Quality Awareness in Data Management and Mining

Laure BERTI-ÉQUILLE

Soutenance pour l'Habilitation à Diriger de Recherches
IRISA - Université de Rennes 1

25 Juin 2007

Plan de l'exposé

Plan

Formation doctorale

1996 : Université de Paris IX-Dauphine

- Diplôme d'Étude Approfondie d'Informatique

1996-1999 : Université de Toulon et du Var

- Doctorat d'Informatique: *Qualité des données et leur recommandation : application à la veille technologique*
- Moniteur C.I.E.S.

Formation post-doctorale

1999-2000 : Université d'Avignon et Pays du Vaucluse

- Attachée Temporaire à l'Enseignement et à la Recherche

Fonction actuelle

2000 - aujourd'hui : Université de Rennes 1 - IRISA

- Maître de Conférences

Spécialités enseignées à l'Université de Rennes 1

Bases de données	DIIC2 et MPRO2
Bases de données avancées	MPRO2 TC
Entrepôts de données	MPRO2 MIAGE
Technologies XML	MPRO1 MIAGE
Analyse-conception objet	MPRO2 MIAGE
Conduite de projet	MPRO1-2 MIAGE

Détails disponibles à http://www.irisa.fr/Laure.Berti-Equille/Enseignement.html

Quelques chiffres

Publications depuis 1996

2 chapitres de livres et 3 actes édités
5 revues internationales et 7 revues nationales
15 conférences et 6 ateliers internationaux
7 conférences et 2 ateliers nationaux

53% en unique auteur

Encadrement

1 thèse soutenue et une en cours
1 ingénieur expert
1 étude post-doctorale en cours
4 stages de D.E.A. ou Masters de recherche et un internship
2 participations à des jurys de thèse (rapporteur)

Quelques chiffres

Publications depuis 1996

2 chapitres de livres et 3 actes édités
5 revues internationales et 7 revues nationales
15 conférences et 6 ateliers internationaux
7 conférences et 2 ateliers nationaux

53% en unique auteur

Encadrement

1 thèse soutenue et une en cours
1 ingénieur expert
1 étude post-doctorale en cours
4 stages de D.E.A. ou Masters de recherche et un internship
2 participations à des jurys de thèse (rapporteur)

Coordination

- **Projet intégré européen (FP-6)**
  ENTHRONE Phase 1, 2003-2005, Coordinatrice INRIA Rennes
- **Projets internationaux**
  - CLINIQ, PHC Italie, Università La Sapienza - IStat, 2006
  - M4, PHC Japon, National Institute of Informatics, 2002
- **Projet national (ANR)**
  QUADRIS, ANR-05-MMSA, Coordinatrice, 2006-2009

Contrats et collaborations

- **Responsabilité scientifique**
  - Contrat avec Genielog, 2005-2006
  - Contrat avec Écoles Militaires de Coëtquidan, 2003-2008
- **Participation**
  Projet inter-EPST avec l'INSERM U522, 2002-2003

Coordination

- **Projet intégré européen (FP-6)**
  ENTHRONE Phase 1, 2003-2005, Coordinatrice INRIA Rennes
- **Projets internationaux**
  - CLINIQ, PHC Italie, Università La Sapienza - IStat, 2006
  - M4, PHC Japon, National Institute of Informatics, 2002
- **Projet national (ANR)**
  QUADRIS, ANR-05-MMSA, Coordinatrice, 2006-2009

Contrats et collaborations

- **Responsabilité scientifique**
  - Contrat avec Genielog, 2005-2006
  - Contrat avec Écoles Militaires de Coëtquidan, 2003-2008
- **Participation**
  Projet inter-EPST avec l'INSERM U522, 2002-2003

Organisation

- **Deux premières éditions de l'atelier national**
  Qualité des données et des connaissances (DKQ)
  en conjonction avec EGC, Paris et Lille, janvier 2005 et 2006
- **Seconde édition du workshop international**
  Information Quality in Information Systems (IQIS)
  en conjonction avec ACM SIGMOD, Baltimore, USA, juin 2005

Participation

- Membre de comités d'organisation :
  BDA'05, JOBIM'02, EDD'01, INFORSID'98
- Membre de comités de programme :
  21 comités de programmes depuis 2005 dont VLDB'07
- Membre de comités éditoriaux de revues internationales :
  - *International Journal of Information Quality (IJIQ)*
  - *Journal of Digital Information Management (JDIM)*

Organisation

- **Deux premières éditions de l'atelier national**
  Qualité des données et des connaissances (DKQ)
  en conjonction avec EGC, Paris et Lille, janvier 2005 et 2006
- **Seconde édition du workshop international**
  Information Quality in Information Systems (IQIS)
  en conjonction avec ACM SIGMOD, Baltimore, USA, juin 2005

Participation

- Membre de comités d'organisation :
  BDA'05, JOBIM'02, EDD'01, INFORSID'98
- Membre de comités de programme :
  21 comités de programmes depuis 2005 dont VLDB'07
- Membre de comités éditoriaux de revues internationales :
  - *International Journal of Information Quality (IJIQ)*
  - *Journal of Digital Information Management (JDIM)*

Plan

Problèmes de qualité des données

Au niveau de la structure

- X Valeurs manquantes
- X Violation de contraintes de domaines
- X Violation de contraintes d'intégrité référentielle
- X Doublons exacts

- X Données catégorielles fausses
- X Données périmées
- X Données incohérentes
- X Conflits de nommage
- X Conflits structurels

Problèmes de qualité des données

Au niveau de la structure

- √ Valeurs manquantes
- √ Violation de contraintes de domaines
- √ Violation de contraintes d'intégrité référentielle
- √ Doublons exacts

- X Données catégorielles fausses
- X Données périmées
- X Données incohérentes
- X Conflits de nommage
- X Conflits structurels

Problèmes de qualité des données

Au niveau de la structure

- √ Valeurs manquantes
- √ Violation de contraintes de domaines
- √ Violation de contraintes d'intégrité référentielle
- √ Doublons exacts

- X Données catégorielles fausses
- X Données périmées
- X Données incohérentes
- X Conflits de nommage
- X Conflits structurels

Problèmes de qualité des données

Au niveau des instances

- X Données non standardisées
- X Valeurs incomplètes
- X Données erronées ou aberrantes
- X Erreurs typographiques
- X Valeurs imbriquées
- X Valeurs ou attributs mal renseignés
- X Données ambigües ou contradictoires
- X Doublons approximatifs

Problèmes de qualité des données

Au niveau des instances

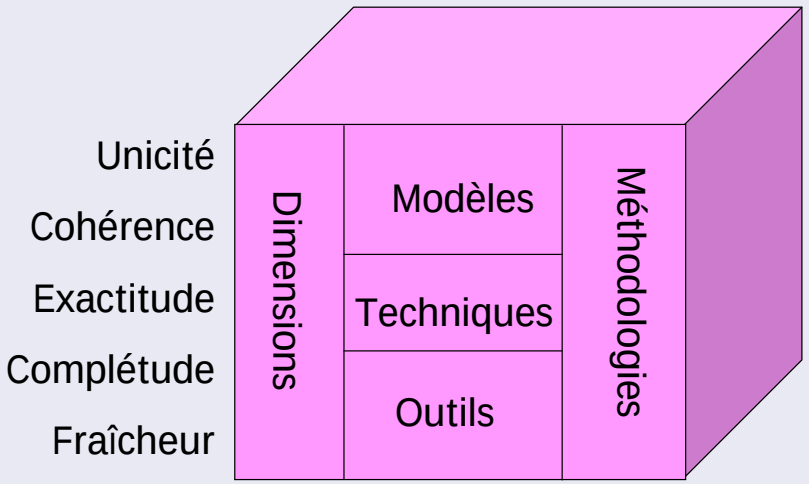
- X Données non standardisées
- X Valeurs incomplètes
- X Données erronées ou aberrantes
- X Erreurs typographiques
- X Valeurs imbriquées
- X Valeurs ou attributs mal renseignés
- X Données ambigües ou contradictoires
- X Doublons approximatifs

Travaux de recherche

Aux confluents de plusieurs disciplines

- Statistiques
- Bases de Données et Systèmes d'Information
- Gestion de projet
- Ingénierie des connaissances

