



**ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ  
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ**

**ΜΕΤΑΠΤΥΧΙΑΚΟ ΔΙΠΛΩΜΑ ΕΙΔΙΚΕΥΣΗΣ (MSc)  
στα ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ**

**ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ**

**Απεικόνιση προτύπων  
σε τεχνολογίες Βάσεων Δεδομένων**

**Κοτσιφάκος Ευάγγελος**

**M3020014**

**ΑΘΗΝΑ, ΦΕΒΡΟΥΑΡΙΟΣ 2004**

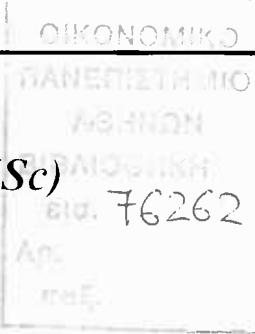




ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΚΑΤΑΛΟΓΟΥ

**ΜΕΤΑΠΤΥΧΙΑΚΟ ΔΙΠΛΩΜΑ ΕΙΔΙΚΕΥΣΗΣ (MSc)  
στα ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ**



**ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ**

**Απεικόνιση προτύπων  
σε τεχνολογίες Βάσεων Δεδομένων**

**Κοτσιφάκος Ευάγγελος**

**M3020014**



**Επιβλέπων Καθηγητής:  
κ. Μιχάλης Βαζιργιάννης**

**Εξωτερικοί Κριτές:  
κοι Γιάννης Θεοδωρίδης, Βασίλειος Βασσάλος**

**ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ  
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ**

**ΑΘΗΝΑ, ΦΕΒΡΟΥΑΡΙΟΣ 2004**





Στήν Κέλλυ



# ΑΠΕΙΚΟΝΙΣΗ ΠΡΟΤΥΠΩΝ ΣΕ ΤΕΧΝΟΛΟΓΙΕΣ ΒΑΣΕΩΝ ΔΕΔΟΜΕΝΩΝ

## Περίληψη

Η ανάγκη για την εξαγωγή χρήσιμων συμπερασμάτων από τις μεγάλες συλλογές δεδομένων διαφόρων ειδών και προέλευσης έχει δώσει ώθηση στην ανάπτυξη συστημάτων και τεχνικών Εξόρυξης Γνώσης (data mining). Οι τεχνικές αυτές εφαρμόζονται στον όγκο των δεδομένων προκειμένου να εξαχθούν κάποια πρότυπα που τα χαρακτηρίζουν. Η αποτελεσματική εξαγωγή και εκμετάλλευση των προτύπων αποτελεί σήμερα πρόβλημα πολλών επιστημονικών πεδίων και η ανάγκη για ένα σύστημα που θα μπορεί αποτελεσματικά και αποδοτικά να διαχειριστεί τα διάφορα πρότυπα γίνεται ολοένα και μεγαλύτερη. Μέχρι σήμερα έχουν γίνει προσπάθειες για την υλοποίηση τέτοιων συστημάτων που κυρίως στηρίζονται στις ήδη επιτυχημένες και αποδεκτές σχεσιακές και αντικειμενο-σχεσιακές βάσεις δεδομένων. Στην εργασία αυτή εξετάζονται οι ιδιαίτεροτήτες που διέπουν τα πρότυπα διαφόρων ειδών, το λογικό μοντέλο και το λειτουργικό πλαίσιο που μπορούν να περιγράψουν σφαιρικά τη μορφή των προτύπων και τις επιθυμητές λειτουργίες πάνω σε αυτά και περιγράφεται ένα Σύστημα Διαχείρισης Βάσεων Προτύπων. Επιπλέον, μια υποτυπώδης βάση προτύπων υλοποιείται χρησιμοποιώντας την ημι-δομημένη προσέγγιση της XML και συγκρίνεται η καταλληλότητα της προσέγγισης αυτής σε σχέση με υλοποιήσεις σε σχεσιακό και αντικειμενο-σχεσιακό μοντέλο. Τα αποτελέσματα δείχνουν ότι εξαιτίας της φύσης των προτύπων η ημι-δομημένη προσέγγιση είναι καταλληλότερη από τις άλλες δύο για την υλοποίηση μιας βάσης προτύπων. Βασικό χαρακτηριστικό της προσέγγισης αυτής και προϋπόθεση για την επιτυχία της αποτελεί η δημιουργία των κατάλληλων XML σχημάτων που θα περιγράφουν τους διάφορους τύπους προτύπων. Τα σχήματα αυτά πρέπει να ενσωματώνουν τα ιδιαίτερα χαρακτηριστικά της δομής των προτύπων, πρέπει όμως ταυτόχρονα να είναι αφαιρετικά και αρκετά ελεύθερα για να μπορέσουν να αναπαραστήσουν όλα τα πρότυπα του συγκεκριμένου τύπου. Η καταλληλότητα της XML να χρησιμοποιηθεί για τη δημιουργία μιας βάσης προτύπων αποτελεί το πρώτο βήμα. Η ικανότητά της να απαντήσει αποτελεσματικά σε θέματα ομοιότητας προτύπων και διαχείρισης μεγάλου όγκου προτύπων θα κρίνει αν υπάρχει η ανάγκη για δημιουργία ενός νέου συστήματος με καινούργια γλώσσα αναπαράστασης, καινούργιες τεχνικές φυσικής αποθήκευσης, ευρετηριοποίησης και οπτικοποίησης προσανατολισμένο αποκλειστικά στα πρότυπα.

# PATTERN REPRESENTATION IN DATABASE TECHNOLOGIES

## Abstract

The need for the extraction of useful knowledge from large collections of data collected from a number of sources has led to a great development of the Data mining systems and techniques. These techniques are applied to the datasets in order to extract patterns that characterize them. A lot of different scientific fields are dealing with this problem and the need for an integrated system that could manage effectively and efficiently the patterns is emerging. Until now, systems like that have been developed, usually based on the truly efficient and commonly adapted relational and object-relational database models. In this project the special characteristics of patterns are presented, along with a logical model for the development of a Pattern Base Management System. A rudimentary pattern base has been developed using the semi-structured approach and XML and its capability of efficiently describe the pattern base is compared to relational and object-relational approaches. Basic issue in the development of an XML pattern base is the design of the proper schemas that describe the different pattern types. Those schemas should be general enough to represent every possible pattern of that type but also should exploit the pattern's special features in order to achieve better storage and retrieval performance. Ensuring the capability of XML to efficiently represent every pattern type is the first step. The capability of XML to also deal with issues of pattern similarity and manipulation of large collection of patterns will point the need for the design and implementation from scratch of a new system with special query and manipulation language, storing, imaging and visualization techniques that will be based on the characteristics of patterns.

# ΠΕΡΙΕΧΟΜΕΝΑ

ΠΕΡΙΕΧΟΜΕΝΑ .....	3
1. ΕΙΣΑΓΩΓΗ.....	5
1.1 Εξόρυξη Γνώσης και Πρότυπα .....	5
1.2 Πρότυπο.....	7
1.3 Πρότυπα διαφόρων ειδών.....	7
1.4 Συστήματα Διαχείρισης Βάσεων Δεδομένων και Συστήματα Διαχείρισης Βάσεων Προτύπων.....	10
1.5 Σύστημα Διαχείρισης Βάσεων Προτύπων (ΣΔΒΠ).....	11
1.6 Ορισμός του προβλήματος – Στόχος της έρευνας .....	15
2. ΜΟΝΤΕΛΟΠΟΙΗΣΗ ΣΥΣΤΗΜΑΤΟΣ ΔΙΑΧΕΙΡΙΣΗΣ ΒΑΣΕΩΝ ΠΡΟΤΥΠΩΝ.....	17
2.1 Το λογικό μοντέλο ενός ΣΔΒΠ.....	17
2.1.1 Τύποι Προτύπων (pattern types) .....	18
2.1.2 Πρότυπα (patterns) .....	19
2.1.3 Κλάσεις (classes).....	19
2.2 Σχέσεις μεταξύ προτύπων.....	20
2.2.1 Εξειδίκευση (specialization) .....	21
2.2.2 Σύνθεση και εκλέπτυνση (Composition and refinement).....	21
2.3 Ομοιότητα προτύπων .....	21
3. ΥΛΟΠΟΙΗΣΗ XML ΒΑΣΗΣ ΠΡΟΤΥΠΩΝ .....	23
3.1 XML (Extensible Markup Language) .....	23
3.1.1 Καλώς-ορισμένα και ορθά XML έγγραφα. DTDs και XMLSchema .....	25
3.2 Βάση προτύπων και XML. Σχήματα για την περιγραφή των προτύπων... .....	28
3.3 Υλοποίηση XML μοντέλου σε ORACLE 9i .....	38
3.3.1 Τρόποι αποθήκευσης XML δεδομένων στην ORACLE 9i .....	39
3.3.2 Δοκιμαστικά δεδομένα. Εργαλεία που χρησιμοποιήθηκαν.....	42
3.3.3 Ερωτήματα (queries) στην XML βάση προτύπων.....	43
3.4 Κριτική – συμπεράσματα εξέτασης συστήματος διαχείρισης προτύπων σε XML .....	61
4. Εναλλακτικές υλοποίησεις .....	63

<b>4.1</b>	<b>Υλοποίηση Σχεσιακής Βάσης Προτύπων .....</b>	<b>63</b>
4.1.1	Ερωτήματα (queries) στη σχεσιακή βάση προτύπων.....	66
4.1.2	Κριτική – συμπεράσματα εξέτασης συστήματος διαχείρισης προτύπων σε σχεσιακό μοντέλο.....	76
<b>4.2</b>	<b>Υλοποίηση Αντικειμενο-σχεσιακής βάσης προτύπων.....</b>	<b>76</b>
4.2.1	Ερωτήματα (queries) στην αντικειμενο-σχεσιακή βάση προτύπων .....	77
4.2.2	Κριτική – συμπεράσματα εξέτασης συστήματος διαχείρισης προτύπων σε αντικειμενο-σχεσιακό μοντέλο .....	82
<b>4.3</b>	<b>Άλλες προσεγγίσεις.....</b>	<b>82</b>
<b>5</b>	<b>ΣΥΓΚΡΙΤΙΚΗ ΜΕΛΕΤΗ ΥΛΟΠΟΙΗΣΕΩΝ.....</b>	<b>85</b>
5.1	Κριτήρια σύγκρισης .....	85
5.2	Σύγκριση υλοποιήσεων .....	85
<b>6.</b>	<b>ΣΥΜΠΕΡΑΣΜΑΤΑ - ΣΥΝΕΙΣΦΟΡΑ.....</b>	<b>89</b>
<b>7.</b>	<b>ΑΝΟΙΚΤΗ ΕΡΕΥΝΑ.....</b>	<b>90</b>
<b>ΠΑΡΑΡΤΗΜΑ Α .....</b>		<b>91</b>
<b>ΠΑΡΑΡΤΗΜΑ Β. Πίνακας εικόνων .....</b>		<b>93</b>
<b>ΠΑΡΑΡΤΗΜΑ Γ. Πίνακας πινάκων .....</b>		<b>95</b>
<b>ΑΝΑΦΟΡΕΣ .....</b>		<b>96</b>
<b>Ευχαριστίες .....</b>		<b>98</b>

# 1. ΕΙΣΑΓΩΓΗ

Η δυνατότητα συλλογής διαφορετικών τύπων δεδομένων από διάφορες πηγές (εικόνες από δορυφόρους, εγγραφές πελατών και αγορών, χρηματιστηριακές συναλλαγές, δεδομένα από αισθητήρες κλπ) έχει οδηγήσει στην ανάγκη για καταχώρηση και αξιοποίηση ενός τεραστίου όγκου δεδομένων. Ένα πρόβλημα σε πολλούς επιστημονικούς αλλά και εμπορικούς τομείς είναι το πώς θα γίνει η διαχείριση του τεράστιου αυτού όγκου δεδομένων που έχει συλλεχθεί έτσι ώστε να μπορούν να εξαχθούν χρήσιμα συμπεράσματα από την ανάλυσή τους. Η ολοκλήρωση και μεταφορά των δεδομένων από διάφορες πηγές αποτελούν επίσης σημαντικά ζητήματα ιδιαίτερα με τη ανάγκη της διαμοίρασης της πληροφορίας (information sharing) και τη χρήση κατανευμημένων εφαρμογών. Ο μεγάλος όγκος των αποθηκευμένων δεδομένων είναι σημαντικό εμπόδιο για την ολοκλήρωση και μεταφορά τους.

Με τη χρήση της Εξόρυξης Γνώσης (Data Mining) μπορούμε να εξάγουμε ενδιαφέροντα συμπεράσματα για τα αποθηκευμένα δεδομένα. Ουσιαστικά το ενδιαφέρον βρίσκεται στην πληροφορία που πηγάζει από τη διαδικασία της εξόρυξης γνώσης και η διαχείριση αυτής της πληροφορίας γίνεται πλέον ο στόχος των Συστημάτων Βάσεων Δεδομένων και Εξόρυξης Γνώσης.

Οι υπάρχουσες τεχνολογίες Βάσεων Δεδομένων είναι πλέον ώριμες και παρέχουν εργαλεία για την αποτελεσματική διαχείριση Βάσεων Δεδομένων οποιασδήποτε εφαρμογής και διαφορετικού όγκου. Επιπλέον ενσωματώνουν τεχνικές Εξόρυξης Γνώσης και παρέχουν τη δυνατότητα εκμετάλλευσης του μεγάλου όγκου των δεδομένων για τη στήριξη αποφάσεων.

## 1.1 Εξόρυξη Γνώσης και Πρότυπα

Οι κυριότερες τεχνικές Εξόρυξης Γνώσης είναι η συσταδοποίηση, η κατηγοριοποίηση-ταξινόμηση (classification), τα δέντρα απόφασης και οι κανόνες συσχέτισης. Τα αποτελέσματα των τεχνικών αυτών είναι διάφορα **πρότυπα** (classification rules, decision trees, clusters κλπ) που αντιστοιχίζουν ή κατηγοριοποιούν τα δεδομένα σε διάφορες ομάδες με βάση κάποιες ιδιότητές τους.

Συνοπτικά αναφέρουμε [1]:

### *Clustering - Συσταδοποίηση*

Με τη διαδικασία της συσταδοποίησης τα δεδομένα χωρίζονται σε ομάδες (clusters) έτσι ώστε σε κάθε ομάδα να βρίσκονται σημεία δεδομένων που είναι πιο πολύ όμοια μεταξύ τους βάση κάποιων κριτηρίων ομοιότητας. Η διαδικασία της συσταδοποίησης μπορεί να έχει διαφορετικά αποτελέσματα ως προς το διαχωρισμό των δεδομένων σε ομάδες ανάλογα με τον αλγόριθμο συσταδοποίησης και τις παραμέτρους εισόδου. Η ποιότητα του κάθε διαφορετικού διαχωρισμού (της συσταδοποίησης) μπορεί να ελεγχθεί με κάποια μέτρα ποιότητας για να επιλεγεί ο καλύτερος.

### *Classification – Κατηγοριοποίηση-ταξινόμηση*

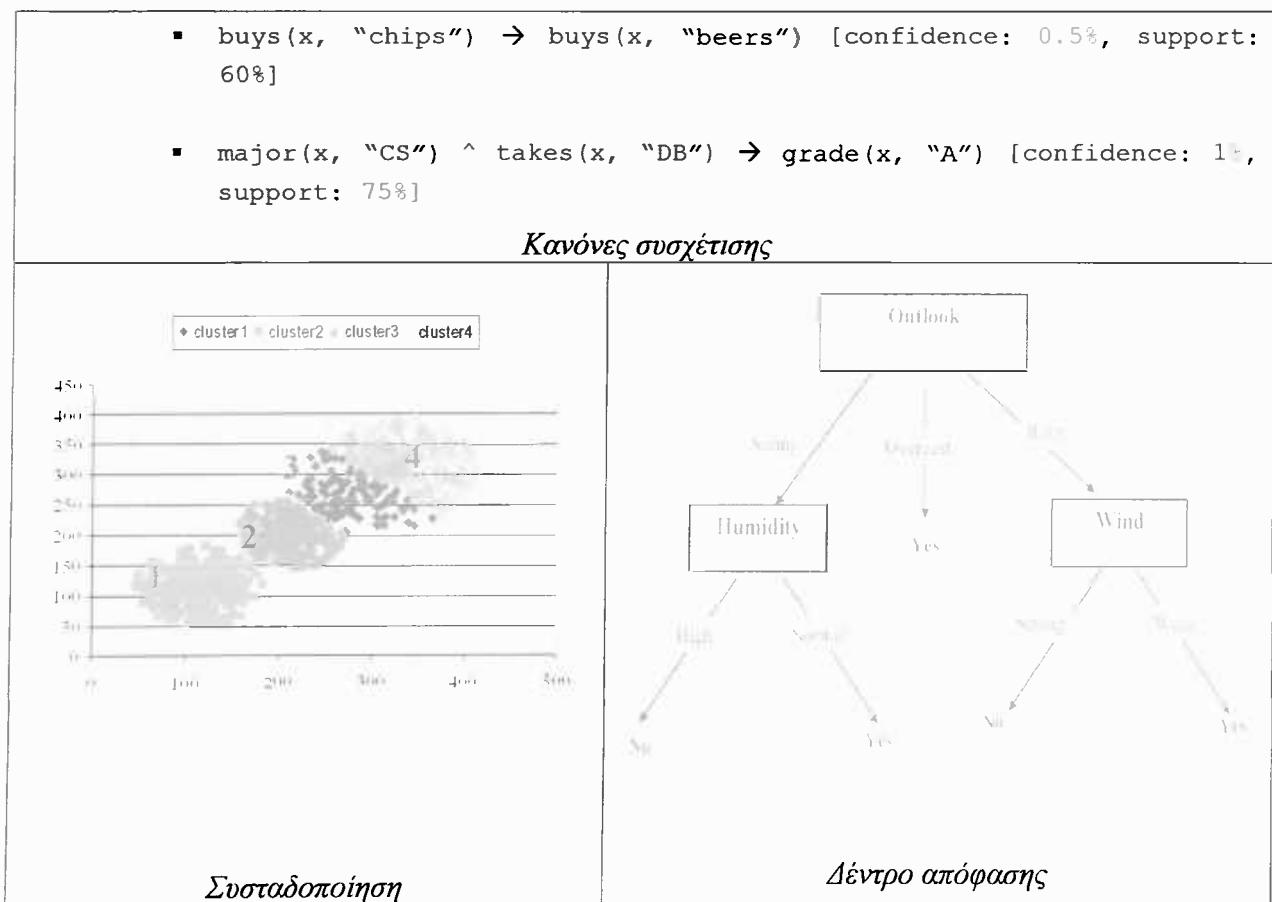
Το πρόβλημα της κατηγοριοποίησης έχει ερευνηθεί αναλυτικά κυρίως στους τομείς της στατιστικής, της αναγνώρισης προτύπων και του machine learning. Ένας μεγάλος αριθμός από τεχνικές κατηγοριοποίησης έχει αναπτυχθεί και κάποιες από αυτές είναι: Bayesian classification, Νευρωνικά Δίκτυα και Δέντρα Αποφάσεων. Η κατηγοριοποίηση είναι μια φόρμα ανάλυσης των δεδομένων που μπορεί να παράγει μοντέλα περιγραφής αυτών ή να προβλέψει τις τάσεις τους. Χρησιμοποιείται για τη δημιουργία ευφυών βάσεων αποφάσεων στον επιχειρηματικό και επιστημονικό χώρο.

### Association Rules – Κανόνες Συσχέτισης

Οι κανόνες συσχέτισης αποκαλύπτουν συσχετίσεις μεταξύ των γνωρισμάτων ενός συνόλου δεδομένων. Οι κανόνες αυτοί μπορούν να παρασταθούν με τη μορφή  $A \rightarrow B$ , όπου τα A, B αναφέρονται σε τιμές γνωρισμάτων των υπό εξέταση δεδομένων. Συγκεκριμένα τα A και B επιλέγονται να είναι συχνά εμφανιζόμενες τιμές. Το βασικό χαρακτηριστικό είναι ότι εγγραφές που περιέχουν την A τιμή περιέχουν επίσης και την B. Με βάση κάποια μέτρα ποιότητας (support, confidence, leverage, lift), οι κανόνες μπορούν να ελεγχθούν και να περιοριστούν. Η πιο τυπική εφαρμογή των κανόνων συσχέτισης είναι η ανάλυση καλαθιού καταναλωτή με την οποία μπορούμε να δούμε τις συνήθειες των καταναλωτών και το ποια προϊόντα σχετίζονται με ποια κατά τις προτιμήσεις τους. Με αναλύσεις αυτού του τύπου οι πωλητές μπορούν να προσαρμόσουν κατάλληλα τις στρατηγικές προώθησης των προϊόντων τους

- buys(x, "chips") → buys(x, "beers") [confidence: 0.5%, support: 60%]
- major(x, "CS") ^ takes(x, "DB") → grade(x, "A") [confidence: 1%, support: 75%]

### Κανόνες συσχέτισης



Πίνακας 1 Τα τρία βασικά πρότυπα που εξάγονται με τεχνικές εξόρυξης γνώσης

Η περαιτέρω ανάλυση των τεχνικών Εξόρυξης Γνώσης ξεφεύγει από τα πλαίσια αυτής της εργασίας. Περισσότερα στοιχεία μπορούν να βρεθούν στις αναφορές [16, 17, 18].

## 1.2 Πρότυπο

Πολλοί ορισμοί έχουν προταθεί από διάφορους ερευνητές για το τι είναι πρότυπο καθότι συναντάται σε πολλούς επιστημονικούς χώρους. Ο όρος πρότυπο στην καθημερινή μας γλώσσα είναι πολλές φορές συνώνυμο της κανονικότητας, της επανάληψης ενός σχήματος (όχι απαραίτητα γεωμετρικού). Χρησιμοποιούμε τον όρο πρότυπο σε συνδυασμό με το χρόνο, το χώρο, τον ήχο και τον τρόπο που δομούμε τη σκέψη μας.

Υπάρχουν δύο αντίθετες απόψεις σχετικά με το πρότυπο. Σύμφωνα με την πρώτη, όταν δεν είναι δυνατή η αναγνώριση μιας κανονικότητας, ενός προτύπου τότε συμπεραίνεται ότι δεν υπάρχει πρότυπο. Σύμφωνα με τη δεύτερη άποψη, όταν δεν υπάρχει κανονικότητα τότε αναφερόμαστε σε τυχαίο πρότυπο. Στους περισσότερους ορισμούς για τα πρότυπα αναφέρεται ότι το πρότυπο πρέπει να ορίζεται ανεξάρτητα από κάποια κλίμακα αναφοράς. Ο ορισμός θα πρέπει να είναι θεωρητικός και να συγκεκριμενοποιείται μόνο ανάλογα με την εφαρμογή που εξετάζουμε. Έτσι, θα μπορούσαμε να καταλήξουμε ότι πρότυπο είναι ένας γενικός όρος για κάθε αναγνωρίσιμη κανονικότητα στα δεδομένα.

Ειδικά στην επιστήμη της πληροφορικής, αναφέρουμε έναν ορισμό που ταιριάζει καλύτερα στις Βάσεις Δεδομένων και στην εξόρυξη γνώσης.

### Ορισμός:

«Το πρότυπο (pattern) αντιπροσωπεύει και περιγράφει ένα μεγάλο όγκο δεδομένων με ένα αποτελεσματικό τρόπο, είναι δηλαδή μια συνοπτική, πλούσια σε σημασιολογία αναπαράσταση ενός μέρους των αρχικών δεδομένων» [1].

Ο ορισμός αυτός όμως μπορεί να περιγράψει και πρότυπα που παρουσιάζονται σε άλλα επιστημονικά πεδία. Στη συνέχεια αναφέρονται κάποια είδη προτύπων από διάφορους επιστημονικούς χώρους προκειμένου να διαπιστωθούν οι ομοιότητες με τα πρότυπα που παράγονται από τις τεχνικές εξόρυξης γνώσης.

## 1.3 Πρότυπα διαφόρων ειδών

### *Ακολουθιακά πρότυπα – Ανάλυση χρονολογικών σειρών*

Η εξόρυξη ακολουθιακών προτύπων είναι η εξόρυξη συχνά εμφανιζόμενων προτύπων που σχετίζονται με το χρόνο ή άλλες ακολουθίες. Οι περισσότερες έρευνες συγκεντρώνονται στα συμβολικά πρότυπα. Το πρόβλημα παρουσιάζεται ως εξής:

Δοσμένης μιας μεγάλης σε μέγεθος συμβολοσειράς, ενδιαφερόμαστε για ακολουθίες-πρότυπα της μορφής  $\alpha \rightarrow \beta$ , όπου  $\alpha, \beta$  είναι υπό-συμβολοσειρές της αρχικής, τέτοια ώστε η συχνότητα εμφάνισης του  $\alpha\beta$  και η πιθανότητα το  $\alpha$  να ακολουθείται από το  $\beta$  να μην είναι χαμηλότερη από ένα προκαθορισμένο όριο. Τέτοιου είδους δεδομένα στην καθημερινή ζωή άλλα και στην επιστήμη συναντώνται συχνά (πχ. κείμενο, παρτιτούρες μουσικής, καιρικά δεδομένα, δορυφορικά δεδομένα, επιχειρηματικές συναλλαγές, εγγραφές τηλεπικοινωνιών, ακολουθίες DNA, ιστορικά ιατρικά αρχεία κ.α.). Η ανακάλυψη επαναλαμβανόμενων ή όμοιων προτύπων βοηθά τον χρήστη ή τον επιστήμονα στην πρόβλεψη κάποιων φαινομένων [1].

### **Επεξεργασία σήματος – Ανάκτηση μουσικού περιεχομένου**

Η μεγάλης κλίμακας αποθήκευση ήχου και μουσικής έγινε εφικτή τις τελευταίες δεκαετίες. Επιπλέον η δυνατότητα κατανομής του πολυμεσικού περιεχομένου στο διαδίκτυο έθεσε νέες απαιτήσεις για μεγάλες και ευέλικτες Βάσεις Δεδομένων Πολυμέσων. Μια από τις βασικές απαιτήσεις είναι η πολύπλοκη αναζήτηση περιεχομένου σε μια μεγάλη βάση με μουσικά δεδομένα. Ερωτήσεις σε μουσικά δεδομένα απευθύνονται σε πληροφορίες που βρίσκονται κρυμμένες στο ίδιο το περιεχόμενο, στο ηχητικό σήμα. Τα πρότυπα και οι εφαρμογές τους στην περύπτωση του μουσικού περιεχομένου μπορεί να είναι διάφορα: πρότυπα στη μουσική δομή, στον εντοπισμό μελωδίας, ρυθμικών προτύπων, αναγνώριση κλίμακας, αναγνώριση συνθέτη, οργάνων κ.α [1].

### **Πρότυπα στην ανάκτηση πληροφοριών (Information Retrieval)**

Η χρήση προτύπων στον τομέα αυτόν είναι προφανής. Σε μια μεγάλη συλλογή προφορικών δεδομένων (που ονομάζεται *corpus*) οι χρήστες κάνουν επερωτήσεις με βάση το ενδιαφέρον τους. Οι ερωτήσεις είναι συνήθως δύσκολο να οριστούν, αντίθετα με τα παραδοσιακά συστήματα Βάσεων Δεδομένων, επειδή υπάρχει έλλειψη αποδοτικής γλώσσας επερωτήσεων. Η πολυσημία των όρων προσθέτει μεγάλη δυσκολία στην ανακάλυψη προτύπων σε τέτοιου είδους εφαρμογές [1].

### **Μαθηματικά**

Τα μαθηματικά είναι η κατ' εξοχήν επιστήμη των προτύπων. Υπάρχουν πρότυπα πολύ οικεία σε όλους μας και σίγουρα αυτά μπορεί να είναι ακολουθιακά, χωρικά, χρονικά ακόμα και γλωσσικά. Τα μαθηματικά πρότυπα μπορούν να ομαδοποιηθούν σε επτά κατηγορίες.

- i. Αριθμητικά πρότυπα (πρώτοι αριθμοί, τετράγωνα αριθμών, αριθμοί Fibonacci κ.α.)
- ii. Πρότυπα γραφικών παραστάσεων
- iii. Πρότυπα σε σχήματα (ομοιότητες στον αριθμό γωνιών, ακμών κλπ.)
- iv. Πρότυπα σε άπειρες ακολουθίες (σειρές Taylor, ακολουθίες Fibonacci, πρώτοι αριθμοί κλπ.)
- v. Πρότυπα αλγεβρικής μορφής
- vi. Γεωμετρικά πρότυπα

vii. Πρότυπα στην κρυπτογραφία (κρυπτογραφικά συστήματα)

*Οπτικοποίηση*

Ανάλυση εικόνων από διάφορες πηγές για εύρεση προτύπων και απεικόνιση αυτών [2].

*Αστρονομία*

Οι βάσεις αστρονομικών δεδομένων περιέχουν τεράστιο όγκο πληροφοριών που όμως είναι δύσκολο να αξιοποιηθούν. Οι Επιστήμονες αναζητούν πρότυπα μέσα στις τεράστιες αυτές βάσεις προσπαθώντας να τις ενοποιήσουν και να εφαρμόσουν αλγόριθμους κατηγοριοποίησης και συσταδοποίησης. Ουσιαστικά πρόκειται για χωροχρονικές βάσεις αφού τα ενδιαφέροντα πρότυπα είναι αυτά οιμάδων αστεριών ή άλλων ουράνιων σωμάτων που μοιάζουν σε κάποιο χαρακτηριστικό τους. Η επεξεργασία των δεδομένων περιλαμβάνει και την επεξεργασία εικόνων υψηλής ανάλυσης που έχουν ληφθεί από τηλεσκόπια και δορυφόρους. Ο χώρος της αστρονομίας δείχνει ιδιαίτερο ενδιαφέρον για τα πρότυπα και για ένα σύστημα εξαγωγής αλλά και διαχείρισης αυτών.

Στον παρακάτω πίνακα συνοψίζονται διάφορα είδη προτύπων και οι εφαρμογές που αυτά παρουσιάζονται [3]:

Εφαρμογή	Αρχικά Δεδομένα	Πρότυπα
Ανάλυση καλαθιού αγορών	Εγγραφές πωλήσεων	Κανόνες Συσχέτισης αντικειμένων
Επεξεργασία σήματος	Σύνθετα σήματα	Επαναλαμβανόμενες κυματομορφές
Παρακολούθηση κινούμενων αντικειμένων	Διανύσματα κίνησης-τροχιάς	Εξισώσεις
Ανάκτηση πληροφοριών	Κείμενα	Συχνότητα εμφάνισης λέξεων
Επεξεργασία εικόνας	Βάση Δεδομένων εικόνας	Χαρακτηριστικά εικόνας
Τμηματοποίηση αγοράς	Profiles χρηστών	Ομάδες χρηστών
Ανάκτηση μουσικών δεδομένων	Παρτιτούρες μουσικής, αρχεία ήχου	Ρυθμικά, μελωδικά, αρμονικά
Παρακολούθηση συστήματος	Αρχεία πληροφοριών εξόδου	Πρότυπα αποτυχίας συστήματος
Οικονομικές εφαρμογές	Εμπορικές εγγραφές	Τάσεις μετοχών
Ανάλυση πληροφοριών “click” σε εφαρμογές web	Αρχεία πρόσβασης server	Ακολουθίες από “click”
Επιδημιολογία	Εγγραφές κλινικών	Συσχέτιση συμπτωμάτων-ασθενειών
Αξιολόγηση κινδύνου	Εγγραφές πελατών	Δέντρα αποφάσεων

Πίνακας 2 Παραδείγματα εφαρμογών και προτύπων

#### 1.4 Συστήματα Διαχείρισης Βάσεων Δεδομένων και Συστήματα Διαχείρισης Βάσεων Προτύπων

Τα Συστήματα Διαχείρισης Βάσεων Δεδομένων (*ΣΔΒΔ*) διαχειρίζονται (αποθηκεύουν, ανακτούν, ταξινομούν και αναπαριστούν) αποτελεσματικά τα δεδομένα που συλλέγονται από διάφορες πηγές. Τα αρχικά αυτά δεδομένα (raw data) είναι για τα *ΣΔΒΔ first-class citizens*, δηλαδή στοιχεία που δεν αναλύονται περαιτέρω και όλες οι πράξεις γίνονται πάνω σε αυτά.

