Laboratoire d'Informatique et d'Automatique pour les Systèmes

Approche guidée par la Valeur et la Variété pour concevoir des entrepôts de données étendus

Ladjel BELLATRECHE

[1]LIAS/ISAE-ENSMA
Futuroscope, Poitiers, France
(bellatreche, selma.khouri)@ensma.fr

Selma KHOURI[1,2], Nabila BERKANI[2]

[2]Ecole nationale Supérieure d'Informatique
Alger, Algérie
(s_khouri, n_berkani)@esi.dz

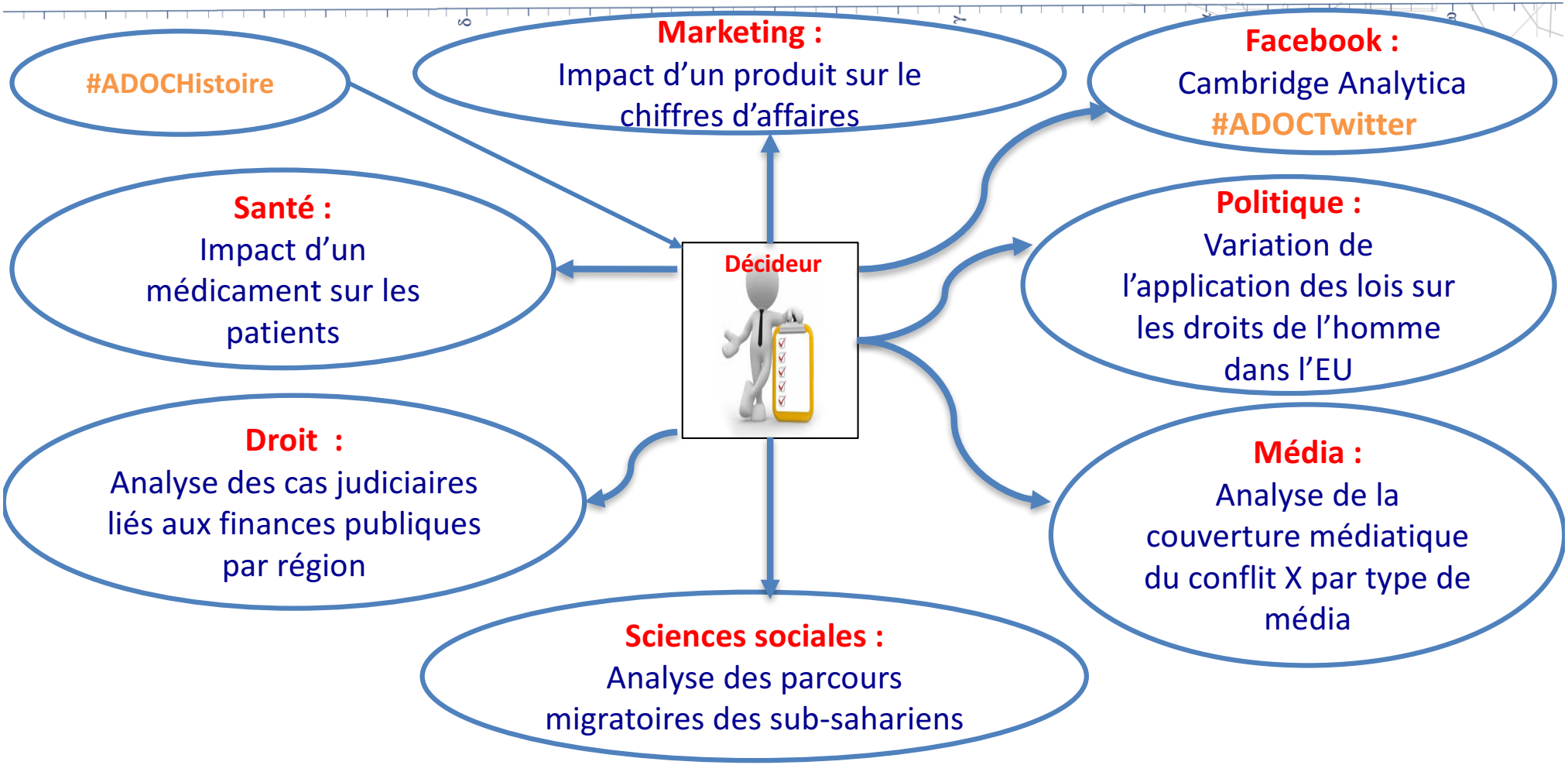
ADOC - MaDICS : Variété des Données SHS, Nantes, 28 Mai, 2018

