BDBIO - Intégration de données

Fabien Duchateau

fabien.duchateau [at] univ-lyon1.fr

Université Claude Bernard Lyon 1

2023 - 2024

https://perso.liris.cnrs.fr/fabien.duchateau/BDBIO/

Rappels

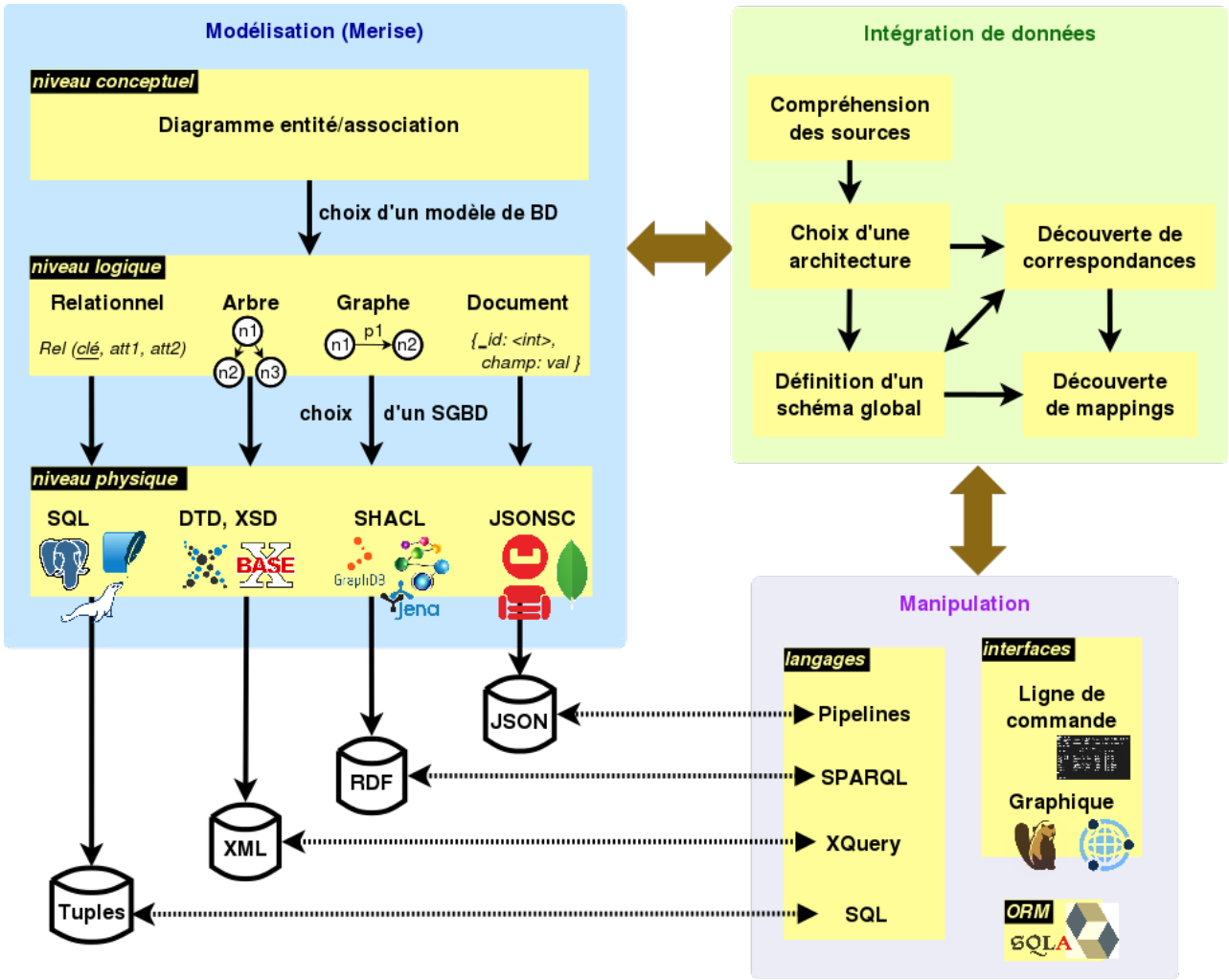
Modèle	Sérialisation	LDD	LMD	Exemples SGBD
Relationnel	n-uplets	SQL	SQL	SQLite, MariaDB, PostgreSQL
Arbre	XML	DTD, XML Schema	XPath, XQuery	XBase, existDB
Document	JSON	?	?	MongoDB, CouchDB
Graphe (triplets)	RDF/XML, N3, Turtle, N-triples	RDFS OWL	SPARQL, RQL	AllegroGraph, BlazeGraph

Vu la quantité exponentielle de données disponibles, comment interroger et agréger des informations issues de plusieurs sources de données aux modèles différents ?

---

http://en.wikipedia.org/wiki/List_of_biological_databases

Aperçu des notions de BDBIO



Ces diapositives utilisent le genre féminin (e.g., chercheuse, développeuses) plutôt que l'écriture inclusive (moins accessible, moins concise, et pas totalement inclusive)

Intégration de données

Fournir un accès uniforme à des sources de données hétérogènes et stockées sur des sites autonomes

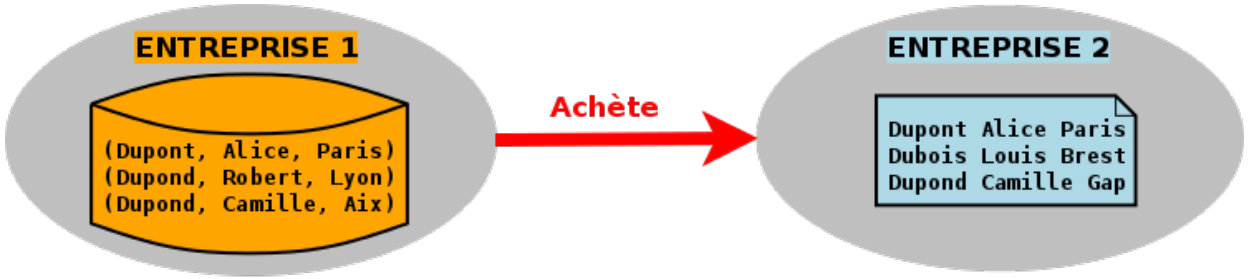
- **Accès** = requêtes d'interrogation principalement
- **Uniforme** = requête sur une seule interface
- **Sources de données** = au moins deux sources distribuées
- **Hétérogénéité des sources** = à différents niveaux
- **Autonomie** = indépendance des sites (pas de modification sur les sources de données)

---

http://en.wikipedia.org/wiki/Data_integration

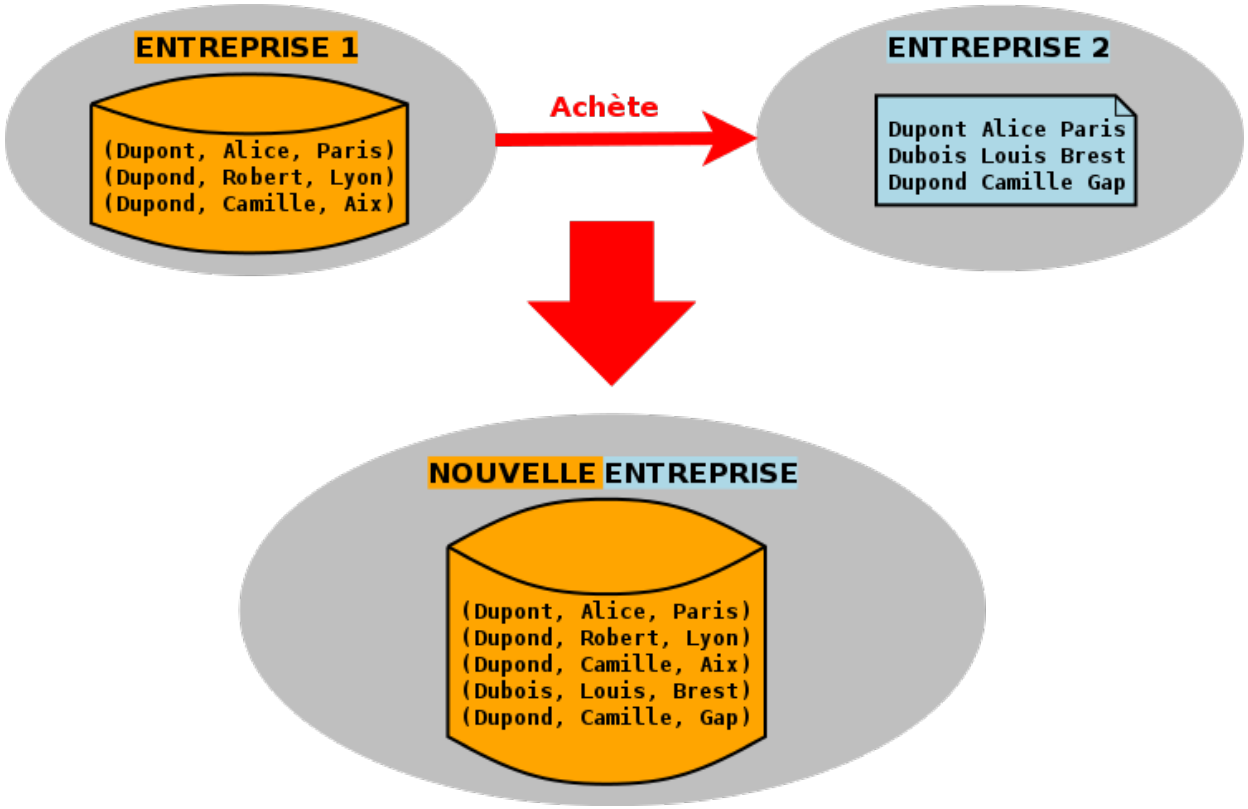
Exemple 1 : nettoyage de données

Rachat d'entreprise : besoin de détecter les entités en doublon (e.g., clientes, produits), puis de construire la BD de la nouvelle entreprise



