Πάντειο Πανεπιστήμιο

Τμήμα Δημόσιας Διοίκησης

Διδακτορική Διατριβή

Μαρίας Βαρδάκη του Εμμανουήλ

Θέμα: *Μεταδεδομένα στη Δημόσια Διοίκηση*

Ιανουάριος 2005

Στη μνήμη του πατέρα μου

ΕΥΧΑΡΙΣΤΙΕΣ

Επιθυμώ να ευχαριστήσω την Επιβλέπουσα της διατριβής μου, Αν. Καθηγήτρια κα Β. Μαλινδρέτου – Οικονομάκη και τον Καθηγητή κ. Ι. Βαβούρα για τη τη βοήθεια και το ενδιαφέρον τους για την ολοκλήρωση της διατριβής.

Ιδιαιτέρα θα ήθελα να ευχαριστήσω τον Καθηγητή κ. Χ. Παπαγεωργίου για τις εποικοδομητικές συζητήσεις, την ενθάρρυνση και τη συμπαράστασή του καθ' όλη τη διάρκεια των τελευταίων ετών.

Επίσης θα ήθελα να εκφράσω τις ευχαριστίες μου στην Καθηγήτρια Κα Ο. Χρυσαφίνου, τον Καθηγητή κ. Α. Τασσόπουλο, τον Αν. Καθηγητή κ. Α. Τσάμη και το Λέκτορα κ. Ν. Καραβίτη για τις εύστοχες παρατηρήσεις τους.

ΠΕΡΙΕΧΟΜΕΝΑ

ΕΙΣΑΓΩΓΗ

Η παρούσα διατριβή εισάγει ένα γενικό, σημασιολογικά πλούσιο μοντέλο μεταδεδομένων το οποίο συμβάλλει στην προτυποποίηση των διαδικασιών συλλογής, επεξεργασίας και διάχυσης της στατιστικής πληροφορίας, αποτελώντας έτσι ένα πολύτιμο εργαλείο για τους φορείς δημόσιας διοίκησης στην εξαγωγή ακριβέστερων, ποιοτικών στατιστικών αποτελεσμάτων και δεικτών οικονομικής πολιτικής, με χαμηλότερο κόστος και μειωμένο φόρτο εργασίας του ανθρώπινου δυναμικού τους. Η χρησιμότητα της δυνατότητας εξαγωγής νέων, ποιοτικών δεικτών γίνεται πιο αισθητή στην προσπάθεια χάραξης ενιαίας οικονομικής πολιτικής μεταξύ ενός συνόλου χωρών με πολλές ιδιαιτερότητες, οπότε τα δεδομένα κάθε χώρας πρέπει να είναι συγκρίσιμα και ποιοτικά, ιδιαίτερα σε μια εποχή ένταξης όλο και περισσότερων χωρών στην Ευρωπαϊκή Ένωση με εμφανείς διαφορές σε κουλτούρα και οικονομική πολιτική.

Εκτός όμως από τις απαιτήσεις σε διεθνές επίπεδο, η ανάγκη για συγκρίσιμα στοιχεία παρατηρείται έντονα και στις λειτουργίες των επιμέρους οργανισμών κάθε χώρας. Στατιστικά αποτελέσματα από διαφορετικές πηγές συλλογής και επεξεργασίας δεδομένων συγκεντρώνονται και χρησιμοποιούνται για την αποτύπωση της οικονομικής κατάστασης του κράτους. Επιπρόσθετα, συγκρίσεις παρόντων στοιχείων με αντίστοιχα παρελθόντων χρονικών περιόδων ζητούνται πλέον συνεχώς για κοινωνικοπολιτικούς και οικονομικούς λόγους από τους φορείς δημόσιας διοίκησης και πολιτικής. Κατά συνέπεια, η χρήση των μεταδεδομένων είναι απαραίτητη για τη δυνατότητα σύγκρισης των στοιχείων λαμβάνοντας υπόψη τις όποιες διαφοροποιήσεις στις μεθόδους, στο στατιστικό πληθυσμό, στις νέες ανάγκες έρευνας, κλπ, τόσο κατά τη διάρκεια των ετών όσο και εξαιτίας των διαφορετικών πηγών που χρησιμοποιήθηκαν.

Διαφορετικές προσεγγίσεις στο θέμα των μεταδεδομένων έχουν παρατηρηθεί κατά τα τελευταία χρόνια. Όλες όμως συγκλίνουν στην αναγκαιότητα της χρήσης μεταδεδομένων για την εξαγωγή ποιοτικών αποτελεσμάτων. Οι διαφοροποιήσεις αυτές σχετίζονται με το 'είδος' των μεταδεδομένων που χρειάζονται, με την 'ποσότητα' αυτών, και κυρίως με το αν πρέπει και είναι εφικτό τα μεταδεδομένα να δημοσιεύονται ταυτόχρονα με τα στατιστικά αποτελέσματα.

Στη διατριβή αυτή βασιστήκαμε στις νέες προσεγγίσεις όπως παρουσιάζονται στη πρόσφατη βιβλιογραφία, σύμφωνα με τις οποίες τα μεταδεδομένα είναι απαραίτητα όχι μόνο στον παραγωγό της στατιστικής πληροφορίας αλλά και σε κάθε χρήστη. Ασφαλώς όλοι οι χρήστες δεν χρειάζεται να έχουν στη διάθεσή τους την ίδια ποσότητα μεταδεδομένων, αυτό είναι κάτι που σχετίζεται με την επικείμενη χρήση των δεδομένων – περαιτέρω ανάλυση, απλή κατανόηση και χρήση τους, δυνατότητα σύγκρισής τους ως δευτερογενείς πηγές αποτελεσμάτων, κλπ – αλλά είναι γεγονός ότι υπό κάποια μορφή και ποσότητα η χρήση των μεταδεδομένων είναι απαραίτητη.

Στη συνέχεια, λαμβάνοντας υπόψη την ανάπτυξη της τεχνολογίας, εξετάσαμε τις νεότερες προσεγγίσεις, σύμφωνα με τις οποίες τα μεταδεδομένα πρέπει να είναι δομημένα σε ένα σχήμα, το οποίο μπορεί να αποθηκευτεί στη βάση δεδομένων του πληροφοριακού συστήματος και έτσι να επιτυγχάνεται η αυτοματοποιημένη διαδικασία

επεξεργασίας των δεδομένων ταυτόχρονα με τα αντίστοιχα μεταδεδομένα τους και να δημοσιεύονται αυτόματα.

Η έρευνα για τη χρήση των μεταδεδομένων είχε ξεκινήσει από μία ομάδα εργασίας στην οποία συμμετείχα. Είχαμε εργαστεί στη δημιουργία δομημένων μοντέλων μεταδεδομένων περιορισμένων δυνατοτήτων, καθώς επίσης και τον ορισμό αυτόματων μετασχηματισμών δεδομένων με τη χρήση κατάλληλων μεταδεδομένων. Σχετικές εργασίες με αποτελέσματα εκείνης της περιόδου, δημοσιευμένα σε περιοδικά και πρακτικά διεθνών συνεδρίων Στατιστικής και Πληροφορικής είναι: *[Papageorgiou et.al (1999a),(1999b), (2000a), (2000b), (2001a), (2001b)] και [Pentaris & Vardaki (1998)].*

Από την έρευνα εκείνης της περιόδου προέκυψε η ανάγκη μοντελοποίησης των μεταδεδομένων για το σύνολο της διαδικασίας των στατιστικών δεδομένων και η επεξεργασία κάποιων περαιτέρω μετασχηματισμών για τη δημιουργία οικονομικών δεικτών.