ENSi Poitiers – Ingénierie pour la Protection de l'Environnement

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

Université de Poitiers – 1431

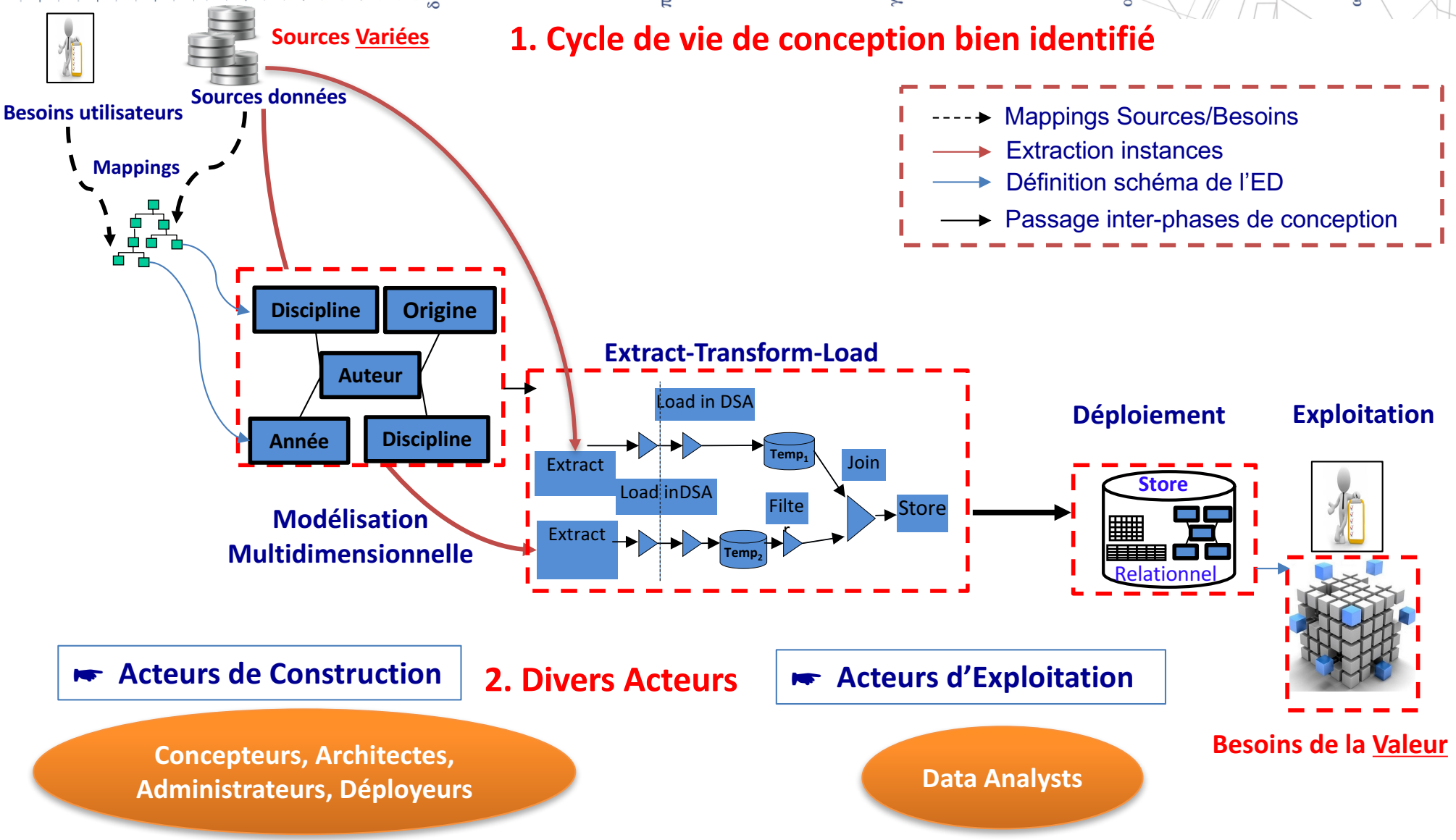
Données : mine d'or à analyser



☛ **Entrepôt de Données (ED) :** collection de données hétérogènes (**variées**), intégrées, historisée, non volatiles, organisées **analyser** les données → création de la **valeur ajoutée** (implicite)

Contexte & Verrous Scientifiques

30 ans d'existence : maturité



Agenda

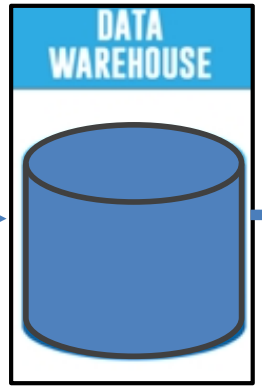
- Evolution de la Valeur et Variété (2V)
- Linked Open Data & 2V
- Conception d'un ED Dirigée par 2V
- Expérimentations
- Conclusion

Compétitivité : Valeur

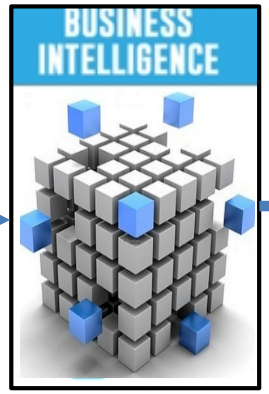
Besoins utilisateurs

Sources données variées

Intégration

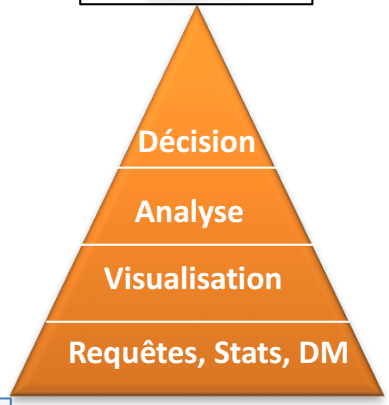


Déploiement



Exploitation

Utilisateur



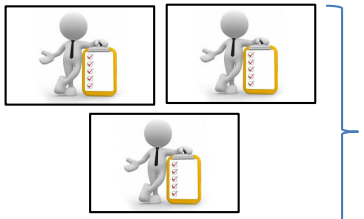
Valeur Ajoutée (implicite)

Tout le travail de conception et d'administration au service de la valeur

Contrepartie (Valeur) : Variété

Variété des besoins

(Entreprises étendues : Airbus)



Projet multidisciplinaire : ADOC …

Variété des sources

❑ **Variété Traditionnelle** (issue des **Systèmes d'Intégration**)

→ Hétérogénéité **syntaxique** et **schématique**

S1

- 2 classes
- 7 propriétés

Livre
- SBN
- type (domaine)
- prix *(HT)*
- TVA
- éditeur

Livre numérique
- Isbn numérique
- stockage (*giga-octets*)

S2

- 1 classe
- 6 propriétés

E-Book
- ISBN
- type (roman, essai, ..)
- price *(TTC)*
- préface
- éditeur
- stockage *(méga-octets)*

❑ **Conflits syntaxiques**

Synonyme

Homonyme

❑ **Conflits sémantiques**

Conflit de contexte

Conflit unité mesure ($, £, €, …)

Causes de la Variété

1. **Les sources de données ne sont pas conçues pour être intégrées dans le futur;**
2. **Absence de sémantique.**

❑ **Exercice :**

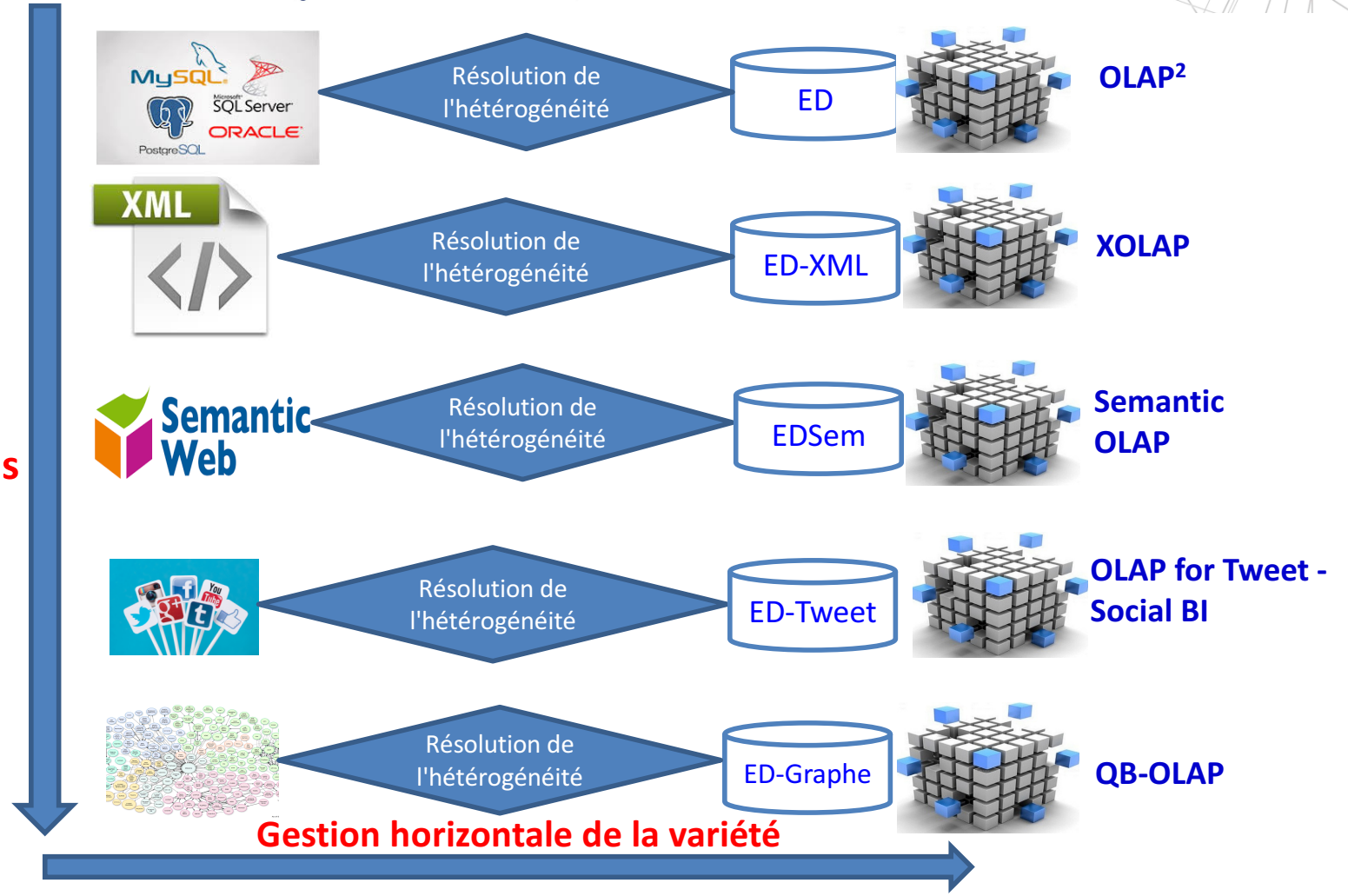
- Proposer un modèle conceptuel (Entité Association) pour ce cahier de charge (**Entités, Propriétés, Relations, Contraintes**)

