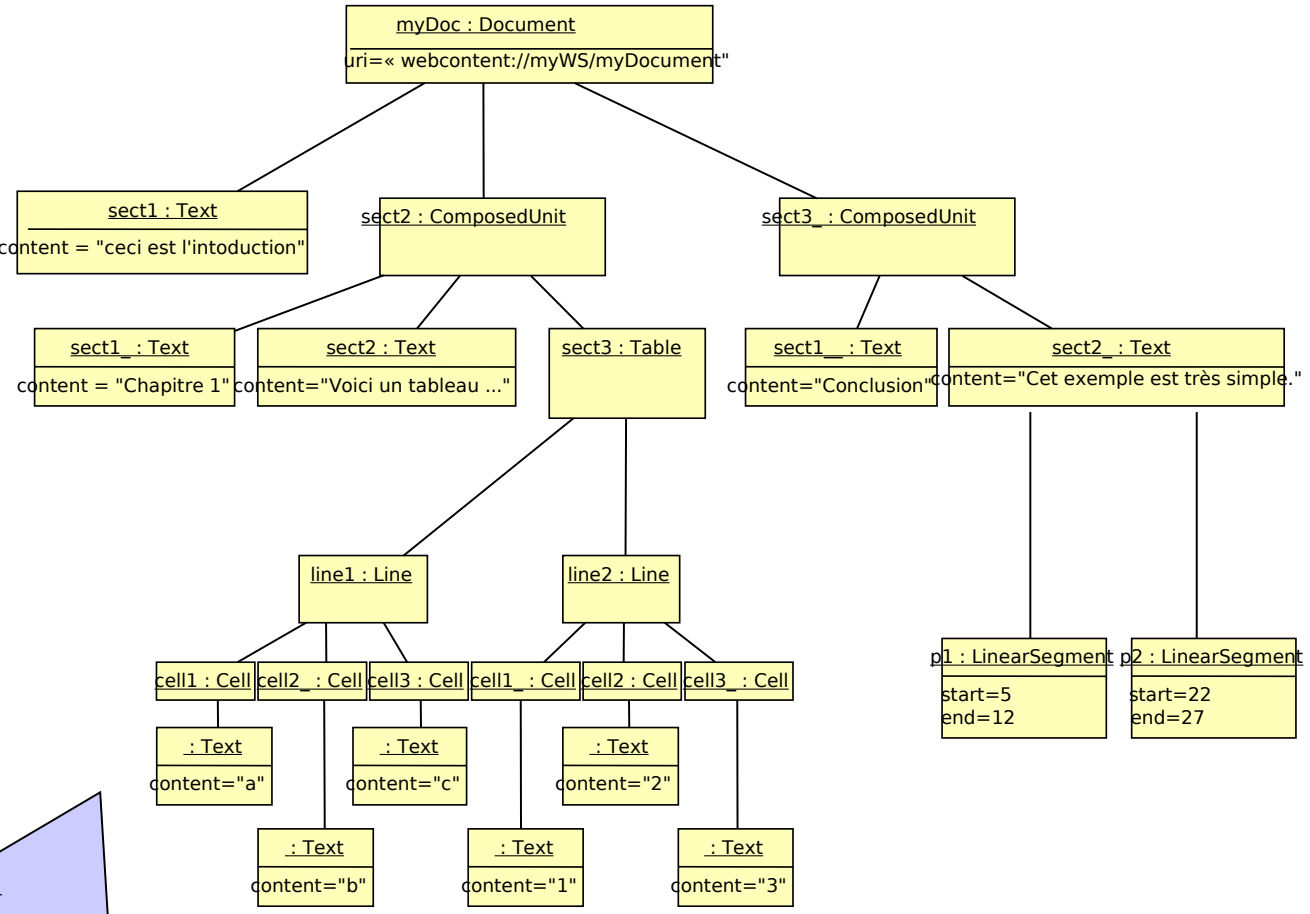
Voici un tableau...

