Construction d'une ontologie par programmation basée sur l'exemple

Quentin Riché-Piotaix

Université de Poitiers - LIAS ENSMA - CHU de Poitiers

4 mai 2017

Introduction

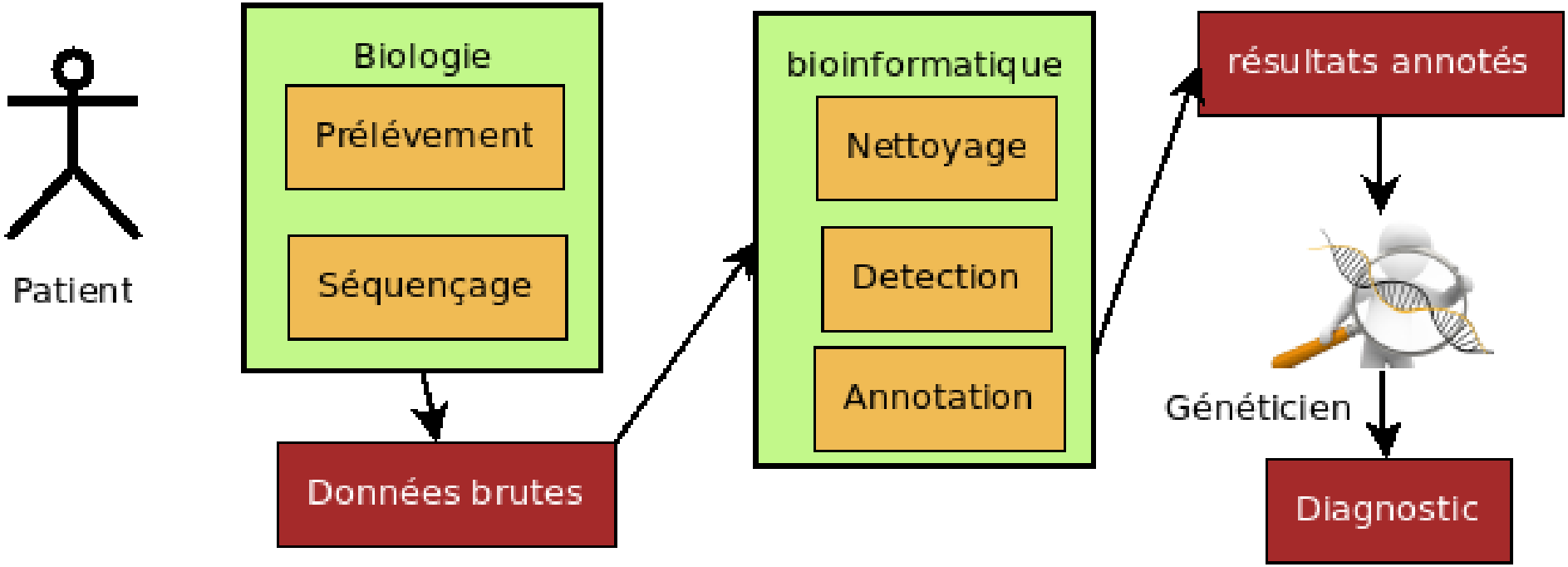
Encadrement

- Patrick Girard, LIAS - Université de Poitiers
- Ladjel Bellatreche, LIAS
- Frédéric Bilan, CHU de Poitiers - Université de Poitiers

