



ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

**Θέματα Διαχείρισης Δεδομένων για Εφαρμογές
Βιοεπιστημών**

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΒΑΣΙΛΑΚΗ ΞΕΦΤΕΡΗ

Επιβλέπων : Τέλης Σαββάλας
Καθηγητής Ε.Μ.Π.

Αθήνα, Μάρτιος 1999

Η σελίδα αυτή είναι σκόπιμα λευκή.



ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Θέματα Διαχείρισης Δεδομένων για Εφαρμογές Βιοεπιστημών

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΒΑΣΙΛΑΚΗ ΞΕΦΤΕΡΗ

Επιβλέπων : Τέλης Σαββάλας
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 18^η Φεβρουαρίου 1999.

(Υπογραφή)

.....
Τέλης Σαββάλας
Καθηγητής Ε.Μ.Π.

(Υπογραφή)

.....
Μάριος Παπαδόπουλος
Καθηγητής Ε.Μ.Π.

(Υπογραφή)

.....
Θωμάς Πίττας
Καθηγητής Ε.Μ.Π.

Αθήνα, Μάρτιος 1999

(Υπογραφή)

.....

ΒΑΣΙΛΑΚΗΣ ΞΕΦΤΕΡΗΣ

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

© 1999 – All rights reserved

Περίληψη

Τα τελευταία δεκαπέντε χρόνια, σημαντικές εξελίξεις στον ερευνητικό χώρο επιστημών του ευρύτερου τομέα της βιολογίας (αποκωδικοποίηση γονιδιωμάτων, αποτελέσματα σύγχρονων πειραμάτων της γενετικής, της μοριακής βιολογίας) έχουν εγείρει νέες προκλήσεις για τη διαχείριση βάσεων δεδομένων και την πληροφορική, αφού έχουν συσσωρεύσει τεράστιο πλήθος πολυειδών δεδομένων. Η παρούσα διπλωματική εργασία ασχολείται με θέματα γύρω από την αποθήκευση και επεξεργασία αυτών. Αρχικά γίνεται προσπάθεια να προσεγγιστούν με τη ματιά του μηχανικού υπολογιστών απαραίτητες έννοιες σχετικές με τις λειτουργίες που λαμβάνουν χώρα στους ζωντανούς οργανισμούς καθώς και τη δομή των υποκειμένων και αντικειμένων τούτων των ενεργειών. Περιγράφονται, έτσι, μεταξύ των άλλων το κεντρικό δόγμα της μοριακής βιολογίας, η μεταγραφή, η μετάφραση, οι DNA sequencers, τα microarray πειράματα, τα νουκλεϊκά οξέα, οι πρωτεΐνες, το κύτταρο. Αναλύονται, επίσης, η ποικιλομορφία και τα ειδικά χαρακτηριστικά των δεδομένων των βιοεπιστημών, ενώ ενδιαφέρουν και τα διάφορα πρότυπα με τα οποία αποθηκεύονται στις υπάρχουσες βάσεις. Απασχολούν λ.χ. οι ιδιότητες ακολουθιών νουκλεοτιδίων ή αμινοξέων, τρισδιάστατων δομών μακρομορίων, βιολογικών μονοπατιών, όπως και το μοντέλο του NCBI, το BIOML και άλλα XML πρότυπα. Τα ερωτήματα που θέτουν οι αντίστοιχοι επιστήμονες και οι εργασίες που χρειάζεται να εκτελούν αποτελούν επιπλέον αντικείμενο μελέτης. Τέτοιες είναι, για παράδειγμα, η σύγκριση ακολουθιών, η φυλογενετική ανάλυση, η sequence assembly, ο προσδιορισμός δομής από ακολουθία. Από τα προηγούμενα γίνεται εφικτό να εντοπιστούν κύρια προβλήματα (π.χ. προέλευση και ενοποίηση δεδομένων), στα οποία πρέπει να δώσει απάντηση η τεχνολογία των βάσεων δεδομένων και προτείνονται ορισμένες πιθανές λύσεις (επεκτάσεις στην SQL, ανάπτυξη νέου μοντέλου και γλώσσας) για περαιτέρω έρευνα. Τέλος, εξετάζονται το εργαλείο BLAST και το Pathways Database System (PathCase) ως προς το σκοπό και το θεωρητικό υπόβαθρο αλλά και πειραματικά.