Τα πρότυπα χαρακτηρίζονται από μεγάλη πολυπλοκότητα, πολυμορφία και μεταβλητότητα. Η δομή τους και τα χαρακτηριστικά τους διαφέρουν ριζικά από τα αρχικά δεδομένα. Τα πρότυπα είναι συνοπτικές αναπαραστάσεις ενός μεγάλου όγκου δεδομένων και έχουν μια δομή, αντίθετα με τα αρχικά δεδομένα που είναι (αν δεν επεξεργαστούν) αδόμητα. Επιπλέον τα πρότυπα περιέχουν σημασιολογία, περιέχουν

δηλαδή σημαντική πληροφορία για τη σημασία και τα χαρακτηριστικά των αρχικών δεδομένων που αναπαριστούν.

Λόγω του μεγάλου όγκου των δεδομένων αλλά και των αυξημένων απαιτήσεων των χρηστών, τα ίδια τα πρότυπα γίνονται όλο πιο σύνθετα και πολύπλοκα αλλά και πιο χρήσιμα. Δημιουργείται λοιπόν η ανάγκη τα πρότυπα αυτά να αποθηκεύονται και να διαχειρίζονται με παρόμοιο τρόπο που μέχρι τώρα διαχειρίζονται τα απλά δεδομένα. Μέχρι τώρα δεν υπάρχουν εργαλεία που να εξυπηρετούν ένα τέτοιο σκοπό, δεν υπάρχουν δηλαδή βάσεις που να αποθηκεύουν και να διαχειρίζονται πρότυπα δεδομένων. Οι πιο συνηθισμένες λύσεις περιλαμβάνουν την επέκταση ήδη υπαρχόντων Συστημάτων Διαχείρισης Βάσεων Δεδομένων (DBMS) έτσι ώστε να μπορούν να υποστηρίζουν τη διαχείριση κάποιων προτύπων που έχουν προκύψει από διαδικασίες Εξόρυξης Γνώσης [1].

Τέτοιες προσεγγίσεις όμως δε λαμβάνουν υπόψιν τις ιδιαίτερες ιδιότητες των προτύπων και για λόγο αυτό δεν είναι ικανές να τα διαχειριστούν αποτελεσματικά. Επιπλέον είναι κατασκευασμένες για ένα συγκεκριμένο τύπο προτύπων ενώ τα τελευταία μπορούν να παρουσιαστούν σε πολλές διαφορετικές μορφές ανά εφαρμογή [1].

Επομένως, μία άλλη προσέγγιση είναι απαραίτητη που να έχει ως βασικό στοιχείο διαχείρισης το πρότυπο και να λαμβάνει υπόψιν τις ιδιαίτεροτήτες αυτού, να είναι όμως ικανή να αντιμετωπίζει πρότυπα όλων των ειδών. Πολλές εφαρμογές που έχουν σχέση με πρότυπα (τηλεπικοινωνίες, ιατρική, περιβαλλοντολογικά πληροφοριακά συστήματα, αστρονομία, γέα-επιστήμες κα.) θα μπορούσαν να ωφεληθούν από ένα ολοκληρωμένο σύστημα διαχείρισης προτύπων στο οποίο θα μπορούν να γίνουν λειτουργίες αποθήκευσης, ανάκτησης και ερώτησης πάνω στα πρότυπα (και σύγκρισης αυτών), αποφεύγοντας τη συνεχή πρόσβαση και χρονοβόρα αναζήτηση στα μεγάλου όγκου αρχικά δεδομένα [2].

## 1.5 Σύστημα Διαχείρισης Βάσεων Προτύπων (ΣΔΒΠ)

Από τα παραπάνω συμπεραίνουμε ότι ένα ξεχωριστό σύστημα διαχείρισης είναι απαραίτητο. Με βάση τις ιδιαίτεροτήτες των προτύπων, ένα τέτοιο σύστημα θα πρέπει να χαρακτηρίζεται από [3]:

- Γενικότητα (generality). Το σύστημα πρέπει να είναι σχεδιασμένο για να δέχεται πρότυπα διαφόρων τύπων ώστε να μπορεί να υποστηρίζει τις διαφορετικές εφαρμογές.
- Επεκτασιμότητα (extensibility). Το σύστημα πρέπει να είναι επεκτάσιμο ούτως ώστε να μπορεί να ενσωματώσει και να διαχειριστεί αποτελεσματικά τύπους προτύπων που είναι πιθανό να εμφανιστούν από νέες εφαρμογές και ανάγκες διαφορετικών χρηστών.
- Εκμετάλλευση των ιδιαιτεροτήτων (χαρακτηριστικών) των προτύπων. Τα χαρακτηριστικά της δομής των προτύπων, σύμφωνα με τη μοντελοποίησή τους (κεφ. 2.1), μπορούν πρέπει να ληφθούν υπόψιν προκειμένου να γίνει

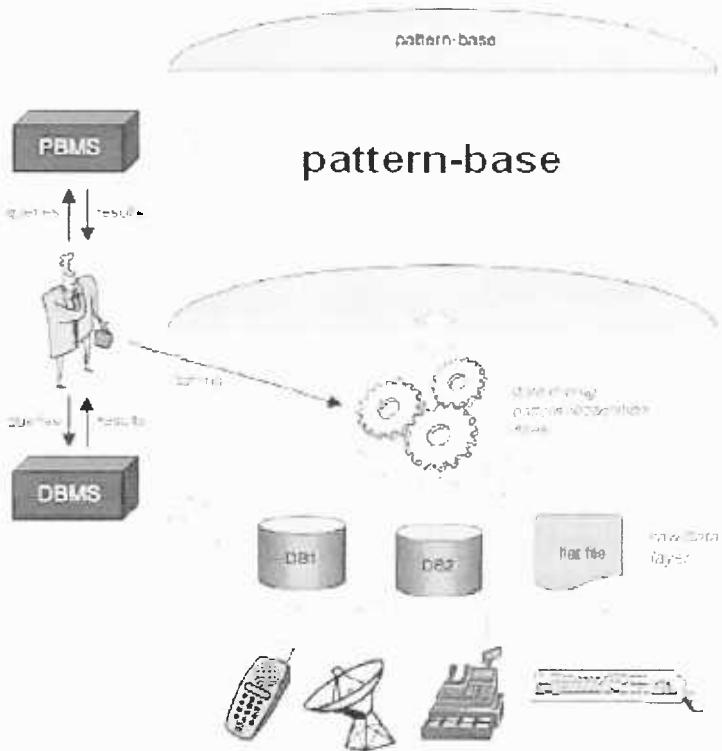
ευκολότερη η υλοποίηση της βάσης, της δημιουργίας ευρετηρίων και η δυνατότητα ερωτήσεων.

- Δυνατότητα υλοποίησης περιορισμών που ορίζονται από το λογικό μοντέλο. Το σύστημα θα πρέπει να διαθέτει τους κατάλληλους μηχανισμούς ελέγχου της ορθότητας των προτύπων-δεδομένων που αποθηκεύονται σε αυτό και των μεταξύ τους σχέσεων.
- Επαναχρησιμοποίηση (reusability). Το σύστημα πρέπει να έχει τη δυνατότητα να επαναχρησιμοποιεί τα ήδη ορισμένα στοιχεία ούτως ώστε να επιτυγχάνεται μείωση της πολυπλοκότητας.

Το πρόβλημα της απευθείας αποθήκευσης και ανάκτησης προτύπων, αντίθετα με αυτό των Data Warehouse συστημάτων, δεν έχει αποσπάσει μέχρι τώρα την απαραίτητη προσοχή των ανθρώπων τόσο του εμπορικού όσο και του επιστημονικού τομέα. Είναι όμως σίγουρο ότι ο τελικός χρήστης έχει να επωφεληθεί από τη χρήση συστημάτων διαχείρισης προτύπων (Pattern-Based Management Systems, PBMS) γιατί προσφέρουν [3]:

- Αφαίρεση (Abstraction). Το ΣΔΒΠ δεν χειρίζεται τα αρχικά δεδομένα αλλά τα πρότυπα που τα χαρακτηρίζουν. Τα πρότυπα μεταχειρίζονται από το ΣΔΒΠ με τον ίδιο τρόπο που τα ΣΔΒΔ μεταχειρίζονται τα αρχικά δεδομένα.
- Αποτελεσματικότητα (Efficiency). Με την ύπαρξη νέας αρχιτεκτονικής γίνεται σαφής διαχωρισμός του ΣΔΒΔ και του ΣΔΒΠ βελτιώνοντας την απόδοση και των δύο ενώ επιτρέπεται παράλληλα η ανάπτυξη προχωρημένων τεχνικών επεξεργασίας προτύπων.
- Δυνατότητα Ερωτήσεων (Querying). Το ΣΔΒΠ θα παρέχει ξεχωριστή και αποτελεσματική γλώσσα για την δημιουργία επερωτήσεων προκειμένου να επιτυγχάνεται η αναζήτηση και η σύγκριση προτύπων.

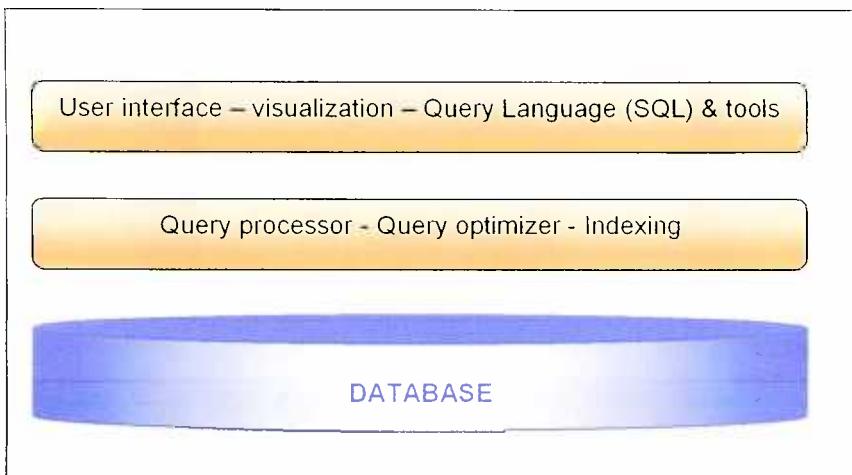
Συνοψίζοντας, ένα ΣΔΒΠ θα είναι ένα σύστημα που θα περιέχει και θα διαχειρίζεται πρότυπα που έχουν προκύψει από αλγόριθμους εξόρυξης γνώσης και είναι αποθηκευμένα σε μια βάση προτύπων (pattern-base). Οι αλγόριθμοι αυτοί εφαρμόζονται σε δεδομένα αποθηκευμένα σε βάσεις δεδομένων που με τη σειρά τους έχουν συλλεχθεί από διάφορες πηγές. Ο χρήστης μέσω ενός ΣΔΒΔ μπορεί να επεξεργαστεί τα δεδομένα αυτά. Με τον ίδιο τρόπο μέσα από ένα ΣΔΒΠ θα μπορεί να αναφέρεται στα πρότυπα. Στην Εικόνα 1 φαίνονται σχηματικά όλες οι σχέσεις.



**Εικόνα 1** Συστήματα διαχείρισης βάσεων δεδομένων και προτύπων

Στο χαμηλότερο επίπεδο, ένα σύνολο συσκευών παράγει δεδομένα τα οποία οργανώνονται και αποθηκεύονται σε βάσεις δεδομένων και αρχεία διαφόρων ειδών για να χρησιμοποιηθούν (συνήθως) από ένα Σύστημα Διαχείρισης Βάσεων Δεδομένων. Αλγόριθμοι εξόρυξης γνώσης εφαρμόζονται στα δεδομένα αυτά για να παράγουν πρότυπα που θα αποθηκευτούν στη Βάση Προτύπων (pattern-base). Οι χρήστες ρυθμίζουν τους αλγόριθμους εξόρυξης και μπορούν να αλληλεπιδράσουν απευθείας με το ΣΔΒΔ όσο και με το ΣΔΒΠ.

Από άποψη αρχιτεκτονικής, ένα ΣΔΒΠ θα είναι παρόμοιο με ένα ΣΔΒΔ. Σε γενικές γραμμές, το χαμηλότερο επίπεδο είναι αυτό της φυσικής αποθήκευσης, το επόμενο επίπεδο περιέχει τους μηχανισμούς ευρετηριοποίησης και τον query optimizer, και στο πιο ψηλό επίπεδο βρίσκεται η γλώσσα ερωτήσεων και η διεπαφή με το χρήστη (Εικόνα 2). Ενώ υπάρχει αντιστοιχία στα επίπεδα αρχιτεκτονικής μεταξύ των ΣΔΒΠ και ΣΔΒΔ, είναι απαραίτητες κάποιες τροποποιήσεις σε καθένα από αυτά ώστε η διαχείριση των προτύπων να είναι προσανατολισμένη στα ιδιαίτερα χαρακτηριστικά τους.

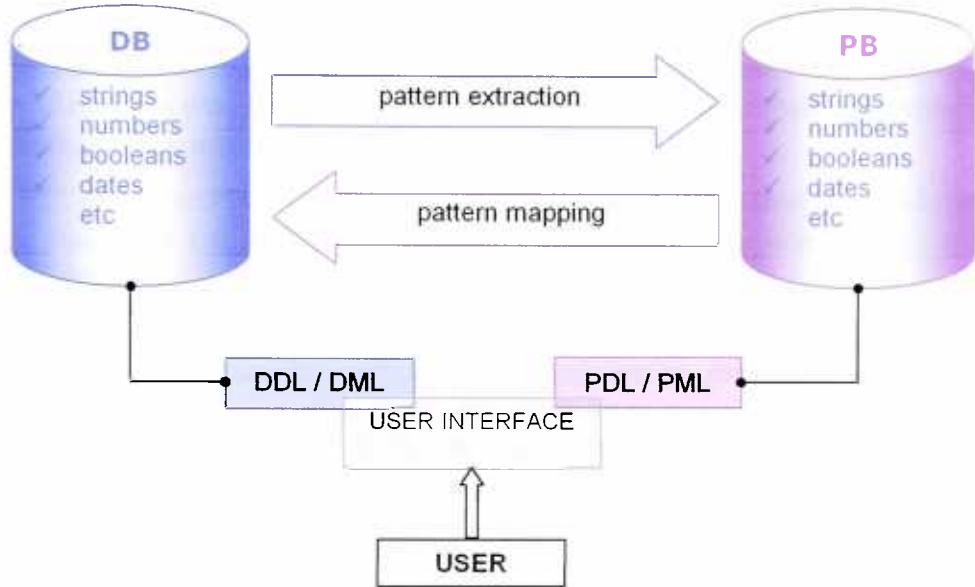


**Εικόνα 2 Αρχιτεκτονική βάσης δεδομένων**

Μια προσέγγιση για τον ορισμό και μοντελοποίηση ενός συστήματος διαχείρισης προτύπων έγινε από τα μέλη του ερευνητικού έργου της Ευρωπαϊκής Κοινότητας, *PANDA* (*PAterns for Next-generation DAtabase Systems*) που προτείνουν ένα λογικό μοντέλο και ένα μοντέλο λειτουργίας για ένα ΣΔΒΠ [3]. Η προτεινόμενη μέσα από το σχεδίαστη έχει στοιχεία που μοιάζουν με αυτά μιας αντικειμενοστραφούς (object-oriented) προσέγγισης όμως έχει βασικές διαφορές από αυτή [2].

Ένα σύστημα διαχείρισης προτύπων διαχειρίζεται (αποθηκεύει, επεξεργάζεται, ανακτά) πρότυπα που έχουν προκύψει από ένα σύνολο αρχικών δεδομένων (raw data) με σκοπό τη δυνατότητα σύγκρισης προτύπων και την υποστήριξη βασικών λειτουργιών σχετικές με τα πρότυπα, για την εξαγωγή χρήσιμων πληροφοριών. Το σύνολο των προτύπων που διαχειρίζεται ένα ΣΔΒΠ λέγεται βάση-προτύπων (pattern-base) [3].

Η ιδέα ενός ΣΔΒΠ, απεικονίζεται στο παρακάτω σχήμα (Εικόνα 3). Τα πρότυπα εξάγονται από ένα κλασικό ΣΔΒΔ και αποθηκεύονται στο ΣΔΒΠ. Κατ' αντιστοιχία με τις γλώσσες ορισμού και διαχείρισης δεδομένων (DDL, DML) θα υπάρχουν γλώσσες ορισμού και διαχείρισης προτύπων (PDL, PML), που θα λαμβάνουν υπόψιν την πολυπλοκότητα και πολυμορφία των προτύπων και θα χρησιμοποιούνται για την αλληλεπίδρασή του χρήστη με το ΣΔΒΠ.



### Εικόνα 3 Σύστημα Διαχείρισης Βάσεων Προτύπων και Βάσεων Δεδομένων

Το ΣΔΒΠ θα συνυπάρχει, όπως φαίνεται και στο σχήμα, με το παραδοσιακό ΣΔΒΔ και θα υπάρχει σύνδεση-αντιστοίχιση των προτύπων με τα δεδομένα από τα οποία εξήχθησαν και αντίστροφα.

Τα βασικά θέματα που προκύπτουν από αυτήν την προσέγγιση, είναι:

- Θέματα φυσικής αποθήκευσης και αναπαράστασης προτύπων, δημιουργία ευρετηρίων κ.α.
- Θέματα γλώσσας ορισμού και διαχείρισης προτύπων.
- Θέματα αντιστοίχισης μεταξύ των προτύπων και των απλών δεδομένων από το οποία εξήχθησαν.

## 1.6 Ορισμός του προβλήματος – Στόχος της έρευνας

Η ύπαρξη ενός συστήματος που να διαχειρίζεται τα πρότυπα φαίνεται να είναι απαραίτητη. Το ερώτημα που τίθεται είναι αν μπορεί ένα τέτοιο σύστημα να υλοποιηθεί χρησιμοποιώντας τις ήδη υπάρχουσες τεχνολογίες βάσεων δεδομένων. Μπορούν οι ιδιαίτερότητες που διέπουν ένα τέτοιο σύστημα να απεικονιστούν σε ένα από τα γνωστά μοντέλα Βάσεων Δεδομένων, ή για την μεγαλύτερη αποτελεσματικότητα θα ήταν προτιμότερο να σχεδιαστεί ένα καινούργιο μοντέλο; Οι μέθοδοι αποθήκευσης, ευρετηριωτοίσης και ανάκτησης (γλώσσα ερωτήσεων κλπ) των Για να απαντηθεί το ερώτημα αυτό είναι απαραίτητο να υλοποιηθεί μια βάση προτύπων α) σε σχεσιακή βάση (relational approach), β) σε αντικειμενο-σχεσιακή βάση (object-relational approach) και γ) σε ημι-δομημένη XML βάση (semistructured approach). Οι διαφορετικές υλοποίησεις θα πρέπει να συγκριθούν ώστε εκτός της δυνατότητας υλοποίησης να μελετηθεί η αποτελεσματικότητα και η απόδοσή τους.

Οι υλοποιήσεις σε σχεσιακή και αντικειμενο-σχεσιακή έχουν υλοποιηθεί [11]. Στόχος της παρούσης εργασίας είναι να υλοποιηθεί μια βάση προτύπων σε XML, να συγκριθεί με τις δύο άλλες υλοποιήσεις, να διαπιστωθεί ποια είναι η καταλληλότερη για την υλοποίηση μιας βάσης προτύπων και να επισημανθούν τα πλεονεκτήματα και μειονεκτήματα που παρουσιάζονται.

Λόγο της φύσης των προτύπων και των αναγκών για ανταλλαγή των δεδομένων μεταξύ διαφόρων εφαρμογών, η λύση της XML βάσης πιστεύεται ότι είναι η καταλληλότερη.

Η εργασία στα επόμενα κεφάλαια θα οργανωθεί ως εξής. Θα περιγραφτεί ένα μοντέλο για το ΣΔΒΠ και θα οριστούν τα στοιχεία και τα χαρακτηριστικά του. Στη συνέχεια θα ακολουθήσει η υλοποίηση μιας βάσης προτύπων σε XML. Στο κεφάλαιο 3 θα παρουσιαστεί η γλώσσα XML, τα βασικά χαρακτηριστικά της, οι επιλογές υλοποίησης που ακολουθήθηκαν καθώς και η υλοποίησή της σε περιβάλλον ORACLE 9i.

Εκτός των συμπερασμάτων που θα εξαχθούν για τη συγκεκριμένη υλοποίηση, θα ακολουθήσει μια σύγκριση με άλλα συστήματα σε σχεσιακό και αντικειμενο-σχεσιακό μοντέλο. Θα εξεταστεί η αποτελεσματικότητά τους σε όρους ευκολίας υλοποίησης, ευκολίας έκφρασης ερωτήσεων, δυνατότητα διαχείρισης διαφορετικών ειδών προτύπων και επεκτασιμότητας. Τέλος θα παρουσιαστούν τα συμπεράσματα της εργασίας και η ανοικτή έρευνα

## 2. ΜΟΝΤΕΛΟΠΟΙΗΣΗ ΣΥΣΤΗΜΑΤΟΣ ΔΙΑΧΕΙΡΙΣΗΣ ΒΑΣΕΩΝ ΠΡΟΤΥΠΩΝ

### 2.1 Το λογικό μοντέλο ενός ΣΔΒΠ

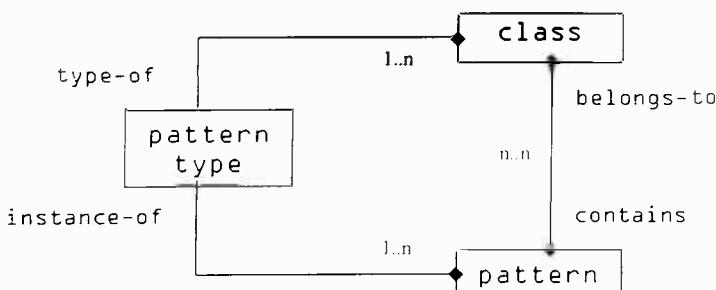
Οι ερευνητές του έργου PANDA έχουν ορίσει το λογικό μοντέλο ενός ΣΔΒΠ υιοθετώντας τους παρακάτω ορισμούς [3].

Ως **πρότυπο** (pattern) ορίζεται μια σύντομη και περιεκτική σημασιολογικά αναπαράσταση ενός συνόλου δεδομένων (raw data). Ένα πρότυπο μπορεί να είναι είτε ένα επαναλαμβανόμενο υποσύνολο των αρχικών δεδομένων είτε σε μορφή είτε σε τιμές, είτε μια επαναλαμβανόμενη σχέση μεταξύ μερών του συνόλου των δεδομένων (πχ. Η αναλυτική φόρμα ενός πολυωνυμικού περιορισμού που ισχύει για κάποια δεδομένα, clustering) [3].

Τα πρότυπα μπορούν να ομαδοποιηθούν σε τύπους προτύπων (pattern types). Ένας **τύπος προτύπου** περιλαμβάνει τη δομή (structure), το μέτρο ποιότητας (measure), την πηγή (source) και την έκφραση (expression). Η δομή χαρακτηρίζει το πρότυπο στο χώρο των προτύπων, περιγράφει απλά τη δομή του. Το μέτρο ποιότητας χαρακτηρίζει το πρότυπο μετρώντας το κατά πόσο η απεικόνιση των δεδομένων αντιπροσωπεύει την πραγματική φύση των αρχικών δεδομένων. Η πηγή περιγράφει τα αρχικά δεδομένα με τα οποία σχετίζεται το πρότυπο ενώ η έκφραση περιγράφει προσεγγιστικά την αντιστοίχιση (mapping) μεταξύ του προτύπου και των αρχικών δεδομένων.

Χαρακτηριστικό των προτύπων είναι ότι το καθένα αναφέρεται σε πολλά δεδομένα αλλά και διαφορετικά πρότυπα (πιθανώς και διαφορετικού τύπου) μπορεί να σχετίζονται με τα ίδια δεδομένα.

Με τον ορισμό του προτύπου και του τύπου προτύπου μπορούμε να ορίσουμε και την κλάση. Μια **κλάση** είναι ένα σύνολο από σημασιολογικά συσχετιζόμενα πρότυπα και αποτελεί το κλειδί για τον ορισμό μιας γλώσσας επερωτήσεων προτύπων. Μία κλάση ορίζεται για ένα συγκεκριμένο τύπο προτύπου και περιλαμβάνει μόνο πρότυπα του τύπου αυτού. Επιπλέον κάθε πρότυπο ανήκει τουλάχιστον σε μία κλάση. Σχηματικά:



Εικόνα 4 Κλάσεις, τύποι προτύπων και πρότυπα

### 2.1.1 Τύποι Προτύπων (pattern types)

Ως απλοί τύποι ορίζονται οι ακέραιοι, οι πραγματικοί αριθμοί, οι Boolean, οι συμβολοσειρές και οι χρονοσφραγίδες (timestamps). Οι σύνθετοι τύποι (ουσιαστικά type constructors) περιλαμβάνουν λίστες, σύνολα, πίνακες, ζεύγη (tuples) και bags. Παραδείγματα τύπων είναι (οι τύποι είναι με κεφαλαία):

- salary: REAL
- SET(INTEGER)
- TUPLE(x: INTEGER, y: INTEGER)
- Personnel: LIST(TUPLE(age: INTEGER, salary: INTEGER))

Με τον ορισμό τύπων δίνεται μια περιγραφή της δομής των προτύπων και της σχέσης τους με τα πρότυπα.

Ένας τύπος προτύπου ορίζεται τυπικά ως εξής:

$pt = (n, ss, ds, ms, f)$

όπου:

n: όνομα τύπου

ss: σχήμα δομής (structure schema)

ds: σχήμα πηγής-αρχικών δεδομένων (source schema)

ms: ποιοτικό μέτρο (measure schema)

f: τύπος αντιστοίχισης (formula)

Το σχήμα δομής ορίζει το χώρο των προτύπων περιγράφοντας τη δομή του κάθε προτύπου.

Το σχήμα αρχικών δεδομένων ορίζει τα αρχικά δεδομένα από τα οποία τα πρότυπα έχουν προκύψει.

Το ποιοτικό μέτρο περιγράφει τα μέτρα που χρησιμοποιούνται για τη μέτρηση της ποιότητας σχετικά με την πιστότητα της αναπαράσταση των προτύπων από τα αρχικά δεδομένα. Το μέτρο αυτό είναι σημαντικό γιατί δίνει στο χρήστη μια εικόνα του πόσο ακριβής είναι η απεικόνιση των δεδομένων σε πρότυπα.

Ο τύπος αντιστοίχισης f περιγράφει τη σχέση μεταξύ των αρχικών δεδομένων και των προτύπων. Τις περισσότερες φορές, αν και περιγράφει ακριβώς τη σχέση μεταξύ των δύο χώρων (αρχικά δεδομένα – πρότυπα), είναι προσεγγιστικός.

Ένα παράδειγμα τύπου προτύπου είναι:

n: AssociationRule

ss: TUPLE(head: SET(STRING), body: SET(STRING))

```

ds: BAG(transaction: SET(STRING))

ms: TUPLE(confidence: REAL, support: REAL)

f: head U body ⊆ transaction

```

### 2.1.2 Πρότυπα (patterns)

Έστω ότι έχουμε ένα τύπο προτύπου  $pt=(n,ss,ds,ms,et)$ . Ένα πρότυπο αυτού του τύπου ορίζεται τυπικά ως εξής:

$p = (pid, s, d, m, e)$

όπου:

$pid$ : το μοναδικό αναγνωριστικό του προτύπου

$s$ : μια τιμή για τον τύπο  $ss$

$d$ : ένα σύνολο δεδομένων του ίδιου τύπου με το  $ds$

$m$ : μια τιμή για τον τύπο  $ms$

$e$ : μια έκφραση-τύπος που δηλώνει την περιοχή των αρχικών δεδομένων που σχετίζονται με το πρότυπο  $p$

Πολλές άλλες πληροφορίες θα μπορούσαν να συσχετιστούν με τον ορισμό ενός προτύπου όπως για παράδειγμα ο αλγόριθμος (εξόρυξης) που το σχημάτισε, οι παράμετροι που είχαν τεθεί κ.α. Ένα παράδειγμα προτύπου σε σχέση με το προηγούμενο παράδειγμα τύπου προτύπου είναι:

```

pid: 413

s: (head={'Boots'}, body={'Socks', 'Hat'})

d: 'SELECT SETOF(article) AS transaction
    FROM sales GROUP BY transactionId'

m: (confidence=0.75, support=0.55)

e: {transaction: {'Boots', 'Socks', 'Hat'} ⊆ transaction}

```

### 2.1.3 Κλάσεις (classes)

Μία κλάση ορίζεται τυπικά ως εξής:

$c = (cid, pt, pc)$

όπου:

$cid$ : το μοναδικό αναγνωριστικό της κλάσης

$pt$ : ένας τύπος προτύπου

pc: ένα σύνολο από πρότυπα τύπου pt

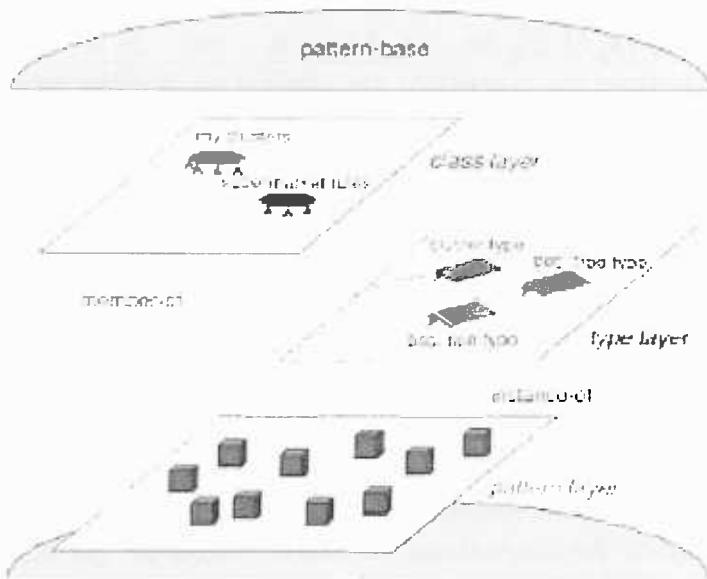
Ένα παράδειγμα κλάσης είναι:

cid: 123

pt: AssociationRule

pc: {413, 414, 415, 416, 417, 418, 419, 420}

Στο παρακάτω σχήμα μπορούμε να δούμε τα περιεχόμενα της βάσης προτύπων η που υπήρχε στην Εικόνα 1.



Εικόνα 5 Η βάση προτύπων και τα συστατικά της

Στο επίπεδο προτύπων (pattern layer) βρίσκονται τα πρότυπα που παράγονται από τα εργαλεία εξόρυξης γνώσης ή από άλλες εφαρμογές. Στο επίπεδο τύπου προτύπου υπάρχουν οι τύποι των διαφόρων προτύπων που είναι ορισμένοι είτε από το σύστημα είτε από τους χρήστες. Τα πρότυπα κάθε τύπου έχουν τα ίδια χαρακτηριστικά δομής. Στο ανώτερο επίπεδο, το επίπεδο κλάσεων υπάρχουν οι ορισμοί των κλάσεων προτύπων δηλαδή σύνολο σημασιολογικά συσχετιζόμενων προτύπων [3].

## 2.2 Σχέσεις μεταξύ προτύπων

Για την αύξηση της λειτουργικότητας, τη βελτιστοποίηση της επαναχρησιμότητας και επεκτασιμότητας του μοντέλου ΣΔΒΠ, κάποιες σχέσεις μεταξύ των προτύπων έχουν προταθεί.

### 2.2.1 Εξειδίκευση (specialization)

Η αφαίρεση (abstraction), η εξειδίκευση (specialization) και οι ιδιότητες της κληρονομικότητας (inheritance) έχουν υιοθετηθεί από πολλές μοντελοποιήσεις δεδομένων για το λόγο ότι βελτιώνουν την επεκτασιμότητα και επαναχρησιμοποίηση επιτρέποντας έτσι την εισαγωγή νέων οντοτήτων που απλά προκύπτουν από ήδη υπάρχουσες.

Για τον ορισμό της έννοιας της εξειδίκευσης πρέπει πρώτα να οριστεί η έννοια του υπο-τύπου μεταξύ βασικών τύπων (ο τύπος `integer` για παράδειγμα είναι υπο-τύπος του τύπου `real`). Η έννοια του υπο-τύπου μπορεί να οριστεί και για σύνθετους τύπους. Όταν μεταξύ δύο σύνθετων τύπων οι πιο «εξωτερικοί» type constructors συμπίπτουν και κάθε συστατικό του ενός τύπου είναι εξειδίκευση των αντίστοιχων συστατικών του άλλου τύπου τότε ο δεύτερος εξειδικεύει τον πρώτο. Όταν ακόμα τα σχήματα δομής, αρχικών δεδομένων και το ποιοτικό μέτρο ενός τύπου `pt1` εξειδικεύει τα αντίστοιχα ενός άλλου `pt2`, τότε ο `pt1` εξειδικεύει τον `pt2` (ο `pt1` είναι υπο-τύπος του `pt2`). Σχετικά με τις κλάσεις, αν ο τύπος `pt1` εξειδικεύει τον τύπο `pt2` και η κλάση `c` ορίζεται για τον `pt2` τότε και τα στιγμιότυπα (τα πρότυπα) του `pt1` μπορεί να είναι μέρος της `c` [3].