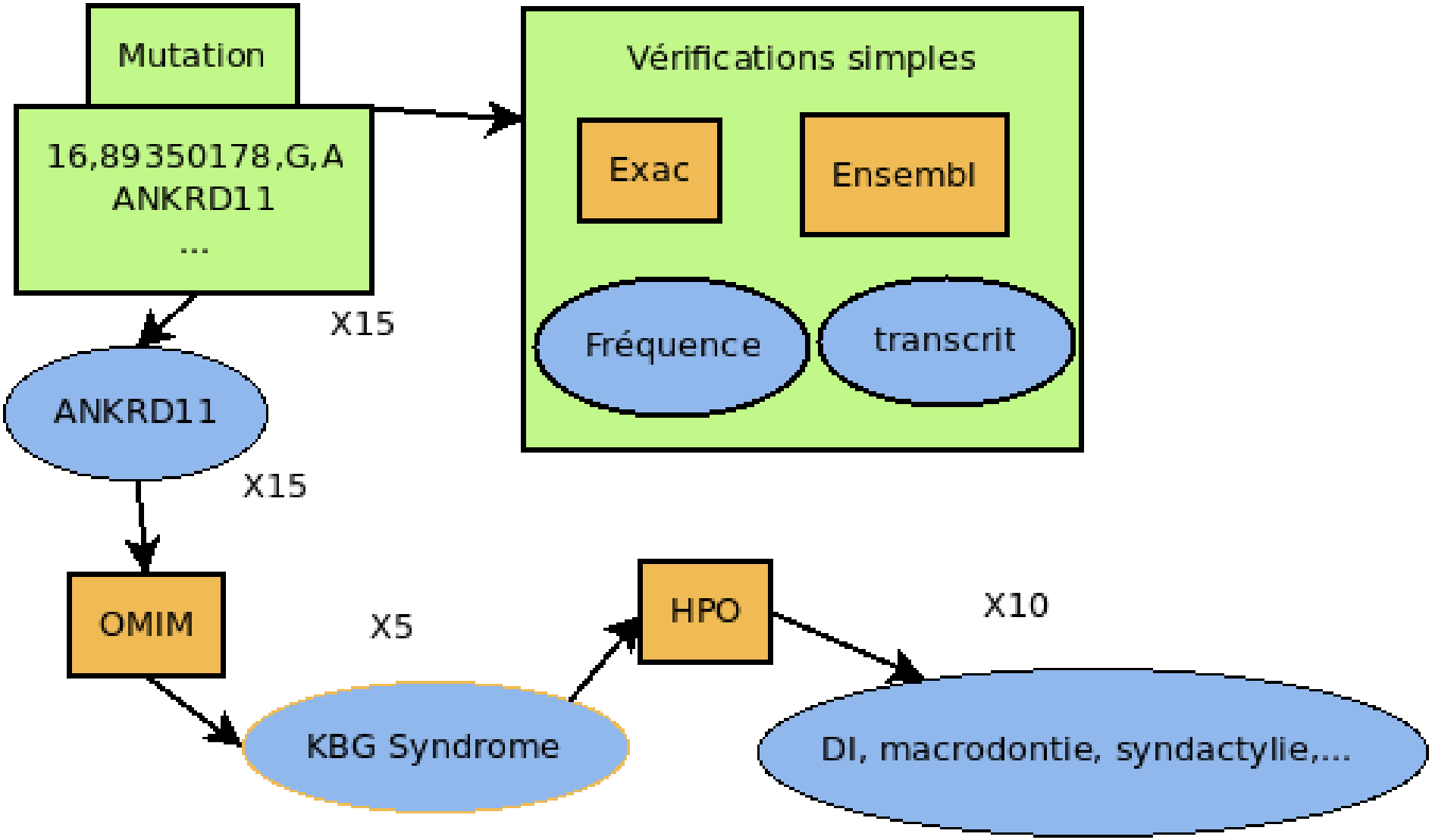
Sujet

Exploration à base ontologique de données issues de patients atteints de maladies génétiques rares.

Contexte : La Génétique



Scénario exemple



$\longrightarrow$ Besoin de rassembler les données

Contexte : La Génétique

Les bases de données génétiques

- sont nombreuses
- sont variées
- sont spécialisées
- se recoupent
- sont sémantiquement inconstantes
- sont disponibles en fichiers semi structurés

Module *ad hoc* de transfert ?
Ontologie ?

Contexte : La Génétique

Spécificité du domaine

Constante évolution, ajout et modification de savoir réguliers et fréquents

$\longrightarrow$ Ontologie globale impossible à maintenir

Spécificité du domaine

Les généticiens n'ont pas besoin d'une ontologie couvrant tout le domaine : Poitiers s'intéresse particulièrement à la déficience intellectuelle

$\longrightarrow$ Créations de petites ontologies spécifiques

Spécificité du domaine

Les généticiens n'ont pas possibilité de travailler avec un ingénieur spécialisé

$\longrightarrow$ Création de sa propre ontologie par le généticien

Contexte : les généticiens

Des experts...

- Formation médicale ou scientifique (Médecine ou doctorat en génétique)
- Différentes sous-spécialités : génétique moléculaire, génétique clinique, cytogénétique,...

...novices

- Pas de formation informatique
- Utilisation fréquente de base de données

Autres contraintes

- Pas de formation possible
- Confiance totale dans le résultat nécessaire

Récapitulatif

2 Contraintes

- Domaine très changeant
- Utilisateurs novices

2 Atouts

- Données semi structurées
- Utilisateurs experts

Sujet

Sujet

Exploration à base ontologique de données issues de patients atteints de maladies génétiques rares.

Notre approche

Peut on permettre à un utilisateur expert de créer sa propre ontologie à partir des données ?

