Recherche de recouvrements dans une collection de schémas de bases de données

Anne-France Brogneaux, *Ravi Ramdoyal*, Julien Vilz, Jean-Luc Hainaut
Laboratoire d'Ingénierie des Bases de Données
Goupe Precise

Institut d'Informatique
Université de Namur, Belgique

L'Institut d'Informatique de Namur

- L'Université de Namur
  - 5000 Etudiants
  - 7 Facultés

- L'Institut d'Informatique
  - 16 Professeurs
  - 80 chercheurs
  - Divers domaines :
    - Bases de données
    - Aide à la décision
    - Interfaces homme-machine
    - Ingénierie logicielle
    - Multimédia
    - Théorie organisationnelle
    - …

Le groupe PRECISE

- Precise is REsearch Center in Information Systems Engineering

- Principaux domaines de recherche :
  - Méthodes et modèles
  - Ingénierie de bases de données
  - Analyse des besoins
  - CASE tools
  - Interopérabilité et re-ingénierie
  - Qualité et mesures

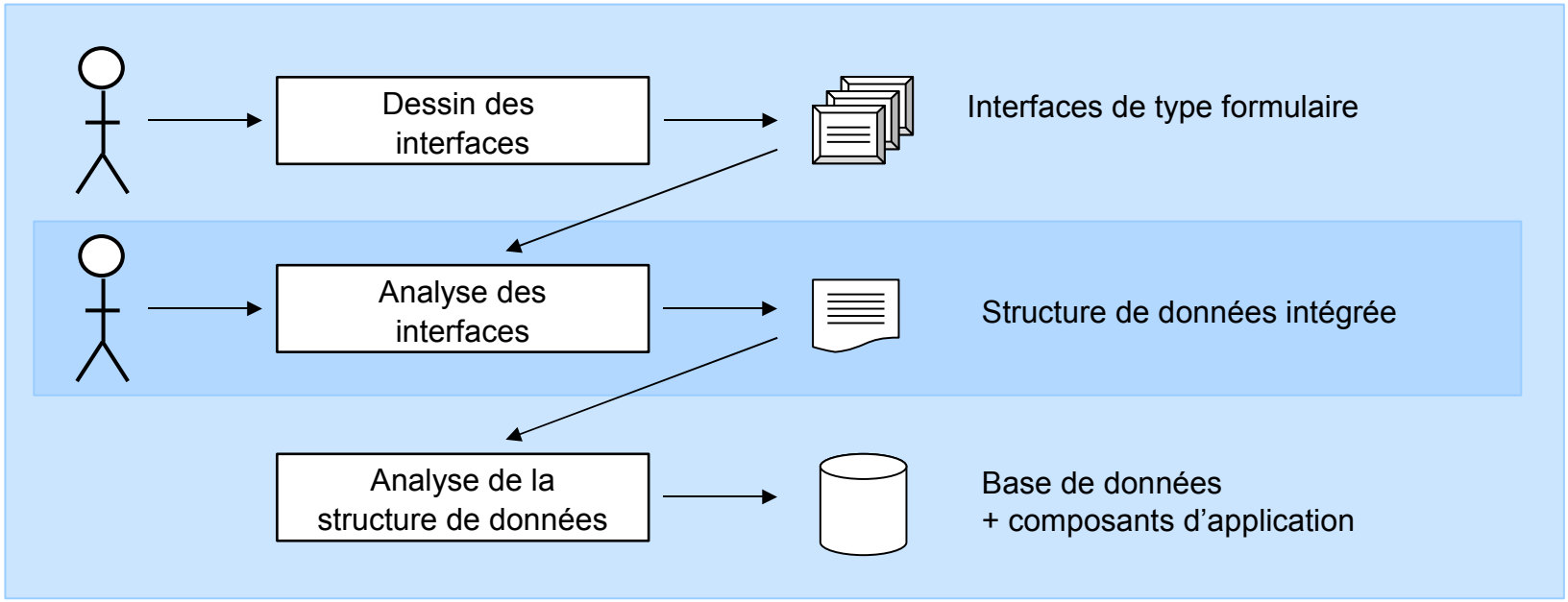
- Equipe
  - 6 Professeurs
    - Vincent ENGLEBERT
    - Naji HABRA
    - Jean-Luc HAINAUT
    - Patrick HEYMANS
    - Michael PETIT
    - Pierre-Yves SCHOBBENS
  - 25 Chercheurs