### 2.2.2 Σύνθεση και εκλέπτυνση (Composition and refinement)

Η δημιουργία πολύπλοκων τύπων με τη σύνθεση άλλων επιτυγχάνεται με την επέκταση του συνόλου των βασικών τύπων με άλλους τύπους προτύπων. Η δυνατότητα που δίνεται στο χρήστη να δηλώσει πολύπλοκους τύπους έχει δύο επιπτώσεις στη μοντελοποίηση του PBMS.

Πρώτον, δίνεται η δυνατότητα να δηλωθεί ένα σχήμα δομής ως ένα σύνθετο σχήμα (να οριστεί στο structure schema ένας τύπος προτύπου) οπότε αναδρομικά να μπορεί να οριστεί ένα είδος ιεραρχίας. Αυτό το χαρακτηριστικό ονομάζεται σύνθεση (composition).

Δεύτερον, ένας πολύπλοκος τύπος μπορεί να εμφανιστεί στο σχήμα πηγής-αρχικών δεδομένων (να οριστεί στο source schema ένας τύπος προτύπου). Αυτό επιτρέπει τη δήλωση προτύπων που έχουν προέρθει από εξόρυξη άλλων προτύπων. Αφού ένα πρότυπο είναι μια συνοπτική αναπαράσταση των αρχικών δεδομένων που αντιπροσωπεύει, αυτό το χαρακτηριστικό μπορεί να ονομαστεί εκλέπτυνση. Στην ουσία με τον τρόπο αυτό επιτυγχάνεται μεγαλύτερο επίπεδο λεπτομέρειας [3].

## 2.3 Ομοιότητα προτύπων

Δύο πρότυπα του ίδιου τύπου μπορεί να συγκριθούν για να υπολογιστεί μια τιμή αμοιβαίας ομοιότητας  $s$ ,  $s \in [0,1]$ . Η ομοιότητα μεταξύ δύο προτύπων μπορεί να υπολογιστεί ως συνάρτηση της ομοιότητας της δομής τους (structure), των μέτρων ποιότητας (measure), της πηγής (source) και της έκφρασης (expression):

*pattern\_similarity =*

*f(structure\_similarity, measure\_similarity, source\_similarity, expression\_similarity)*

Αν δύο πρότυπα έχουν ακριβώς την ίδια δομή, τότε το μέτρο ομοιότητας αντιστοιχεί στη σύγκριση των υπόλοιπων στοιχείων. Στη γενική περίπτωση όμως τα πρότυπα διαφέρουν στη δομή τους, οπότε ένα αρχικό βήμα χρειάζεται για να γίνουν συμβατά ως προς τη δομή τους για να μπορούν να συγκριθούν.

Οι πιο ενδιαφέρουσες περιπτώσεις στην ομοιότητα προτύπων είναι η ανάκτηση των όμοιων προτύπων προς ένα πρότυπο που δίνεται ως είσοδος και η ανάκτηση των προτύπων που η ομοιότητά τους ξεπερνά ένα συγκεκριμένο όριο.

Τα βασικότερα θέματα σχετικά με την ομοιότητα προτύπων είναι η χρησιμοποίηση της σημασιολογίας των δεδομένων, δηλαδή η ομοιότητα βασισμένη όχι μόνο σε απλές τιμές κάποιων μέτρων αλλά η σημασία των δεδομένων και οι συσχετίσεις μεταξύ της δομής ή/ και του μέτρου ποιότητας των προτύπων.

Η σύγκριση των προτύπων θα είναι μια από τις βασικές διαδικασίες που θα επιθυμούν τη χρήση ενός συστήματος διαχείρισης προτύπων για αυτό το λόγο πρέπει να ληφθεί υπόψιν στην υλοποίηση του συστήματος αυτού. Η έρευνα για τον ορισμό της ομοιότητας των προτύπων δεν έχει πλήρως ολοκληρωθεί έχουν και δεν εξετάζεται στην παρούσα εργασία.

### 3. ΥΛΟΠΟΙΗΣΗ XML ΒΑΣΗΣ ΠΡΟΤΥΠΩΝ

#### 3.1 XML (Extensible Markup Language)

Η XML έχει πολλά κοινά με την HTML, αφού είναι ένα μέρος του SGML (Standard Generalized Markup Language) [21]. Ο βασικός όμως σκοπός της XML δεν είναι να περιγράφει τη μορφοποίηση του κειμένου, αλλά να μεταδίδει δομημένα δεδομένα. Ενώ η HTML περιγράφει τη μορφή ενός κειμένου, η XML περιγράφει το περιεχόμενο. Με την XML λύνονται πολλά προβλήματα που υπάρχουν για την μετάδοση και αξιοποίηση των δεδομένων που βρίσκονται στον παγκόσμιο ιστό αλλά και σε διαφορετικές εφαρμογές [12]. Όμως η εξάπλωση της XML και η αξιοποίηση πολλών δυνατοτήτων της, οδήγησαν στο να χρησιμοποιείται όχι μόνο για μετάδοση αλλά και για αποθήκευση και διαχείριση δεδομένων. Έτσι, τώρα αναφερόμαστε σε XML βάσεις δεδομένων που στην ουσία είναι η συλλογή πολλών XML εγγράφων τα οποία ακολουθούν μια «χαλαρή» δομή και περιγράφουν το περιεχόμενό τους ανάλογα με την εφαρμογή. Σημαντική ωστόσο είναι η έλλειψη προτυποποιημένης και κοινά αποδεκτής γλώσσας ερωτήσεων (query language) για XML βάσεις δεδομένων, χωρίς αυτό να σημαίνει ότι δεν υπάρχουν τέτοιες (XQuery) [15, 19]. Επίσης το θέμα της αποτελεσματικής αποθήκευσης των XML εγγράφων δεν έχει ακόμα λυθεί, αν και έχουν προταθεί διάφορες εναλλακτικές πρακτικές. Ο χώρος των XML βάσεων και εφαρμογών γνωρίζει μεγάλη ανάπτυξη και η έρευνα σε θέματα αποθήκευσης και γλώσσας ερωτήσεων XML εγγράφων είναι εκτεταμένη.

Μεγάλες εταιρείες συστημάτων βάσεων δεδομένων (πχ. ORACLE, MICROSOFT) ενσωματώνουν πλέον την υποστήριξη XML εγγράφων, δίνοντας τη δυνατότητα όχι μόνο της αποθήκευσης και αναπαράστασης XML εγγράφων αλλά και της σύνδεσής τους με δεδομένα και πίνακες του παραδοσιακού σχεσιακού ή αντικειμενο-σχεσιακού μοντέλου (στην ουσία στηρίζονται σε αυτά). Από την άλλη μεριά υπάρχουν τα αμιγώς XML συστήματα βάσεων δεδομένων (πχ. TAMINO, eXist) που αντιθέτως με τα προαναφερθέντα δε βασίζονται σε δομές και τεχνικές παραδοσιακών συστημάτων (σχεσιακού κλπ) για την αποθήκευση XML εγγράφων (και τη δυνατότητα ερωτήσεων σε αυτά), αλλά αποτελούν αυτόνομα XML συστήματα με τεχνικές αποθήκευσης, αναπαράστασης και γλώσσα ερωτήσεων αποκλειστικά για XML έγγραφα.

Τα βασικά χαρακτηριστικά της XML είναι:

- XML ετικέτες (tags):
  - Δυνατότητα ορισμού από τον χρήστη
  - Περιγράφουν τη δομή και τη σημασία των δεδομένων
- Ποικιλία εφαρμογών:
  - Ανταλλαγή δεδομένων (Data exchange)
  - Εξαγωγή δεδομένων (Data extraction)

- Μετασχηματισμός δεδομένων (Data transformation)
- Ολοκλήρωση δεδομένων (Data integration)

Στην XML η δομή των δεδομένων δεν είναι πολύ αυστηρή. Για παράδειγμα είναι δυνατόν να ορίσει κάποιος μια νέα ετικέτα χωρίς να επηρεάζει αυτό την εφαρμογή που λαμβάνει τα δεδομένα που απλά την αγνοεί. Μπορεί ο κάθε χρήστης να τροποποιεί (να εμπλουτίζει) τη δομή των δεδομένων ανάλογα με τις ανάγκες του χωρίς να επηρεάζει τους άλλους χρήστες που μπορούν να αγνοούν τις τροποποιήσεις του.

Το χαρακτηριστικό αυτό είναι πολύ σημαντικό στην περίπτωση των προτύπων καθότι τα χαρακτηριστικά τους και η δομή τους διαφέρουν από σε τομέα σε τομέα (επεξεργασία σήματος, αναγνώριση εικόνας, εξόρυξη γνώσης, χωρικές βάσεις δεδομένων κλπ) αλλά και μεταξύ εφαρμογών του ίδιου τομέα (πχ. διαφορετική αναπαράσταση κανόνων συσχέτισης σε διαφορετικές εφαρμογές).

Σε αντίθεση με τις παραδοσιακές βάσεις δεδομένων, σε μια XML βάση, τα έγγραφα (documents) δεδομένων μπορούν να δημιουργηθούν χωρίς να υπάρχει κάποιο σχήμα που να τα περιορίζει. Ένα πεδίο (element) μπορεί να έχει οποιοδήποτε υπο-πεδίο (sub-element) ή γνώρισμα (attribute). Η ελευθερία αυτή προέρχεται από το ότι στην XML τα δεδομένα είναι αυτοπεριγραφόμενα και δεν χρησιμοποιείται μεγάλος αριθμός συσχετίσεων προκειμένου να αναπαρασταθούν πολύπλοκοι τύποι δεδομένων, όπως στο σχεσιακό μοντέλο. Η χρήση φωλιασμένων πεδίων (nested elements) στην XML μειώνουν τον αριθμό των συσχετίσεων που χρειάζεται να αναπαρασταθούν.

Η ευκολία της XML στην ανταλλαγή εγγράφων-δεδομένων είναι καταλυτική για την εφαρμογή των προτύπων. Διαφορετικοί χρήστες μπορούν να χρησιμοποιούν διαφορετική αναπαράσταση και μορφή για τα ίδια πρότυπα, διαφορετικά ονόματα για τα ίδια δεδομένα ή και ακόμα ίδια ονόματα για διαφορετικά δεδομένα. Μπορούν να ανταλλάσσουν και να επεξεργάζονται τα έγγραφα χρησιμοποιώντας μόνο τα στοιχεία για τα οποία ενδιαφέρονται. Τα θέματα μετασχηματισμού και μορφοποίησης των δεδομένων μεταξύ διαφορετικών XML αναπαραστάσεων, αντιμετωπίζονται μέσω γλωσσών όπως είναι η XSLT και η XQuery [11].

```

<PurchaseOrder orderDate="1999-05-20">
    <shipTo type="US">
        <name>Alice Smith</name>
        <street>123 Maple Street</street>
        <city>Mill Valley</city>
        <state>CA</state>
        <zip>90952</zip>
    </shipTo>
    <billTo type="UK">
        <name>Trevor Mostyn</name>
        <street>12, The Gables</street>
        <city>Bourton-on-the-Water</city>
        <state>Glous.</state>
        <zip>GL3 2BB</zip>
    </billTo>
    <shipDate>1999-05-25</shipDate>
    <comment>Get these things to me in a hurry, my lawn
going wild!</comment>
    <Items>
        <Item pno="333-333">
            <productName>Lawnmower,
                model BUZZ-1</productName>
            <quantity>1</quantity>
            <price>148.95</price>
            <comment>Please confirm this is the electri-
model</comment>
        </Item>
        <Item pno="444-444">
            <productName>Baby Monitor,
                model SNOOZE-2</productName>
            <quantity>1</quantity>
            <price>39.98</price>
        </Item>
    </Items>
</PurchaseOrder>

```

**Εικόνα 6 Παράδειγμα XML εγγράφου**

### 3.1.1 Καλώς-ορισμένα και ορθά XML έγγραφα. DTDs και XMLSchema

Αν και η XML είναι ευέλικτη και δεν απαιτεί κάποια προκαθορισμένη δομή, υπάρχουν κάποιοι βασικοί κανόνες που πρέπει να πληρούν όλα τα XML έγγραφα. Οι κανόνες αυτοί αφορούν στο ποια ονόματα είναι επιτρεπτά για τις ετικέτες (tags), τα γνωρίσματα που περιέχουν αυτές κλπ. Τα έγγραφα που όντως πληρούν τους απλούς

αυτούς κανόνες λέγεται ότι είναι καλώς ορισμένα (well-formed) και μπορούν να διερμηνευτούν από οποιαδήποτε XML διερμηνευτή.

Επιπλέον, η δομή ενός XML εγγράφου μπορεί να συγκριθεί και να επικυρωθεί σε σχέση με ένα αρχείο ορισμού τύπου εγγράφου (Document Type Definition, DTD) ή ένα XML σχήμα (XMLSchema). Ένα έγγραφο XML λέγεται ότι είναι έγκυρο όταν συμφωνεί με το αντίστοιχο ορισμό ή σχήμα του.

Ενώ και το DTDs και το XMLSchema είναι μηχανισμοί για τον ορισμό και έλεγχο της δομής των XML εγγράφων, διαφέρουν σε πολλά σημεία.

Στο DTD δηλώνονται τα πεδία (elements) και τα γνωρίσματα (attributes) που μπορούν να χρησιμοποιηθούν μέσα στο XML έγγραφο, ή ακόμα και αυτά που δεν μπορούν να χρησιμοποιηθούν. Δεν ορίζονται όμως περιορισμοί για το πόσα ίδια πεδία μπορούν να υπάρχουν, τι τύπου δεδομένα θα περιέχουν αυτά κα.

Ενώ το DTD είναι υποσύνολο του SGML, το XMLSchema στηρίζεται στην ίδια την XML, με αποτέλεσμα το σύνολο των υποστηριζόμενων δομών να είναι κι αυτό επεκτάσιμο. Το XMLSchema υποστηρίζει περισσότερες και πιο περίπλοκες δομές από το DTD και υπάρχει η δυνατότητα για περιγραφή περισσότερων και πιο ισχυρών περιορισμών γιατί υποστηρίζονται κάποιοι αρχικοί τύποι όπως συμβολοσειράς (string), ακεραίου (integer) κλπ [13].

Το XMLSchema παρέχει περισσότερη ελευθερία στον ορισμό ενός σχήματος ενώ ταυτόχρονα δίνει τη δυνατότητα για αυστηρότερο έλεγχο. Σε εφαρμογές δεδομενοκεντρικές συνεπώς, το XMLSchema είναι προτιμότερο [13].

```

<schema
    targetNamespace='http://.../PurchaseOrder'
    xmlns:po='http://.../PurchaseOrder'
    xmlns='http://www.w3.org/1999/XMLSchema'>

    <element name='PurchaseOrder'
        type='po:PurchaseOrderType'/>
    <element name='comment' type='string'/>
    <type name='PurchaseOrderType'>
        <element name='shipTo' type='po:Address'/>
        <element name='billTo' type='po:Address'/>
        <element name='shipDate' type='date'/>
        <element ref='po:comment' minOccurs='0' />
        <element name='Items' type='po:Items'/>
        <attribute name='orderDate' type='date' />
    </type>
    <type name='Address'>
        <element name='name' type='string'/>
        <element name='street' type='string'/>
        <element name='city' type='string'/>
        <element name='state' type='string'/>
        <element name='zip' type='integer'/>
        <attribute name='type' type='string' />
    </type>
    <type name='Items'>
        <element name='Item'
            minOccurs='0' maxOccurs='*'>
            <type>
                <element name='productName'
                    type='string'/>
                <element name='quantity'>
                    <datatype source='integer'>
                        <minExclusive value='0' />
                    </datatype>
                </element>
                <element name='price' type='decimal'/>
                <element ref='po:comment' minOccurs='0' />
                <attribute name='pno' type='string' />
            </type>
        </element>
    </type>
</schema>

```

### Εικόνα 7 Παράδειγμα XML Σχήματος

### 3.2 Βάση προτύπων και XML. Σχήματα για την περιγραφή των προτύπων.

Για το σχεδιασμό της Βάσης Προτύπων σε XML βάση δεδομένων, προτιμήθηκε για η περιγραφή του βασικού σχήματος με την XMLSchema για τους λόγους που αναφέρθηκαν.

Η δημιουργία ενός καλού σχήματος επηρεάζεται από πολλούς παράγοντες. Λόγω της ευελιξίας της XML είναι δυνατόν κάποιος να κατασκευάσει για την ίδια εφαρμογή περισσότερα του ενός σχήματα. Όμως δεν είναι σίγουρο ότι θα είναι όλα το ίδιο αποτελεσματικά. Ένα σχήμα δεν μπορεί να είναι σωστό ή λάθος, μπορεί όμως να είναι καλό ή κακό [13]. Αφενός, το σχήμα πρέπει να περιγράφει καλά και πλήρως την εφαρμογή και αφετέρου να ταιριάζει με τα περισσότερα πιθανά ερωτήματα (most-likely queries) ώστε να βελτιώνει την αποδοτικότητά της. Με βάση τις αρχές αυτές και το λογικό μοντέλο του ΣΔΒΠ κατασκευάστηκαν τα παρακάτω XML σχήματα. Σύμφωνα με το λογικό μοντέλο ενός ΣΔΒΠ [3], το κάθε πρότυπο ανήκει σε έναν «τύπο προτύπου» που το περιγράφει. Εξ ορισμού λοιπόν ο «τύπος προτύπου» παίζει το ρόλο του XML σχήματος. Ετσι για κάθε διαφορετικό τύπο προτύπου θα πρέπει να υπάρχει ένα XML σχήμα και τα πρότυπα θα είναι τα XML έγγραφα (XML instances) που θα ελέγχονται από το σχήμα αυτό.

Για παράδειγμα το σχήμα “association\_rule.xsd” περιγράφει τη μορφή των κανόνων συσχέτισης, είναι ο τύπος προτύπου (Pattern Type) των κανόνων αυτών. Το “pattern-association\_rules.xml” περιέχει πρότυπα του τύπου κανόνων συσχέτισης και ελέγχονται από το σχήμα “association\_rule.xsd”.

Ενα οποιοδήποτε πρότυπο (στη γενική μορφή) αποτελείται από τα πεδία “name”, “structure”, “source”, “measure” και “expression”. Συγκεκριμένα, στο πρότυπο κανόνα συσχέτισης το μέρος της δομής του (structure) αποτελείται πάντα από τα μέρη “head” και “body”, η πηγή προέλευσης (source) απλά περιγράφει την προέλευση των δεδομένων, το μέτρο ποιότητας (measure) περιέχει τα μέτρα ποιότητας του κανόνα (τα οποία συχνά διαφέρουν ανά εφαρμογή) και η έκφραση (expression) περιγράφει τη σχέση των κανόνων με τα δεδομένα από τα οποία προήλθε.

Πρέπει να τονιστεί ότι για κάθε είδος προτύπου (pattern type) θα πρέπει να υπάρχει ξεχωριστό XMLSchema. Το κάθε όμως σχήμα θα πρέπει να είναι αρκετά γενικό ώστε να μπορεί να περιγράψει όλα τα δυνατά πρότυπα του τύπου που εκφράζει. Ή α πρέπει δηλαδή το “association\_rule.xsd” να μπορεί να περιγράψει όλα τα πιθανά πρότυπα κανόνων συσχέτισης. Γι αυτό και στην παρούσα υλοποίηση το σχήμα σχεδιάστηκε ώστε να μπορεί να δεχθεί πρότυπα κανόνων συσχέτισης με διαφορετικό αριθμό μέτρων ποιότητας, γνωρισμάτων, τιμών και διαφορετικής δομής.

Πιο συγκεκριμένα, στο αντίστοιχο xsd (association\_rule.xsd), καθένα από τα “head”, “body”, “measure” χωρίζονται σε «προτάσεις» (clauses) για να εξασφαλιστεί η ελευθερία του κάθε χρήστη στο να εισάγει το δικό του τύπο κανόνα συσχέτισης χωρίς περιορισμό. Ο περιορισμός υπάρχει μόνο στη δομή, που πρέπει να είναι ίδια για κάθε πρότυπο και στο ότι ο κανόνας συσχέτισης αποτελείται από “head” και “body”.

Πρέπει να ακολουθείται ένα σχήμα προκειμένου να είναι αποτελεσματικά τα ερωτήματα (queries) πάνω στα XML έγγραφα.

Επιπλέον, το κάθε πρότυπο που ανήκει σε ένα τύπο προτύπου αναγνωρίζεται μοναδικά από ένα γνώρισμα “id” που περιέχει. Ένα ακόμα γνώρισμα προσδιορίζει τον τύπο προτύπου στον οποίο ανήκει. Από το “association\_rule.xsd” φαίνεται ότι κάθε XML έγγραφο μπορεί να περιέχει περισσότερα του ενός πρότυπα του ίδιου τύπου.

Παρακάτω, στην Εικόνα 9 απεικονίζεται σχηματικά το “association\_rule.xsd”. Το σχήμα παράχθηκε από το XMLSpy.

Αντίστοιχα θα πρέπει να δημιουργείται ένα διαφορετικό XMLSchema για κάθε τύπο προτύπου που θέλουμε να αποθηκεύσουμε (clusters κλπ). Δεν υπάρχει μόνο ένας τρόπος δημιουργίας ενός XML σχήματος για κάποιο τύπο προτύπου, αλλά πρέπει αυτό να είναι αφενός αρκετά γενικό και αφετέρου να προβάλει τις ιδιαιτερότητες του τύπου που περιγράφει.

```

<?xml version="1.0" encoding="UTF-8"?>
<!--W3C Schema generated by XMLSPY v5 rel. 2 U (http://www.xmlspy.com)-->
<xs:schema xmlns="http://www.w3.org/2001/XMLSchema"
  xmlns:xi="http://www.w3.org/2001/XMLSchema-instance"
  elementFormDefault="qualified">
  <xs:element name="attrib_name" type="xs:string"/>
  <xs:element name="attrib_value" type="xs:string"/>
  <xs:element name="body">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="s_clause" maxOccurs="unbounded"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
  <xs:element name="expression" type="xs:string"/>
  <xs:element name="head">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="s_clause" maxOccurs="unbounded"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
  <xs:element name="m_clause">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="measure_name"/>
        <xs:element ref="measure_value"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
  <xs:element name="measure">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="m_clause" maxOccurs="unbounded"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
  <xs:element name="measure_name" type="xs:string"/>
  <xs:element name="measure_value" type="xs:string"/>
  <xs:element name="name" type="xs:string"/>
  <xs:element name="assoc_rules" type="patternsType"/>
  <xs:element name="s_clause">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="attrib_name"/>
        <xs:element ref="attrib_value"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
  <xs:element name="source" type="xs:string"/>
  <xs:element name="structure">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="head"/>
        <xs:element ref="body"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
  <xs:complexType name="patternType">
    <xs:attribute name="ptype" type="xs:string" use="required"/>
    <xs:attribute name="description" type="xs:string" use="optional"/>
  </xs:complexType>

```

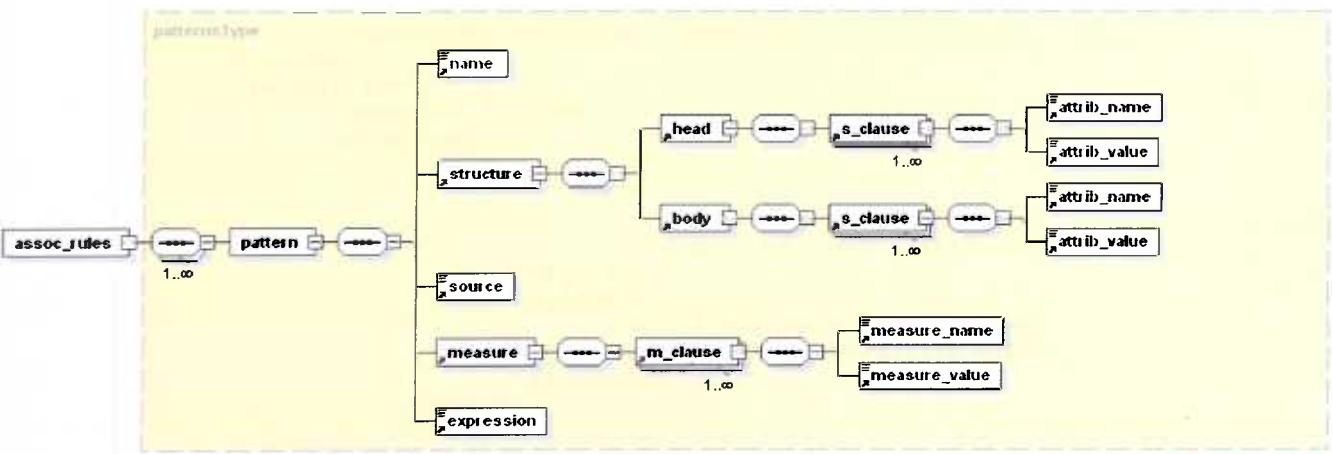
```

<xss:complexType name="patternsType">
    <xss:complexContent>
        <xss:extension base="patternType">
            <xss:sequence maxOccurs="unbounded">
                <xss:element name="pattern">
                    <xss:complexType>
                        <xss:sequence>
                            <xss:element ref="name"/>
                            <xss:element
                                ref="structure"/>
                            <xss:element ref="source"/>
                            <xss:element ref="measure"/>
                            <xss:element
                                ref="expression"/>
                        </xss:sequence>
                        <xss:attribute name="id"
                            type="xs:int" use="required"/>
                    </xss:complexType>
                </xss:element>
            </xss:sequence>
        </xss:extension>
    </xss:complexContent>

```

**Εικόνα 8 association\_rule.xsd**

Σχήμα (σε XMLSchema) για την μορφή του τύπου (pattern type) κανόνα συσχέτισης



Generated with XMLSpy Schema Editor <http://www.xmlspy.com>

**Εικόνα 9 Σχηματική απεικόνιση του association\_rule.xsd**

```

<assoc_rules ptype="association_rule" description="some sample association rule
patterns" :mlns: :si="http://www.w3.org/2001/XMLSchema-instance"
:si:noNamespaceSchemaLocation="association_rule.xsd">
    <pattern id="1">
        <name>rule 1</name>
        <structure>
            <head>
                <s_clause>
                    <attrib_name>buys</attrib_name>
                    <attrib_value>scarf</attrib_value>
                </s_clause>
                <s_clause>
                    <attrib_name>buys</attrib_name>
                    <attrib_value>cap</attrib_value>
                </s_clause>
            </head>
            <body>
                <s_clause>
                    <attrib_name>buys</attrib_name>
                    <attrib_value>gloves</attrib_value>
                </s_clause>
            </body>
        </structure>
        <source>SELECT * FROM orders</source>
        <measure>
            <m_clause>
                <measure_name>support</measure_name>
                <measure_value>0.35</measure_value>
            </m_clause>
            <m_clause>
                <measure_name>confidence</measure_name>
                <measure_value>0.75</measure_value>
            </m_clause>
        </measure>
        <expression>{buys="hat",buys="cap",buys="gloves"}</expression>
    </pattern>
    <pattern id="2">
        <name>rule 1</name>
        <structure>
            <head>
                <s_clause>
                    <attrib_name>buys</attrib_name>
                    <attrib_value>scarf</attrib_value>
                </s_clause>
                <s_clause>
                    <attrib_name>buys</attrib_name>
                    <attrib_value>cap</attrib_value>
                </s_clause>
            </head>
            <body>
                <s_clause>
                    <attrib_name>buys</attrib_name>
                    <attrib_value>gloves</attrib_value>
                </s_clause>
            </body>
        </structure>
        <source>SELECT * FROM orders</source>

```

```

<measure>
    <m_clause>
        <measure_name>support</measure_name>
        <measure_value>0.35</measure_value>
    </m_clause>
    <m_clause>
        <measure_name>confidence</measure_name>
        <measure_value>0.75</measure_value>
    </m_clause>
</measure>
<expression>{buys="hat",buys="cap",buys="gloves"}</expression>
</pattern>
<pattern id="3">
    <name>rule 3</name>
    <structure>
        <head>
            <s_clause>
                <attrib_name>buys</attrib_name>
                <attrib_value>mobile</attrib_value>
            </s_clause>
        </head>
        <body>
            <s_clause>
                <attrib_name>buys</attrib_name>
                <attrib_value>leather case</attrib_value>
            </s_clause>
        </body>
    </structure>
    <source>SELECT * FROM orders</source>
    <measure>
        <m_clause>
            <measure_name>lift</measure_name>
            <measure_value>1.9</measure_value>
        </m_clause>
        <m_clause>
            <measure_name>leverage</measure_name>
            <measure_value>1.2</measure_value>
        </m_clause>
    </measure>
    <expression>{buys="mobile",buys="leather case"}</expression>
</pattern>
</assoc_rules>

```

**Εικόνα 10 pattern\_association\_rule.xml**  
**XML έγγραφο τύπου κανόνα συσχέτισης (association rule)**

Η ανάλυση των συστάδων (cluster analysis) είναι οι τεχνικές που οδηγούν την κατάταξη των δεδομένων σε ομάδες. Βασικό χαρακτηριστικό των περισσοτέρων τεχνικών είναι ότι κάθε ομάδα πρέπει να περιέχει τουλάχιστον ένα στοιχείο και κάθε στοιχείο θα πρέπει να βρίσκεται σε μια και μόνο ομάδα [16]. Η πιο συνηθισμένη αναπαράσταση συστάδων χρησιμοποιεί ένα σημείο ως κέντρο και μια ακτίνα, σχηματίζοντας έναν νοητό κύκλο. Τα σημεία των δεδομένων που βρίσκονται μέσα στην απόσταση αυτή, μέσα στον κύκλο δηλαδή, ανήκουν στην συγκεκριμένη συστάδα (cluster). Με βάση των απλό αυτό ορισμό μπορεί να σχεδιαστεί και ένα XMLSchema όπως φαίνεται παρακάτω.

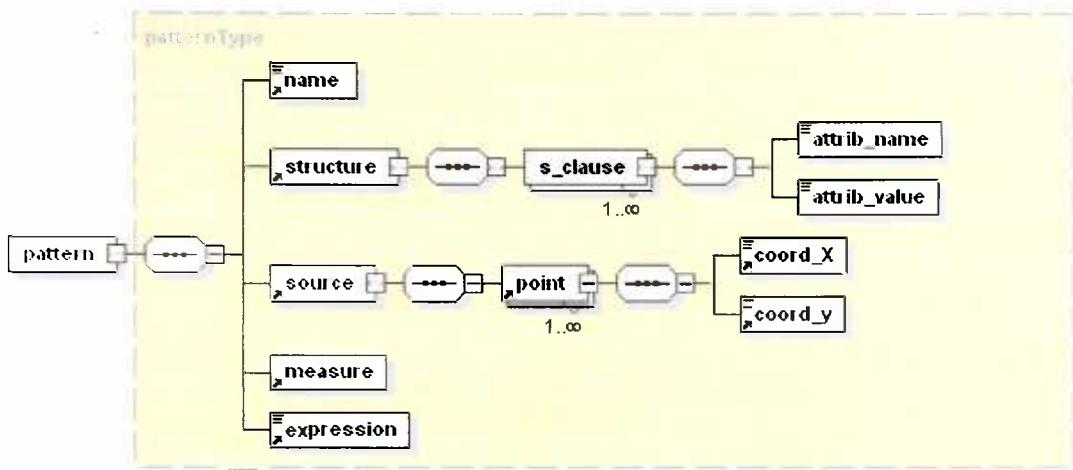
```

<?xml version="1.0" encoding="UTF-8"?>
<!--W3C Schema generated by XMLSPY v5 rel. 2 U (http://www.xmlspy.com)-->
<x: schema xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
    <x:element name="center_x_coord" type="xs:decimal"/>
    <x:element name="center_y_coord" type="xs:decimal"/>
    <x:element name="coord_X" type="xs:decimal"/>
    <x:element name="coord_y" type="xs:decimal"/>
    <x:element name="expression" type="xs:string"/>
    <x:element name="measure">
        <x:complexType/>
    </x:element>
    <x:element name="name" type="xs:string"/>
    <x:element name="pattern">
        <x:complexType>
            <x:sequence>
                <x:element ref="name"/>
                <x:element ref="structure"/>
                <x:element ref="source"/>
                <x:element ref="measure"/>
                <x:element ref="expression"/>
            </x:sequence>
            <x:attribute name="ptype" type="xs:string" use="required"/>
            <x:attribute name="id" type="xs:int" use="required"/>
        </x:complexType>
    </x:element>
    <x:element name="point">
        <x:complexType>
            <x:sequence>
                <x:element ref="coord_X"/>
                <x:element ref="coord_y"/>
            </x:sequence>
        </x:complexType>
    </x:element>
    <x:element name="radius" type="xs:decimal"/>
    <x:element name="source">
        <x:complexType>
            <x:sequence>
                <x:element ref="point" maxOccurs="unbounded"/>
            </x:sequence>
        </x:complexType>
    </x:element>
    <x:element name="structure">
        <x:complexType>
            <x:sequence>
                <x:element name="s_clause" maxOccurs="unbounded">
                    <x:complexType>
                        <x:sequence>
                            <x:element name="attrib_name" type="xs:string"/>
                            <x:element name="attrib_value" type="xs:string"/>
                        </x:sequence>
                    </x:complexType>
                </x:sequence>
            </x:complexType>
        </x:element>
    </x:complexType>

```

### Εικόνα 11 cluster.xsd

Σχήμα (σε XMLSchema) για την μορφή του τύπου (pattern type)



Generated with XMLSpy Schema Editor [www.xmlspy.com](http://www.xmlspy.com)

**Εικόνα 12** Σχηματική απεικόνιση του cluster.xsd

Αν και το σχήμα αυτό αναπαριστά ένα γνωστό και συχνά εμφανιζόμενο τύπο συστάδων, σίγουρα δεν είναι αρκετά γενικό για να περιγραφούν όλοι οι πιθανοί τύποι συστάδων. Ο βαθμός γενικότητας του σχήματος καθορίζεται από τον κατασκευαστή του σχήματος και έχει σχέση και με την εκάστοτε εφαρμογή.

