Traitement et Exploitation de Masse de Données

Rencontre JIRC – 16 avril 2014

Ladjel BELLATRECHE

bellatreche@ensma.fr

http://www.lias-lab.fr/members/bellatreche

ENSi Poitiers – INGÉNIERIE pour la PROTECTION de L'ENVIRONNEMENT

isae – Institut Supérieur de l'Aéronautique et de l'Espace – ENSMA

1431 Université de Poitiers

Agenda

- **Presentation of the LIAS**
- **Domain Ontologies**
- **Management of Big Data & Queries**
- **Collaboration between LI & LIAS**

Paysage académique

Directeur : Patrick COIRAULT

Directeur Adjoint : Emmanuel GROLLEAU

Tutelles

Fédération

Transfert

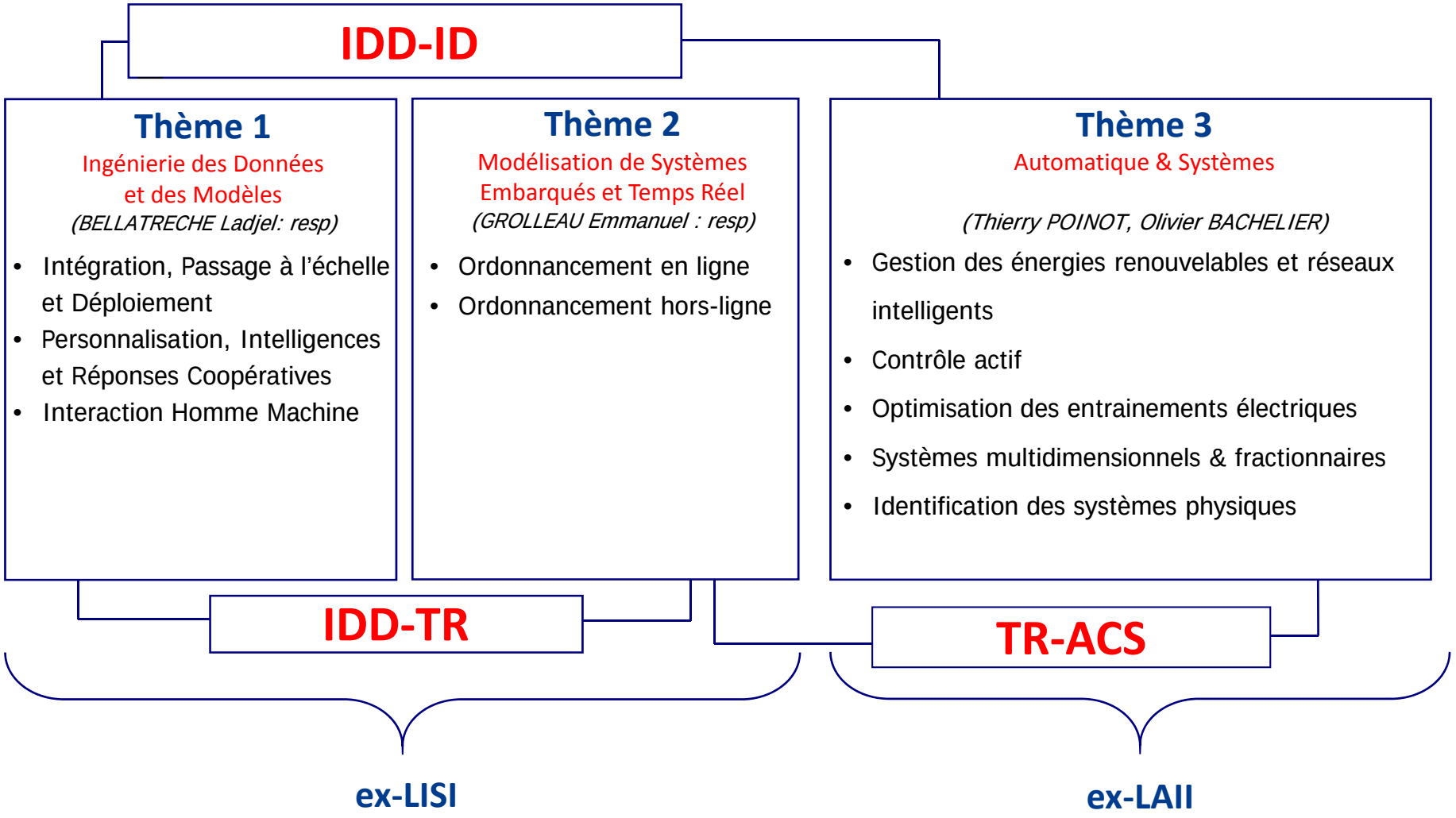
Quelques chiffres

- **43 permanents**
  - 12 Professeurs
  - 23 Maîtres de Conférences (7 HDR)
  - 2 Professeurs Agrégés
  - 1 Ingénieur de Recherche
  - 5 Administratifs & Techniques
- **33 doctorants**
- **~ 35 stagiaires**
- **Surfaces : 1319 m²**

Partenaires

- **Industriels ou semi-publics**
  - EADS, Airbus, CFCA, Dassault, Orange, Barco, ACCO, ITRON, Leroy Somer, EDF, RTE, SOCOMEC, Kapteos, Renault, Alstom Power, Areva, IFP, Teradata – USA, IBM – India, etc.
- **Académiques nationaux**
  - INRIA, ONERA, LRI, IRCCyN, LORIA, IFP, CEA, ENSEIHT, etc.
- **Internationaux**
  - Belgique, Hongrie, Italie, Islande, Hollande, Pologne, Espagne, Royaume Uni, Allemagne, etc.
  - Tunisie, Algérie, Maroc, USA, Canada, Vietnam, Venezuela, Australie, Brésil, etc.

Organisation Scientifique



Ingénierie des Données et des Modèles (IDD)

Ladjel BELLATRECHE (Resp. Thème)

bellatreche@ensma.fr

http://www.lias-lab.fr/members/bellatreche

ENSi Poitiers – INGÉNIERIE pour la PROTECTION de L'ENVIRONNEMENT

isae – Institut Supérieur de l'Aéronautique et de l'Espace – ENSMA

Université de Poitiers – 1431

Permanents (10 +1)