Ontology learning

Processus de création ou d'augmentation d'une ontologie de façon automatique ou semi-automatique

Approches

- Regrouper des ontologies
- Spécialiser une ontologie de haut niveau
- Construire une ontologie depuis le départ

Bibliographie

- Principalement approches machine learning
- Quelques approches participatives à grande échelles

Limites des approches actuelles

- Interfaces faites pour des ingénieurs base de données
  $\longrightarrow$ Réponse habituelle : proximité avec des standards du domaine (SPARQL)
  $\longrightarrow$ Réponse non applicable pour nous
- Dépendance au format : malformation rendant complexe l'extraction d'information de fichiers semi-structurés
  $\longrightarrow$ Aide de l'utilisateur et parsing intelligent
- Duplication d'entité lorsqu'elles sont présentes dans deux sources différentes
  $\longrightarrow$ Nous allons être très concernés
- Évaluation de l'ontologie produite est compliquée
  $\longrightarrow$ Solution classique : revue manuelle par un expert
  $\longrightarrow$ Nous ne sommes pas concernés puisque l'ontologie est construite par l'expert

Hypothèses

Observations

- Les généticiens utilisent plusieurs bases de données
- Ils font des requêtes complexes en faisant plusieurs requêtes simples sur plusieurs bases

Première hypothèse :

- Les généticiens ont un modèle mental du domaine

Seconde hypothèse :

- Il est possible de trouver des correspondances entre leur modèle mental et le modèle réel

Démarche

Test vocabulaire métier

- Valider première et seconde hypothèses
- Produire des données pour la phase suivante

Test prototype

- Construction d'un prototype à partir des résultats précédent
- Évaluation du prototype avec des utilisateurs

Détermination du vocabulaire métier

Objectifs

- Vérifier nos hypothèses
- Déterminer le vocabulaire utilisé, les structures grammaticales,...
- Vérifier la présence d'homonymie ou de synonymie

5 participants :

- un médecin généticien moléculaire
- un médecin clinicien
- une ingénieure en biologie moléculaire
- une médecin cytogénéticienne
- une biologiste moléculaire

Détermination du vocabulaire métier

3 bases de données :

- OMIM : gène, maladie
- HPO : maladie, phénotype, gène
- dbSNP : variation, gène

Aperçu des données :

```
#Format: diseaseId<tab>gene-symbol<tab>gene-id(entrez)<tab>HPO-ID<tab>HPO-term-name
OMIM:608980     FREM1   158326  HP:0000414      Bulbous nose
OMIM:608980     FREM1   158326  HP:0000077      Abnormality of the kidney
OMIM:608980     FREM1   158326  HP:0000322      Short philtrum
OMIM:608980     FREM1   158326  HP:0011803      Bifid nose
OMIM:608980     FREM1   158326  HP:0000143      Rectovaginal fistula
OMIM:608980     FREM1   158326  HP:0001545      Anteriorly placed anus
OMIM:608980     FREM1   158326  HP:0000007      Autosomal recessive inheritance
OMIM:611528     JUP     3728    HP:0004756      Ventricular tachycardia
OMIM:611528     JUP     3728    HP:0000006      Autosomal dominant inheritance
```

Résultats

- Les concepts sont assez facilement isolés
- Si les concepts sont bien isolés, les attributs sont bien répartis
- Le type des attributs n'est pas spontanément précisé
- Beaucoup de synonymes sont employés – jusqu'à 7 pour une personne
- Plusieurs homonymie, mais jamais chez un seul participant
- Les cardinalités peuvent être spontanément précisé via des modaux

Remarques

- Importance des exemples
- Attribut piège : identifiant OMIM
- Expertise réellement indispensable

Hypothèses

Conclusions du test

- Il existe un modèle mental pour les experts
- Il y a des correspondances avec le modèle réel

Prochaine étape

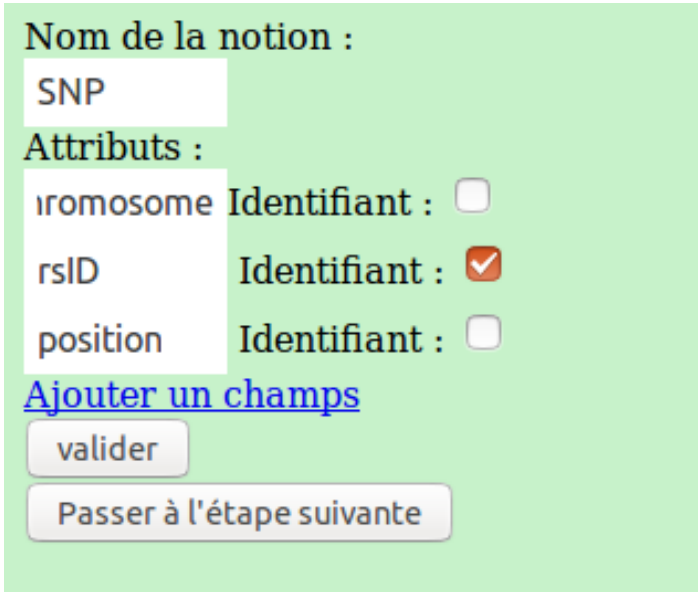
Peut on, à partir du modèle mental des experts, obtenir les informations permettant de bâtir l'ontologie ?