Ακολουθεί ένα παράδειγμα εγγράφου δεδομένων που στηρίζεται στο σχήμα “cluster.xsd”.

```

<pattern ptype="cluster" id="2" xmlns:xsi="http://www.w3.org/2001/XMLSchema-
instance"
xsi:noNamespaceSchemaLocation="cluster.xsd">
    <name>sales cluster</name>
    <structure>
        <s_clause>
            <attrib_name>radius</attrib_name>
            <attrib_value>5</attrib_value>
        </s_clause>
        <s_clause>
            <attrib_name>center_coord_x</attrib_name>
            <attrib_value>2</attrib_value>
        </s_clause>
        <s_clause>
            <attrib_name>center_coord_y</attrib_name>
            <attrib_value>3</attrib_value>
        </s_clause>
    </structure>
    <source>
        <point>
            <coord_X>0</coord_X>
            <coord_y>1</coord_y>
        </point>
        <point>
            <coord_X>2</coord_X>
            <coord_y>3</coord_y>
        </point>
        <point>
            <coord_X>1</coord_X>
            <coord_y>4</coord_y>
        </point>
    </source>
    <measure/>
    <expression>{f: (x-cx)^2+(y-cy)^2<= radius^2}</expression>
</pattern>

```

### Εικόνα 13 cluster.xml XML έγγραφο τύπου συστάδας (cluster)

Βασική έννοια στο μοντέλο του PBMS για το σύστημα διαχείρισης προτύπων αποτελεί η **κλάση**. Μια κλάση, όπως προαναφέρθηκε αντιπροσωπεύει ένα σύνολο προτύπων ενός συγκεκριμένου τύπου. Μπορεί για παράδειγμα να οριστεί μια κλάση κανόνων συσχέτισης που αφορούν σε κάποια συγκεκριμένα δεδομένα (για παράδειγμα τις πωλήσεις ενός καταστήματος για μια χρονική περίοδο). Μια κλάση εκτός από το αναγνωριστικό της, περιέχει και τα αναγνωριστικά των προτύπων που περιέχει. Κάθε πρότυπο επιπλέον πρέπει να ανήκει σε τουλάχιστον μια κλάση. Με βάση αυτές τις προϋποθέσεις το παρακάτω XMLSchema περιγράφει την κλάση η οποία έχει ως γνωρίσματα το όνομά της (name), το αναγνωριστικό της (id), τον τύπο προτύπων που περιέχει (ptype) και μια περιγραφή και σαν πεδία (elements) τα αναγνωριστικά των προτύπων (pids).

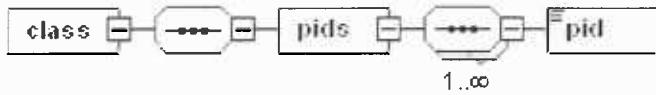
```

<?xml version="1.0" encoding="UTF-8"?>
<xss:schema xmlns:xss="http://www.w3.org/2001/XMLSchema"
elementFormDefault="qualified" attributeFormDefault="unqualified">
<xss:element name="class">
    <xss:complexType>
        <xss:sequence>
            <xss:element name="pids">
                <xss:complexType>
                    <xss:sequence maxOccurs="unbounded">
                        <xss:element name="pid"
type="xss:string"/>
                    </xss:sequence>
                </xss:complexType>
            </xss:element>
        </xss:sequence>
        <xss:attribute name="id" type="xss:string" use="required"/>
        <xss:attribute name="name" type="xss:string" use="required"/>
        <xss:attribute name="ptype" type="xss:string" use="required"/>
        <xss:attribute name="description" type="xss:string"
use="optional"/>
    </xss:complexType>
</xss:element>
</xss:schema>

```

#### **Εικόνα 14 class.xsd**

Σχήμα (σε XML-Schema) για την μορφή της κλάσης (class)



Generated with XMLSpy Schema Editor [www.xmlspy.com](http://www.xmlspy.com)

**Εικόνα 15 Σχηματική απεικόνιση του class.xsd**

```

<?xml version="1.0" encoding="UTF-8"?>
<class xmlns:nsi="http://www.w3.org/2001/XMLSchema-instance"
nsi:noNamespaceSchemaLocation="class.xsd" id="1" name="class1"
ptype="association_rule" description="this is a class containing association rules
for application X.">
  <pids>
    <pid>14</pid>
    <pid>12</pid>
    <pid>11</pid>
    <pid>16</pid>
    <pid>15</pid>
    <pid>13</pid>
  </pids>
</class>

```

**Εικόνα 16 class1.xml XML έγγραφο κλάσης (παράδειγμα).**

Είναι φανερό ότι τα πρότυπα που ανήκουν σε μια κλάση κάποιου τύπου, για παράδειγμα κανόνων συσχέτισης, πρέπει και αυτά να ανήκουν στον ίδιο τύπο. Για αυτό το λόγο θα ήταν επιθυμητό να ελέγχεται ότι το pid (το αναγνωριστικό του προτύπου) που υπάρχει στην κλάση, αντιστοιχεί σε πρότυπο του ίδιου τύπου. Η XML γενικά δεν παρέχει ευκολίες στη δήλωση περιορισμών ιδιαίτερα όταν αυτοί αφορούν σε διαφορετικά XML έγγραφα.

### 3.3 Υλοποίηση XML μοντέλου σε ORACLE 9i

Η βάση προτύπων όπως παρουσιάστηκε θα πρέπει να υλοποιηθεί και να «δοκιμαστεί» σε ένα σύστημα βάσεων δεδομένων που να υποστηρίζει διαχείριση XML σχημάτων και εγγράφων. Τα συστήματα αυτά χωρίζονται σε δύο κατηγορίες. Τα παραδοσιακά σχεσιακά ή αντικειμενο-σχεσιακά που πλέον υποστηρίζουν XML και τα αμιγώς XML συστήματα που έχουν δημιουργηθεί αποκλειστικά για τη διαχείριση XML εγγράφων. Τα πιο γνωστά συστήματα της πρώτης κατηγορίας είναι αυτά της IBM (υποστήριξη XML στο DB2), της ORACLE και της Microsoft (υποστήριξη XML στο SQL server 2000). Στη δεύτερη κατηγορία ανήκουν λιγότερα συστήματα από τα οποία τα πιο δημοφιλή είναι το TAMINO και το eXist.

Η υλοποίηση της XML βάσης προτύπων έγινε στην ORACLE 9i για τους εξής λόγους:

- Παρέχει γρήγορη αποθήκευση και ανάκτηση XML εγγράφων
- Παρέχει εύκολη ολοκλήρωση των διαφόρων εφαρμογών που είναι εγκατεστημένες στη βάση.
- Ενοποιεί τα δεδομένα που βρίσκονται σε σχεσιακούς πίνακες με τα XML έγγραφα με τη χρήση του XMLType και της SQL/XML.

Οι δυνατότητες της ORACLE ως βάσης δεδομένων είναι πάρα πολλές και η υποστήριξη της XML είναι πολύ αξιόπιστη. Περισσότερα για τις δυνατότητες της ORACLE στη διαχείριση XML δεδομένων μπορούν να βρεθούν στην [14].

Ένας ακόμα λόγος για την επιλογή της ORACLE για τη βάση προτύπων είναι ότι θα είναι δυνατή η σύγκριση με την αντικειμενο-σχεσιακή βάση προτύπων που έχει ήδη υλοποιηθεί [11].

Επιπλέον, όπως αναφέρθηκε, σε ένα σύστημα διαχείρισης προτύπων τα αρχικά δεδομένα (raw data) και τα πρότυπα που ανακαλύπτονται σε αυτά συνυπάρχουν έτσι ώστε να μπορεί να γίνεται η άμεση αντιστοίχιση μεταξύ τους. Στην ORACLE παρέχεται η δυνατότητα αυτή της συνύπαρξης δηλαδή των δεδομένων και η ενοποίησή τους με μηχανισμούς μετατροπής. Έχοντας δηλαδή τα δεδομένα σε σχεσιακούς πίνακες υπάρχει η δυνατότητα μετατροπής τους σε XML μορφή και αποθήκευσή τους σε πίνακες τύπου XML. Παράλληλα εκτός του ότι υπάρχει η ειδικά τροποποιημένη SQL για την υποστήριξη ερωτήσεων σε XML έγγραφα, χρησιμοποιούνται τεχνικές ευρετηριοποίησης (indexing) ειδικά προσαρμοσμένες στη δομή της XML. Αυτό έχει ως αποτέλεσμα την αύξηση της αποτελεσματικότητας και τη μείωση χρόνου αποθήκευσης και ανάκτησης.

Για τους παραπάνω λόγους επιλέχθηκε ένα σύστημα που δεν είναι αμιγώς XML. Μια όμως ενδιαφέρουσα μελλοντική έρευνα θα ήταν η αντίστοιχη υλοποίηση της XML βάσης προτύπων σε ένα αμιγώς XML σύστημα όπως το TAMINO.

Λαμβάνοντας υπόψιν αυτή τη μελλοντική πιθανότητα, τα XML σχήματα που δημιουργήθηκαν και ο τρόπος υλοποίησής τους στην ORACLE 9i για τη βάση προτύπων σχεδιάστηκαν κατάλληλα. Στη συνέχεια αυτό θα γίνει πιο σαφές εξετάζοντας τις εναλλακτικές λύσεις υλοποίησης στην ORACLE.

### 3.3.1 Τρόποι αποθήκευσης XML δεδομένων στην ORACLE 9i

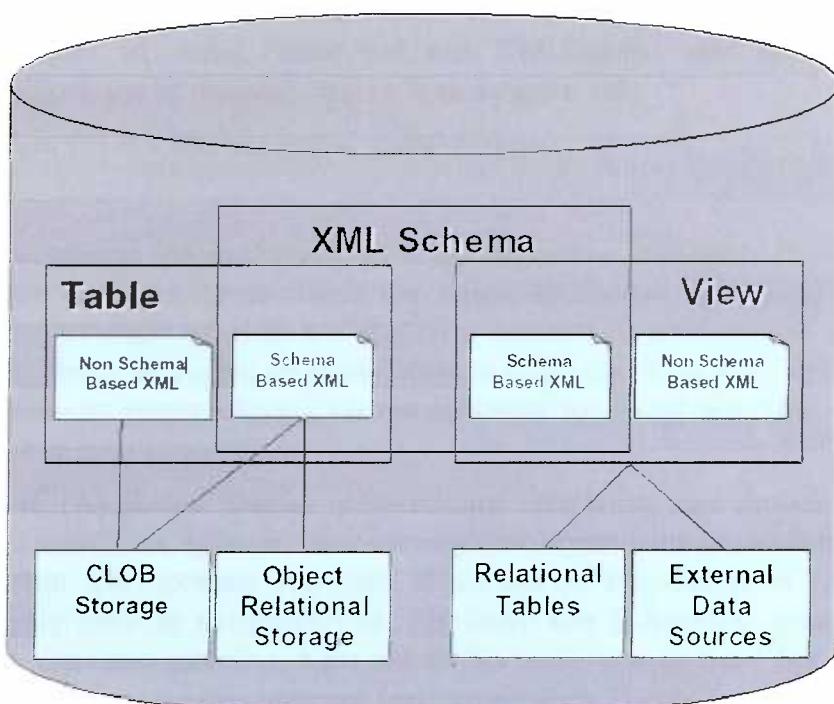
Για την αποθήκευση των XML εγγράφων η ORACLE δημιούργησε ένα νέο τύπο δεδομένων το **XMLType**. Ο XMLType είναι ένας «καθαρός» τύπος όπως για παράδειγμα είναι ο τύπος DATE. Με τη χρήση αυτού του τύπου είναι δυνατή η δήλωση σχεσιακών πινάκων με στήλες που περιέχουν XML δεδομένα με τον ίδιο τρόπο που δηλώνεται ότι μια στήλη περιέχει δεδομένα τύπου DATE. Επίσης ο τύπος αυτός μπορεί να χρησιμοποιηθεί στη δήλωση PL/SQL μεταβλητών ή συναρτήσεων και διαδικασιών. Σημαντική είναι η δυνατότητα ορισμού πινάκων αποκλειστικά τύπου XMLType και ο ορισμός όψεων (views) που περιέχουν στοιχεία τόσο από πίνακες με απλά δεδομένα όσο και δεδομένα τύπου XMLType.

Εξ ορισμού μια στήλη ή ένας πίνακας τύπου XMLType μπορεί να περιέχει οποιοδήποτε καλώς ορισμένο XML έγγραφο. Το περιεχόμενο του XML εγγράφου αποθηκεύεται με τη χρήση του τύπου *CLOB* (*Character Large Object*) επιτρέποντας μεγάλη ευελιξία στο σχήμα της δομής του XML εγγράφου και καλή απόδοση στην εισαγωγή και ανάκτηση.

Όμως ένας πίνακας ή στήλη μπορεί να περιορίζεται από ένα XMLSchema. Αυτό έχει αρκετά πλεονεκτήματα [14].

- Θα είναι βέβαιο ότι μόνο έγκυρα έγγραφα θα εισάγονται στη βάση.
- Από τη στιγμή που τα δεδομένα (τα XML έγγραφα) θα συμφωνούν με κάποιο γνωστό σχήμα, η πληροφορία που περιέχεται στο XMLSchema μπορεί να χρησιμοποιηθεί για να γίνεται πιο αποτελεσματική επεξεργασία των ερωτημάτων και της ενημέρωσης των XML εγγράφων.
- Περιορίζοντας τα δεδομένα με ένα XMLSchema δίνεται η δυνατότητα του διαχωρισμού των διαφόρων τμημάτων του εγγράφου και την αποθήκευσή τους σε SQL αντικείμενα αντί για την απλή αποθήκευσή τους σαν κείμενο σε ένα CLOB.

Το παρακάτω σχήμα απεικονίζει τις διαφορετικές επιλογές που υπάρχουν σε σχέση με την αποθήκευση και αναπαράσταση XML δεδομένων στην ORACLE 9i [14].



**Εικόνα 17** Επιλογές αποθήκευσης XMLType

Η χρήση κάποιου XMLSchema είναι σαφώς προτιμότερη για τη βάση προτύπων για τους λόγους που προαναφέρθηκαν, οπότε αποκλείεται η αποθήκευση της XML σαν κείμενο τύπου CLOB. Εξάλλου τα θέματα της απόδοσης στην ανάκτηση και την ενημέρωση είναι βασικής σημασίας στη βάση προτύπων. Συγκεντρωτικά τα πλεονεκτήματα και μειονεκτήματα της αποθήκευσης σε αδόμητη μορφή (σαν CLOB) και της δομημένης (με XMLSchema) φαίνονται στο παράρτημα A.

Για τη βάση προτύπων θα πρέπει να κατασκευαστεί ένα XMLSchema για κάθε τύπο προτύπου (pattern type) και ένα XMLSchema για τις κλάσεις. Τα έγγραφα που θα

καταχωρίζονται κάτω από κάθε XMLSchema για τους τύπους προτύπων θα αποτελούν πρότυπα του τύπου αυτού. Σε ένα XML έγγραφο μπορεί να υπάρχουν περισσότερα του ενός πρότυπα. Κάθε πρότυπο πρέπει να ξεχωρίζει μοναδικά με το αναγνωριστικό του, να έχει ένα όνομα και να ανήκει σε μια κλάση κάποιου συγκεκριμένου τύπου προτύπου.

Κάθε κλάση θα αντιστοιχεί σε ένα XML έγγραφο που θα περιέχει το όνομά της, το αναγνωριστικό της, τον τύπο προτύπου που αναπαριστά και φυσικά τα αναγνωριστικά των προτύπων που περιέχει. Όλα τα παραπάνω στοιχεία υπαγορεύουν τους περιορισμούς που πρέπει να υπάρχουν στη βάση για την αποφυγή εισαγωγής λανθασμένων δεδομένων. Οι εναλλακτικές λύσεις για την κατασκευή αυτής της βάσης στην ORACLE είναι οι εξής:

- A. Χρήση πινάκων με στήλες απλού τύπου (πχ. INTEGER ή VARCHAR) για την αποθήκευση των αναγνωριστικών και των ονομάτων και τύπου XMLType για την αποθήκευση των XML δεδομένων.
- B. Χρήση μόνο πινάκων τύπου XMLType για την αποθήκευση των XML εγγράφων τα οποία, όπως και στα XMLSchema που περιγράφθηκαν, περιέχουν και τα αναγνωριστικά και τα ονόματά τους.
- C. Χρήση και των δύο παραπάνω τύπων πινάκων.

Για την υλοποίηση οποιασδήποτε από τις παραπάνω επιλογές, θα πρέπει τα XMLSchema που χρησιμοποιούνται να «εγκατασταθούν» στη βάση (Register Schema). Στη συνέχεια στον ορισμό στηλών ή πινάκων θα επιλέγεται το αντίστοιχο σχήμα με το οποίο θα πρέπει να συμφωνούν τα δεδομένα. Ετσι κατά την εισαγωγή των δεδομένων θα γίνεται έλεγχος για την ορθότητα της δομής των XML εγγράφων σύμφωνα με το αντίστοιχο σχήμα.

Στην επιλογή (A) δίνεται εύκολα η δυνατότητα συσχέτισης των πινάκων με ξένα κλειδιά έτσι ώστε κάθε πρότυπο να αντιστοιχεί τουλάχιστον σε μια κλάση, η κλάση να αποτελείται από πρότυπα του ίδιου τύπου και να πληρούνται οι περιορισμοί μοναδικότητας (από τα αναγνωριστικά, ids) όλων των δεδομένων, όπως ακριβώς γίνεται στο σχεσιακό μοντέλο. Κάτι τέτοιο δεν είναι εφικτό όταν δεν υπάρχουν ξεχωριστά πεδία για τα αναγνωριστικά (στις περιπτώσεις Β και Γ) και η πληροφορία αυτή βρίσκεται μέσα στα XML έγγραφα. Είναι πολύ δύσκολος έτσι ο έλεγχος των κλειδιών και ο έλεγχος των περιορισμών. Μετά και από δοκιμές και πειραματισμό, αποφασίστηκε η επιλογή (B). Ο λόγος είναι ότι παρότι καθίσταται δύσκολος ο έλεγχος ορθότητας των προς εισαγωγή δεδομένων, το σύστημα θα είναι καθαρά XML, αφού θα αποτελείται μόνο από XML έγγραφα και όχι από πεδία του παραδοσιακού σχεσιακού τύπου. Με τον τρόπο αυτό θα είναι εύκολη η μεταφορά του μοντέλου και των δεδομένων σε άλλο σύστημα, πιθανώς αμιγώς XML, για δοκιμή της αποτελεσματικότητας και ευκολία στη σύγκριση των ερωτημάτων στα δύο συστήματα.

Καταλήγοντας λοιπόν στα σχήματα που αναφέρθηκαν (παράγραφος 3.2) και στην επιλογή (B) για την αποθήκευση των σχημάτων και δεδομένων στην ORACLE,

δημιουργήθηκαν οι κατάλληλοι πίνακες (ένας για τις κλάσεις, που ελέγχεται από το XML Schema “class.xsd” και ένας για κάθε τύπο προτύπου που ελέγχεται από το “association\_rule.xsd” για παράδειγμα.

### 3.3.2 Δοκιμαστικά δεδομένα. Εργαλεία που χρησιμοποιήθηκαν

Για την συμπλήρωση της XML βάσης προτύπων χρησιμοποιήθηκε ένα μικρό υποσύνολο (50 εγγραφές από τις 814 που παραχωρήθηκαν) των δεδομένων μιας ιατρικής βάσης δεδομένων που παραχώρησε το Οικονομικό Πανεπιστήμιο Αθηνών (ΟΠΑ). Τα δεδομένα αυτά είναι κανόνες συσχέτισης, αποτέλεσμα διαδικασίας εξόρυξης γνώστης και παραχωρήθηκαν σε μορφή αρχείου Microsoft Excel και η μορφή τους φαίνεται στην Εικόνα 18.

ID	STRUCTURE: HEAD , BODY	SOURCE	MEASURE: LIST_OF(Coverage, Strength, Lift, Leverage)	EXPRESSION: HEAD, BODY
1	head={SPEC='st'},body={ORG='sa-'}	SELECT * FROM auebRawData	Coverage=0.002(31),Strength=1.000,Lift=54.20,Leverage=0.0 {SPEC='st',ORG='sa-'}	
2	head={HOSP='avm',SERV='out'},body={ORG SELECT * FROM auebRawData		Coverage=0.002(29),Strength=1.000,Lift=54.20,Leverage=0.0 {HOSP='avm',SERV='out',ORG='sa-'}	
3	head={HOSP='avm',SPEC='st'},body={ORG =SELECT * FROM auebRawData		Coverage=0.001(19),Strength=1.000,Lift=54.20,Leverage=0.0 {HOSP='avm',SPEC='st',ORG='sa-'}	
4	head={HOSP='avm',SPEC='st'},body={ORG =SELECT * FROM auebRawData		Coverage=0.001(16),Strength=1.000,Lift=54.20,Leverage=0.0 {HOSP='avm',SPEC='st',ORG='sa-'}	
5	head={SPEC='st'},body={ORG='sa-'}	SELECT * FROM auebRawData	Coverage=0.002(28),Strength=0.964,Lift=52.26,Leverage=0.0 {SPEC='st',ORG='sa-'}	
6	head={SPEC='st'},body={ORG='sa-'}	SELECT * FROM auebRawData	Coverage=0.002(30),Strength=0.933,Lift=50.59,Leverage=0.0 {SPEC='st',ORG='sa-'}	
7	head={HOSP='avm',SPEC='st'},body={ORG =SELECT * FROM auebRawData		Coverage=0.001(22),Strength=0.909,Lift=49.27,Leverage=0.0 {HOSP='avm',SPEC='st',ORG='sa-'}	
8	head={HOSP='avm',SERV='med'},body={OR SELECT * FROM auebRawData		Coverage=0.003(51),Strength=0.902,Lift=48.89,Leverage=0.0 {HOSP='avm',SERV='med',ORG='sa-'}	
9	head={HOSP='avm'},body={ORG='sa-'}	SELECT * FROM auebRawData	Coverage=0.009(130),Strength=0.900,Lift=48.78,Leverage=0 {HOSP='avm',ORG='sa-'}	
10	head={SERV='ped',SPEC='st'},body={ORG =SELECT * FROM auebRawData		Coverage=0.001(19),Strength=0.895,Lift=48.50,Leverage=0.0 {SERV='ped',SPEC='st',ORG='sa-'}	
11	head={HOSP='avm',SERV='ped'},body={ORG SELECT * FROM auebRawData		Coverage=0.003(47),Strength=0.894,Lift=48.43,Leverage=0.0 {HOSP='avm',SERV='ped',ORG='sa-'}	
12	head={SPEC='st'},body={ORG='ssp'}	SELECT * FROM auebRawData	Coverage=0.021(309),Strength=0.052,Lift=48.06,Leverage=0 {SPEC='st',ORG='ssp'}	
13	head={SERV='out'},body={ORG='sa-'}	SELECT * FROM auebRawData	Coverage=0.003(43),Strength=0.884,Lift=47.90,Leverage=0.0 {SERV='out',ORG='sa-'}	
14	head={SERV='med'},body={ORG='sa-'}	SELECT * FROM auebRawData	Coverage=0.008(122),Strength=0.836,Lift=45.32,Leverage=0 {SERV='med',ORG='sa-'}	
15	head={SPEC='st'},body={ORG='sa-'}	SELECT * FROM auebRawData	Coverage=0.002(24),Strength=0.833,Lift=45.17,Leverage=0.0 {SPEC='st',ORG='sa-'}	
16	head={SERV='med',SPEC='st'},body={ORG =SELECT * FROM auebRawData		Coverage=0.002(27),Strength=0.815,Lift=44.16,Leverage=0.0 {SERV='med',SPEC='st',ORG='sa-'}	
17	head={HOSP='tgh',SERV='med'},body={ORG SELECT * FROM auebRawData		Coverage=0.005(71),Strength=0.789,Lift=42.75,Leverage=0.0 {HOSP='tgh',SERV='med',ORG='sa-'}	
18	head={SPEC='st'},body={ORG='sa-'}	SELECT * FROM auebRawData	Coverage=0.021(309),Strength=0.783,Lift=42.45,Leverage=0 {SPEC='st',ORG='sa-'}	
19	head={SERV='ped'},body={ORG='sa-'}	SELECT * FROM auebRawData	Coverage=0.009(138),Strength=0.725,Lift=39.28,Leverage=0 {SERV='ped',ORG='sa-'}	
20	head={SPEC='st'},body={ORG='sa-'}	SELECT * FROM auebRawData	Coverage=0.003(47),Strength=0.723,Lift=39.21,Leverage=0.0 {SPEC='st',ORG='sa-'}	
21	head={HOSP='tgh'},body={ORG='sa-'}	SELECT * FROM auebRawData	Coverage=0.012(179),Strength=0.698,Lift=37.85,Leverage=0 {HOSP='tgh',ORG='sa-'}	
22	head={HOSP='tgh',SERV='ped'},body={ORG SELECT * FROM auebRawData		Coverage=0.006(91),Strength=0.637,Lift=34.55,Leverage=0.0 {HOSP='tgh',SERV='ped',ORG='sa-'}	
23	head={HOSP='tgh',SPEC='st'},body={ORG =SELECT * FROM auebRawData		Coverage=0.002(34),Strength=0.618,Lift=33.48,Leverage=0.0 {HOSP='tgh',SPEC='st',ORG='sa-'}	
24	head={HOSP='avm',SPEC='th'},body={ORG =SELECT * FROM auebRawData		Coverage=0.004(62),Strength=1.000,Lift=27.71,Leverage=0.0 {HOSP='avm',SPEC='th',ORG='sa-'}	

**Εικόνα 18** Υποσύνολο των δεδομένων που παραχωρήθηκαν από το ΟΠΑ. Κανόνες συσχέτισης από εξόρυξη σε βάση ιατρικών δεδομένων.

Τα δεδομένα αυτά χρειάστηκε να μετασχηματιστούν σε XML έγγραφα της μορφής που περιγράφει το XML Schema “association\_rule.xsd”. Για τη διαδικασία αυτή χρησιμοποιήθηκαν τα παρακάτω εργαλεία:

Microsoft Excel έκδοση XP, Microsoft Access έκδοση XP, XMLSpy 5.0 της Altova και UltraEdit-32 professional Text/HEX editor v.8.2.

Η διαδικασία περιελάμβανε το διαχωρισμό των στηλών των δεδομένων σε αυτές που ήταν απαραίτητες για το σχήμα, τη μετατροπή του αρχείου Excel σε XML στο XMLSpy μέσω της Microsoft Access και τη χρήση ενός πολύπλοκου query και την αντικατάσταση κάποιων πεδίων του XML εγγράφου ώστε να ταιριάζει ακριβώς στο σχήμα. Το αποτέλεσμα είναι η δημιουργία XML εγγράφων όπως αυτά που περιγράφονται στο 3.2 στην Εικόνα 8 και Εικόνα 10.

Η παραπάνω διαδικασία αν και σχετικά αυτοματοποιημένη στην τελική της μορφή, είναι σχετικά χρονοβόρα ειδικά όσο αυξάνεται ο όγκος των δεδομένων. Θα μπορούσε

δε να αποφευχθεί είτε αν τα αρχικά δεδομένα ήταν ήδη εγκατεστημένα στην ORACLE σε σχεσιακό μοντέλο είτε αν γινόταν η εξαγωγή τους κατευθείαν σε XML μορφή, σε περίπτωση που κάτι τέτοιο ήταν εφικτό από τη βάση που ήταν αρχικά αποθηκευμένα.

Σε ένα ολοκληρωμένο σύστημα διαχείρισης προτύπων τα πρότυπα θα πρέπει να συνυπάρχουν με τα αρχικά δεδομένα. Θα πρέπει να υπάρχει σύνδεση μεταξύ τους ώστε με τη χρήση των μερών “source” και “expression” του προτύπου να μπορούν να γίνουν ερωτήσεις σχετικές με τα αρχικά δεδομένα και φυσικά ενημέρωση τόσο των αρχικών δεδομένων όσο και των προτύπων που προέρχονται από αυτά.

Η υλοποίηση έγινε σε έναν Intel Pentium III στα 1000 MHz με 256 MB RAM, 20GB σκληρό δίσκο με λειτουργικό σύστημα Windows XP Professional.

### 3.3.3 Ερωτήματα (queries) στην XML βάση προτύπων

Βασικό κριτήριο για το αν είναι εφικτή η υλοποίηση μιας βάσης προτύπων σε XML εκτός από τη δημιουργία των σχημάτων είναι η δυνατότητα να εκφραστούν τα κατάλληλα ερωτήματα (queries) σε αυτή. Τα κατάλληλα ερωτήματα είναι αυτά που ένας χρήστης θα ήταν περισσότερο πιθανό να εκφράσει σε μια βάση προτύπων. Τα περισσότερα ερωτήματα που ακολουθούν έχουν βρεθεί από την [11] όπου αναφέρονται εκτενώς κάποια πιθανά ερωτήματα για μια βάση προτύπων. Στη συνέχεια παρατίθενται κάποια από αυτά (επιλέχθηκαν τα σημαντικότερα) σε απλή γλώσσα, η αντίστοιχη έκφραση σε SQL/XML και η απάντηση που επιστρέφει η βάση δεδομένων.

Η ανάκτηση από το κομμάτι της πηγής αρχικών δεδομένων του προτύπου (source schema) μπορεί να απαιτεί και αναζήτηση στα αρχικά δεδομένα. Τέτοιου είδους ερωτήσεις ονομάζονται cross-over ερωτήσεις. Για να απαντηθούν τα cross-over ερωτήματα, χρειάζεται πρόσβαση, απομακρυσμένη, στο σύστημα αποθήκευσης των αρχικών δεδομένων προκειμένου να εκτελέσουμε τα ερωτήματα αυτά απευθείας στα αρχικά δεδομένα (raw data). Η αρχιτεκτονική που προτείνεται επιτρέπει την απλή σύνδεση μεταξύ των αρχικών δεδομένων και των προτύπων.

Τα ερωτήματα ομαδοποιούνται ως εξής:

Ερωτήματα E1 και E4: Ερωτήματα επιλογής στο σύνολο των προτύπων.

Ερωτήματα E2, E3, E5 – E11: Ερωτήματα επιλογής σε σχέση με τον τύπο προτύπων. Επιλογή με βάση τα δομικά στοιχεία (δομή, πηγή, μέτρα ποιότητας κλπ) των προτύπων.

Ερωτήματα E12 – E14: Ερωτήματα επιλογής συνόλων προτύπων (UNION, INTERSECT κλπ.)