a	b	c
1	2	3

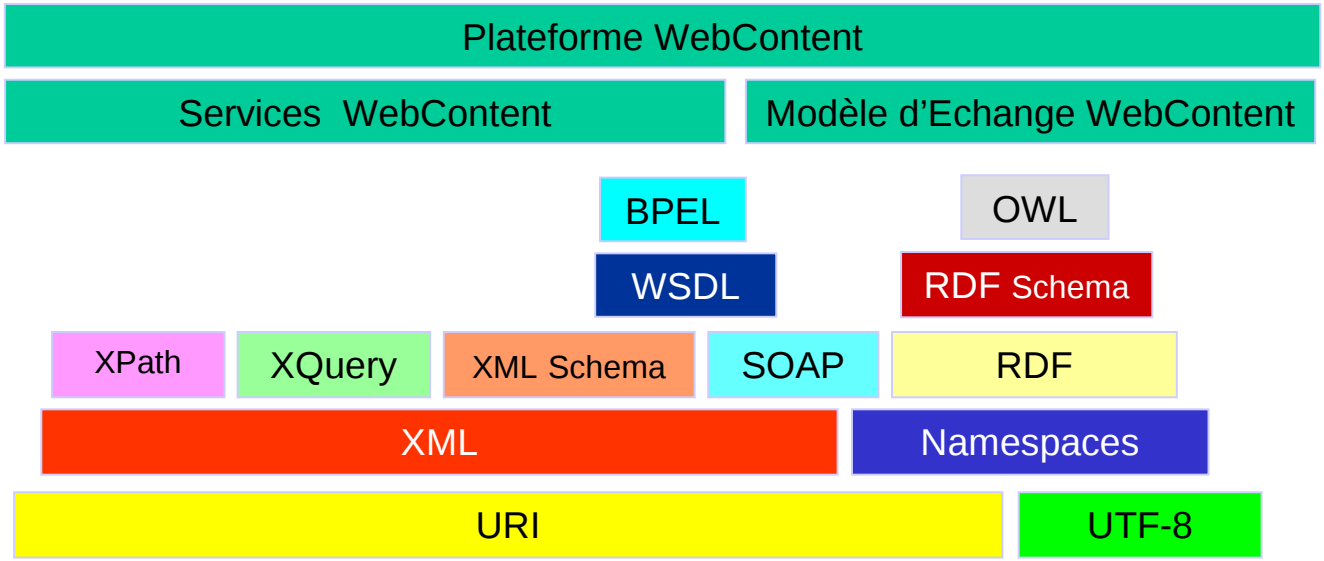
Conclusion

Cet exemple est très simple





Importance des Standards



Résultats Disponible

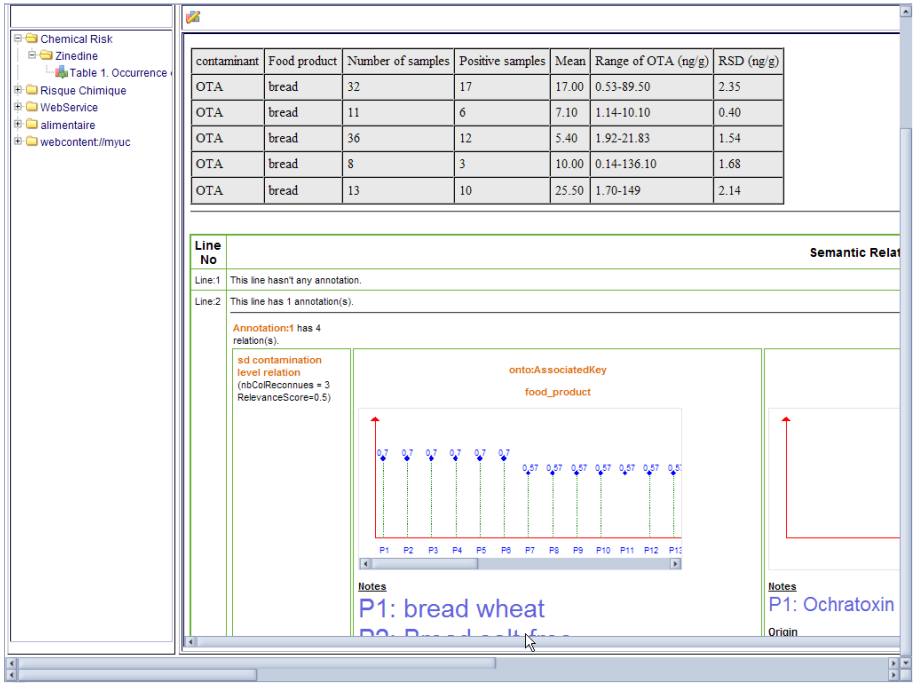
- Accessibles depuis http://www.webcontent-project.org
- Spécifications de services (XX)
  - Descriptions
  - APIs
- Modèle de données
  - Spécifications
  - Schémas XML
- Tutoriels Java pour réaliser, intégrer, déployer, orchestrer, tester, ... les services.
- Bibliothèques et tutoriels C++ pour réaliser et déployer des services (LGPL)