Selon 5 modalités



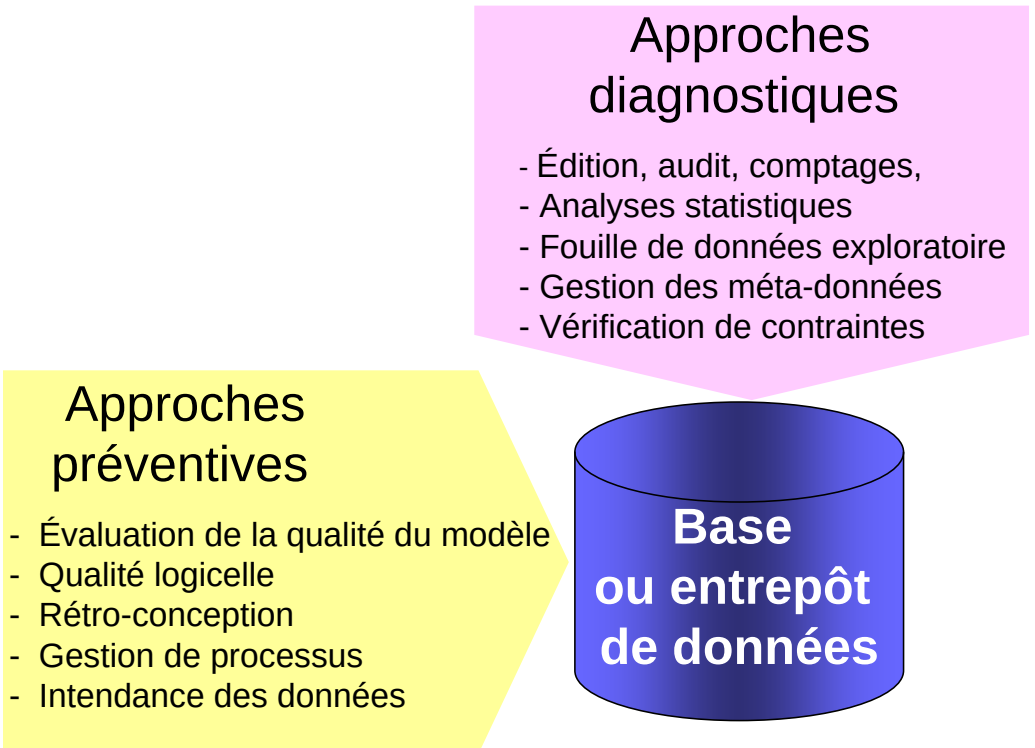
Principales approches

Principales approches

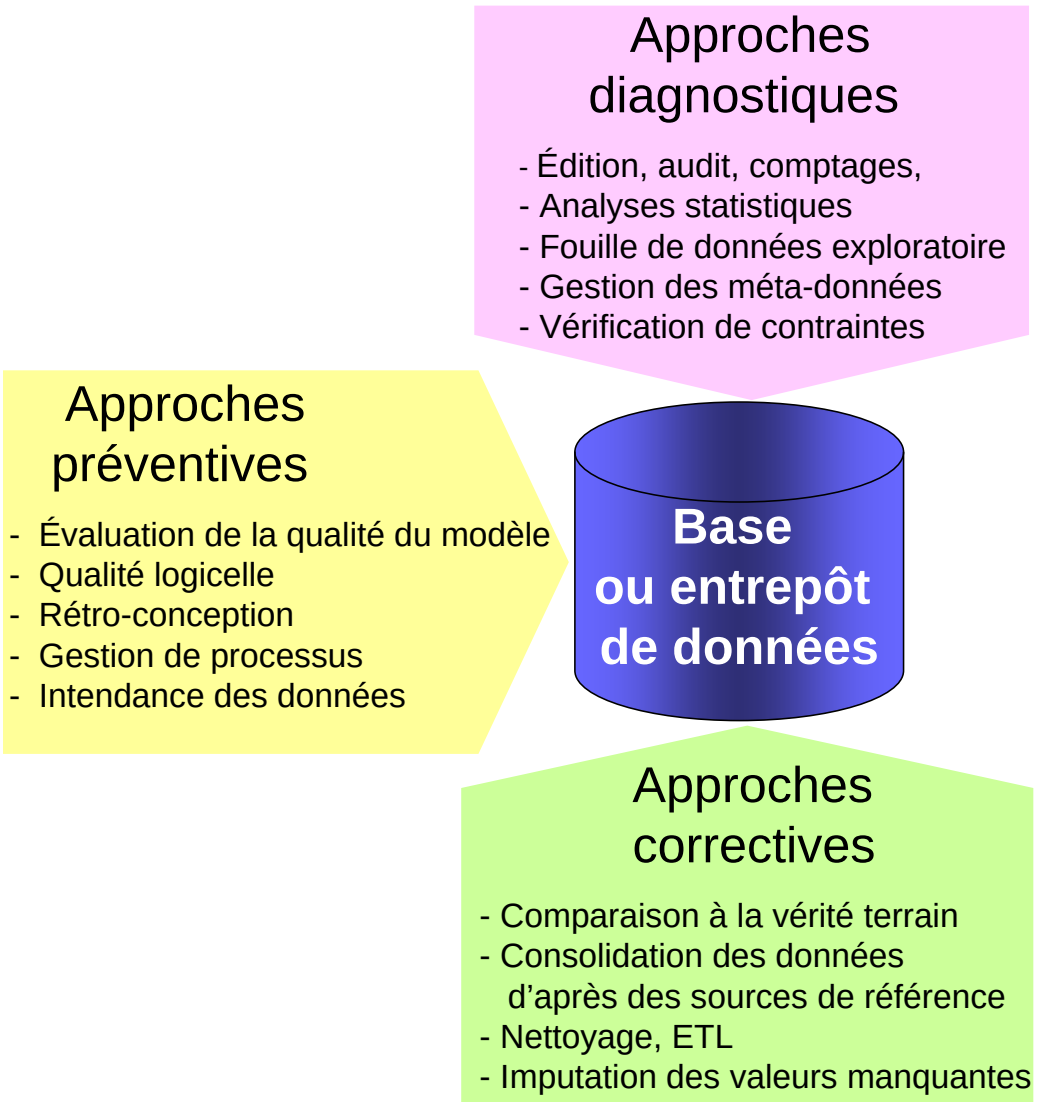
Approches préventives

- Évaluation de la qualité du modèle
- Qualité logicelle
- Rétro-conception
- Gestion de processus
- Intendance des données

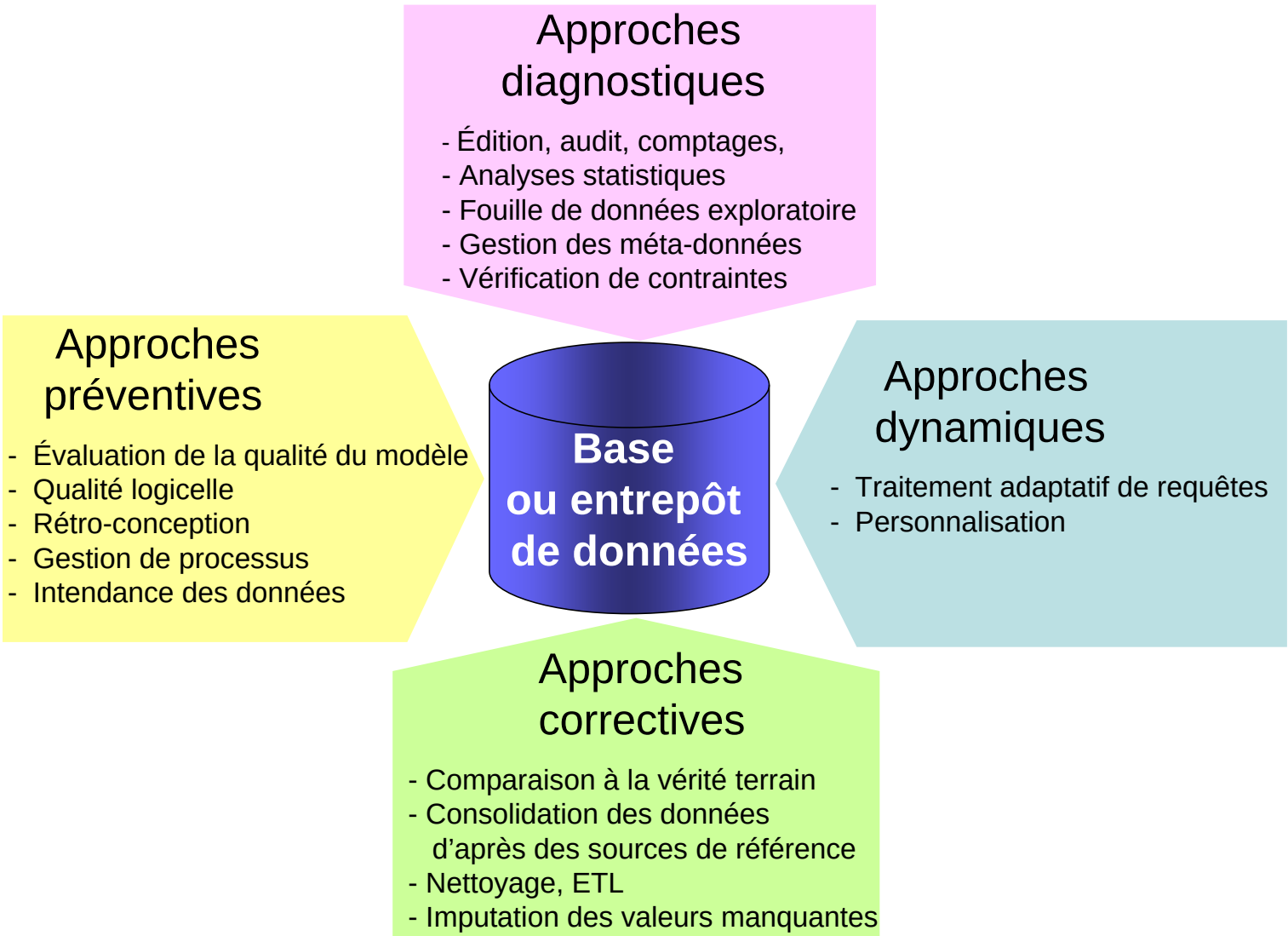
Principales approches



Principales approches



Principales approches



Principaux défis

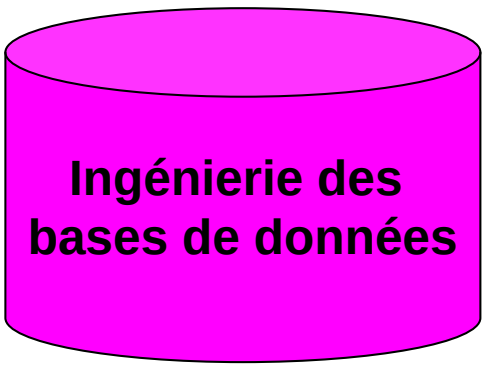
- **Au niveau méthodologique**
  - Unification et standardisation
  - Benchmarks
- **Au niveau de l'ingénierie des systèmes d'information**
  - Patterns d'architecture pour intégrer le contrôle de la qualité
- **Au niveau des langages (LDD et LMD)**
  - Déclaration et gestion intégrée des méta-données
  - Développement et optimisation de langages étendus
- **Au niveau algorithmique**
  - Volumétrie des données et méta-données
  - Indexation des données et méta-données associées
  - Optimisation des calculs de méta-données statistiques
  - Prise en compte dynamique de la qualité dans le traitement des données

Approche adoptée

Apports mutuels de deux disciplines

Axe 1 Utilisation de techniques de fouille de données pour évaluer la qualité des données

Axe 2 Exploitation des méta-données décrivant la qualité des données pour évaluer et conforter la qualité des connaissances extraites à des fins décisionnelles

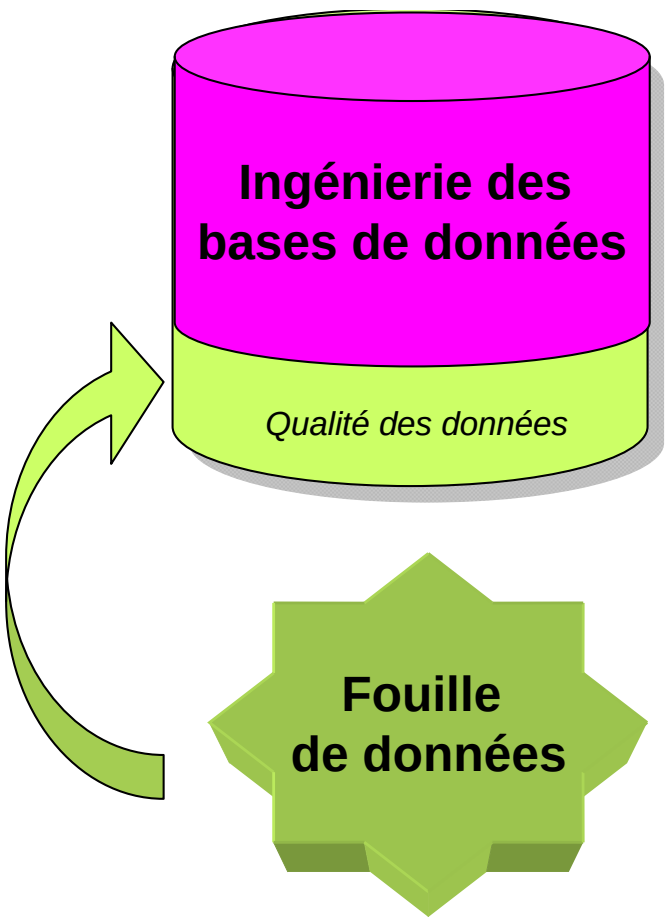


Approche adoptée

Apports mutuels de deux disciplines

Axe 1 Utilisation de techniques de fouille de données pour évaluer la qualité des données