Ερωτήματα E15 – E17: Ερωτήματα συγκεντρωτικών συναρτήσεων (Max, Min, Average, Count κλπ)

Σημ. Χρησιμοποιείται η μορφή  $A.B.x$  για να δηλώσει την προσπέλαση του πεδίου  $x$  που βρίσκεται κάτω από το  $B$  το οποίο βρίσκεται κάτω από το  $A$ .

**E1)** Ανάκτηση των ονομάτων των προτύπων που ανήκουν στην κλάση class1

➤ Αφηρημένη μορφή:

```
SELECT A.name  
FROM Association_rules A, classes C  
WHERE C.name="class1"
```

➤ Υλοποίηση SQL/XML στη βάση προτύπων:

```
select distinct  
extractValue(value(y), '//pattern[@id=""]||extract(value(e), 'p  
id/text()')||'"]/name/text()' ) as pattern_name from  
assoc_rules y, classes x,  
TABLE(XMLsequence(extract(value(x), 'class[@name="class1"]//p  
ids/pid')))) e
```

The screenshot shows the Oracle SQL\*Plus Worksheet interface. The query entered in the worksheet is:

```
select distinct extractValue(value(y), '//pattern[@id=""]||extract(value(e), pid/text()'||'"]/name/text()' ) as  
x, TABLE(XMLsequence(extract(value(x), 'class[@name="class1"]//p  
ids/pid')))) e
```

The results are displayed in a table with the following data:

PATTERN_NAME
pattern11
pattern12
pattern13
pattern14
pattern15
pattern16
pattern17

7 rows selected.

**Εικόνα 19** Ερώτημα 1 (E1)

**E2)** Ανάκτηση των διαφορετικών ποιοτικών μέτρων (measures) που υπάρχουν στους κανόνες συσχέτισης.

➤ Αφηρημένη μορφή:

```
SELECT A.measure.name  
FROM Association_Rules A
```

➤ Υλοποίηση SQL/XML στη βάση προτύπων:

```
select distinct extractValue(value(r),  
'//m_clause/measure_name/text()') as measures from  
assoc_rules y, classes  
x, TABLE(XMLsequence(extract(value(x), '//pids/pid'))) e,  
TABLE(XMLsequence(extract(value(y), '//pattern[@id=""||extract  
value(e), "pid/text()"]||""]//m_clause'))) r;
```

The screenshot shows the Oracle SQL\*Plus Worksheet interface. The query window contains the XML query provided above. The results pane shows the output:

```
MEASURES  
Coverage  
Leverage  
Lift  
Strength  
4 rows selected.
```

Εικόνα 20 Ερώτημα 2 (E2)

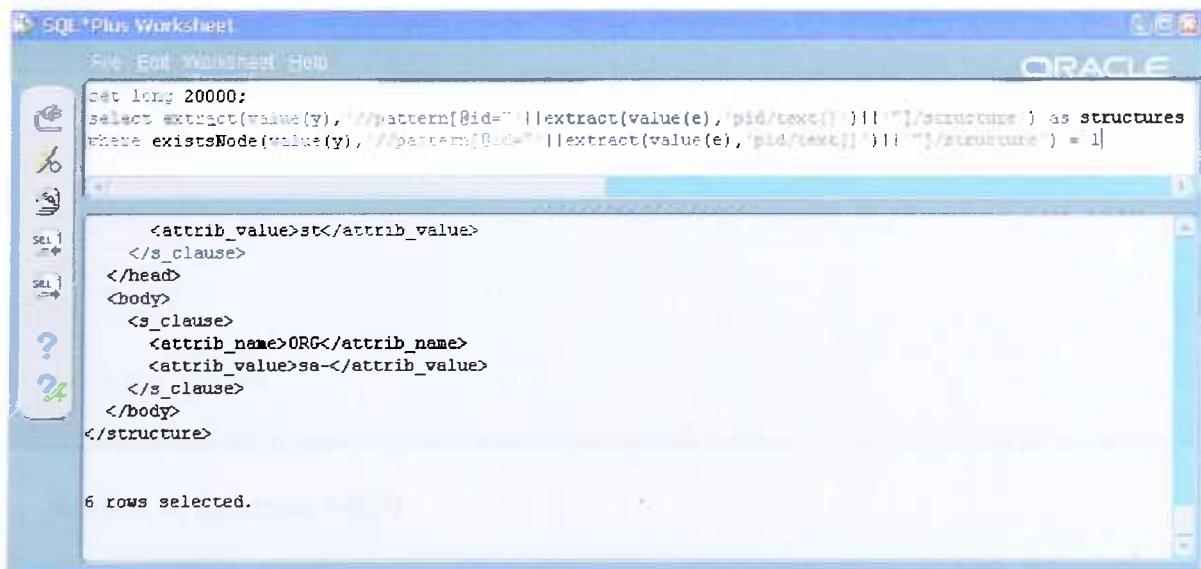
**E3)** Ανάκτηση της δομής (ομοίως της πηγής, του μέτρου ποιότητας ή την έκφραση) των κανόνων συσχέτισης που ανήκουν στην κλάση class1.

➤ Αφηρημένη μορφή:

```
SELECT A.structure  
FROM Association_Rules A
```

➤ Υλοποίηση SQL/XML στη βάση προτύπων:

```
select  
extract(value(y), '//pattern[@id=""]||extract(value(e), "pid/text()")||"/structure') as structures from assoc_rules y,  
classes x,  
TABLE(XMLsequence(extract(value(x), 'class[@name="class1"]//p  
ids/pid'))) e  
  
where  
existsNode(value(y), '//pattern[@id=""]||extract(value(e), "pid  
/text()")||"/structure") = 1
```



```
SQL*Plus Worksheet  
File Edit Worksheet Help  
ORACLE  
set long 20000;  
select extract(value(y), '//pattern[@id=""]||extract(value(e), "pid/text()")||"/structure") as structures  
where existsNode(value(y), '//pattern[@id=""]||extract(value(e), "pid/text()")||"/structure") = 1  
  
<attrib_value>st</attrib_value>  
</s_clause>  
</head>  
<body>  
<s_clause>  
<attrib_name>ORG</attrib_name>  
<attrib_value>sa-</attrib_value>  
</s_clause>  
</body>  
</structure>  
  
6 rows selected.
```

**Εικόνα 21** Ερώτημα 3 (E3)

**E4)** Ανάκτηση όλων των προτύπων ενός συγκεκριμένου τύπου.

➤ Αφηρημένη μορφή:

```
SELECT *
FROM classes
WHERE patternType="AssociationRule"
```

➤ Υλοποίηση SQL/XML στη βάση προτύπων:

```
select distinct
extractValue(value(y), '//pattern[@id=""]||extract(value(e), 'p
id/text()')||'"]/name/text()') as pattern_name from
assoc_rules y, classes x,
TABLE(XMLsequence(extract(value(x), 'class[@ptype="associatio
n_rule"]//pids/pid'))) e
```

```
SQL*Plus Worksheet
```

```
File Edit Advanced Help
```

```
ORACLE
```

```
select distinct extractValue( e  , 'y', '//pattern[@id=""]||extract(value(e), pid/text()')||'"]/name/text()')
pattern_name from assoc_rules y, classes x,
TABLE(XMLsequence(extract(value(x), 'class[@ptype="association_rule"]//pids/pid'))) e
```

```
pattern15
pattern16
pattern20
pattern21
pattern24
pattern27
pattern30
pattern4
pattern6
pattern7
```

```
15 rows selected.
```

**Εικόνα 22** Ερώτημα 4 (E4)

**E5)** Ανάκτηση εκείνων των κανόνων συσχέτισης που ανήκουν στην κλάση class1 των οποίων το μέτρο ποιότητας lift είναι μεγαλύτερο από 48.

➤ Αφηρημένη μορφή:

```
SELECT A
FROM Association_Rules A, Classes C
WHERE A.measure.lift > 48
AND C.name="class1"
```

➤ Υλοποίηση SQL/XML στη βάση προτύπων:

```
select
extract(value(y), '//pattern[@id=""]||extract(value(e), "pid/text()")||"/m_clause[measure_name="Lift"] [measure_value>"48"]") as pattern_name

from assoc_rules y, classes x,
TABLE(XMLsequence(extract(value(x), '//pids/pid'))) e

where
existsNode(value(y), '//pattern[@id=""]||extract(value(e), "pid/text()")||"/m_clause[measure_name="Lift"] [measure_value>"48"]') = 1
```

```
select extract(value(y), "/pattern[@id='']||extract(value(e), pid/text())||"/m_clause[measure_name='Lift']) as pattern_name
from assoc_rules y, classes x, TABLE(XMLsequence(extract(value(x), '//pids/pid'))) e
where existsNode(value(y), "/pattern[@id='']||extract(value(e), pid/text())||"/m_clause[measure_name='Lift']) = 1
```

```
<m_clause>
<measure_name>lift</measure_name>
<measure_value>49.27</measure_value>
</m_clause>

<m_clause>
<measure_name>lift</measure_name>
<measure_value>50.59</measure_value>
</m_clause>
```

5 rows selected.

Εικόνα 23 Ερώτημα 5 (E5)

**E6)** Ανάκτηση εκείνων των κανόνων συσχέτισης που ανήκουν στην κλάση class1 και που το μέρος body περιέχει το “ORG”.

➤ Αφηρημένη μορφή:

```
SELECT A
FROM Association_Rules A, Classes C
WHERE "ORG" IN A.structure.body
AND C.name="class1"
```

➤ Υλοποίηση SQL/XML της βάσης προτύπων:

```
select distinct
extractValue(value(y), '//pattern[@id=""]||extract(value(e), 'p
id/text()')||'"]/name/text()') as pattern_name from
assoc_rules y, classes x,
TABLE(XMLsequence(extract(value(x), '//pids/pid'))) e
where
existsNode(value(y), '//pattern[@id=""]||extract(value(e), 'pid
/text()')||'"]/structure/body/s_clause[attrib_name="ORG"]')=
1
```

```
SQL*Plus Worksheet
File Edit Favorites Help
ORACLE
Select distinct extractValue(value(y), '/pattern[@id=""||extract(value(e), pid/text())||""]/name/text()') as pattern_name from assoc_rules y, classes x, TABLE(XMLsequence(extract(value(x), '//pids/pid'))) e where existsNode(value(y), '/pattern[@id=""||extract(value(e), 'pid/text()')||""]/structure/body/s_clause[attrib_name="ORG"]')=1
pattern14
pattern15
pattern16
pattern20
pattern21
pattern24
pattern27
pattern30
pattern4
pattern6
pattern7
14 rows selected.
```

Εικόνα 24 Ερώτημα 6 (E6)

**E7)** Ανάκτηση εκείνων των κανόνων συσχέτισης που ανήκουν στην κλάση class1 και περιέχουν μια συγκεκριμένη τιμή στη δομή τους (structure) ("ORG").

- Αφηρημένη μορφή:

```
SELECT A
FROM Association_Rules A, Classes C
WHERE A.structure contains attribute ORG
AND C.name="class1"
```

- Υλοποίηση SQL/XML στη βάση προτύπων:

```
select distinct
extractValue(value(y), '//pattern[@id=""]||extract(value(e), 'p
id/text()')||'"]/name/text()' ) as pattern_name from
assoc_rules y, classes x,
TABLE(XMLsequence(extract(value(x), '//pids/pid'))) e
where
existsNode(value(y), '//pattern[@id=""]||extract(value(e), 'pid
/text()')||'"]/structure/body/s_clause[attrib_name="ORG"]')=
1 OR
existsNode(value(y), '//pattern[@id=""]||extract(value(e), 'pid
/text()')||'"]/structure/head/s_clause[attrib_name="ORG"]')=
1
```

```
--> select distinct extractValue(value(y), '//pattern[@id=""]||extract(value(e), pid/text()')||"name/text()' ) as
--> where existsNode(value(y), '//pattern[@id=""]||extract(value(e), pid/text()')||'"]/structure/body/s_clause[attrib
--> name="ORG"]')=1 OR
--> existsNode(value(y), '//pattern[@id=""]||extract(value(e), pid/text()')||'"]/structure/head/s_clause[attrib
--> name="ORG"]')=1
pattern14
pattern15
pattern16
pattern20
pattern21
pattern24
pattern27
pattern30
pattern4
pattern6
pattern7
14 rows selected.
```

Εικόνα 25 Ερώτημα 7 (E7)

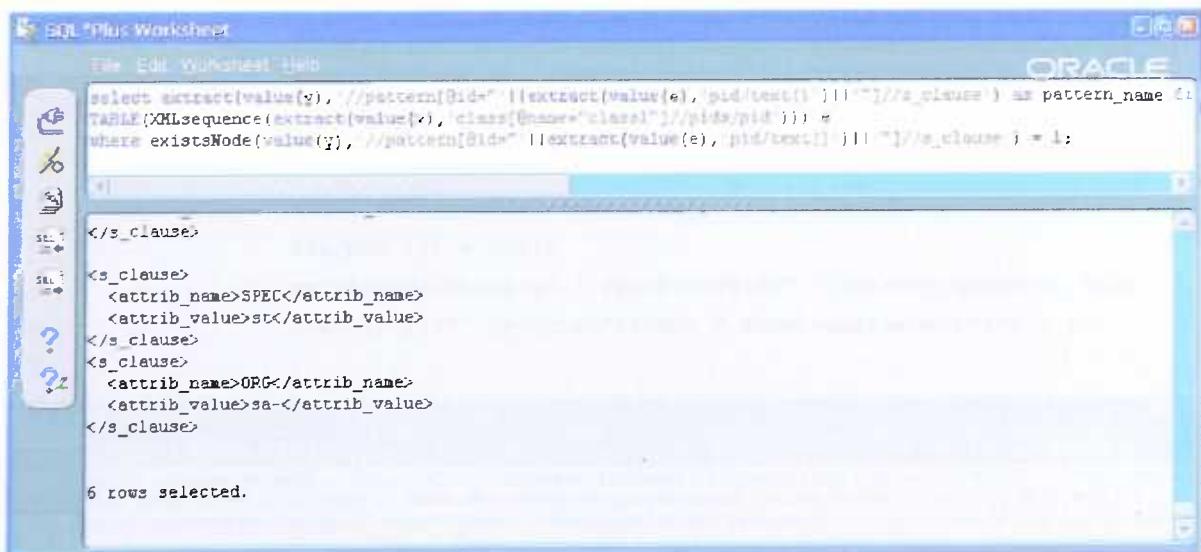
**E8)** Ανάκτηση του μέρους head και του μέρους body των προτύπων που ανήκουν στην κλάση class1.

➤ Αφηρημένη μορφή:

```
SELECT A.structure.body, A.structure.head  
FROM Association_rules A, classes C  
WHERE C.name="class1"
```

➤ Υλοποίηση SQL/XML στη βάση προτύπων:

```
select  
    extract(value(y), '//pattern[@id="'||extract(value(e), 'pid/te  
xt()' )||'"']//s_clause') as pattern_name from assoc_rules y,  
    classes x,  
    TABLE(XMLsequence(extract(value(x), 'class[@name="class1"]//p  
ids/pid'))) e  
  
where  
    existsNode(value(y), '//pattern[@id="'||extract(value(e), 'pid  
/text()' )||'"']//s_clause') = 1;
```



The screenshot shows the Oracle SQL\*Plus Worksheet interface. The query window contains the following XML query:

```
select extract(value(y), '//pattern[@id="'||extract(value(e), 'pid/text()' )||'"']//s_clause') as pattern_name from assoc_rules y,  
    TABLE(XMLsequence(extract(value(x), 'class[@name="class1"]//pids/pid'))) e  
where existsNode(value(y), '//pattern[@id="'||extract(value(e), 'pid/text()' )||'"']//s_clause') = 1;
```

The results pane displays the output of the query, which includes several XML fragments representing clauses:

```
</s_clause>  
<s_clause>  
  <attrib_name>SPEC</attrib_name>  
  <attrib_value>stx</attrib_value>  
</s_clause>  
<s_clause>  
  <attrib_name>ORG</attrib_name>  
  <attrib_value>sa-</attrib_value>  
</s_clause>
```

At the bottom of the results pane, it says "6 rows selected."

**Εικόνα 26** Ερώτημα 8 (E8)

- E9)** Ανάκτηση των κανόνων συσχέτισης που ανήκουν στην κλάση class1 και που έχουν εξορυχτεί από ένα σύνολο δεδομένων που αναφέρονται σε ιστορικά δεδομένα από την Αθήνα.

Για να αναπαρασταθεί το ερώτημα αυτό, υποθέτουμε ότι τα δεδομένα auebRawData περιέχουν ένα πεδίο 'city' που περιέχει την πόλη που τα δεδομένα έχουν εξαχθεί.

- Αφηρημένη μορφή:

```
SELECT A FROM Association_Rules A, Classes C
WHERE C.name= 'class1" AND EXIST (A.source INTERSECT
Athens_data)
Athens_data:      SELECT * FROM auebRawData
WHERE city = 'Athens'
```

To Athens\_data είναι ένα ερώτημα που εκτελείται πάνω σε ένα πίνακα Σχεσιακής βάσης (δηλ. auebRawData), και άρα είναι ένα cross-over ερώτημα.

- Υλοποίηση SQL/XML στη βάση προτύπων:

```
select
extract(value(y) , '//pattern[@id=""]||extract(value(e) , 'pid/te
xt()')||"") [source="SELECT * FROM auebRawData"] ) as
pattern_name from assoc_rules y, classes x,
TABLE(XMLsequence(extract(value(x) , 'class[@name="class1"]//p
ids/pid'))) e where
existsNode(value(y) , '//pattern[@id=""]||extract(value(e) , 'pid
/text()')||"") [source="SELECT * FROM auebRawData"] ) = 1;
```

```
SQL*Plus Worksheet
File Edit Worksheet Help
ORACLE
select extract(value(y) , '/pattern[@id=""]||extract(value(e) , 'pid/te
xt()')||"") [source="SELECT * FROM auebRawData"] ) as
pattern_name from assoc_rules y, classes x,
TABLE(XMLsequence(extract(value(x) , 'class[@name="class1"]//p
ids/pid'))) e where
existsNode(value(y) , '/pattern[@id=""]||extract(value(e) , 'pid
/text()')||"") [source="SELECT * FROM auebRawData"] ) = 1;

<measure_name>Lift</measure_name>
<measure_value>45.17</measure_value>
</m_clause>
<m_clause>
<measure_name>Leverage</measure_name>
<measure_value>0.0013(19)</measure_value>
</m_clause>
</measure>
<expression>(SPEC="st",ORG="sa-")</expression>
</pattern>

6 rows selected.
```

Εικόνα 27 Ερώτημα 9 (E9)

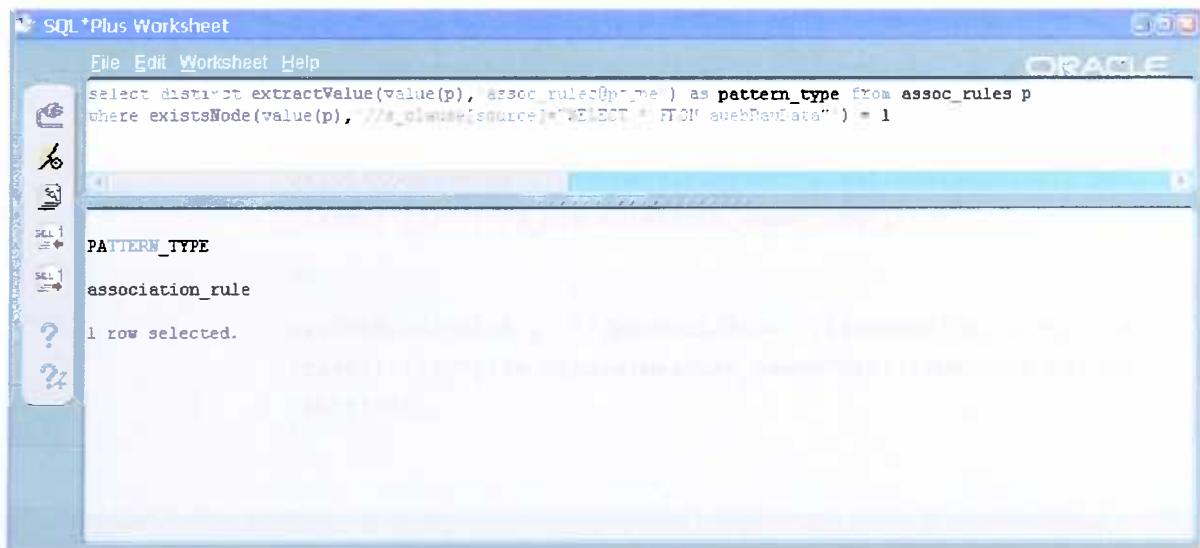
**E10)** Ανάκτηση όλων των διαφορετικών τύπων προτύπων που έχουν εξαχθεί από ένα συγκεκριμένο σύνολο αρχικών δεδομένων.

- Αφηρημένη μορφή:

```
SELECT C.ptype  
FROM classes C, Association_rules A  
WHERE A.source = "auebRAWDATA"
```

- Υλοποίηση SQL/XML στη βάση προτύπων:

```
select distinct extractValue(value(p), 'assoc_rules@ptype')  
from assoc_rules p  
  
where existsNode(value(p), '//s_clause[source] = "SELECT * FROM  
auebRawData"') = 1
```



The screenshot shows the Oracle SQL\*Plus Worksheet interface. The query window contains the following XQuery code:

```
select distinct extractValue(value(p), 'assoc_rules@ptype') as pattern_type from assoc_rules p  
where existsNode(value(p), '//s_clause[source] = "SELECT * FROM auebRawData"') = 1
```

The results pane displays the output:

PATTERN_TYPE
association_rule

1 row selected.

**Εικόνα 28** Ερώτημα 10 (E10)

**E11)** Ανάκτηση όλων των κανόνων συσχέτισης που ανήκουν στην κλάση class1 και που το μέρος body της δομής περιέχει το "org" ή που το μέτρο ποιότητας lift είναι μεγαλύτερο από 48

- Αφηρημένη μορφή:

```
SELECT A
FROM Association_Rules A, Classes C
WHERE "org" IN A.structure.body OR A.measure.lift > 48
AND C.name="class1"
```

- Υλοποίηση SQL/XML στη βάση προτύπων:

```
select distinct
extractValue(value(y), '//pattern[@id=""]||extract(value(e), 'p
id/text()'||'"]/name/text()' ) as pattern_name from
assoc_rules y, classes x,
TABLE(XMLsequence(extract(value(x), '//pids/pid'))) e
where
existsNode(value(y), '//pattern[@id=""]||extract(value(e), 'pid
/text()'||'"]//s_clause[attrib_name="ORG"]')=1
OR
existsNode(value(y), '//pattern[@id=""]||extract(value(e), 'pid
/text()'||'"]//m_clause[measure_name="Lift"] [measure_value>
"48"]')=1
```

The screenshot shows the Oracle SQL\*Plus Worksheet interface. The query window contains the SQL code provided above. The results window shows the output of the query, which lists 14 rows selected, corresponding to the pattern names: pattern14, pattern15, pattern16, pattern20, pattern21, pattern24, pattern27, pattern30, pattern4, pattern6, and pattern7.

```
select distinct extractValue(' value(y), '//pattern[@id=""]||extract(value(e), pid/text()'||'"]/name/cnvct()' ) as
pattern_name from assoc_rules y, classes x, TABLE(XMLsequence(extract(value(x), '//pids/pid'))) e
where
existsNode(value(y), '//pattern[@id=""]||extract(value(e), 'pid/text()'||'"]//s_clause[attrib_name="ORG"]')=1
OR
existsNode(value(y), '//pattern[@id=""]||extract(value(e), 'pid/text()'||'"]//m_clause[measure_name="Lift"] [measure_value>
"48"]')=1
```

pattern14  
pattern15  
pattern16  
pattern20  
pattern21  
pattern24  
pattern27  
pattern30  
pattern4  
pattern6  
pattern7

14 rows selected.

Εικόνα 29 Ερώτημα 11 (E11)

**E12)** Ανάκτηση όλων των κανόνων συσχέτισης που ανήκουν στην κλάση class1 ή στην κλάση class2

- Αφηρημένη μορφή:

```
SELECT A.name
FROM Association_Rules A, Classes C
WHERE C.name="class1"
UNION
SELECT A
FROM Association_Rules A, Classes C
WHERE C.name="class2"
```

- Υλοποίηση SQL/XML στη βάση προτύπων:

```
select distinct
extractValue(value(y) , '//pattern[@id=""]||extract(value(e) , 'p
id/text()'||'"]/name/text()' ) as pattern_name
from assoc_rules y, classes x,
TABLE(XMLSequence(extract(value(x) , 'class[@name="class1" or
@name="class2"]//pids/pid'))) e
```

```
SQL*Plus Worksheet - Emp
```

```
select distinct extractValue(value(y) , '//pattern[@id=""]||extract(value(e) , pid/text()'||"")/name/text()' ) as pattern_name
from assoc_rules y, classes x, TABLE(XMLSequence(extract(value(x) , 'class[@name="class1" or @name="class2"]//pids/pid'))) e
```

```
pattern14
pattern15
pattern16
pattern20
pattern21
pattern30
pattern4
pattern6
pattern7
```

```
13 rows selected.
```

Εικόνα 30 Ερώτημα 12 (E12)

**E13)** Ανάκτηση όλων των κανόνων συσχέτισης που ανήκουν στην κλάση class1 και στην κλάση class2

- Αφηρημένη μορφή:

```
SELECT A.name
FROM Association_Rules A, Classes C
WHERE C.name="class3"
UNION
SELECT A
FROM Association_Rules A, Classes C
WHERE C.name="class2"
```

- Υλοποίηση SQL/XML στη βάση προτύπων:

```
select distinct
extractValue(value(a), '//pattern[@id=""||e.q1||"]/name/text()')
as pattern_name from assoc_rules a, classes x, (select
extractValue(value(y), '//pid/text()') as q1 FROM classes c,
TABLE(XMLSequence(extract(value(c), 'class[@name="class3"]//pid'))) y
intersect
select extractValue(value(y), '//pid/text()') FROM classes c,
TABLE(XMLSequence(extract(value(c), 'class[@name="class1"]//pid'))) y) e
```

```
File Edit Worksheet Help
File Edit Worksheet Help
select distinct extractValue(value(a), '//pattern[@id=""||e.q1||"]/name/text()')
as pattern_name
from assoc_rules a, classes x, (select extractValue(value(y), '//pid/text()') as q1 FROM classes c,
TABLE(XMLSequence(extract(value(c), 'class[@name="class3"]//pid'))) y
intersect
select extractValue(value(y), '//pid/text()') FROM classes c,
TABLE(XMLSequence(extract(value(c), 'class[@name="class1"]//pid'))) y) e
```

PATTERN_NAME
pattern15
pattern16

3 rows selected.

Εικόνα 31 Ερώτημα 13 (E13)

**E14)** Ανάκτηση όλων των κανόνων συσχέτισης που ανήκουν στην κλάση class1 και όχι στην κλάση class3

➤ Αφορημένη μορφή:

```
SELECT A.name
  FROM Association_Rules A, Classes C
 WHERE C.name="class1"
EXCEPT
SELECT A
  FROM Association_Rules A, Classes C
 WHERE C.name="class3"
```

➤ Υλοποίηση SQL/XML στη βάση προτύπων:

```
(select extractValue(value(y), '//pid') as pattern_ids
  FROM classes c,
  TABLE(XMLSequence(extract(value(c), 'class[@name="class3"]//pid'))) y)
minus
(select extractValue(value(y), '//pid')
  FROM classes c,
  TABLE(XMLSequence(extract(value(c), 'class[@name="class1"]//pid'))) y)
```

The screenshot shows the Oracle SQL\*Plus Worksheet interface. The query is displayed in the workspace, and the results are shown in the output window. The results are as follows:

PATTERN_IDS
20
21
24
27

4 rows selected.

Εικόνα 32 Ερώτημα 14 (E14)

**E15)** Ανάκτηση του αριθμού των προτύπων μιας συγκεκριμένης κλάσης (class1)

- Αφηρημένη μορφή:

```
SELECT count(C.pid)
FROM classes C
WHERE C.name="class1"
```

- Υλοποίηση SQL/XML στη βάση προτύπων:

```
select count(*) as number_of_patterns from classes c,
table(xmlsequence(extract(value(c), 'class[@name="class1"]//p
id')))) p
```

The screenshot shows the Oracle SQL\*Plus Worksheet interface. The command entered is:

```
select count(*) as number_of_patterns from classes c, table(xmlsequence(extract(value(c), 'class[@name="class1"]//p
id')))) p
```

The output window displays the result:

NUMBER_OF_PATTERNS
6

1 row selected.

**Εικόνα 33** Ερώτημα 15 (E15)

**E16)** Ανάκτηση της μεγαλύτερης (max, ομοίως της ελάχιστης ή της μέσης κλπ) τιμής του μέτρου ποιότητας lift από τα πρότυπα της κλάσης class1

➤ Αφηρημένη μορφή:

```
SELECT max(A.measure.lift)
FROM Association_Rules A, Classes C
WHERE C.name="class1"
```

➤ Υλοποίηση SQL/XML στη βάση προτύπων:

```
select max(extractvalue(value(val), '//text()')) as
maximum_lift from assoc_rules a,
TABLE(xmlsequence(extract(value(a), '/m_clause[measure_name=
"Lift"]/measure_value'))) val
```

The screenshot shows the Oracle SQL\*Plus Worksheet interface. The query window contains the following SQL/XML code:

```
select max(extractvalue(value(val), '//text()')) as maximum_lift from assoc_rules a,
TABLE(xmlsequence(extract(value(a), '/m_clause[measure_name=
"Lift"]/measure_value'))) val
```

The results window displays the output:

MAXIMUM_LIFT
54.20

1 row selected.

**Εικόνα 34** Ερώτημα 16 (E16)

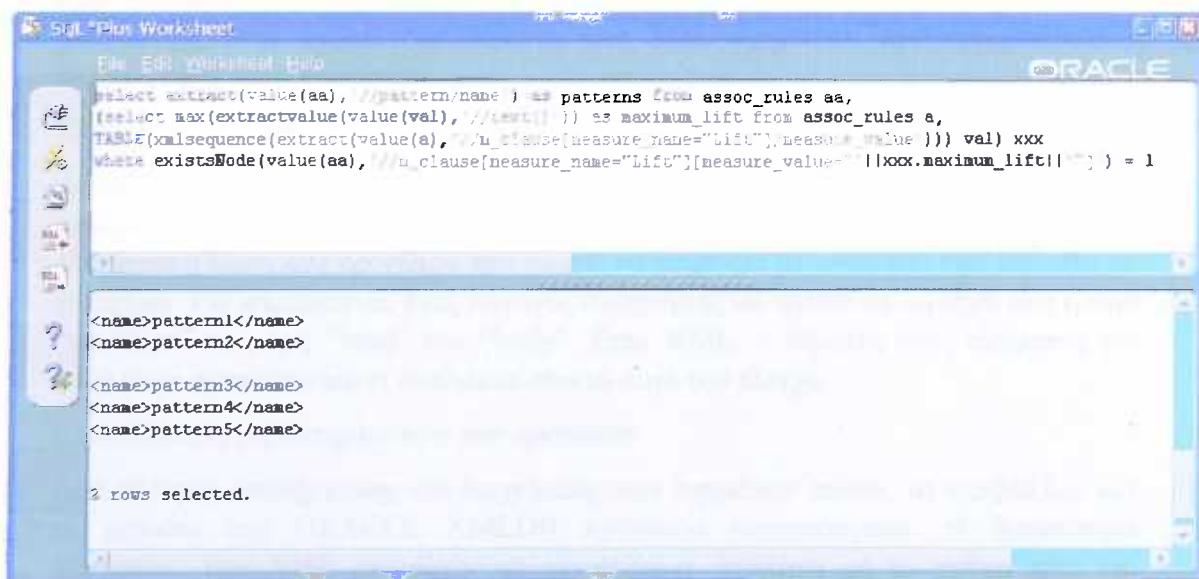
**E17)** Ανάκτηση των ονομάτων των προτύπων που έχουν τη μεγαλύτερη (max, ομοίως της ελάχιστης ή της μέσης κλπ) τιμή του μέτρου ποιότητας lift από τα πρότυπα της κλάσης class1

- Αφηρημένη μορφή:

```
SELECT A.name  
FROM Association_Rules A, classes C  
WHERE max(A.measure.lift)  
AND C.name="class1"
```

- Υλοποίηση SQL/XML στη βάση προτύπων:

```
select extract(value(aa), '//pattern/name/text()') as  
pattern_names from assoc_rules aa, (select  
max(extractvalue(value(val), '//text()')) as maximum_lift  
from assoc_rules a,  
TABLE(xmlsequence(extract(value(a), '//m_clause[measure_name=  
"Lift"]/measure_value'))) val) xxx  
  
where  
existsNode(value(aa), '//m_clause[measure_name="Lift"] [measur  
e_value=""||xxx.maximum_lift||""]') = 1
```



```
File Edit View Insert Help ORACLE  
select extract(value(aa), '//pattern/name ) as patterns from assoc_rules aa,  
(select max(extractvalue(value(val), '//text()')) as maximum_lift from assoc_rules a,  
TABLE(xmlsequence(extract(value(a), '//m_clause[measure_name="Lift"]/measure_value ))) val) xxx  
where existsNode(value(aa), '//m_clause[measure_name="Lift"] [measur  
e_value=""||xxx.maximum_lift||""]') = 1
```

```
<name>pattern1</name>  
<name>pattern2</name>  
  
<name>pattern3</name>  
<name>pattern4</name>  
<name>pattern5</name>
```

2 rows selected.