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.

▶ **Résultat : 20 (# participants) de schémas conceptuels**

▶ **Schéma conceptuel transformé à un modèle physique (hétérogénéité propagée)**

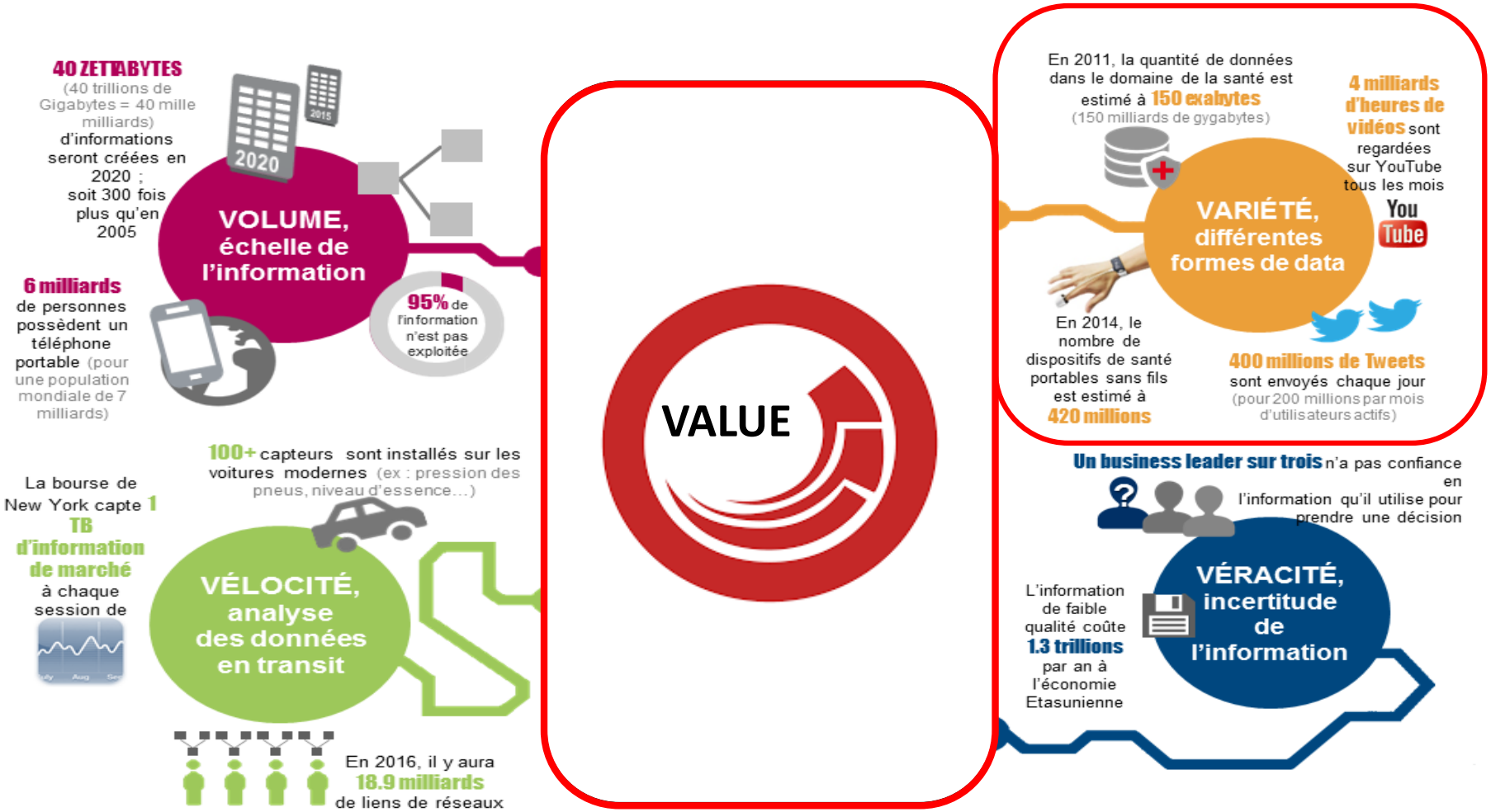
Forte Variété mais traitement isolé



- Résolution de l'hétérogénéité (**concepteur**) → appel à la **multidisciplinarité** (e.g., **experts**, ou **ontologies** : médicales, environnementales, **dettes italiennes**, …)

Big Data : mise en avant de 2V

Traitement global de la Variété



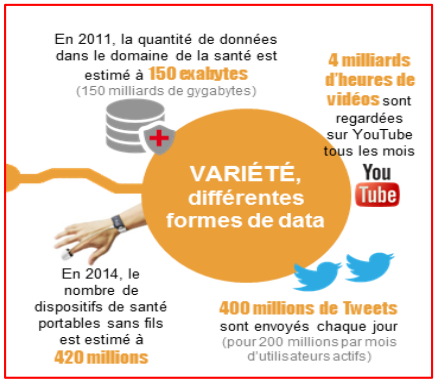
Interaction Valeur & Variété

''Valeuriste'' (Exploitation) [DaWaK'2016]

- **Affichage des besoins liés à la Valeur :**
  - Exemples[1] :
  - ***Media: Has the coverage of media changed over time?***
  - ***Politics: Speeches EU parliament that contain ''human rights'' per country***
  - ***Finance: Evolution of debates related to Greece crisis by country***
- **Difficile à la capter (en fonction du domaine)**
- **Interaction entre Concepteurs et ''Valeuristes''**

  : multidisciplinarité

Concepteurs



BD

Library

Journaux

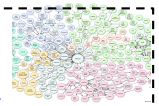
Sources Internes + Sources Externes

→ **Recours aux Linked Open Data :**

Gestion Traditionnelle de la Variété

+

Variété des Formalismes (graphes)