Ladjel BELLATRECHE
Pr - ENSMA

Mickael BARON
IR à l'ENSMA

Allan FOUSSE
MCF - UP

Selma KHOURI
MCF ESI - Alger

Patrick GIRARD
Pr - UP

Brice CHARDIN
MCF – ENSMA

Laurent GUITTET
MCF - ENSMA

David MARCHEIX
MCF - UP

Allel HADJ ALI
Pr - ENSMA

Zoé FAGET
MCF - ENSIP / UP

Stéphane JEAN
MCF - UP

Doctorants : Situation Actuelle

Okba BARKAT

Fethi BELABDELLI

Paule BONDIOMBOUY

Selma BOUARAR

Géraud FOKOU

Thomas LACHAUME

Ahcène BOUKORCA

Samia BOULKRINA

Zahira CHOUIREF

Zouhir DJIHANI

Nadir GUETMI

Rima BOUCHAKRI

Soumia BENKRID

Bery MBAIOSSOUM

Linda MOHAND OUSSAID

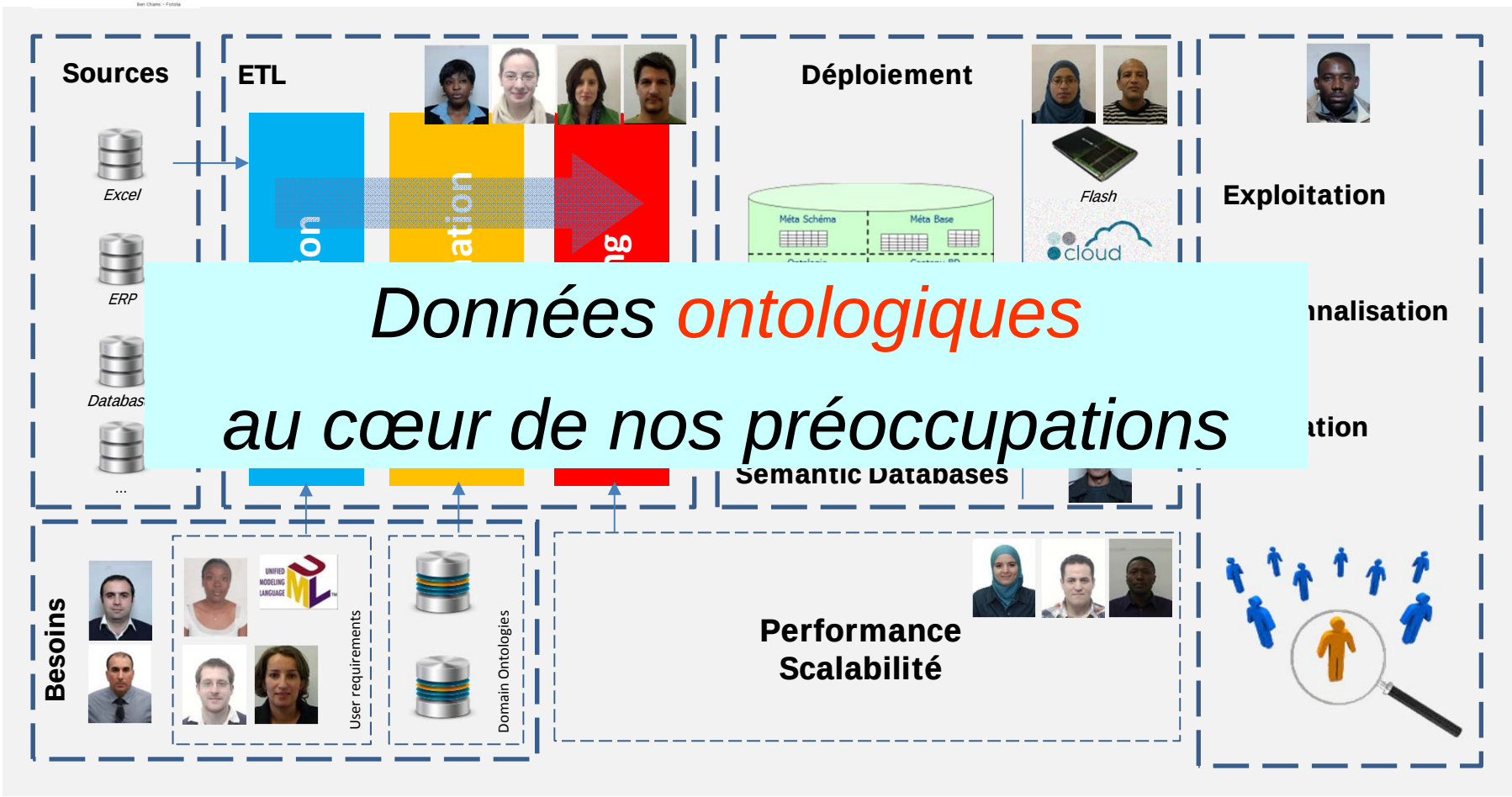
Kevin ROYER

Cifre

Abdel ALI ZAZOU

Cheikh SALMI

Cycle de Vie des Applications Avancées



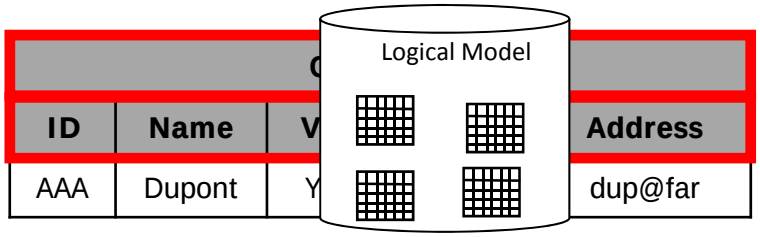
Données ontologiques

au cœur de nos préoccupations

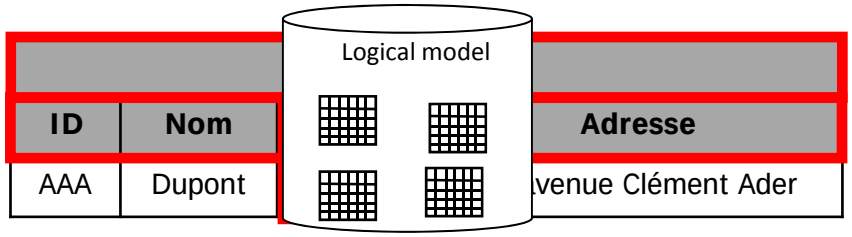
Data Source Integration: Problem

Semantic heterogeneity

- Conflicts of naming
- Conflicts of structures
- Conflicts of units of measure



Bank 1



Bank 2

Survey:

[Wache and al. 2001] *Ontology-based Integration of Information — a Survey of Existing Approaches*.

[Noy 2004] *Semantic Integration: A Survey Of Ontology-Based Approaches.*

Database Design : Problem

- ↗ **Data Sources are not Designed to be Integrated in the Future**
- ↗ **Lack of Semantic**

Exercise:

→ **Propose a Conceptual Model (Peter Chen) for the following application:**

Cahier des Charges

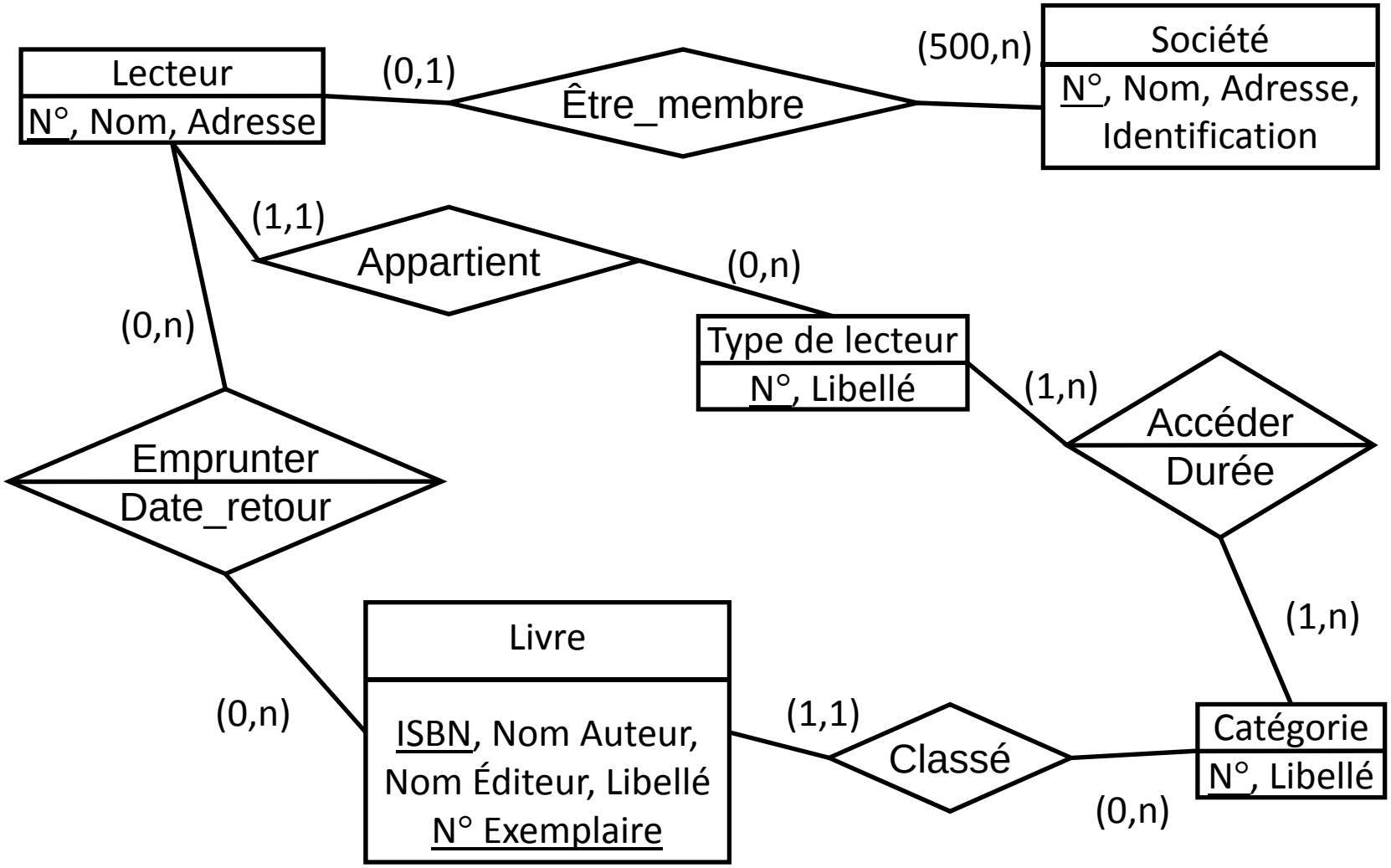
La bibliothèque enregistre chaque lecteur à qui elle donne un numéro de lecteur. Elle lui prend son nom et son adresse. Le lecteur peut éventuellement être membre d'une société adhérente. On enregistre alors l'identification de cette société.

Un lecteur peut emprunter plusieurs livres chaque jour. A chaque prêt, on associe une «date de retour au plus tard ». Un lecteur appartient à un « type de lecteur ». Ce type lui permet d'avoir ou non accès à certaines catégories de livres. La durée du prêt dépend de la catégorie du livre et du type de lecteur. Elle est la même pour tous les livres d'une catégorie donnée empruntés par un quelconque lecteur d'un type donné. Un livre est caractérisé par son numéro d'inventaire.

Il est nécessaire de connaître sa catégorie, le nom de son auteur, son éditeur, ainsi que le nombre de ses différents exemplaires disponibles. L'édition, lorsqu'elle existe, est également à connaître.

La catégorie d'un livre se repère par un numéro et possède un libellé. Il en est de même pour le type de lecteur. Une société adhérente possède un nom et une adresse ; elle s'engage à envoyer un minimum de 500 lecteurs.

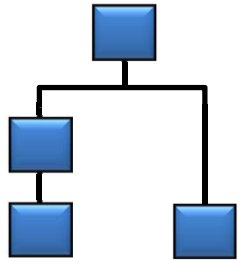
Une Solution…



Ontologies as a Solution to Reduce Heterogeneities

An ontology = conceptualization of a domain

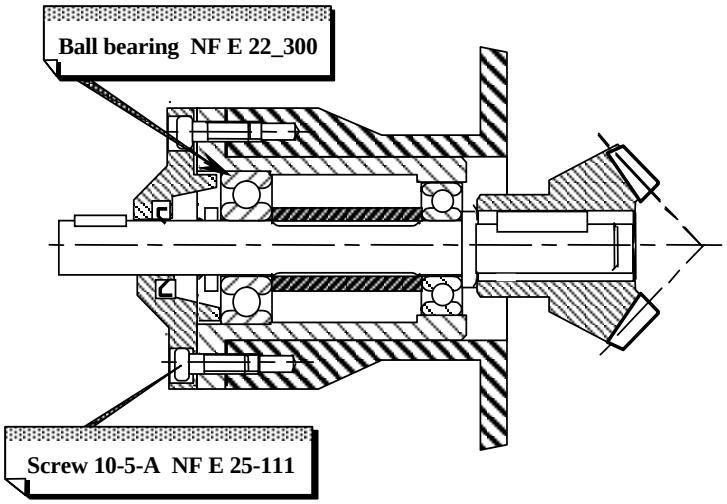
classes and properties



- **Formal**
  - _ Reasoning
  - _ Consistency
- **Consensual**
  - _ Covers a wide range of applications
  - _ Many partners involved (e.g, Gene / Product ontologies)
- **Capability to be referenced** (concepts dictionary)
  - _ Identifier for each concept
  - _ Independent of any environment

Modèle d'ontologie : PLIB

- **Origine : Fin '80, projet international de normalisation en ingénierie**



But initial de la série ISO 13584 (PLIB) :