Mise en place d'une stratégie IHM
$\longrightarrow$ Développement d'un prototype pour validation

Développement du prototype

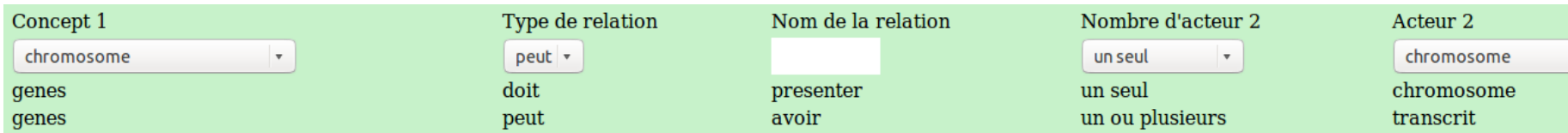
4 étapes :

- Phase de parsing des données : import de fichiers semi structurés, parsing et affichage
- Extraction des concept : à partir des données affichées, extraction des concepts et des attributs



Développement du prototype

- Définition des relations : création des relations entre les concepts, avec leurs cardinalités



- Visionnage des résultats : récapitulatifs des concepts et relations saisis

Test du prototype

Objectif :
Évaluer la faisabilité de l'approche

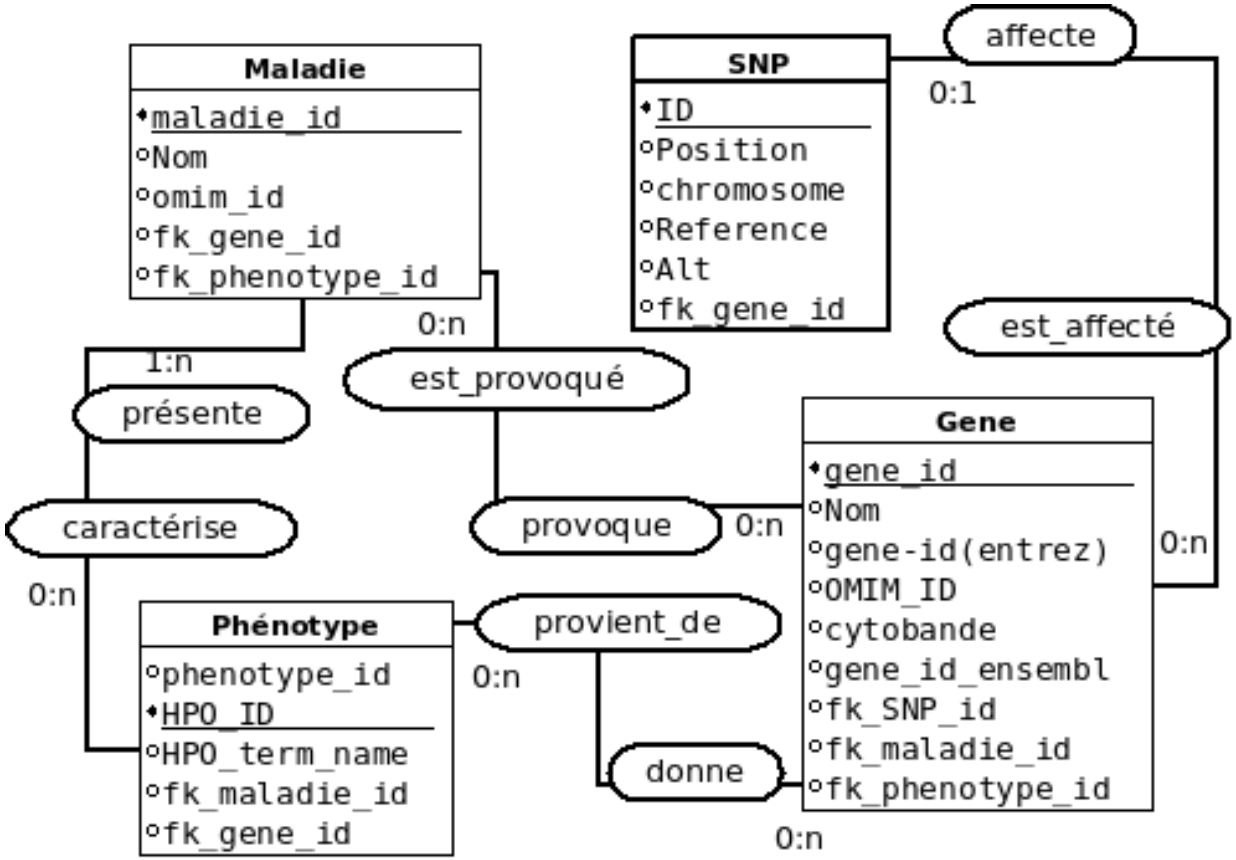
Scénario

Le participant doit rassembler trois bases de données biologiques

4 Participants :

- un médecin généticien moléculaire
- un médecin clinicien
- une ingénieure en biologie moléculaire
- nouveau participant : interne en génétique