Επίσης, ένα μοντέλο μεταδεδομένων θα έπρεπε να μπορεί να εφαρμοστεί όχι μόνο στα 'κλασσικά δεδομένα' (classical data) αλλά και στα 'συμβολικά δεδομένα' (symbolic data), επειδή η Συμβολική Ανάλυση (βλ. Bock & Diday, 2000) είναι πλέον διαδεδομένη, κυρίως στους Ακαδημαϊκούς κύκλους (δες Κεφάλαιο 6 για περισσότερες λεπτομέρειες).

Το σημαντικότερο επίτευγμα αυτής της διδακτορικής διατριβής είναι η δημιουργία ενός ολοκληρωμένου, σημασιολογικά πλούσιου στατιστικού μοντέλου μεταδεδομένων, το οποίο σχεδιάστηκε για να καλύψει τα σημαντικότερα στάδια της επεξεργασίας στατιστικών πληροφοριών (συλλογή και ανάλυση δεδομένων συμπεριλαμβανομένης της εναρμόνισης, της επεξεργασίας των δεδομένων και των μεταδεδομένων τους και της διάχυσης των αποτελεσμάτων). Το μοντέλο αυτό κρίθηκε απαραίτητο γιατί μέχρι τώρα στη βιβλιογραφία παρουσιάζονται επιμέρους μοντέλα τα οποία έχουν δημιουργηθεί για να εξυπηρετήσουν συγκεκριμένη δραστηριότητα και στάδιο επεξεργασίας της πληροφορίας, όπως κατασκευή ερωτηματολογίου, διάχυση πληροφορίας, κλπ.

Το προτεινόμενο μοντέλο μπορεί να ελαχιστοποιήσει την πολυπλοκότητα των δεδομένων που αποθηκεύονται σε πληροφοριακά συστήματα καθώς και των προβλημάτων συμβατότητας μεταξύ απομακρυσμένων συστημάτων παραγωγής στατιστικής πληροφορίας.

Επιπλέον, εισάγεται ένα πλήθος *μετασχηματισμών/διαδικασιών (operations/ transformations) για τον αυτόματο χειρισμό και των δεδομένων όσο και των μεταδεδομένων*, καθώς επίσης και *μεταδεδομένα που καθορίζουν τη θέση και τύπο αρχείου των στοιχείων στη βάση δεδομένων του συστήματος, ώστε να επιτυγχάνεται ο αυτόματος εντοπισμός και εξόρυξη της ζητούμενης πληροφορίας.*

Βασικό επίσης μέρος της διατριβής αποτελεί και η *μεθοδολογική προσέγγιση για την εναρμόνιση των στατιστικών μεταδεδομένων,* όπου αναπτύσσονται μεθοδολογίες και μετασχηματισμοί για την εναρμόνιση της πληροφορίας όταν παρουσιάζονται ασυνέχειες στις χρονοσειρές ή ασυνέχειες στο χώρο, λόγω διαφορετικών ταξινομήσεων, μεθόδων συλλογής και επεξεργασίας της πληροφορίας, νέων

νομοθετικών ρυθμίσεων, εθνικές διαφορές, κλπ. Οι μετασχηματισμοί αυτοί *συμβάλλουν στη δυνατότητα ομογενοποίησης των δεδομένων τόσο από διαφορετικές χώρες όσο και στην ίδια χώρα αλλά για διαφορετικές χρονικές περιόδους.*

Τα αποτελέσματα της έρευνας για το μοντέλο μεταδεδομένων και τους μετασχηματισμούς παρουσιάστηκαν αρχικά στις εργασίες των *[Papageorgiou et.al (2001c), (2001d)]* και ακολούθως στις δημοσιεύσεις *[Papageorgiou et.al (2002), Vardaki & Papageorgiou (2004), Vardaki (2004a)].*

Στη συνέχεια το προτεινόμενο μοντέλο προσαρμόζεται στις απαιτήσεις της συμβολικής ανάλυσης, προστίθενται αρκετά νέα μεταδομένα που σχετίζονται αποκλειστικά και μόνο με την συμβολική ανάλυση και εντοπίζονται οι σχέσεις των κλασσικών και συμβολικών δεδομένων. *Η χρησιμότητα αυτού του τροποποιημένου μοντέλου έγκειται στο ότι ακόμα και αν χρησιμοποιείται η μέθοδος της συμβολικής ανάλυσης, το βασικό κομμάτι του προτεινόμενου μοντέλου παραμένει με κάποιες προσθήκες μεταδεδομένων και σχέσεις αυτών.* Επίσης, δημιουργούνται και ενσωματώνονται στο μοντέλο, περαιτέρω μετασχηματισμοί ειδικά για τον χειρισμό των συμβολικών δεδομένων.

Επιπρόσθετα, αυτή η δυνατότητα προσθήκης μεταδεδομένων και μετατροπής/ αναπροσαρμογής του μοντέλου για να ικανοποιήσει μία τελείως διαφορετική μορφή ανάλυσης δεδομένων αποδεικνύει την ευελιξία του προτεινόμενου μοντέλου να προσαρμοστεί σε νέες απαιτήσεις. Σχετικές δημοσιεύσεις είναι οι [*Vardaki (2004b), Papageorgiou & Vardaki (2004)].*

Πιο αναλυτικά η διατριβή αποτελείται από επτά κεφάλαια, η διάρθρωση των οποίων έχει ως εξής:

Κεφάλαιο 1: Στο κεφάλαιο αυτό εισάγονται αναγκαίοι ορισμοί για την κατανόηση των μεταδεδομένων, αναλύεται η χρησιμότητά τους και παρατίθενται εφαρμογές χρήσης τους στη δημόσια διοίκηση. Για την καλύτερη κατανόηση της αναγκαιότητας των μεταδεδομένων εξετάζεται η ροή πληροφορίας στους φορείς Δημόσιας Διοίκησης καθώς και οι διαδικασίες λήψης αποφάσεων. Επίσης παρουσιάζονται παλαιότερες και νεότερες προσεγγίσεις σχετικά με τις μεθόδους αποθήκευσης και χρήσης των μεταδεδομένων από τους παραγωγούς στατιστικής πληροφορίας (free-text, μορφή ενιαίων φορμών, προηγμένες δομημένες μορφές). Αξιοποιώντας τις δυνατότητες των πληροφοριακών συστημάτων καταλήγουμε ότι σήμερα ευνοείται η δημιουργία μοντέλου μεταδεδομένων και εξηγούμε τη διαδικασία επιλογών που πρέπει να ακολουθηθεί προκειμένου να προβεί κανείς σε μοντελοποίηση των μεταδεδομένων.

Κεφάλαιο 2: Περιλαμβάνονται η μεθοδολογία εναρμόνισης δεδομένων που αναπτύχθηκε, ώστε να αντιμετωπίσουν περιπτώσεις ασυνεχειών στις χρονοσειρές και στο χώρο. Αναφέρουμε επιγραμματικά τους δύο άξονες εργασίας:

- *Ασυνέχειες σε χρονοσειρές (breaks in time series):* Αντιμετώπιση προβλημάτων που παρατηρούνται κατά την προσπάθεια σύγκρισης δεδομένων της ίδιας χώρας αλλά σε διαφορετικές χρονικές περιόδους λόγω μεταβολής της χρησιμοποιούμενης ταξινόμησης ή της μεθόδου υπολογισμού δεικτών (πχ. αν σε μια χώρα στους

εθνικούς πόρους που διατίθενται για έρευνα προστεθούν, λόγω αλλαγής μεθόδου υπολογισμού, και οι κοινοτικοί πόροι, τότε οι δείκτες πριν και μετά από την αλλαγή αυτή της μεθοδολογίας δεν είναι απολύτως συγκρίσιμοι). Παρατίθενται οι κατάλληλοι μετασχηματισμοί απεικόνισης.