Axe 2 Exploitation des méta-données décrivant la qualité des données pour évaluer et conforter la qualité des connaissances extraites à des fins décisionnelles

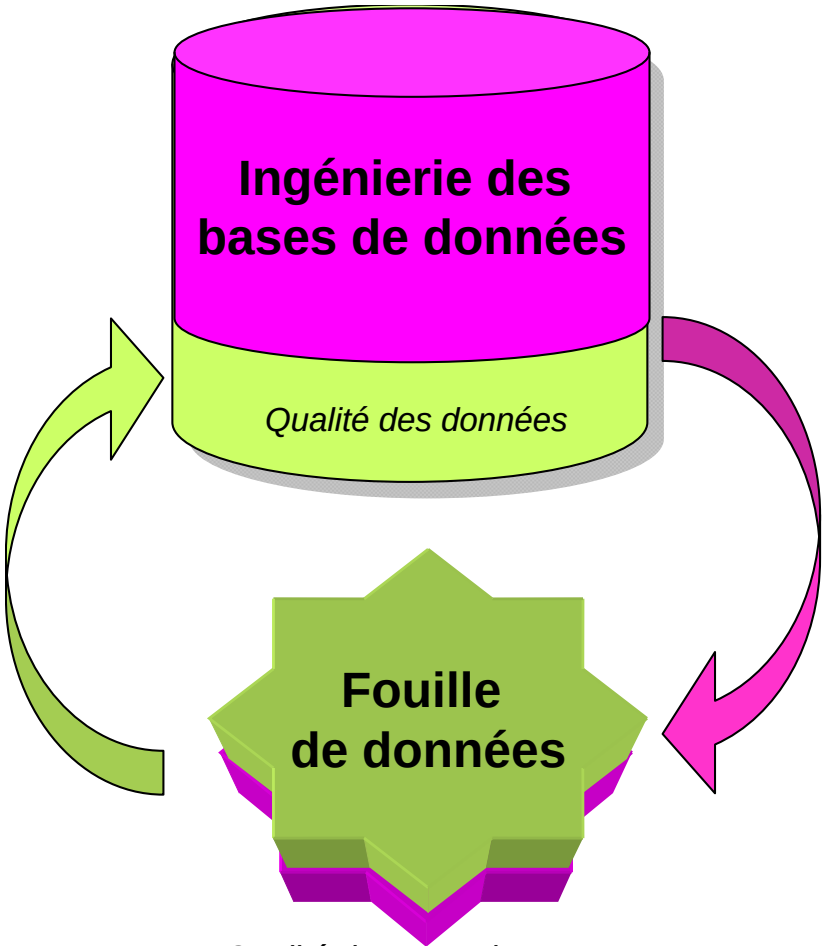


Approche adoptée

Apports mutuels de deux disciplines

Axe 1 Utilisation de techniques de fouille de données pour évaluer la qualité des données

Axe 2 Exploitation des méta-données décrivant la qualité des données pour évaluer et conforter la qualité des connaissances extraites à des fins décisionnelles



Axe 1 : Prise en compte de la qualité dans la gestion des données

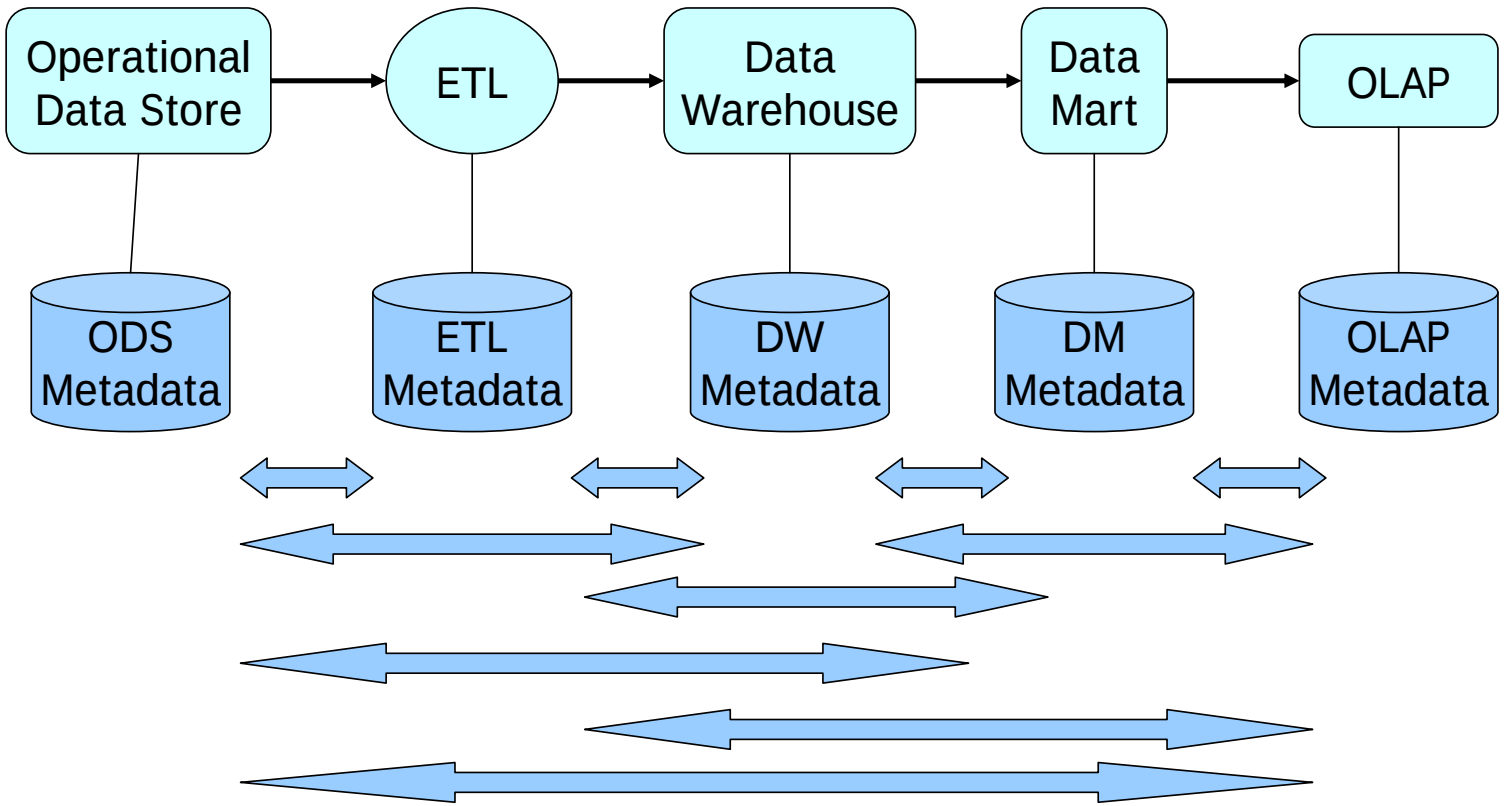
Objectif : Calculer et gérer des méta-données décrivant des facteurs mesurables de la qualité des données

Contributions :

1. Modélisation des méta-données et gestion conjointe des données et méta-données
2. Utilisation et adaptation de méthodes statistiques et de techniques de fouille de données spécifiques à la détection d'anomalies sur les données
3. Extension d'un langage pour exploiter ces méta-données lors du requêtage des données

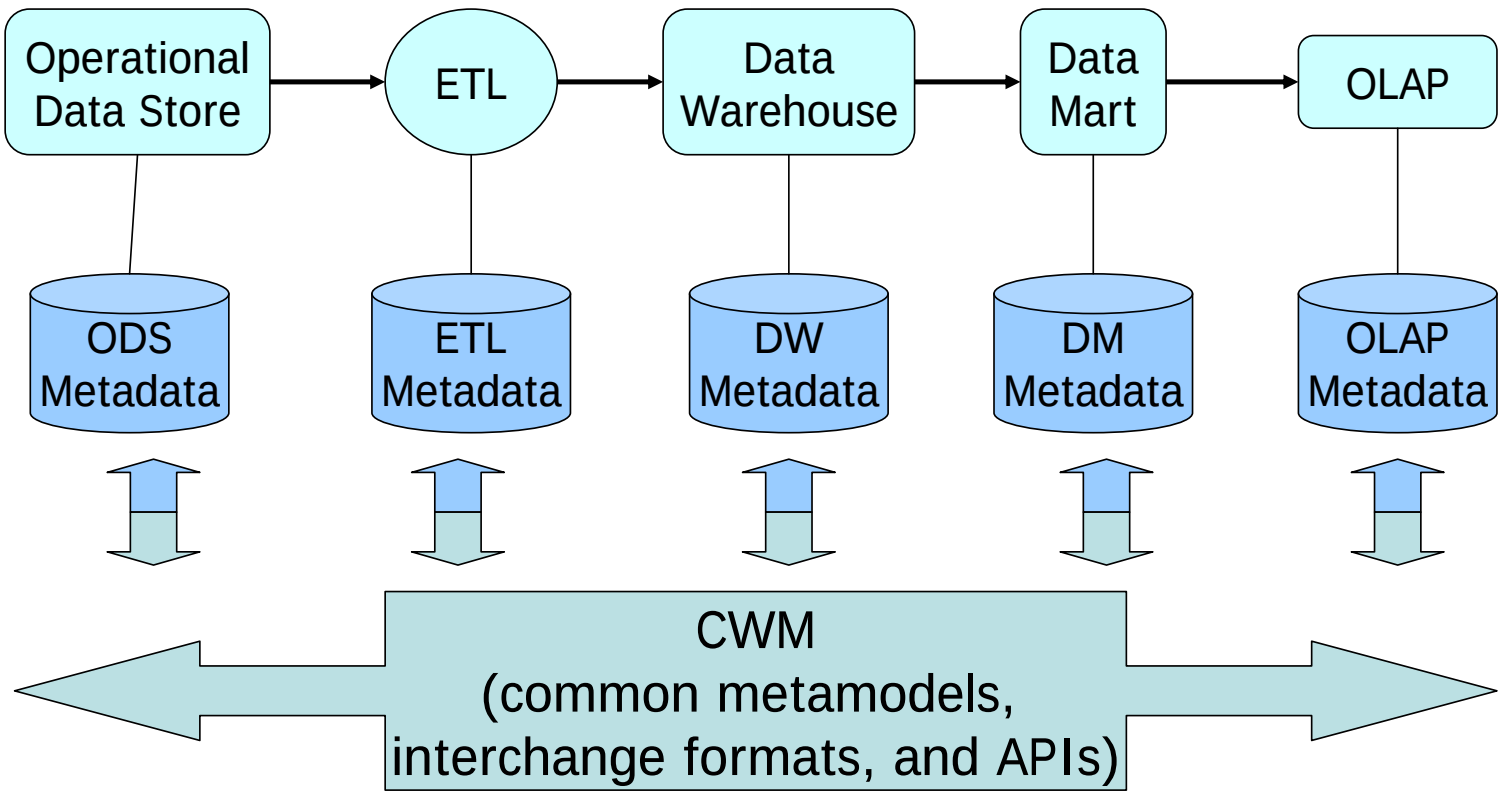
Extension de CWM - *Common Warehouse Metamodel* (OMG)