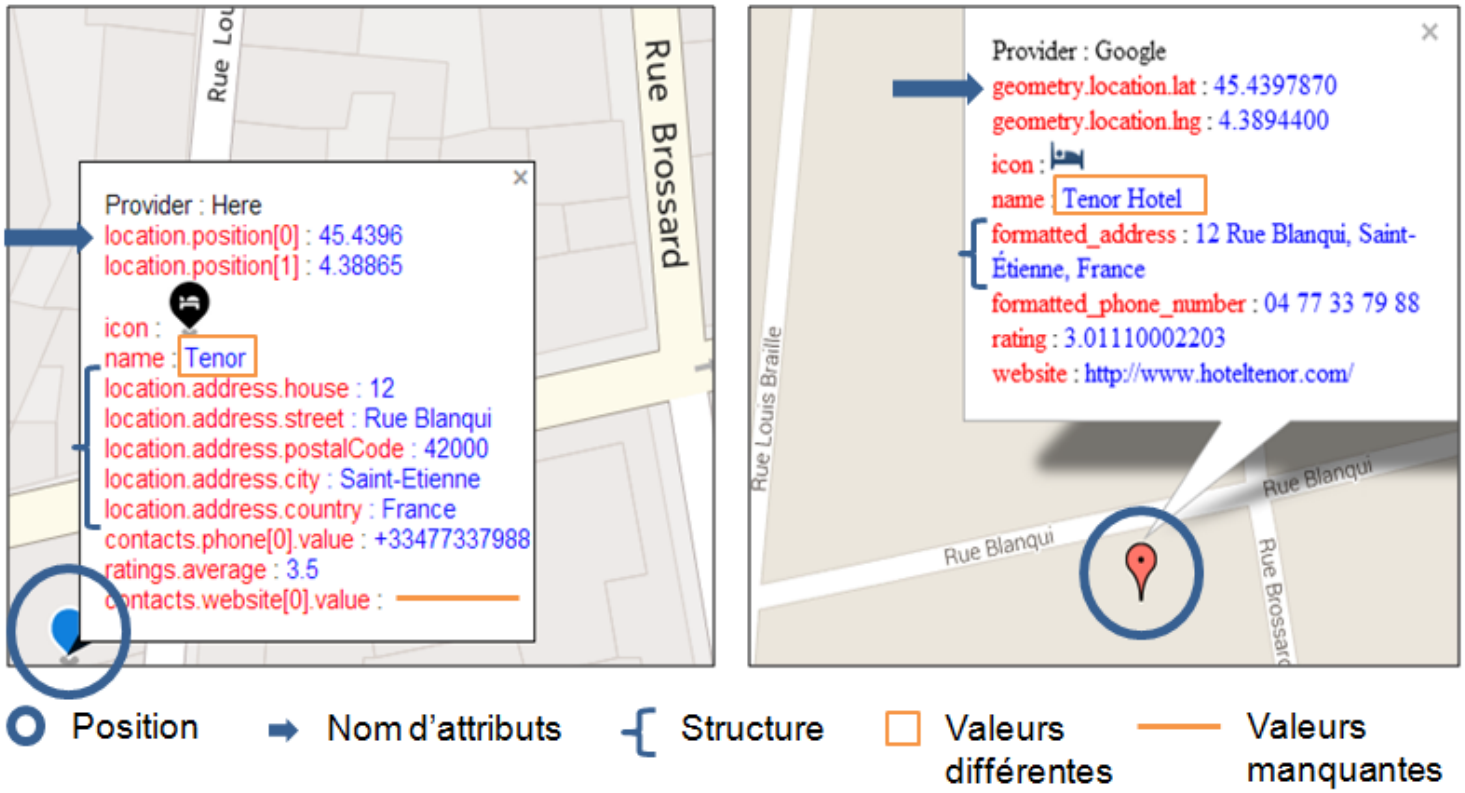
Exemple 1 : nettoyage de données

Rachat d'entreprise : besoin de détecter les entités en doublon (e.g., clientes, produits), puis de construire la BD de la nouvelle entreprise



Exemple 2 : fusion d'entités

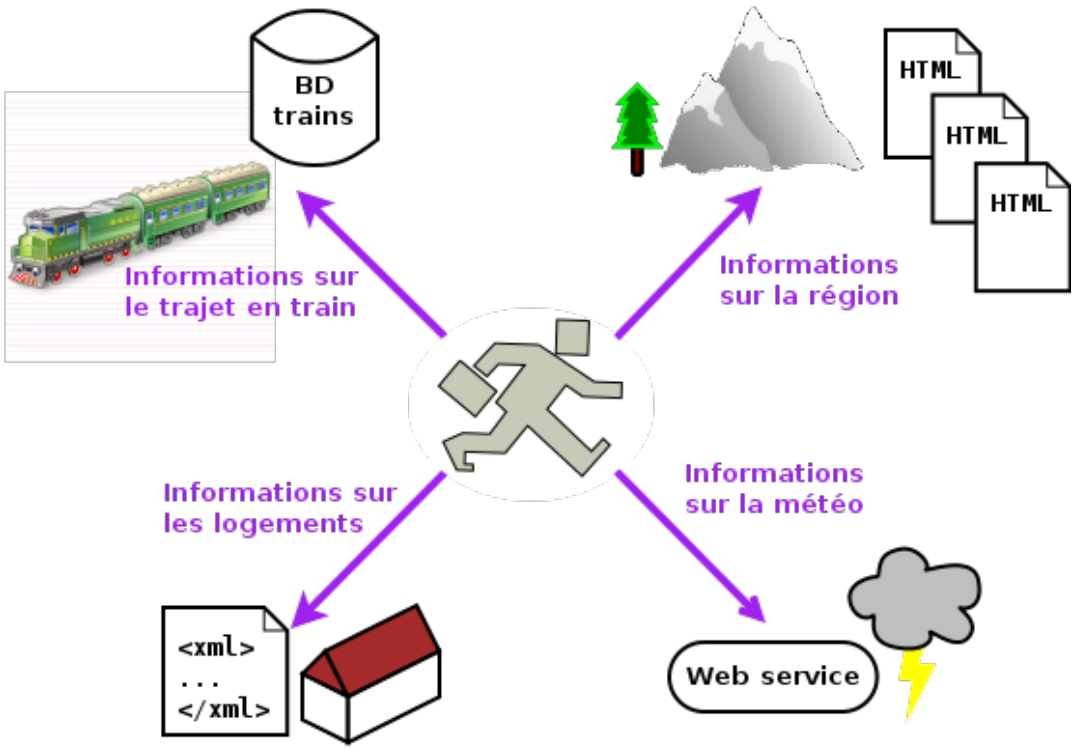
Recherche d'informations cartographiques exactes et complètes : besoin de fusionner les informations de plusieurs entités



GeoBench, http://geobench.liris.cnrs.fr/

Exemple 3 : interrogation distribuée

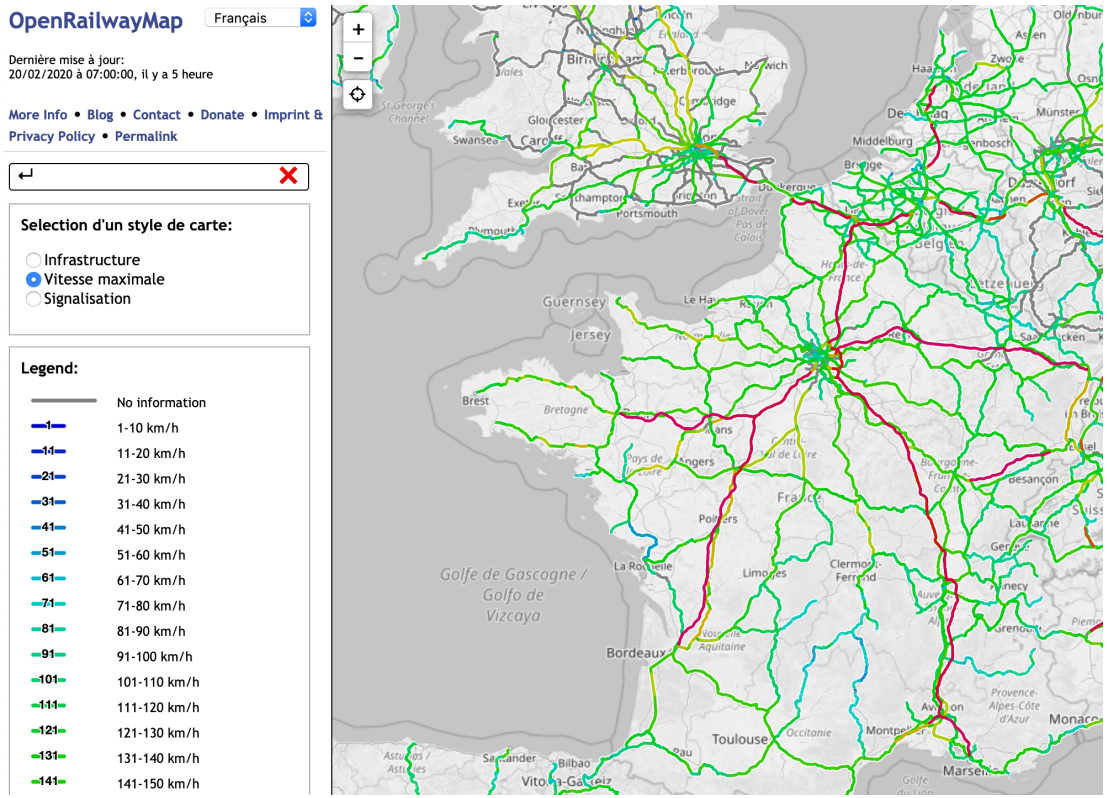
Départ en vacances : à partir d'une requête, besoin d'informations sur la météo, les horaires de train, les logements disponibles, etc.



http://en.wikipedia.org/wiki/Aggregator

Exemple 4 : mashup

OpenRailwayMap : informations sur les trains (infrastructure, vitesse, etc.)



https://www.openrailwaymap.org/
https://wiki.openstreetmap.org/wiki/List_of_OSM-based_services

Intégration de données en bioinformatique

Objectifs : consolider, enrichir et partager des données, comparer des observations ou expérimentations, etc.

- UniProt (220m de protéines, 75M d'acides aminés), GenBank (940M de bases, 230m séquences), PubMed (32m d'articles)
- Exemples : identification des éléments fonctionnels et des circuits régulatoires chez la drosophile, analyse des changements dans les processus métaboliques des cancers

Formatage/fusion/merge/enrichissement de données/tableurs $\Rightarrow$ généralement de l'intégration de données (plus ou moins complexe)

---

http://www.ebi.ac.uk/uniprot/TrEMBLstats

http://www.ncbi.nlm.nih.gov/genbank/statistics

Gomez-Cabrero et al. *Data integration in the era of omics: current and future challenges*. BMC (2014)

Lapatas et al., *Data integration in biological research: an overview*. J. Biol. Res. (2015)

Plan

Le problème d'intégration

Architectures

Définition de correspondances

Définition d'un schéma global

Définition de mappings

Définition

Un système d'intégration de données est un triplet $< \mathcal{G}, \mathcal{S}, \mathcal{M} >$:

- $\mathcal{G}$ le schéma global (ou intégré)
- $\mathcal{S}$ un ensemble de sources (de données)
- $\mathcal{M}$ un ensemble de mappings (fonctions de transformation entre des concepts similaires). L'ensemble de correspondances $\mathcal{C}$ est inclus dans $\mathcal{M}$

À partir de $\mathcal{S}$, comment définir $\mathcal{G}$ et/ou $\mathcal{M}$ pour accéder aux données des sources en requêtant $\mathcal{G}$?

Scénarios

Selon le scénario, l'intégration concerne **uniquement les métadonnées** (schéma) ou les **métadonnées + instances** :

- Construction d'un vocabulaire commun
- Estimation du coût d'un projet intégration
- Découverte de schémas pertinents
- Interrogation distribuée (médiation)
- Déduplication (nettoyage d'une BD)
- Migration de données (transfert des données d'une source vers un modèle cible, sous-ensemble du problème d'intégration)
- Fusion de données (e.g., entrepôt, enrichissement)

Caractéristiques des sources de données

En plus de l'indépendance et l'autonomie, les sources sont hétérogènes à différents niveaux :

- Modèle de représentation des données
- Format des données
- Langage pour accéder aux données
- Schéma (ou ontologie) :
  - sémantique (signification ou interprétation des concepts, e.g., synonymie, polysémie, subsomption)
  - structurel (e.g., différences de structures, types de données, contraintes)
- Instances :
  - sémantique (signification ou interprétation des valeurs, e.g., synonymie, polysémie, subsomption)
  - syntaxique (e.g., permutations de mots, abréviations, typos, alias, multilinguisme, différences de standard)

Exemples d'hétérogénéité

À quel niveau se trouve l'hétérogénéité ?

- SPARQL / RQL
- Turtle / N-triples
- Relationnel / graphe
- Adresse / {rue, code postal, ville}
- Richard Peyzaret / F'murr
- Auteures / écrivaines
- Tolkien / J.R.R. Tolkien

Exemples d'hétérogénéité

À quel niveau se trouve l'hétérogénéité ?

- SPARQL / RQL $\Rightarrow$ langage
- Turtle / N-triples
- Relationnel / graphe
- Adresse / {rue, code postal, ville}
- Richard Peyzaret / F'murr
- Auteures / écrivaines
- Tolkien / J.R.R. Tolkien

Exemples d'hétérogénéité

À quel niveau se trouve l'hétérogénéité ?