[1] http://www.talkofeurope.eu/data

Contributions & Résultats : Variété dans les Données SHS

Valeur dans LOD

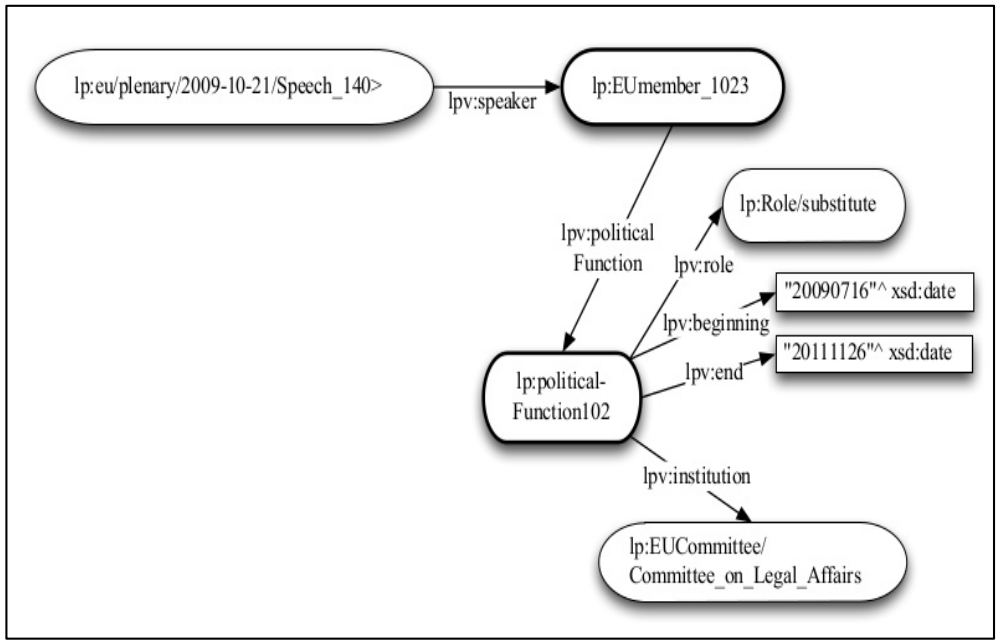
▶ *Un Exemple d'un Besoin porteur de Valeur (tiré du Portail LOD : http://www.talkofeurope.eu/data/)*

- *For a particular agenda item (the fourth item of 16 December 2010), find the frequency distribution of the speaking slots over the EU parties of the speakers involved*

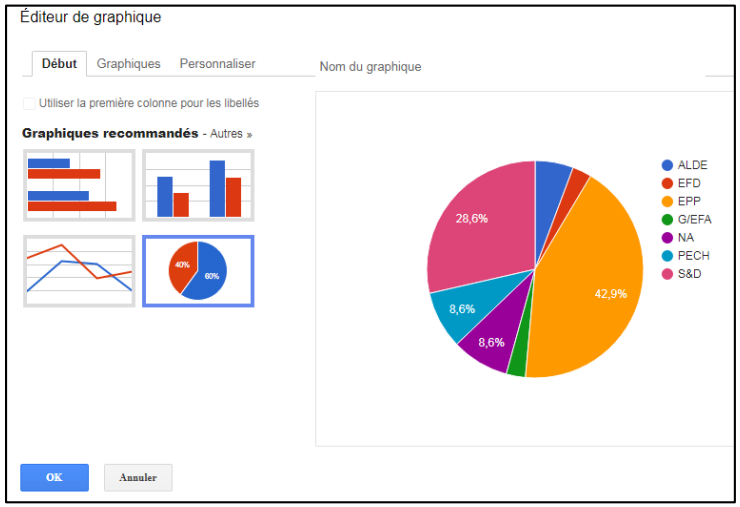
❑ Sparql Endpoint

- http://linkedpolitics.ops.few.vu.nl/yasgui/index.html

```
SELECT ?partyname (COUNT(DISTINCT ?speech) AS ?speechno)
WHERE
{ <http://purl.org/linkedpolitics/eu/plenary/2010-12-
16_AgendaItem_4> dcterms:hasPart ?speech.
 ?speech lpv:spokenAs ?function.
?function lpv:institution ?party.
?party rdf:type lpv:EUParty.
?party lpv:acronym ?partyname. }
GROUP BY ?partyname
```



❑ Visualisation du Résultat



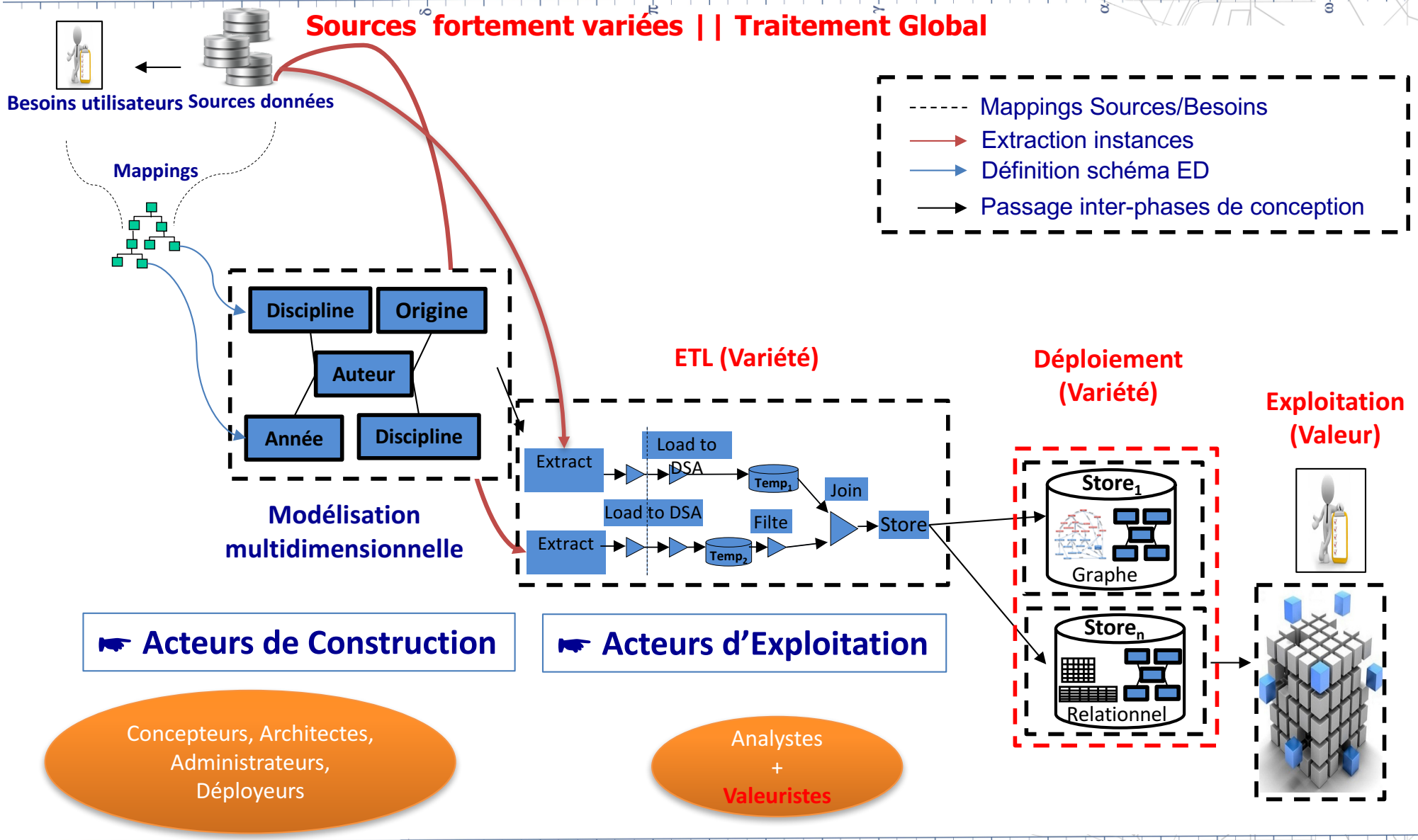
Triplet: <Speech_140, speaker, Eumember_1023>
→ Graphe (et ontologies)

