FSEGS

Faculté des Sciences Économiques et de Gestion de Sfax-Tunisie

MIRACL

Multimedia, InfoRmation systems and Advanced Computing Laboratory

Du XML au multidimensionnel : Conception de magasins de données

Yasser Hachaichi, Jamel Feki, Hanêne Ben-Abdallah

PLAN

1. Contexte
2. État de l'art & motivations
3. Présentation de l'approche
4. Prétraitement de la source
5. Extraction des concepts multidimensionnels
6. Bilan & Perspectives

Contexte

- Mondialisation, la concurrence croissante
  - la **prise de décision** est devenue **cruciale** pour les dirigeants
- Systèmes d'information décisionnels (SID)
  - Dédiés au pilotage des entreprises
  - Basés sur des structures particulières de stockage spécifiques **ED & MD**
  - Traditionnellement alimentés par des informations issues de sources internes à l'entreprise
- Ouverture des entreprises sur l'Internet
  - les sources de données englobent des données échangées avec les partenaires et/ou issues du Web
  - Les documents XML : source de données de plus en plus répandues

Approche de construction de schémas multidimensionnels à partir de sources XML

- Approches de conception des SID :
  - Ascendantes, Descendantes et Mixtes
  - Ascendantes pour les sources XML
    - Semi-automatiques
    - Requièrent une intense intervention pour l'identification de certains concepts multidimensionnels
    - Intervention qui nécessite une double expertise :
      - Dans le domaine des sources XML
      - Dans celle du domaine décisionnel
    - Génération d'un grand nombre de schémas MD
    - Pas de règles pour dériver automatiquement les représentations logiques

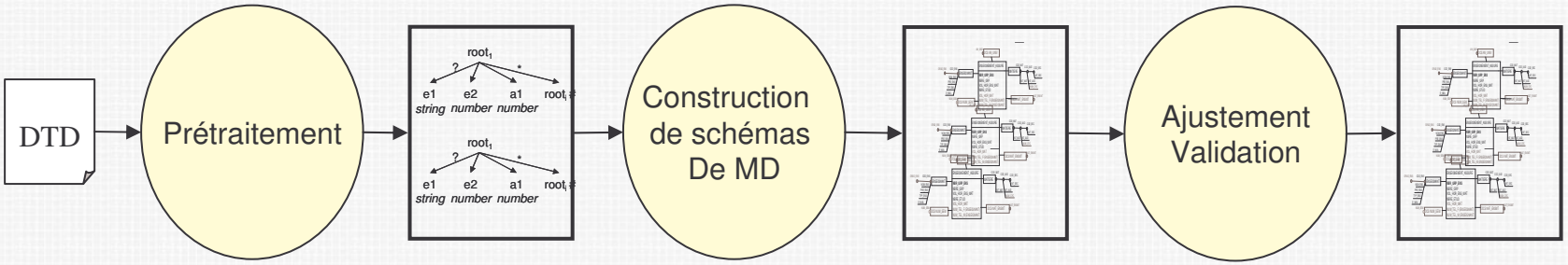
État de l'art & motivations

- Objectif : Méthode et outil de conception d'ED
  - Concevoir automatiquement les ED
  - Partir de sources XML
  - Préparer le passage automatique vers le niveau logique

- Proposition
  - Méthode automatisable d'aide à la conception de MD
  - Des règles d'extractions (source XML décrite par une DTD),
  - Classement des concepts extraits par niveaux de pertinence
  - Association concept-source