Λέξεις Κλειδιά: <<.....>>

Η σελίδα αυτή είναι σκόπιμα λευκή.

Abstract

During the last fifteen years, the remarkable developments in life sciences research (sequenced genomes, results of modern experiments in molecular biology and genetics) have brought the database community and computer science to new challenges, because of the abundance and diversity of the data in such sciences. This diploma thesis deals with storage and management issues about these data types. First of all, basic principles about the functions that take place inside living organisms and the structure of the entities involved are approached, with the eye of a computer engineer. The central dogma of molecular biology, transcription, translation, DNA sequencers, microarray experiments, nucleic acids, proteins and the cell are described among others. Then, the heterogeneity and the special characteristics of life sciences' data are analysed, while various standards in which these data are stored in current databases are also considered. For instance, the properties of nucleotide or amino acid sequences and those of the structure of macromolecules, biological pathways, the NCBI data model, BIOML and other XML standards are discussed. What is more, the queries needed by the related researchers and main processes done by them on data are examined. These processes include alignment, phylogenetic analysis, sequence assembly and determination of 3D structure based on sequence. As a consequence of the preceding study, main problems (like data provenance and integration) which the database community has to handle are detected, while probable solutions that have already been proposed (such as SQL extensions or the introduction of a new data model and language) are mentioned. Finally, BLAST and the Pathways Database System (PathCase) are examined concerning not only their purpose and theoretical background, but also their experimental response. This methodology can be used as a guide for carrying out both outdoor and indoor measurements in any spectrum, by making simple changes at the function keys of the Spectrum Analyzer. The usefulness of the methodology is the specification of the procedure of the measurements in steps, from the protection of the personnel and the equipment up to the analytical stages of the measurements procedure.

Keywords: <<.....>>

Η σελίδα αυτή είναι σκόπιμα λευκή.

Πίνακας περιεχομένων

1	Εισαγωγή.....	1
1.1	<Τίτλος που έχει σχέση με τον γενικότερο θέμα που διαπραγματεύεται η διπλωματική>.....	1
1.2	Αντικείμενο διπλωματικής	1
1.2.1	Συνεισφορά	2
1.3	Οργάνωση κειμένου	3
2	Υπόβαθρο.....	4
2.1	< τίτλος έννοιας/θέματος 1>	4
2.2	< τίτλος έννοιας/θέματος 2>	4
3	<τίτλος θεματικής περιοχής 1 στην οποία αναζητούμε να βρούμε τεχνικές και μεθοδολογίες που θα επικεντρωθεί η μελέτη μας>	5
3.1	Προβλήματα και απαιτήσεις.....	5
3.2	<Τίτλος τεχνικής/μεθοδολογίας 1>	6
3.3	<Τίτλος τεχνικής/μεθοδολογίας 2>	6
3.4	Σύνοψη συγκρίσεων	6
4	Μελλοντικές κατευθύνσεις έρευνας	7
4.1	<τίτλος θεματικής περιοχής 1 >.....	7
4.2	<τίτλος θεματικής περιοχής 2 >.....	7
5	Επίλογος.....	8
5.1	Σύνοψη και συμπεράσματα	8
5.2	Μελλοντικές επεκτάσεις.....	8
6	Βιβλιογραφία.....	9

1

Εισαγωγή

1.1 <Τίτλος που έχει σχέση με τον γενικότερο θέμα που διαπραγματεύεται η διπλωματική>