Εικόνα 35 Ερώτημα 17 (E17)

### **3.4 Κριτική – συμπεράσματα εξέτασης συστήματος διαχείρισης προτύπων σε XML.**

Βασική αρχή για την υλοποίηση της βάσης προτύπων σε XML ήταν η διατήρηση της γενικότητας καθώς και η υποστήριξη των λειτουργιών και των προδιαγραφών που περιγράφθηκαν στο λογικό μοντέλο για ένα σύστημα διαχείρισης προτύπων [3]. Το σύστημα είναι υποτυπώδες και δε σκοπεύει στην παρουσίαση ολοκληρωμένης βάσης προτύπων αλλά στην εξέταση της δυνατότητας κατασκευής αυτής. Τα δεδομένα που χρησιμοποιήθηκαν ήταν ενδεικτικά και σε ιδανικότερες συνθήκες θα βρισκόντουσαν αποθηκευμένα στην ίδια βάση με τη βάση προτύπων ώστε να είναι άμεση η εκτέλεση των cross-over ερωτήσεων. Η προεργασία των δεδομένων για εισαγωγή και επεξεργασία είναι μια χρονοβόρα διαδικασία και ευαίσθητη στα λάθη. Στην καλύτερη περίπτωση η εξαγωγή των XML εγγράφων θα γινόταν απευθείας από τη βάση με τα αρχικά δεδομένα και θα αποθηκευόντουσαν σε αυτή.

#### *Γενικότητα-επεκτασιμότητα*

Βασικό χαρακτηριστικό της XML βάσης προτύπων είναι η δυνατότητα ορισμού οποιασδήποτε μορφής τύπου προτύπου (επεκτασιμότητα). Ο ορισμός του τύπου προτύπου γίνεται με τη δημιουργία ενός XML σχήματος κατασκευασμένου ώστε να μπορεί να αναπαραστήσει όλα τα πρότυπα του συγκεκριμένου τύπου. Για κάθε νέο τύπο προτύπου που χρειάζεται να αποθηκευτεί πρέπει να οριστεί ένα νέο σχήμα. Ενώ το χαρακτηριστικό αυτό δίνει μεγάλη ευελιξία και ελευθερία, ταυτόχρονα προϋποθέτει την κατασκευή καλών XML σχημάτων. Το πρόβλημα λοιπόν μετατίθεται στην αποτελεσματικότητα των XML σχημάτων να αναπαραστήσουν πλήρως κάθε τύπο προτύπου.

#### *Έλεγχος εγκυρότητας*

Λόγω της δήλωσης ενός σχήματος για κάθε τύπο προτύπου, είναι εφικτός ο έλεγχος εγκυρότητας κάθε προτύπου που πρόκειται να αποθηκευτεί. Αυτό σημαίνει ότι ελέγχεται η δομή του προτύπου που πρέπει να ταιριάζει με αυτή που έχει δηλωθεί με το σχήμα. Για παράδειγμα, ένας κανόνας συσχέτισης θα πρέπει να περιέχει στο τμήμα “structure” τα μέρη “head” και “body”. Στην XML, η δήλωση ενός σχήματος για κάθε τύπο προτύπου κάνει αυτόματα εφικτό αυτό τον έλεγχο.

#### *Εκμετάλλευση χαρακτηριστικών των προτύπων*

Από πλευράς αποθήκευσης και διαχείρισης των σχημάτων αυτών, το περιβάλλον και οι μέθοδοι της ORACLE XMLDB κρίνονται ικανοποιητικά. Η δυνατότητα διάσπασης των XML εγγράφων σε αντικείμενα, ανάλογα με το σχήμα τους (το XMLSchema που τα αναπαριστά) και τη διαχείριση αυτών των αντικειμένων, κάνει το σύστημα πιο αποτελεσματικό ιδιαίτερα όσον αφορά την ανάκτηση και την ενημέρωση, τις σημαντικότερες δηλαδή λειτουργίες μιας βάσης προτύπων.

#### *Ευκολία δημιουργίας βάσης-ερωτήσεων*

Η δημιουργία ενός συστήματος σαν αυτό που παρουσιάστηκε δεν παρουσιάζει ιδιαίτερες δυσκολίες. Ωστόσο, η σύνταξη των ερωτημάτων είναι σχετικά πολύπλοκη και απαιτεί καλή γνώση της XPath [19] και των ιδιαίτερων διαδικασιών που παρέχει

η XMLDB. Ο συνδυασμός της XPath με την SQL δίνει το μεγάλο πλεονέκτημα συνδυασμού σχεσιακών πινάκων και πινάκων XMLType στα ίδια ερωτήματα και τη δημιουργία απαντήσεων είτε σε απλά σχεσιακά στοιχεία (στήλες και πίνακες) είτε σε μορφή XML εγγράφων που μπορούν να αποθηκευτούν ή να χρησιμοποιηθούν για την ανταλλαγή δεδομένων μεταξύ εφαρμογών.

#### *Δυνατότητα υλοποίησης περιορισμών*

Ένα μειονέκτημα της XML βάσης προτύπων είναι ότι δεν είναι εύκολη η υλοποίηση των περιορισμών. Για παράδειγμα είναι απαραίτητο σύμφωνα με το λογικό μοντέλο της βάσης προτύπων [3] το κάθε πρότυπο να ανήκει σε τουλάχιστον μια κλάση, ενώ μια κλάση πρέπει να περιέχει μόνο πρότυπα του ιδίου τύπου. Επειδή η πληροφορία για τα αναγνωριστικά των προτύπων (ids) βρίσκεται μέσα στα XML έγγραφα, δεν είναι δυνατός ο περιορισμός των τιμών τους με βάση τιμές που βρίσκονται σε άλλα έγγραφα. Για το λόγο αυτό θα είναι απαραίτητος ο έλεγχος των περιορισμών σε ένα άλλο επίπεδο, με κάποια εφαρμογή για παράδειγμα.

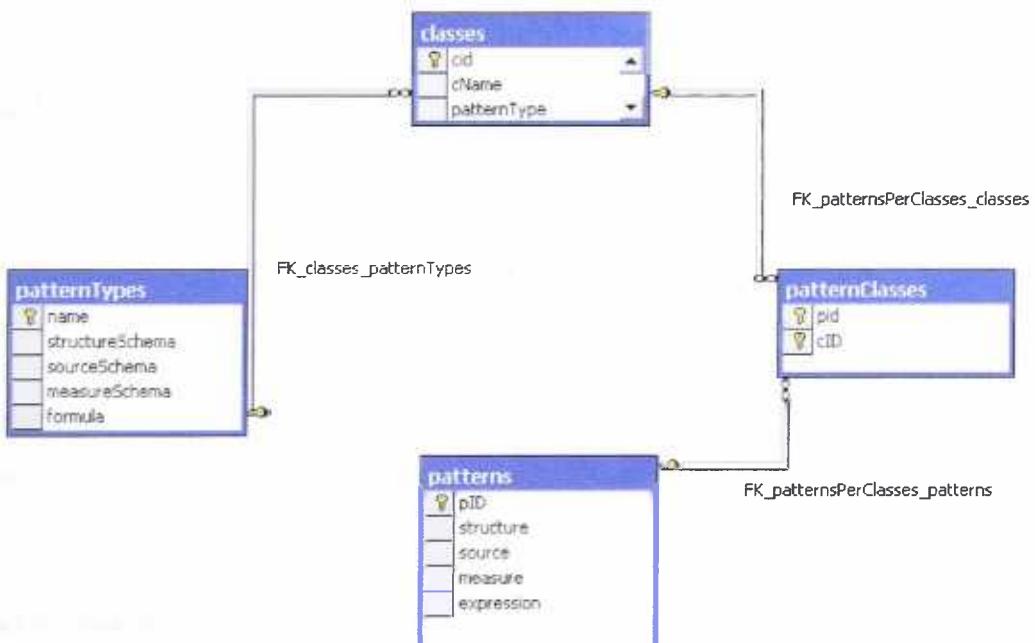


## 4. Εναλλακτικές υλοποιήσεις

Στην εργασία [11] η βάση προτύπων έχει υλοποιηθεί και στο σχεσιακό μοντέλο και στο αντικειμενο-σχεσιακό. Παρουσιάζονται τα δύο μοντέλα και κάποια βασικά ερωτήματα πάνω σε αυτά ώστε να είναι εφικτή η σύγκριση με την XML υλοποίηση.

### 4.1 Υλοποίηση Σχεσιακής Βάσης Προτύπων

Η υλοποίηση του προτεινόμενου λογικού μοντέλου σε σχεσιακή βάση έγινε σε SQL Server 2000 της Microsoft. Σύμφωνα με τους ορισμούς που δόθηκαν παραπάνω, το σχεσιακό μοντέλο που υιοθετήθηκε φαίνεται στην Εικόνα 36 [11].



Εικόνα 36 Σχεσιακό μοντέλο Βάσης Προτύπων

Κατά τη διάρκεια της σχεδίασης του σχήματος και των ερωτήσεων που θα εξεταζόντουσαν πάνω σε αυτό, δόθηκε ιδιαίτερη προσοχή ούτως ώστε να πληρούνται οι βασικές απαιτήσεις για γενικότητα, επεκτασιμότητα και επαναχρησιμοποίηση [11].

Παρακάτω δίνεται η περιγραφή του κάθε πίνακα και των περιεχομένων του.

Πίνακας 3 Πίνακες Σχεσιακού μοντέλου για Βάση προτύπων

**Table: patternTypes**

Όνομα πεδίου	Περιγραφή
Name	Το όνομα του τύπου προτύπου.
StructureSchema	Η δομή των προτύπων του συγκεκριμένου τύπου.
SourceSchema	Η περιγραφή των αρχικών δεδομένων από τα οποία προήλθαν τα πρότυπα.
MeasureSchema	Μέτρα ποιότητας για την ορθή αναπαράσταση των αρχικών δεδομένων από τα πρότυπα.
Formula	Ο τύπος που εκφράζει τη σχέση των προτύπων με τα αρχικά δεδομένα.

**Table: patterns**

Όνομα πεδίου	Περιγραφή
PID	Το αναγνωριστικό του προτύπου
Structure	Η δομή του προτύπου
Source	Περιγραφή των αρχικών δεδομένων από τα οποία προήλθαν τα πρότυπα
Measure	Οι τιμές για τα μέτρα ποιότητας
Expression	Ένας τύπος που σχετίζει τα πρότυπα με τα δεδομένα από τα οποία προήλθαν.

**Table: classes**

Όνομα πεδίου	Περιγραφή
cID	Το αναγνωριστικό της κλάσης
PatternType	Ο τύπος του προτύπου που αντιπροσωπεύει η κλάση
CName	Το όνομα-περιγραφή της κλάσης

**Table: patternClasses**

Όνομα πεδίου	Περιγραφή
cID	Το αναγνωριστικό της κλάσης
Pid	Το αναγνωριστικό του προτύπου

Τα δεδομένα που χρησιμοποιήθηκαν για τον έλεγχο των ερωτήσεων στη σχεσιακή βάση προτύπων είναι δύο ειδών, κανόνες συσχέτισης και συστάδες και προήλθαν από ιατρικά δεδομένα και από δεδομένα κατάτμησης πελατών (κατηγοριοποίηση πελατών) αντίστοιχα. Δυστυχώς, τα ποιοτικά μέτρα και ο τύπος αντιστοίχισης δεν υπάρχουν για τις συστάδες. Στην Εικόνα 37 φαίνονται παραδείγματα των δεδομένων που περιέχουν οι διάφοροι πίνακες του σχεσιακού μοντέλου.

SQL Server Enterprise Manager - [2>Data in Table 'patternTypes' in 'patternBaseNew' on 'NTOUTSI\NETSDC']				
name	structureSchema	sourceSchema	measureSchema	formula
AssociationRule	TUPLE(head:SET(STRING), body:SET(STRI BAG/transaction: SET(STRING))		TUPLE(confidence:REAL, support:REAL)	For every x (x belongs to head or x belongs to body) ms:confidence = 1.000, ms:support = 1.000
Cluster	TUPLE(radius:ROOT TYPE, center:TUPLE(c:SET(x:ROOT TYPE,y:ROOT TYPE))		ms: EMPTY SET	f: (x-cx)^2+(y-cy)^2 <= radius^2
ClusterOfInteger	TUPLE(radius:ROOT TYPE, center:TUPLE(cx:INTEGER, cy:INTEGER))		ms:avgIntraClusterDistance:REAL	f: (x-cx)^2+(y-cy)^2 <= radius^2
ClusterOfRules	representative:AssociationRule	SET(rule:AssociationRule)	ms: TUPLE(deviationOnConfidence:REAL, f:rule.ss.head=representative.ss.head)	

(a) Περιεχόμενα του πίνακα 'patternTypes'

SQL Server Enterprise Manager - [2>Data in Table 'patterns' in 'patternBaseNew' on 'NTOUTSI\NETSDC']				
pid	structure	source	measure	expression
1	head={SPEC='st'},body={ORG='sa'};	SELECT * FROM auebRawData	Coverage=0.002(31),Strength=1.000,Lift=54.20,Lever=0.002(29),Support=1.000,Confidence=1.000	{SPEC='st'},ORG='sa'
2	head={HOSP='avm',SERV='out'},body={ORG='sa'}	SELECT * FROM auebRawData	Coverage=0.002(29),Strength=1.000,Lift=54.20,Lever=0.002(28),Support=1.000,Confidence=1.000	{HOSP='avm',SERV='out'}
3	head={HOSP='avm',SPEC='st'},body={ORG='sa'}	SELECT * FROM auebRawData	Coverage=0.001(19),Strength=1.000,Lift=54.20,Lever=0.001(16),Support=1.000,Confidence=1.000	{HOSP='avm',SPEC='st'}
4	head={HOSP='avm',SPEC='st'},body={ORG='sa'}	SELECT * FROM auebRawData	Coverage=0.001(16),Strength=1.000,Lift=54.20,Lever=0.001(15),Support=1.000,Confidence=1.000	{HOSP='avm',SPEC='st'}
5	head={SPEC='st'},body={ORG='sa'}	SELECT * FROM auebRawData	Coverage=0.002(28),Strength=0.954,Lift=52.26,Lever=0.002(27),Support=0.954,Confidence=0.954	{SPEC='st'},ORG='sa'
6	head={SPEC='st'},body={ORG='sa'}	SELECT * FROM auebRawData	Coverage=0.002(30),Strength=0.933,Lift=50.59,Lever=0.002(29),Support=0.933,Confidence=0.933	{SPEC='st'},ORG='sa'
7	head={HOSP='avm',SPEC='st'},body={ORG='sa'}	SELECT * FROM auebRawData	Coverage=0.001(22),Strength=0.909,Lift=49.27,Lever=0.001(21),Support=0.909,Confidence=0.909	{HOSP='avm',SPEC='st'}
8	head={HOSP='avm',SERV='med'},body={ORG='sa'}	SELECT * FROM auebRawData	Coverage=0.003(51),Strength=0.932,Lift=48.89,Lever=0.003(50),Support=0.932,Confidence=0.932	{HOSP='avm',SERV='med'}

(b) Περιεχόμενα του πίνακα 'patterns'

SQL Server Enterprise Manager - [2>Data in Table 'classes' in 'patternBaseNew' on 'NTOUTSI\NETSDC']		
cid	Show/Hide Diagram Pane	patternType
1	AUEB Association Rule 1	AssociationRule
2	MIT Clusters 1	Cluster

(c) Περιεχόμενα του πίνακα 'classes'

SQL Server Enterprise Manager - [2>Data in Table 'patternClasses' in 'patternBaseNew' on 'NTOUTSI\NETSDC']	
pid	cID
1	1
2	1
3	1
4	1
5	1

(d) Περιεχόμενα του πίνακα 'patternClasses'

Εικόνα 37 Περιεχόμενα πινάκων στη σχεσιακή βάση προτύπων

#### 4.1.1 Ερωτήματα (queries) στη σχεσιακή βάση προτύπων

Στη σχεσιακή βάση προτύπων εφαρμόστηκαν περίπου 60 ερωτήσεις σχετικά με προβολή (projection), επιλογή (selection), order-by, ένωση, τομή και διαφορά, συγκεντρωτικές συναρτήσεις, ομοιότητα προτύπων, join, semi-join και ερωτήσεις σχετικά με τις ιδιότητες της σύνθεσης, εκλέπτυνσης και εξειδίκευσης. Οι ερωτήσεις απευθύνονται τόσο στο επίπεδο προτύπων όσο και στο επίπεδο τύπου προτύπων και κλάσης [11].

Επιλεκτικά παρουσιάζονται τα πιο χαρακτηριστικά ερωτήματα που μπορεί να συναντηθούν σε μια βάση προτύπων με την περιγραφή τους στη φυσική γλώσσα, την υλοποίηση τους σε SQL και τα αποτελέσματα του ερωτήματος. Τα ερωτήματα αυτά αντιστοιχούν σε αυτά που υλοποιήθηκαν στην XML βάση.

*Σημ. Χρησιμοποιείται η μορφή  $A.B.x$  για να δηλώσει την προσπέλαση του πεδίου  $x$  που βρίσκεται κάτω από το  $B$  το οποίο βρίσκεται κάτω από το  $A$ .*



**E18)** Ανάκτηση της δομής (ομοίως της πηγής, του μέτρου ποιότητας ή την έκφραση) των κανόνων συσχέτισης που ανήκουν στην κλάση Association\_Rule\_1.

Το ερώτημα αυτό είναι αντίστοιχο του ερωτήματος (E3)

Υλοποίηση SQL στην βάση προτύπων:

```

SQL Server Enterprise Manager - [2 Data in Table 'classes' in 'patternbaseview' on 'MILIAUTEST\NETBIOS']

Console Window Help
File SQL Object Tools View Insert Run Reports Options Help
SELECT patterns.structure
FROM classes INNER JOIN patternClasses ON classes.cid = patternClasses.cID
INNER JOIN patterns ON patternClasses.pID = patterns.pID
WHERE (classes.cName = 'AUEB Association Rule 1')

```

structure
head={SPEC='st'},body={ORG='sa-'}
head={HOSP='avm',SERV='out'},body={ORG='sa-'}
head={HOSP='avm',HOSP='avm',SPEC='th'},body={ORG='sa-'}
head={SERV='out',SPEC='th'},body={ORG='bsa'}
head={HOSP='avm',SPEC='th'},body={ORG='bsa'}
head={HOSP='avm',HOSP='avm',SPEC='th'},body={ORG='bsa'}
head={SERV='out',SPEC='th'},body={ORG='bsa'}
head={HOSP='avm',HOSP='avm',SPEC='th'},body={ORG='bsa'}
head={HOSP='avm',SPEC='th'},body={ORG='bsa'}
head={HOSP='avm',SPEC='th'},body={ORG='bsa'}
head={SERV='out',SPEC='th'},body={ORG='bsa'}
head={HOSP='avm',HOSP='avm',SPEC='th'},body={ORG='bsa'}
head={HOSP='avm',HOSP='avm',SPEC='th'},body={ORG='bsa'}

Εικόνα 38 Ερώτημα 18 (Ε18)

**E19)** Ανάκτηση όλων των προτύπων ενός συγκεκριμένου τύπου.

Το ερώτημα αυτό είναι αντίστοιχο του ερωτήματος (E4)

Υλοποίηση SQL στη βάση προτύπων:

```

SQL Server Enterprise Manager - [2 Data in Table 'bases' in 'patternbaseview' on 'MILIAUTEST\NETBIOS']

Console Window Help
File SQL Object Tools View Insert Run Reports Options Help
SELECT patterns.*
FROM patterns INNER JOIN
patternClasses ON patterns.pID = patternClasses.pid INNER JOIN
classes ON classes.cid = patternClasses.cID
WHERE (classes.patternType = 'AssociationRule')

```

pID	structure	source	measure	expression
1	head={SPEC='st'},body={SELECT * FROM aub Coverage=0.002(31),Strength=1.0 {SPEC='st',ORG='sa-'}}			
2	head={HOSP='avm',SERV='out'},body={SELECT * FROM aub Coverage=0.002(29),Strength=1.0 {HOSP='avm',SERV='out',ORG='sa-'}}			
27	head={HOSP='avm',HC SELECT * FROM aub Coverage=0.004(53),Strength=1.0 {HOSP='avm',HOSP='avm',SPEC='th'}}			
28	head={SERV='out',SPEC SELECT * FROM aub Coverage=0.003(51),Strength=1.0 {SERV='out',SPEC='th',ORG='bsa'}}			
29	head={HOSP='avm',SPI SELECT * FROM aub Coverage=0.003(50),Strength=1.0 {HOSP='avm',SPEC='th',ORG='bsa'}}			

Εικόνα 39 Ερώτημα 19 (Ε19)

**E20)** Ανάκτηση εκείνων των κανόνων συσχέτισης που ανήκουν στην κλάση Association\_Rule\_1 των οποίων το μέτρο ποιότητας confidence είναι μεγαλύτερο από 0.7.

Το ερώτημα αυτό είναι αντίστοιχο του ερωτήματος (E5)

Υλοποίηση SQL στη βάση προτύπων:

Στην παρούσα βάση προτύπων, το μέτρο ποιότητας είναι μια ενιαία εγγραφή που αποτελείται από confidence και support, έτσι πρέπει να επιλέξουμε το κατάλληλο μέρος του πεδίου, που αναφέρεται στο confidence και στη συνέχεια να εφαρμόσουμε τον περιορισμό.

```

SELECT measure, LEFT(RIGHT(LEFT(measure, CHARINDEX(';', measure) - 1), 9), CHARINDEX(';', RIGHT(LEFT(measure, CHARINDEX(';', measure) - 1), 9)) - 1)
AS coverage
FROM patterns
WHERE (LEFT(RIGHT(LEFT(measure, CHARINDEX(';', measure) - 1), 9), CHARINDEX(';', RIGHT(LEFT(measure, CHARINDEX(';', measure) - 1), 9)) - 1) > 0.007)

```

measure	coverage
Coverage=0.009(130),Strength=0.900,Lift=48.78,Leverage=0.0077(114)	.009
Coverage=0.021(309),Strength=0.052,Lift=48.06,Leverage=0.0011(15)	.021
Coverage=0.008(122),Strength=0.836,Lift=45.32,Leverage=0.0067(99)	.008
Gravitee=0.017(39),Strength=0.783,Lift=47.45,Leverage=0.0159(236)	.021

**Εικόνα 40** Ερώτημα 20 (E20)

**E21)** Ανάκτηση εκείνων των κανόνων συσχέτισης που ανήκουν στην κλάση Association\_Rule\_1 και που το μέρος body περιέχει το "ORG".

Το ερώτημα αυτό είναι αντίστοιχο του ερωτήματος (E6)

Υλοποίηση SQL της βάσης προτύπων:

Από τα δύο μέρη, το head και το body, πρέπει να επιλεγεί το μέρος του body και μετά να εφαρμοστεί ο περιορισμός

```

SELECT patterns.structure, LEFT(patterns.structure, CHARINDEX('body=', patterns.structure) - 2) AS headExpression, RIGHT(patterns.structure,
LEN(patterns.structure) - CHARINDEX('body', patterns.structure) + 1) AS bodyExpression, classes.cName
FROM patterns INNER JOIN
patternClasses ON patterns.pID = patternClasses.pID INNER JOIN
classes ON patternClasses.cID = classes.cID
WHERE (RIGHT(patterns.structure, LEN(patterns.structure) - CHARINDEX('body', patterns.structure) + 1) LIKE '%ORG%') AND
(classes.cName = 'AUEB Association Rules 1')

```

structure	headExpression	bodyExpression
head={SPEC='st'},body={ORG='sa-'}	head={SPEC='st'}	body={ORG='sa-'}
head={HOSP='avm',SERV='out'},body={ORG='sa-'}	head={HOSP='avm',SERV='out'}	body={ORG='sa-'}
head={HOSP='avm',SPEC='st'},body={ORG='sa-'}	head={HOSP='avm',SPEC='st'}	body={ORG='sa-'}
head={HOSP='avm',SPEC='st'},body={ORG='sa-'}	head={HOSP='avm',SPEC='st'}	body={ORG='sa-'}
head={SPEC='st'},body={ORG='sa-'}	head={SPEC='st'}	body={ORG='sa-'}
head={SPEC='st'},body={ORG='sa-'}	head={SPEC='st'}	body={ORG='sa-'}

**Εικόνα 41** Ερώτημα 21 (E21)

**E22)** Ανάκτηση εκείνων των κανόνων συσχέτισης που ανήκουν στην κλάση Association\_Rule\_1 και περιέχουν μια συγκεκριμένη τιμή στη δομή τους (structure) («ORG»).

*To ερώτημα αυτό είναι αντίστοιχο του ερωτήματος (E7)*

Υλοποίηση SQL στη βάση προτύπων:

```

SELECT patterns.*
FROM   patterns INNER JOIN
       patternClasses ON patterns.pID = patternClasses.pid INNER JOIN
       classes ON patternClasses.cID = classes.cid
WHERE  (classes.cName = 'AUEB Association Rules 1') AND (patterns.structure LIKE '%org=%')

```

pID	structure	source	measure	expression
1	head={SPEC='st'},body={ORG='sa-'}	SELECT * FROM auebRawData	Coverage=0.002(31),Strength {SPEC='st',ORG='sa-'}	
2	head={HOSP='avm',SERV='out'},body={ORG='sa-'}	SELECT * FROM auebRawData	Coverage=0.002(29),Strength {HOSP='avm',SERV='out',OR	
3	head={HOSP='avm',SPEC='st'},body={ORG='sa-'}	SELECT * FROM auebRawData	Coverage=0.001(19),Strength {HOSP='avm',SPEC='st',ORG	
4	head={HOSP='avm',SPEC='st'},body={ORG='sa-'}	SELECT * FROM auebRawData	Coverage=0.001(16),Strength {HOSP='avm',SPEC='st',ORG	
5	head={SPEC='st'},body={ORG='sa-'}	SELECT * FROM auebRawData	Coverage=0.002(28),Strength {SPEC='st',ORG='sa-'}	

**Εικόνα 42** Ερώτημα 22 (E22)

**E23)** Ανάκτηση του μέρους head και του μέρους body των προτύπων που ανήκουν στην κλάση Association\_Rule\_1.

*To ερώτημα αυτό είναι αντίστοιχο του ερωτήματος (E8)*

Υλοποίηση SQL στη βάση προτύπων:

```

SELECT LEFT(patterns.structure, CHARINDEX('body', patterns.structure) - 2) AS headExpression, RIGHT(patterns.structure, LEN(patterns.structure) - CHARINDEX('body', patterns.structure) + 1) AS bodyExpression
FROM   patterns INNER JOIN
       patternClasses ON patterns.pID = patternClasses.pid INNER JOIN
       classes ON patternClasses.cID = classes.cid
WHERE  (classes.cName = 'AUEB Association Rules 1')

```

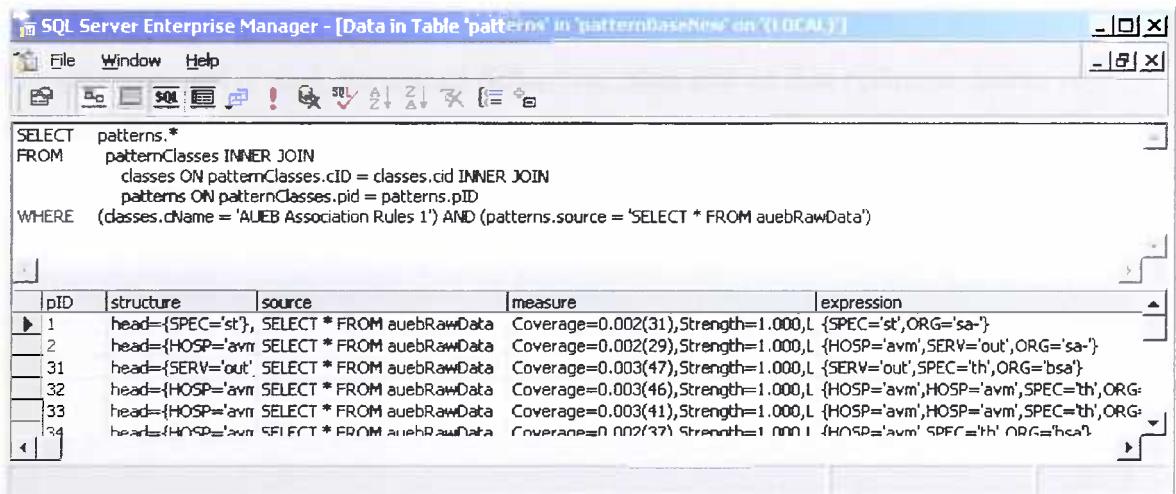
headExpression	bodyExpression
head={SPEC='st'}	body={ORG='sa-'}
head={HOSP='avm',SERV='out'}	body={ORG='sa-'}
head={HOSP='avm',SPEC='st'}	body={ORG='sa-'}
head={HOSP='avm',SPEC='st'}	body={ORG='sa-'}
head={SPEC='st'}	body={ORG='sa-'}

**Εικόνα 43** Ερώτημα 23 (E23)

**E24)** Ανάκτηση των κανόνων συσχέτισης που ανήκουν στην κλάση Association\_Rule\_1 και που έχουν εξορυχτεί από ένα σύνολο δεδομένων που αναφέρονται σε ιστορικά δεδομένα από την Αθήνα.

Το ερώτημα αυτό είναι αντίστοιχο του ερωτήματος (E9)

Υλοποίηση SQL στη βάση προτύπων:



The screenshot shows the SQL Server Enterprise Manager interface with a query window open. The query is:

```
SELECT patterns.*  
FROM patternClasses INNER JOIN  
    classes ON patternClasses.cid = classes.cid INNER JOIN  
    patterns ON patternClasses.pid = patterns.pID  
WHERE (classes.cName = 'AUEB Association Rules 1') AND (patterns.source = 'SELECT * FROM auebRawData')
```

The results grid displays the following data:

pID	structure	source	measure	expression
1	head={SPEC='st'}, SELECT * FROM auebRawData		Coverage=0.002(31),Strength=1.000,L	{SPEC='st',ORG='sa-'}
2	head={HOSP='avm' SELECT * FROM auebRawData		Coverage=0.002(29),Strength=1.000,L	{HOSP='avm',SERV='out',ORG='sa-'}
31	head={SERV='out', SELECT * FROM auebRawData		Coverage=0.003(47),Strength=1.000,L	{SERV='out',SPEC='th',ORG='bsa'}
32	head={HOSP='avm' SELECT * FROM auebRawData		Coverage=0.003(46),Strength=1.000,L	{HOSP='avm',HOSP='avm',SPEC='th',ORG='bsa'}
33	head={HOSP='avm' SELECT * FROM auebRawData		Coverage=0.003(41),Strength=1.000,L	{HOSP='avm',HOSP='avm',SPEC='th',ORG='bsa'}
34	head={HOSP='avm' SELECT * FROM auebRawData		Coverage=0.002(37),Strength=1.000,L	{HOSP='avm',SPEC='th',ORG='bsa'}

Εικόνα 44 Ερώτημα 24 (E24)

**E25)** Δοσμένων δύο προτύπων, να γίνει ανάκτηση των αρχικών δεδομένων που είναι κοινά και στα δύο

- Αφηρημένη μορφή:

```
SELECT P.source.BAG(transaction: SET(STRING))
FROM pattern P
WHERE P.pID=1
INTERSECT
SELECT P.source.BAG(transaction: SET(STRING))
FROM pattern P
WHERE P.pID=2
```

Για να επιτευχθεί η τομή, τα αρχικά δεδομένα, πίσω από τα δύο πρότυπα πρέπει να είναι συμβατά.