- SPARQL / RQL $\Rightarrow$ langage
- Turtle / N-triples $\Rightarrow$ format
- Relationnel / graphe
- Adresse / {rue, code postal, ville}
- Richard Peyzaret / F'murr
- Auteures / écrivaines
- Tolkien / J.R.R. Tolkien

Exemples d'hétérogénéité

À quel niveau se trouve l'hétérogénéité ?

- SPARQL / RQL $\Rightarrow$ langage
- Turtle / N-triples $\Rightarrow$ format
- Relationnel / graphe $\Rightarrow$ modèle
- Adresse / {rue, code postal, ville}
- Richard Peyzaret / F'murr
- Auteures / écrivaines
- Tolkien / J.R.R. Tolkien

Exemples d'hétérogénéité

À quel niveau se trouve l'hétérogénéité ?

- SPARQL / RQL $\Rightarrow$ langage
- Turtle / N-triples $\Rightarrow$ format
- Relationnel / graphe $\Rightarrow$ modèle
- Adresse / {rue, code postal, ville} $\Rightarrow$ schéma - structurel
- Richard Peyzaret / F'murr
- Auteures / écrivaines
- Tolkien / J.R.R. Tolkien

Exemples d'hétérogénéité

À quel niveau se trouve l'hétérogénéité ?

- SPARQL / RQL $\Rightarrow$ langage
- Turtle / N-triples $\Rightarrow$ format
- Relationnel / graphe $\Rightarrow$ modèle
- Adresse / {rue, code postal, ville} $\Rightarrow$ schéma - structurel
- Richard Peyzaret / F'murr $\Rightarrow$ instances - sémantique
- Auteures / écrivaines
- Tolkien / J.R.R. Tolkien

Exemples d'hétérogénéité

À quel niveau se trouve l'hétérogénéité ?

- SPARQL / RQL $\Rightarrow$ langage
- Turtle / N-triples $\Rightarrow$ format
- Relationnel / graphe $\Rightarrow$ modèle
- Adresse / {rue, code postal, ville} $\Rightarrow$ schéma - structurel
- Richard Peyzaret / F'murr $\Rightarrow$ instances - sémantique
- Auteures / écrivaines $\Rightarrow$ schéma (ou instances) - sémantique
- Tolkien / J.R.R. Tolkien

Exemples d'hétérogénéité

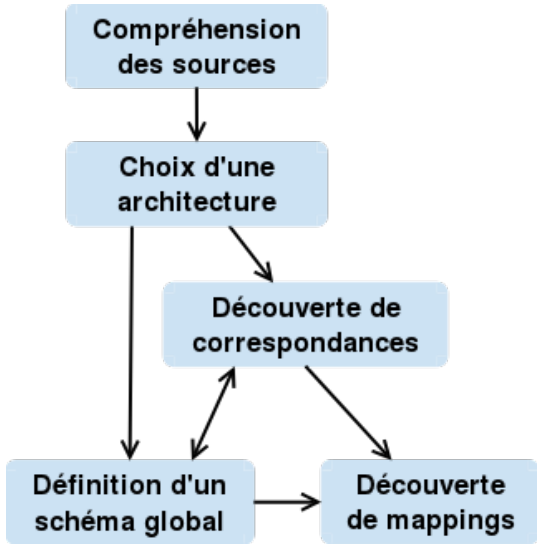
À quel niveau se trouve l'hétérogénéité ?

- SPARQL / RQL ⇒ langage
- Turtle / N-triples ⇒ format
- Relationnel / graphe ⇒ modèle
- Adresse / {rue, code postal, ville} ⇒ schéma - structurel
- Richard Peyzaret / F'murr ⇒ instances - sémantique
- Auteures / écrivaines ⇒ schéma (ou instances) - sémantique
- Tolkien / J.R.R. Tolkien ⇒ instances - syntaxique

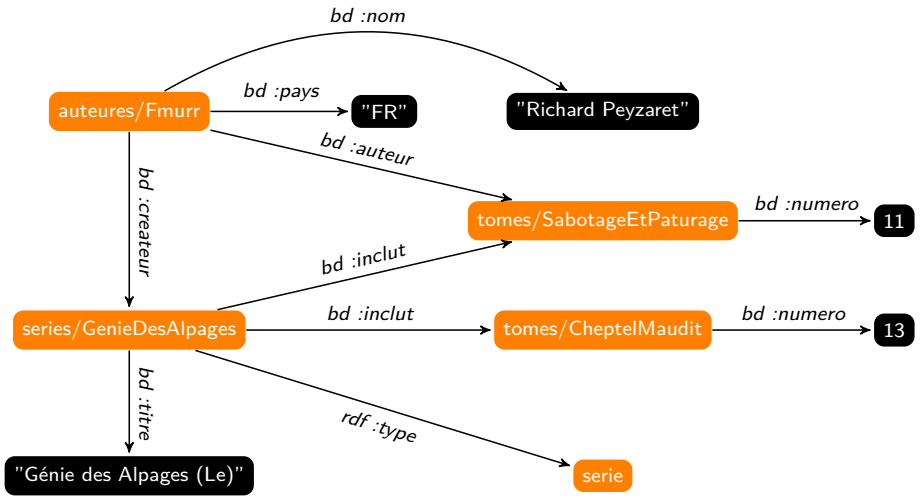
Étapes de l'intégration

Intégration généralement manuelle ou semi-automatique :

- Compréhension de la connaissance dans les sources
- Choix d'une architecture d'intégration
- Définition de correspondances (soit entre sources, soit entre une source et le schéma global) :
  - au niveau du modèle (schéma, ontologie)
  - au niveau des données (n-uplets, entités)
- Définition d'un schéma global
- Définition de mappings



Un cas concret



id	cat	coll	titre	idA	annee	num
1	BD	Le génie des alpages	Les intondables	1	1980	03/14
2	roman	Le Disque-monde	Timbré	2	2004	
3	BD	Le génie des alpages	Cheptel maudit	1	2004	13/14

Table Livres

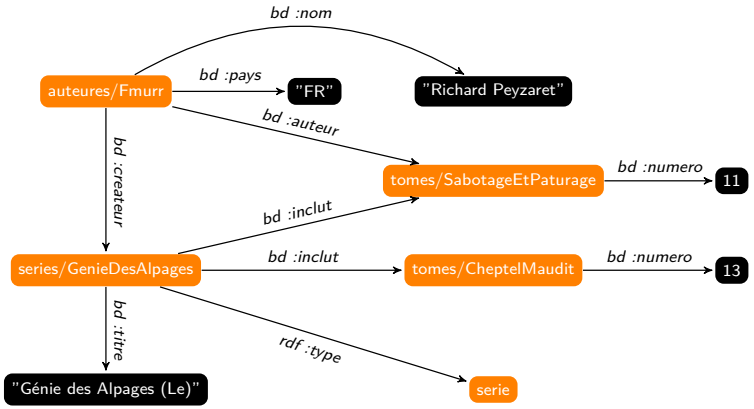
idA	auteur	nationalite	anneeNaissance
1	Fmurr	France	1946
2	Terry Pratchett	Royaume-Uni	1948

Table Auteures

Une source S1 (graphe RDF) et une source S2 (tables)

- Correspondances au niveau modèle (e.g., $serie \subseteq collection$)
- Correspondances au niveau données (e.g., $Richard\ Peyzaret \equiv Fmurr$)
- Schéma global avec tous les concepts
- Mappings (e.g., $bd:numero \models num.remove\_suffix('/')$)

Un cas concret - résultat



id	cat	coll	titre	idA	annee	num
1	BD	Le génie des alpages	Les intondables	1	1980	03/14
2	roman	Le Disque-monde	Timbré	2	2004	
3	BD	Le génie des alpages	Cheptel maudit	1	2004	13/14

Table Livres

idA	auteur	nationalite	anneeNaissance
1	Fmurr	France	1946
2	Terry Pratchett	Royaume-Uni	1948

Table Auteures

Un résultat (idéal) de fusion des données :

id	cat	coll	titre	annee	num
1	BD	Le génie des alpages	Les intondables	1980	3
2	roman	Le Disque-monde	Timbré	2004	
3	BD	Le génie des alpages	Cheptel maudit	2004	13
1001	BD	Le génie des alpages	Sabotage et pâturage		11

Table Livres

idA	auteur	nationalité	annéeNaissance	alias
1	Richard Peyzaret	France	1946	Fmurr
2	Terry Pratchett	Royaume-Uni	1948	

Table Auteures

idA	idL
1	1
2	2
1	3
1	1001

Table Creatrice

Plan

Types d'architecture pour l'intégration

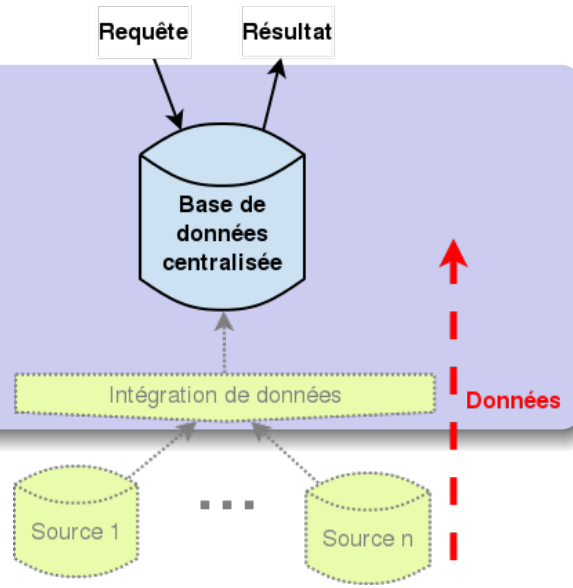
- Centralisation de données
- Intégration de jeux de données
- Entrepôt de données (data warehouses)
- Système fédéré
- Système médiateur
- Linked data (liens entre objets)
- Architecture P2P (intégration à la volée en environnement dynamique et large échelle, mais peu utilisé)

Centralisation de données

- Migration en amont de toutes les données des sources dans une seule BD
- Les sources de données initiales n'existent plus
- En général, interface ou API pour interroger la BD centralisée

☺ Données à jour
☺ Un seul point d'accès

☹ Forte dépendance à la source
☹ Effort d'intégration en amont

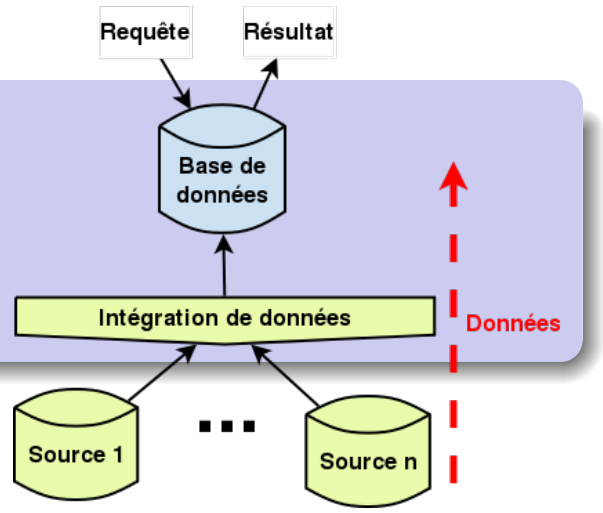


Exemples : Uniprot, GenBank

Intégration de jeux de données

- Duplication de données pertinentes dans une seule BD
- Les sources de données existent toujours
- Intégration en amont, avec un outil ou un script maison

☺ Personnalisation des données
☺ Un seul point d'accès
☹ Obsolescence des données



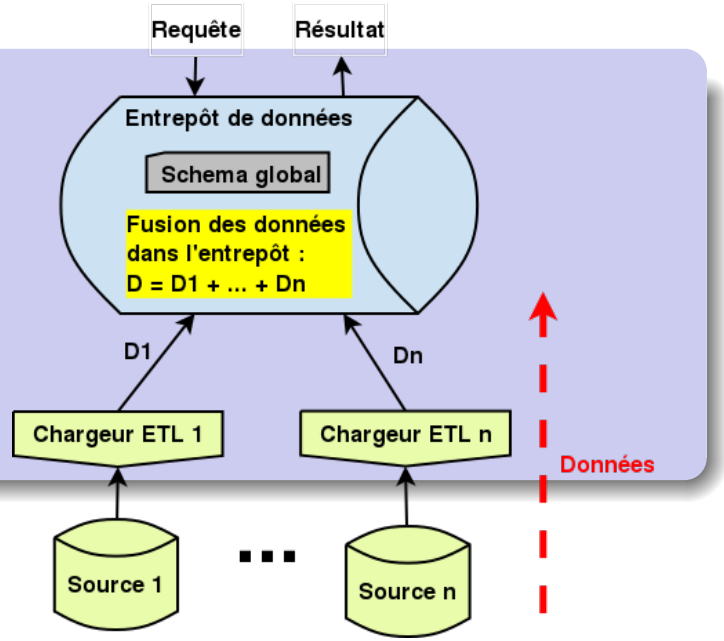
Exemple : le projet de BDBIO...