- Autres Exemples de LOD :
  - PoliMedia: linking media archives to political event data: http://polimedia.nl/
  - Plenary debates of the European Parliament: http://www.talkofeurope.eu/data/

ED dans l'ère de la variété et valeur

- Variété & Valeur
- LOD & 2V
- Notre Approche
- Expérimentations
- Conclusion

Sources fortement variées | | Traitement Global



Formalisation sensible aux 2 V

- **Entrées :**

1. Un ensemble de sources internes : $S_{Internes} = \{S_{I1}, S_{I2}, \ldots, S_{Im}\}$
2. Un ensemble de sources externes LOD : $S_{Externes} = = \{S_{E1}, S_{E2}, \ldots, S_{En}\}$
3. Chaque source interne et externe possède:
   - Son propre format Format ($S_{Interne}$)
   - Son modèle conceptuel CM_I décrivant sa syntaxe/sémantique
   - Est liée à une discipline D
4. Ensemble de besoins B à satisfaire
5. Un ED (**à définir [DaWaK'2016] ou opérationnel [Ravat et al. 2017]**) possédant:
   - Un ou plusieurs formats Format (S_{ED}) = $\{f_1, f_2, \ldots, f_k\}$
   - Un modèle conceptuel CM_{ED} décrivant sa syntaxe/sémantique

- **Objectif :**
  - Augmenter la valeur pour une discipline donnée D

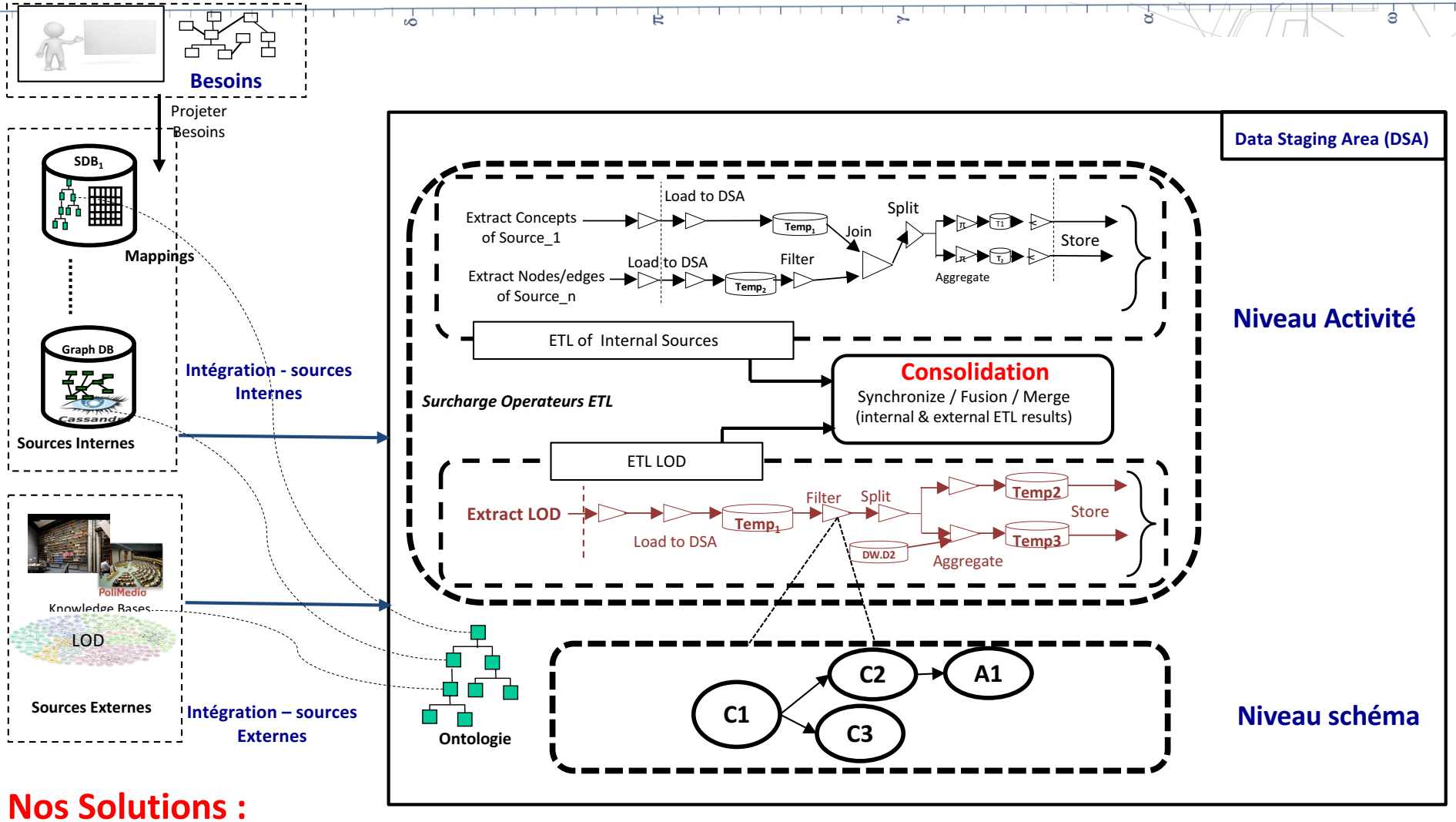
- **Enjeux :**
  - Mesures de la valeur

    Valeur = Poids(S_i , D) * Valeur (S_i); $S_i \in S_{internes} \cup S_{externes}$ [Ballou et al.]*

*Ballou, D. P., & Tayi, G. K. (1999). Enhancing data quality in data warehouse environments. *Com of the ACM*, *42*(1), 73-78

ETL - Variété

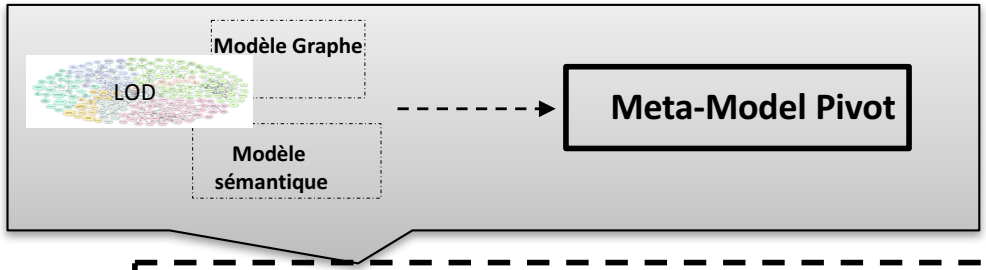
- Variété & Valeur
- LOD & 2V
- Notre Approche
- Expérimentations
- Conclusion