Problème d'intégration des méta-données : $\frac{n \times (n-1)}{2}$ échanges

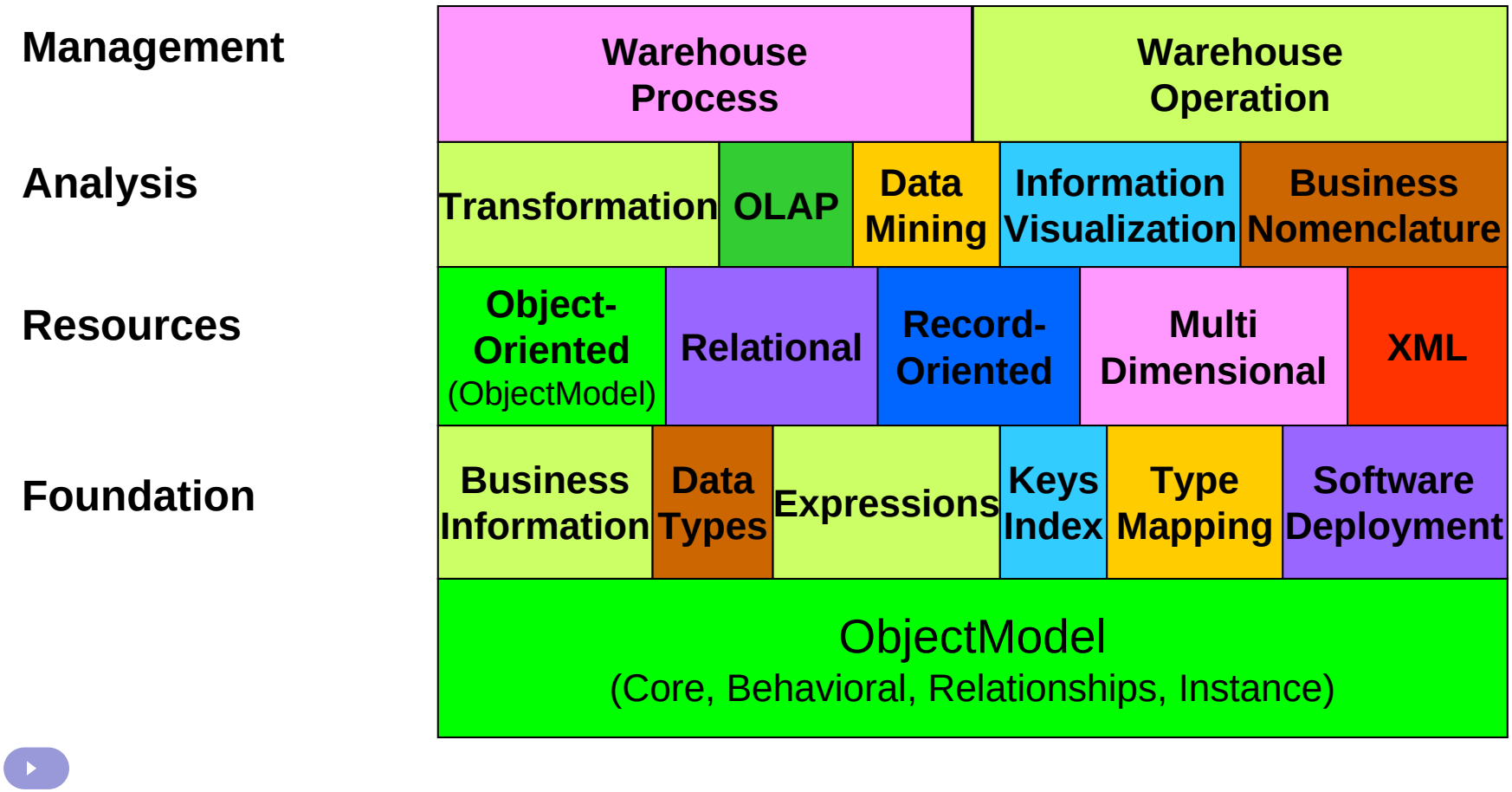


Extension de CWM - *Common Warehouse Metamodel* (OMG)

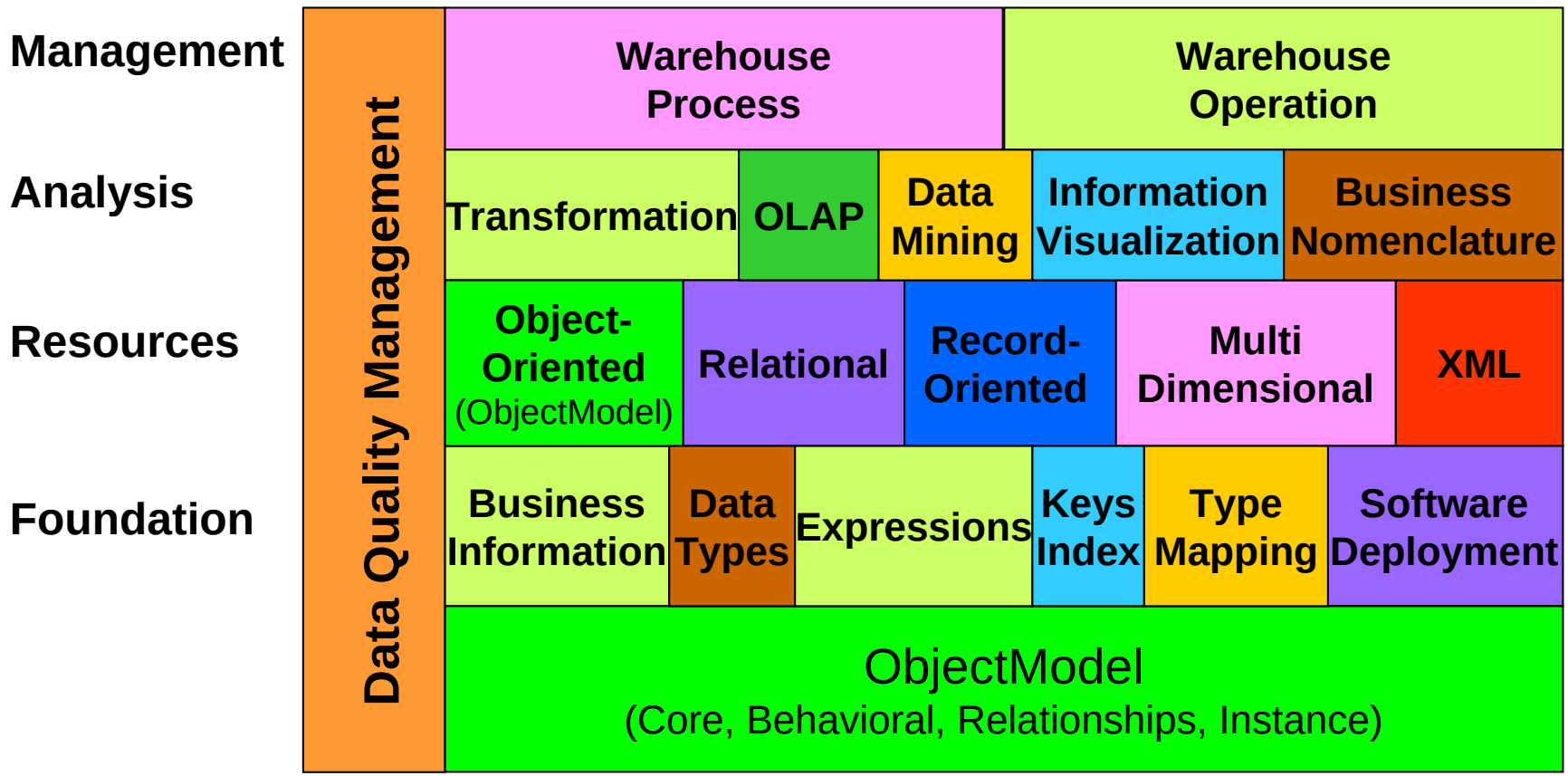
n adaptateurs CWM pour l'intégration des méta-données