Plan de la présentation

Recherche de recouvrements dans une
collection de schémas de bases de données

- Mise en contexte
- Problématique
- Définitions
- Approches
- Conclusion

- La démarche ReQuest
  - Production de sites de commerce électronique (PME)
  - Implication « active » des utilisateurs finaux

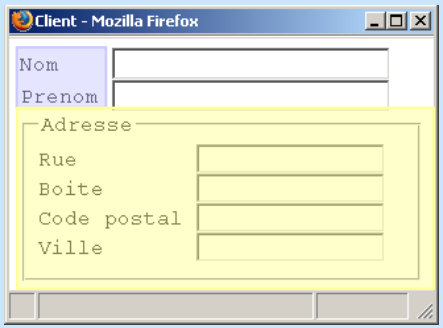


Vue globale de la démarche ReQuest

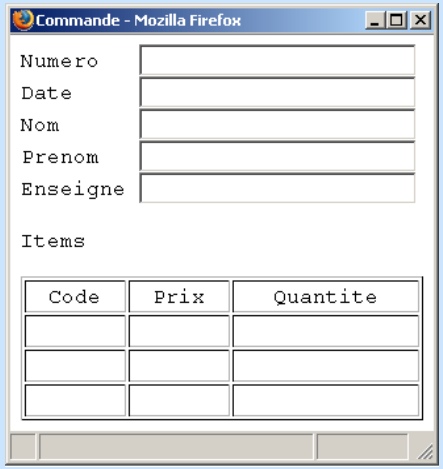
2. Problématique

- Phase d'analyse des interfaces
  - Les interfaces contiennent de l'information
  - Cette information est redondante
- Représentation des interfaces sous la forme d'un modèle de données
  - Choix du modèle Entité-Association
- Traitement des redondances via le modèle de données
  - Identification
  - Validation
  - Résolution
- Exemple
  - Système simplifié de gestion pour la commande de produits dans une chaîne de magasin