Test du prototype



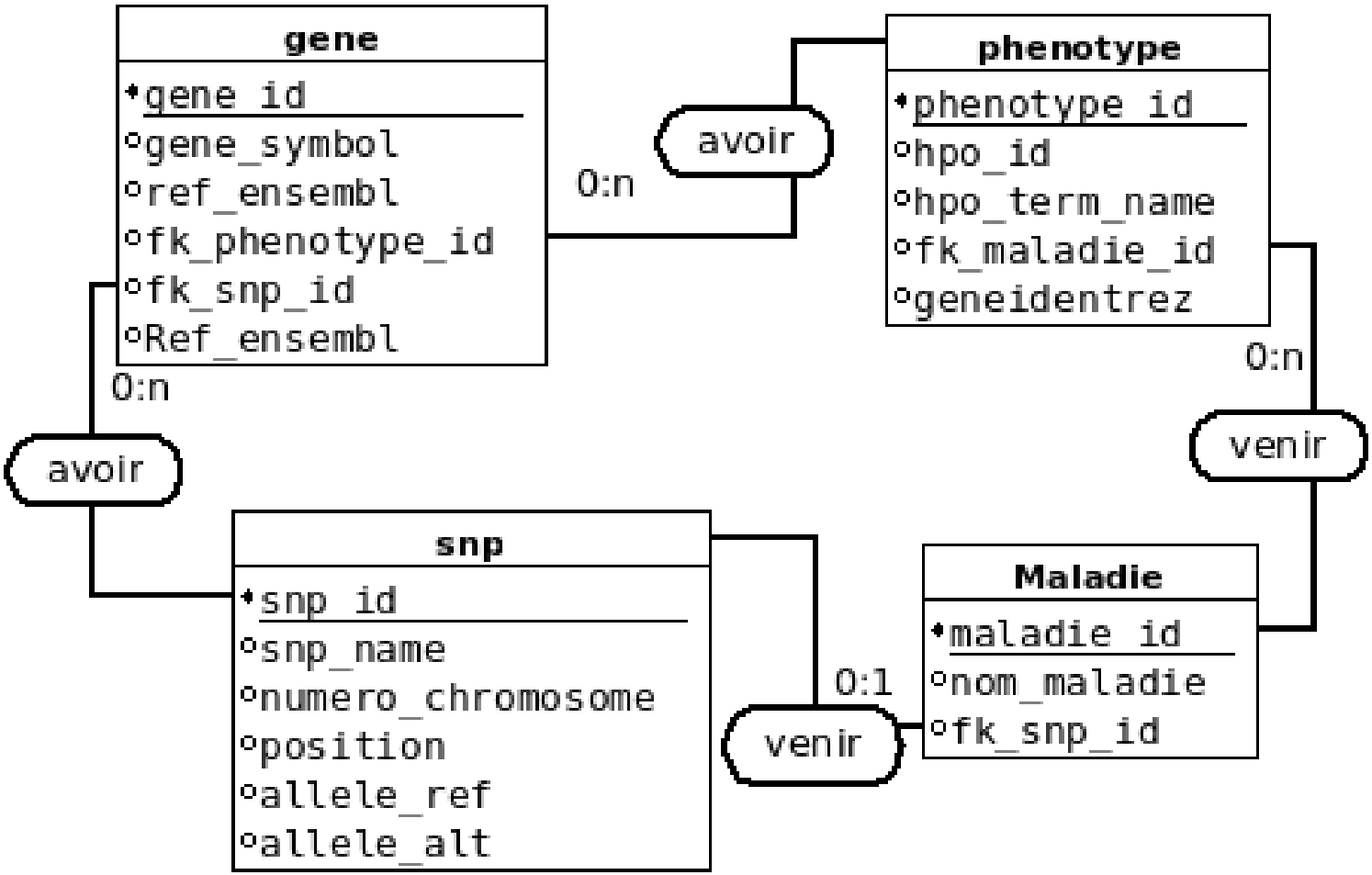
Évaluation

Comparaison des schémas entités associations obtenus avec l'objectif

