DESFEUX Nicolas
TEXIER Aurélien

TP4 : Analyse de Concepts Formelle

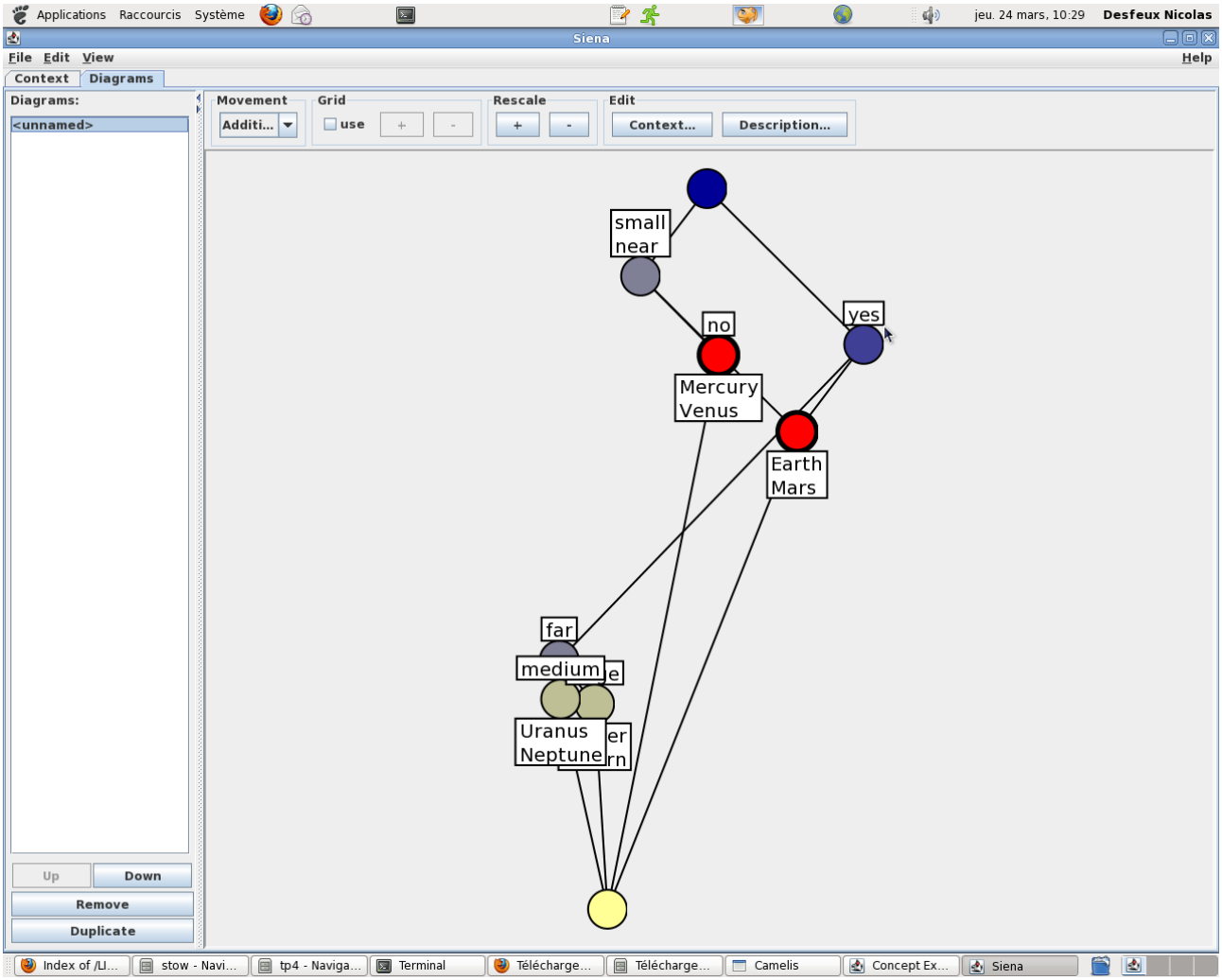
1 Présentation des outils

L'objectif de cette partie de nous présenter des outils d'analyse de concepts formelle. Parmi ces outils, on retrouve ToscanaJ, Conexp, Camelis. On nous décrit leur fonctionnement, et l'intérêt que nous aurons à les utiliser.