- *Ασυνέχειες στο χώρο (breaks in space):* Αντιμετώπιση προβλημάτων που παρατηρούνται όταν επιχειρούμε να συγκρίνουμε δεδομένα από διαφορετικές χώρες εξαιτίας μεθοδολογικών διαφορών στη δειγματοληπτική έρευνα. Αναλύονται οι περιπτώσεις ασυνεχειών και εξάγονται συντελεστές για την αντιμετώπισή τους.

Τέλος, στο ίδιο κεφάλαιο εξετάζουμε τις συνιστώσες της ποιότητας της στατιστικής πληροφορίας και πραγματοποιείται μία διερεύνηση της δυνατότητας διασφάλισης της ποιότητας βάσει των κριτηρίων της Eurostat και την εξαγωγή δεικτών κυρίως για την αξιολόγηση της συγκρισιμότητας και συνοχής των αποτελεσμάτων. Η έρευνα αυτή χρησιμοποιεί πορίσματα από τα προβλήματα που παρουσιάζονται στην εναρμόνιση δεδομένων για την εξαγωγή δεικτών αξιολόγησης της συγκρισιμότητας και συνοχής δεδομένων από διαφορετικές χώρες και για διαφορετικές χρήσεις.

Κεφάλαιο 3: Στο κεφάλαιο αυτό εξετάζονται βήμα προς βήμα οι προϋποθέσεις, τα κριτήρια και απαραίτητες διαδικασίες και επιλογές προκειμένου να δημιουργήσουμε ένα μοντέλο μεταδεδομένων χρήσιμο στους φορείς δημόσιας διοίκησης, το οποίο θα συμβάλει στην αυτοματοποίηση των διαδικασιών εξαγωγής ποιοτικών στατιστικών αποτελεσμάτων. Οι διαδικασίες αυτές λαμβάνονται υπόψη στο κεφάλαιο 4 στο οποίο περιγράφεται το προτεινόμενο μοντέλο μεταδεδομένων. Εξετάζονται διεξοδικά και με τη σειρά με την οποία πρέπει να ληφθούν υπόψη οι παρακάτω προϋποθέσεις/ διαδικασίες:

- Εξέταση της ροής δεδομένων και μεταδεδομένων σε έναν οργανισμό
- Συστηματική μελέτη του σχήματος δεδομένων των βάσεων ορισμένων ενδεικτικών φορέων. Επιλέξαμε να μελετήσουμε το σχήμα δεδομένων της βάσης σε τρεις αντιπροσωπευτικούς δημόσιους οργανισμούς της χώρας μας: i) Γενική Γραμματεία Πληροφοριακών Συστημάτων του Υπουργείου Οικονομίας και Οικονομικών, ii) Εθνική Στατιστική Υπηρεσία Ελλάδος και iii) Οργανισμό Εκπαίδευσης και Επαγγελματικής Κατάρτισης του ΥΠΕΠΘ. Στη συνέχεια δημιουργήσαμε ένα συνολικό διάγραμμα που περιλαμβάνει τις χρησιμοποιούμενες μεταβλητές και από τους τρεις οργανισμούς. Στόχος είναι να εργαστούμε για την ικανοποίηση του συγκεντρωτικού αυτού διαγράμματος και να αποδείξουμε ότι το μοντέλο μεταδεδομένων και οι μετασχηματισμοί μπορούν να ικανοποιήσουν και τους τρεις οργανισμούς χωρίς να επηρεάζεται από τις διαφορετικές λειτουργίες και ιδιαιτερότητες του κάθε οργανισμού.
- Συγκεκριμενοποίηση των βασικών ιδιοτήτων/κριτηρίων που πρέπει να πληρεί το μοντέλο μεταδεδομένων τόσο από άποψη σχεδίασης όσο και περιεχομένου
- Επιλογή κατηγοριών μεταδεδομένων (ποια μεταδεδομένα θα μοντελοποιηθούν)
- Επιλογή τεχνικής μοντελοποίησης
- Επιλογή γλώσσας μοντελοποίησης

Κεφάλαιο 4: Είναι το βασικό κεφάλαιο της διατριβής στο οποίο αναλύεται και επεξηγείται πλήρως το μοντέλο στατιστικών δεδομένων/μεταδεδομένων που δημιουργήθηκε. Η σημασιολογία του μοντέλου αναλύεται, περιγράφοντας κάθε μέρος της στατιστικής επεξεργασίας. Για την καλύτερη κατανόησή του, το μοντέλο εξετάζεται κατά στάδιο της πληροφορίας που σχηματοποιεί: συλλογή και ανάλυση δεδομένων συμπεριλαμβανομένης της εναρμόνισης, της επεξεργασίας των δεδομένων και των μεταδεδομένων τους και της διάδοσης των αποτελεσμάτων.

Επιπλέον στο ίδιο κεφάλαιο, εισάγονται επτά αλγόριθμοι (μετασχηματισμοί/ διαδικασίες) για τον αυτόματο χειρισμό και των δεδομένων και των μεταδεδομένων και δίνονται οι ιδιότητες και οι περιορισμοί τους, καθώς επίσης και οι εκφράσεις ενεργοποίησής τους στη βάση δεδομένων.

Κεφάλαιο 5: Στο κεφάλαιο αυτό περιγράφονται οι κανόνες που ακολουθούνται προκειμένου να μετατραπεί το μοντέλο μεταδεδομένων σε ένα σχεσιακό σχήμα βάσεων δεδομένων και η δυνατότητα ενσωμάτωσής του, απεικονίζοντας κατάλληλα τα μεταδεδομένα του. Επιπλέον, κατασκευάζεται βάσει αυτών των κανόνων το σχήμα της βάσης δεδομένων/μεταδεδομένων για το συγκεκριμένο μοντέλο μεταδεδομένων που μπορεί να αποτελέσει τη βάση ενός ολοκληρωμένου, βασισμένου σε μεταδεδομένα, πληροφοριακού συστήματος.

Κεφάλαιο 6: Η επέκταση του στατιστικού μας μοντέλου μεταδεδομένων στα συμβολικά δεδομένα εξετάζεται στο παρόν κεφάλαιο. Δημιουργείται ένα τροποποιημένο μοντέλο που συνδυάζει κλασσικά και συμβολικά δεδομένα και παρακολουθεί κάθε διαδικασία δημιουργίας και ανάλυσης συμβολικών δεδομένων, μέχρι και την τελική παρουσίαση αυτών. Τέλος δίνεται ένα παράδειγμα δυνατής παρουσίασης των μεταδεδομένων σε ένα συμβολικό πίνακα (symbolic data table) κατόπιν ενσωμάτωσης του μοντέλου σε μία βάση παραγωγής συμβολικών δεδομένων, καθώς και η δυνατότητα εμφάνισης των μεταδεδομένων σε γραφήματα συμβολικών αντικειμένων.

Κεφάλαιο 7: Περιλαμβάνει τα *συμπεράσματα, προτάσεις και προοπτικές μελλοντικής σχετικής έρευνας*. Παρατίθεται ένας αριθμός προβλημάτων που παρατηρήθηκαν κατά την εκπόνηση της διατριβής και δίνονται ορισμένες προτάσεις επίλυσής τους. Εξετάζεται σε ποιο βαθμό τα αποτελέσματα της διατριβής συνέβαλαν στην επίλυση αυτών καθώς και ποιες είναι οι δυνατότητες περαιτέρω έρευνας στην περιοχή. Τέλος, η συμβολή των αποτελεσμάτων στην προτυποποίηση των διαδικασιών καθώς και της τεχνολογίας περιγράφεται αναλυτικά.