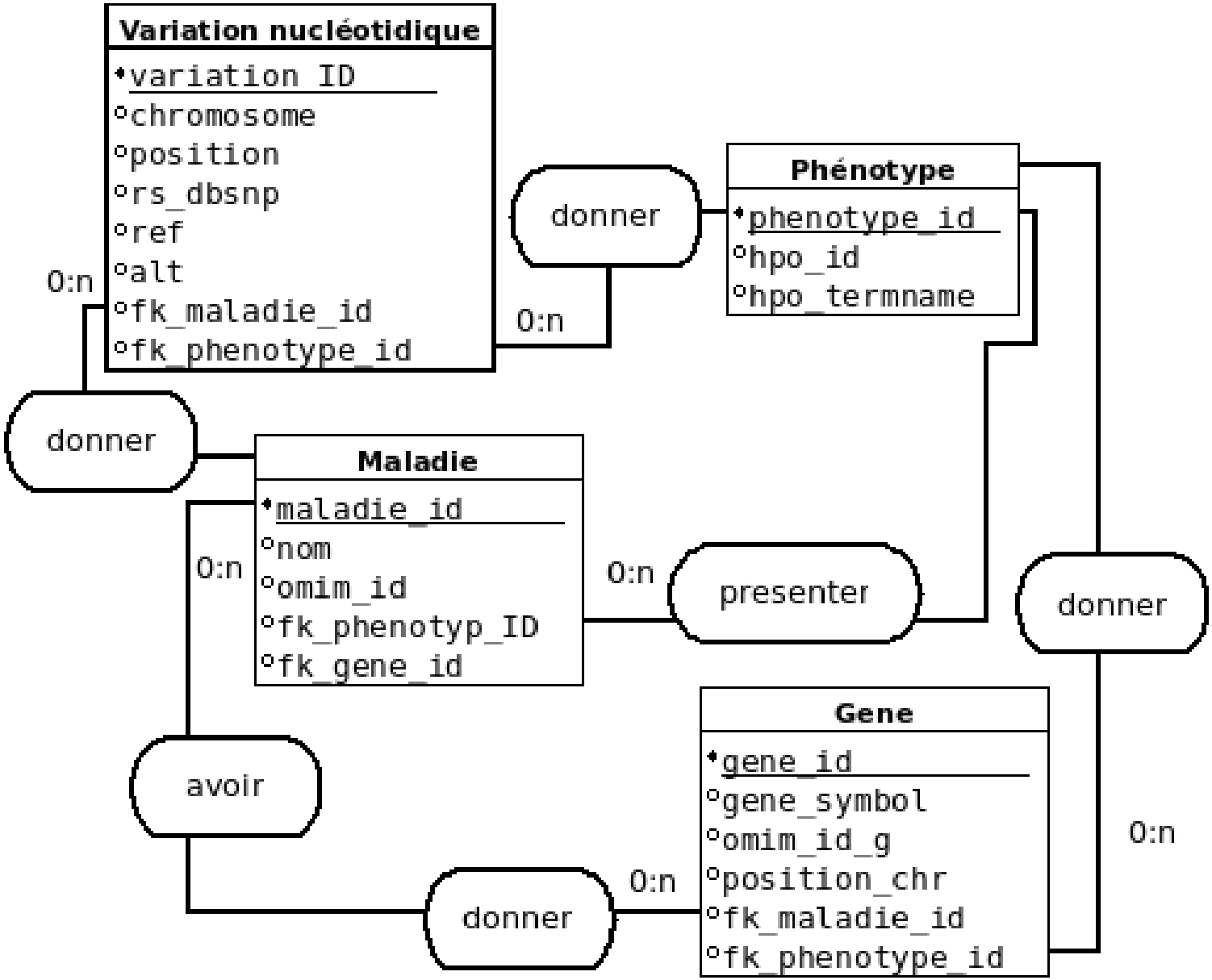
Observation des participants : points de blocages, réflexions, temps