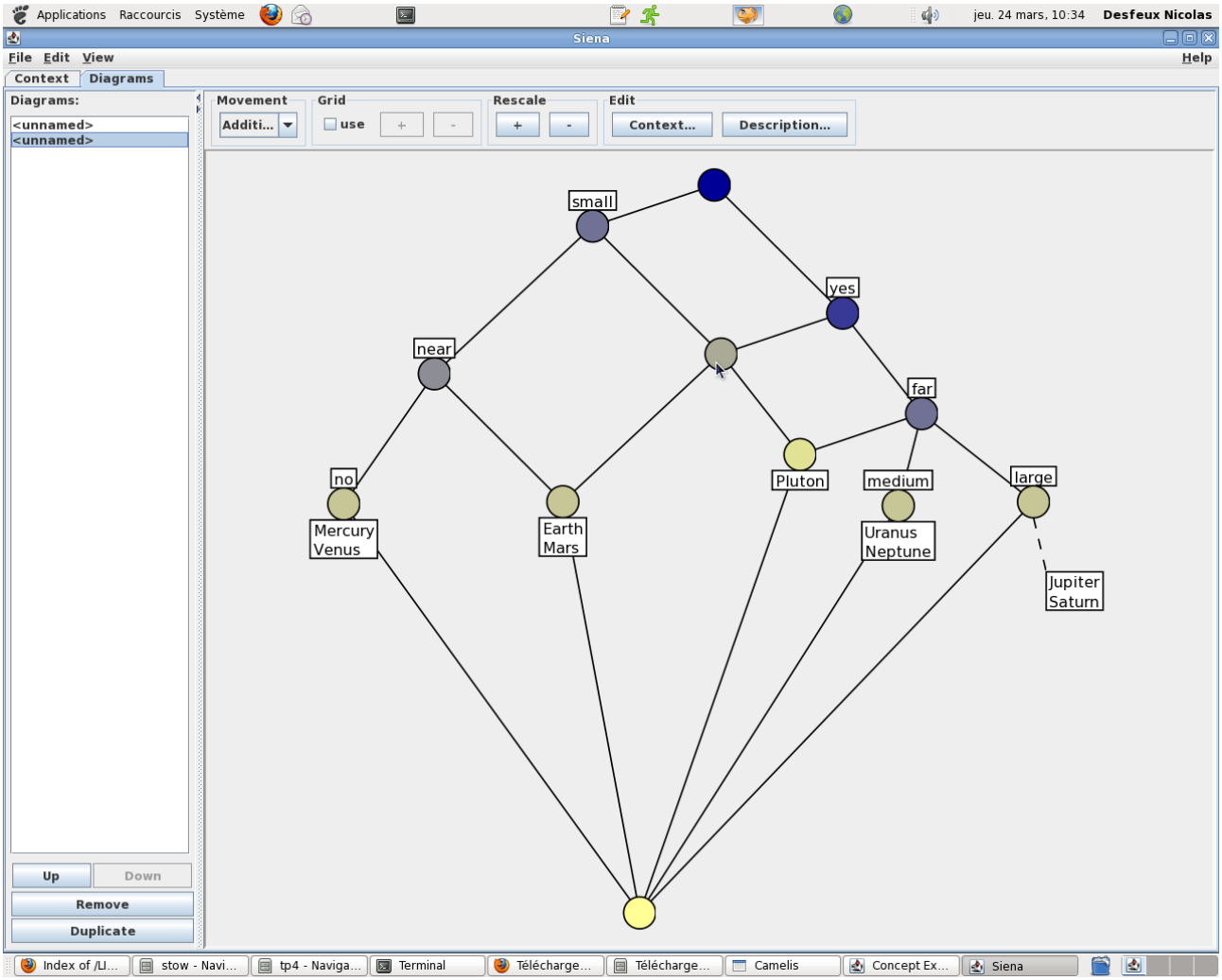
2 Le système solaire avec Conexp

2.1 Question 1

Voici le treillis généré :