- modéliser
- échanger
- archiver
- publier

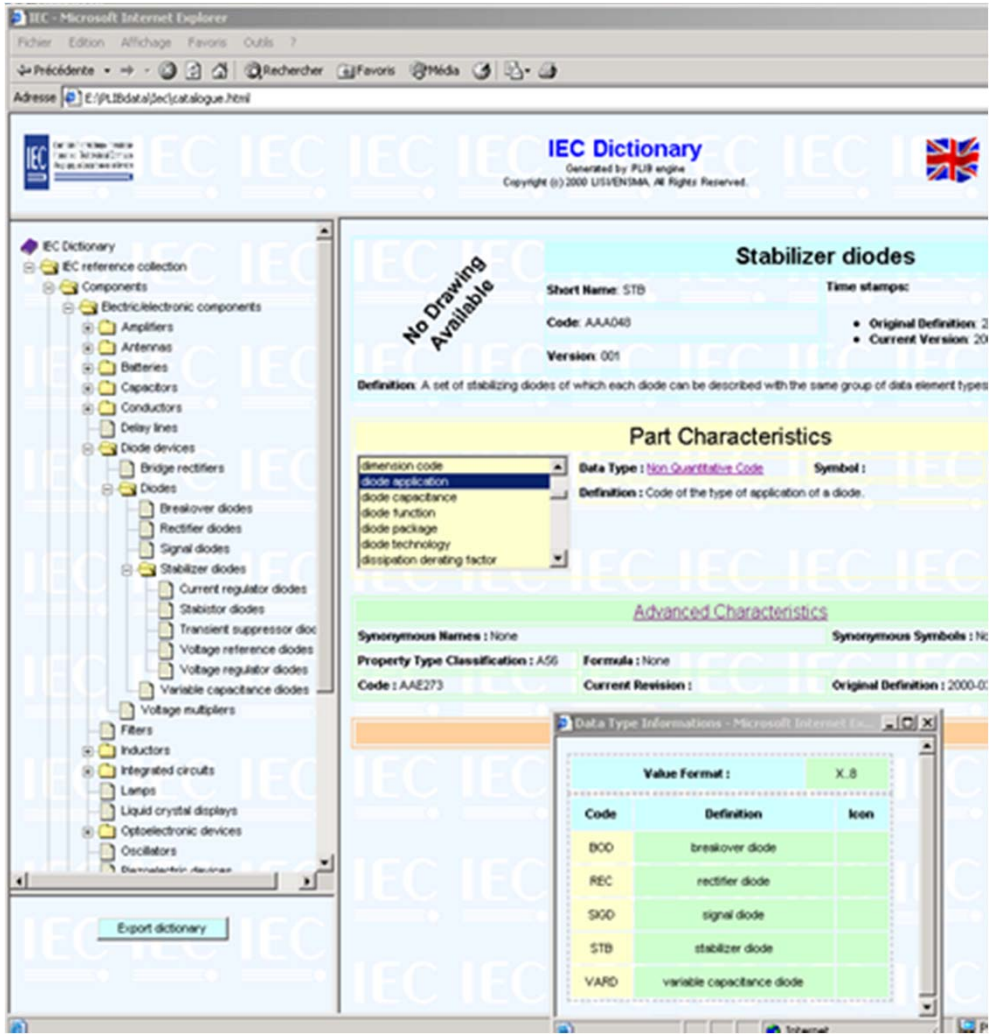
les catalogues de composants industriels:

passer des documents aux données

- **Principes de base**
  - ➢ Un catalogue définit une ontologie et des données à base ontologique

Exemple d'ontologie

- **Electronique : Ontologie IEC 61360-4**
  - Description de composants électroniques
  - Standardisée (nombreux pays impliqués)
  - 190 classes
  - 1026 propriétés



Modularity

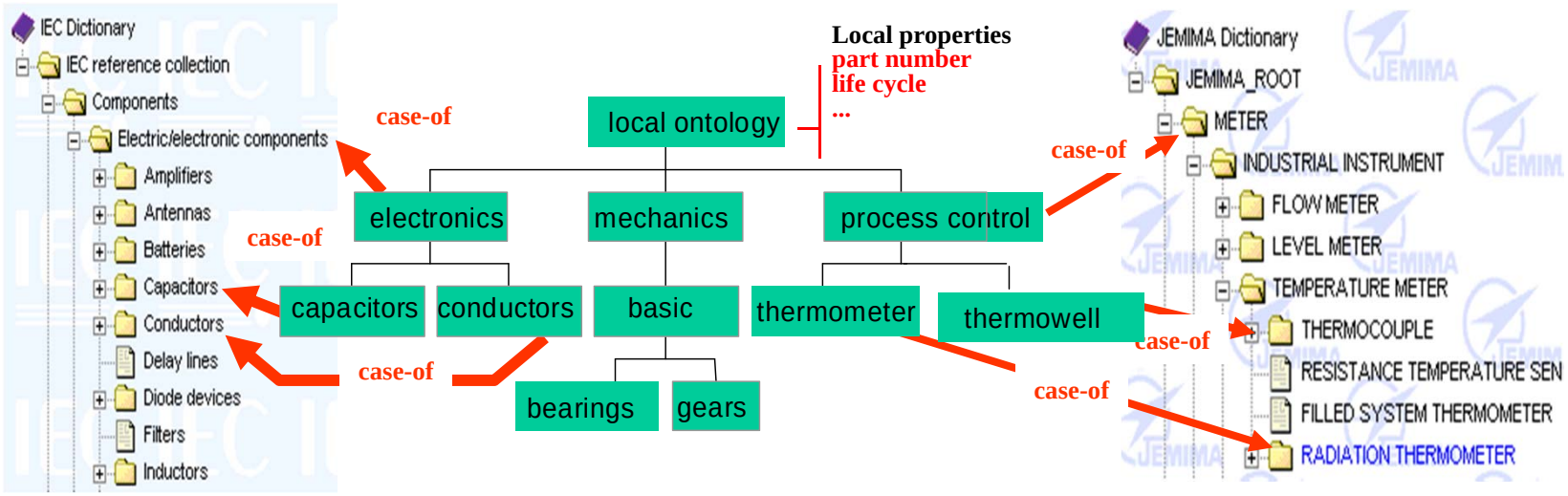
- Hierarchies of shared ontologies may not fulfill user requirements
- Some properties may not have been defined