Question aux participants

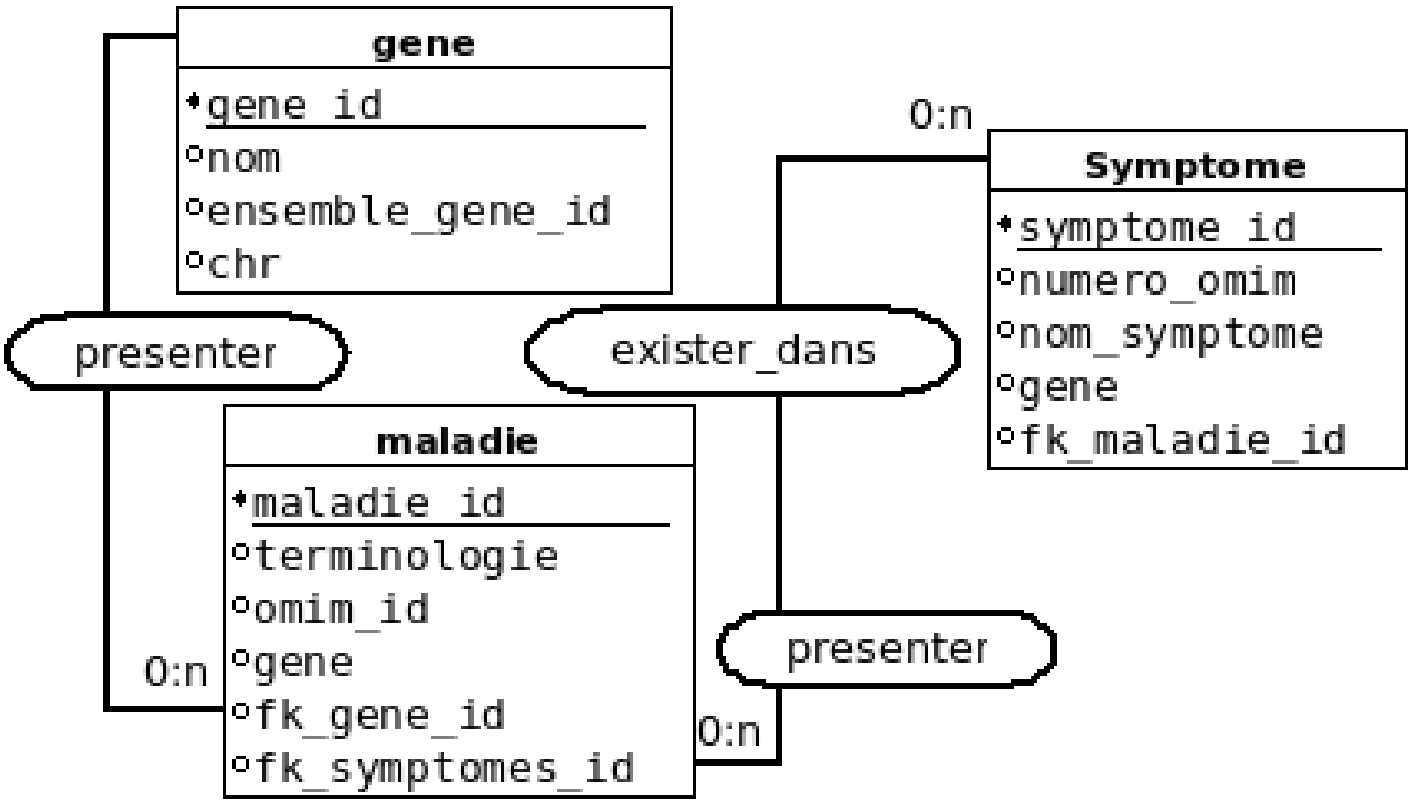
Résultats


gene
gene_id
gene_symbol
ref_ensembl
fk_phenotype_id
fk_snp_id
Ref_ensembl
0:n
avoir
0:n
phenotype
phenotype_id
hpo_id
hpo_term_name
fk_maladie_id
geneidentrez
0:n
venir
avoir
snp
snp_id
snp_name
numero_chromosome
position
allele_ref
allele_alt
0:1
venir
Maladie
maladie_id
nom_maladie
fk_snp_id