ΚΕΦΑΛΑΙΟ 1

ΜΕΤΑΔΕΔΟΜΕΝΑ ΚΑΙ ΣΤΑΤΙΣΤΙΚΑ ΜΟΝΤΕΛΑ ΜΕΤΑΔΕΔΟΜΕΝΩΝ: ΟΡΙΣΜΟΙ, ΧΡΗΣΙΜΟΤΗΤΑ ΚΑΙ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΔΗΜΟΣΙΑ ΔΙΟΙΚΗΣΗ

1.1 Στατιστικά μεταδεδομένα: ορισμοί και χρησιμότητα

Ο όρος Μεταδεδομένα (metadata) χρησιμοποιείται όλο και συχνότερα στις διάφορες επιστήμες. Ο ευρέως αποδεκτός ορισμός της έννοιας των μεταδεδομένων είναι εκείνος ο οποίος τα θεωρεί ως δεδομένα που περιγράφουν δεδομένα (data about data) [Grossmann (1997), Sundgren (1996)]. Όσον αφορά τη στατιστική, ο ανωτέρω όρος χρησιμοποιείται για να δηλώσει *κάθε πληροφορία σχετικά με πραγματικά στατιστικά δεδομένα*. Ο συγκεκριμένος όρος χρησιμοποιείται τις τρεις τελευταίες δεκαετίες όχι μόνο από φορείς συλλογής και ανάλυσης στατιστικής πληροφορίας, αλλά είναι ευρέως διαδεδομένα και στην επιστήμη της πληροφορικής [Ghosh, 1988, [Kent & Schuerhoff, 1997], [Karge, 1998], [Muller et.al., 1999], [Ozsoyoglou et.al., 1989], [Stonebraker, 1994], [Poole et al., 2002], [Westlake, 1997]..

Τα μεταδεδομένα περιγράφουν αριθμητικά δεδομένα και τις ιδιότητές τους και αποτελούν πληροφορία σχετικά με τις πραγματικές τιμές των μεταβλητών, όπως για παράδειγμα ποια είναι ακριβώς η έννοια των τιμών που παρουσιάζονται στις αναλύσεις των δεικτών, πώς προέκυψαν αυτά τα αποτελέσματα, κλπ. Αποτελούν τη βάση για ανάλυση δεδομένων [Grossmann & Papageorgiou, 1997]. Για παράδειγμα, η μέθοδος συλλογής των στοιχείων, το μέγεθος του δείγματος, το ποσοστό του στατιστικού πληθυσμού που δεν απάντησε, η διαδικασία αντιμετώπισης των σφαλμάτων, οι μεθοδολογικές ιδιαιτερότητες και οι διαδικασίες εναρμόνισης των στοιχείων με τις διεθνείς επιταγές είναι ένα ελάχιστο παράδειγμα των αναγκαίων μεταδεδομένων για την ποιοτική ανάλυση των δεδομένων, την εξαγωγή δεικτών και τη σύγκρισή τους με άλλων χωρών.

Ανάμεσα στις χρήσεις των μεταδεδομένων, μπορούμε να θεωρήσουμε ως αντιπροσωπευτικές τις παρακάτω [Hand, 1993]:

- Την ερμηνεία των αποτελεσμάτων
- Την εγκυρότητα των δεδομένων
- Κατευθύνσεις για την ανάλυση των αποτελεσμάτων. Μεταδεδομένα όπως οι ορισμοί των μεταβλητών για παράδειγμα, μας καθοδηγούν στο ποιες μεταβλητές μάς ενδιαφέρει να αναλύσουμε και ποιες οι ιδιότητές τους.

Επιπρόσθετα, τα μεταδεδομένα έχουν μεγάλη σημασία και για δευτερογενή ανάλυση πληροφορίας, η οποία πραγματοποιήθηκε από άλλους ερευνητές πλην αυτών που έκαναν τη συλλογή και αρχική ανάλυση των δεδομένων. Αυτή είναι κυρίως η περίπτωση των φορέων δημόσιας διοίκησης και ερευνητικών κέντρων [Froeschl & Grossmann, 2001]. Στις περιπτώσεις αυτές, τα μεταδεδομένα προσδίδουν ακριβώς αυτή την αναγκαία πληροφορία για την επεξεργασία δεδομένων που έχουν συλλεγεί

από άλλες (πρωτογενείς) πηγές και δίνουν την δυνατότητα κατανόησης της κωδικοποίησης της πληροφορίας ώστε να καθίσταται εφικτή η σύγκρισή της με άλλες αντίστοιχες πηγές δεδομένων σε εθνικό ή διεθνές επίπεδο.

Επίσης, η δυνατότητα πρόσβασης στις πηγές δεδομένων καθώς και η εμπιστευτικότητα των δεδομένων προστατεύονται με την χρήση κατάλληλων μεταδεδομένων που συνδέουν την πρωτογενή με τη δευτερογενή πηγή (χρήστη) και τη διαδικασία που καλείται να ακολουθήσει [Marsh et.al., 1994].

Διεθνείς οργανισμοί, φορείς δημόσιας διοίκησης και γενικότερα χρήστες και επεξεργαστές δεδομένων συλλέγουν και αποθηκεύουν προς επεξεργασία τεράστιο αριθμό στατιστικών δεδομένων. Λόγω του μεγάλου αριθμού των στατιστικών στοιχείων που τίθενται προς επεξεργασία αλλά και των διαφορετικών αποδεκτών των αναλύσεων, η ταυτόχρονη συλλογή και επεξεργασία των αντίστοιχων μεταδεδομένων είναι πλέον επιτακτική ανάγκη για τη διασφάλιση της ποιότητας των στατιστικών αποτελεσμάτων [Froeschl & Grossmann,2000].

Εξαιτίας του κόστους και επιπρόσθετου φόρτου εργασίας που απαιτεί η σωστή χρήση και δημοσίευση των μεταδεδομένων, παλαιότερα παρέβλεπαν την αναγκαιότητά τους, με αποτέλεσμα να παρουσιάζονται σήμερα σημαντικά προβλήματα στην ανάλυση των χρονοσειρών και την ποιότητα των εξαγομένων δεικτών. Ακόμα και σήμερα αρκετές φορές, η πληροφορία που δίνουν τα μεταδεδομένα επιλεκτικά καταγράφεται ως 'υποσημειώσεις' ή 'μεθοδολογικές αναφορές' που συνοδεύουν στατιστικούς πίνακες.

Οι αναλυτές όμως της στατιστικής πληροφορίας απαιτούν τη μέγιστη δυνατή παροχή πληροφοριών για την ανάλυση και σύγκριση των δεδομένων. Σημαντικές επενδύσεις τόσο σε οικονομικούς πόρους όσο και σε ανθρώπινο δυναμικό έχουν πραγματοποιηθεί τα τελευταία χρόνια για την ικανοποίηση των παραπάνω αναγκών με στόχο τη διαρκή βελτίωση της ποιότητας και κατά συνέπεια της ακρίβειας της στατιστικής πληροφορίας.

Το πρόβλημα εντείνεται σε επίπεδο ανάλυσης στατιστικής πληροφορίας και μεταπληροφορίας από τους διεθνείς οργανισμούς. Στις περιπτώσεις αυτές, δεδομένα τα οποία συλλέχθηκαν από διάφορες χώρες, σε διαφορετικό χρόνο, με διαφορετική μεθοδολογία, με τη χρήση διαφορετικών ορισμών και στατιστικών μεγεθών και μετρήσεων, και διάφορες ταξινομήσεις, χρειάζεται να εναρμονιστούν σύμφωνα με διεθνείς και αποδεκτές ταξινομήσεις, λαμβάνοντας ταυτόχρονα υπόψη το νομικό και μεθοδολογικό πλαίσιο κάθε χώρας.