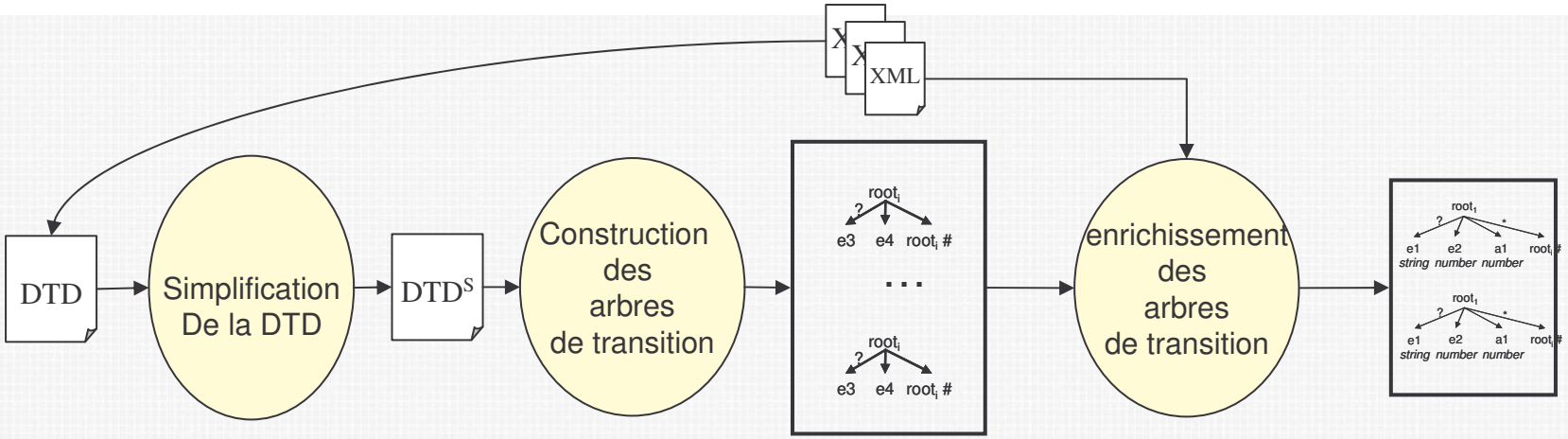
Présentation de l'approche

- Méthode de conception d'ED



  - Prétraitement
    - Simplifier la DTD → Réduire la complexité
    - Construire les arbres de transition → Préparer l'extraction
  - Construction de schémas de MD
    - Extraire les concepts multidimensionnels (règles)
    - Associer un niveau de pertinence (potentiel analytique de chaque concept)
  - Ajustement/Validation
    - Adapter les MD aux besoins
      - Opérateurs de manipulations

Prétraitement



- Suppression des éléments vides

  <!ELEMENT test EMPTY> → <test></test> ou <test/>

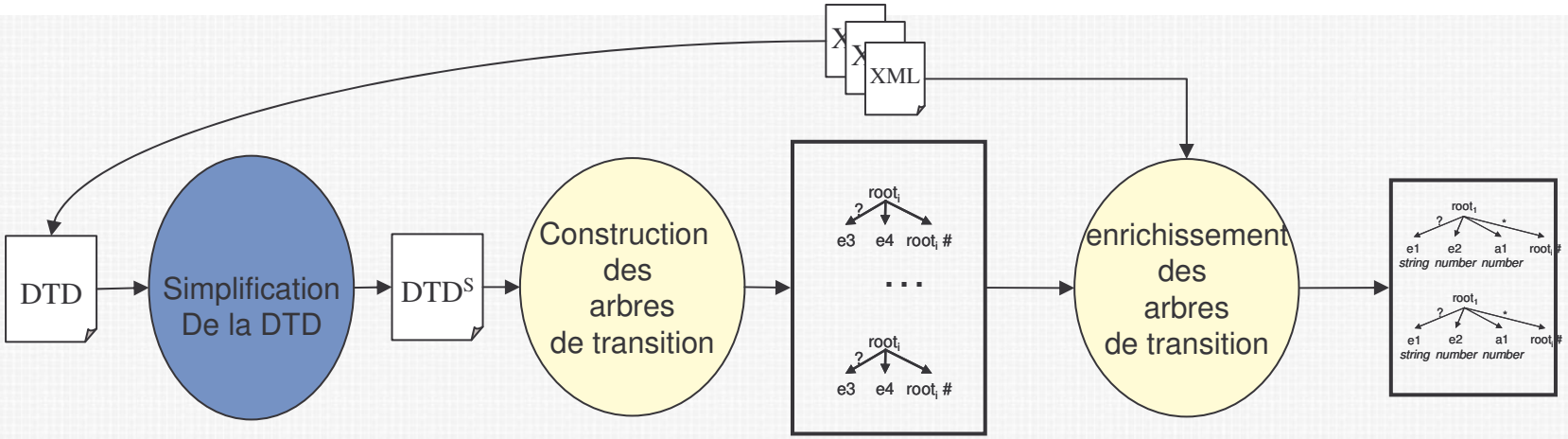
- Substitution des références
  - remplacer toute référence à l'entité par sa valeur
  - supprimer la déclaration de l'entité

  Exemple : déclarer un groupe d'éléments une seule fois dans la DTD et le référencer afin d'éviter sa répétition

  ⇒ garantir la désignation des concepts multidimensionnels identifiés par leur noms au lieu des références non significatives