2.2 Question 2



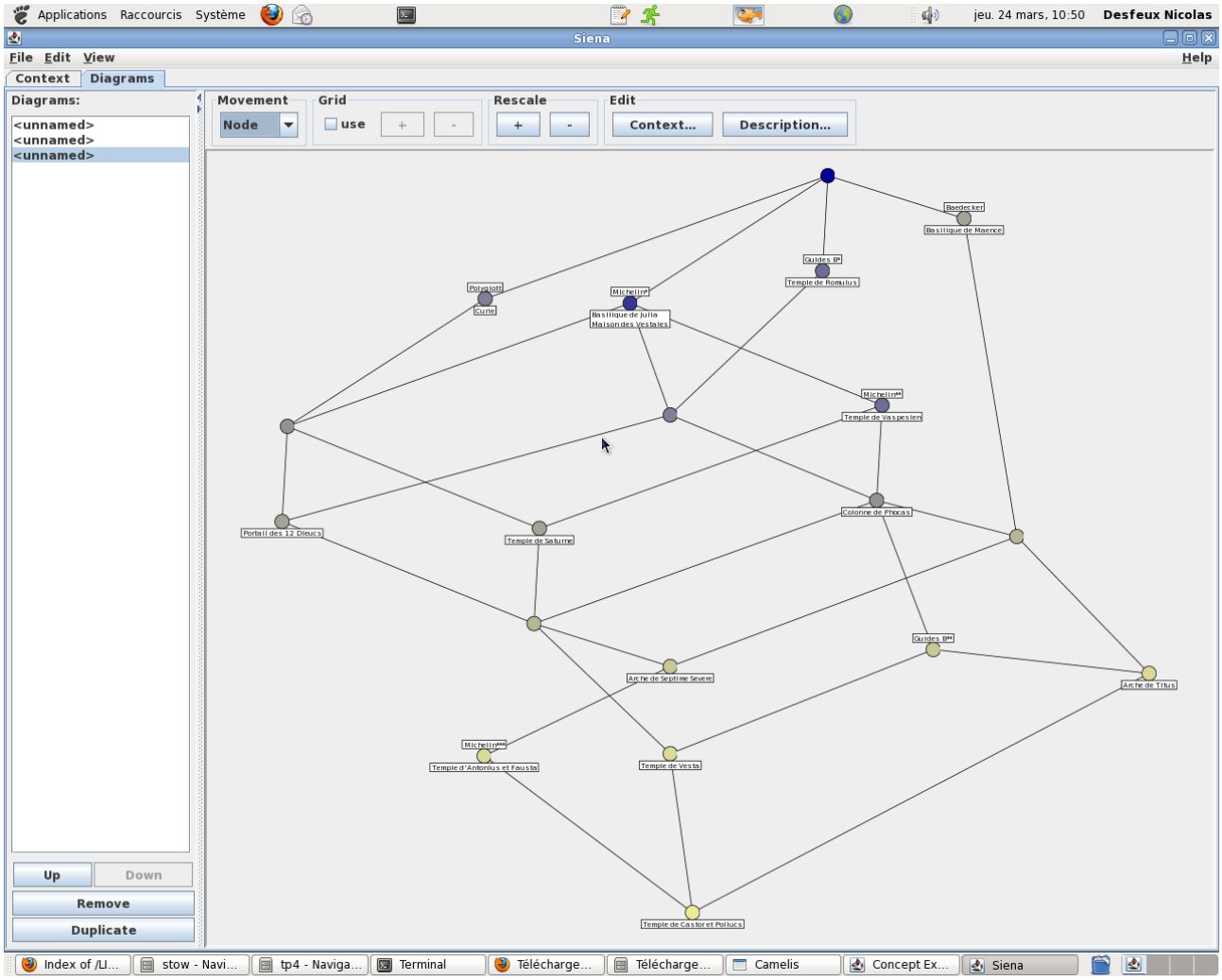
Pluton n'est plus considéré comme une planète de notre système solaire. Par rapport au treillis précédent, il y a deux nœud de plus, dont un avec Pluton seul dedans. Ce nœud est relié à « far », à la bottom clause, et au nœud représentant « petite planète avec satellite ». C'est ce nœud qui n'existe pas dans le treillis précédent. L'introduction de Pluton entraîne la création de nœud pour le moins singulier. Pluton est très isolé dans le treillis, c'est un argument pour considérer que Pluton n'est pas une planète du système solaire.