Εδώ αυτή κάνουμε μια γενική περιγραφή του θέματος που διαπραγματεύεται η διπλωματική. Αναφέρουμε τα κύρια χαρακτηριστικά, προσδιορίζουμε σύντομα τις επιμέρους περιοχές και εξηγούμε γιατί έχουν ενδιαφέρον. Η συζήτηση των περιοχών θα πρέπει να προϋποθέτει τον αναγνώστη για το που θα επικεντρωθεί η βιβλιογραφική μελέτη της διπλωματικής, χωρίς ακόμα να αναφερόμαστε συγκεκριμένα στα θέματα της μελέτης.

1.2 Αντικείμενο διπλωματικής

Εδώ αναφερόμαστε συγκεκριμένα στο που θα επικεντρωθεί η βιβλιογραφική μελέτη της διπλωματικής. Αναφέρουμε λεπτομερώς γιατί είναι ενδιαφέρουσα η συγκεκριμένη μελέτη, πάντα σε σχέση με τα θέματα που έχουμε θίξει στην προηγούμενη ενότητα. Δίνουμε επίσης τη μέθοδο με την οποία θα γίνει η μελέτη αυτή.

Μια καλή μέθοδος που προτείνουμε για τέτοιες διπλωματικές εργασίες, αποτελείται από τα εξής βήματα:

α) Σύντομη περιγραφή των θεματικών περιοχών στις οποίες αναζητούμε να βρούμε τεχνικές και μεθοδολογίες στις οποίες θα επικεντρωθεί η μελέτη μας.

β) Σύντομη καταγραφή των προβλημάτων που αντιμετωπίζει κάθε περιοχή.

γ) Λεπτομερής καταγραφή των απαιτήσεων των τεχνικών και μεθοδολογιών (προσοχή: ακόμα δεν έχουμε αναφερθεί στις τεχνικές καθ'αυτές! Θα το κάνουμε αμέσως μετά).

δ) Λεπτομερής περιγραφή των τεχνικών και μεθοδολογιών. Καλό είναι η παρουσίαση να γίνει ανά θεματική περιοχή, όπως αυτές έχουν προσδιοριστεί στο α). Η περιγραφή των τεχνικών/μεθοδολογιών θα πρέπει να δείχνει κατά πόσο μια τεχνική/μεθοδολογία ικανοποιεί τις απαιτήσεις του γ).

ε) Σύνοψη του δ), με παρουσίαση των αποτελεσμάτων σε μορφή με την οποία ο αναγνώστης εύκολα να εντοπίζει τα χαρακτηριστικά των τεχνικών/μεθοδολογιών. Μια τέτοια μορφή παρουσίασης επιτυγχάνεται με χρήση πίνακα, όπου οι γραμμές αντιστοιχούν σε τεχνικές/μεθοδολογίες, οι στήλες σε χαρακτηριστικά, και μέσα στα κελιά συνοψίζονται τα σχόλια για κάθε τεχνική/μεθοδολογία.

Στην ουσία, όπως θα δείτε, η οργάνωση των κεφαλαίων που θα ακολουθήσουν εξυπηρετεί τα α), β), γ),δ) και ε).

Είναι σημαντικό κάποιος που θα διαβάσει την ενότητα αυτή να καταλάβει σε σημαντικό βαθμό τον σκοπό της διπλωματικής σας και τις δυσκολίες οργάνωσης της βιβλιογραφικής μελέτης, χωρίς να είναι αναγκαίο να δει όλα τα άλλα κεφάλαια. Η ενότητα αυτή θέλει πολύ προσοχή και καλύτερα να τη γράψετε αφού έχετε γράψει όλα τα υπόλοιπα κεφάλαια.