2. Problématique

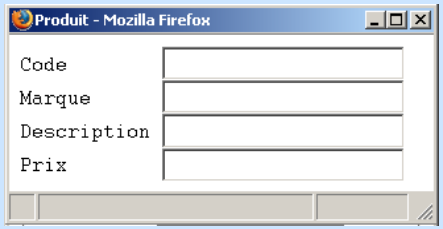


CLIENT
Nom
Prénom
Adresse
Rue
Boîte
Code postal
Ville



COMMANDE
Numéro
Date
Nom
Prénom
Enseigne
Items[0-N]
Code
Prix
Quantité

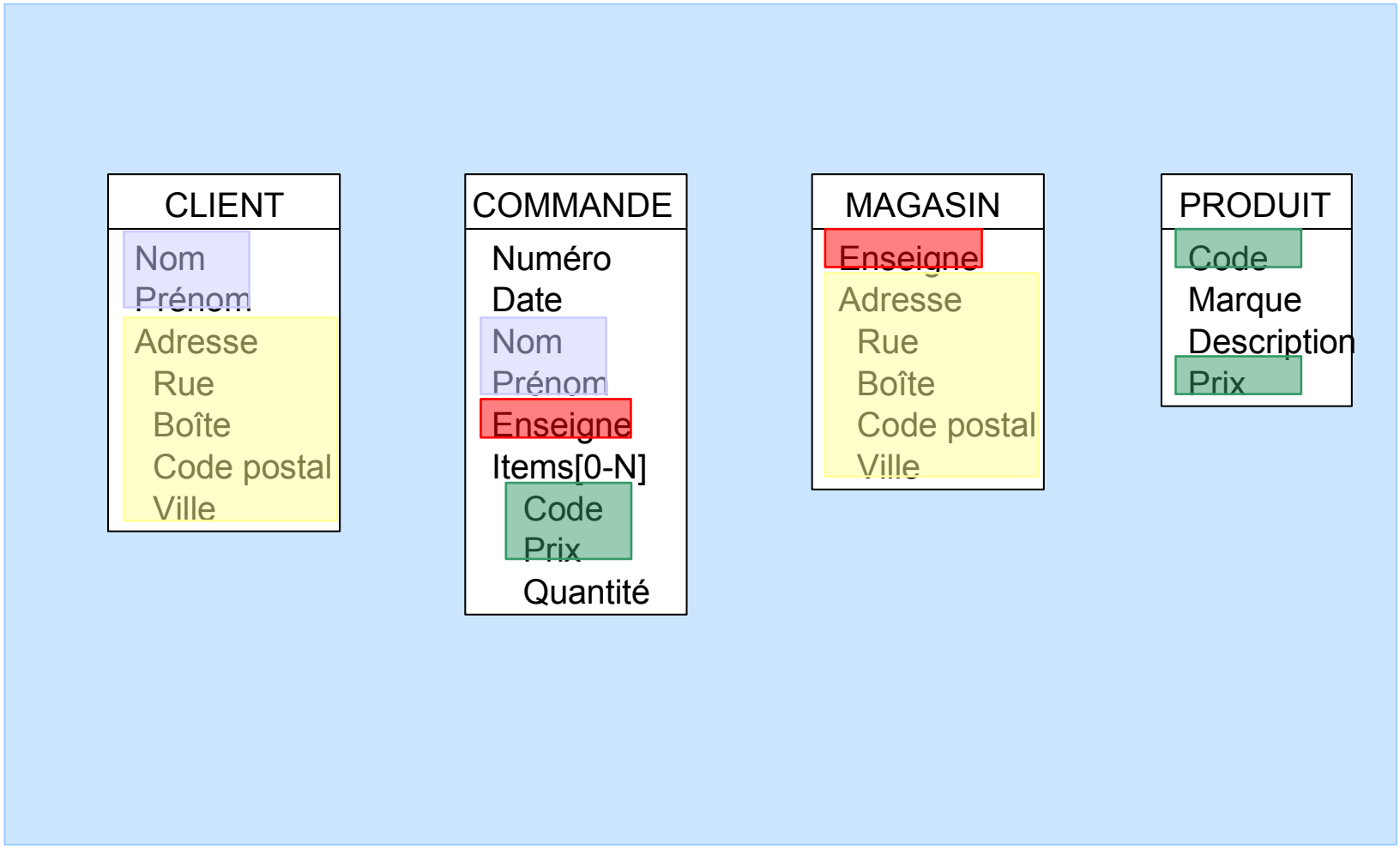
MAGASIN
Enseigne
Adresse
Rue
Boîte
Code postal
Ville