3 Vestiges Romains avec ToscanaJ

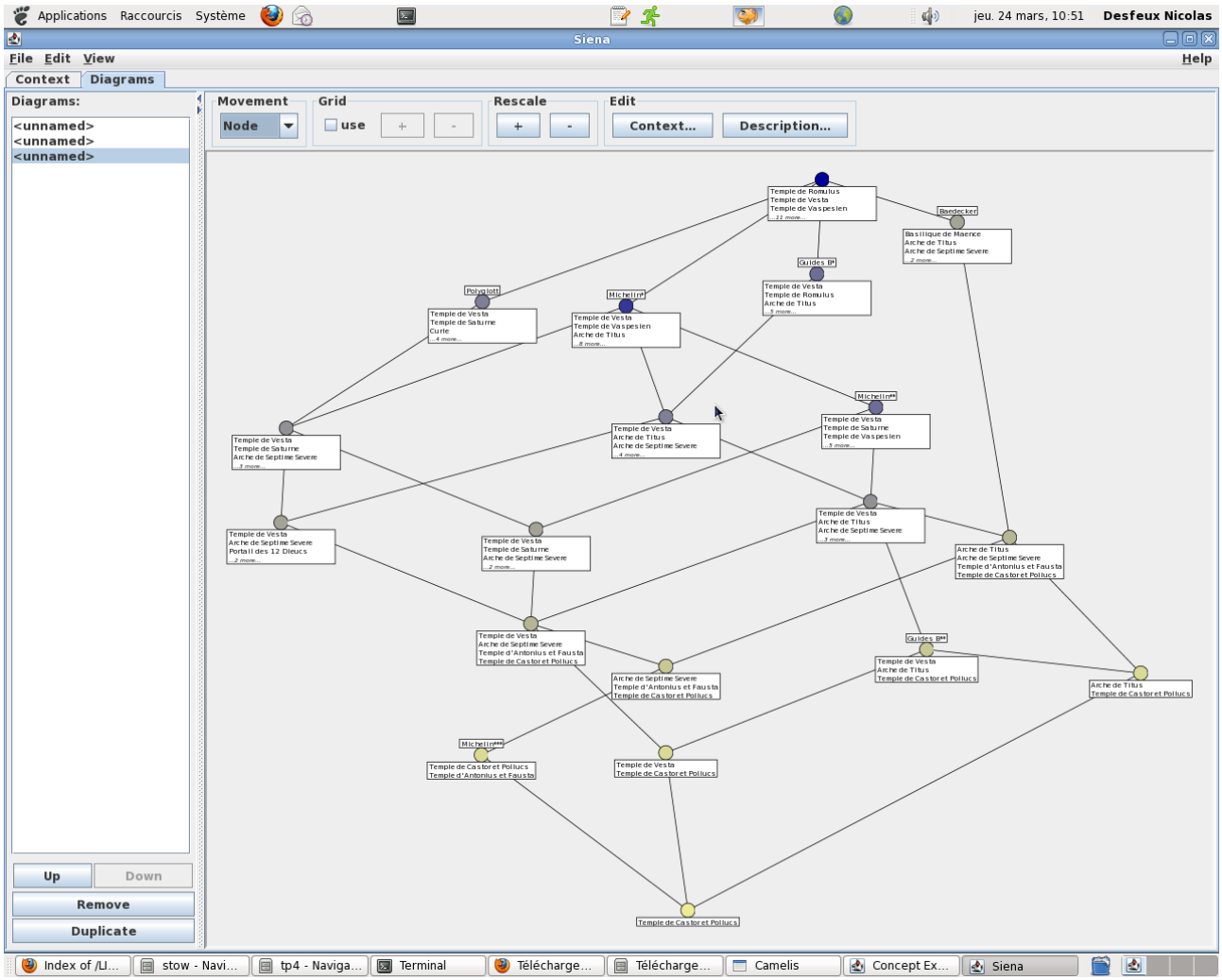
3.1 Question 4

Voici le treillis associé aux données sur les vestiges romains :

Avec l'option « Show only exact matches » :



Avec l'option « Show all matches » :