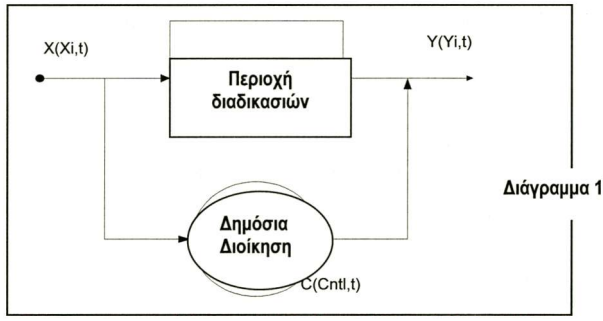
Επομένως είναι απαραίτητη η απαλοιφή αυτής της αδυναμίας και η δυνατότητα καλύτερου χειρισμού των μεταδεδομένων. Η εναρμόνιση των Στατιστικών Μεταδεδομένων είναι ο βασικός πλέον στόχος κάθε Οργανισμού (σε επίπεδο policy-making ή απλά σε επίπεδο χρήστη τελικής πληροφορίας). Επιπρόσθετα, χρήστες της στατιστικής πληροφορίας είναι και όλοι οι ερευνητές και αναλυτές στατιστικών αποτελεσμάτων, όπως Πανεπιστήμια, εταιρείες, ιδιώτες, κλπ. Επειδή όμως τα δεδομένα είναι στενά συνδεδεμένα με τα αντίστοιχα μεταδεδομένα, η ποιότητα των αποτελεσμάτων προϋποθέτει υψηλή ποιότητα των μεταδεδομένων. Είναι σαφές ότι τα μεταδεδομένα θεωρούνται υψηλής ποιότητας όταν δίνουν κάθε πληροφορία

απαραίτητη για την πλήρη και σαφή κατανόηση της έννοιας των δεδομένων. Αυτή η αδυναμία είναι δυνατόν να περιοριστεί με την κατάλληλη χρήση των μεταδεδομένων και τη δημιουργία αλγορίθμων για να γίνονται αυτοματοποιημένα οι αντίστοιχες μετατροπές.

Το βασικό θέμα το οποίο προκύπτει κατά τη χρήση των μεταδεδομένων στη στατιστική πληροφορία είναι να επικεντρώσουμε το ενδιαφέρον, κατ' αρχήν στην επίτευξη συλλογής, παρουσίασης και χρήσης των ίδιων των μεταδεδομένων από τους περισσότερους οργανισμούς ώστε τα αποτελέσματα να είναι συγκρίσιμα αλλά και εύκολο να επεξεργαστούν ξανά από δευτερεύουσα πηγή στατιστικής πληροφορίας μέσα από την υπάρχουσα ροή πληροφορίας και τις σχετικές διαδικασίες.

1.2 Ροή Πληροφορίας στη Δημόσια Διοίκηση και διαδικασίες λήψης αποφάσεων

Η δημόσια διοίκηση και πολιτική μπορεί να γίνει κατανοητή ως σύνολο διαδικασιών με στόχο να μετασχηματίσει τις ιδιότητες μιας δεδομένης ροής εισαγωγής X(Xi, t) των χαρακτηριστικών του εργατικού δυναμικού, σε ένα βελτιωμένο σύνολο ροής παραγωγής Y(Yi, t) (βελτιωμένες δεξιότητες των εργαζομένων). Έτσι ο ρόλος της δημόσιας διοίκησης είναι να ενεργήσει με βάση ένα σύνολο μεταβλητών που λειτουργούν ως μηχανισμοί ελέγχου C(Cntl, t), κατά το χρόνο *t*.



Σύμφωνα με αυτήν την προσέγγιση ο μηχανισμός χάραξης δημόσιας πολιτικής μπορεί να εξεταστεί από την άποψη των ροών εισαγωγής και παραγωγής καθώς και των μεταβλητών ελέγχου. Ο μηχανισμός δημόσιας πολιτικής πρέπει να εξετάσει ένα δεδομένο σύνολο χαρακτηριστικών γνωρισμάτων των ροών εισαγωγής, όπως για παράδειγμα:

- Τα χαρακτηριστικά του εργατικού δυναμικού (δημογραφικά χαρακτηριστικά γνωρίσματα, λαμβάνοντας υπόψη το σύνολο δεξιοτήτων, τη διανομή των επαγγελματικών ειδικοτήτων, το επίπεδο γνώσης και εμπειρίας, τα πολιτισμικά χαρακτηριστικά, κλπ.).
- Τα χαρακτηριστικά γνωρίσματα και τις βραχυπρόθεσμες και μακροπρόθεσμες τάσεις της αγοράς εργασίας (παρούσα κατάσταση στην αγορά, όροι εργασίας,

υπάρχοντα επίπεδα αμοιβών, τεχνολογική πρόοδο και αλλαγές, μετατροπές της δομής της εργασίας, κλπ...);

- Τις τάσεις και μεταβολές του διεθνούς ανταγωνισμού και τις τρέχουσες ανάγκες των επιχειρήσεων.

Τα χαρακτηριστικά των ανωτέρω μεταβλητών αποτελούν ***μεταδεδομένα*** υπό εξέταση, τα οποία είναι αναγκαία για τη χάραξη της πολιτικής μιας χώρας. Επομένως, για τη χάραξη ενιαίας πολιτικής όπως για παράδειγμα στην Ευρωπαϊκή Ένωση, κάθε στοιχείο που θα καταδεικνύει τις ιδιαιτερότητες μίας χώρας και, κατ' επέκταση, μιας οικονομίας είναι πολύτιμη. Αυτά τα στοιχεία παρέχονται με τη μορφή μεταδεδομένων και πρέπει να δίνονται σε ενιαία για όλες τις χώρες μορφή και περιεχόμενο ώστε να είναι συγκρίσιμες.

Αυτές οι πληροφορίες αξιολογούνται για τη λήψη αποφάσεων και χάραξης πολιτικής – βραχυπρόθεσμης και μακροπρόθεσμης, για μία ή σύνολο χωρών – οπότε αντιλαμβάνεται κανείς τη σπουδαιότητα τους.

Η λήψη αποφάσεων ξεκινάει έχοντας όλες τις ανωτέρω μεταβλητές. Διακρίνεται δε σε φάσεις στις οποίες τα μεταδεδομένα παίζουν σημαντικό ρόλο:

Στάδιο του προγραμματισμού της δημόσιας πολιτικής, όπου ένα σύνολο πιθανών εναλλακτικών πολιτικών αξιολογείται και τεκμηριώνεται. Η προετοιμασία είναι μέρος ενός λεπτομερούς σχεδίου για την πολιτική, τον προσδιορισμό των στόχων, τη μεθοδολογία, και τις απαραίτητες ενέργειες που απαιτούνται προκειμένου να εφαρμοστεί το σχέδιο. Δίνεται μια εκτίμηση των αναγκαίων πόρων (ανθρώπινο δυναμικό και υλικά) και προετοιμάζεται μια αξιολόγηση των απαραίτητων δαπανών. Επίσης προετοιμάζεται μια προκαταρκτική αξιολόγηση των δαπανών και του κέρδους της εναλλακτικής λύσης.

Η έκβαση αυτού του σταδίου για ολόκληρο το σύνολο των εναλλακτικών λύσεων λαμβάνεται υπόψη προκειμένου να γίνει αποδεκτό το καλύτερο μίγμα των πολιτικών και για να απορρίψει αυτές που είναι λιγότερο αποτελεσματικές. Τα κριτήρια για την επιλογή μιας πολιτικής συσχετίζονται με τη γενική αποτελεσματικότητα των εναλλακτικών λύσεων, της κόστους/κέρδους εκτίμησης, της γενικής στρατηγικής του φορέα καθώς επίσης και των συγκεκριμένων απαιτήσεων μιας πολιτικής περιοχής. Η επιλογή των πολιτικών υπόκειται στους εκάστοτε περιορισμούς προϋπολογισμών. Η εκ των προτέρων αξιολόγηση των επιλεγμένων πολιτικών, στοχεύει στην προδιαγραφή των κατ' εκτίμηση αποτελεσμάτων, τα οποία αναμένονται από την εφαρμογή της προγραμματισμένης πολιτικής. Θα χρησιμεύσει ως μια βασική γραμμή για την αξιολόγηση της πραγματικών αποτελεσματικότητας και της αποδοτικότητας ολόκληρης της προσπάθειας.