PRODUIT
Code
Marque
Description
Prix

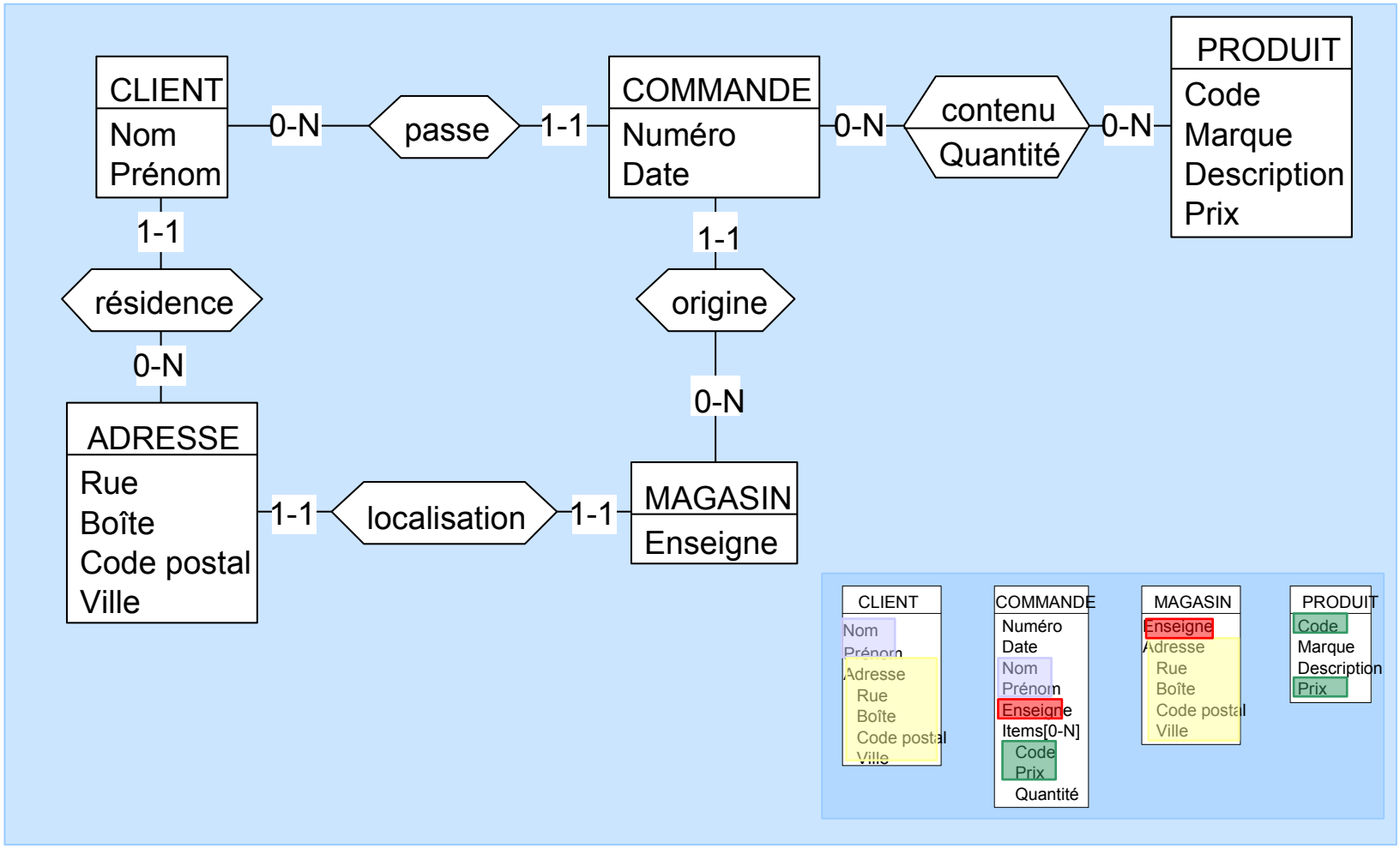
Représentation de blocs fonctionnels d'interface sous la forme de types d'entité