Entrepôt de données

- Duplication de données pertinentes dans un entrepôt
- Les sources de données existent toujours
- Intégration en amont avec un processus ETL (Extract - Transform - Load)
- Interrogation comme avec une base de données classique + data mining, analyse multidimensionnelle, etc.

☺ Personnalisation des données
☺ Temps de réponse immédiat
☺ Pas d'impact sur les sources

☹ Complexité de l'intégration
☹ Obsolescence des données



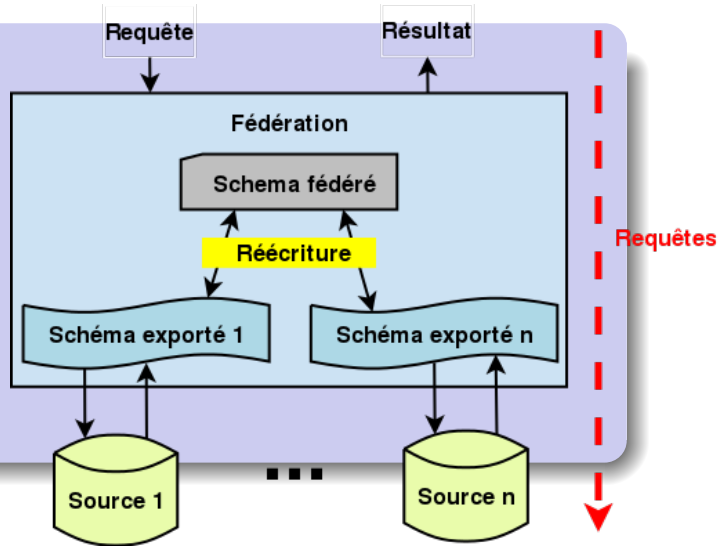
Exemples : String, Pathway commons

Système fédéré

- Intégration à la volée de sources avec une faible hétérogénéité (e.g., plusieurs BD relationnelles)
- Les données ne sont stockées que dans les sources
- Interrogation sur le schéma fédéré (décomposition de requêtes et recomposition des résultats)

☺ Processus minimum d'intégration
☺ Données à jour
☺ Effort distribué

☹ Réservé à des données faiblement hétérogènes



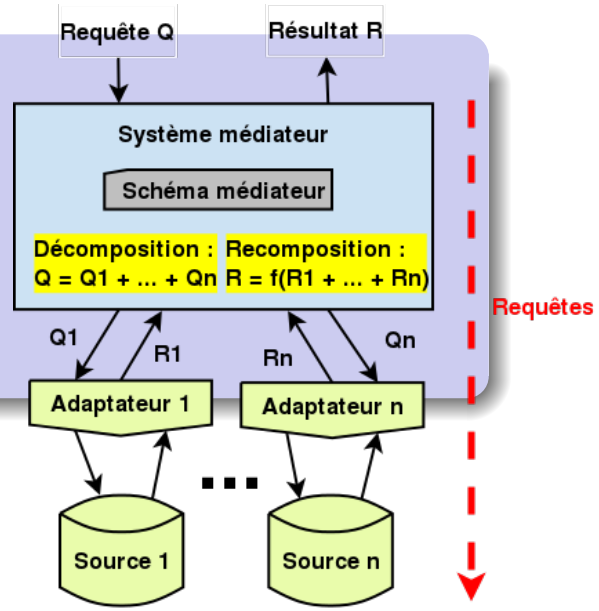
Exemple : PSIQUIC et son langage MIQL

Système médiateur

- Intégration à la volée (réécriture de requête)
- Les données ne sont stockées que dans les sources
- Interrogation sur le schéma médiateur (décomposition de requêtes et recomposition des résultats)

☺ Données à jour
☺ Effort distribué

☹ Réécriture de requêtes parfois complexe
☹ Temps de réponse variable



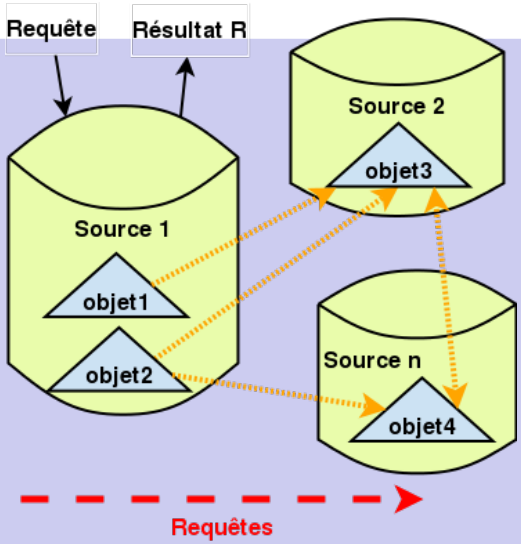
Exemples : ExPaSy, Entrez

Linked data

- Pas d'intégration, mais des liens (sémantiques) entre objets
- Les données ne sont stockées que dans les sources
- Interrogation propagée vers les différentes sources

☺ Données à jour
☺ Effort distribué

☹ Temps de réponse variable
☹ Agrégation manuelle des données
☹ Interrogation limitée ou écriture de requêtes complexes



Exemples : Bio2RDF

En résumé

Architectures :

- Intégration de jeux de données
- Entrepôt de données ou centralisation
- Système fédéré ou médiateur
- Linked data

---

http://fr.wikipedia.org/wiki/Fosse_Arenberg
http://fr.wikipedia.org/wiki/Tour_financi%C3%A8re_Bitexco
http://fr.wikipedia.org/wiki/R%C3%B8ros

Plan

Définition

Une **correspondance** est un lien sémantique entre deux éléments de sources de données différentes :

- Modèle : entre les propriétés, attributs, prédicats, tables, etc.
- Données : entre les n-uplets, entités, ressources, etc.

Une correspondance se représente (généralement) par un triplet $< e_1, e_2, \mathcal{R} >$ avec :

- e_1 et e_2 les deux éléments provenant de différentes sources
- $\mathcal{R}$ la relation sémantique entre ces éléments (e.g., équivalence, subsomption)

En réalité, une correspondance peut impliquer plus de deux éléments (correspondance complexe) :

- $<\{adresse, commune\}, \{numéro, rue, code\ postal, ville\}, \equiv >$

Motivation

Pourquoi découvrir des correspondances ?

- Pour aider à construire le schéma global ou évaluer l'intersection entre les sources de données (correspondances au niveau modèle)
- Pour la déduplication (correspondances aux niveaux modèle et données)
- Pour préparer la fusion des données (correspondances aux niveaux modèle et données)
- Pour la réécriture de requêtes (correspondances aux niveaux modèle et données)

Bellahsene et al., *Schema Matching and Mapping*, Springer (2011)

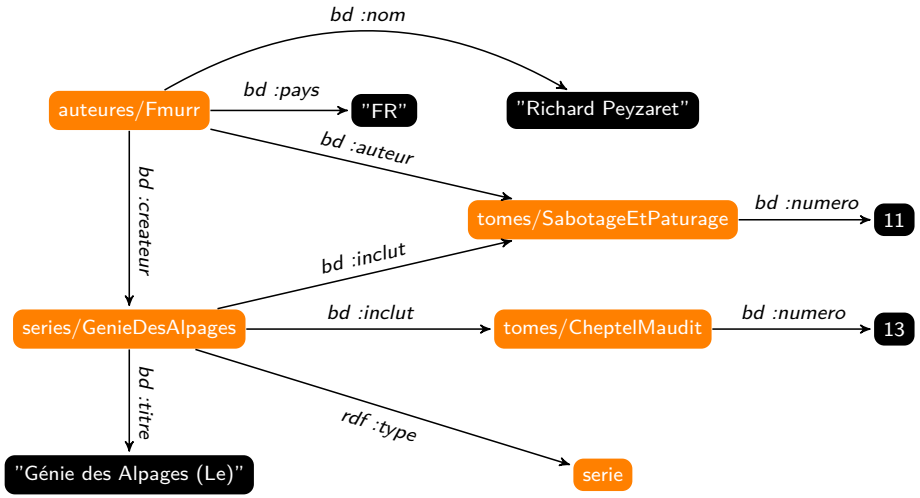
Découverte de correspondances

Découverte manuelle ou semi-automatique :

- En visualisant les sources de données (connaissance du domaine requise)
- En utilisant un outil d'appariement (e.g., basé sur des mesures de similarité)

http://dilbert.com/

Un cas concret - découverte des correspondances



Quel ensemble de correspondances entre ces deux sources de données ?

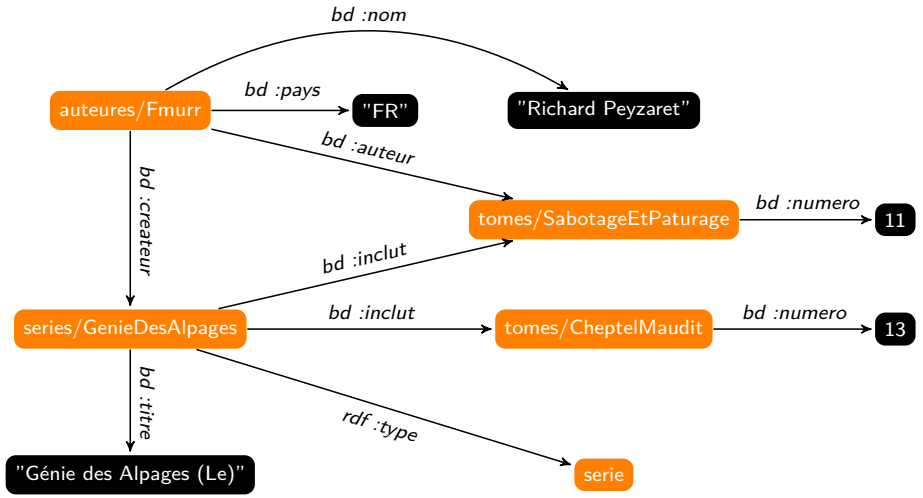
id	cat	coll	titre	idA	annee	num
1	BD	Le génie des alpages	Les intondables	1	1980	03/14
2	roman	Le Disque-monde	Timbré	2	2004	
3	BD	Le génie des alpages	Cheptel maudit	1	2004	13/14

Table Livres

idA	auteur	nationalite	anneeNaissance
1	Fmurr	France	1946
2	Terry Pratchett	Royaume-Uni	1948

Table Auteures

Un cas concret - découverte des correspondances



id	cat	coll	titre	idA	annee	num
1	BD	Le génie des alpages	Les intondables	1	1980	03/14
2	roman	Le Disque-monde	Timbré	2	2004	
3	BD	Le génie des alpages	Cheptel maudit	1	2004	13/14

Table Livres

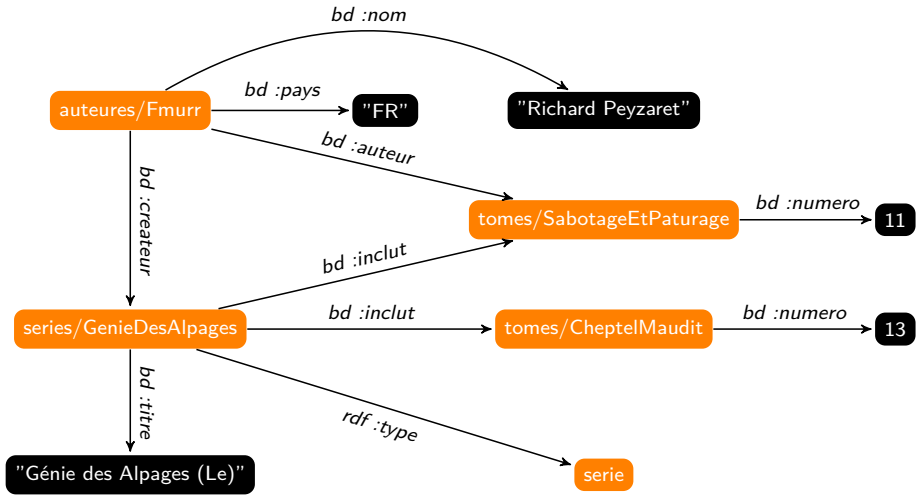
idA	auteur	nationalite	anneeNaissance
1	Fmurr	France	1946
2	Terry Pratchett	Royaume-Uni	1948

Table Auteures

Quel ensemble de correspondances entre ces deux sources de données ?