  ⇒ faciliter les opérations ultérieures de chargement

Prétraitement

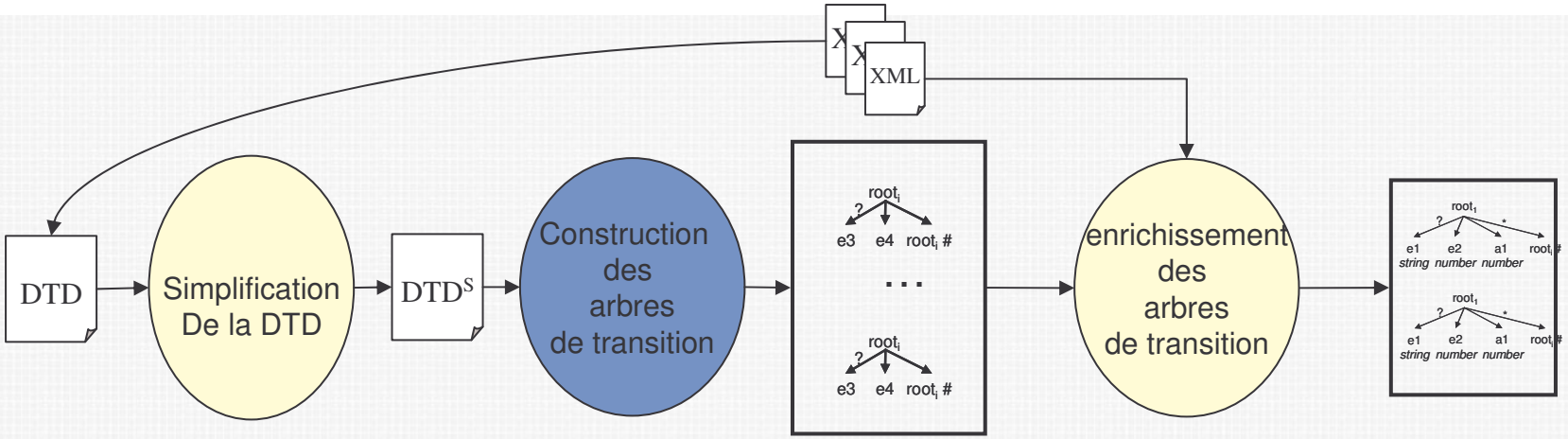


- *Simplification des déclarations d'éléments*

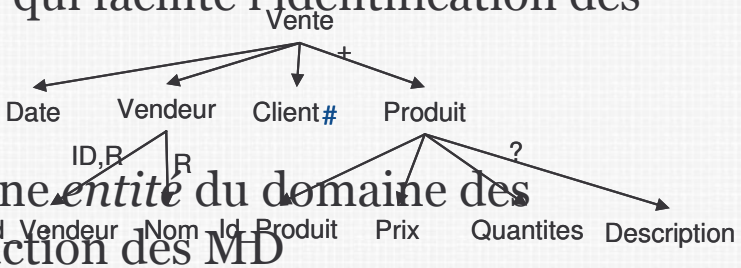
(a) Aplatissement	(b) Réduction	(c) Groupement
(e1, e2)*→e1*, e2*	e1**→ e1*	…,e1*,…,e1*, …→ e1*,…
(e1, e2)?→e1?, e2?	e1*?→ e1*	…,e1*,…,e1?, …→ e1*,…
(e1|e2) →e1?, e2?	e1?*→ e1*	…,e1?, …,e1*, …→e1*,…
	e1??→ e1?	…,e1?,…,e1?, …→ e1*,…
		…, e1, …, e1, …→ e1+,…