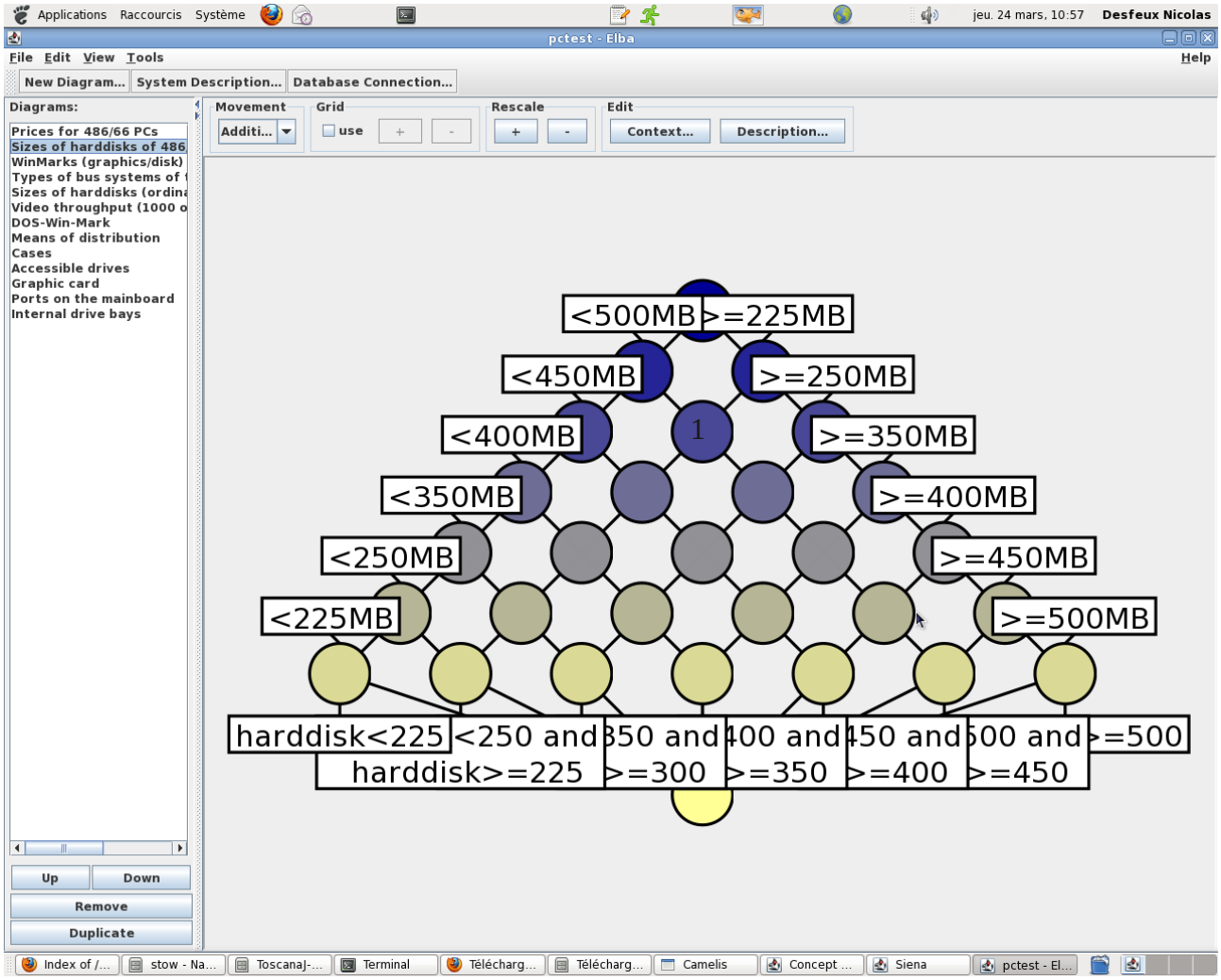


3.2 Question 5

L'option « show only exact matches » ne montre que les objets correspondant aux noeuds en question. « Show all matches » montre tous les descendants en plus des objets correspondant au noeud. C'est une extension du concept.