- $\mathcal{C}_{modele} = \{$
  $<serie, coll, \subseteq>,$
  $<bd\ :nom, auteur, \equiv>,$
  $<bd\ :numero, num, \equiv>$
  $<bd\ :pays, nationalité,$
  $\equiv>$
  $\}$

Un cas concret - découverte des correspondances



id	cat	coll	titre	idA	annee	num
1	BD	Le génie des alpages	Les intondables	1	1980	03/14
2	roman	Le Disque-monde	Timbré	2	2004	
3	BD	Le génie des alpages	Cheptel maudit	1	2004	13/14

Table Livres

idA	auteur	nationalite	anneeNaissance
1	Fmurr	France	1946
2	Terry Pratchett	Royaume-Uni	1948

Table Auteures

Quel ensemble de correspondances entre ces deux sources de données ?

- $\mathcal{C}_{modele} = \{$
  $<serie, coll, \subseteq>$,
  $<bd:nom, auteur, \equiv>$,
  $<bd:numero, num, \equiv>$
  $<bd:pays, nationalité,$
  $\equiv>$
  $\}$
- $\mathcal{C}_{donnees} = \{$
  $<auteures/Fmurr$, n-uplet Auteures $1, \equiv>$,
  $<tomes/CheptelMaudit$,
  n-uplet Livres $3, \equiv>$,
  $\}$

En résumé

La découverte de correspondances :

- Processus coûteux et rarement automatisé
- Nécessite une bonne connaissance du domaine
- Permet par la suite de construire le schéma global ou de découvrir les mappings

Plan

Définition

Un **schéma global** ou schéma intégré est un schéma qui regroupe (tous) les concepts des sources

Pourquoi définir un schéma global ?

- Pour stocker les données en respectant ce schéma
- Pour accéder aux données via ce schéma global (requêtes sur une seule interface)
- Pour évaluer le recoupement entre les concepts des différentes sources

---

Lenzerini. *Data integration: a theoretical perspective*. PODS (2002)
Halevy et al. *Data integration: the teenage years*. VLDB (2006)

Critères

Quels critères pour définir un schéma global ?

- Expressivité du modèle ou du format (e.g., une DTD est moins expressive que XML Schema)
- Flexibilité (garantir ou non la transformation de tous les concepts des sources et de toutes leurs données/contraintes)
- Complexité (en terme de structure à modéliser, selon le volume de données à traiter)

Chiticariu et al. *Interactive generation of integrated schemas*. SIGMOD (2008)

Construction d'un schéma global

Construction d'un schéma global (version simplifiée) :

1. Compréhension des connaissances et de la sémantique des sources
2. Choix d'un modèle, format, vocabulaire
3. Détection manuelle ou semi-automatique de correspondances entre les concepts (similaires) de différentes sources
4. Ajout des concepts dans le schéma global (minimalité, complétude)
5. Vérification par des expertes du domaine (cohérence, contraintes, etc.)

Construction d'un schéma global (2)

Un schéma global peut être construit manuellement ou semi-automatiquement

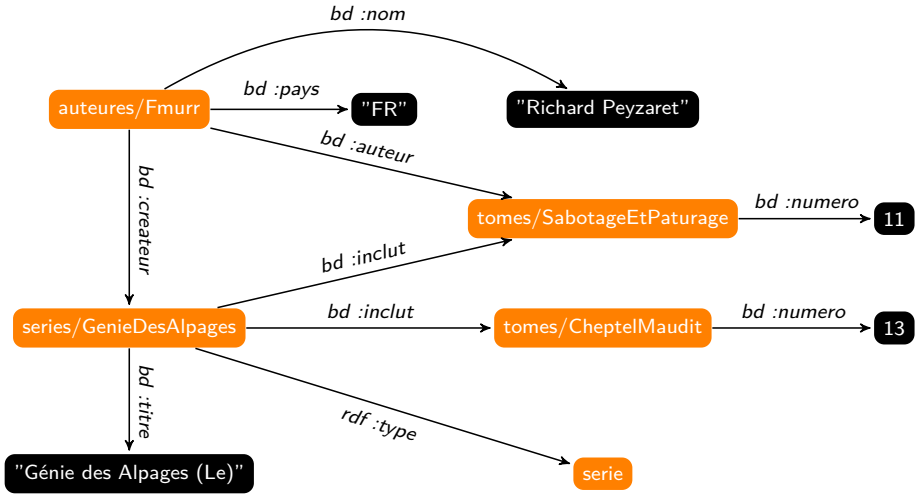
Le schéma / ontologie de l'une des sources de données peut servir de schéma global :

- Soit parce qu'il couvre tous les concepts des autres sources
- Soit parce que le schéma de cette source convient pour le stockage, pour l'interrogation ou aux utilisatrices

En général, choix parmi plusieurs schémas globaux possibles selon :

- Des contraintes techniques (e.g., logicielles, réseau)
- Des contraintes de modélisation
- Des préférences utilisateurs

Un cas concret - schéma global



Quel schéma global pour ces deux sources ?

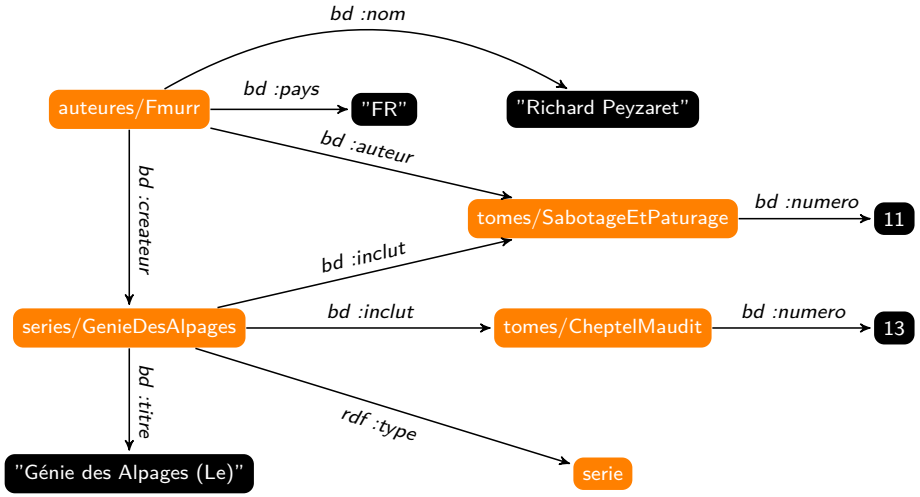
id	cat	coll	titre	idA	annee	num
1	BD	Le génie des alpages	Les intondables	1	1980	03/14
2	roman	Le Disque-monde	Timbré	2	2004	
3	BD	Le génie des alpages	Cheptel maudit	1	2004	13/14

Table Livres

idA	auteur	nationalite	anneeNaissance
1	Fmurr	France	1946
2	Terry Pratchett	Royaume-Uni	1948

Table Auteures

Un cas concret - schéma global



id	cat	coll	titre	idA	annee	num
1	BD	Le génie des alpages	Les intondables	1	1980	03/14
2	roman	Le Disque-monde	Timbré	2	2004	
3	BD	Le génie des alpages	Cheptel maudit	1	2004	13/14

Table Livres

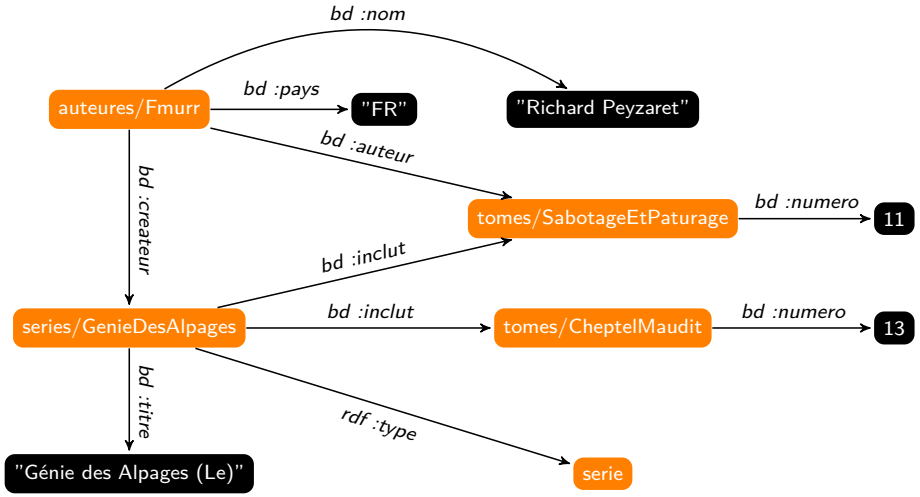
idA	auteur	nationalite	anneeNaissance
1	Fmurr	France	1946
2	Terry Pratchett	Royaume-Uni	1948

Table Auteures

Quel schéma global pour ces deux sources ?