2. Problématique



Identification de redondances de noms d'attributs

2. Problématique

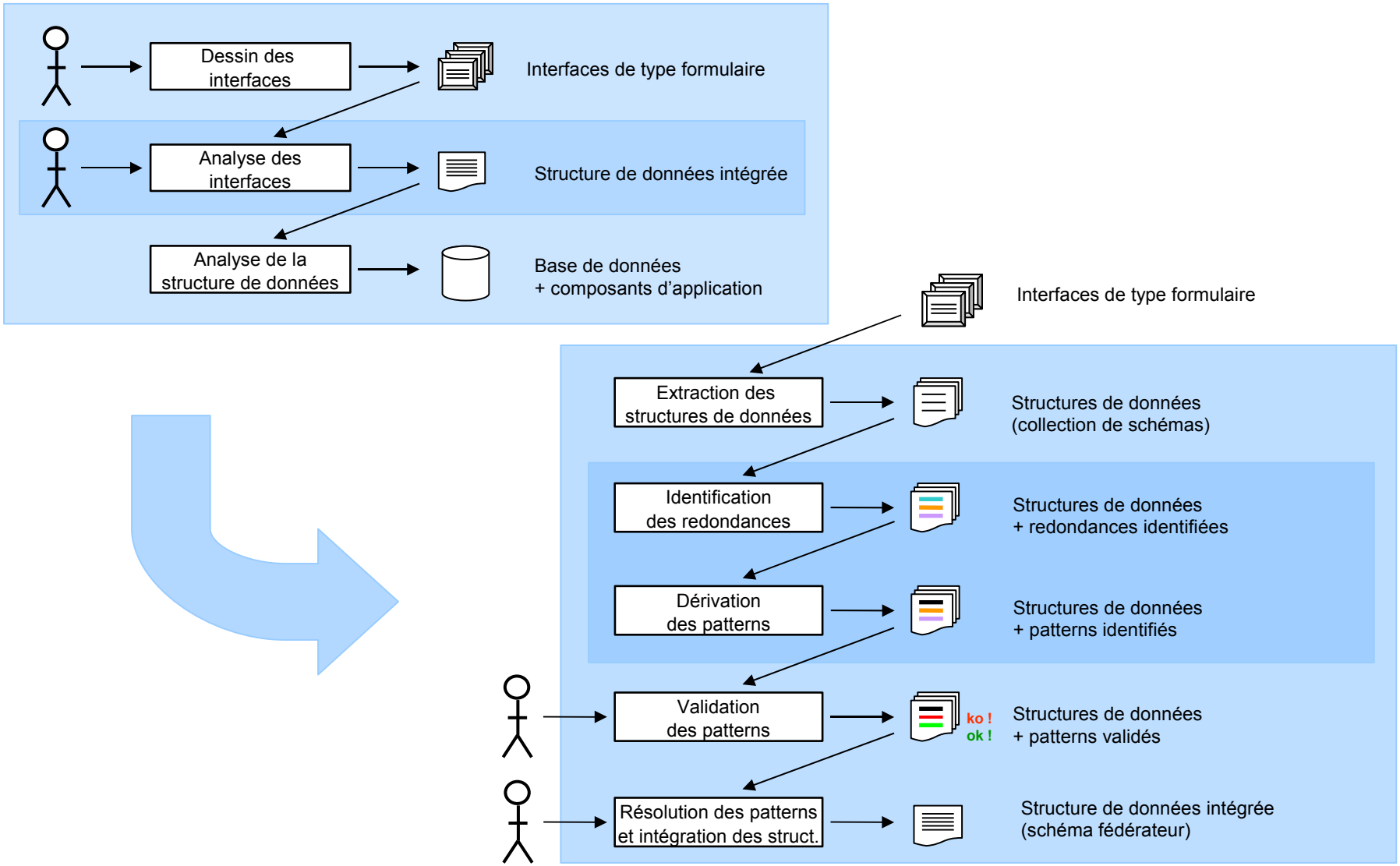


Résolution de redondances de noms d'attributs.

3. Définitions

- Redondance
  - Exprimée par ensembles d'attributs similaires
- Indicateurs de similarité
  - Nom, taille, type de valeur, ...
  - Indice compris entre 0 et 1
- Similarité de deux attributs
  - Moyenne pondérée d'indices de similarités
- Similarité de deux groupes d'attributs
  - Moyenne des similarités d'attributs pris deux à deux
- Pattern
  - Redondance particulière (« significative »)
  - Classe d'équivalence définie par :
    - la relation de similarité de groupes d'attributs
    - un seuil de similarité pour les membres de la classe (instances)
  - Instance représentative calculée