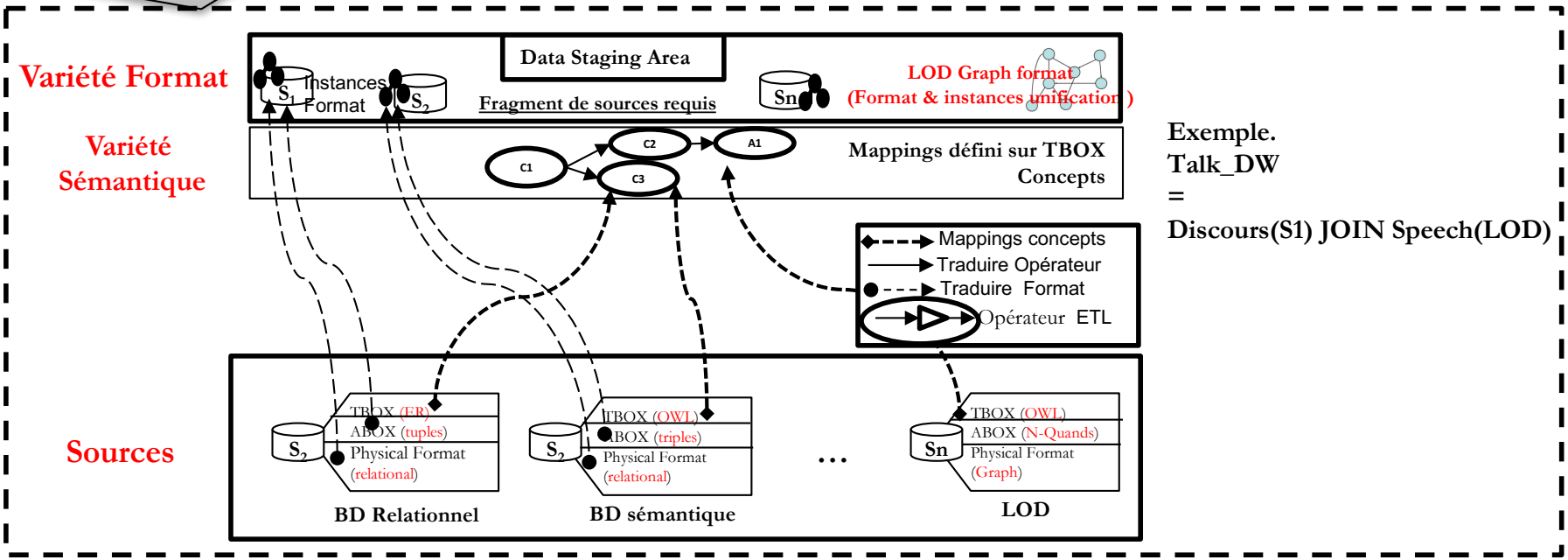


Nos Solutions :

1. **Schéma Pivot** : sources internes et externes [DEXA'2017]
2. **Synchronisation** des données externes et internes [ER'2018]
3. **Redéfinition** des opérateurs traditionnels (overloading) : [DEXA'2017}

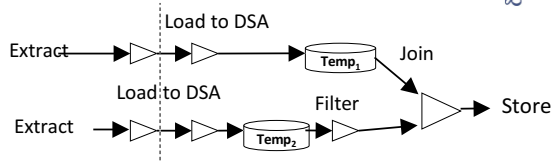
ETL niveau schéma :
Choix du LOD - Pivot