Une DTD^S ≠ DTD

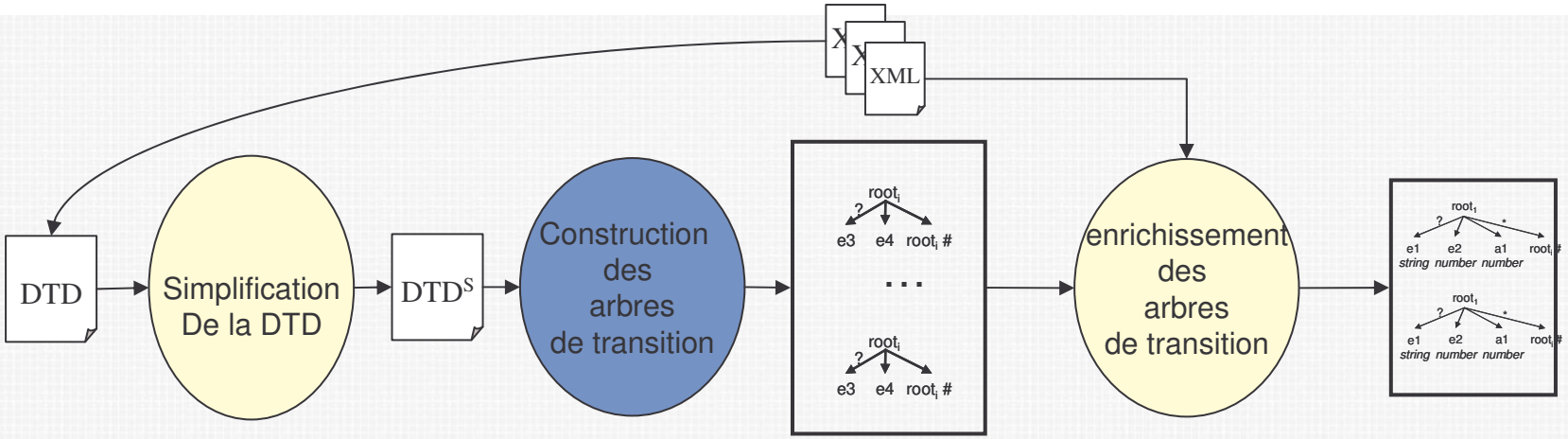
Prétraitement



- Arbres de transition
  - Les arbres de transition modélisent les liens structurels entre les éléments de la DTD qui facilite l'identification des concepts multidimensionnels.
  - chaque feuille est soit
    - un élément PCDATA
    - un attribut
    - une référence vers le sommet d'un arbre de transition
  - Chaque arbre de transition représente une *entité* du domaine des documents XML utilisés pour la construction des MD



Prétraitement



- *Détermination des sommets des arbres de transition*

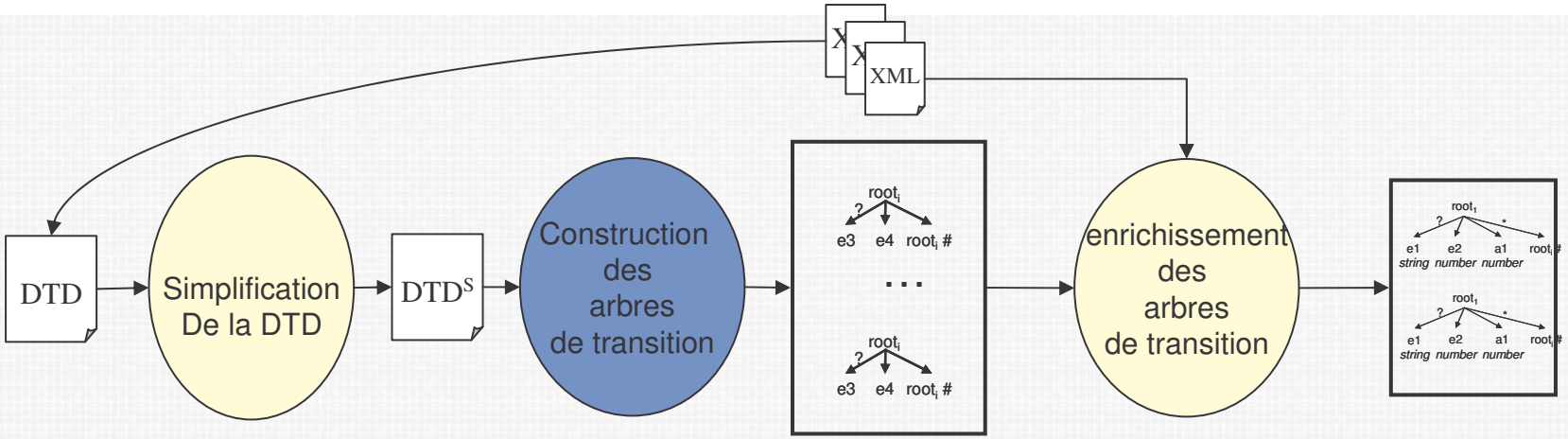
  RD1 : élément qui n'apparaît pas dans d'autres déclarations d'éléments

  RD2 : élément contenant au moins un élément non PCDATA

  RD3 : élément se trouvant dans la déclaration de n≥2 autres éléments

  RD4 : élément qui apparaît récursivement (directement ou indirectement) dans sa propre déclaration