→ Need to "customize" ontologies

❑ A specific relationship: case-of

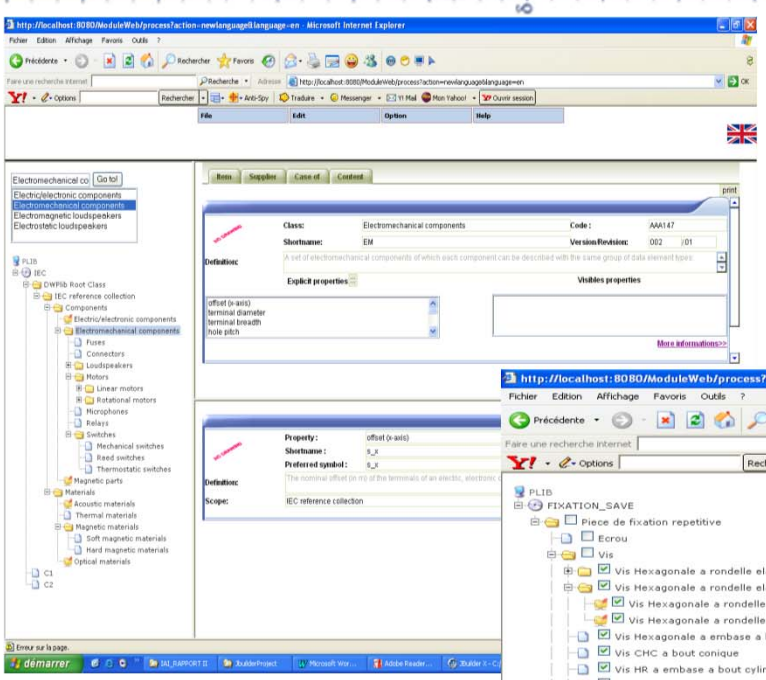
✓ Subsomption without inheritance

✓ The local ontology imports properties

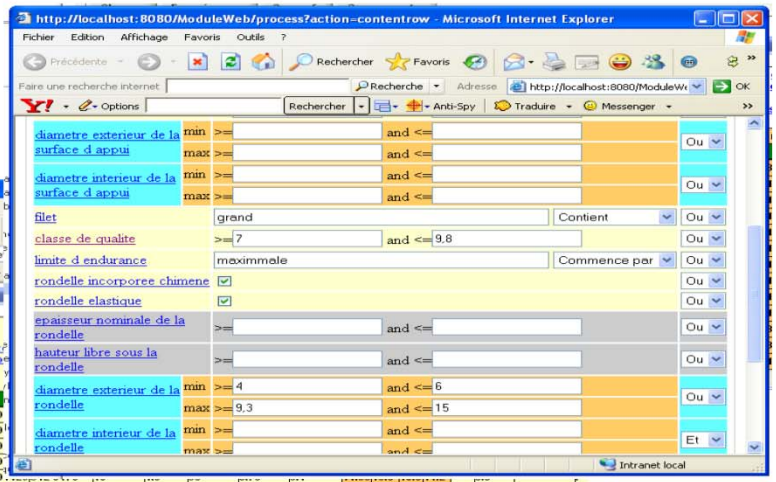


PLIB Tools

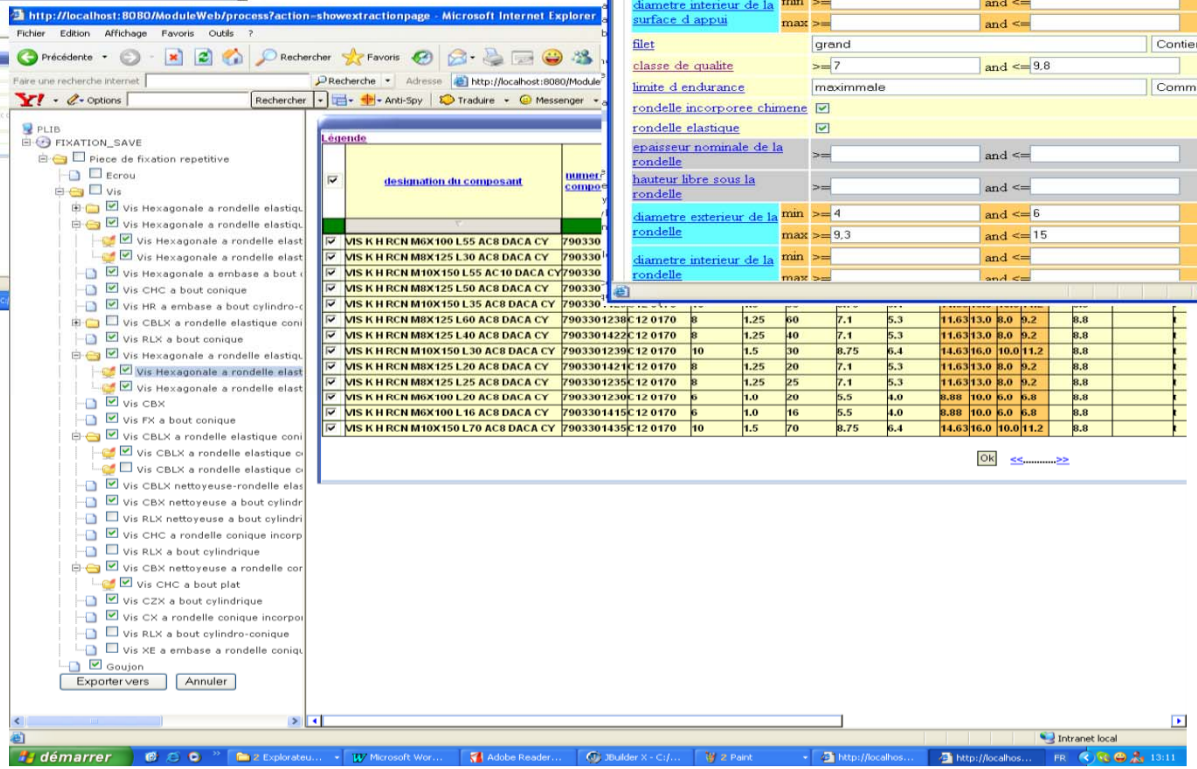
Ontology Editor



Query interface



Class extension management (creation, extraction, ...)



Ontologies in Industry: Renault Project

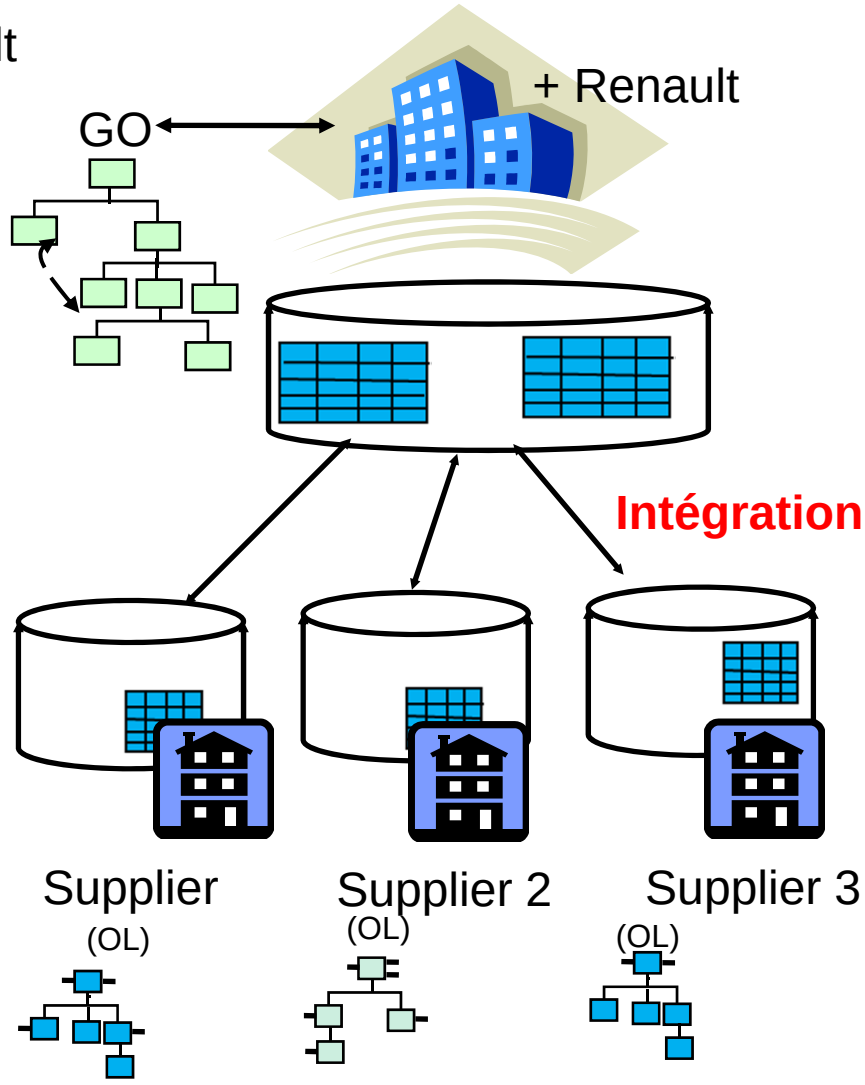
(1) Definition of a global ontology (GO) Renault

(2) Definition of a Local Ontology (LO) by each supplier

(3) Mapping of LOs on OP

(4) Export by suppliers
- in terms of the GO
- In terms of the LO

Put in practice with **170 subcontractors and suppliers**



Ontologies dans Patrimoine

- **Contribution des Ontologies dans la Numérisation et l'usage des Images du Patrimoine**
- **Collaborations**
  - Centre d'Etudes supérieures de Civilisation médiévale (CESCM)
  - Laboratoire LIAS (LIAS – ENSMA)
  - Département SIC du laboratoire XLIM (XLIM – SIC)

- **Financement** : CPER

Problématique

- **Corpus des inscriptions de la France Médiévale (notices)**

46. **Saint-Gildas-de-Rhuys**, **ancienne abbatiale Saint-Gildas** – Demande de prière pour Geoffroi.

Demande de prière.
Église, chapelle d'axe, mur est, à 1,49 m du sol. Taille du bloc de granit : 32 × 38,5 cm. Hauteur des lettres : de 5 à 6 cm. L'inscription a été repassée à la peinture noire à une date indéterminée.
Datation proposée : XII^e siècle [datation paléographique].

Inscription vue en place par le CIFM en 1997. Texte établi d'après l'original.
GRAND, « Saint-Gildas-de-Rhuys », 1919, p. 374 [reproduction et commentaire].

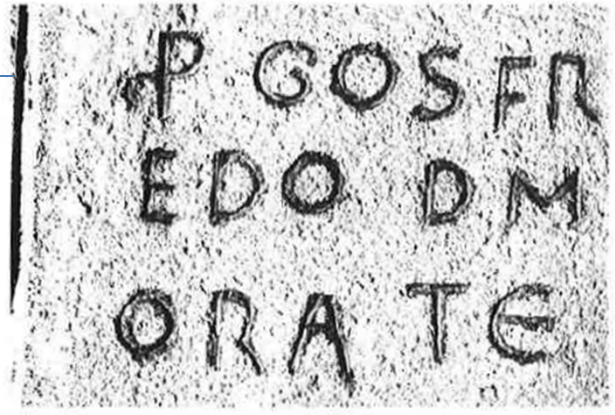
Disposition horizontale sur trois lignes, sans cadre ni réglure. Écriture majoritairement capitale, assez irrégulière et de grand module. Deux onciales : *G* de *Gosfredo* et *E* de *orate*. Pas de ponctuation. Abréviation : *P* de *pro* et *DM* pour *Deum* (pas de tilde).

Localisation

Monument

Datation

Image



Ecriture

P GOSFR
EDO DM
ORATE

P(ro) Gosfredo D(eu)m orate.

Priez Dieu pour Geoffroi.

Exemple (Extrait du Volume 23)

- **Besoin de représentation informatique (données et images)**
  - ➢ Pour rechercher, échanger, archiver…

Solution Proposée

- **Enrichissement de la base de données par des modèles de connaissances : des ontologies**

Id	Localisation	Datation	Monument	Ecriture	Image
46	Saint-Gildas	XIIème siècle	Ancienne Abbaye	Capitale assez irrégulière …	

- **Intérêt pour la recherche de notices**
  - Quelles sont les inscriptions localisées en Bretagne ?
  - Quelles sont les inscriptions datées entre 1050 et 1250
  - Quelles sont les inscriptions faites dans un monastère ?