Το επόμενο βήμα περιλαμβάνει τον έλεγχο της πραγματικής εφαρμογής των πολιτικών από την άποψη των χρησιμοποιούμενων πόρων, των πραγματικών δαπανών και κερδών. Αυτό το βήμα στοχεύει να ελέγξει την εφαρμογή των πολιτικών και να κάνει τις απαραίτητες τροποποιήσεις προκειμένου να εξακριβωθεί η επιτυχής αίτησή του.

Η συνεχής αξιολόγηση στοχεύει να ελέγξει και να μετρήσει τις πραγματικές εκβάσεις από τους εφαρμοσμένους στόχους και να τις συγκρίνει με τα αναμενόμενα αποτελέσματα της φάσης προγραμματισμού. Η εξέταση λαμβάνει υπόψη τους πραγματικούς χρησιμοποιούμενους πόρους, το πραγματικό κόστος της χρήσης των πόρων και τα οφέλη από την εφαρμογή των προγραμματισμένων ενεργειών. Ο κύριος στόχος της τρέχουσας αξιολόγησης είναι να αποφασίσει εάν η συνέχεια των προγραμματισμένων δραστηριοτήτων θα παραγάγει τα αναμενόμενα αποτελέσματα ή εάν οι τροποποιήσεις ή ακόμα και η αναστολή του σχεδίου απαιτούνται.

Σαν τελικό στάδιο της εφαρμογής μιας δημόσιας πολιτικής, η εκ των υστέρων αξιολόγηση στοχεύει να εξετάσει το εφαρμοσμένο πολιτικό σχέδιο μετά από την ολοκλήρωση όλων των προγραμματισμένων ενεργειών. Ο στόχος της εκ των υστέρων αξιολόγησης είναι να αναθεωρηθεί ολόκληρο το σχέδιο από την άποψη των κερδών, των δαπανών, και της πραγματικής χρήσης των πόρων (ανθρώπινων και υλικών) και των γενικών εκβάσεων του πολιτικού σχεδίου από την άποψη της πραγματοποίησης των στόχων που έχουν τεθεί. Η αναθεώρηση στοχεύει να συγκρίνει τα πραγματικά αποτελέσματα ενάντια στα αναμενόμενα, όπως αυτά περιγράφηκαν στο αρχικό πολιτικό σχέδιο και στις τροποποιήσεις του. Ένας άλλος στόχος είναι να τεκμηριωθούν τα εφαρμοσμένα σχέδια προκειμένου να δημιουργηθεί ένα αρχείο ιστορίας που μπορεί να χρησιμοποιηθεί ως εργαλείο αναφοράς για τις πιο πρόσφατες πολιτικές πρωτοβουλίες.

Η διαδικασία διάδοσης της πολιτικής είναι ένας 'οριζόντιος στόχος' που λαμβάνει υπόψη όλους τους προηγούμενους στόχους. Εστιάζει στην προώθηση των αποτελεσμάτων δημόσιας πολιτικής στο ευρύ κοινό και στις ενδιαφερόμενες δημόσιες και ιδιωτικές υπηρεσίες. Η κατάλληλη διάδοση αυξάνει τη δημόσια ευαισθητοποίηση σχετικά με το πρόβλημα, ενισχύει τη δημόσια συμμετοχή στη διαδικασία δημόσιας πολιτικής, εξασφαλίζει τη νομιμότητα των ενεργειών που λαμβάνονται από τις δημόσιες υπηρεσίες και βοηθά τις προσπάθειες ελέγχου και αξιολόγησης

Οι μηχανισμοί πολιτικών ανατροφοδότησης και επανασχεδιασμών εκτείνονται σε ολόκληρη τη διαδικασία και ενεργούν ως διορθωτικός παράγοντας στις ανεπάρκειες, οι αλλαγές των παραμέτρων που υπολογίζονται κατά τη διάρκεια των εμποδίων φάσης προγραμματισμού καθώς επίσης και εφαρμογής που προκύπτουν κατά τη διάρκεια της φάσης εφαρμογής δημόσιας πολιτικής. Οι διορθωτικές ενέργειες στοχεύουν να ξανασχεδιάσουν τις πολιτικές για να συναντήσουν τις αναμενόμενες εκβάσεις και για να αυξήσουν την αποδοτικότητα και την αποτελεσματικότητα.

Η χάραξη δημόσιας πολιτικής όμως είναι βασισμένη σε μεγάλη έκταση στη διαθεσιμότητα αξιόπιστων στατιστικών στοιχείων. Σε όλα αυτά τα στάδια πρέπει να κρατούνται τα αρχικά μεταδεδομένα, καθώς και τα νέα μεταδεδομένα που προκύπτουν από την κάθε διαδικασία και να είναι στη διάθεση του δημόσιου φορέα δημιουργίας της πολιτικής, άλλα μεταδεδομένα να είναι διαθέσιμα στον πολιτικό και οικονομικό αναλυτή και άλλα σε άλλες κατηγορίες χρηστών.

Η κατηγοριοποίηση των αναγκαίων μεταδεδομένων που χρησιμοποιούνται από εθνικούς και διεθνείς φορείς δημόσιας διοίκησης μπορεί να πραγματοποιηθεί κατά τρεις τρόπους:

Α) Βάσει του αριθμού των πηγών που λαμβάνονται υπόψη για την εφαρμογή πολιτικής.

Στην περίπτωση αυτή, αν λαμβάνονται υπόψη τα στοιχεία που παράγει μόνο μία πηγή δεδομένων - όπως για παράδειγμα η Στατιστική Υπηρεσία της Ελλάδος - για την εξαγωγή στατιστικών αποτελεσμάτων, τα μεταδεδομένα που χρησιμοποιούνται είναι αυτά που συλλέγονται και διατίθενται από τη συγκεκριμένη πηγή και μόνο. Στην περίπτωση όμως που περισσότερες από μία πηγές δεδομένων χρησιμοποιούνται (για παράδειγμα οι Στατιστική Υπηρεσία Ελλάδος και η Γενική Γραμματεία Πληροφοριακών Συστημάτων του Υπ. Οικονομίας και Οικονομικών για την εξαγωγή συνολικών αποτελεσμάτων για το διασυνοριακό εμπόριο), ενδέχεται να παρατηρηθούν προβλήματα σύγκρισης των δεδομένων εξαιτίας μεθοδολογικών διαφορών στη μέθοδο συλλογής δεδομένων κάθε πηγής, διαφορετικών νομοθετικών ρυθμίσεων και γενικά μεθοδολογικές διαφορές που ενδέχεται να καταστήσουν προβληματική τη σύγκριση.

Β) Ανάλογα με το χρήστη της πληροφορίας κατηγοριοποιώντας τους ανάλογα με τον αν παράγουν ή αν απλά χρησιμοποιούν την πληροφορία.

Γ) Βάσει της χρησιμότητάς τους σε διάφορες φάσεις της διαδικασίας

Το μοντέλο που δημιουργήθηκε και αναλύεται στο κεφάλαιο 4 καθώς και το αναπροσαρμοζόμενο του κεφαλαίου 6 στηρίζεται στον τρίτο αυτό δυνατό τρόπο κατηγοριοποίησης των μεταδεδομένων.