1.2.1 Συνεισφορά

Εδώ παραθέτουμε αριθμητικά συγκεκριμένες ενέργειες που κάναμε κατά τη βιβλιογραφική μελέτη και που στην ουσία συνιστούν την συνεισφορά της μελέτης αυτής για τους αναγνώστες. Συνήθως η υποενότητα αυτή έχει την παρακάτω μορφή:

Η συνεισφορά της διπλωματικής συνοψίζεται ως εξής:

1. Μελετήσαμε συστήματα κ.λ.π.
2. Προσδιορίσαμε τα προβλήματα τους.
3. Προτείναμε μια μορφή κατάταξης των τεχνικών σε κατηγορίες με βάση κοινά χαρακτηριστικά.
4. Συζητήσαμε πλεονεκτήματα και μειονκτήματα των τεχνικών, κ.λ.π.
5. ...

1.3 Οργάνωση κειμένου

Εδώ περιγράφουμε τα κεφάλαια της διπλωματικής: 1 πρόταση για το τι θα έχει κάθε κεφάλαιο. Συνήθως η ενότητα αυτή έχει την παρακάτω μορφή (δεν θα σας πάρει πάνω από 1 μεγάλη παράγραφο):

Εργασίες σχετικές με το αντικείμενο της διπλωματικής παρουσιάζονται στο Κεφάλαιο 2 . Το Κεφάλαιο 3 συζητά μπλα μπλα. Στο Κεφάλαιο 4 αναπτύσσουμε κ.λ.π.

2

Υπόβαθρο

Εδώ περιγράφουμε έννοιες/θέματα χρήσιμες για τη διπλωματική που ενδεχομένως ο αναγνώστης να πρέπει να εξοικειωθεί προτού ξεκινήσουμε τη βιβλιογραφική μελέτη. Για παράδειγμα, αν η μελέτη αφορά Θέματα Διαχείρισης Δεδομένων για Εφαρμογές Βιοεπιστημών, θα θέλαμε να υπάρχει μια σύντομη συζήτηση για βιολογικές οντότητες (π.χ. dna, rna), βιολογικές διαδικασίες, κ.λ.π.

2.1 < τίτλος έννοιας/θέματος 1>

<<.....>>

2.2 < τίτλος έννοιας/θέματος 2>

<<.....>>

3

***<τίτλος θεματικής περιοχής 1 στην οποία
αναζητούμε να βρούμε τεχνικές και μεθοδολογίες
που θα επικεντρωθεί η μελέτη μας>***

Εδώ γράφουμε ότι θα μελετήσουμε τεχνικές και μεθοδολογίες στη συγκεκριμένη θεματική περιοχή.

Θα έχουμε τόσα κεφάλαια σαν αυτό, όσα και οι θεματικές περιοχές που εξετάζουμε. Για παράδειγμα, αν η μελέτη αφορά Θέματα Διαχείρισης Δεδομένων για Εφαρμογές Βιοεπιστημών, θα θέλαμε να υπάρχουν 2 τέτοια κεφάλαια: α) Τρόποι αποθήκευσης βιολογικών δεδομένων, β) γλώσσες ερωτήσεων βιολογικών δεδομένων.

3.1 Προβλήματα και απαιτήσεις

Περιγράφουμε τη θεματική περιοχή. Αναλύουμε λεπτομερώς τα προβλήματα της, και τέλος προσδιορίζουμε τα χαρακτηριστικά που πρέπει να έχουν οι τεχνικές και μεθοδολογίες για να λύσουν τα προβλήματά της. Καλό είναι τα χαρακτηριστικά αυτά να είναι συγκεκριμένα και σαφή.

3.2 <Τίτλος τεχνικής/μεθοδολογίας 1>

Εδώ ξεκινάμε την λεπτομερή συζήτηση για κάθε μια από τις τεχνικές/μεθοδολογίες της θεματικής περιοχής. Εκτός από την απλή περιγραφή τους, πρέπει να προσδιορίσουμε τα πλεονεκτήματα και τα μειονεκτήματά τους και επίσης να απαντήσουμε στο ερώτημα κατά πόσο έχουν τα χαρακτηριστικά που θέσαμε πριν.