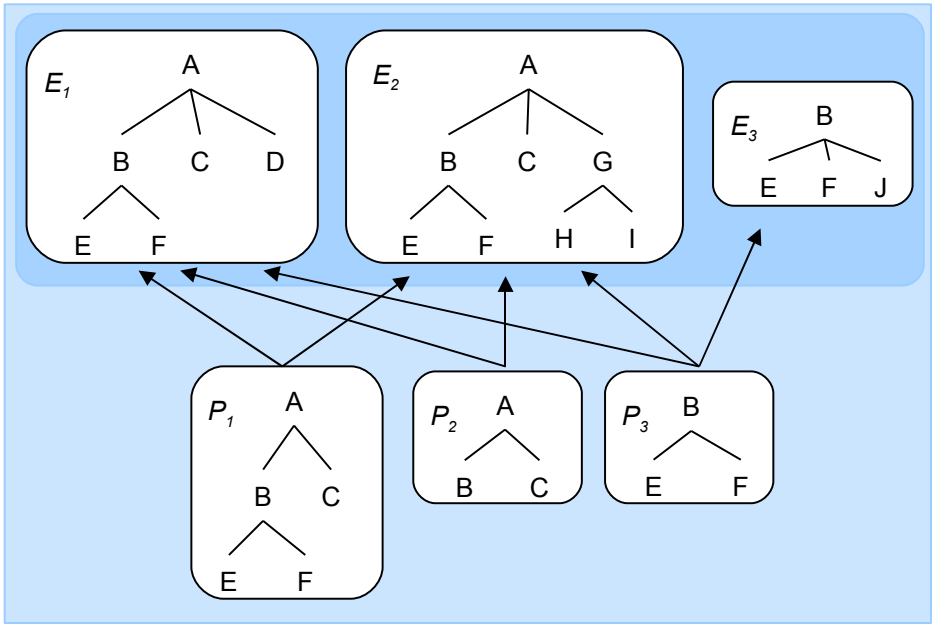
3. Définitions



Analyse des redondances

Recherche de patterns arborescents (tree mining)

- (1/2) Identification des redondances
  - Considérer chaque type d'entité comme un arbre et procéder à la recherche des différents sous-arbres communs.
  - Choix de l'algorithme FreqT