- Υλοποίηση SQL στη βάση προτύπων:

The screenshot shows the SQL Server Enterprise Manager interface with a query results grid. The grid displays data from a table named 'patterns'. The columns are labeled: SEX, SPEC#NO#, WARD, SERV, SP#DATE, SPEC, ORG, GRAM, AMP, AUG, TIC, PIP, AZL. The data consists of 12 rows, each representing a different record from the table. The first few rows show entries like 'm 72 ap med 4/1/1995 ur eco - 22 22 <NULL 27 <NULL' and 'f 6 cp med 1/6/1995 ur eco - 23 23 <NULL 31 <NULL'. The last row shows 'c ~ bx sur 2/6/1995 ur cfr - 22 22 <NULL 25 <NULL'.

SEX	SPEC#NO#	WARD	SERV	SP#DATE	SPEC	ORG	GRAM	AMP	AUG	TIC	PIP	AZL
m	72	ap	med	4/1/1995	ur	eco	-	22	22	<NULL	27	<NULL
f	6	cp	med	1/6/1995	ur	eco	-	23	23	<NULL	31	<NULL
m	3	cp	med	1/6/1995	ur	eco	-	20	20	<NULL	29	<NULL
f	19	ped	ped	1/6/1995	ur	eco	-	6	13	<NULL	6	<NULL
f	4	ort	sur	1/6/1995	ur	eco	-	22	22	<NULL	27	<NULL
m	1	ort	sur	1/6/1995	ea	mmo	-	24	21	<NULL	29	<NULL
f	29	ort	sur	2/6/1995	ur	bsb	+	32	32	<NULL	<NULL	<NULL
m	31	bx	sur	2/6/1995	ur	cfr	-	22	22	<NULL	25	<NULL
c	~											

Εικόνα 45 Ερώτημα 25 (E25)

**E26)** Ανάκτηση όλων των διαφορετικών τύπων προτύπων που έχουν εξαχθεί από ένα συγκεκριμένο σύνολο αρχικών δεδομένων.

*To ερώτημα αυτό είναι αντίστοιχο του ερωτήματος (E10)*

Υλοποίηση SQL στη βάση προτύπων:

```

SELECT DISTINCT classes.patternType
FROM      patterns INNER JOIN
          patternClasses ON patterns.pID = patternClasses.pid INNER JOIN
          classes ON patternClasses.cID = classes.cid
WHERE     (patterns.source LIKE '% FROM auebRawData%')
  
```

**Εικόνα 46** Ερώτημα 26 (E26)

**E27)** Ανάκτηση όλων των κανόνων συσχέτισης που ανήκουν στην κλάση Association\_Rule\_1 και που το μέρος body της δομής περιέχει το "org" ή που το μέτρο ποιότητας coverage είναι μεγαλύτερο από 0.01

*To ερώτημα αυτό είναι αντίστοιχο του ερωτήματος (E11)*

Υλοποίηση SQL στη βάση προτύπων:

```

SELECT measure, LEFT(RIGHT(LEFT(measure, CHARINDEX(' ', measure) - 1), 9), CHARINDEX(' ', RIGHT(LEFT(measure, CHARINDEX(' ', measure) - 1), 9)) - 1)
AS Expr1, structure
FROM      patterns
WHERE     (RIGHT(structure, LEN(structure) - CHARINDEX('body', structure) + 1) LIKE '%org%') OR
          (LEFT(RIGHT(LEFT(measure, CHARINDEX(' ', measure) - 1), 9), CHARINDEX(' ', RIGHT(LEFT(measure, CHARINDEX(' ', measure) - 1), 9)) - 1) > 0.007)
  
```

measure	Expr1	structure
Coverage=0.002(31),Strength=1.000,Lift=54.20,Leverage=0.0020(30)	0.002	head={SPEC='st'},body={ORG='sa-'}
Coverage=0.002(29),Strength=1.000,Lift=54.20,Leverage=0.0019(28)	0.002	head=(HOSP='avm',SERV='out'),body={ORG='sa-'}
Coverage=0.001(19),Strength=1.000,Lift=54.20,Leverage=0.0013(18)	0.001	head=(HOSP='avm',SPEC='st'),body={ORG='sa-'}
Coverage=0.001(16),Strength=1.000,Lift=54.20,Leverage=0.0011(15)	0.001	head=(HOSP='avm',SPEC='st'),body={ORG='sa-'}
Coverage=0.002(28),Strength=0.964,Lift=52.26,Leverage=0.0018(26)	0.002	head=(SPEC='st'),body={ORG='sa-'}
Coverage=0.002(30),Strength=0.933,Lift=50.59,Leverage=0.0018(27)	0.002	head=(SPEC='st'),body={ORG='sa-'}
Coverage=0.001(22),Strength=0.909,Lift=49.27,Leverage=0.0013(19)	0.001	head=(HOSP='avm',SPEC='st'),body={ORG='sa-'}
Coverage=0.003(51),Strength=0.902,Lift=48.89,Leverage=0.0030(45)	0.003	head=(HOSP='avm',SERV='med'),body={ORG='sa-'}
Coverage=0.009(130),Strength=0.900,Lift=48.78,Leverage=0.0077(114)	0.009	head=(HOSP='avm'),body={ORG='sa-'}
Coverage=0.001(19),Strength=0.895,Lift=48.50,Leverage=0.0011(16)	0.001	head=(SERV='ped',SPEC='st'),body={ORG='sa-'}
Coverage=0.002(47),Strength=0.894,Lift=48.43,Leverage=0.002(41)	0.002	head=(HOSP='avm',SERV='hrt'),body={ORG='sa-'}

**Εικόνα 47** Ερώτημα 27 (E27)

**E28)** Ανάκτηση όλων των κανόνων συσχέτισης που ανήκουν στην κλάση Association\_Rule\_1 ή στην κλάση Association\_Rule\_2

Το ερώτημα αυτό είναι αντίστοιχο του ερωτήματος (E12)

Υλοποίηση SQL στη βάση προτύπων:

Η χρήση της UNION ALL είναι δυνατή αν είναι επιθυμητή η διπλή εμφάνιση όλων των κοινών προτύπων.

```

SELECT classes.cname AS ClassName, patterns.*
FROM patterns INNER JOIN
      patternClasses ON patterns.pID = patternClasses.pID INNER JOIN
      classes ON patternClasses.cID = classes.cid
WHERE (classes.cName = 'AUEB Association Rules 1')
UNION
SELECT classes.cname AS ClassName, patterns.*
FROM patterns INNER JOIN
      patternClasses ON patterns.pID = patternClasses.pID INNER JOIN
      classes ON patternClasses.cID = classes.cid
WHERE (classes.cName = 'Association Rules 2')
    
```

ClassName	pID	structure	source	measure
AUEB Association Rules 1	1	head={SPEC='st'}	body={ SELECT * FROM auebRawData }	Coverage=0.002(31),Strength=1.000,Lift=54.2
AUEB Association Rules 1	2	head={HOSP='avm'}	SERV= SELECT * FROM auebRawData	Coverage=0.002(29),Strength=1.000,Lift=54.2
AUEB Association Rules 1	3	head={HOSP='avm'}	SPEC= SELECT * FROM auebRawData	Coverage=0.001(19),Strength=1.000,Lift=54.2

**Εικόνα 48** Ερώτημα 28 (E28)

**E29)** Ανάκτηση όλων των κανόνων συσχέτισης που ανήκουν στην κλάση Association\_Rule\_1 και στην κλάση Association\_Rule\_2

Το ερώτημα αυτό είναι αντίστοιχο του ερωτήματος (E13)

Υλοποίηση SQL στη βάση προτύπων:

```

SELECT patterns.*
FROM patterns INNER JOIN
      patternClasses ON patterns.pID = patternClasses.pID INNER JOIN
      classes ON patternClasses.cID = classes.cid
WHERE (classes.cName = 'AUEB Association Rules 1') AND EXISTS
      (SELECT patterns.*
       FROM patterns INNER JOIN
             patternClasses ON patterns.pID = patternClasses.pID INNER JOIN
             classes ON patternClasses.cID = classes.cid
       WHERE (classes.cName = 'Association Rules 2'))
    
```

pID	structure	source	measure	expression
1				

**Εικόνα 49** Ερώτημα 29 (E29)

**E30)** Ανάκτηση όλων των κανόνων συσχέτισης που ανήκουν στην κλάση Association\_Rule\_1 και όχι στην κλάση Association\_Rule\_2

*To ερώτημα αυτό είναι αντίστοιχο του ερωτήματος (E14)*

Υλοποίηση SQL στη βάση προτύπων:

```

SELECT patterns.*  

FROM patterns INNER JOIN  

    patternClasses ON patterns.pid = patternClasses.pid INNER JOIN  

        classes ON patternClasses.cID = classes.cid  

WHERE (classes.cName = 'AUEB Association Rules 1') AND (NOT EXISTS  

    (SELECT patterns.*  

        FROM patterns INNER JOIN  

            patternClasses ON patterns.pid = patternClasses.pid INNER JOIN  

                classes ON patternClasses.cID = classes.cid  

        WHERE (classes.cName = 'Association Rules 2')))

```

pID	structure	source	measure	expression
1	head={SPEC='st'},body={ORG= SELECT * FROM auebRawDat Coverage=0.002(31),Strength=1.000,Lift=54.20, {SPEC='st',ORG='sa-'}			
2	head={HOSP='avm',SERV='out'},SELECT * FROM auebRawDat Coverage=0.002(29),Strength=1.000,Lift=54.20, {HOSP='avm',SERV='out',ORG='sa-'}			
3	head={HOSP='avm',SPEC='st'},t SELECT * FROM auebRawDat Coverage=0.001(19),Strength=1.000,Lift=54.20, {HOSP='avm',SPEC='st',ORG='sa-'}			
4	head={HOSP='avm',SPEC='st'},t SELECT * FROM auebRawDat Coverage=0.001(16),Strength=1.000,Lift=54.20, {HOSP='avm',SPEC='st',ORG='sa-'}			
5	head={SPEC='st'},body={ORG= SELECT * FROM auebRawDat Coverage=0.002(28),Strength=0.964,Lift=52.26, {SPEC='st',ORG='sa-'}			

**Εικόνα 50** Ερώτημα 30 (E30)

**E31)** Ανάκτηση του αριθμού των προτύπων μιας συγκεκριμένης κλάσης

*To ερώτημα αυτό είναι αντίστοιχο του ερωτήματος (E15)*

Υλοποίηση SQL στη βάση προτύπων:

```

SELECT classes.cName, COUNT(classes.cName) AS numOfPatternsPerClass  

FROM patterns INNER JOIN  

    patternClasses ON patterns.pid = patternClasses.pid INNER JOIN  

        classes ON patternClasses.cID = classes.cid  

GROUP BY classes.cName

```

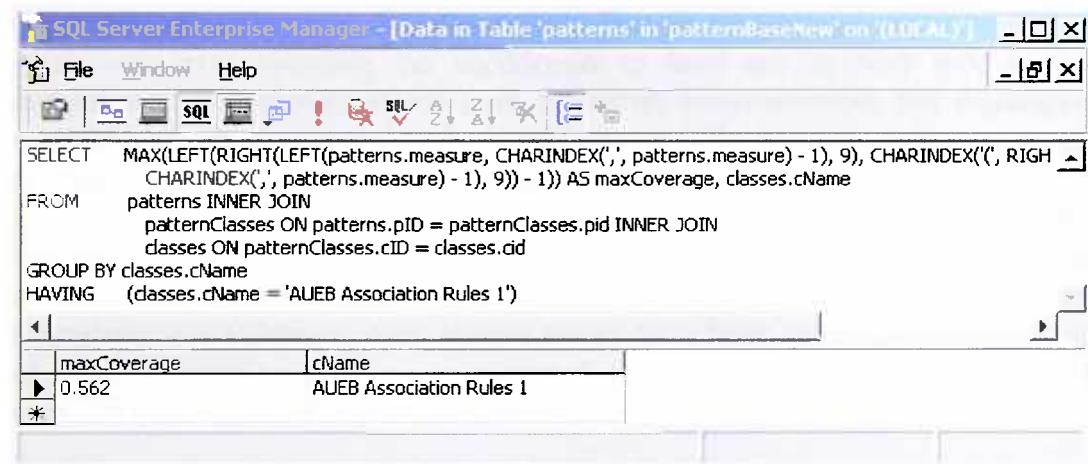
cName	numOfPatternsPerClass
AUEB Association Rules 1	814
MIT Cluster Rules 1	7

**Εικόνα 51** Ερώτημα 31 (E31)

**E32)** Ανάκτηση της μεγαλύτερης (max, ομοίως της ελάχιστης ή της μέσης κλπ) τιμής του μέτρου ποιότητας confidence από τα πρότυπα της κλάσης Association\_Rule\_1

Το ερώτημα αυτό είναι αντίστοιχο του ερωτήματος (E16)

Υλοποίηση SQL στη βάση προτύπων:



The screenshot shows the SQL Server Enterprise Manager interface with a query window open. The query is as follows:

```
SELECT MAX(LEFT(RIGHT(LEFT(patterns.measure, CHARINDEX(' ', patterns.measure) - 1), 9), CHARINDEX(' ', RIGHT(patterns.measure, CHARINDEX(' ', patterns.measure) - 1), 9)) - 1)) AS maxCoverage, classes.cName
FROM patterns INNER JOIN
    patternClasses ON patterns.pID = patternClasses.pid INNER JOIN
    classes ON patternClasses.cID = classes.cid
GROUP BY classes.cName
HAVING (classes.cName = 'AUEB Association Rules 1')
```

The results grid displays one row:

maxCoverage	cName
0.562	AUEB Association Rules 1

**Εικόνα 52** Ερώτημα 32 (E32)

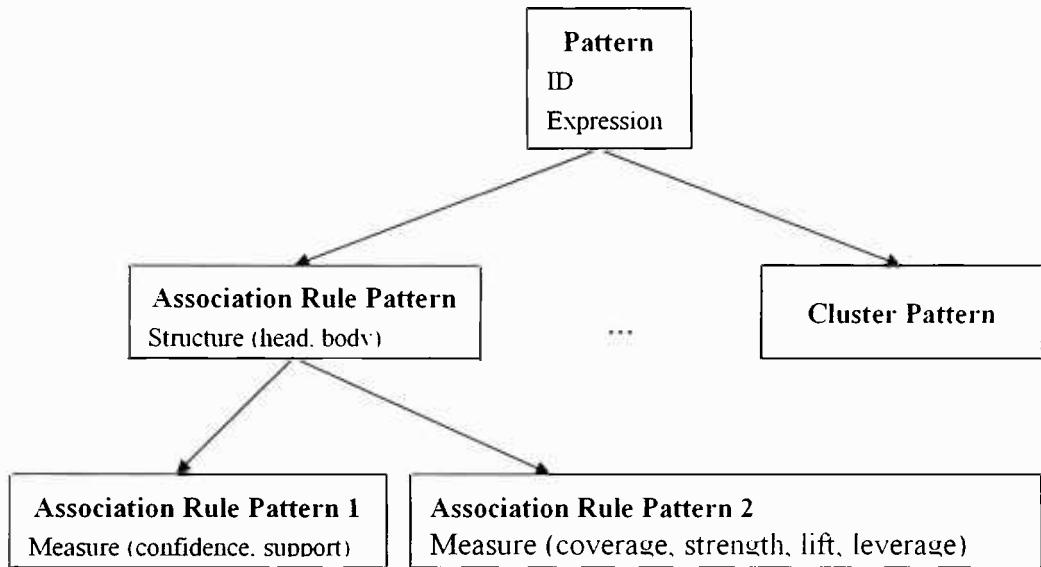
#### **4.1.2 Κριτική – συμπεράσματα εξέτασης συστήματος διαχείρισης προτύπων σε σχεσιακό μοντέλο**

Το βασικό μειονέκτημα του σχεσιακού μοντέλου είναι ότι κάθε συστατικό του (*structure, source, measure, etc*), υλοποιήθηκε σαν ένα γενικής μορφής πεδίο (συμβολοσειρά), οπότε η εκτέλεση των ερωτήσεων γίνεται ιδιαίτερα πολύπλοκη (απαιτείται αλφαριθμητική ανάλυση κάθε πεδίου για να χωριστεί στα στοιχεία που χρειάζονται στην ερώτηση, για παράδειγμα το *head* από το *body* ενός κανόνα συσχέτισης). Το χαρακτηριστικό αυτό προσθέτει πολυπλοκότητα στη δημιουργία ερωτημάτων και δεν επιτρέπει την διάκριση της δομής των προτύπων ή την εκμετάλλευση του τύπου τους. Ενώ το μοντέλο παρέχει απόλυτη ελευθερία στην ενσωμάτωση διαφορετικών τύπων προτύπων ταυτόχρονα μεταχειρίζεται τα πρότυπα σαν απλές συμβολοσειρές παραβλέποντας τις ιδιαιτερότητές τους. Η υλοποίηση που παρουσιάστηκε δεν αποτελεί σίγουρα την μοναδική δυνατή υλοποίηση ούτε απαραίτητα την καλύτερη, όμως γίνεται σαφές ότι η βάση προτύπων δεν μπορεί να υλοποιηθεί αποτελεσματικά με το σχεσιακό μοντέλο γιατί χάνεται ή γίνεται πολύ δύσκολη η διατήρηση των ιδιαιτεροτήτων και της σημασιολογίας των προτύπων.

## **4.2 Υλοποίηση Αντικειμενο-σχεσιακής βάσης προτύπων**

Στο αντικειμενο-σχεσιακό μοντέλο (*object-relational*) μπορεί να αποφευχθεί το βασικό μειονέκτημα του σχεσιακού ορίζοντας διαφορετικά αντικείμενα (*object*) και ιδιότητες (*attributes*) για κάθε ξεχωριστή συνιστώσα του προτύπου, εκμεταλλεύομενοι και τη σημαντική ιδιότητα της αντικειμενοστρέφιας, την κληρονομικότητα. Με τον τρόπο αυτό πετυχαίνουμε σημαντική μείωση της πολυπλοκότητας και αύξηση της αποδοτικότητας αφού πλέον τα ερωτήματα γίνονται απλούστερα.

Το λογικό μοντέλο αποτυπώθηκε μερικώς σε αντικειμενο-σχεσιακή βάση δεδομένων, συγκεκριμένα στο ORDBMS της ORACLE database 9i, και η κεντρική του ιδέα φαίνεται στο παρακάτω σχήμα [11].



Στη ρίζα υπάρχει ένα αντικείμενο «πρότυπο» γενικού τύπου που περιέχει τα κοινά χαρακτηριστικά όλων των τύπων προτύπων, όπως το αναγνωριστικό (id), τον τύπο αντιστοίχισης (expression) και την πηγή(source), τα αρχικά δεδομένα. Το γενικό αυτό αντικείμενο αναλύεται σε διάφορα υπο-αντικείμενα ανάλογα τον τύπο προτύπου (κανόνες συσχέτισης, συστάδες κλπ). Τα αντικείμενα αυτά διαφέρουν ως προς τη δομή τους και τα μέτρα ποιότητας και κληρονομούν από τον πατέρα τους, που είναι το αντικείμενο «pattern», τις ιδιότητές του (και τις μεθόδους του εάν υπάρχουν). Για παράδειγμα, το αντικείμενο «association rule pattern» περιέχει τις ιδιότητες που έχει κληρονομήσει από το αντικείμενο «pattern» (ID, expression και source) και επιπλέον περιέχει την ιδιότητα «structure» που αποτελείται από το «head» και «body». Το αντικείμενο αυτό αναλύεται περαιτέρω λόγω των διαφορετικών μέτρων ποιότητας που μπορεί να έχουν δύο διαφορετικοί κανόνες συσχέτισης. Έτσι, το μεν αντικείμενο «Association Rule Pattern 1» έχει εκτός από τις ιδιότητες που κληρονομεί, μια ιδιότητα «measure» που περιλαμβάνει δύο μέτρα ποιότητας, «confidence» και «support», και το δε αντικείμενο «Association Rule Pattern 2» μια ιδιότητα «measure» που περιλαμβάνει τα μέτρα ποιότητας coverage, strength, lift, leverage.

#### 4.2.1 Ερωτήματα (queries) στην αντικειμενο-σχεσιακή βάση προτύπων

Παρατίθενται παρακάτω μερικά βασικά ερωτήματα που υλοποιήθηκαν στην αντικειμενο-σχεσιακή βάση προτύπων αντίστοιχα με τις προηγούμενες υλοποιήσεις.

**E33)** Ανάκτηση της έκφρασης (expression) (ομοίως το ID, της πηγής (source), το μέτρο ποιότητας) από όλα τα πρότυπα.

Το ερώτημα αυτό είναι παρόμοιο του ερωτήματος (E3)

Υλοποίηση:

```
select expression from hr.tbl_patterns p;
```

The screenshot shows the Oracle SQL\*Plus Worksheet interface. In the command window, the query `select expression from hr.tbl_patterns p;` is entered. Below the command window, the results are displayed in a table with one column labeled "EXPRESSION". The results list 10 rows of data, each consisting of three values: SERV, SPEC, and ORG. The data includes various combinations such as {SERV=out, SPEC=th, ORG=bsa}, {HOSP=avm, SERV=out, ORG=bsa}, and {HOSP=avm, SPEC=sp, ORG=ac}. The results are preceded by the message "10 γραμμές επιλέχθηκαν."

EXPRESSION
{SERV=out, SPEC=th, ORG=bsa}
{HOSP=avm, SERV=out, ORG=bsa}
{HOSP=avm, SPEC=sp, ORG=ac}
{SPEC=bl, ORG=scn}
{SERV=med, SPEC=bl, ORG=scn}
{SERV=out, SPEC=th, ORG=bsa}
{HOSP=tgh, SPEC=ur, ORG=ci}
{SPEC=ea, ORG=seu}
{HOSP=avm, SPEC=ur, ORG=eco}
{HOSP=avm, ORG=eco}

10 γραμμές επιλέχθηκαν.

Εικόνα 53 Ερώτημα 33 (E33)

**E34)** Ανάκτηση της δομής των προτύπων κανόνων συσχέτισης.

Το ερώτημα αυτό είναι αντίστοιχο του ερωτήματος (E3)

Υλοποίηση:

```
select p.id, treat(value(p) as
hr.assrule_pattern).structureschema from hr.tbl_patterns p;
```

Η συνάρτηση *treat* χρησιμοποιείται για να προσδιοριστεί το γνώρισμα ενός υποτύπου.

The screenshot shows the Oracle SQL\*Plus Worksheet interface. The query window contains the following SQL code:

```
select p.id, treat(value(p) as hr.assrule_pattern).structureschema from hr.tbl_patterns p;
```

The results pane displays the output of the query, which is a large JSON-like structure representing the *STRUCTURESCHEMA* of the patterns. The structure is deeply nested, showing multiple levels of *ASSRULE\_STRUCTURE* and *ASSRULE\_LIST\_OF\_PAIRS* elements. The output is numbered from 1 to 7, indicating the sequence of the structure's components.

**Εικόνα 54** Ερώτημα 34 (E34)

**E35)** Ανάκτηση του μέρους body (ή head) της δομής των προτύπων κανόνων συσχέτισης.

Το ερώτημα αυτό είναι αντίστοιχο του ερωτήματος (E8)

Υλοποίηση:

```
select p.id,value(e) from hr.tbl_patterns p,
table(treat(value(p) as
hr.assrule_pattern).structureschema.body) e;
```

The screenshot shows the Oracle SQL\*Plus Worksheet interface. The query is:

```
select p.id,value(e)
from hr.tbl_patterns p,
table(treat(value(p) as hr.assrule_pattern).structureschema.body) e;
```

The output displays the body of the pattern, which consists of a list of ASSRULE\_PAIR entries:

- 1 ASSRULE\_PAIR('ORG', 'BSA')
- 2 ASSRULE\_PAIR('ORG', 'BSA')
- 3 ASSRULE\_PAIR('ORG', 'ac')
- 4 ASSRULE\_PAIR('ORG', 'scn')
- 5 ASSRULE\_PAIR('ORG', 'scn')
- 6 ASSRULE\_PAIR('ORG', 'BSA')
- 7 ASSRULE\_PAIR('ORG', 'ci')
- 8 ASSRULE\_PAIR('ORG', 'seu')

Εικόνα 55 Ερώτημα 35 (E35)

**E36)** Ανάκτηση του μέτρου ποιότητας confidence (ή support) των προτύπων κανόνων συσχέτισης.

➤ Αφηρημένη μορφή:

```
SELECT p.measure.confidence  
FROM association rule patterns p
```

➤ Υλοποίηση:

```
select p.id, treat(value(p) as  
hr.assrule_pattern_1).measureschema.confidence as confidence  
from hr.tbl_patterns p;
```

The screenshot shows the Oracle SQL\*Plus Worksheet interface. The query is:

```
select p.id,  
       treat(value(p) as hr.assrule_pattern_1).measureschema.confidence as confidence  
  from hr.tbl_patterns p;
```

The results are displayed in a grid:

ID	CONFIDENCE
1	,002
2	,0011
3	,007
4	,413
5	,362
6	
7	
8	
9	
10	

Below the grid, a message reads: "10 γραμμές επιλέχθηκαν."

**Εικόνα 56** Ερώτημα 36 (E36)

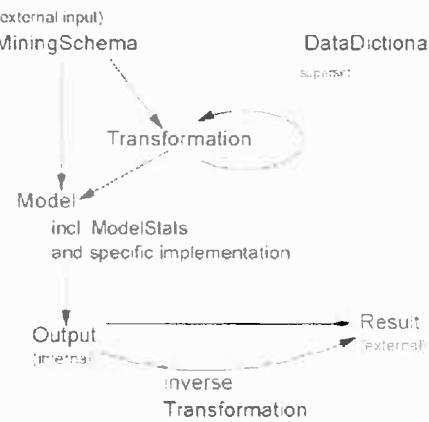
#### **4.2.2 Κριτική – συμπεράσματα εξέτασης συστήματος διαχείρισης προτύπων σε αντικειμενο-σχεσιακό μοντέλο**

Η αντικειμενο-σχεσιακή υλοποίηση φαίνεται ότι καταφέρνει να αναπαραστήσει καλύτερα τα πρότυπα και να χρησιμοποιήσει τα ιδιαίτερα δομικά του στοιχεία, βελτιώνοντας έτσι τα ερωτήματα. Η εξ' αρχής όμως δήλωση του τύπου κάθε προτύπου περιορίζει την ελευθερία δημιουργίας ενός γενικού τύπου προτύπου. Για παράδειγμα, ένας κανόνας συσχέτισης που έχει δηλωθεί με δύο μέτρα ποιότητας δεν είναι δυνατόν να αναπαραστήσει ένα πρότυπο με τέσσερα μέτρα ποιότητας και απαιτείται η δημιουργία νέων αντικειμένων (objects). Αυτό δυσκολεύει και την εκτέλεση πολλών ερωτημάτων γιατί πρότυπα του ίδιου τύπου προτύπου (association rules) μπορεί να βρίσκονται σε διαφορετικά αντικείμενα (association\_rule1 με δύο μέτρα ποιότητας και association\_rule2 με τέσσερα μέτρα ποιότητας).

### **4.3 Άλλες προσεγγίσεις**

Η ανάγκη διαχείρισης των αποτελεσμάτων της εξόρυξης γνώσης δεν είναι πρόσφατη. Διάφορα συστήματα έχουν υλοποιηθεί για τη λύση του προβλήματος αυτού και έχουν ευρέως χρησιμοποιηθεί αποδοτικά. Όλα αυτά τα συστήματα έχουν όμως σα στόχο τη διαχείριση μόνο προτύπων που είναι αποτέλεσμα της εξόρυξης γνώσης.

Η βασικότερη και πιο δημοφιλής προσπάθεια για την μοντελοποίηση προτύπων είναι το Predictive Model Markup Language (PMML) που έχει αναπτύξει η DMG (Data Mining Group)[6]. Η PMML είναι μια σύλλογή από ορισμούς XML εγγράφων που περιγράφουν τις διαδικασίες της εξόρυξης γνώσης και τα αποτελέσματα αυτών. Η PMML ορίζει διάφορα μοντέλα εξόρυξης όπως δέντρο κατηγοριοποίησης, νευρωνικά δίκτυα κα. Επίσης υπάρχουν ορισμοί κοινοί για όλα τα μοντέλα προκειμένου να μπορούν να περιγραφούν τα δεδομένα εισόδου και να εφαρμοστούν οι κατάλληλοι μετασχηματισμοί πριν εκτιμηθεί το μοντέλο. Τα βασικά στοιχεία ενός μοντέλου εξόρυξης και η ροή των δεδομένων μεταξύ τους φαίνονται στο παρακάτω σχήμα.



**Εικόνα 57** Βασικά δομικά στοιχεία μοντέλου εξόρυξης στην PMML

Το DataDictionary περιγράφει τα δεδομένα όπως είναι δηλαδή τα αρχικά δεδομένα εισόδου. Αναφέρεται στα αρχικά δεδομένα και ορίζει τον τρόπο που το μοντέλο εξόρυξης τα ερμηνεύει πχ. κατηγορικά ή αριθμητικά και τις τιμές που μπορούν να εισαχθούν. Τα αρχικά δεδομένα δεν περιλαμβάνονται στην PMML, αλλά είναι αποθηκευμένα σε εξωτερικές πηγές. Το DataDictionary ορίζει την αντιστοίχιση μεταξύ των γνωρισμάτων των αρχικών δεδομένων και των πεδίων του μοντέλου της εξόρυξης. Το MiningSchema ορίζει τη διεπαφή στο χρήστη των προτύπων εξόρυξης. Παρουσιάζει όλα τα πεδία που χρησιμοποιούνται σαν είσοδος για τους υπολογισμούς του μοντέλου εξόρυξης. Το μοντέλο εξόρυξης μπορεί να απαιτεί τη χρήση περισσοτέρων τιμών που εξαρτώνται από τις τιμές εισόδου, αλλά οι τιμές αυτές ορίζονται στο συστατικό του Transformation και δε βρίσκονται στο μοντέλο εξόρυξης. Στο μοντέλο εξόρυξης ορίζονται επίσης ποιες τιμές εκλαμβάνονται σαν τιμές εκτός της σωστής θέσης (outliers) και τι βάρη αντιστοιχούν σε ποια πεδία. Οι μετασχηματισμοί, μετατρέποντας τις αρχικές τιμές σε εσωτερικές τιμές, όπως είναι απαραίτητες για το μοντέλο εξόρυξης. Η έξοδος εξαρτάται από το συγκεκριμένο μοντέλο και το τελικό αποτέλεσμα (μια κλάση πρόβλεψης ή μια πιθανότητα) υπολογίζεται από τις τιμές εξόδου του μοντέλου. Η προσέγγιση της PMML είναι η πιο κοντινή στο στόχο της έρευνας αυτής και στην δημιουργία ενός γενικού συστήματος για τη διαχείριση των προτύπων.

Μία άλλη προσέγγιση βασισμένη στην SQL είναι η SQL/MM [7]. Τα μοντέλα που υποστηρίζονται περιγράφονται από δομημένους τύπους από τύπους SQL που είναι προσβάσιμοι από το SQL:1999 βασικό συντακτικό.

Το Common Warehouse Model (CWM) [9] προτείνει μια μέθοδο αναπαράστασης μεταδεδομένων με σκοπό να επιτρέψει την εύκολη ανταλλαγή μεταδεδομένων μεταξύ διαφορετικών και ετερογενών αποθηκών δεδομένων, αλλά όχι την αποτελεσματική διαχείριση αυτών.

Η πρόταση της JAVA για την υποστήριξη των μοντέλων εξόρυξης γνώσης, το JDM API, επιτρέπει την δημιουργία, αποθήκευση, πρόσβαση και συντήρηση δεδομένων και μεταδεδομένων που αναπαριστούν αποτελέσματα της εξόρυξης γνώσης [9].

Τέλος η PQL (Pattern Query Language) που προτάθηκε από το χώρο του Information Discovery, είναι μια γλώσσα επερωτήσεων στο ίδιο πρότυπο με την SQL αλλά τα πρότυπα τα διαχειρίζεται με τον παραδοσιακό τρόπο και τα αποθηκεύει σε σχεσιακούς πίνακες [10].

Από πλευράς αρχιτεκτονικής, όλες οι προσεγγίσεις χωρίζονται σε δύο κύριες κατηγορίες. Στην πρώτη, ένα επιπλέον επίπεδο προστίθεται στην κορυφή του DBMS και χειρίζεται τα πρότυπα με βάση την παραδοσιακή object-relational προσέγγιση. Στην δεύτερη κατηγορία, επεκτάσεις επιτυγχάνονται επιτρέποντας την ενσωμάτωση νέων συστατικών (components) στον πυρήνα του DBMS.