Packages de CWM



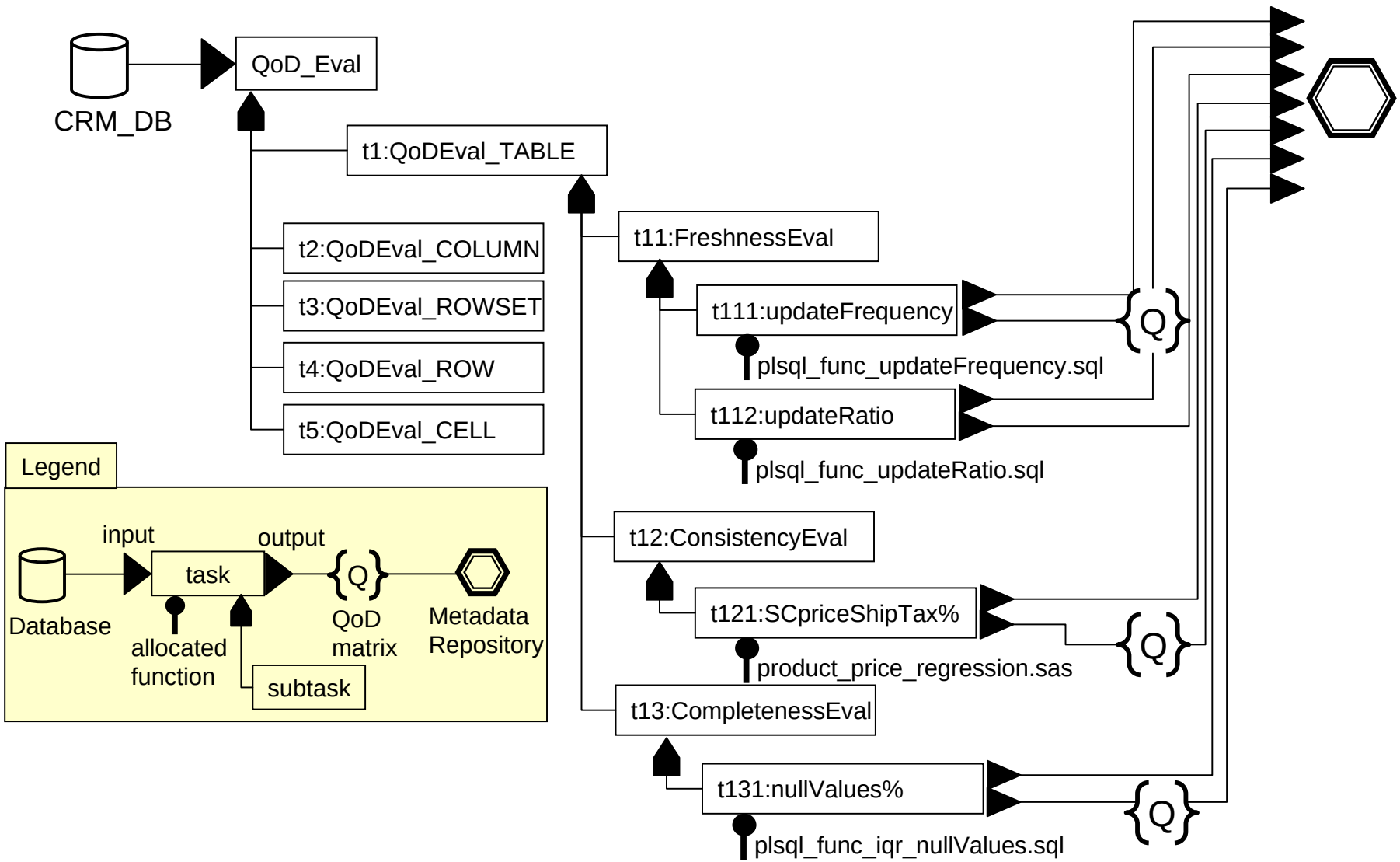
Packages de CWM



Calcul de méta-données par des fonctions analytiques

- Recensement des fonctions utiles pour mesurer des facteurs de la qualité des données pour différents niveaux de granularité
  - **I :** Fonctions de profilage
  - **II :** Fonctions utilisant des contraintes notamment statistiques
  - **III :** Fonctions utilisant des résumés, histogrammes et techniques d'échantillonnage
  - **IV :** Fonctions utilisant des techniques de fouille de données
- Composition de fonctions au sein de *workflows analytiques*
- Stockage et indexation des méta-données

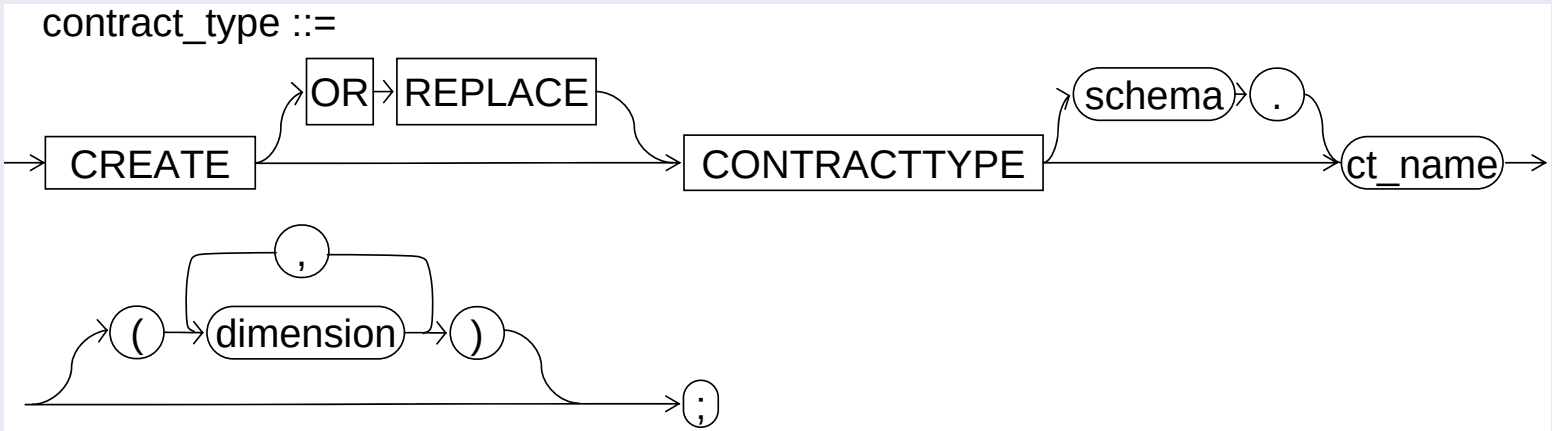
Exemple de workflow analytique



Extension d'un langage de requêtes de type SQL

Démarche préalable au requêtage

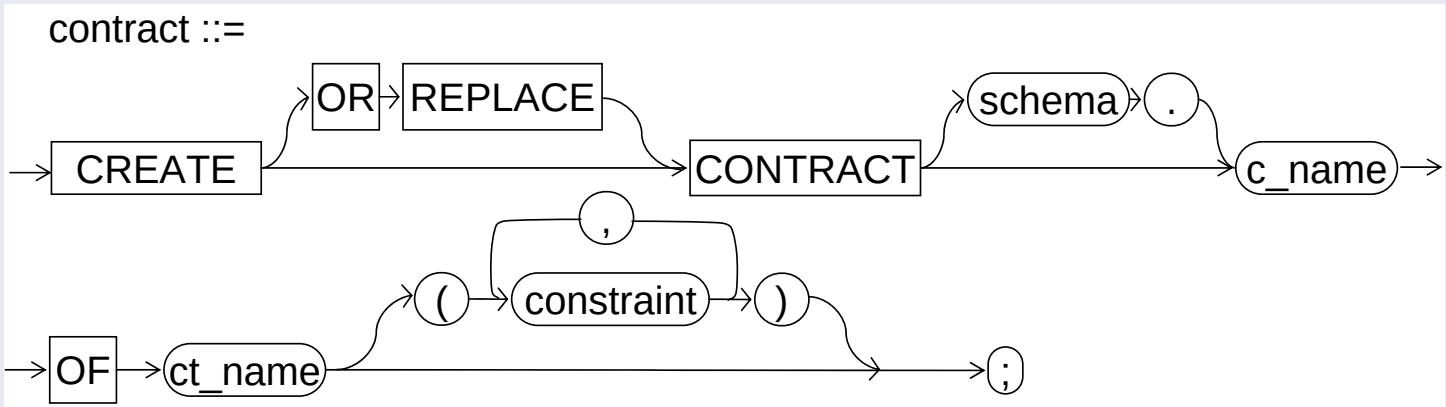
1. Création de types de contrats composés de dimensions associées à un ou plusieurs niveaux de granularité
2. Création de contrats avec spécification de contraintes sur chaque dimension



Extension d'un langage de requêtes de type SQL

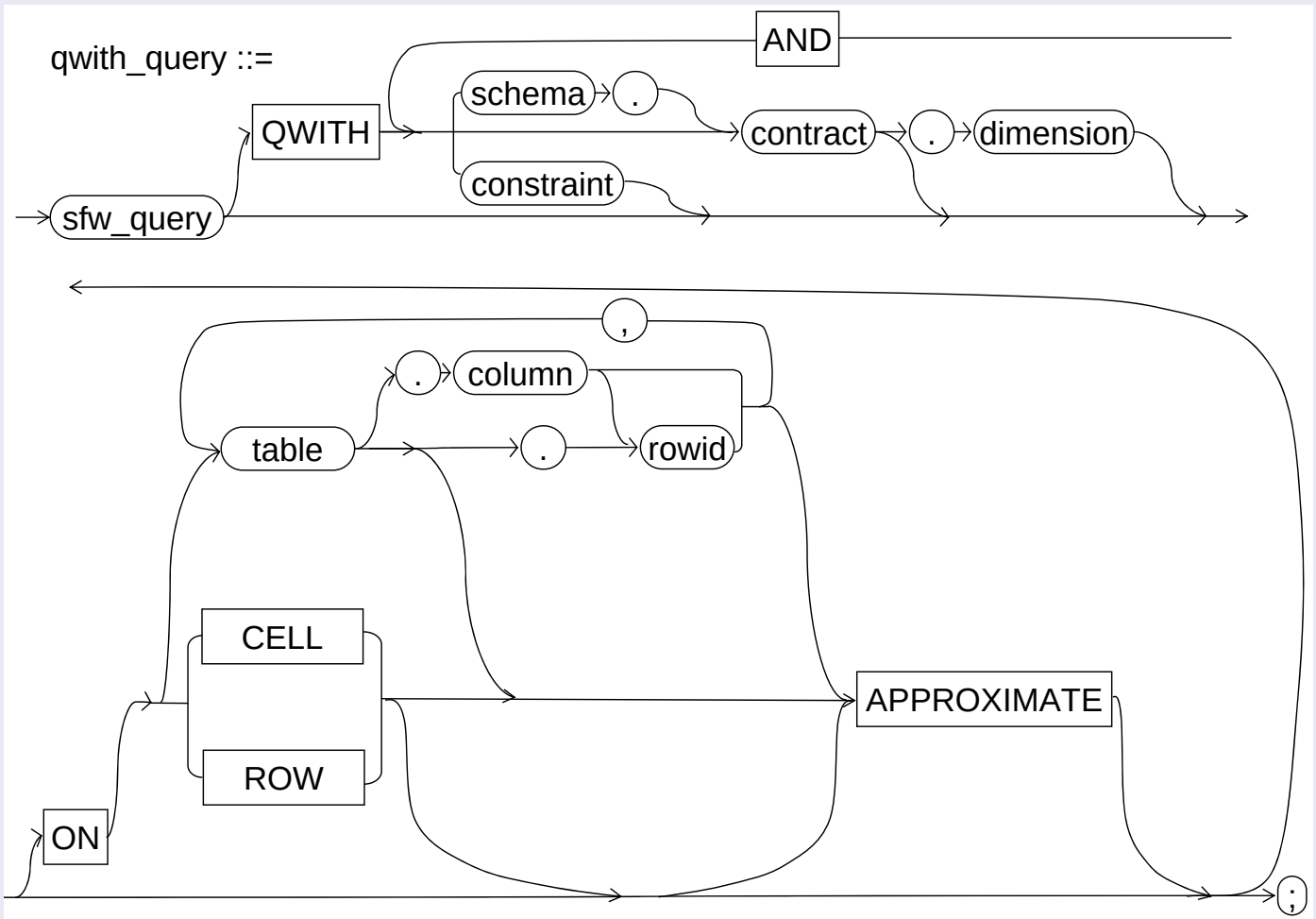
Démarche préalable au requêtage

1. Création de types de contrats composés de dimensions associées à un ou plusieurs niveaux de granularité
2. Création de contrats avec spécification de contraintes sur chaque dimension