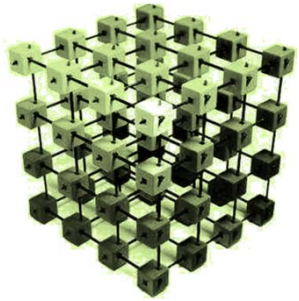
Big Data

Competitive

- ➔ Acquisition of data
  - ❑ Data Warehousing
- ➔ Storage of data
  - ❑ Evolution and diversity of Storage Devices
- ➔ Efficient Query Processing
  - ❑ Query interaction
- ➔ Personnalisation & recommandation

❑ **Two Tendencies**

- ❑ **Big Data outside DBMS**
- ❑ **Big Data inside DBMS**



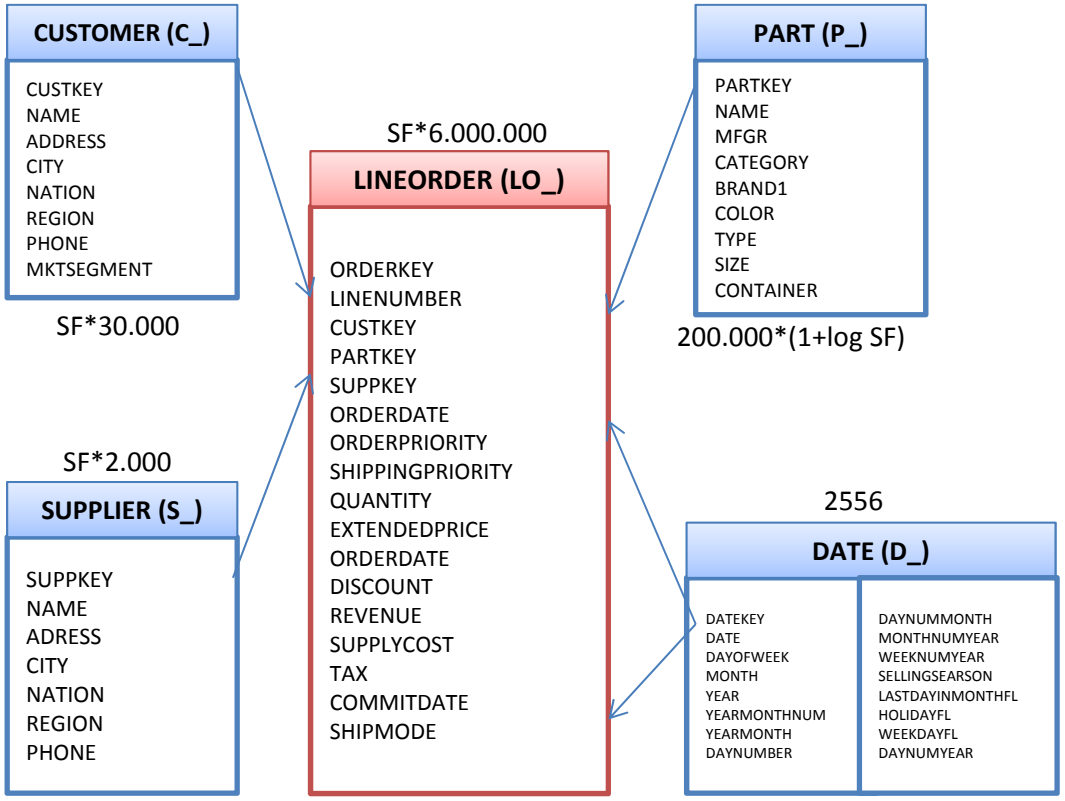
Relational Data Warehouses

➔ Revisit data warehousing Technology

➔ The revisited phase: Physical Design

Data Warehousing & Business Intelligence

Star Schema Benchmark



Star join query

```
select d_year, c_nation,
sum(lo_revenue - lo_supplycost)
 as profit
from date, customer, supplier,
 part, lineorder
 where lo_custkey = c_custkey
 and lo_suppkey = s_suppkey
 and lo_partkey = p_partkey
 and lo_orderdate = d_datekey
 and c_region = 'AMERICA'
 and s_region = 'AMERICA'
 and (p_mfgr = 'MFGR#1'
     or p_mfgr = 'MFGR#2')
 group by d_year, c_nation
 order by d_year, c_nation ;
```

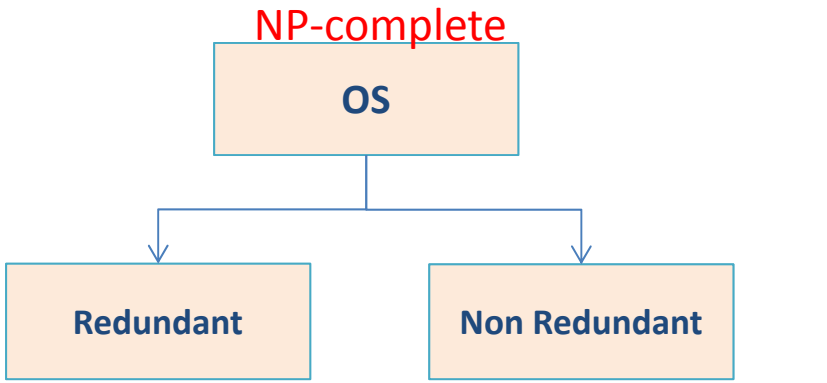
→ **Each join operation passes through the fact table**

→ **Presence of selection on dimension tables**

→ **Aggregations**

Physical Design

- A crucial issue for query performance [Chaudhuri'07]
  - ➔ Selection of optimisation structures (OS)



Indexes (Ix)
Materialized Views (MV)
…

Horizontal Fragmentation (HF)
…

Constraints
Storage space
Maintenance

Constraints:
Fragmentation Threshold
…

Physical Design Problem:

Given:

- A DW schema
- Query Workload Q
- A Set Optimization Structures S
- A Set of Constraints related to C

Objective: select schemes of SO Optimizing Q and Satisfying C

Selection Modes:

✓Isolated: |OS| = 1. Ex. index selection [Microsoft]

✓Multiple : |OS| > 1. Ex. joint selection of indexes and materialized views [IBM DB2].

Problème de la Fragmentation Horizontale

Problème d'optimisation à contrainte

Entrées :

- Entrepôt de données
  - Tables de dimension $D=\{D_1, D_2, \ldots, D_d\}$
  - Une table des faits F
- Charge de requêtes les plus fréquentes
- **W : seuil (fixé par l'administrateur)**

Sorties :

- Ensemble $D' \subseteq D$ des tables de dimension fragmentées
- Ensemble de N fragments de faits $F_1, \ldots, F_N$

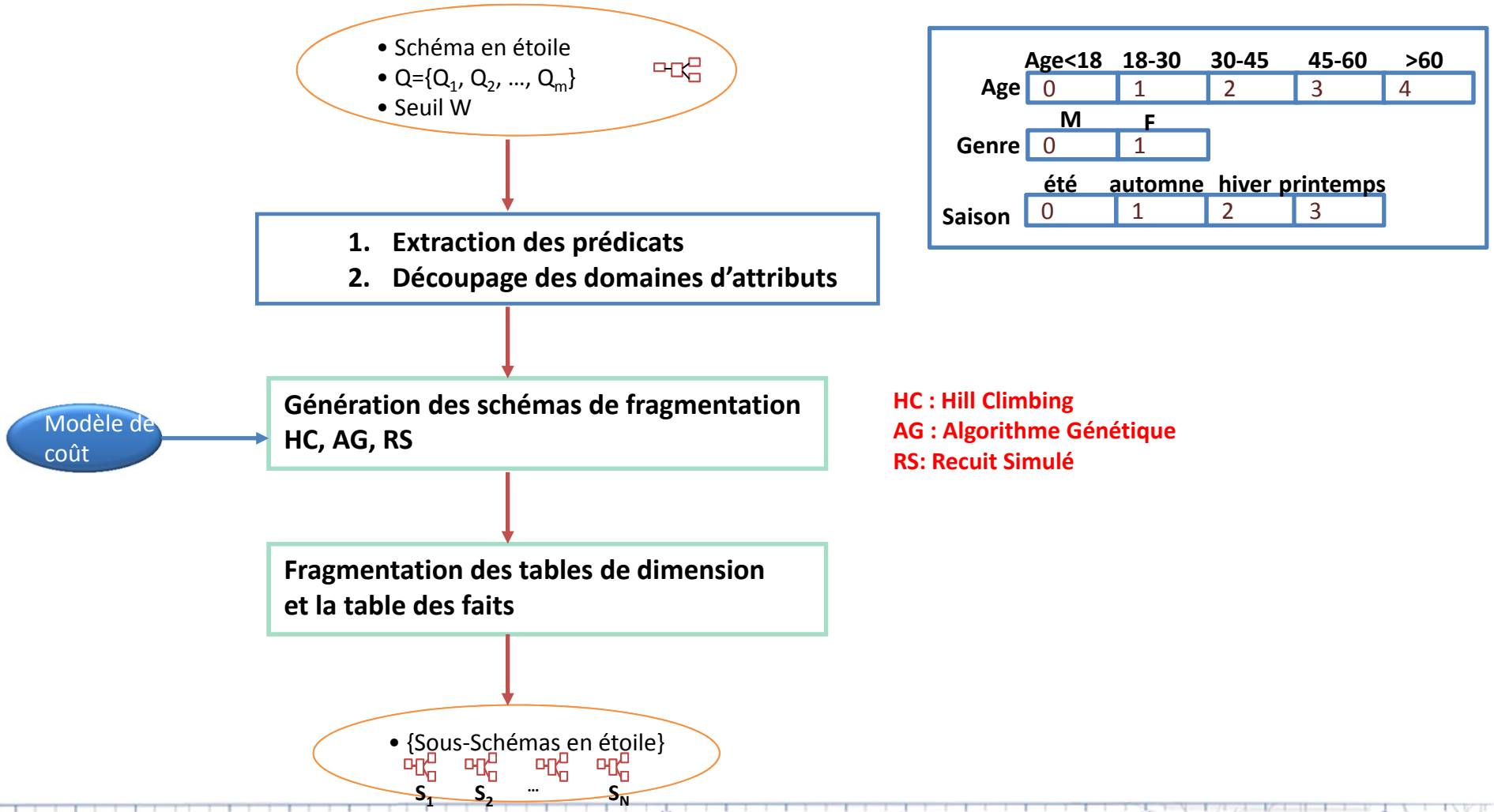
Objectifs :

- Réduire le temps d'exécution de Q
- $N \leq W$

Etude de complexité (IJDWM 2009): P. Richard (TR)

- **Problème de fragmentation horizontale à un seul domaine (PFHSD)**
  - Une seule table de dimension D
  - Un seul attribut A dans D
- **Réduction à partir du problème 3-Partition**
  - 3-Partition NP-Complet
  - PFHSD NP-Complet
- **Notre problème de fragmentation est plus compliqué**
  - Plusieurs tables de dimension
  - Plusieurs attributs par table de dimension