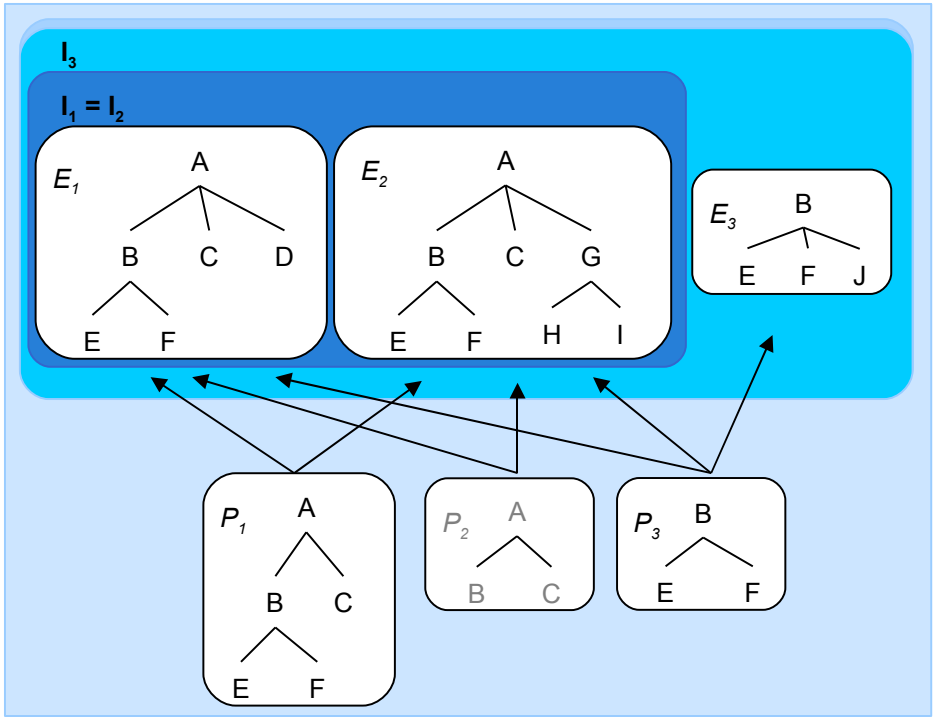


Recherche de patterns arborescents

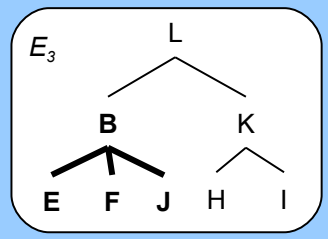
- (2/2) Identification des patterns
  - Filtrage (règle de pertinence)

Si Pi sous-arbre de Pk,
Si Ii contient strictement Ik
=> Alors on conserve Pi.

- Limites : similarité des nœuds parents



Et si ...

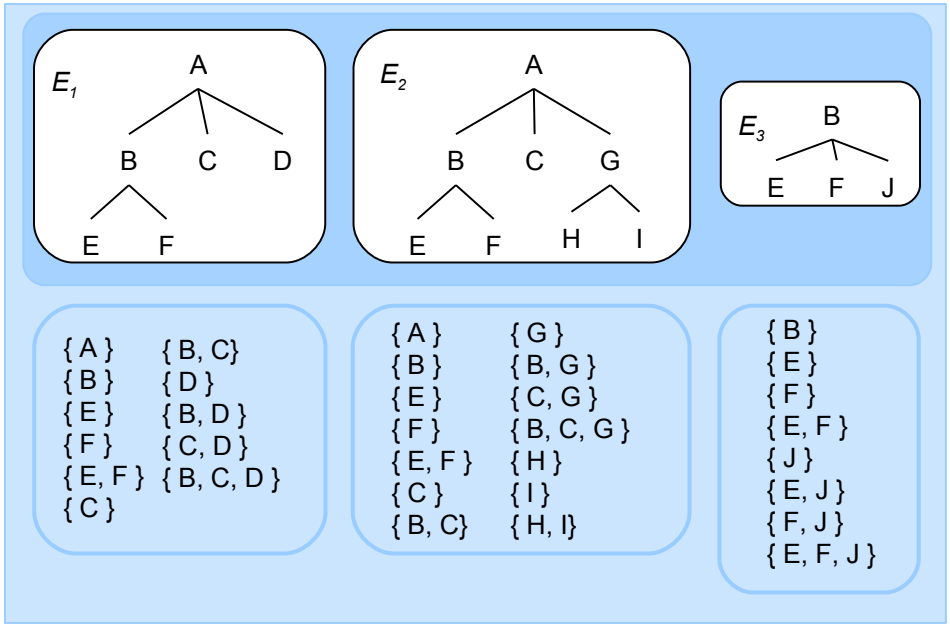


G = K ?

L = A ?

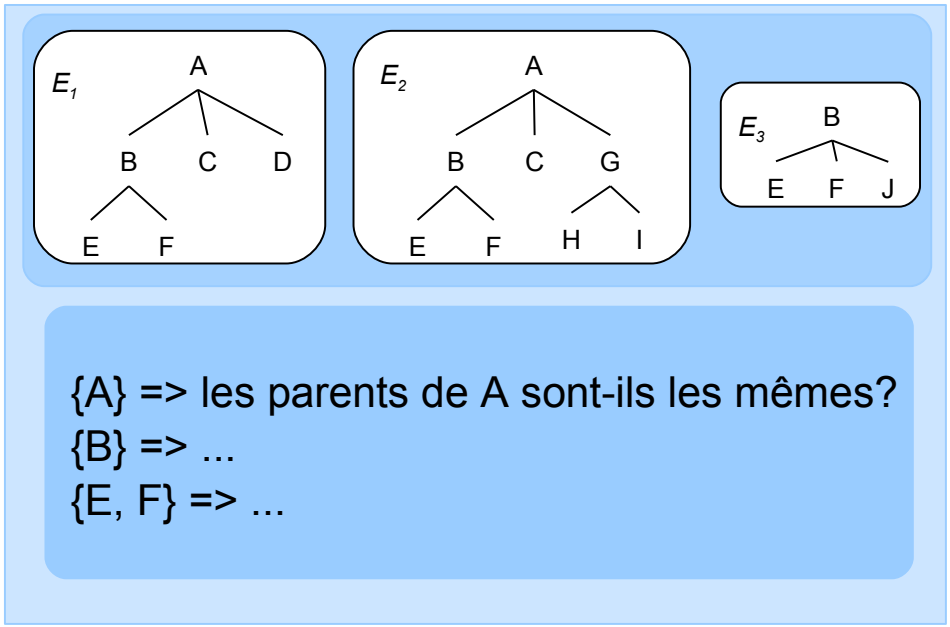
Génération exhaustive de patterns non arborescents