Extension d'un langage de requêtes de type SQL

Requêtage contraint par l'appel de contrats



Exemples

Création de types de contrats

```
CREATE CONTRACTTYPE FRESHNESS(
   timeliness FLOAT ON CELL,ROW BY FUNCTION func_timeliness
        IS LANGUAGE JAVA NAME './XQLib/func_timeliness.java');
CREATE CONTRACTTYPE COMPLETENESS(
   nullValues% FLOAT ON ROW, TABLE BY FUNCTION plsql_nullValues%);
CREATE CONTRACTTYPE CONSISTENCY(
   SCprice FLOAT ON PRODUCT BY FUNCTION price_regression
    IS LANGUAGE SAS NAME './XQLib/price_regression.sas');
```

Création de contrats

```
CREATE CONTRACT fresh OF FRESHNESS(timeliness > .50);
CREATE CONTRACT complete OF COMPLETENESS(nullValues% <= .80);
CREATE CONTRACT consistent OF CONSISTENCY(SCprice< .05);
```

Requête étendue

```
SELECT PROD_ID, CUST_ID, FN, LN
FROM CUSTOMER C, PRODUCT P WHERE P.CUST_ID=C.CUST_ID
QWITH fresh ON CELL AND complete ON ROW AND consistent;
```

Bilan

- Modélisation des méta-données décrivant la qualité des données
- Constitution d'une bibliothèque de fonctions dédiées à l'évaluation des principaux facteurs de qualité
- Conception de *workflows* analytiques pour l'évaluation de la qualité des données
- Exploitation les méta-données par un langage étendu

Perspectives

- Optimisation du langage : approximation et relaxation
- Extension de la bibliothèque et développement d'un outil d'aide à la conception de workflows analytiques

Axe 2 : Prise en compte de la qualité des données pour la fouille

- Évaluer le coût de la non-qualité des données sur les résultats de fouille de données : cas de la découverte de règles d'association
- Proposer un modèle de coût basé sur la qualité des données analysées
- Filtrer les connaissances légitimement intéressantes pour le décisionnel sur la base de la qualité des données analysées

Mesures d'intérêt des règles : généralités

Pour une règle d'association $R : A \rightarrow B$ où A et B sont deux ensembles d'items tels que $A \cap B = \emptyset$, les mesures sont :

Support :	$\frac{N_A - N_{A\bar{B}}}{N}$
Confiance :	$1 - \frac{N_{A\bar{B}}}{N_A}$

- Une règle est dite valide si sa confiance est supérieure à un seuil de confiance σ_C, et son support est supérieur au seuil de support σ_S
- Une règle est dite exacte si sa confiance est de 1, sinon la règle est partielle.

Constat : Ignorance de la qualité des données analysées

Mesure de la qualité d'une règle

La qualité de la règle R: $A \rightarrow B$ est définie telle que :

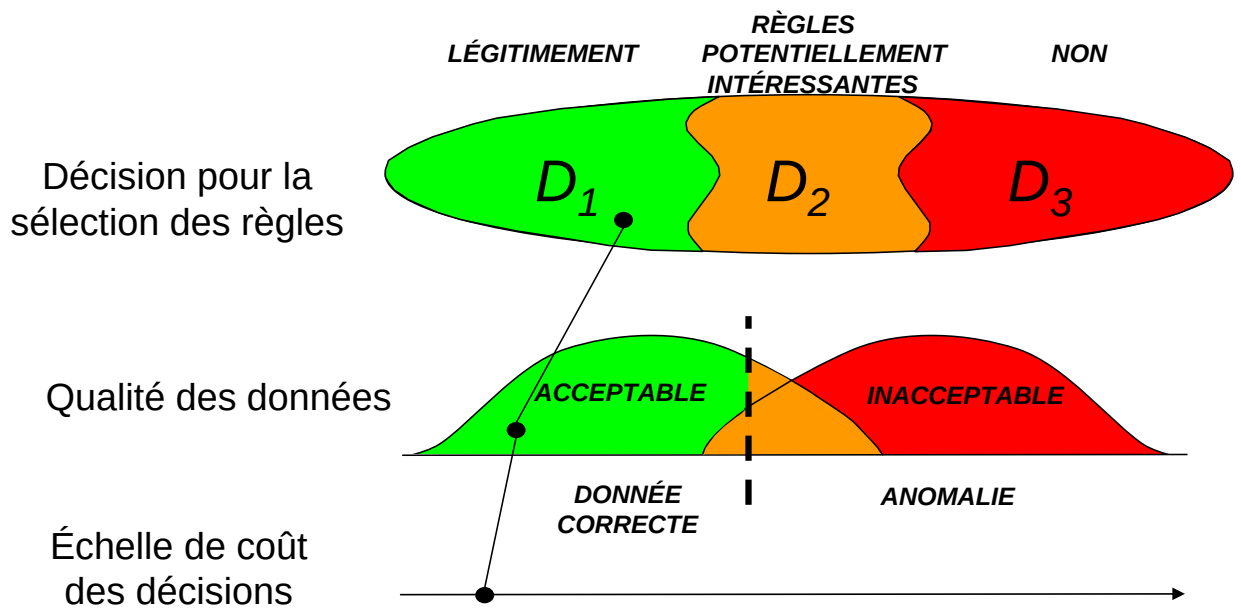
$$Q(R) = \begin{pmatrix} q_1(R) \\ q_2(R) \\ \dots \\ q_k(R) \end{pmatrix} = \begin{pmatrix} q_1(A) \otimes_1 q_1(B) \\ q_2(A) \otimes_2 q_2(B) \\ \dots \\ q_k(A) \otimes_k q_k(B) \end{pmatrix}$$

avec $q_j(A)$ et $q_j(B)$ les mesures associées à la dimension de qualité j calculées sur A et B composant la règle R et $\otimes_j$ une fonction de fusion particulière à chaque dimension, par exemple :

j	*Dimension*	*Fonction de fusion* $\otimes_j$
1	Fraîcheur	$\min[q_1(A), q_1(B)]$
2	Cohérence	$q_2(A) \cdot q_2(B)$
3	Complétude	$q_3(A) + q_3(B) - q_3(A) \cdot q_3(B)$

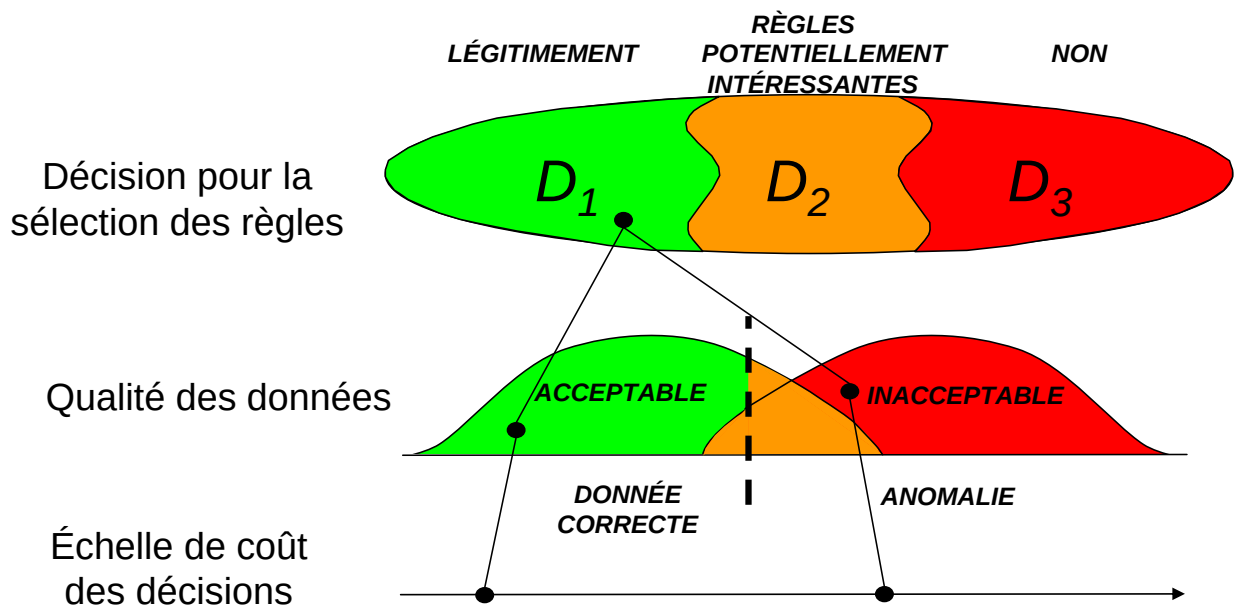
Objectifs

1. Évaluer le coût moyen d'une décision de sélection de règles d'après les mesures d'intérêt en ignorant la qualité des données analysées
2. Minimiser le coût moyen en prenant en compte les probabilités que les méta-données reflètent bien la qualité effective des données



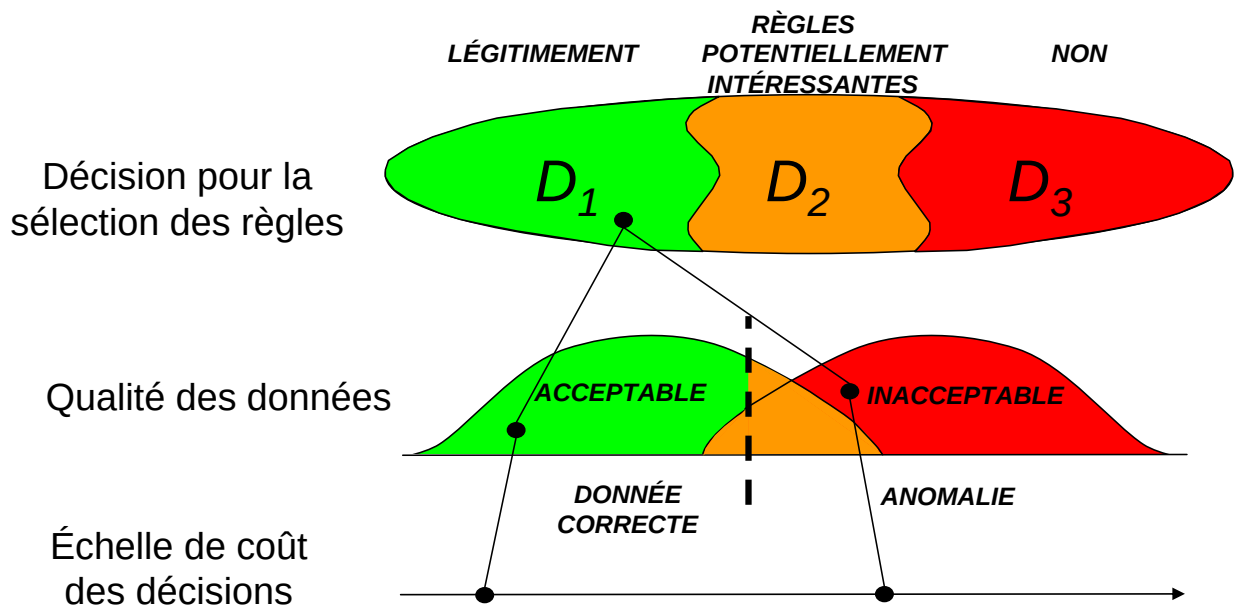
Objectifs

1. Évaluer le coût moyen d'une décision de sélection de règles d'après les mesures d'intérêt en ignorant la qualité des données analysées
2. Minimiser le coût moyen en prenant en compte les probabilités que les méta-données reflètent bien la qualité effective des données