Approches de fragmentation

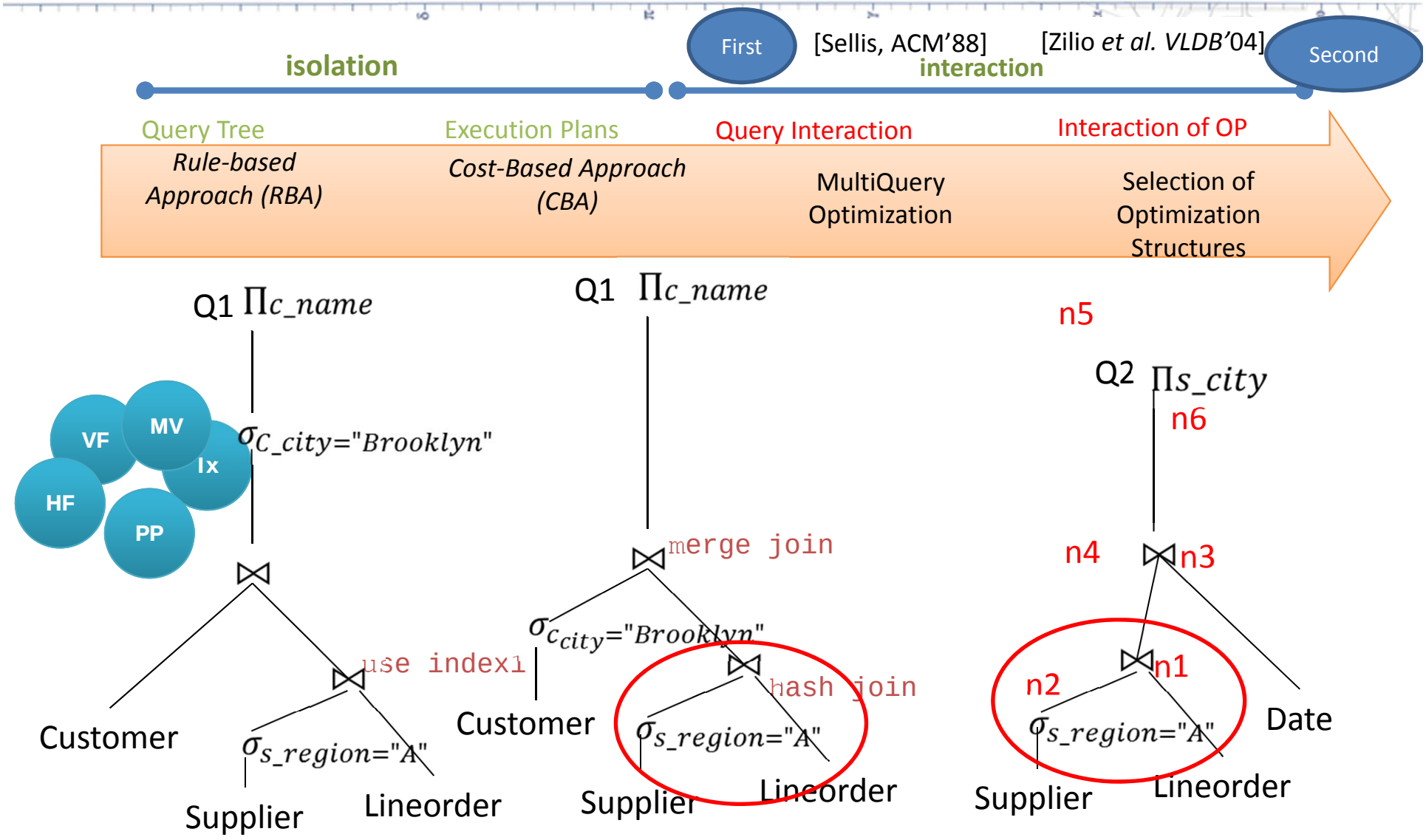


Management of Big Queries

- New Applications:
  - Routinely and mixed queries
  - Alibaba system daily runs more than 20,000 queries due to the business needs and 40% of the statements share similar data operations
  - Star Join Queries

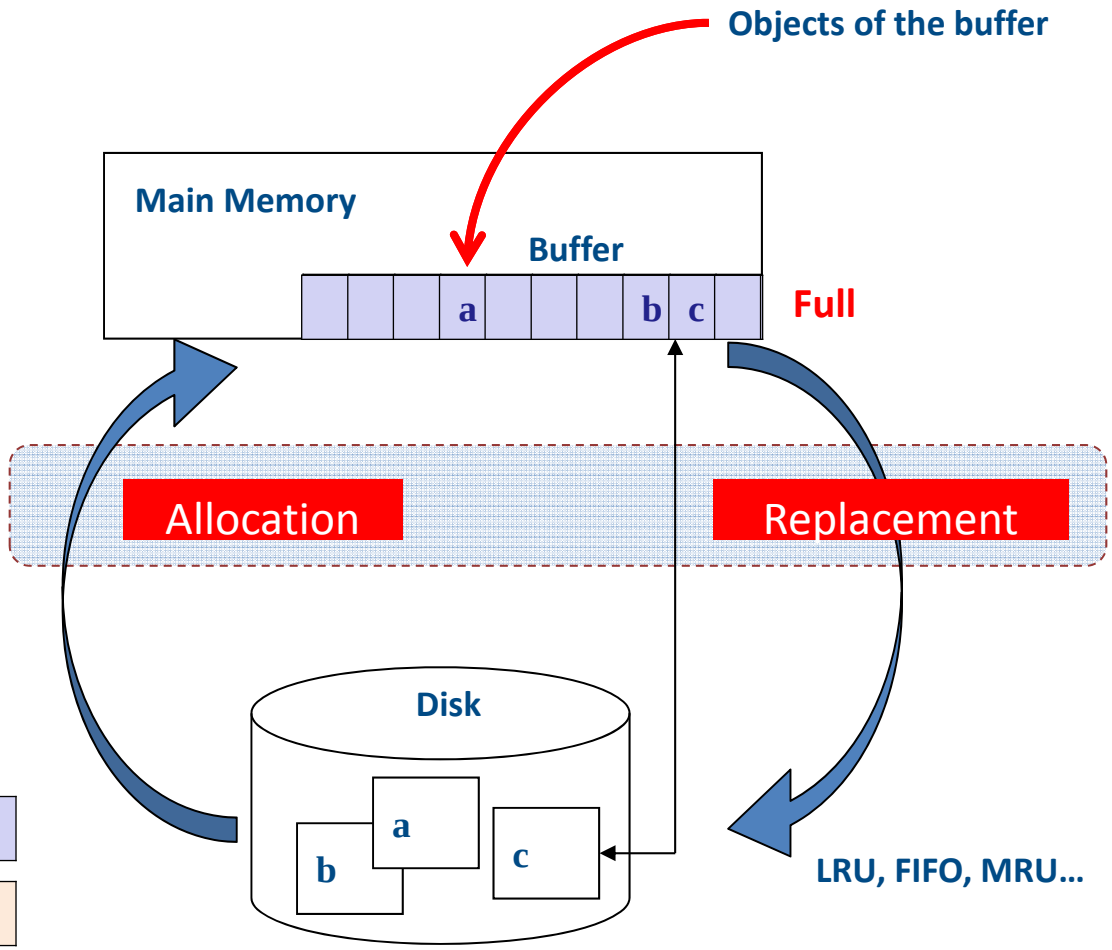
- Recommendation increases the interaction between queries (sharing operations)

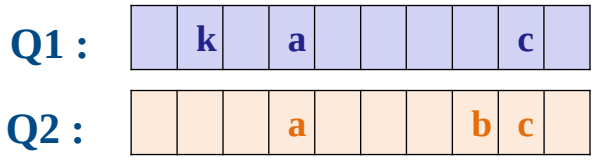
Evolution of Query Processing



Interaction in Buffer Management Problem

Principle:

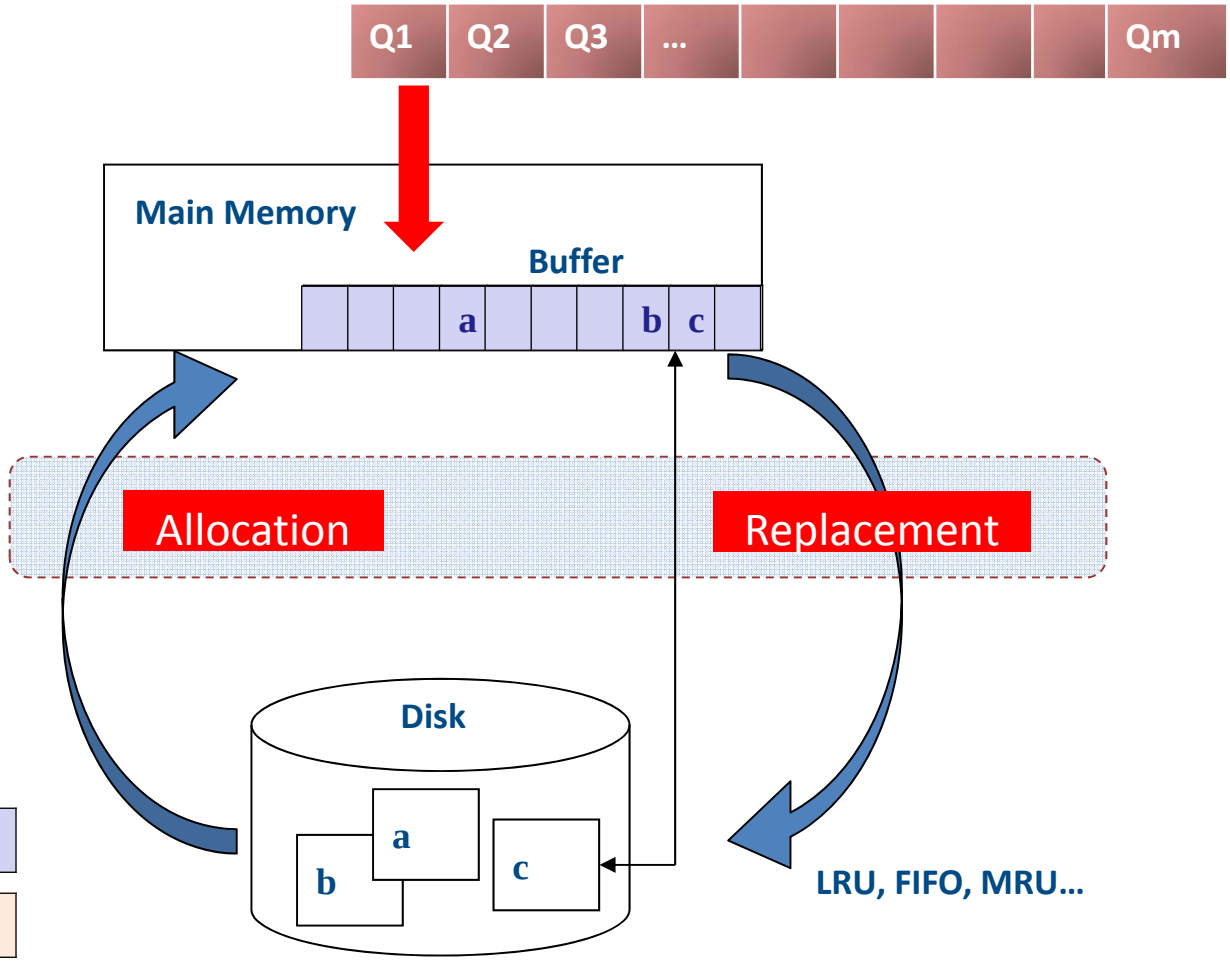




[Härder,84],[Chou,85], [Cornel,89],[Roy,00], [Marco,08]…

Interaction in Query Scheduling Problem

Principle :