1.3 Προσεγγίσεις της μεταπληροφορίας

1.3.1 Απλά (free-text) μεταδεδομένα υπό μορφή υποσημειώσεων

Στην αρχή τα μεταδεδομένα δεν παρουσιάζονταν καθόλου στο χρήστη, απλά ο παραγωγός της στατιστικής πληροφορίας τα γνώριζε για να διευκολύνει τη δική του εργασία ανάλυσης της πληροφορίας. Πολλές φορές όμως τα παρουσίαζε υπό μορφή υποσημειώσεων (footnotes) στους εξαγόμενους πίνακες. Τα free-text μεταδεδομένα αποτελούσαν μία εύκολη μορφή παρουσίασης της επιπλέον αυτής αναγκαίας πληροφορίας αλλά δεν αποσκοπούσαν στην επεξεργασία αυτής από τα Στατιστικά Πληροφοριακά Συστήματα (ΣΠΣ).

Επειδή όμως σε μεγάλες βάσεις δεδομένων η εισαγωγή πολλών υποσημειώσεων σε ένα πίνακα δεν είναι εφικτή και μπερδεύει το χρήστη, κρατούσαν μόνο τα μεταδεδομένα τα οποία θεωρούνταν απαραίτητα για ένα συνηθισμένο χρήστη, χωρίς ιδιαίτερες απαιτήσεις. Ακόμα όμως και σε αυτή την περίπτωση, όταν κάποιος χρήστης ήθελε να συγκρίνει πίνακες παρόμοιων ερευνών από διαφορετικές έρευνες, χρονικές περιόδους και πηγές, συνήθως οι υποσημειώσεις των πινάκων δεν αναφέρονταν στα

ίδια μεταδεδομένα αλλά σε αυτά που ο εκάστοτε παραγωγός πληροφορίας έκρινε ότι έπρεπε να χορηγηθούν στο χρήστη.

Αργότερα, στην περίπτωση που τα αποτελέσματα στέλνονταν στη Eurostat [Eurostat, 2000a], τον OECD [OECD, 1999] ή άλλους διεθνείς οργανισμούς, έγινε η προσπάθεια να είναι κοινά τα μεταδεδομένα που δημοσιεύονταν από κάθε χώρα. Ακόμα όμως και τότε, η ορολογία δεν ήταν ίδια, ή υπήρχαν διαφορές στους ορισμούς, κάτι που δυσχέραινε την ανάλυση των αποτελεσμάτων.

Παρέμενε πρόβλημα ασφαλώς το γεγονός ότι τα free-text μεταδεδομένα ένα πληροφοριακό σύστημα απλά τα παρουσιάζει στα αποτελέσματά του αλλά δεν τα επεξεργάζεται ώστε να παράγονται αυτόματα τα συγκριτικά αποτελέσματα, κάτι που ακόμη έπρεπε να κάνει το ανθρώπινο δυναμικό του οργανισμού, με κίνδυνο ανθρώπινων λαθών ή παρερμηνείας εννοιών και φυσικά με αυξημένο κόστος.

1.1.1 Πινακοποιημένη μορφή με τη χρήση ενιαίων φορμών (Templates)

Την τελευταία 15ετία, η βιβλιογραφία παρουσιάζει ενδιαφέρουσες αρχικές προτάσεις για την επίλυση του προβλήματος [Papageorgiou et.al., 2000b]. Ο Sundgren (1996, 1999, 2004) προτείνει τη ***χρήση ενιαίων φορμών – πινάκων (templates)*** για τη δομημένη εισαγωγή των μεταδεδομένων, παρατηρώντας ότι «τα μεταδεδομένα που περιγράφουν μία δειγματοληπτική έρευνα και τα δεδομένα που απορρέουν από την έρευνα αυτή, είναι ένας συνδυασμός από δομημένα μεταδεδομένα, όπως για παράδειγμα λίστες κωδικών και περιγραφές καταγεγραμμένων στοιχείων, καθώς και απλών, μη δομημένων μεταδεδομένων, όπως περιγραφές μεταβλητών και διαδικασιών».

Κάθε μεταδεδομένο το οποίο κωδικοποιείται σε μία φόρμα μπορεί εύκολα να αποθηκευτεί σε ένα πληροφοριακό σύστημα με μία σχεσιακή βάση δεδομένων (Relational DataBase Management System (RDBMS)), οπότε απλοποιείται η διαδικασία για τους οργανισμούς και μειώνεται το κόστος [Malvestuto, 1993]. Η μεταπληροφορία που αποθηκεύεται, κωδικοποιείται μέσω μιας τυποποιημένης φόρμας και τέλος όταν εξαχθεί αυτοματοποιημένα μπορεί επίσης να διευκολύνει το χρήστη στην κατανόηση των αποτελεσμάτων και των δεδομένων και να πραγματοποιήσει εύκολα συγκρίσεις [Sundgren, 2000]. Επιπρόσθετα αυτή η μεταπληροφορία μπορεί στη συνέχεια να χρησιμοποιηθεί από μηχανές αναζήτησης (όπως Excite, Google, Infoseek, κλπ) ώστε να συντομευτεί η διαδικασία εντόπισης πληροφοριών.

Κατά συνέπεια, οι φόρμες αυτές καταγράφουν τα μεταδεδομένα με κωδικοποιημένο, ενιαίο τρόπο για όλες τις δειγματοληπτικές έρευνες, επιτρέπουν την αποθήκευση αυτών σε βάσεις δεδομένων με καθορισμένο τρόπο και, στη συνέχεια δίνουν τη δυνατότητα αυτοματοποιημένης εξαγωγής των μεταδεδομένων όταν ο χρήστης το ζητήσει.

Η τεχνική των templates αποτέλεσε μία σαφή βελτίωση από την απλή παρουσίαση των μεταδεδομένων σε υποσημειώσεις που συνόδευαν τους αντίστοιχους πίνακες, λόγω του ότι ενέχει κάποια σχετική δόμηση των μεταδεδομένων υπό κοινό σχεδιασμό

[Froeschl et.al, 2002]. Παρόλα αυτά η πρόταση αυτή παρουσιάζει σχετικούς περιορισμούς [Papageorgiou et.al (2000a)]. Ο πιο σημαντικός είναι ότι αυτές οι φόρμες χρησιμοποιούνται για την απλή εισαγωγή των μεταδεδομένων και όχι για την αυτοματοποιημένη χρήση τους από Πληροφοριακά Συστήματα ανάλυσης δεδομένων, αφού δεν υπάρχει η πληροφορία του *«πώς να χειριστούν τα συστήματα αυτή τη μεταπληροφορία».* Επιπρόσθετα, τα πληροφοριακά συστήματα δεν έχουν την ικανότητα να αντιλαμβάνονται το 'νόημα' της αποθηκευμένης πληροφορίας, οπότε τη χρησιμοποιούν απλά σαν free-text πληροφορία (δες επίσης [Deutsch et al., 1995]). Για παράδειγμα, οι υπολογιστές αντιλαμβάνονται τις ενδείξεις "Euro" και "US Dollar" σαν απλούς χαρακτήρες, χωρίς να συλλαμβάνουν καμία σχέση μεταξύ τους. Σε παρόμοιες λοιπόν περιπτώσεις οι υπολογιστές δεν μπορούν να βοηθήσουν το χρήστη σε οποιαδήποτε ανάλυση δεδομένων, ούτε να τον ειδοποιήσουν για το σφάλμα αν προσπαθήσει, για παράδειγμα, να προσθέσει τα δεδομένα μίας στήλης με μονάδα μέτρησης σε "Euro" και μία στήλη σε "US Dollars".

1.3.2 Μοντέλα μεταδεδομένων (metadata models)

Για την επίλυση των αδυναμιών που παρουσιάζει η χρήση των templates, η ***μοντελοποίηση των μεταδεδομένων*** αποδείχθηκε ως μέσο επίτευξης δυναμικής ανάλυσής τους. Αν η μεταπληροφορία[1] συλλαμβάνεται χρησιμοποιώντας ένα μοντέλο μεταδεδομένων, τότε οι υπολογιστές μπορούν να χρησιμοποιήσουν τη μεταπληροφορία καθ' όλη τη διάρκεια ανάλυσης και συστηματοποίησης των δεδομένων [Froeschl, 1999]. Επιπρόσθετα, κάθε σύγκριση δεικτών θα είναι πιο αποτελεσματική εφόσον θα αποτελεί το υπόβαθρο όλων των επί μέρους αναλύσεων και μελετών.