Prétraitement

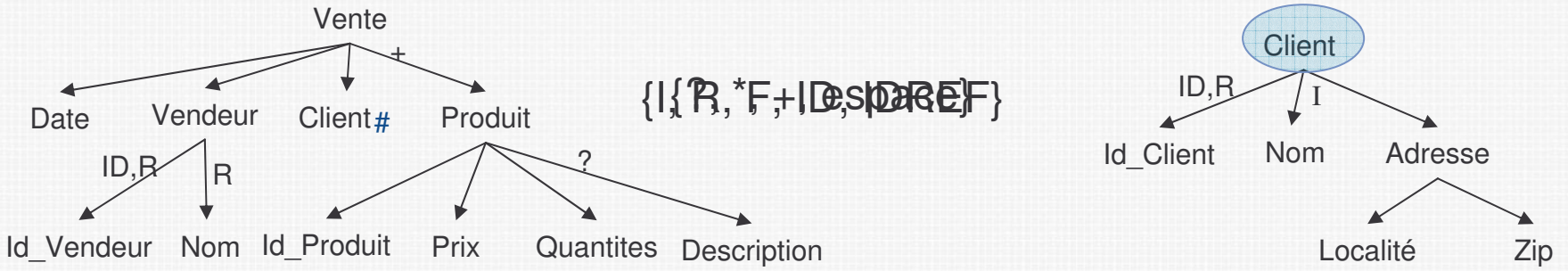


- Construction

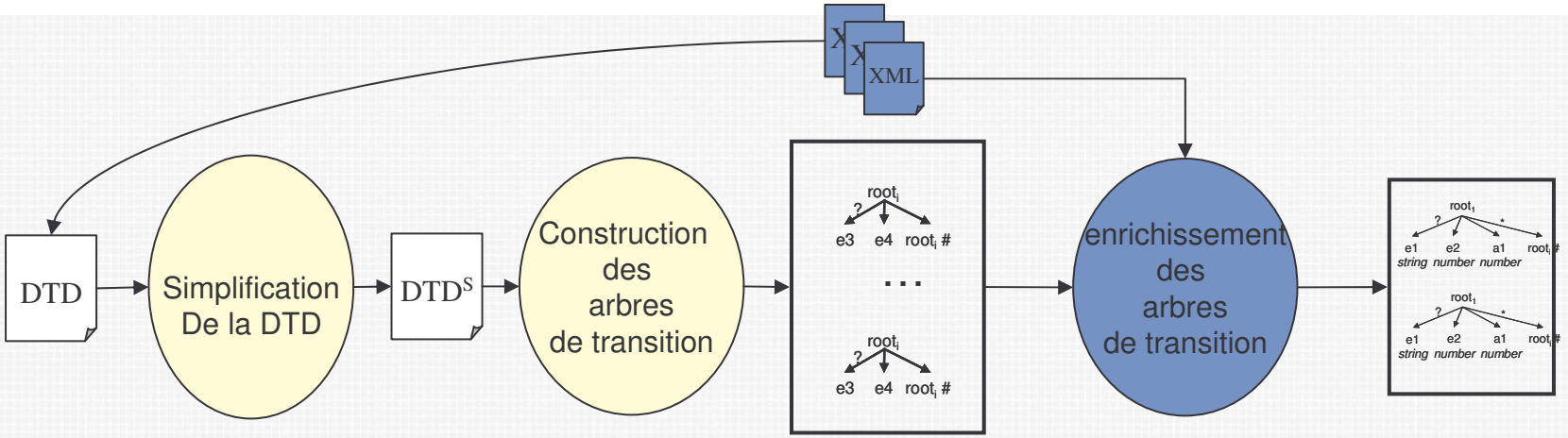
***CreateTree*(*E*, DTD^S)** // *E* le nœud actuel du DTD^S
{
pour chaque élément *e* dans la déclaration de *E* **faire**
{
addChildNode(*E, e*) //ajoute un enfant *e* au nœud *E*
markArcCard (*E, e*) // annote l'arc de *E* vers *e* avec la cardinalité de *e* **{?, *, +, espace}**.
si (*e* est identifié comme racine) **alors**
markNode(*e*, #) //racine déjà identifiée avec **RD1-RD4**
sinon si (*e* contient autres éléments ou attributs) **alors**
CreateTree(*e*, *DTD*^{S})
}
pour chaque attribut *a* dans la déclaration de *E* **faire**
{
addChildNode(*E, a*)
markArcType(*E, a*) //annote l'arc de *E* vers *a* avec le type de *a* **{I, R, F, ID, IDREF}**.
}}

Prétraitement


```
<?xml version="1.0" encoding="UTF-8"?>
<!ELEMENT vente (date, client, produit+, vendeur)>
<!ELEMENT date (#PCDATA)>
<!ELEMENT client (adresse)>
<!ATTLIST client id_client ID #REQUIRED nom CDATA #IMPLIED >
<!ELEMENT adresse (localité, ZIP)>
<!ELEMENT localité (#PCDATA)>
<!ELEMENT ZIP (#PCDATA)>
<!ELEMENT produit (id_produit, prix, quantites, description?)>
<!ELEMENT id_produit (#PCDATA)>
<!ELEMENT prix (#PCDATA)>
<!ELEMENT quantites (#PCDATA)>
<!ELEMENT vendeur EMPTY>
<!ATTLIST vendeur id_vendeur ID #REQUIRED nom CDATA #REQUIRED >
<!ELEMENT description (#PCDATA)>
```