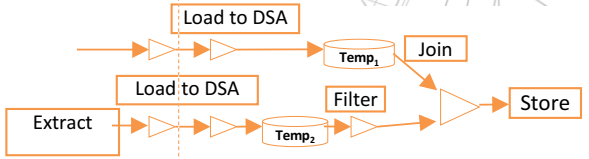


ETL : activité

- Variété & Valeur
- LOD & 2V
- Notre Approche
- Expérimentations
- Conclusion

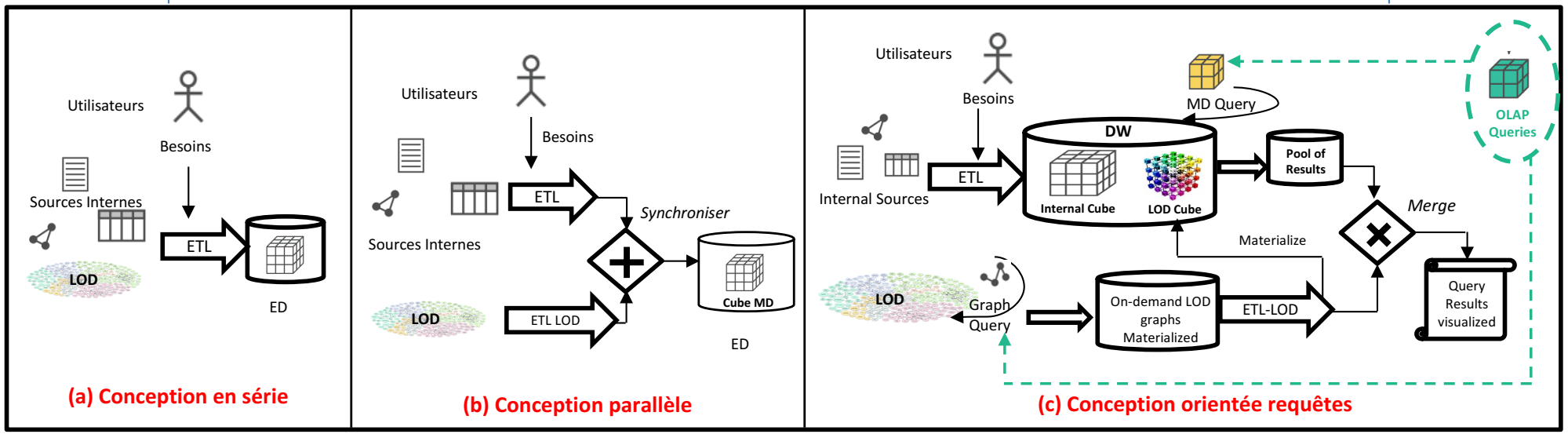


ETL Interne

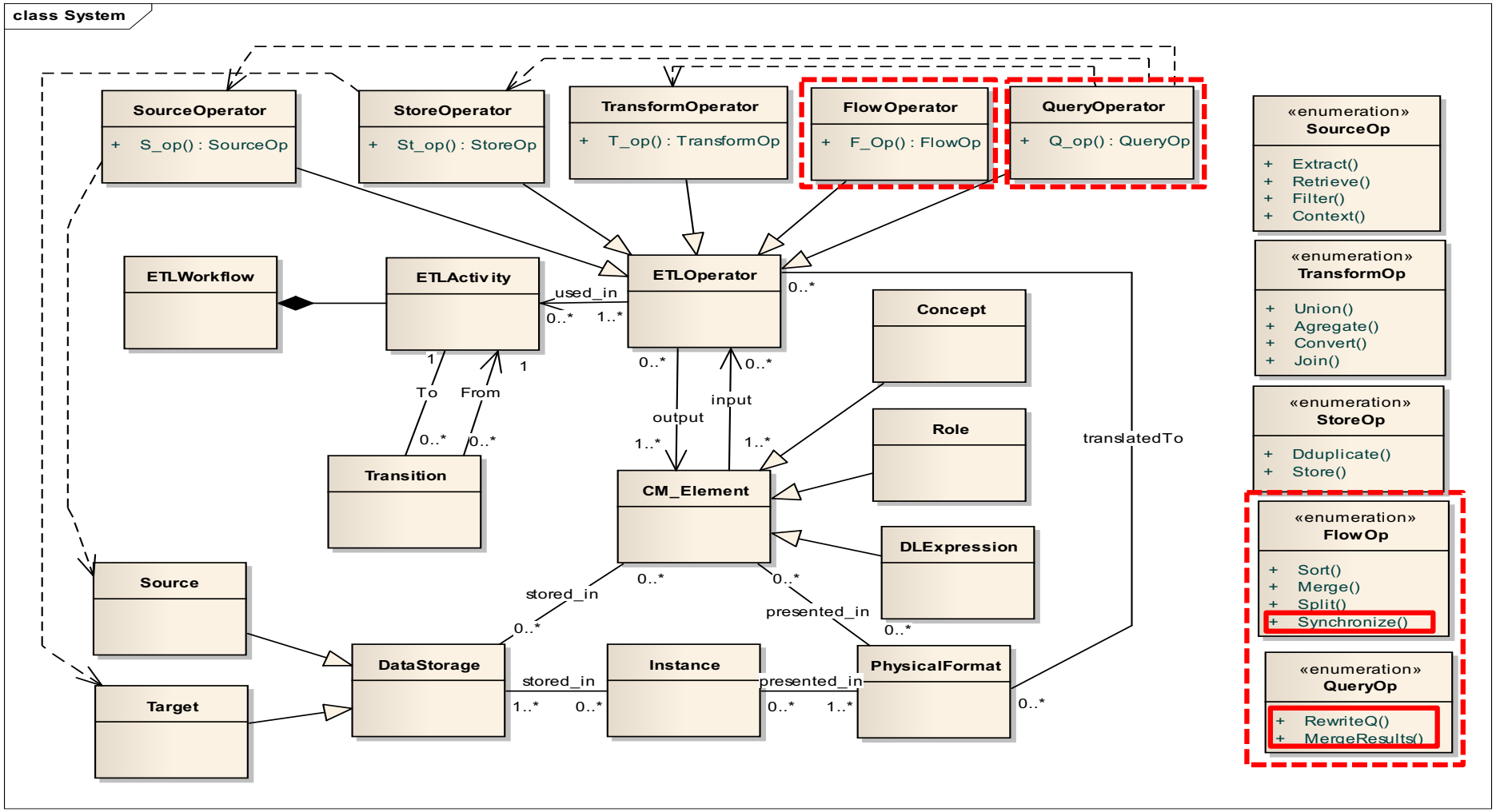


ETL Externe

Trois Scénarii



Modèle Pivot : ETL



Capturer la valeur

Valeur = Poids(S_i, D) * Valeur (S_i) tel que : $S_i \in S_{internes} \cup S_{Externes}$

❑ Métriques d'évaluation de la valeur:

- **Taux de satisfaction des besoins par une Source S_i**
- $$Valeur(Si, MD) = \frac{Nombre\ de\ Réponses\ de\ Besoins(Si)}{Nombre\ de\ réponses\ de\ l'ensemble\ des\ Besoins}$$
- **Multidimensionnel (MD) augmente le pouvoir décisionnel:**
  - **Taux de concepts MD**

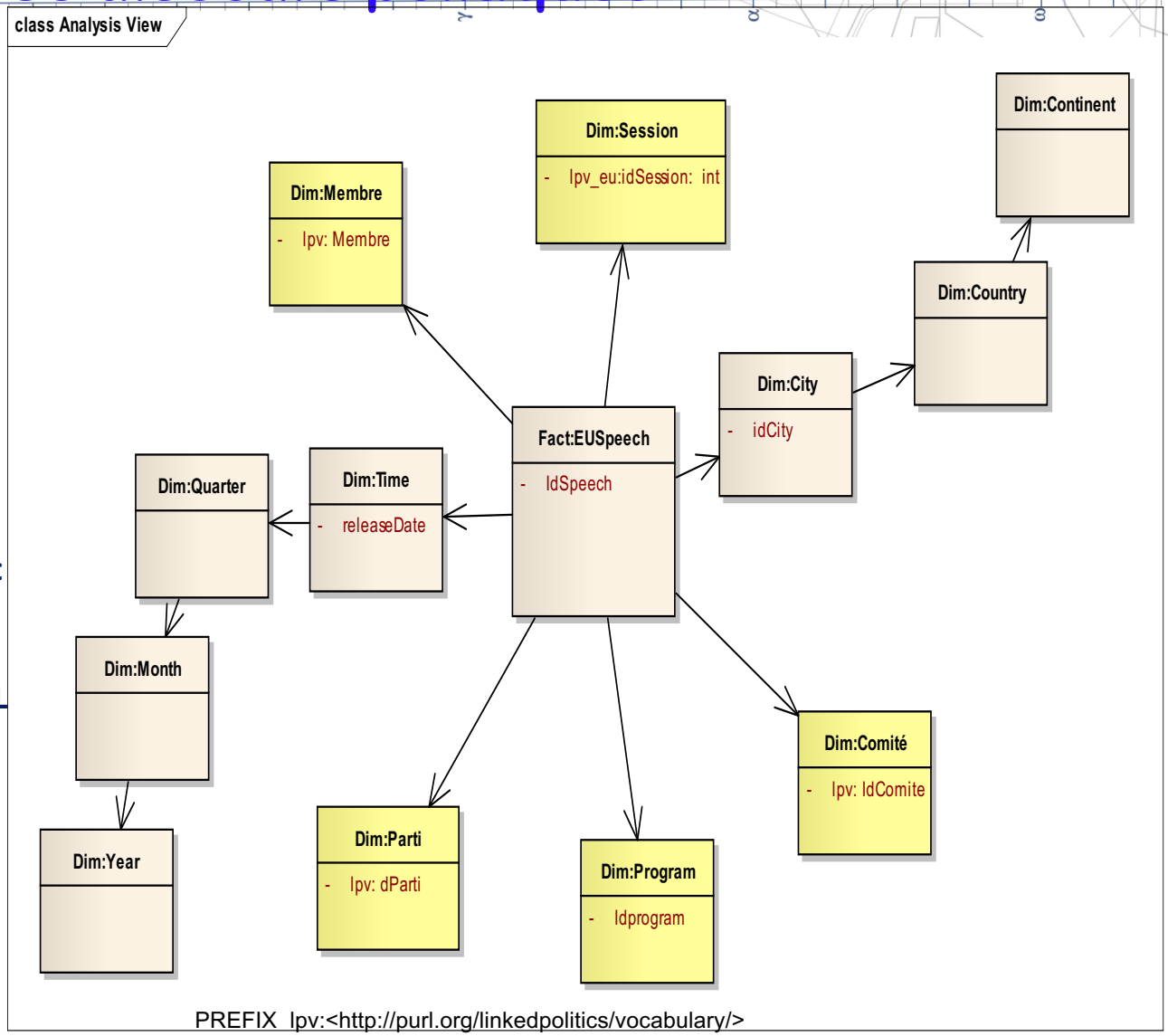
$$Valeur(Si, MD) = \frac{Nombre\ de\ Concepts\ MD\ (Si)}{Nombre\ Total\ de\ Concepts\ de\ l'ED}$$

  - **Taux d'instances intégrées**

$$Valeur(Si, I) = \frac{Nombre\ d'instances\ (Si)}{Nombre\ Total\ d'instances\ de\ l'ED}$$