4 Connexion à une base de données relationnelle sous ToscanaJ

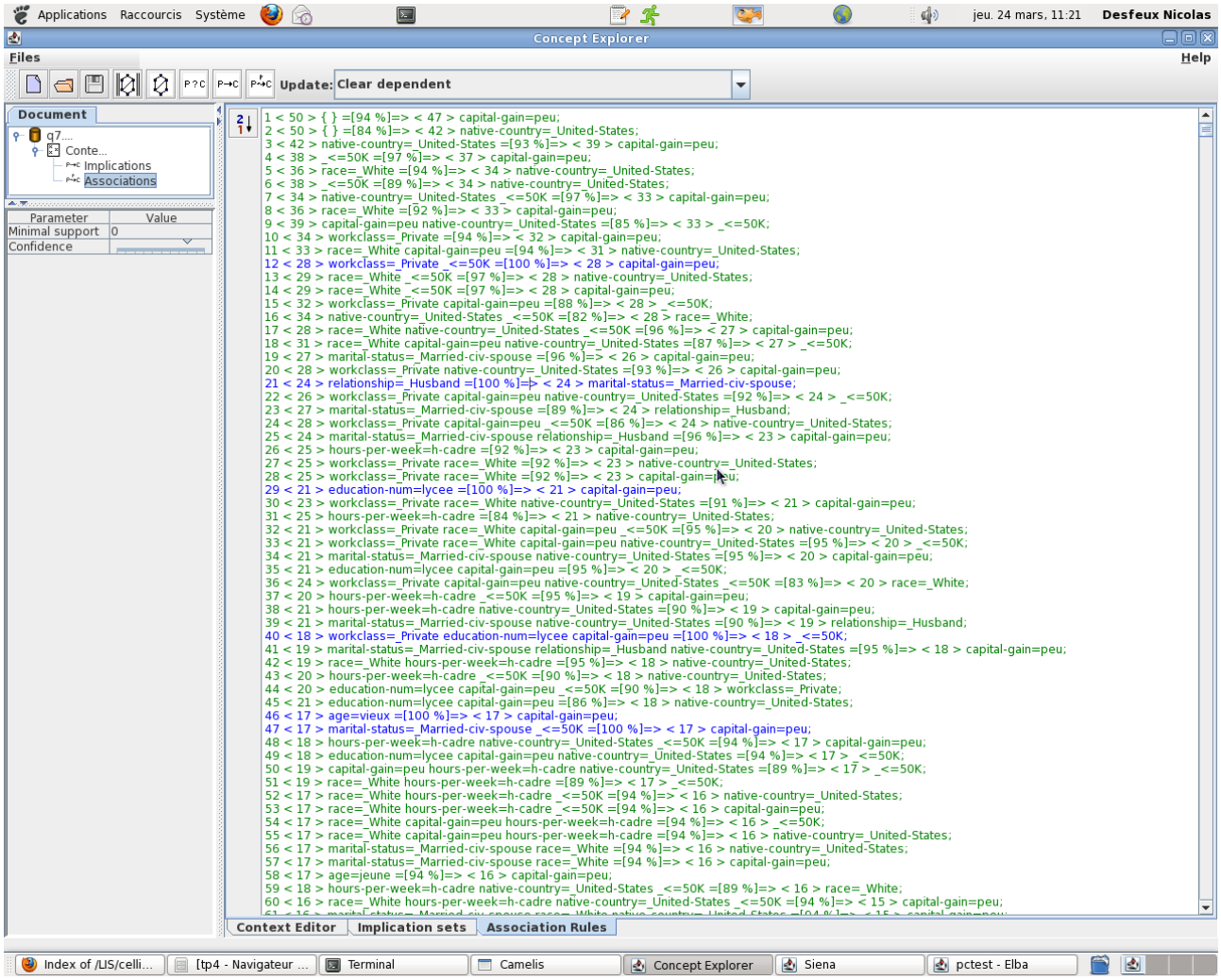
4.1 Question 6



En regardant l'échelle conceptuelle de la taille des disques durs, on peut en déduire que l'attribut qui a été discrétisé est de type continu. Il s'agit de la taille des disques durs. On pose donc comme question à l'exemple : est-ce que la taille du disque dur est inférieur ou supérieur égal à X ? où X représente une capacité. Tout les nœuds représente toutes les combinaisons possibles des classes discrètes. Par exemple, pour le nœud « 1 », il contient les disques durs ayant une capacité inférieure à 450MB et supérieure ou égale à 250MB.