Q1 : k a c

Q2 : a b c

[Zhou,09], [Schild,98] , [Phan,08]…

BMP-QSP: Similarities & Complementarities

BMP

- Inputs: RDW, Workload Q
- Output: Bufferization Strategy
- Objective : Reducing the cost of Q
- Constraint : Buffer size B

QSP

- Inputs: RDW, Workload Q, Bufferization Strategy
- Output: Scheduled Queries
- Objective : Reducing the cost of Q

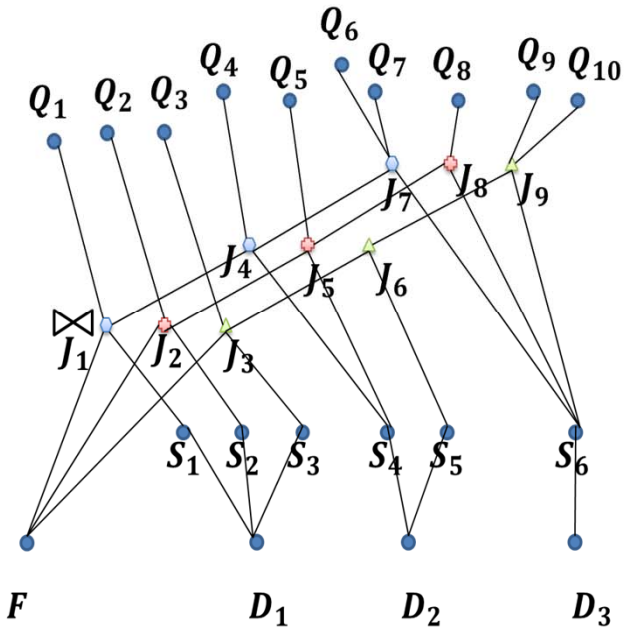
⇨ Two Hard Problems
⇨ Same Data Warehouse Schema
⇨ Same Workload
⇨ Same Objective Function

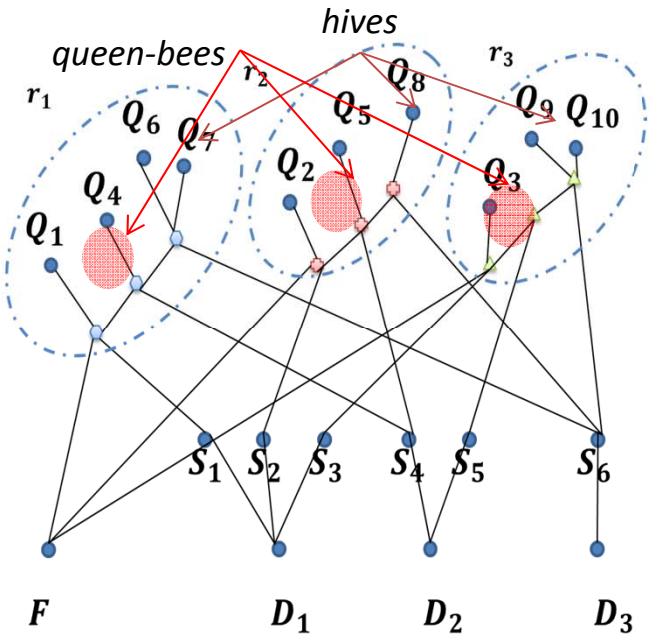
➔ Usually they are studied in isolation way, except the work of [Diwan et al. 06]

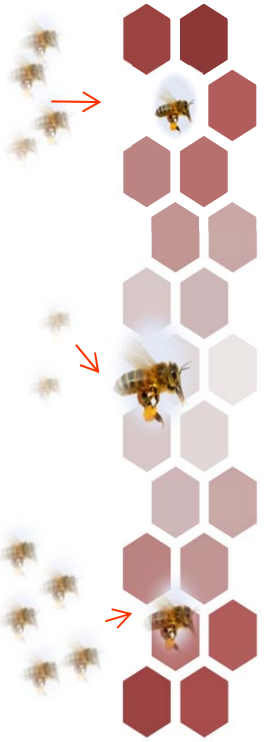
Formalization of the Joint Problem (BMQSP):

- **Inputs:** RDW, Workload Q = $\{Q_1;Q_2; \ldots;Q_m\}$
- **Constraint**: Buffer Size B
- **Outputs**:
  1. Scheduled Queries $SQ=\{SQ_1; SQ_2; \ldots SQ_m\}$
  2. Corresponding Buffer Management scenario

Bee-Inspired Algorithm for BMP



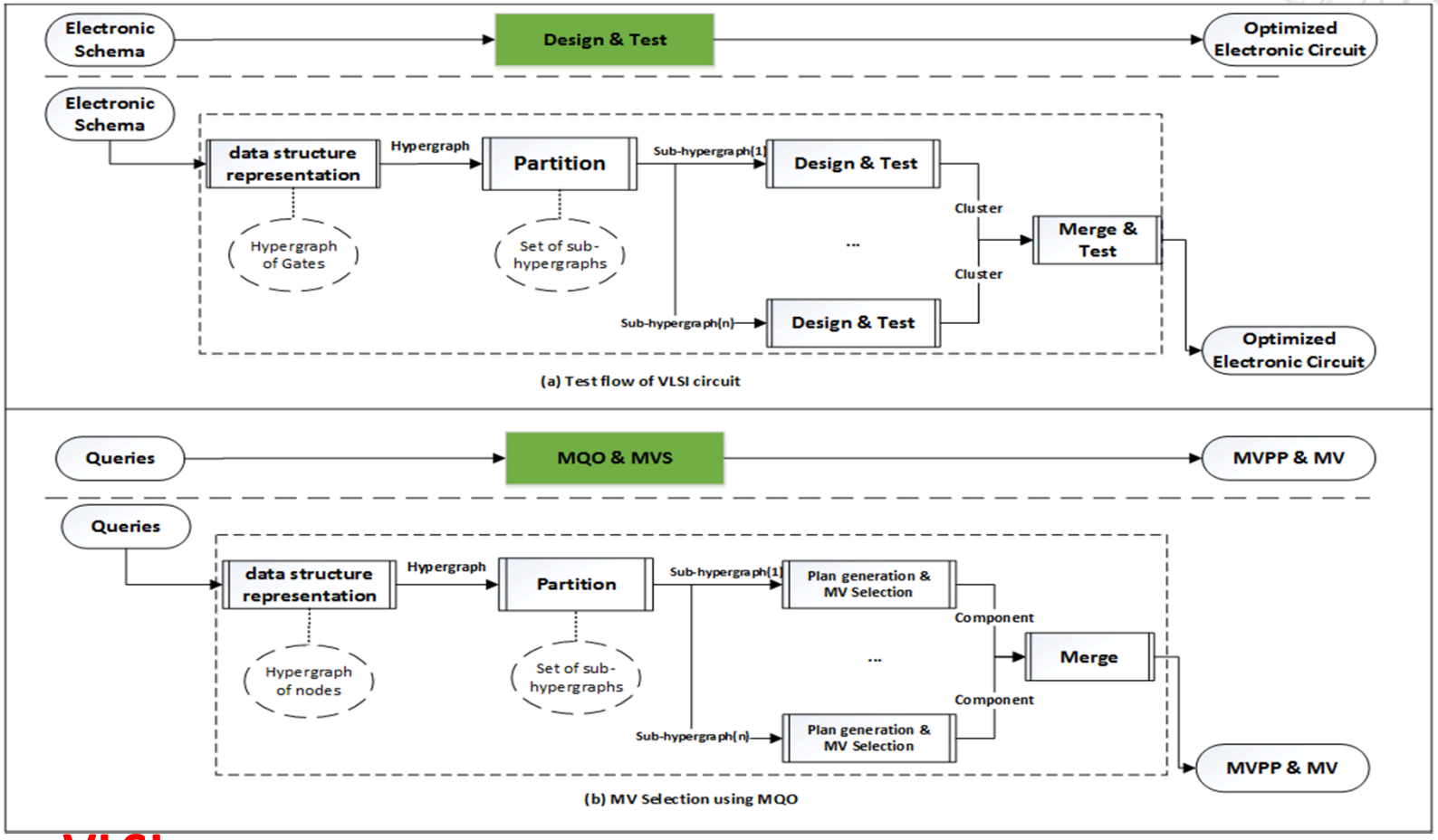




→ Q1, Q2 and Q3

1) *Identification of correlated queries*
2) *Query Grouping (hives)*
3) *Choice of the queen-bee*
4) *Local scheduling (hive level)*

Graph Theory as Solution for Big Queries



VLSI

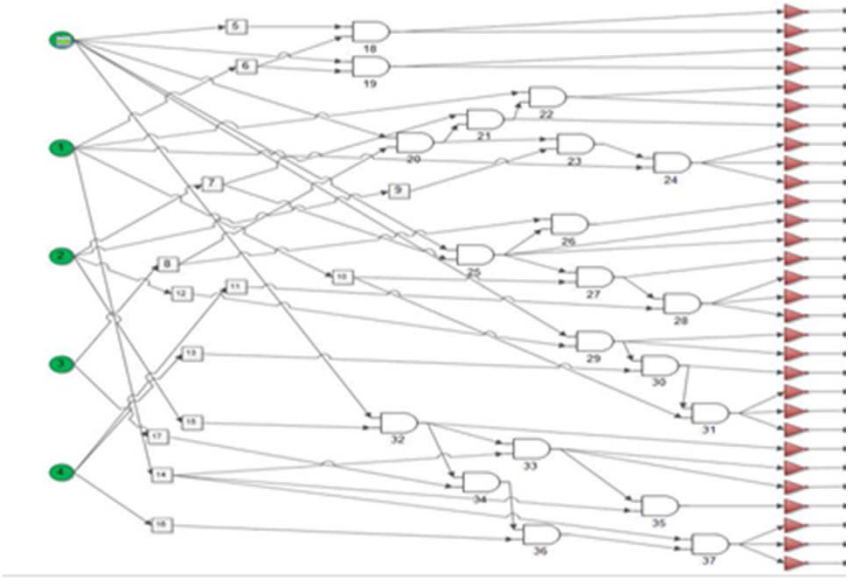
→Management of millions of Gates.