- Choix du modèle relationnel (simplicité et plus grande couverture des concepts)

Un cas concret - schéma global



id	cat	coll	titre	idA	annee	num
1	BD	Le génie des alpages	Les intondables	1	1980	03/14
2	roman	Le Disque-monde	Timbré	2	2004	
3	BD	Le génie des alpages	Cheptel maudit	1	2004	13/14

Table Livres

idA	auteur	nationalite	anneeNaissance
1	Fmurr	France	1946
2	Terry Pratchett	Royaume-Uni	1948

Table Auteures

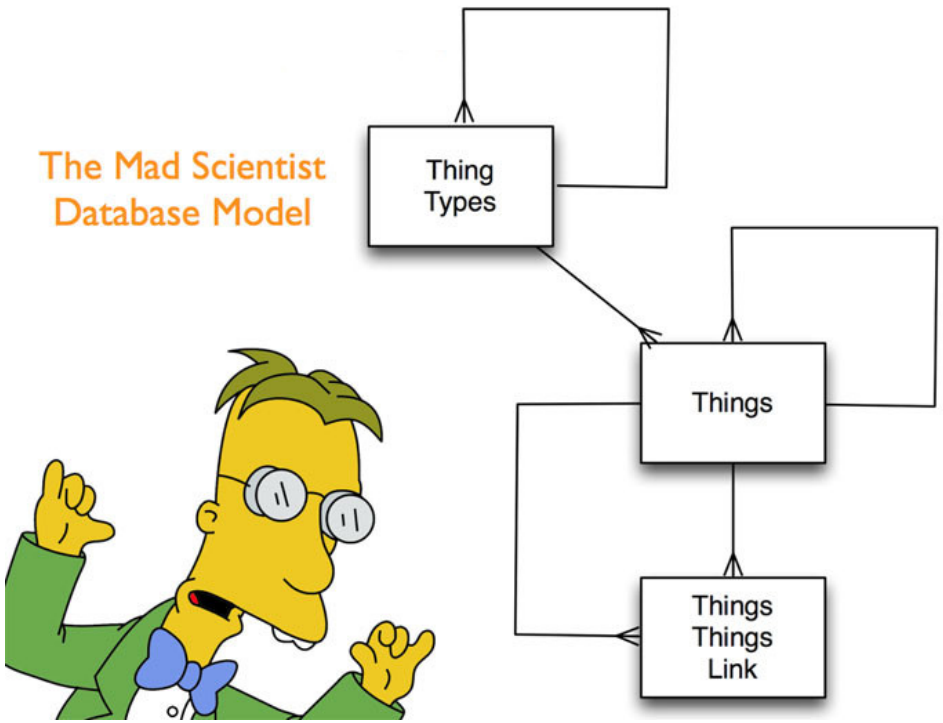
Quel schéma global pour ces deux sources ?

- Choix du modèle relationnel (simplicité et plus grande couverture des concepts)
- Un schéma possible :
  `Livres (idL, titre, coll, annee, num)`
  `Auteures (idA, nom, pays, anneeNaiss)`
  `Créateur(#idA, #idL)`

En résumé

Définition du schéma global :

- Schéma utilisé pour l'interrogation ou le stockage
- En général, nécessite de découvrir les correspondances entre les sources (i.e., concepts similaires)
- Validation du schéma par les expertes du domaine



Plan

Définition

Un **mapping** $m =< e_1, e_2, \mathcal{R}, f >$ est une correspondance directionnelle qui spécifie en plus la fonction de transformation f entre deux éléments du modèle, i.e., la technique pour transformer une valeur conforme à e_1 en une instance conforme à e_2

Un mapping permet :

- De transformer les données des sources afin de se conformer au modèle du schéma global
- De transformer les données d'une source afin de se conformer au modèle d'une autre source

Fagin et al. *Data exchange: getting to the core*. TODS (2005)

Motivation

Pourquoi découvrir des mappings ?

- Pour éviter les redondances lors du stockage et/ou de l'interrogation, et maintenir la cohérence
- Pour garantir la complétude en terme de résultat d'une requête (e.g., sur les attributs d'un concept)
- Pour corriger les éventuels problèmes d'hétérogénéité
- Pour fusionner les instances contenues dans les différentes sources

Problèmes d'hétérogénéité au niveau modèle

Conflits de label :

- Homonymie : le même label est utilisé pour désigner des concepts différents (e.g., *titre*)
- Synonymie : le même concept est décrit par plusieurs labels différents (e.g., *auteures* / *écrivaines*)

Conflits structurels :

- Le même concept est représenté par des structures différentes dans le modèle (e.g., attributs manquants et/ou différents)
- Deux concepts ne possèdent pas les mêmes contraintes (e.g., type de données, cardinalité)

Conflits sur les relations sémantiques (e.g., sous-classe, agrégation)

Problèmes d'hétérogénéité au niveau données

Conflit sémantique :

- La même entité est représentée par différents objets/valeurs dans les sources
- La même entité est interprétée différemment dans les différentes sources

Conflit d'identité (clés différentes) :

- Le même concept est identifié par différent(s) attribut(s) dans les schémas sources
- La même valeur de clé dans deux sources différentes ne signifie pas forcément que les deux objets soient les mêmes

Résolution des problèmes d'hétérogénéité

Conflits de label :

- Renommage des concepts

Conflits structurels - attributs manquants :

- Certains modèles (RDF) sont flexibles au niveau du schéma !
- Ajout de valeurs (e.g., en SQL, produit cartésien avec une table intermédiaire contenant des NULL)

Conflits structurels - attributs différents :

- Fonction de transformation (e.g., unités de mesure)
- Jointure avec une table de correspondances (e.g., nom_pays / code_pays)

nom pays	code pays
Allemagne	D
France	FR
Espagne	ES
...	...

Résolution des problèmes d'hétérogénéité (2)

Conflit sémantique :

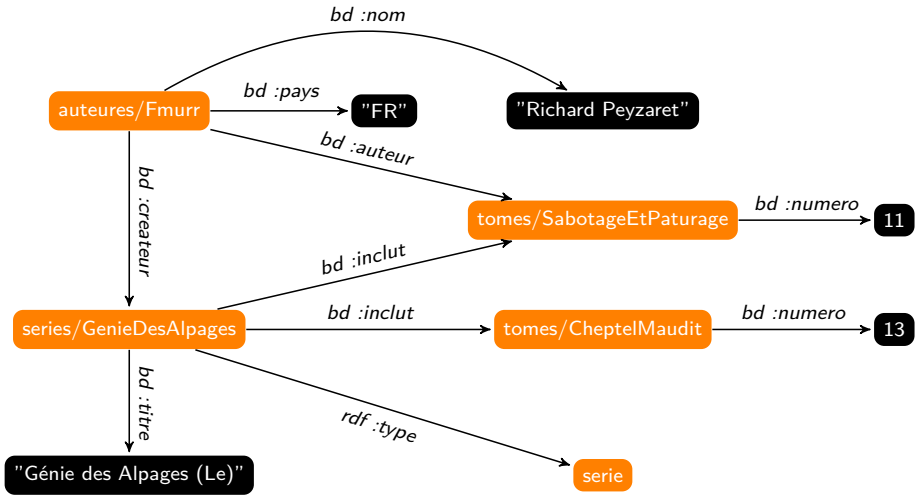
- Sélection des valeurs d'une source (provenance, degré de confiance, qualité)
- Alignement d'entité avec base de données externe

Conflit d'identité :

- Définition d'une fonction de conversion des clés (e.g. `conversion(clé) =` α `+ clé`)
- Jointure avec une table de correspondances

C'est au niveau de la fonction de transformation du mapping que l'on résout ces conflits

Un cas concret - conflits pour S1



Quels conflits entre la source S1 et le schéma global SG ?

Schéma global (SG) :

```
Livres (idL, titre, coll, annee, num)
Auteures (idA, nom, pays, anneeNaiss)
Créatrice(#idA, #idL)
```

Un cas concret - conflits pour S1

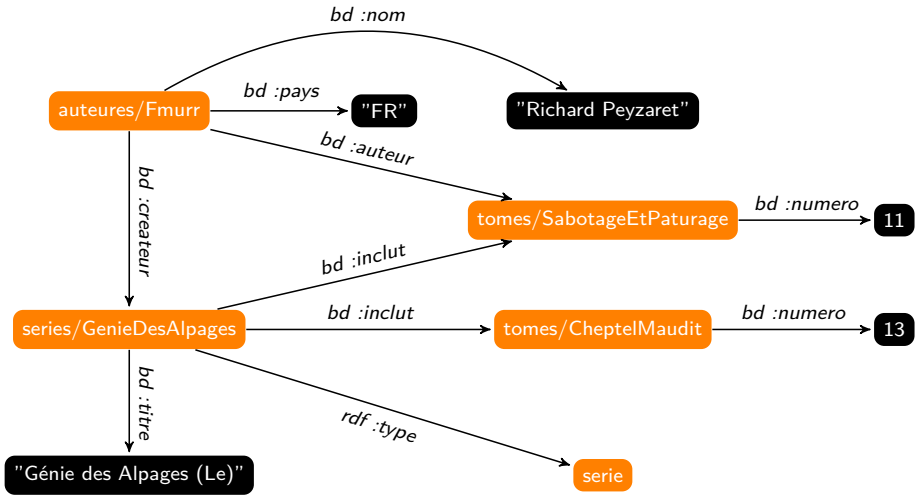


Schéma global (SG) :

```
Livres (idL, titre, coll, annee, num)
Auteures (idA, nom, pays, anneeNaiss)
Créatrice(#idA, #idL)
```

Quels conflits entre la source S1 et le schéma global SG ?

- De label
- Structurels (pas de *titre*, *année* et *annéeNaissance* dans S1)
- D'identité
- Sémantiques (valeurs différentes pour *SG.coll*, *SG.nom*)

Un cas concret - conflits pour S2

id	cat	coll	titre	idA	annee	num
1	BD	Le génie des alpages	Les intondables	1	1980	03/14
2	roman	Le Disque-monde	Timbré	2	2004	
3	BD	Le génie des alpages	Cheptel maudit	1	2004	13/14

Table Livres

idA	auteur	nationalite	anneeNaissance
1	Fmurr	France	1946
2	Terry Pratchett	Royaume-Uni	1948

Table Auteures

Quels conflits entre la source S2 et le schéma global SG ?

Schéma global (SG) :

`Livres (idL, titre, coll, annee, num)`
`Auteures (idA, nom, pays, anneeNaiss)`
`Créatrice(#idA, #idL)`

Un cas concret - conflits pour S2

id	cat	coll	titre	idA	annee	num
1	BD	Le génie des alpages	Les intondables	1	1980	03/14
2	roman	Le Disque-monde	Timbré	2	2004	
3	BD	Le génie des alpages	Cheptel maudit	1	2004	13/14

Table Livres

idA	auteur	nationalite	anneeNaissance
1	Fmurr	France	1946
2	Terry Pratchett	Royaume-Uni	1948

Table Auteures

Schéma global (SG) :

```
Livres (idL, titre, coll, annee, num)
Auteures (idA, nom, pays, anneeNaiss)
Créatrice(#idA, #idL)
```

Quels conflits entre la source S2 et le schéma global SG ?

- De label
- Sémantiques (valeurs différentes pour *SG.num*, *SG.nom* et *SG.nationalite*)

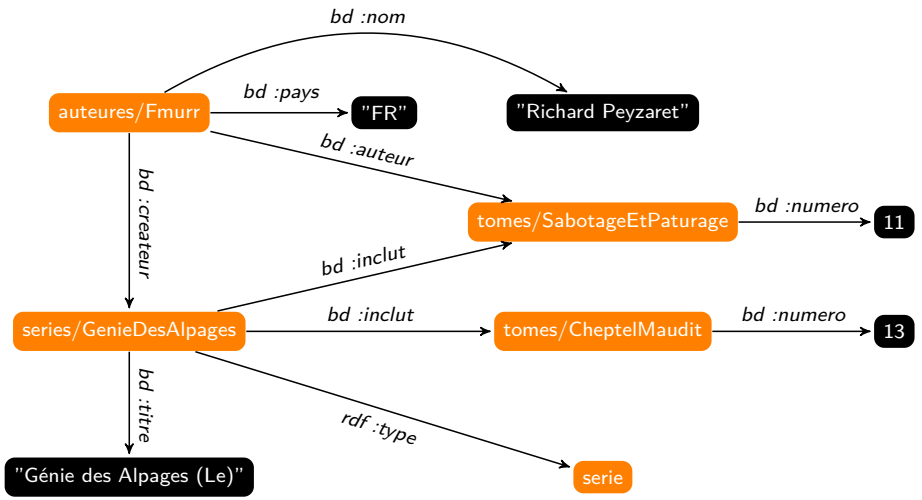
Formalisme des mappings

Plusieurs approches pour le formalisme des mappings :

- Global As View (GAV) = le schéma global est défini comme un ensemble de vues sur les sources
- Local As View (LAV) = les sources sont définies comme des vues sur le schéma global
- Global Local As View (GLAV) ou Both As View (BAV) = bidirectionnel

Un mapping correspondra à une ou plusieurs requêtes / programmes (SQL, SPARQL, XQuery, etc.) sur les sources (GAV) ou sur le schéma global (LAV)

Un cas concret - mappings GAV



id	cat	coll	titre	idA	annee	num
1	BD	Le génie des alpages	Les intondables	1	1980	03/14
2	roman	Le Disque-monde	Timbré	2	2004	
3	BD	Le génie des alpages	Cheptel maudit	1	2004	13/14

Table Livres

idA	auteur	nationalite	anneeNaissance
1	Fmurr	France	1946
2	Terry Pratchett	Royaume-Uni	1948

Table Auteures

Schéma global (SG) :

`Livres (idL, titre, coll, annee, num)`
`Auteures (idA, nom, pays, anneeNaiss)`
`Créatrice(#idA, #idL)`

Quels mappings GAV ?