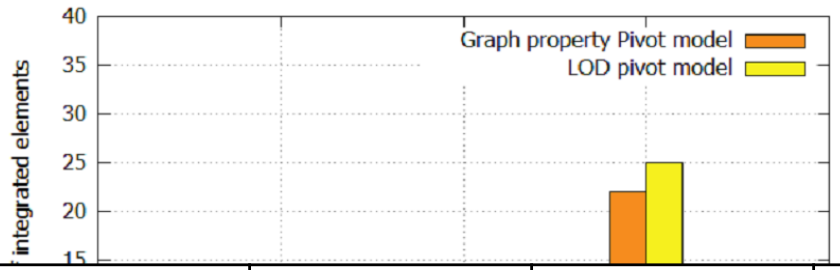
Étude de cas:
Analyse discours politiques

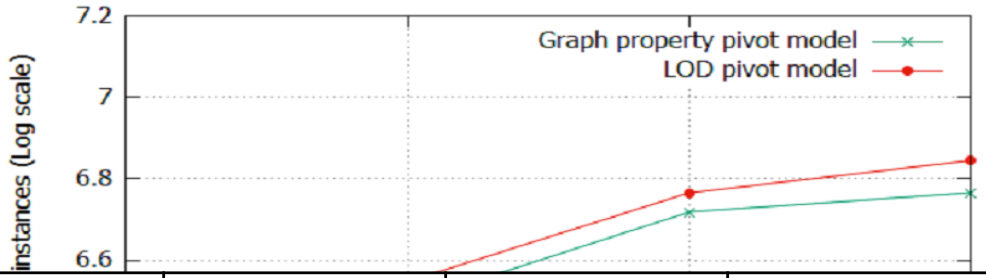
- SGBD Oracle 12c release 1 (RDF Semantic);
- Discours du parlement de l'union européenne (http://www.talkofeurope.eu/data/)
- Oracle pour le déploiement (déploiement Semantic Vs déploiement Gaphe N-Quads)
- Oracle inference engine basé sur TrOWL et Pellet



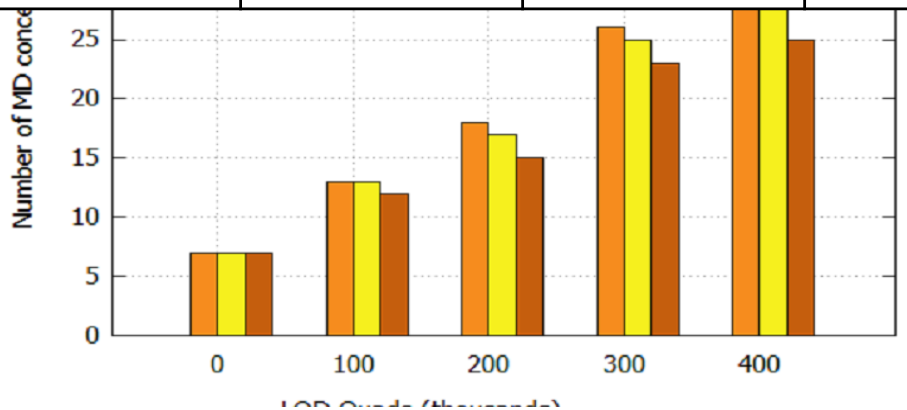
PREFIX lpv:<http://purl.org/linkedpolitics/vocabulary/>
PREFIX lpv_eu:<http://purl.org/linkedpolitics/eu/plenary/>

Expérimentations

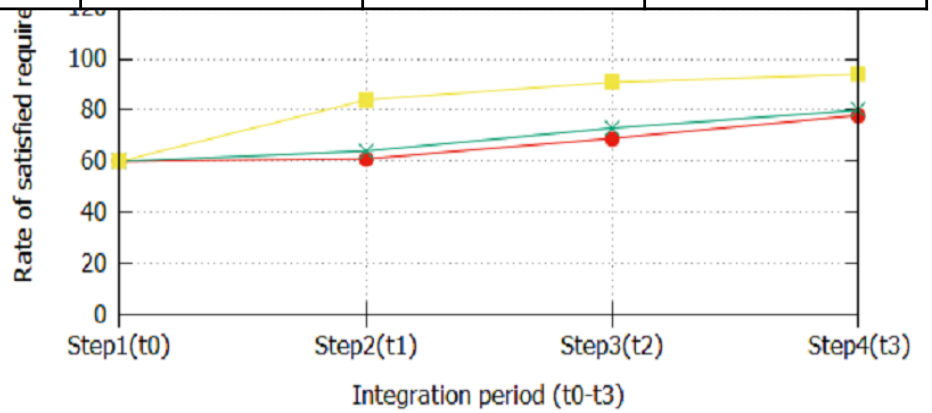




Métriques Sources	Dimensions/ Mesures	Valeur $(Si)_{MD}$	Valeur $(Si)_{Besoins}$	Valeu $(Si)_{Instances}$	Instances	Temps de réponse
Sources internes	6/1	~ 31%	6%	~10%	550K	1.1
Conception en série	10/7	71%	80%	94%	$7,7x10^6$	3.2
Conception parallèle	11/8	73%	84%	85%	$3,1x10^6$	2.6
Conception à la demande	12/8	74%	96%	84%	$2,9x10^6$	1.7



Nombre de concepts multidimensionnels intégrés



Taux de besoins satisfaits

Conclusions & Actions Futures

Conclusion

- ❑ Variété & Valeur pour la renaissance des ED
- ❑ Variété – Valeur = Vivier pour la multidisciplinarité
- ❑ Revisite du Cycle de vie de Conception des ED
- ❑ Formalisation
- ❑ ED pour les Données SHS
- ❑ Expérimentations

- ❑ Formalisation de la valeur
- ❑ Outillage
- ❑ Passage à l'action : projet multidisciplinaire

Productions

1. S. Khouri, L. Bellatreche, Consolidation of BI Efforts in the LOD Era for African Context, in ICSE Workshops, Gothenburg, Sweden, May 2018.
2. I. Boukhari, S. Jean, I. Ait-Sadoune, L. Bellatreche, The Role of User Requirements in Data Repository Design, Software Tools for Technology Transfer Journal (STTT), Springer, 2017
3. N. Berkani, L. Bellatreche: A Variety-Sensitive ETL Processes. DEXA (2), 201-216, 2017
4. Y. Aït Ameur, M. Baron, L. Bellatreche, S. Jean, E. Sardet: Ontologies in engineering: the OntoDB/OntoQL platform. Soft Computing, Springer, 21(2): 369-389, 2017
5. O. Barkat, S. Khouri, L. Bellatreche, B. Narhimene: Bridging context and data warehouses through ontologies. ACM SAC, 336-341, 2017
6. S. Khouri, N. Berkani, L. Bellatreche: Tracing data warehouse design lifecycle semantically. Computer Standards & Interfaces Journal, Elsevier 51: 132-151, 2017
7. Z. Djilani, N. Berkani, L. Bellatreche: Towards Functional Requirements Analytics. ISoLA (1) , 358-373, 2016
8. N. Berkani, L. Bellatreche, B. Benatallah: A Value-Added Approach to Design BI Applications. DaWaK, 361-375, 2016
9. L. Bellatreche, A. Cuzzocrea, I-Y. Song: Advances in data warehousing and OLAP in the Big Data Era. Information Systems Journal, Elsevier, 53: 39-40, 2015

www.lias-lab.fr/members/bellatreche