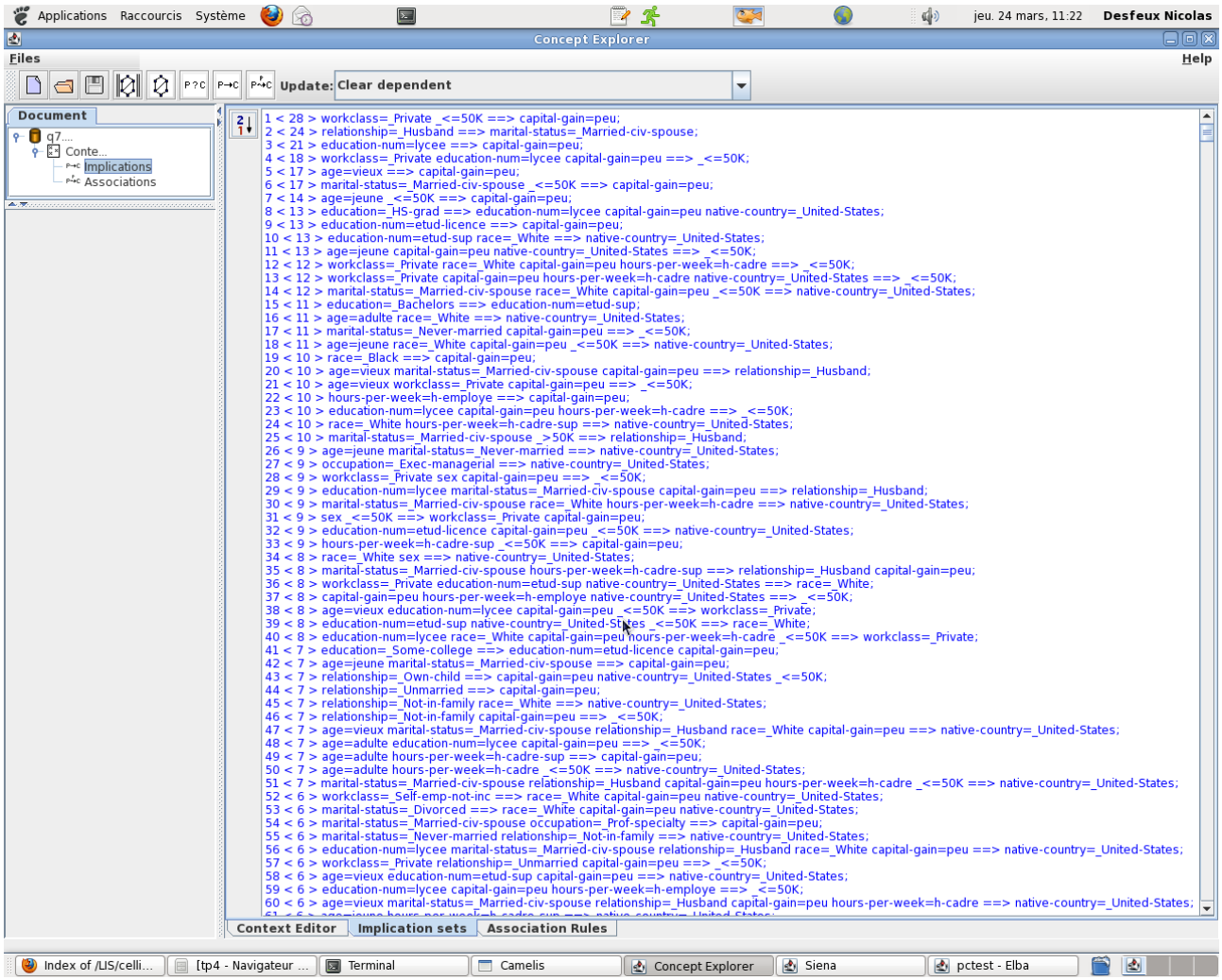
5 Recensement américain : Implications-Règles d'association

5.1 Question 7



La base d'implication (les règles en bleu)est un sous-ensemble des règles d'associations (les règles vertes ci-dessus). Les règles d'association sont des faits sur à 100%. Le pourcentage des autres règles est inférieur à 100%.