- (1/2) Identification des redondances
  - sous forme de série (attributs de même niveau)
  - type d'entité par type d'entité

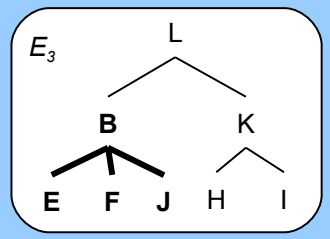


Génération exhaustive de patterns non arborescents

- (2/2) Identification des patterns
  - Filtres en fréquence, largeur et généricité
- Limites : temps d'exécution !



Et si ...

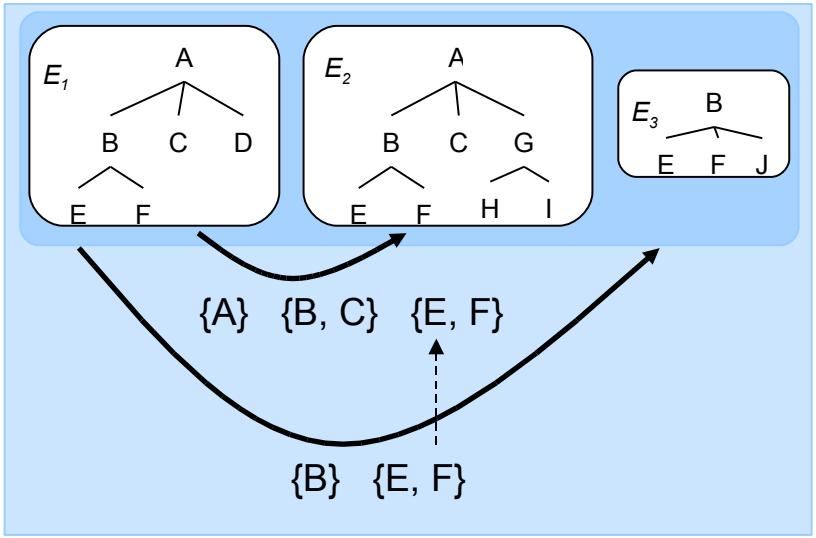


{H, I} => G = K ?

{B} => L = A ?

Génération sélective de patterns non arborescents

- (1/2) Identification des redondances
  - Amélioration de l'approche précédente
  - par comparaison de type d'entité
- (2/2) Identification des patterns
  - Filtres précédents
- En cours d'expérimentation...



Conclusion

- Démarche ReQuest
  - Extraction d'un schéma conceptuel intégré à partir d'IHM
- Méthode
  - Identification de redondances
  - Identification de patterns
  - + validation, résolution, ...
- Passage à l'échelle gérable
  - Taille et structure des *pages* web
  - Taille et structure des *sites* web orientés e-business

Merci pour votre attention

Des questions ?

Université de Namur

http://www.fundp.ac.be

Institut d'Informatique

http://www.info.fundp.ac.be

Groupe PRECISE

http://www.software-engineering.be/

Contact :

- Courrier électronique
  ravi.ramdoyal@fundp.ac.be
- Site du laboratoire :
  http://www.info.fundp.ac.be/libd