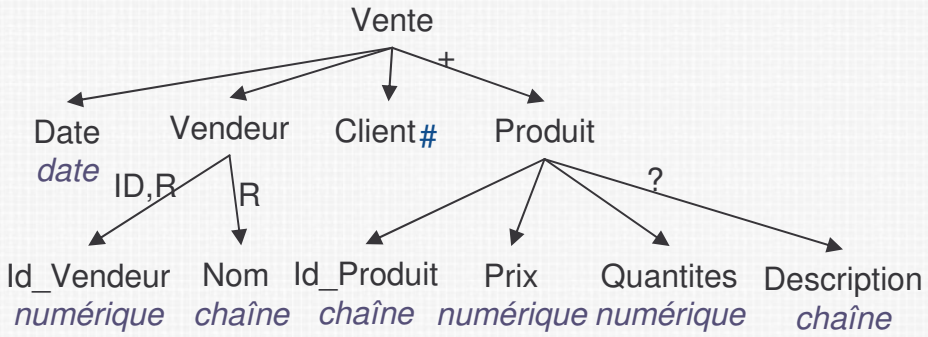


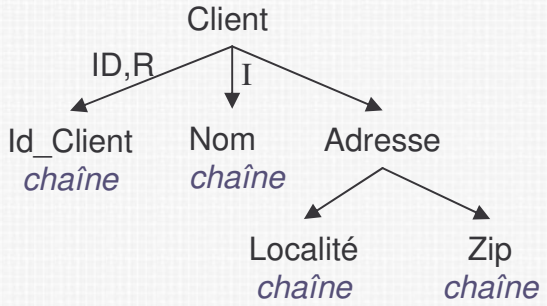
Prétraitement



- Problème
  - l'identification de certains CM se base sur les types de données
  - DTD : Tout est défini comme chaîne de caractères (PCDATA et CDATA).
- Solution
  - interroger un ensemble de documents XML conformes à la DTD
  - pour chaque feuille (non marquée par #), scanner le texte contenu & identifier le type approprié

Prétraitement



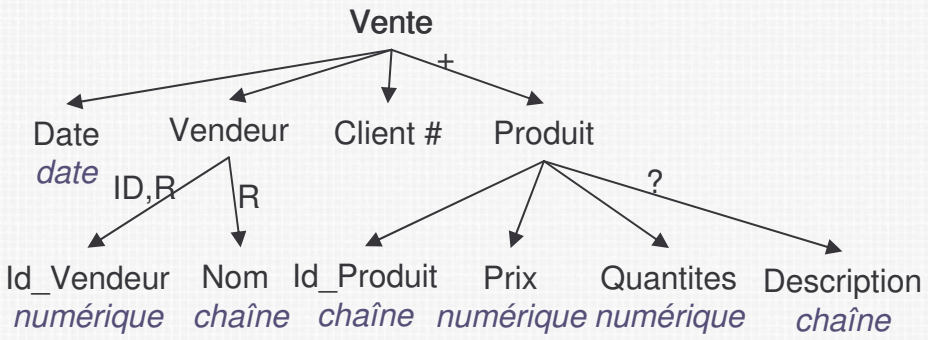


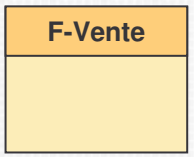
- Identification des faits

Rf : Le sommet de tout arbre de transition contenant au moins une feuille
  - non marquée par #
  - de type numérique
  - dont l'arc de transition est sans étiquette ou étiqueté *R*

Rcf : Un fait est considéré *pertinent* à la prise de décision si l'arbre de transition correspondant possède au moins une feuille annotée par #.
  - La complexité d'un arbre (annotés par #) garantit pour le fait construit une multiplicité d'axes d'analyses (privilège)

- Identification des faits





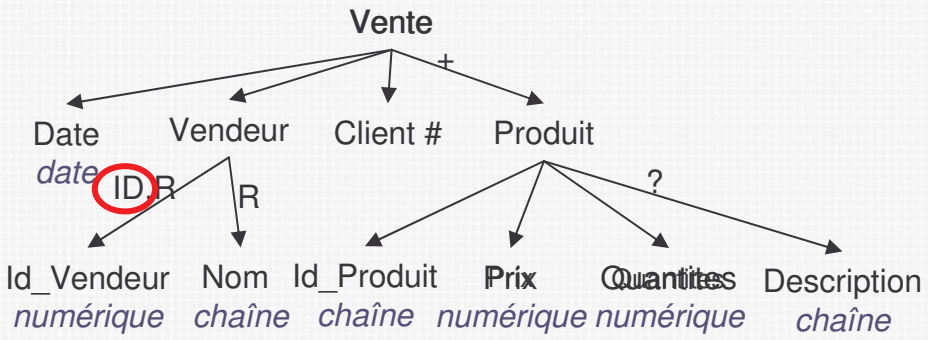
- Identification des mesures

Rm : feuille

- Appartenant à un arbre de transition dont le sommet est identifié comme fait,
- non marquée par #,
- de type numérique et
- reliée par un arc non marqué {I, F, ID, IDREF}

Rcm : Une mesure identifiée est pertinente à la prise de décision si elle est liée directement au nœud fait.