Objectifs

1. Évaluer le coût moyen d'une décision de sélection de règles d'après les mesures d'intérêt en ignorant la qualité des données analysées
2. Minimiser le coût moyen en prenant en compte les probabilités que les méta-données reflètent bien la qualité effective des données



Expériences

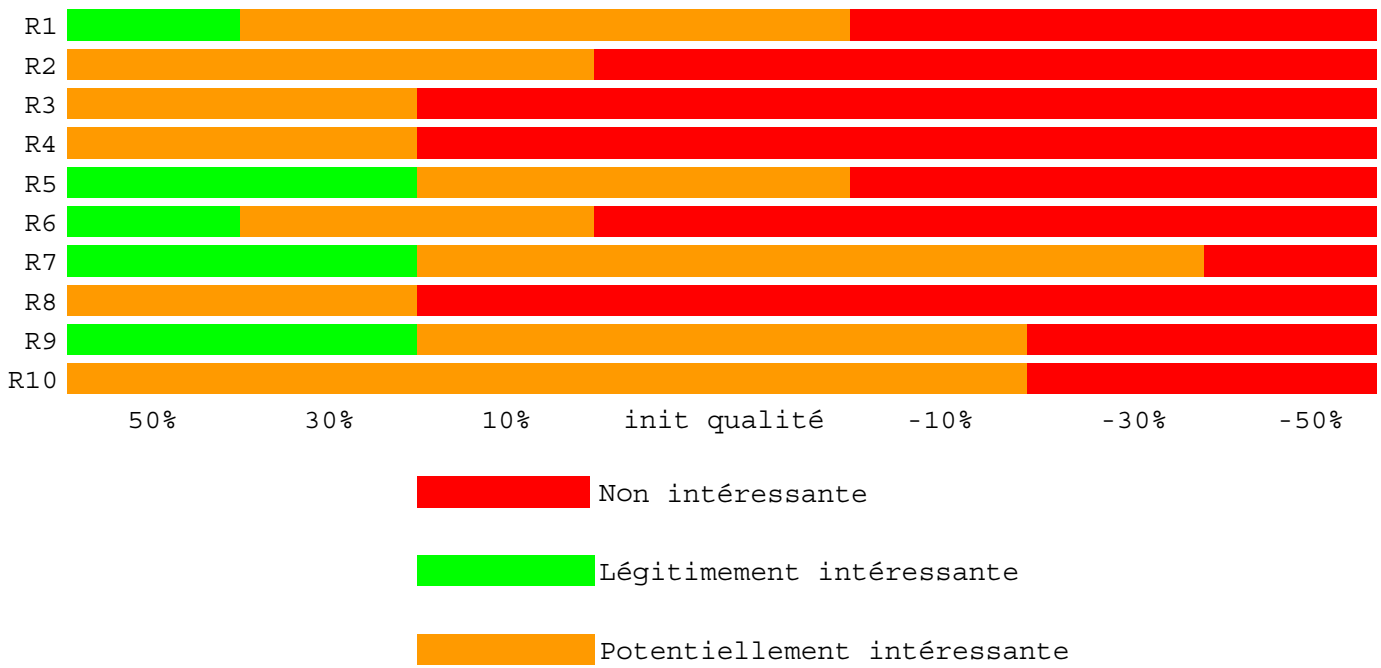
À partir des données de la KDD Cup-98 :

- Extraction des meilleures règles d'association
- Génération de méta-données synthétiques pour décrire la qualité des données
- Évaluation du statut des règles et du coût des décisions prises d'après les mesures d'intérêt
- Variations de la qualité des données

Expériences

Résultats intéressants

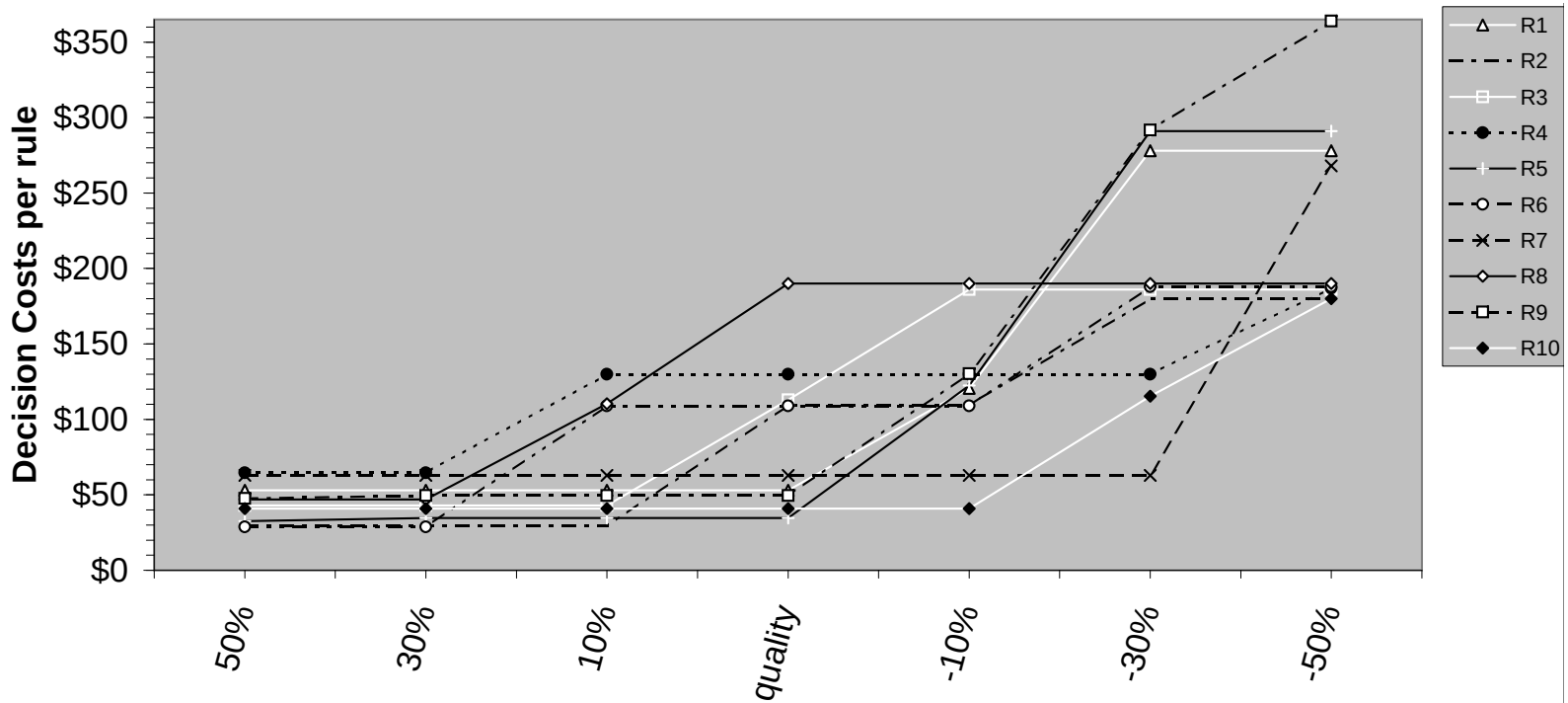
- les meilleures règles ne sont pas toujours légitimement intéressantes
- la dégradation de la qualité de données s'accompagne d'une augmentation significative du coût



Expériences

Résultats intéressants

- les meilleures règles ne sont pas toujours légitimement intéressantes
- la dégradation de la qualité de données s'accompagne d'une augmentation significative du coût



Bilan et perspectives

- Exploitation des méta-données décrivant la qualité des données pour :
  - l'évaluation de la qualité des règles et leur validation
  - le post-filtrage des règles d'association
- Rétro-analyse et ciblage des actions correctives sur les données utilisées en analyse à des fins décisionnelles
- Application à d'autres techniques de fouille

Plan

3 Applications

- Intégration de données génomiques et biomédicales
- Médiation de données CRM
- Monitoring de flux de données de télécommunication

Projet en collaboration avec l'INSERM U522

Contexte

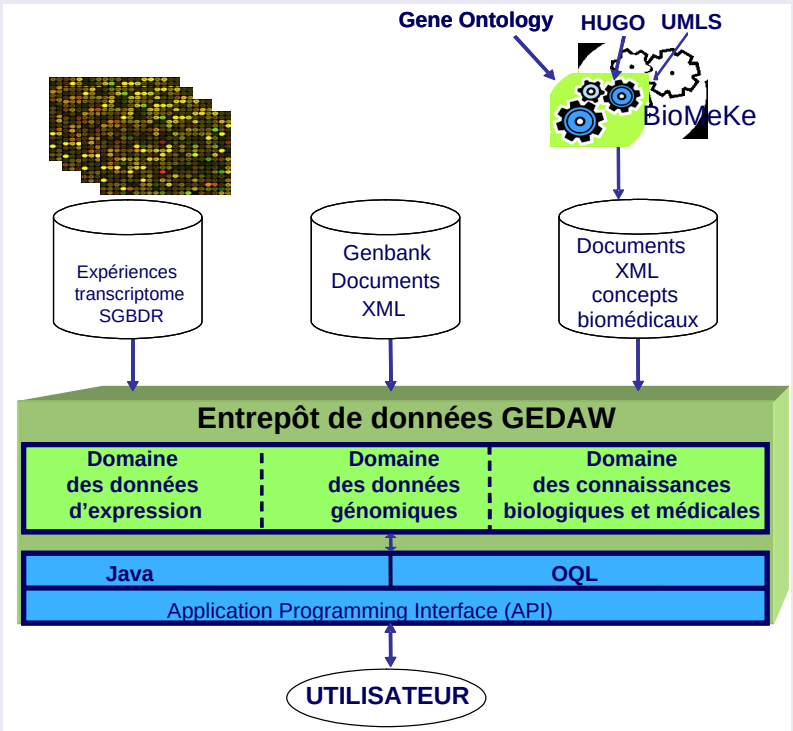
Collecte de toutes les connaissances génomiques et biomédicales disponibles dans les banques de données publiques sur les gènes impliqués dans les pathologies du foie

Contributions

- Modélisation des données du domaine génomique
- Conception d'un processus ETL dédié (XML $\rightarrow$ OODW)
- Evaluation de la qualité des données biomédicales
- Développement d'outils d'exploration de l'entrepôt, de requêtage par navigation et de profilage de la qualité des données util(isabl)es par les biologistes

Architecture : système d'intégration de données

- Extraction et nettoyage des données XML issues des principales banques de données publiques (GenBank, SwissProt)
- Intégration au sein d'un entrepôt orienté objet de données: GEDAW (*Gene Expression DAta Warehouse*)