Services en Ligne

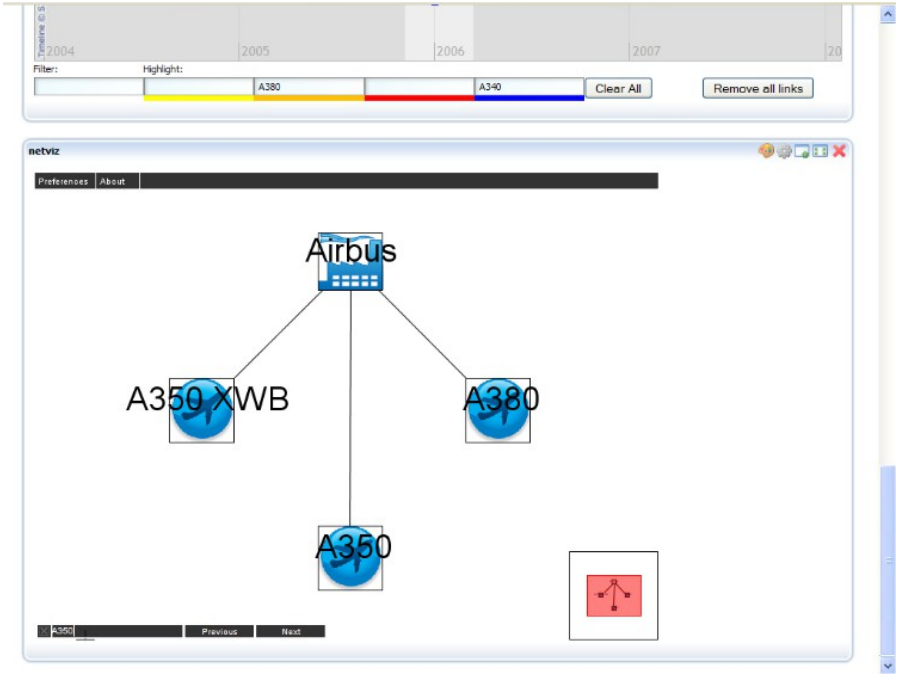
Service / Portlet	Implémenté par
Named entities	CEA LIST, EADS, Exalead
Crawling	CEA LIST, Exalead
Formating/Normalisation	CEA LIST, INRA, Exalead
SyntacticAnalysis	CEA LIST
Indexing	CEA LIST, EADS, Exalead, LIP6
Full text search	CEA LIST, EADS, Exalead, LIP6
Semantic search (SPARQL)	GEMO
Structured search (xQuery)	GEMO
Semantic annotation	CEA LIST, INRA, Exalead
Ontology enrichment	CEA LIST
Ontology alignment	EXMO
Language identification	EADS
Summarization	CEA LIST
Storage & Indexing in P2P	GEMO
Generic visualizations portlets	In Situ
Semantic Web data presentation	In Situ
Open source persistent storage	PRISM
Service Catalog	LIP6
Classification	LIP6

Applications en cours de Finalisation

@Web de l'INRA pour Bongrain



EADS pour Airbus



Actualité et Futur de la Plateforme WebContent

- Support du cœur par EADS au travers de son offre WebLab
- Développement collaboratif des bibliothèques C++ (sur gna.org)
- Interopérabilité WebContent/UIMA
- Utilisation dans de nombreux projets français ou européens : Vitalas , Virtuoso, VIGIEs, Scribo,...
- A finaliser encore:
  - mode collaboratif de développement et de support du cœur
  - rapprochement avec un organisme tel que Oasis, W3C ou OW2
  - standardisation des spécifications
- A l'automne: workshop

Conclusion

La plateforme permet de développer des applications

- Traitant de **gros volumes** de données hétérogènes et non-structurées
- Dans des **domaines variés**
  - Veilles technique ou stratégique
  - Veille économique
  - Open Source Intelligence (OSINT)
  - Entreprise Information Portals (EIP)
  - Content Management Systems (CMS)
  - Indexation d'archives multimédia
  - Gestion de connaissances « Métier »
  - etc.
- Fondé sur les **technologies du Web Sémantique**