- Identification des mesures



F-Vente
Prix Quantites

- ## Identification des dimensions

Rd1 : nœud non terminal de profondeur deux d'un arbre-fait est une dimension pour le fait de cet arbre.

Rd2 : nœud *N* de profondeur deux, annoté par # et appartenant à un arbre-fait est une dimension (*D-N)* pour ce fait.

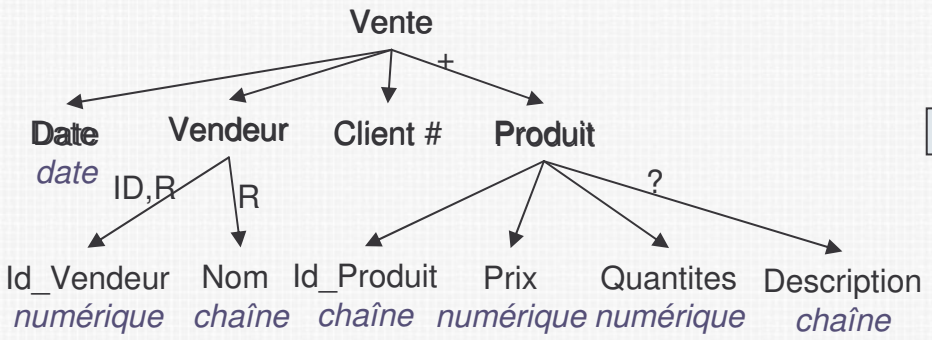
Rd3 : nœud *N1* ayant un arc sortant annoté ID vers un nœud référencé par *N2* dont le parent est un fait *F* est une dimension (*D-N1*) pour *F*.

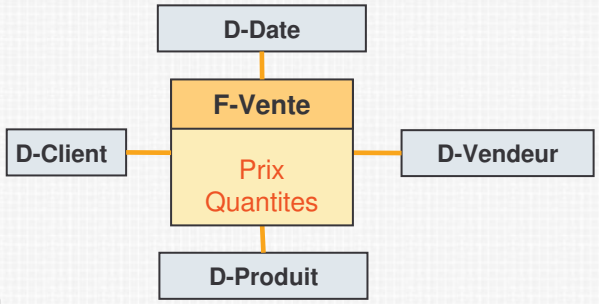
Rd4 : nœud *N* de profondeur deux de type *Date* et appartenant à un arbre-fait F estampille F et lui construit une dimension temporelle *D-Date-N*.

Rdc : Une dimension construite sur un nœud *N* est considérée pertinente si l'arc entrant à *N* est soit non étiqueté soit étiqueté +.