→Presence of testing tools

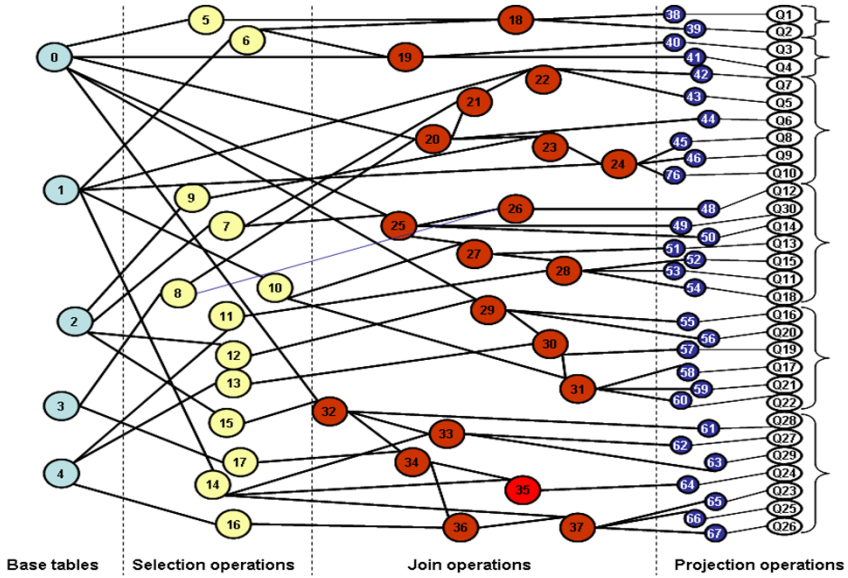
Mentors Graphics, Grenoble

Analogy between MQP & VLSI

Electronic Circuit



Multi Query Plan (MQP)



Objective (VLSI):

Circuit Design is split into clusters of gates:

→ Important connection between gates inside a cluster

→ Minimal connection between clusters.

Objective (MQP):

Grouping queries in disjoint components

→ Strong interaction between queries inside each component

→ Minimal sharing of results between components.

Tools

- Forge @ LIAS: http://www.lias-lab.fr/forge/projects

- **OntoDB Platform**
  - OntoDB Schema: a ontology based database tool
  - OntoQL: an exploitation language for ontology based databases
  - OntoQLConsole: an OntoQL interactive terminal
  - OntoQLPlus: an OntoQL editor. It is a command line
- **OntoQL** language interface. It provides a syntax highlighting and a history of the executed commands.

- **OntoDBench**: estimate the structuredness of her/his dataset and evaluate the scalability of the three main storage layouts of ontology-based data on its real datasets and workload

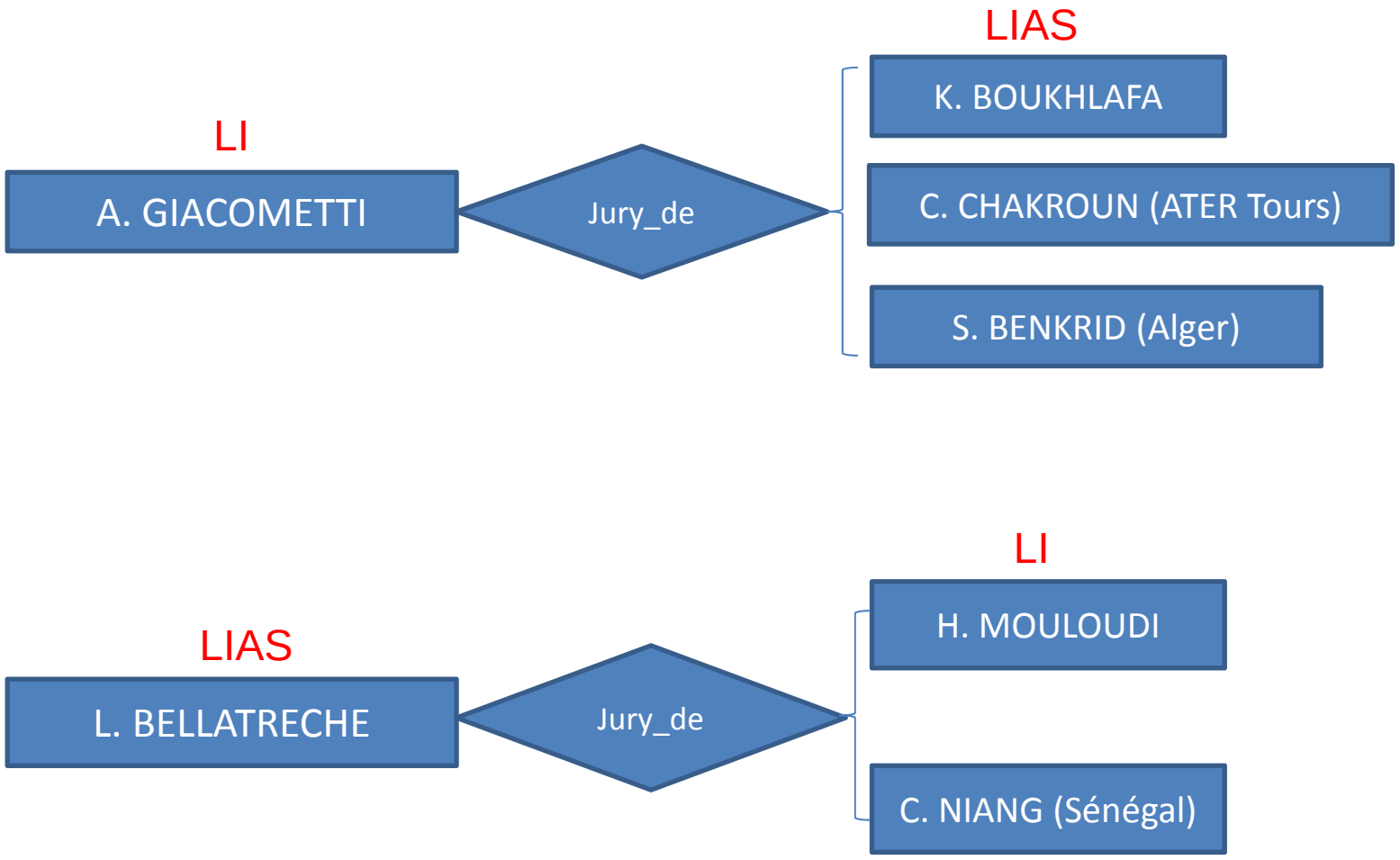
- **KMAD Platform**: tool for contributing to the incorporation of ergonomics into the interactive systems design process

LI ~LIAS : Longue Histoire

Thème	Equipes LI / LIAS	Type Financement	Cadre	Année
Personnalisation des Requêtes OLAP	BDTLN/IDD	Fonts Propres	Thèse H. MOULOUDI (LI)	03-06
Sélection des structures d'optimisation	OC/IDD-TR	CS UFR/ENSMA	PFE M. SERVAIS (PolyTech)	09-10
Besoins & Personnalisation	BDTLN/IDD	Fonts Propres	Thèse S. KHOURI (LIAS)	09-13
Big Data & Queries	BDTLN/IDD	Actions de Recherches Collaboratives		En cours

Enseignement/Encadrement : Erasmus Mundus Master in IT4BI

Jurys de Thèse



Production Scientifique Commune

1. Selma Khouri, Ladjel Bellatreche, Patrick Marcel : Une démarche de conception d'un entrepôt sémantique matérialisant les données et les besoins. Ingénierie des Systèmes d'Information, 17(5): 9-34 (2012)
2. Oscar Romero, Patrick Marcel, Alberto Abelló, Verónika Peralta, Ladjel Bellatreche: Describing Analytical Sessions Using a Multidimensional Algebra. DaWaK, pp. 224-239, LNCS, Springer, 2011.
3. Selma Khouri, Ladjel Bellatreche, Patrick Marcel : Embedding User's Requirements in Data Warehouse Repositories. OTM Workshops, pp. 35-36, LNCS, Springer, 2011.
4. Ladjel Bellatreche, Arnaud Giacometti, Patrick Marcel (Eds.) : Actes des 3èmes journées francophones sur les Entrepôts de Données et l'Analyse en ligne, EDA 2007, Poitiers, France, Juin 2007. RNTI B-3, Cépaduès 2007,
5. Hassina Mouloudi, Ladjel Bellatreche, Arnaud Giacometti, Patrick Marcel: Personalization of MDX Queries. BDA, Lille, 2006
6. *Ladjel Bellatreche, Arnaud Giacometti, Patrick Marcel, Hassina Mouloudi, Dominique Laurent: A personalization framework for OLAP queries. ACM DOLAP, pp. 9-18, 2005 (76 citations Google Scholar)*
7. Ladjel Bellatreche, Arnaud Giacometti, Dominique Laurent: A Framework for Combining Rule-Based and Cost-Based Approaches to Optimize OLAP Queries. EDA 2005, pp. 177-196

Collaborons….