Données de Gestion de la Relation Client

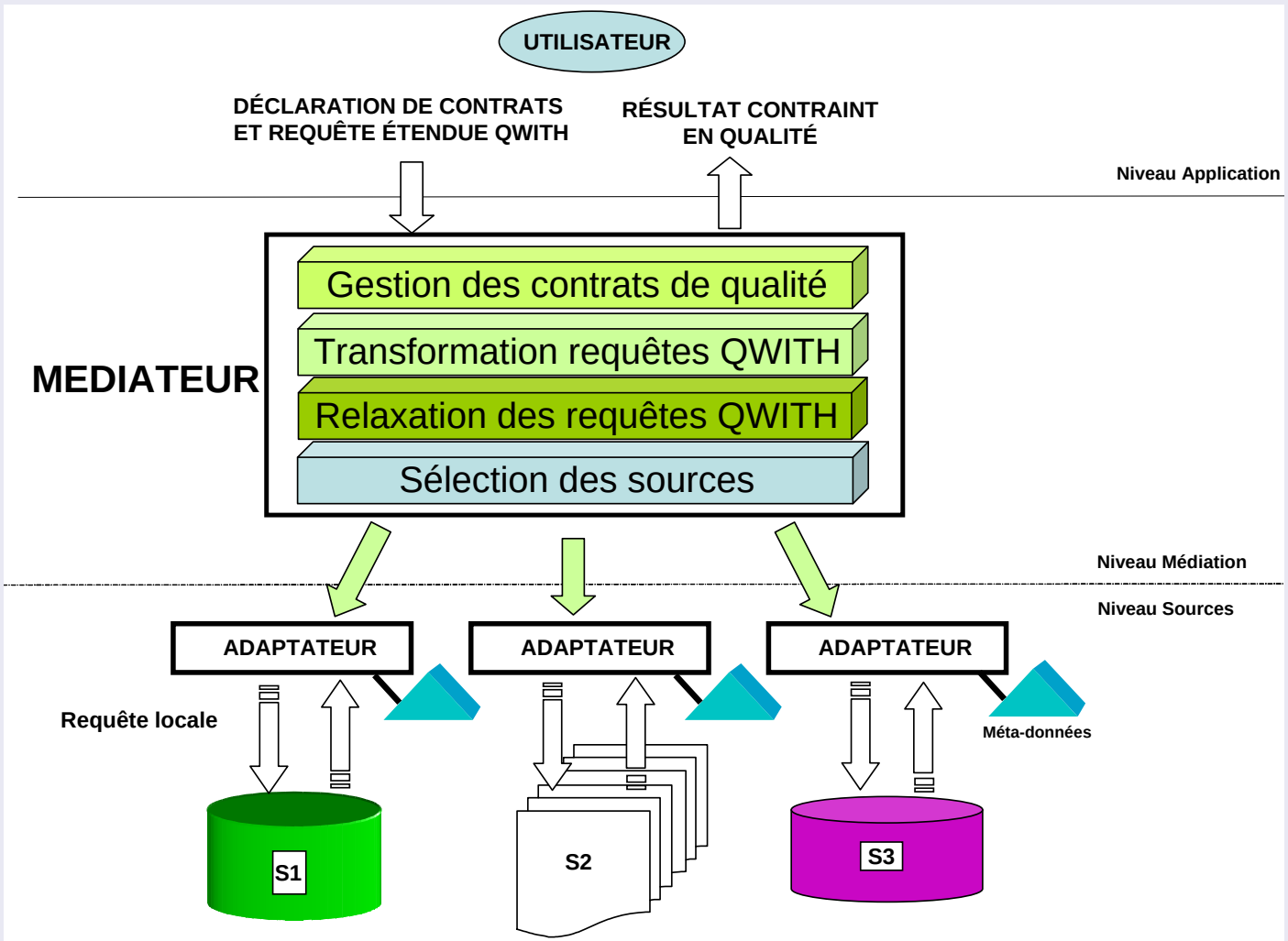
Contexte

Requêtage étendu de données intégrant des contraintes de qualité dans un environnement de médiation

Contributions

- Déclaration et propagation des contrats
- Langage de requêtes étendu par la clause `QWITH`
- Transformation des requêtes globales étendues (`SFW-QWITH`) en requêtes locales étendues
- Sélection de sources selon leur capacité à répondre à la requête et à satisfaire au mieux des contraintes imposées sur la qualité des données
- Négociation ou relaxation des contraintes de qualité

Architecture : système de médiation de données



Données de téléphonie

Contexte

Travail prospectif pour la société Genielog/SFR-Cegetel sur les techniques de fouille applicables à l'évaluation de la qualité des flux de données de téléphonie

Problématique

- Analyses et traitements "au fil de l'eau"
- Contraintes algorithmiques fortes
- Approximation et fenêtrage nécessaires

Contributions

- Étude des techniques de fouille de flux de données
- Spécification des premiers workflows analytiques pour le contrôle de la qualité des flux de données

Plan

Principales contributions

Intégrer la gestion de la qualité des données à l'ingénierie et la gestion des bases de données

- Modélisation de méta-données décrivant la qualité
- Spécification de fonctions et workflows analytiques
- Extension d'un langage de requêtes permettant la déclaration et la manipulation de contraintes sur la qualité

Évaluer la qualité des règles d'association

- Exploitation des méta-données décrivant la qualité des données analysées complémentaire aux mesures d'intérêt
- Modèle de décision pour filtrer les règles légitimement intéressantes

Principales contributions

Intégrer la gestion de la qualité des données à l'ingénierie et la gestion des bases de données

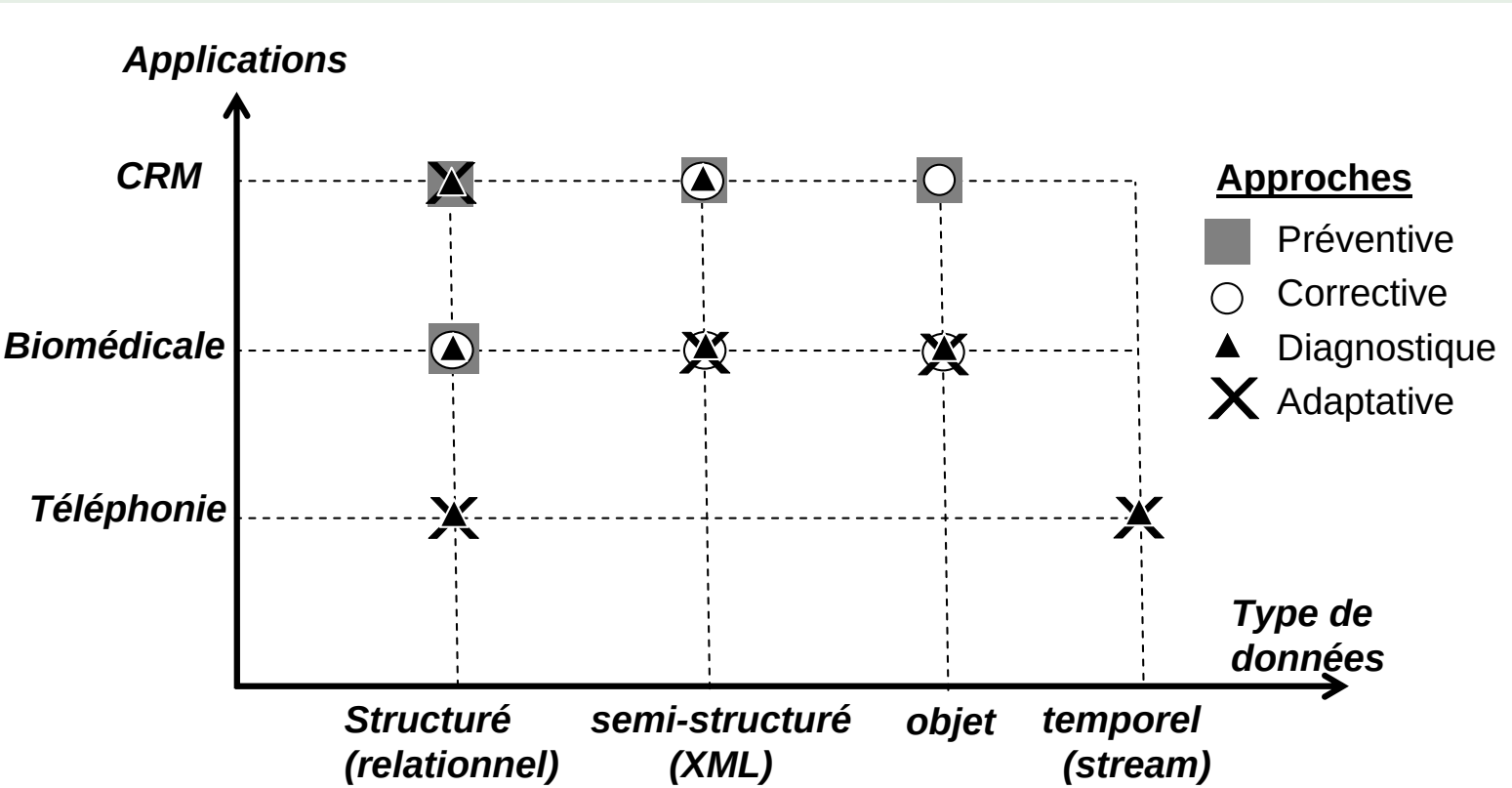
- Modélisation de méta-données décrivant la qualité
- Spécification de fonctions et workflows analytiques
- Extension d'un langage de requêtes permettant la déclaration et la manipulation de contraintes sur la qualité

Évaluer la qualité des règles d'association

- Exploitation des méta-données décrivant la qualité des données analysées complémentaire aux mesures d'intérêt
- Modèle de décision pour filtrer les règles légitimement intéressantes

Application des contributions

- à des domaines variés
- à différents types de données
- selon les différentes approches et types d'architecture



Directions de recherche

À court terme

- Optimisation des requêtes étendues
- Conception de patterns de workflows analytiques dédiés à l'évaluation de la qualité des données
- Étude de la sensibilité des techniques de clustering à des problèmes de qualité des données surajoutés

Directions de recherche

À moyen terme

- Analyse des inter-dépendances entre les dimensions de la qualité des données : **projet QUADRIS**
- Conception de systèmes de gestion de données introspectifs : **projet de mobilité financé par la Commission Européenne** pour 2 ans à AT&T Labs Research, New Jersey, USA, équipe de D. Srivastava

Directions de recherche

À moyen terme

- Analyse des inter-dépendances entre les dimensions de la qualité des données : **projet QUADRIS**
- Conception de systèmes de gestion de données introspectifs : **projet de mobilité financé par la Commission Européenne** pour 2 ans à AT&T Labs Research, New Jersey, USA, équipe de D. Srivastava

À long terme

Élargir ma couverture de la problématique de la qualité des données à :

- d'autres domaines d'application
- des volumétries très nettement supérieures

Merci à tous !