- Identification des dimensions

Rd2



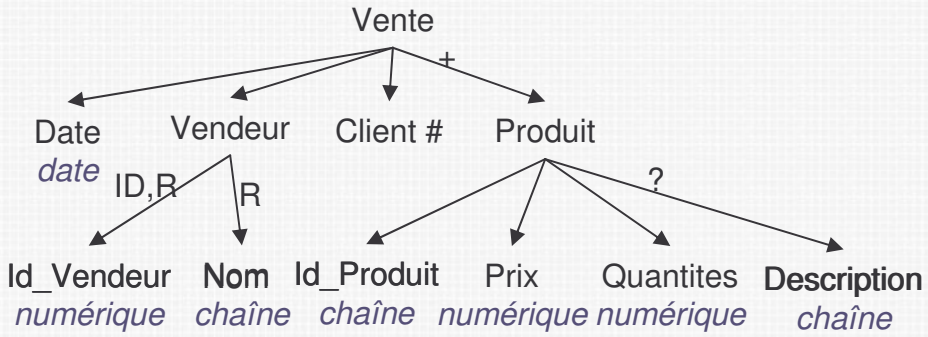


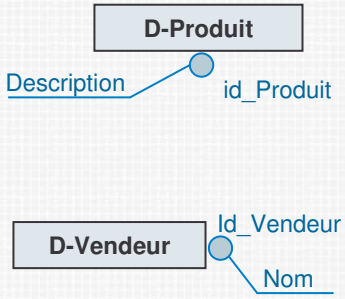
Extraction de Concepts MD

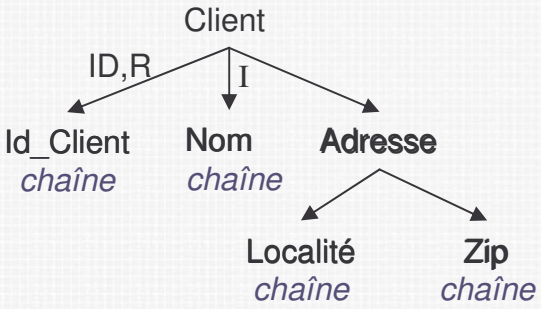
- Identification des hiérarchie
  - **Extraction des identifiants**
    - Deux règles (Rid1, Rid2)
  - **Identification des paramètres de rang 2**
    - Quatre règles (Rp1, Rp2, Rp3 et Rp4)
  - **Identification des paramètres de rang > 2**
    - l'application récursive des règles Rp1 à Rp4 sur les nœuds identifiés par Rp1 à Rp3 produit des paramètres de rang supérieur à deux.

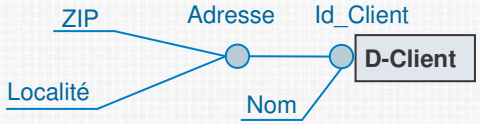
- Identification des attributs faibles
  - **Raf :** Tout nœud terminal, non marqué par # et lié à un nœud-paramètre *P* (*i.e.,* identifié comme paramètre) via un arc non annoté par ID ou IDREF est un attribut faible pour *P*.

- Identification des hiérarchie & des attributs faibles

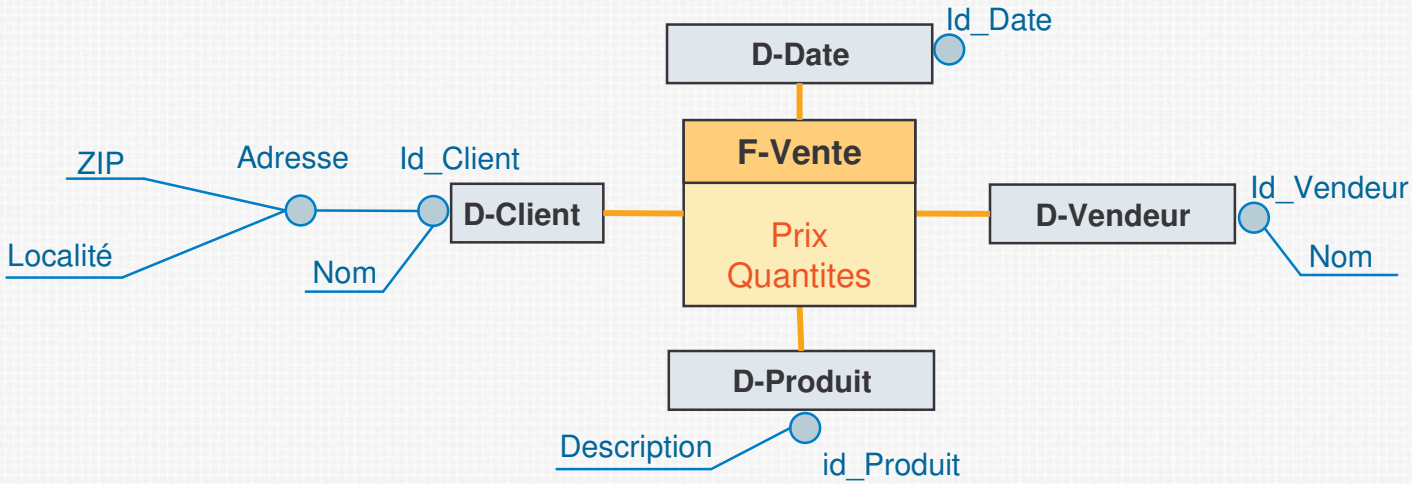








Extraction de Concepts MD



Bilan & Perspective

- Bilan
  - Méthode automatisable, ascendante de construction de schémas en étoile à partir d'une source XML
    - se base sur des règles d'identifications
    - affecte un niveau de pertinence aux concepts extraits
    - assiste le passage au niveau logique (concept-source)
  - Les schémas en étoile obtenus peuvent être adaptés aux besoins analytiques du système de pilotage (Validation)
- Travaux en cours
  - Outil XML-CAME (en cours de développement)
  - Évaluation sur des exemples complexes

Merci pour votre attention

Yasser.Hachaichi@fsegs.rnu.tn