SG.Livres(idl, titre, coll, an, num) $\supseteq$ S1(-, -, -, idl, num, -, coll), S2.Livres(idl, -, coll, titre, -, an, num)

SG.Auteures(ida, nom, pays, an) $\supseteq$ S1(ida, nom, pays, -, -, -, -), S2.Auteures(ida, nom, pays, an)

SG.Creatrice(ida, idl) $\supseteq$ S1(ida, -, -, idl, -, -, -), S2.Livres(idl, -, -, -, ida, -, -)

Un cas concret - mappings GAV (2)

Pour le mapping GAV de SG.Livres :

SG.Livres(idl, titre, coll, an, num) $\supseteq$ S1(-, -, -, idl, num, -, coll), S2.Livres(idl, -, coll, titre, -, an, num)

```
3  SELECT ?idl ('null' as ?titre) ?coll ('null' as ?an) ?num
4  FROM <http://example.org/S1>
5  WHERE {
6    ?idl bd:numero ?num .
7    _:s bd:inclut ?idl .
8    _:s bd:titre ?coll .
9  }
10
11 SELECT id, titre, coll, annee, SUBSTRING_INDEX(num, '/', 1) AS
       num
12 FROM S2.Livres ;
```

Pour S1.idl, il faut en plus créer une table de correspondances (e.g., "tomes/CheptelMaudit" $\leftrightarrow$ 1001)

Un cas concret - mappings GAV (3)

Pour le mapping GAV de SG.Auteures :

SG.Auteures(ida, nom, pays, an) $\supseteq$ S1(ida, nom, pays, -, -, -, -),
S2.Auteures(ida, nom, pays, an)

```
16 SELECT ?ida ?nom ?pays ('null' as ?an)
17 FROM <http://example.org/S1>
18 WHERE {
19   ?ida bd:nom ?nom .
20   ?ida bd:pays ?pays .
21 }
22
23 SELECT idA, auteur AS nom, codePays AS pays, anneeNaiss AS an
24 FROM S2.Auteures a INNER JOIN CorrespPaysNat c ON a.nationalite
       = c.codePays ;
```

Pour `?ida`, il faut créer une table de correspondances (e.g., "/auteures/Fmurr" $\leftrightarrow$ 1234). Pour S2, il faut utiliser une table de correspondances entre les codes pays et les nationalités

Un cas concret - mappings GAV (4)

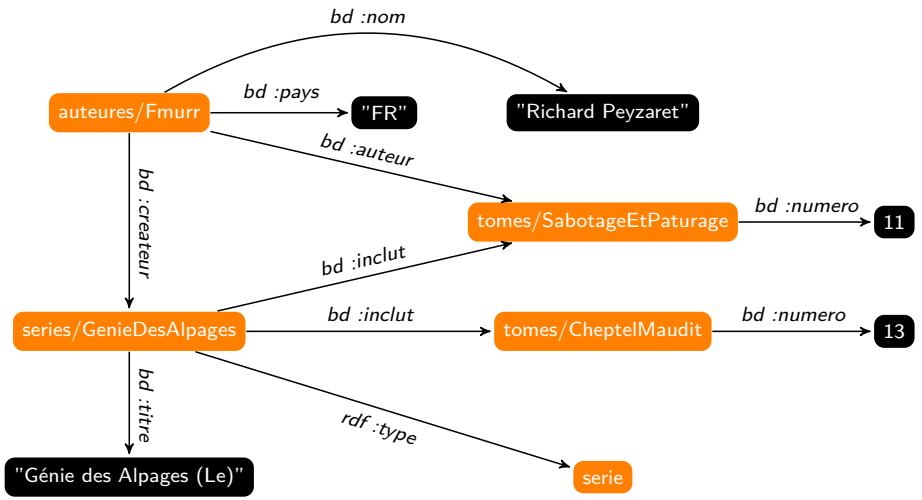
Pour le mapping GAV de SG.Creatrice :

SG.Creatrice(ida, idl) $\supseteq$ S1(ida, -, -, idl, -, -, -), S2.Livres(idl, -, -, -, ida, -, -)

```
28 SELECT ?ida ?idl
29 FROM <http://example.org/S1>
30 WHERE {
31   ?ida bd:auteur ?idl .
32 }
33
34 SELECT a.idA AS ida, id AS idl
35 FROM S2.Livres;
```

Pour ?ida et ?idl, il faut utiliser les tables de correspondances pour retrouver l'identifiant numérique qui correspond à la ressource

Un cas concret - mappings LAV



id	cat	coll	titre	idA	annee	num
1	BD	Le génie des alpages	Les intondables	1	1980	03/14
2	roman	Le Disque-monde	Timbré	2	2004	
3	BD	Le génie des alpages	Cheptel maudit	1	2004	13/14

Table Livres

idA	auteur	nationalite	anneeNaissance
1	Fmurr	France	1946
2	Terry Pratchett	Royaume-Uni	1948

Table Auteures

Schéma global (SG) :

`Livres (idL, titre, coll, annee, num)`
`Auteures (idA, nom, pays, anneeNaiss)`
`Créatrice(#idA, #idL)`

Quels mappings LAV ?

S1(ida, nom, pays, idl, num, ids, titre) $\subseteq$ SG.Livres(idl, -, titre, -, num), SG.Auteures(ida, nom, pays, -), SG.Creatrice(ida, idl)

S2.Livres(idl, cat, coll, titre, ida, an, num) $\subseteq$ SG.Livres(idl, titre, coll, an, num), SG.Creatrice(ida, idl)

S2.Auteures(ida, aut, nat, an) $\subseteq$ SG.Auteures(ida, aut, nat, an)

Un cas concret - mappings LAV (2)

Pour le mapping LAV de S1 :

S1(ida, nom, pays, idl, num, ids, titre) $\subseteq$ SG.Livres(idl, -, titre, -, num), SG.Auteures(ida, nom, pays, -), SG.Creatrice(ida, idl)

```
39 SELECT ressourceIdA, a.nom, a.pays, ressourceIdL, l.num, 'series
       /' || REPLACE(l.coll, ' ', ''), l.coll
40 FROM SG.Livres l NATURAL JOIN SG.creatrice NATURAL JOIN SG.
       Auteures a INNER JOIN CorrespAuteures ca ON a.idA = ca.
       ressourceIdA INNER JOIN CorrespLivres cl ON l.idL = cl.
       ressourceIdL;
```

Pour les ressources, on peut utiliser une table de correspondances (e.g., `idA` et `idL`) ou générer une URI de ressource (e.g., ressource série, fabriquée à partir de l'attribut `coll` auquel on supprime les espaces)

Un cas concret - mappings LAV (3)

Pour le mapping LAV de S2 :

S2.Livres(idl, cat, coll, titre, ida, an, num) $\subseteq$ SG.Livres(idl, titre, coll, an, num), SG.Creatrice(ida, idl)

S2.Auteures(ida, aut, nat, an) $\subseteq$ SG.Auteures(ida, aut, nat, an)

```
44 SELECT l.idL, NULL, coll, titre, c.idA, annee, num || '/?'
45 FROM SG.Livres l NATURAL JOIN SG.Creatrice c ;
46
47 SELECT idA, nom, c.nationalite, anneeNaiss
48 FROM SG.Auteures a INNER JOIN CorrespPaysNat c ON c.nationalite
       = a.pays;
```

Pour l'attribut `num`, on pourrait compter le nombre de n-uplets dans la série (et concaténer cette valeur au `num`). L'attribut `cat` ne peut être renseigné (information non disponible dans le SG)

Conclusion sur le formalisme des mappings

Avantages / inconvénients de GAV :

- Traduction directe des requêtes utilisatrices
- Difficulté pour prendre en compte l'évolution des sources
- Difficulté pour le passage à l'échelle

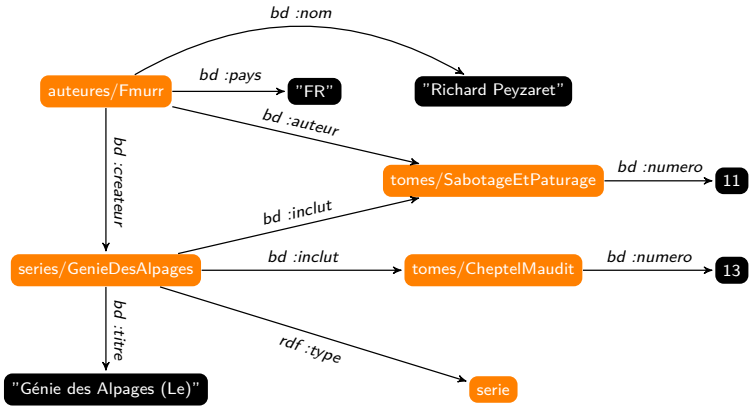
Avantages / inconvénients de LAV :

- Prise en compte de l'évolution
- Aisance pour la prise en compte de l'hétérogénéité
- Nécessité de réécrire les requêtes utilisatrices (algorithmes *Bucket*, *Minicon*, *Inverse Rules*)
- Si le schéma global évolue, nécessité de redéfinir les mappings des schémas source

---

Halevy A, *Answering queries using views: A survey*. VLDBJ (2001)

Exemple de fusion de données



id	cat	coll	titre	idA	annee	num
1	BD	Le génie des alpages	Les intondables	1	1980	03/14
2	roman	Le Disque-monde	Timbré	2	2004	
3	BD	Le génie des alpages	Cheptel maudit	1	2004	13/14

Table Livres

idA	auteur	nationalite	anneeNaissance
1	Fmurr	France	1946
2	Terry Pratchett	Royaume-Uni	1948

Table Auteures

Un résultat (idéal) de fusion des données :

id	cat	coll	titre	annee	num
1	BD	Le génie des alpages	Les intondables	1980	3
2	roman	Le Disque-monde	Timbré	2004	
3	BD	Le génie des alpages	Cheptel maudit	2004	13
1001	BD	Le génie des alpages	Sabotage et pâturage		11

Table Livres

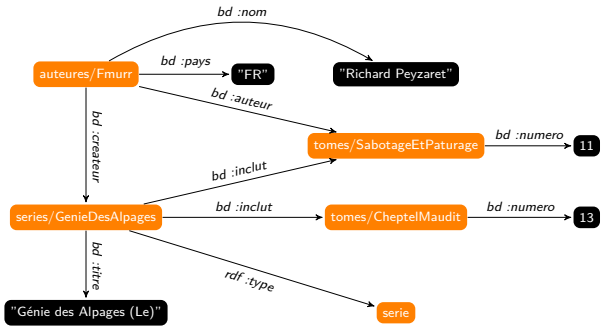
idA	auteur	nationalité	annéeNaissance	alias
1	Richard Peyzaret	France	1946	Fmurr
2	Terry Pratchett	Royaume-Uni	1948	