3.3 <Τίτλος τεχνικής/μεθοδολογίας 2>

<<.....>>

3.4 Σύνοψη συγκρίσεων

Σύνοψη της παρουσίασης των τεχνικών και μεθοδολογιών, με παρουσίαση των αποτελεσμάτων σε μορφή με την οποία ο αναγνώστης εύκολα να εντοπίζει τα χαρακτηριστικά των τεχνικών/μεθοδολογιών. Μια τέτοια μορφή παρουσίασης επιτυγχάνεται με χρήση πίνακα, όπου οι γραμμές αντιστοιχούν σε τεχνικές/μεθοδολογίες, οι στήλες σε χαρακτηριστικά, και μέσα στα κελιά συνοψίζονται τα σχόλια για κάθε τεχνική/μεθοδολογία.

4

Μελλοντικές κατευθύνσεις έρευνας

Εδώ λέμε ότι θα ακολουθήσει παρουσίαση μελλοντικών κατευθύνσεων έρευνας ανά θεματική περιοχή.

4.1 <τίτλος θεματικής περιοχής 1 >

Εδώ περιγράφουμε προβλήματα που δεν έχουν λυθεί από τις τεχνικές/μεθοδολογίες που παρουσιάσαμε στο προηγούμενο κεφάλαιο. Κάνουμε το ίδιο και για τις άλλες θεματικές περιοχές. Τα άλυτα αυτά προβλήματα, αποτελούν στην ουσία προκλήσεις για περαιτέρω έρευνα.

Ακόμα καλύτερα, θα ήταν ωραία να προτείνουμε τρόπους επίλυσης των προβλημάτων αυτών έστω και ως γενική ιδέα!

4.2 <τίτλος θεματικής περιοχής 2 >

<<.....>>

5

Επίλογος

Εδώ λέμε ότι θα συνοψίσουμε την παρουσίαση της διπλωματικής.

5.1 Σύνοψη και συμπεράσματα

Εδώ συνοψίζουμε τα αποτελέσματα της διπλωματικής και περιγράφουμε τα συμπεράσματα που προέκυψαν, αρνητικά και θετικά. Επιβεβαιώνουμε τη συνεισφορά της διπλωματικής στα προβλήματα που αναφέραμε στην εισαγωγή.

5.2 Μελλοντικές επεκτάσεις

Εδώ δίνουμε ιδέες για επέκταση της διπλωματικής.

6

Βιβλιογραφία

- [BBC+99] P.A. Bernstein, Th. Bergstraesser, J. Carlson, S. Pal, P. Sanders, D. Shutt. Microsoft Repository Version 2 and the Open Information Model. To appear in Information Systems 24(2), 1999.
- [BCR94] V. R. Basili, G.Caldiera, H. D. Rombach. The Goal Question Metric Approach. Encyclopedia of Software Engineering - 2 Volume Set, pp. 528-532, John Wiley & Sons, Inc., available at <http://www.cs.umd.edu/users/basili/papers.html>, 1994
- [Dea97] E. B. Dean, "Quality Functional Deployment from the Perspective of Competitive Advantage", available at <http://mijuno.larc.nasa.gov/dfc/qfd.html>
- [JJQV98] M. Jarke, M.A.Jeusfeld, C. Quix, P. Vassiliadis: Architecture and quality in data warehouses, Proceedings CAiSE 98, Pisa, Italy, 1998.
- [JV97] M. Jarke, Y. Vassiliou. Foundations of data warehouse quality – a review of the DWQ project. In Proc. 2nd Intl. Conference Information Quality (IQ-97), Cambridge, Mass., 1997.
- [Orr98] K. Orr. Data quality and systems theory. In Communications of the ACM, 41, 2, pp. 54-57, Feb. 1998.