Voici un aperçu de la base d'implication :



5.2 Question 8

Le treillis formé à partir de ces données est très dur à lire et à interpréter. Comme dans la réalité, on ne peut analyser de gros tableaux directement, sans découpage, c'est la même chose ici pour les treillis. On ne peut pas les analyser quand ils sont trop grand. Cependant, on peut s'en servir comme structures de données et de ce fait naviguer dans un treillis de nœud en nœud. Pour utiliser cette structure de donnée, on va donc pouvoir utiliser l'analyse des concepts.

6 Base de comics gérée avec Camelis

6.1 Chargement de données personnelles

6.1.1 Question 9

On se trouve dans le treillis, on peut faire des requêtes pour naviguer dedans. Les requêtes sont faites soient avec des clics, soit textuellement, soit grâce à des propriétés extrinsèques.

6.1.2 Question 10

On voit :

- Nombre d'objets : 627,
- le dessinateur Franck Miller a 11 comics,
- Il y a 40 comics de genre « aventure »,
- Il y a 72 comics avec 4/5 en note, avec 1 avis.

6.2 Création de propriétés extrinsèque

6.3 Question 11

On ajoute des propriétés dites « extrinsèques ». Par exemple, nous avons ajouté une entité personnage et on glisse des comics dedans. On décide en fait de ce que contient cette entité.

6.4 Apprentissage et Navigation dans les données

6.4.1 Apprentissage

6.4.1.1 Question 12

Pour obtenir la réponse à ces questions, on peut utiliser différentes techniques : soit on utilise des réquêtes, soit grâce aux propriétés, ou alors on peut utiliser une simple navigation dans les données.

Pour trouver qui dessine et scénarise les comics les plus appréciés, nous avons utilisé une réquête sélectionnant les notations supérieur à 3.5, et ensuite nous avons navigué dans les données afin de trouver les scénaristes et les dessinateurs suivant : Bryan Hitch, Kurt Busiek,...

Le fonctionnement est le même pour les dessinateurs et scénaristes des comics les moins appréciés : Dave Gibbons, Franck Miller,...

Pour connaître le nombre de séries par éditeur, nous avons choisi d'utiliser la navigation. Pour chaque éditeur, on a ainsi accès au nombre de série qu'il possède.

Le dessinateur ayant le plus de comics dans cette sélection est Richard Corben.

6.4.2 Navigation

6.4.2.1 Question 13

Pour faire cette sélection, nous avaons écrit une requete :

'Nom série' contains "Batman" and not Editeur is "Panini Comics" and not Editeur is "Semic"

Ainsi, on obtient la liste correct.

'Blondie' n'a pas de genre. Voici la requête que nous avons utilisé :

« not Genre ? »

Nous avons de nouveau créé une requête pour obtenir la liste de comics. Voici la requête utilisée :

('Nom série' contains "blanche" or 'Nom série' contains "noir" or 'Nom série' contains "vie " or 'Nom série' contains "mort") and not 'Nom série' contains "une" and not 'Nom série' contains "la"

Voici la liste :
- A l'ombre des tours mortes
- Mort@17
- Parlez-moi de Mort
- Veuve Noire

6.4.3 Comparaison avec les systèmes existants

6.4.3.1 Question 14

L'intérêt de ce genre de système réside dans son efficacité. Le fonctionnement en arborescence permet de faire des recherches rapide, et facilite grandement la navigation.

Par contre, mettre en place ce genre de système est assez compliqué, puisque cela nécessite d'utiliser un système de création « assisté » de requêtes. Cela nécessite aussi un gros travail pour entrer les données dans l'application, puisqu'il faut classifier chaque exemples.

L'idée est d'associer à ces systèmes un système de classification automatique. Par exemple, si l'on regarde le système de gestion de photos Picasa, il est aujourd'hui capable de classer les photos par date, mais également capable de reconnaître les gens sur les photos. On a donc automatiquement deux types de filtres : sur la date et sur les personnes présentes ou absentes de la photo !

Par rapport à une base de données classique, il y a un gain de vitesse, mais il est indispensable de connaître tout les caractéristiques des objets que l'on veut classer.