Όλες οι υπάρχουσες προσεγγίσεις δε φαίνεται να αντιμετωπίζουν με αποτελεσματικό τρόπο την αποθήκευση και διαχείριση των προτύπων κι αυτό γιατί μια αρχιτεκτονική σχεδιασμένη για απλά δεδομένα δεν μπορεί να ταιριάξει στις διαφορετικές ανάγκες διαχείρισης των προτύπων διαφόρων ειδών[2].

Μία άλλη προσέγγιση που βρίσκεται πολύ κοντά σε αυτήν του έργου PANDA είναι αυτή του CINQ (Consortium on Discovering Knowledge with Inductive Queries). Ο σκοπός του CINQ είναι η μελέτη και ανάπτυξη κατάλληλων τεχνικών ερωτήσεων πάνω σε επαγγεικές βάσεις δεδομένων (inductive databases) που περιέχουν και τα αρχικά δεδομένα και τα πρότυπα που προκύπτουν από αυτά μέσω της διαδικασίας της εξόρυξης γνώσης. Η επιλογή και διαχείριση (select and manipulation) των αρχικών δεδομένων, των προτύπων αλλά και μεταξύ των δύο αυτών (cross-over queries) αποτελούν το βασικό ενδιαφέρον του CINQ που θα μελετήσει θεωρητικά γνωστά (κανόνες συσχέτισης, datalog queries) και νέα πρότυπα (γράφοι) και θα αναπτύξει πρωτότυπα συστήματα για μια ποικιλία εφαρμογών όπως web mining, βιοπληροφορική και ανάλυση τηλεπικοινωνιακών δεδομένων [25].

## 5. ΣΥΓΚΡΙΤΙΚΗ ΜΕΛΕΤΗ ΥΛΟΠΟΙΗΣΕΩΝ

### 5.1 Κριτήρια σύγκρισης

Τα κριτήρια σύγκρισης των υλοποιήσεων που παρουσιάστηκαν είναι ποιοτικά και είναι τα παρακάτω:

- a. Ευκολία δημιουργίας βάσης προτύπων.
- b. Δυνατότητα υλοποίησης περιορισμών που ορίζονται από το λογικό μοντέλο
- c. Εκμετάλλευση των ιδιαιτεροτήτων (χαρακτηριστικών) των προτύπων
- d. Ευκολία δημιουργίας ερωτήσεων και αποτελεσματικότητά τους
- e. Επεκτασιμότητα (γενικότητα)
- f. Ελεγχος εγκυρότητας προτύπων

Ποσοτικά κριτήρια (πχ. χρόνος απόκρισης ή χώρος αποθήκευσης) δεν χρησιμοποιήθηκαν καθότι ο στόχος είναι να ερευνηθεί η καταλληλότητα των τριών μοντέλων (XML, σχεσιακό και αντικειμενο-σχεσιακό) για την υλοποίηση της βάσης προτύπων και όχι η απόδοση αυτών. Η απόδοσή τους μετρημένη με ποσοτικά κριτήρια θα γινόταν στην περίπτωση που και τα τρία μοντέλα κρίνονταν εξίσου κατάλληλα.

### 5.2 Σύγκριση υλοποιήσεων

Αναφορικά με τα κριτήρια που αναφέρθηκαν ακολουθεί η συγκριτική μελέτη.

- a. Ευκολία δημιουργίας βάσης προτύπων.

Και τα τρία μοντέλα (XML, σχεσιακό και αντικειμενο-σχεσιακό) δεν παρουσιάζουν ιδιαίτερες δυσκολίες στη δημιουργία της βάσης. Το σχεσιακό μοντέλο είναι το πιο απλούστερο και η εισαγωγή των δεδομένων σε αυτό είναι επίσης απλή. Στο αντικειμενο-σχεσιακό μοντέλο η δημιουργία είναι λίγο πιο πολύπλοκη αφού πρέπει να διαχωριστούν τα πρότυπα σε αντικείμενα που το καθένα έχει διαφορετική δομή και η εισαγωγή των δεδομένων απαιτεί κάποιον ειδικό μετασχηματισμό ή εργαλείο. Το XML μοντέλο φαίνεται να έχει τη μεγαλύτερη δυσκολία στη δημιουργία του, καθότι τα σχήματα που αντιπροσωπεύουν τους τύπους προτύπων πρέπει να δημιουργηθούν κατάλληλα. Η εισαγωγή των δεδομένων σε αυτό είναι σχετικά εύκολη όταν τα αρχικά δεδομένα βρίσκονται στην ίδια βάση με τα πρότυπα (που θα είναι και η πιθανότερη εκδοχή) και λιγότερο εύκολη όταν αυτά βρίσκονται είτε σε αρχεία είτε σε άλλη βάση δεδομένων (στην περίπτωση αυτή πρέπει πρώτα να εξαχθούν, να τροποποιηθούν κατάλληλα σε XML έγγραφα και να αποθηκευτούν στη βάση προτύπων).

b. Δυνατότητα υλοποίησης περιορισμών που ορίζονται από το λογικό μοντέλο  
Οι βασικότεροι περιορισμοί που προκύπτουν από το λογικό μοντέλο της βάσης προτύπων [3] είναι οι εξής:

1. Μία κλάση πρέπει να περιέχει πρότυπα του ίδιου τύπου
2. Κάθε πρότυπο είναι συγκεκριμένου τύπου και
3. Κάθε πρότυπο ανήκει τουλάχιστον σε μια κλάση

Στο σχεσιακό μοντέλο η υλοποίηση των περιορισμών είναι απλή. Μπορεί να οριστούν πάνω στα ξένα κλειδιά που υπάρχουν στους αντίστοιχους πίνακες. Στο αντικειμενο-σχεσιακό οι περιορισμοί γίνονται πιο περύπλοκοι γιατί τα πρότυπα (ως αντικείμενα) κάποιου τύπου δεν βρίσκονται κάτω από το ίδιο αντικείμενο και επομένως οι περιορισμοί θα πρέπει να εφαρμόζονται για το καθένα ξεχωριστά.

Στο μοντέλο XML ο ορισμός των περιορισμών είναι ιδιαίτερα δύσκολος καθότι τα δεδομένα βρίσκονται σε XML έγγραφα και η συσχέτιση των πεδίων διαφορετικών εγγράφων δεν είναι εφικτή. Ο έλεγχος των περιορισμών λοιπόν θα πρέπει να γίνεται με τη βοήθεια κάποιας εφαρμογής. Στην XML βάση είναι ωστόσο εφικτός ο έλεγχος της ορθότητας των προτύπων σχετικά με τον τύπο που αυτά ανήκουν. Για παράδειγμα το αν ένα πρότυπο τύπου κανόνα συσχέτισης έχει πεδίο “head” στη δομή του, ελέγχεται κατά την εισαγωγή του προτύπου στη βάση από το XMLSchema που έχει οριστεί για τον τύπο αυτόν.

c. Εκμετάλλευση των ιδιαιτεροτήτων (χαρακτηριστικών) των προτύπων

Όλα τα πρότυπα σύμφωνα με τον ορισμό που υιοθετήθηκε αποτελούνται από πέντε βασικά στοιχεία. Το όνομα, τη δομή, την πηγή, τα μέτρα ποιότητας και την έκφραση. Η συγκεκριμένη αυτή δομή μπορεί να χρησιμοποιηθεί για την καλύτερη διαχείριση των προτύπων. Στο σχεσιακό μοντέλο δε γίνεται καμία διάκριση των στοιχείων αυτών αφού αυτά θεωρούνται απλές συμβολοσειρές. Στο αντικειμενο-σχεσιακό και στο XML μοντέλο υπάρχει η βασική διάκριση στα στοιχεία αυτά και χρησιμοποιούνται και για το indexing. Γίνεται εκμετάλλευση δηλαδή των διαφορετικών χαρακτηριστικών για την αύξηση της αποδοτικότητας (καλύτερη ανάκτηση και ενημέρωση των δεδομένων).

d. Ευκολία δημιουργίας ερωτήσεων και αποτελεσματικότητά τους

Σημαντικό κριτήριο είναι η ευκολία σύνταξης ερωτήσεων πάνω στη βάση. Όπως φάνηκε και από την παρουσίαση των υλοποίησεων, στο σχεσιακό μοντέλο η δημιουργία των ερωτήσεων είναι αρκετά πολύπλοκη και στηρίζεται κυρίως σε συναρτήσεις χειρισμού συμβολοσειρών. Επειδή στα άλλα δύο μοντέλα γίνεται εκμετάλλευση της δομής των προτύπων, οι ερωτήσεις είναι πιο κοντά στη φυσική γλώσσα. Αν και για την XML απαιτείται μεγαλύτερη προσπάθεια, δεν υπάρχουν ερωτήματα που δεν μπορούν να απαντηθούν και επιπλέον με τη χρήση της XPath, υπάρχουν ερωτήματα που είναι εφικτά μόνο σε αυτήν. Για παράδειγμα η εύρεση ενός

εγγράφου που στη διαδρομή από τη ρίζα σε ένα κόμβο φύλλο υπάρχει μια συγκεκριμένη τιμή.

Πρέπει εδώ να αναφερθεί ότι ιδιαίτερη δυσκολία παρουσιάστηκε στα ερωτήματα order-by, group-by και στη χρήση της distinct.

#### e. Επεκτασιμότητα

Επεκτασιμότητα σημαίνει την ικανότητα εγκατάστασης στη βάση διαφορετικών τύπων προτύπων. Όσο πιο εύκολη είναι αυτή τόσο πιο επεκτάσιμο είναι το σύστημα. Στο σχεσιακό μοντέλο, κάθε νέος τύπος προτύπου απλά προστίθεται σαν μια νέα απλή εγγραφή στον αντίστοιχο πίνακα. Επομένως το σύστημα είναι εύκολα επεκτάσιμο.

Στο αντικειμενο-σχεσιακό μοντέλο όμως για τον ορισμό ενός νέου τύπου προτύπου είναι απαραίτητη η δημιουργία νέων αντικειμένων αφού πρώτα ελεγχθεί και διαπιστωθεί ότι κάτι τέτοιο δεν υπάρχει ήδη. Η επεκτασιμότητά του κρίνεται μέτρια.

Στο XML μοντέλο, ναι μεν χρειάζεται να ορισθεί ένα νέο σχήμα για κάθε νέο τύπο προτύπου, όμως επειδή ο κάθε τύπος είναι σχεδιασμένος για να περιλαμβάνει όλα τα πιθανά πρότυπα, μια τέτοια περίπτωση (δημιουργίας νέων τύπων) δεν θα είναι συχνή. Η ελευθερία δήλωσης οποιουδήποτε τύπου με οποιαδήποτε μορφή καθιστά το XML μοντέλο ιδιαίτερα επεκτάσιμο.

#### f. Έλεγχος εγκυρότητας προτύπων.

Τα πρότυπα αποτελούνται όπως αναφέρθηκε από κάκοια βασικά δομικά στοιχεία (δομή, πηγή, μέτρα ποιότητας και έκφραση). Επιπλέον, κάθε τύπος προτύπου διαφοροποιείται ως προς κάποιο από αυτά. Είναι πιθανό ο τύπος προτύπου pt1 να πρέπει να περιέχει στη δομή του δύο μέρη (για παράδειγμα “head” και “body”). Είναι επιθυμητό από τη βάση προτύπων, να ελέγχει αν τα πρότυπα που εισάγονται σε αυτή είναι έγκυρα, αν δηλαδή ακολουθούν τον τύπο τους. Λόγω της δυνατότητας της XML να ορίζει σχήματα για κάθε τύπο προτύπου, είναι η μόνη από τις τρεις υλοποιήσεις που μπορεί να εφαρμόσει έλεγχο εγκυρότητας των προτύπων.

Τα συμπεράσματα μπορούν να συνοψιστούν στον παρακάτω πίνακα:

	XML βάση	Σχεσιακή βάση	Αντικ/κή βάση
Ευκολία δημιουργίας βάσης	Μέτρια	Μεγάλη	Μέτρια
Δυνατότητα υλοποίησης περιορισμών	Δύσκολη	Εύκολη	Μέτρια
Έλεγχος εγκυρότητας προτύπων	Ναι	Όχι	Όχι
Εκμετάλλευση των χαρακτηριστικών των προτύπων	Ναι	Όχι	Ναι
Ευκολία δημιουργίας ερωτήσεων	Μέτρια	Μικρή	Μέτρια
Επεκτασιμότητα	Μεγάλη	Μεγάλη	Μέτρια

Πίνακας 4 Συγκριτικά αποτελέσματα

Παρατηρείται ότι αν και η βάση σε XML υπερτερεί στα περισσότερα σημεία, πράγμα που την κάνει προτιμότερη σαν υλοποίηση βάσης προτύπων, υπάρχουν σημεία που δεν είναι πλήρως αποδοτική. Για παράδειγμα, στο θέμα της γλώσσας ερωτήσεων υπάρχουν δυσκολίες. Αυτό δείχνει ότι ίσως είναι αναγκαία η δημιουργία μιας νέας γλώσσας ερωτήσεων εξειδικευμένη στα πρότυπα ή η υιοθέτηση μιας υπάρχουσας κάποιου άλλου μοντέλου (όπως κάποια από αυτές των μοντέλων στο κεφ. 4.3). Βέβαια αυτό θα ήταν περισσότερο εφικτό κατά τη δημιουργία ενός νέου ολοκληρωμένου συστήματος διαχείρισης βάσεων προτύπων.

## 6. ΣΥΜΠΕΡΑΣΜΑΤΑ - ΣΥΝΕΙΣΦΟΡΑ

Με την υλοποίηση της βάσης προτύπων με τρία διαφορετικά μοντέλα διαπιστώθηκε η καταλληλότητα της ημι-δομημένης προσέγγισης. Εξαιτίας της ιδιαιτερότητας των προτύπων να αποτελούνται από διαφορετικά συστατικά ανά εφαρμογή, οι παραδοσιακές τεχνικές αδυνατούν να τα αναπαραστήσουν και να τα χειριστούν αποτελεσματικά.

Συνήθως τα δεδομένα που αποθηκεύονται σε παραδοσιακές βάσεις δεδομένων όπως σχεσιακές και αντικειμενο-σχεσιακές χρειάζεται να είναι (εξαρχής) δομημένα και η δομή τους να περιορίζεται από ένα σχήμα ενώ δεν περιέχουν σημασιολογία.

Η πληθώρα όμως των δεδομένων που πλέον κυριαρχούν στις διάφορες εφαρμογές είναι αδόμητα και δεν ακολουθούν πάντα κάποιο σχήμα. Χαρακτηριστικό παράδειγμα είναι τα δεδομένα του διαδικτύου. Βρίσκονται σε διάφορες μορφές με διαφορετική δομή (ανάλογα με το περιβάλλον που δημιουργήθηκαν και τους κανόνες που το διέπουν), μπορεί όμως σημασιολογικά να είναι τα ίδια.

Τα πρότυπα, ως μια συνοπτική αναπαράσταση ενός μεγάλου όγκου δεδομένων έχουν κοινά στοιχεία και είναι πλούσια σε σημασιολογία. Η κατάσταση είναι πολύπλοκη, αφού τα πρότυπα δεν έχουν όλα την ίδια δομή (διαφορετική η δομή των κανόνων συσχέτισης, διαφορετική των συστάδων κλπ), αλλά και πρότυπα ίδιου τύπου μπορεί να περιλαμβάνουν διαφορετικά γνωρίσματα, ανάλογα με το πώς τα αντιλαμβάνονται και χρησιμοποιούν διάφοροι ερευνητές. Χαρακτηριστικό παράδειγμα αποτελεί αυτό των κανόνων συσχέτισης και των μέτρων ποιότητας που χρησιμοποιούνται για αυτούς. Κάποιοι ερευνητές χρησιμοποιούν τα support και confidence, ενώ κάποιοι άλλοι κάποια διαφορετικά, όπως το lift, το coverage, το strength κλπ. Φυσικά αυτό εξαρτάται άμεσα και από την κατά περίπτωση εφαρμογή.

Τα πρότυπα είναι εξ' ορισμού ημι-δομημένα και ως εκ τούτου η αναπαράστασή τους σε σχεσιακές ή αντικειμενο-σχεσιακές βάσεις δεν είναι καθόλου ευέλικτη, όπως φάνηκε και από τις υλοποιήσεις που αναφέρθηκαν. Θα πρέπει λοιπόν να στραφούμε σε ημι-δομημένες προσεγγίσεις για την αποθήκευση και διαχείριση των προτύπων. Σε μια ημι-δομημένη υλοποίηση δεν είναι υποχρεωτική η δήλωση ενός αυστηρού μοντέλου-σχήματος εκ των προτέρων. Εξαιτίας της φύσης των προτύπων, μια προσανατολισμένη στα δεδομένα (data-oriented) προσέγγιση θα είναι καλύτερη από μια προσανατολισμένη στη δομή προσέγγιση (structure-oriented) όπως αυτή των παραδοσιακών βάσεων). Η χρήση της XML ως της πλέον αποδεκτής ημι-δομημένης προσέγγισης είναι απαραίτητη για την κατασκευή ενός ολοκληρωμένου συστήματος διαχείρισης προτύπων. Τέλος η βάση δεδομένων της ORACLE με την προσθήκη της XLMDB για τη διαχείριση XML δεδομένων κρίνεται ικανοποιητική όμως υπάρχουν ελλείψεις (πχ. η υλοποίηση περιορισμών σε XML έγγραφα) που κάνουν απαραίτητη την εξέταση κι άλλων εναλλακτικών λύσεων. Εξάλλου το XLMDB δεν είναι προσανατολισμένο στα πρότυπα αλλά στη γενική διαχείριση XML εγγράφων.

## 7. ΑΝΟΙΚΤΗ ΕΡΕΥΝΑ

Παρότι η ORACLE έχει δημιουργήσει ένα αρκετά συνεκτικό και αποτελεσματικό περιβάλλον για τη διαχείριση XML εγγράφων, η εξέταση αμιγώς XML συστημάτων (πχ. Tamino) για την υλοποίηση μιας βάσης προτύπων σε XML είναι απαραίτητη, προκειμένου να διαπιστωθεί αν η XML όντως είναι η καλύτερη αναπαράσταση.

Συνοπτικά τα σημαντικότερα θέματα που πρέπει να εξεταστούν είναι τα εξής:

- Θέματα ομοιότητας προτύπων και ενσωμάτωσης λειτουργιών σύγκρισης προτύπων σε ένα Σύστημα Διαχείρισης Βάσεων Προτύπων.
- Εξέταση της δυνατότητας ενσωμάτωσης των χαρακτηριστικών εξειδίκευσης, σύνθεσης και εκλέπτυνσης και επαναχρησιμοποίησης που προτείνονται από το λογικό μοντέλο προτύπων που υιοθετήθηκε και υλοποίηση αυτών σε ένα ΣΔΒΠ.
- Υλοποίηση μιας βάσης προτύπων με τη χρήση αμιγώς XML.συστήματος και σύγκριση αυτής με αυτές που ήδη παρουσιάστηκαν.

Εάν αποδειχθεί ότι η υλοποίηση ενός ΣΔΒΠ στις υπάρχουσες τεχνολογίες δεν μπορεί να καλύψει τα βασικά αυτά χαρακτηριστικά τότε ίσως πρέπει να στραφούμε στη σχεδίαση και δημιουργία ενός νέου συστήματος προσανατολισμένο αποκλειστικά στις ανάγκες διαχείρισης των προτύπων. Αυτό προϋποθέτει την υιοθέτηση μιας ειδικής γλώσσας αναπαράστασης και ερωτήσεων και νέες τεχνικές αποθήκευσης, φυσικής αναπαράστασης και ευρετηριοποίησης καθώς και οπτικοποίησης.

Παράλληλα με τη μελλοντική αυτή εργασία, θα μπορούσε να γίνει μια θεωρητική έρευνα για το αν τα πρότυπα όλων των ειδών μπορούν να αναχθούν σε κάποιους θεμελιώδης τύπους προτύπων. Σε περίπτωση που μια τέτοια έρευνα έδινε θετική απάντηση, τότε ίσως έπρεπε να επανεξεταστεί η μοντελοποίηση των προτύπων και η ανάγκη για τη δημιουργία εκ νέου ενός ΣΔΒΠ.

## ΠΑΡΑΡΤΗΜΑ Α

**Πίνακας 5** Πλεονεκτήματα και μειονεκτήματα αδόμητης και δομημένης αποθήκευσης XML εγγράφων σε ORACLE 9i

	<b>UNSTRUCTURED STORAGE</b>	<b>STRUCTURED STORAGE</b>
Throughput	Highest possible throughput when ingesting and retrieving the entire content of an XML document.	The decomposition process results in slightly reduced throughput when ingesting or retrieving the entire content of an XML document.
Flexibility	Provides the maximum amount of flexibility in terms of the structure of the XML documents that can be stored in an XMLType column or table.	Limited Flexibility. Only document that conform to the XML Schema can be stored in the XMLType table or column. Changes to the XML Schema may require data to be unloaded and reloaded
XML Fidelity	Delivers Document Fidelity: Maintains the original XML byte for byte, which may be important to some applications	DOM Fidelity: A DOM created from an XML document that has been stored in the database will be identical to a DOM created from the original document. However trailing new lines, white space characters between tags and some data formatting may be lost
Optimized Update Operations	Optimized update operation are not possible. When any part of the document is updated the entire document must be written back to disk.	The majority of update operations can be optimized using Query rewrite. This allows in-place, piece-wise update, leading to significantly reduced response times and greater throughput.
XPath based queries.	XPath operations evaluated by constructing DOM from CLOB and using functional	Where possible, XPath operations are evaluated using query-rewrite, leading

	<b>evaluations.</b> This can be very expensive when performing operations on large collections of documents.	to significantly improved performance, particularly with large collections of documents.
<b>SQL Constraint Support</b>	SQL constraints are not currently available.	SQL constraints are supported.
<b>Indexing Support</b>	Text and Functional indexes.	B-Tree, Text and Functional Indexes.
<b>Optimized Memory Management</b>	XML operations of the document require creating a DOM from the document	XML operations can be optimized to reduce memory requirements.

*Πηγή: Oracle9i. XML Database Developer's Guide - Oracle XML DB Release 2 (9.2) October 2002  
Part No. A96620-02*

## **ΠΑΡΑΡΤΗΜΑ Β. Πίνακας εικόνων**

Εικόνα 1 Συστήματα διαχείρισης βάσεων δεδομένων και προτύπων.....	13
Εικόνα 2 Αρχιτεκτονική βάσης δεδομένων.....	14
Εικόνα 3 Σύστημα Διαχείρισης Βάσεων Προτύπων και Βάσεων Δεδομένων .....	15
Εικόνα 4 Κλάσεις, τύποι προτύπων και πρότυπα .....	17
Εικόνα 5 Η βάση προτύπων και τα συστατικά της .....	20
Εικόνα 8 association_rule.xsd Σχήμα (σε XMLSchema) για την μορφή του τύπου (pattern type) κανόνα συσχέτισης.....	31
Εικόνα 9 Σχηματική απεικόνιση του association_rule.xsd.....	31
Εικόνα 10 pattern_association_rule.xml XML έγγραφο τύπου κανόνα συσχέτισης (association rule).....	33
Εικόνα 11 cluster.xsd Σχήμα (σε XMLSchema) για την μορφή του τύπου (pattern type) .....	34
Εικόνα 12 Σχηματική απεικόνιση του cluster.xsd .....	35
Εικόνα 13 cluster.xml XML έγγραφο τύπου συστάδας (cluster).....	36
Εικόνα 14 class.xsd Σχήμα (σε XML-Schema) για την μορφή της κλάσης (class)....	37
Εικόνα 15 Σχηματική απεικόνιση του class.xsd .....	37
Εικόνα 16 class1.xml XML έγγραφο κλάσης (παράδειγμα) .....	38
Εικόνα 17 Επιλογές αποθήκευσης XMLType .....	40
Εικόνα 18 Υποσύνολο των δεδομένων που παραχωρήθηκαν από το ΟΠΑ. Κανόνες συσχέτισης από εξόρυξη σε βάση ιατρικών δεδομένων.....	42
Εικόνα 19 Ερώτημα 1 (E1) .....	44
Εικόνα 20 Ερώτημα 2 (E2) .....	45
Εικόνα 21 Ερώτημα 3 (E3) .....	46
Εικόνα 22 Ερώτημα 4 (E4) .....	47
Εικόνα 23 Ερώτημα 5 (E5) .....	48
Εικόνα 24 Ερώτημα 6 (E6) .....	49
Εικόνα 25 Ερώτημα 7 (E7) .....	50
Εικόνα 26 Ερώτημα 8 (E8) .....	51
Εικόνα 27 Ερώτημα 9 (E9) .....	52
Εικόνα 28 Ερώτημα 10 (E10) .....	53
Εικόνα 29 Ερώτημα 11 (E11) .....	54

Εικόνα 30 Ερώτημα 12 (E12) .....	55
Εικόνα 31 Ερώτημα 13 (E13) .....	56
Εικόνα 32 Ερώτημα 14 (E14) .....	57
Εικόνα 33 Ερώτημα 15 (E15) .....	58
Εικόνα 34 Ερώτημα 16 (E16) .....	59
Εικόνα 35 Ερώτημα 17 (E17) .....	60
Εικόνα 36 Σχεσιακό μοντέλο Βάσης Προτύπων.....	63
Εικόνα 37 Περιεχόμενα πινάκων στη σχεσιακή βάση προτύπων.....	65
Εικόνα 38 Ερώτημα 18 (E18) .....	67
Εικόνα 39 Ερώτημα 19 (E19) .....	67
Εικόνα 40 Ερώτημα 20 (E20) .....	68
Εικόνα 41 Ερώτημα 21 (E21) .....	68
Εικόνα 42 Ερώτημα 22 (E22) .....	69
Εικόνα 43 Ερώτημα 23 (E23) .....	69
Εικόνα 44 Ερώτημα 24 (E24) .....	70
Εικόνα 45 Ερώτημα 25 (E25) .....	71
Εικόνα 46 Ερώτημα 26 (E26) .....	72
Εικόνα 47 Ερώτημα 27 (E27) .....	72
Εικόνα 48 Ερώτημα 28 (E28) .....	73
Εικόνα 49 Ερώτημα 29 (E29) .....	73
Εικόνα 50 Ερώτημα 30 (E30) .....	74
Εικόνα 51 Ερώτημα 31 (E31) .....	74
Εικόνα 52 Ερώτημα 32 (E32) .....	75
Εικόνα 53 Ερώτημα 33 (E33) .....	78
Εικόνα 54 Ερώτημα 34 (E34) .....	79
Εικόνα 55 Ερώτημα 35 (E35) .....	80
Εικόνα 56 Ερώτημα 36 (E36) .....	81
Εικόνα 57 Βασικά δομικά στοιχεία μοντέλου εξόρυξης στην PMML .....	83

## **ΠΑΡΑΡΤΗΜΑ Γ. Πίνακας πινάκων**

Πίνακας 1 Τα τρία βασικά πρότυπα που εξάγονται με τεχνικές εξόρυξης γνώσης.....	6
Πίνακας 2 Παραδείγματα εφαρμογών και προτύπων .....	10
Πίνακας 3 Πίνακες Σχεσιακού μοντέλου για Βάση προτύπων.....	64
Πίνακας 4 Συγκριτικά αποτελέσματα.....	88
Πίνακας 5 Πλεονεκτήματα και μειονεκτήματα αδόμητης και δομημένης αποθήκευσης XML εγγράφων σε ORACLE 9i .....	91



## ΑΝΑΦΟΡΕΣ

- [1] M. Vazirgiannis, E. Vrahnos, M. Halkidi. "Motivating Pattern Management", PANDA Workshop on Pattern-Base Management Systems, April, 10<sup>th</sup> 2003 Como, Italy
- [2] M. Vazirgiannis, M. Halkidi, G. Tsatsaronis, E. Vrachnos, D. Keim, P. Xeros, Y. Theodoridis, A. Pikrakis, S. Theodoridis, "A Survey on Pattern Application Domains and Pattern Management Approaches". PANDA Technical Report PANDA-TR-2003-01. <http://dke.cti.gr/panda>
- [3] S. Rizzi, E. Bertino, B. Catania, Matteo Golfarelli, M. Halkidi, M. Terrovitis, P. Vassiliadis, M. Vazirgiannis, E. Vrahnos. "Towards a logical model for patterns", proc. of the ER, conference 2003
- [4] PANDA Technical details. <http://dke.cti.gr/panda>
- [5] CINQ (Consortium on Discovering Knowledge with Inductive Queries). <http://www.cinq-project.org>
- [6] Predictive Model Markup Language (PMML).  
[http://www.dmg.org/pmmlspecs\\_v2/pmml\\_v2\\_0.html](http://www.dmg.org/pmmlspecs_v2/pmml_v2_0.html), 2003
- [7] ISO SQL/MM Part 6.  
[http://www.sql-99.org/SC32/WG4/Progression\\_Documents/FCD/fcd-datamining-2001-05.pdf](http://www.sql-99.org/SC32/WG4/Progression_Documents/FCD/fcd-datamining-2001-05.pdf), 2001
- [8] Java Data Mining API, <http://www.jcp.org/jsr/detail/73.prt>, 2003
- [9] Common Warehouse Model (CWM). <http://www.omg.org/cwm>, 2001
- [10] Information Discovery Data Mining Suite.  
<http://www.patternwarehouse.com/dmsuite.htm>, 2002
- [11] I. Bartolini, P. Ciaccia, I. Duci, M. Patella, Y. Theodoridis, M. Vazirgiannis, Euripides Vrachnos. "Advances on PBMS physical representation and query processing", PANDA Internal Report
- [12] Serge Abiteboul, Peter Buneman, Dan Suciu. *Data on the Web. From relational to semistructured data and XML*. Morgan Kaufmann Publishers, San Francisco, California 2000
- [13] Akmal B. Chaudhri, Awais Rashid, Roberto Zicari, editors. *XML Data management. Native XML and XML-Enabled Database Systems*. Addison-Wesley, March 2003
- [14] Oracle9i. XML Database Developer's Guide - Oracle XML DB Release 2 (9.2) October 2002 Part No. A96620-02
- [15] Xquery from the Experts. A guide to the W3C XML Query Language. Howard Katz Editor. Addison Wesley editions



- [16] Leonard Kaufman, Peter J. Rousseeuw. *Finding Groups in Data. An introduction to cluster analysis*. Wiley Interscience publications, USA 1990
- [17] Jiawei Han, Micheline Kamber. *Data Mining. Concepts and Techniques*. Morgan Kaufmann Publishers, USA 2001
- [18] Margaret H. Dunham. *Data Mining. Introductory and Advanced Topics*. Prentice Hall, USA 2003
- [19] XQuery 1.0 and XPath 2.0 Functions and Operators. W3C Working Draft 12 November 2003. <http://www.w3.org/TR/2003/WD-xpath-functions-20031112/>
- [20] Extensible Markup Language (XML) 1.0 (Second Edition). W3C Recommendation 6 October 2000. <http://www.w3.org/TR/2000/REC-xml-20001006>
- [21] XML Schema Part 0, Part 1, Part 2. W3C Recommendation, 2 May 2001. <http://www.w3.org/TR/2001/REC-xmlschema-1-20010502/>



## Ευχαριστίες

Θα ήθελα να ευχαριστήσω ιδιαίτερα τους καθηγητές μου κ. Μιχάλη Βαζιργιάννη και κ. Γιάννη Θεοδωρίδη που με καθοδήγησαν με τις συμβουλές τους και με υποστήριξαν στην εκπόνηση της εργασίας αυτής. Επίσης θα ήθελα να εκφράσω τις ευχαριστίες μου στον πρόεδρο του μεταπτυχιακού μας προγράμματος κ. Ευάγγελο Κιουντούζη που με συνέπεια και σεβασμό στις επιστημονικές αρχές μας τις δίδαξε και καλλιεργώντας την επιστημονική και φιλοσοφική σκέψη, μας έδωσε την ώθηση για τη συνέχεια των σπουδών μας. Τέλος, θέλω να ευχαριστήσω την ΥΔ Ειρήνη Ντούτση για τη βοήθεια και το υλικό που μου παραχώρησε καθώς και τον φίλο και συμφοιτητή μου Κωνσταντίνο για τις χρήσιμες συμβουλές του.

Ειδική αναφορά για την υποστήριξη που μου παρείχε θα πρέπει να γίνει στο χρηματοδοτούμενο από την Ευρωπαϊκή Ένωση έργο της INFORMATION SOCIETY TECHNOLOGIES (IST), PANDA.



BWPEA



80025 75540