Table Auteures

idA	idL
1	1
2	2
1	3
1	1001

Table Creatrice

Exemple de migration - relationnel vers RDF



id	cat	coll	titre	idA	annee	num
1	BD	Le génie des alpages	Les intondables	1	1980	03/14
2	roman	Le Disque-monde	Timbré	2	2004	
3	BD	Le génie des alpages	Cheptel maudit	1	2004	13/14

Table Livres

idA	auteur	nationalite	anneeNaissance
1	Fmurr	France	1946
2	Terry Pratchett	Royaume-Uni	1948

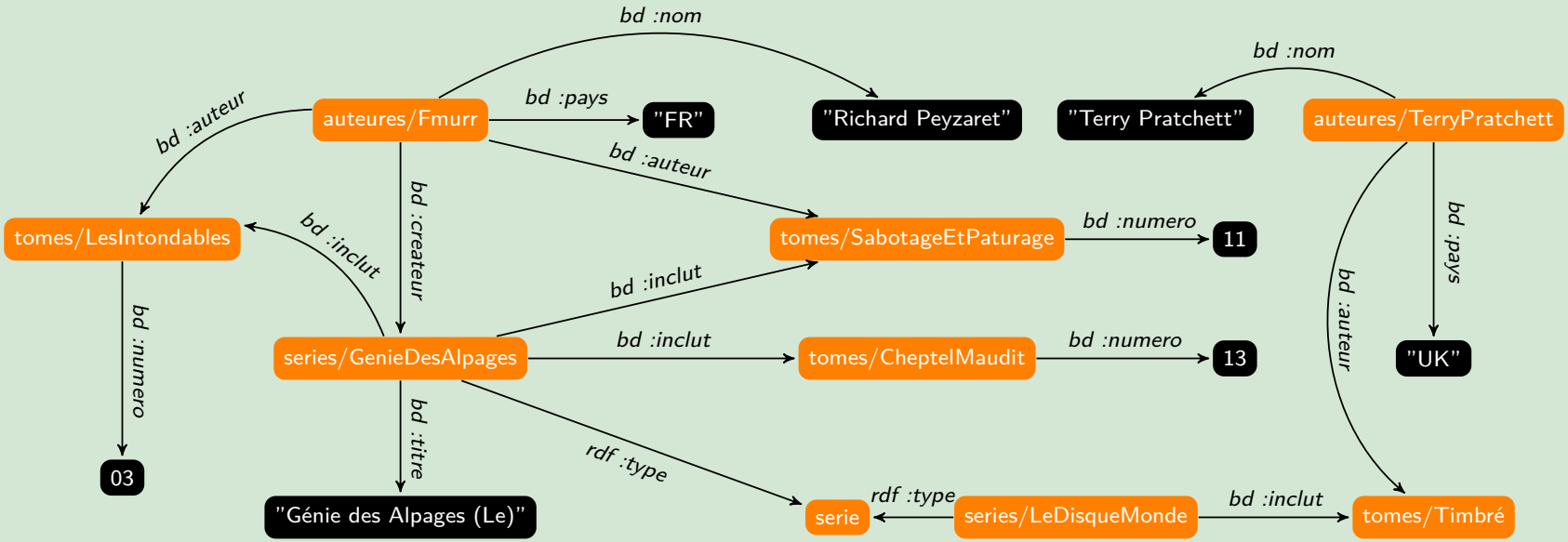
Table Auteures

```
1  bdrdf = connexion(S1) # graphe RDF
2  bdrel = connexion(S2) # BD relationnelle
3
4  instances = bdrel.executer('select * from
       Auteures a join Livres l on l.idA=a.
       idA')
5  pour chaque i de instances:
6   res_a = bdrdf.executer("select ?r where
       {{?r bd:nom i['auteur'] .} UNION {
       auteures/i['auteur'] bd:nom ?v .} }")
7   if !res_a:  # auteur inexistant
8    res_a = "auteures/" + i['auteur'].
       replace(' ', '')
9    pays = get_code_pays(i['pays'])  #
       France -> FR
10   bdrdf.executer("insert data {res_a bd:
       nom i['auteur']}")
11   bdrdf.executer("insert data {res_a bd:
       pays pays}")
12  res_c = bdrdf.executer("select ?r where
       {?r bd:titre clean(i['coll'])}")
```

```
13  if !res_c:  # coll. inexistante
14   res_c = "series/" + i['coll'].
       replace(' ', '')
15   bdrdf.executer("insert data {res_c
       bd:titre clean(i['coll'])}")
16   bdrdf.executer("insert data {res_c
       rdf:type serie")
17  # insertion du tome
18  res_t = bdrdf.executer("select ?r
       where {?r bd:numero ?n}")
19  if !res_t:  # tome inexistant
20   res_t = "tomes/" + i['titre'].
       replace(' ', '')
21   num = i['num'].split('/')[0]
22   bdrdf.executer("insert data {res_t
       bd:numero num .}")
23   bdrdf.executer("insert data {res_a
       bd:auteur res_t .}")
24   bdrdf.executer("insert data {res_c
       bd:inclut res_t .}")
```

Exemple de migration - relationnel vers RDF (2)

Résultat de la migration :



Algorithme perfectible :

- Perte de données (catégorie, années, livres sans auteur, etc.)
- Pas de mise à jour des données (si auteure ou tome existant)
- Choix de créer ou non certains liens (e.g., `bd:createur`)
- Redondance possible : détection des entités correspondantes (e.g., *Richard Peyzaret* et *Fmurr*), etc.

Exemple d'interrogation distribuée

Requête : la liste des titres de livre

```
1  bdrdf = connexion(S1) # graphe RDF
2  bdrel = connexion(S2) # BD
       relationnelle
3
4  results = bdrel.executer('select titre
       from Livres')
5  tomes_rdf = bdrdf.executer("select ?r
       where {?r bd:numero ?n}")
6
7  # ajout des résultats RDF si non
       existants
8  pour chaque t de tomes_rdf:
9   titre = clean(t)
10  corresp = entity_matching(results, t)
        # cherche si un titre
       équivalent existe déjà (eg, par
       comparaison approximative des
       titres)
11  if !corresp:  # pas de correspondance
       dans results, donc ajout du
       titre
12   results.add(titre)
```

Résultat de l'interrogation :

Les intondables
Timbré
Cheptel maudit
Sabotage et paturage

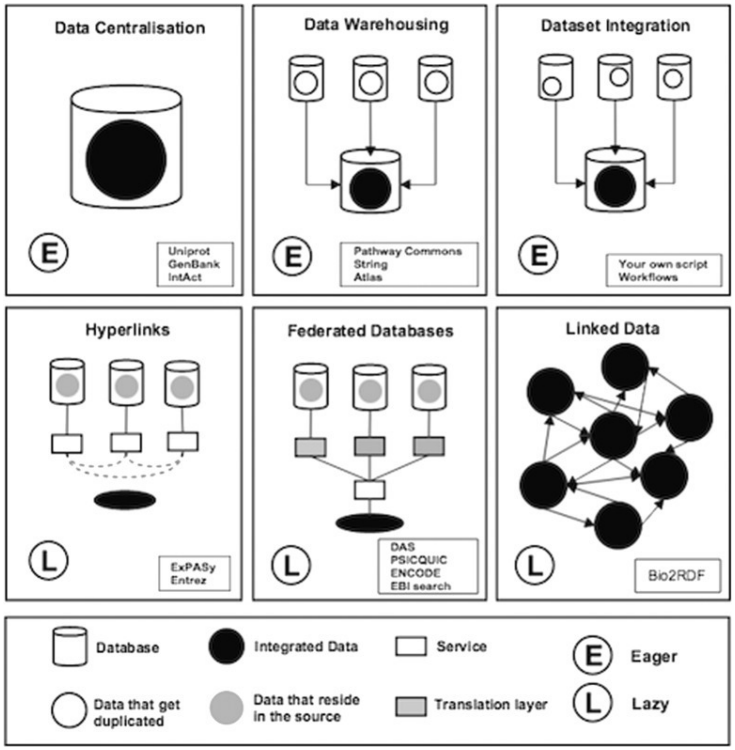
Algorithme perfectible :

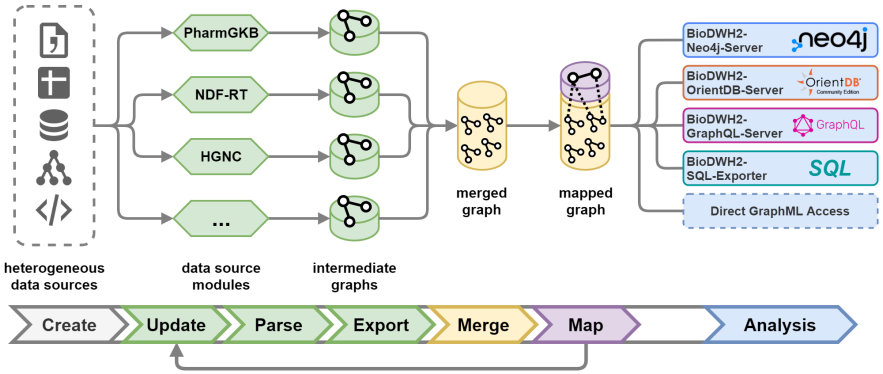
- Algorithme non générique (ici pour une seule requête)
- Redondance possible (malgré l'appariement d'entités)

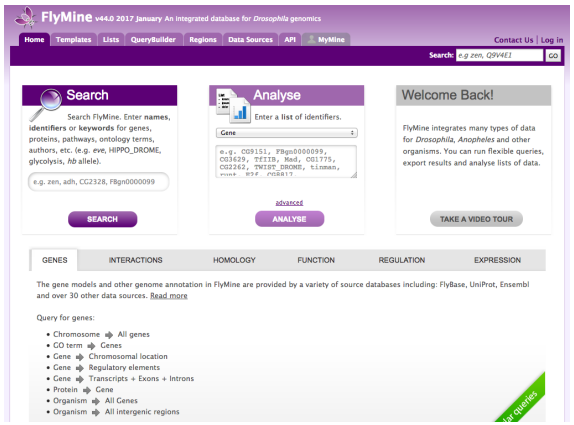
Outils d'intégration de données

- Outils industriels (souvent limités) : Talend Data Integration, OpenRefine, Pentaho Data Integration, IBM Information Integrator, Clover DX, XAware, Data Mapper
- Prototypes de recherche (contexte d'utilisation parfois spécifiques, e.g., large échelle, correspondances complexes) :
  - appariement de schéma / ontologie (e.g., BigGorilla, Harmony, COMA, AgreementMaker)
  - appariement d'instances (e.g., BigGorilla, FEBRL, Karma)
- Benchmarks pour évaluer et comparer les outils

Outils d'intégration de données en bioinformatique







Schneider et al., *Teaching the Fundamentals of Biological Data Integration Using Classroom Games*. PLoS Comput Biol (2012)

Projets bio.tools et jib.tools (Integrative Bioinformatics), outils BioDWH2, Intermine, CellBase, Ondex, Flymine

Conclusion

Intégration de données : fournir un accès uniforme à plusieurs sources hétérogènes stockées sur des sites autonomes

Processus pouvant être utilisés pour l'intégration :

- Définition de correspondances (liens entre concepts)
- Définition d'un schéma global (schéma regroupant tous les concepts des sources)
- Définition de mappings (écrits sous forme de vues intégrantes)

Ces processus sont utilisés différemment selon l'architecture d'intégration envisagée (fédération, entrepôt, médiateur, etc.)

Fillinger et al. *Challenges of big data integration in the life sciences*. Analytical and Bioanalytical Chemistry (2019)

La suite en M2

- MIF24 - bases de données non relationnelles : ORM, NoSQL avancé, algèbre de collection, RDF avancé
- AD - analyse de données : apprentissage, datamining, OLAP, recommandation
- GGMD - gestion de grandes masses de données : distribution des données et des traitements, gestion de flux
- IQD - intégration et qualité des données : intégration de données avancée