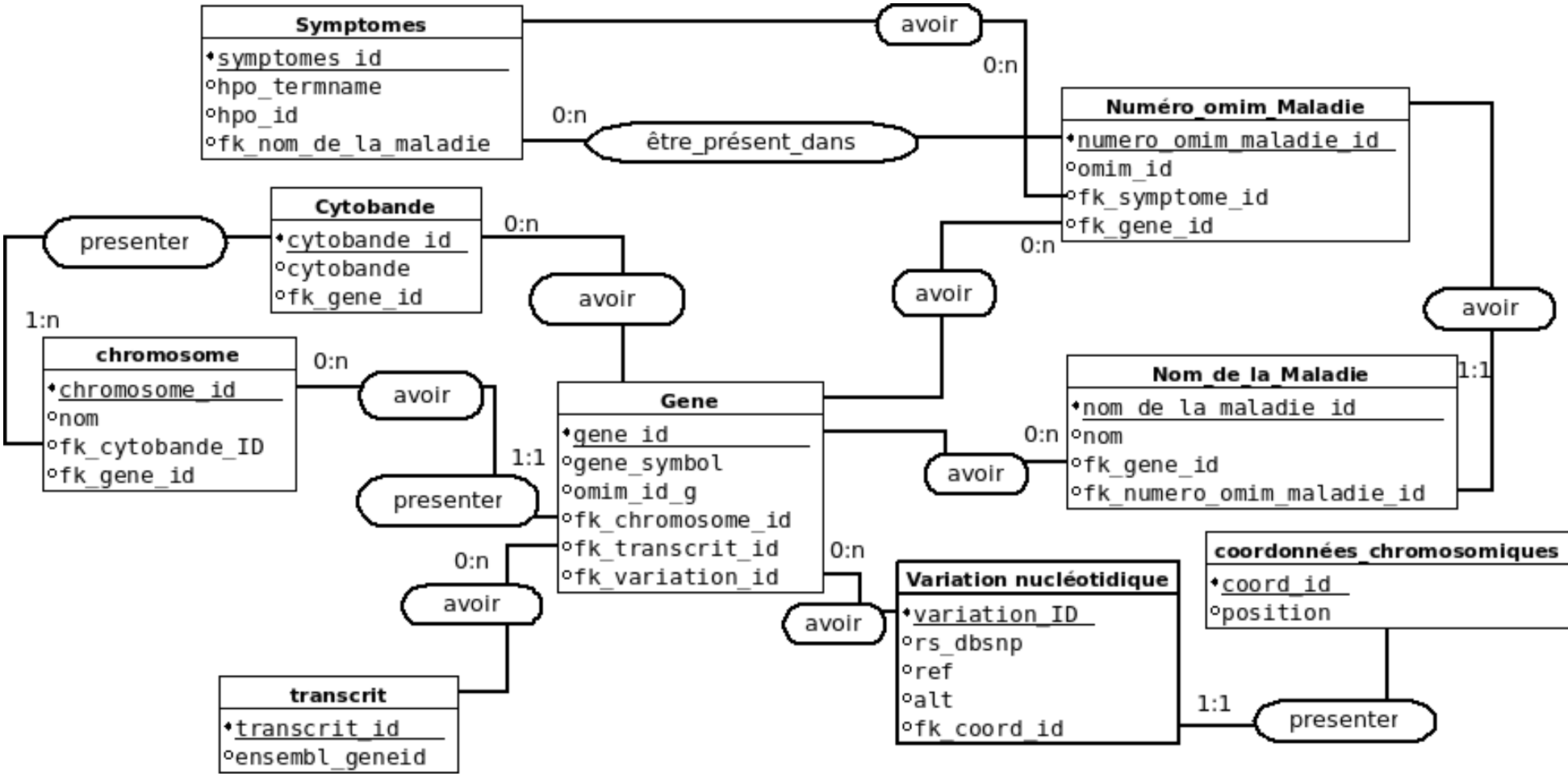
Résultats


Variation nucléotidique
variation_ID
chromosome
position
rs_dbsnp
ref
alt
fk_maladie_id
fk_phenotype_id
donner
0:n
0:n
Phénotype
phenotype_id
hpo_id
hpo_termname
donner
0:n
Maladie
maladie_id
nom
omim_id
fk_phenotyp_ID
fk_gene_id
0:n
presenter
0:n
donner
avoir
donner
0:n
Gene
gene_id
gene_symbol
omim_id_g
position_chr
fk_maladie_id
fk_phenotype_id
0:n

Résultats


gene
gene_id
nom
ensemble_gene_id
chr
Symptome
symptome_id
numero_omim
nom_symptome
gene
fk_maladie_id
maladie
maladie_id
terminologie
omim_id
gene
fk_gene_id
fk_symptomes_id
presenter
exister_dans
presenter
0:n
0:n
0:n
0:n

Résultats


Symptomes
symptomes_id
hpo_termname
hpo_id
fk_nom_de_la_maladie
avoir
0:n
0:n
être_présent_dans
Numéro_omim_Maladie
numero_omim_maladie_id
omim_id
fk_symptome_id
fk_gene_id
Cytobande
cytobande_id
cytobande
fk_gene_id
presenter
0:n
avoir
0:n
avoir
avoir
1:n
chromosome
chromosome_id
nom
fk_cytobande_ID
fk_gene_id
0:n
avoir
Gene
gene_id
gene_symbol
omim_id_g
fk_chromosome_id
fk_transcrit_id
fk_variation_id
Nom_de_la_Maladie
nom_de_la_maladie_id
nom
fk_gene_id
fk_numero_omim_maladie_id
1:1
0:n
avoir
1:1
presenter
0:n
avoir
0:n
avoir
Variation nucléotidique
variation_ID
rs_dbsnp
ref
alt
fk_coord_id
coordonnées_chromosomiques
coord_id
position
1:1
presenter
transcrit
transcrit_id
ensembl_geneid

Bilan résultats

Concepts

- 4 sur 4 pour 2 participants
- 1 manque
- une explosion

Relations

- des oublis...
- ...et des relations supplémentaires
- pas de concepts oubliés
- cardinalité parfois élargie

Bilan résultats

Remarques :

- Tous les systèmes sont implémentables
- Plusieurs synonymes pour les concepts et les attributs
- Cas du "faux" concept
- Réponses aux questions :

Question	F.	G.	S.	X.	Total sur 20
plausibilité du scénario	5	5	3	5	18
simplicité des tâches	3	3	4	3	13
clarté de l'interface	5	3	2	3	13

Situations particulières

- Relation 1 : 1
  $\longrightarrow$ Détection et suggestion à l'utilisateur
- "Faux" concept
  $\longrightarrow$ Impossible à détecter, mais résolution par l'utilisateur
- Oubli de concept
  $\longrightarrow$ Vérifier l'attribution de chaque colonne, peut être volontaire
- Oubli de relation
  $\longrightarrow$ Concept isolé, présence de plusieurs concepts dans un seul fichier
- Erreur de cardinalité
  $\longrightarrow$ vérifiable pour certain cas (0 :1), pas pour d'autres (0 :n)

Conclusion générale

Conclusions

- Il existe un modèle mental du domaine chez les généticiens
- Ce modèle a des similitudes avec le modèle réel
- Il est possible de reconstruire une ontologie avec l'aide de l'expert

Prochaines étapes

- Implémentation des heuristiques d'aides
- Résolution des conflits sémantiques
- Système de requête adapté à l'utilisateur