Σύμφωνα με το παραπάνω σκεπτικό, οι [Papageorgiou et.al, 2000a] και [Grossmann et.al, 1998] έδειξαν ότι ακόμα και ένα απλό minimal μοντέλο είναι αρκετό για να μπορέσουν οι ηλεκτρονικοί υπολογιστές να χειριστούν τα δεδομένα. Ο Froeschl (1997) επίσης, δημιούργησε ένα object-oriented μοντέλο για αποθήκευση και χειρισμό μεταδεδομένων. Τα οφέλη από τη χρήση μοντέλων μεταδεδομένων σε διεθνείς οργανισμούς, φορείς δημόσιας διοίκησης και γενικότερα χρήστες και επεξεργαστές δεδομένων, είναι ποικίλα. Μερικά από τα πιο σημαντικά συνοψίζονται ως εξής:

i) Μειώνει τις περιπτώσεις ανθρώπινου σφάλματος επειδή περιορίζεται στο ελάχιστο η ανάγκη ανθρώπινης παρέμβαση στην ανάλυση των δεδομένων, βελτιώνοντας έτσι την ποιότητα των υπηρεσιών που προσφέρονται από τους δημόσιους φορείς. Για παράδειγμα, το πληροφοριακό σύστημα, έχοντας στη μνήμη το αντίστοιχο μοντέλο μεταδεδομένων, μπορεί να προειδοποιήσει το χρήστη εάν προσπαθήσει να προσθέσει δύο στήλες πινάκων που περιέχουν τιμές σε διαφορετικές μονάδες μέτρησης.

ii) Μειώνει το φόρτο εργασίας του ανθρώπινου δυναμικού των οργανισμών,

---

[1] Στη διατριβή οι όροι «μεταδεδομένα» και «μεταπληροφορία» θα χρησιμοποιούνται εφεξής εναλλάξ με την ίδια έννοια.

iii) Μακροπρόθεσμα μειώνει το κόστος της επεξεργασίας με την ποιοτική και αυτοματοποιημένη εξαγωγή των δεικτών από πληροφοριακά συστήματα.

iv) Επιτρέπει στους χρήστες που δεν έχουν ιδιαίτερες γνώσεις στατιστικής να χρησιμοποιήσουν με ευκολία τα αποτελέσματα αφού επιτρέπει αυτοματοποιημένη ανάλυση και εξαγωγή ποιοτικών αποτελεσμάτων. Αυτή είναι μία σημαντική προσφορά αν σκεφτεί κανείς ότι οι χρήστες του internet δεν είναι συνήθως στατιστικοί.

v) Μειώνει σημαντικά το χρόνο που απαιτείται για την επεξεργασία των δεδομένων, αφού οι υπολογιστές αναλαμβάνουν και την επεξεργασία των μεταδεδομένων.

vi) Αποφεύγονται σφάλματα από παρανόηση των επεξηγήσεων επειδή τα δομημένα μεταδεδομένα είναι αδιαμφισβήτητα θεσπισμένα υπό εθνικούς ή διεθνείς ορισμούς [Froeschl, (1997)].

Θα πρέπει παρόλα αυτά να τονιστεί ότι τα οφέλη της χρήσης των μεταδεδομένων εξαρτώνται από την ποιότητά τους, όπως έχει εξεταστεί από τους Papageorgiou et.al (1999β) και θα αναλυθεί εκτενέστερα στο Κεφάλαιο 2 της διατριβής

ΚΕΦΑΛΑΙΟ 2

ΕΝΑΡΜΟΝΙΣΗ ΣΤΑΤΙΣΤΙΚΩΝ ΜΕΤΑΔΕΔΟΜΕΝΩΝ: ΜΕΘΟΔΟΛΟΓΙΚΗ ΠΡΟΣΕΓΓΙΣΗ ΓΙΑ ΤΗΝ ΕΠΙΤΕΥΞΗ ΠΟΙΟΤΗΤΑΣ ΤΩΝ ΣΤΑΤΙΣΤΙΚΩΝ ΑΠΟΤΕΛΕΣΜΑΤΩΝ.

2.1 Εισαγωγή

Διεθνείς οργανισμοί, κυβερνήσεις και γενικότερα χρήστες και επεξεργαστές δεδομένων άρχισαν να αντιμετωπίζουν τα τελευταία χρόνια το πρόβλημα της ποιότητας των στατιστικών δεδομένων τα οποία αναλύουν.

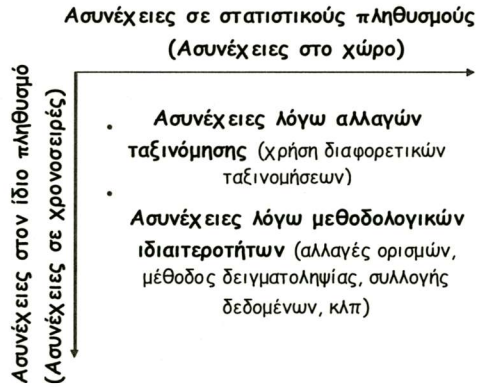
Το πρόβλημα εντείνεται σε επίπεδο ανάλυσης στατιστικής πληροφορίας και μεταπληροφορίας από τους διεθνείς οργανισμούς. Στις περιπτώσεις αυτές, δεδομένα τα οποία συλλέχτηκαν από διάφορες χώρες, σε διαφορετικό χρόνο, με διαφορετική μεθοδολογία, με τη χρήση διαφορετικών ορισμών και στατιστικών μεγεθών και μετρήσεων, και υπό ποικίλες ταξινομήσεις, χρειάζεται να αξιολογηθούν υπό διεθνείς και αποδεκτές μεθόδους και κριτήρια, ώστε τα συλλογικά αποτελέσματα να είναι αξιόπιστα και εναρμονισμένα, χωρίς να υπάρχουν ασυνέχειες στις χρονοσειρές των δεδομένων ώστε να επηρεάσουν την ανάλυση [Papageorgiou et.al., (1999a), (2001d)].

Στην παρούσα ενότητα εξετάζονται μεθοδολογίες που μελετήθηκαν για την εξάλειψη των ασυνεχειών στις χρονοσειρές, στοχεύοντας στη βελτίωση της ποιότητας των αποτελεσμάτων. Οι μετασχηματισμοί που αναπτύχθηκαν για την εναρμόνιση των δεδομένων διακρίνονται στις επόμενες κατηγορίες και περιγράφονται συνοπτικά ως εξής:

i) *Ασυνέχειες σε χρονοσειρές (breaks in time series):* Αντιμετώπιση προβλημάτων που παρατηρούνται κατά την προσπάθεια σύγκρισης δεδομένων της ίδιας χώρας αλλά σε διαφορετικές χρονικές περιόδους λόγω αλλαγής σε κάποια δεδομένα της χώρας (π.χ. κατά την ένωση της Αν. με Δ. Γερμανία τα στοιχεία πριν και μετά την ένωση δεν είναι πλήρως συγκρίσιμα).
Η ύπαρξη ποικίλων ταξινομήσεων και ονοματολογιών καθώς επίσης και των αναθεωρήσεών τους, φέρνει στην επιφάνεια το πρόβλημα της συμβατότητας και της συγκρισιμότητας των στοιχείων που συλλέγονται και που παρέχονται στους διεθνείς οργανισμούς και την EUROSTAT κατά τη διάρκεια των ετών. Στο κεφάλαιο αυτό εισάγεται μία μεθοδολογία με τη βοήθεια ενός διαγράμματος με διεθνείς ταξινομήσεις που χρησιμοποιούνται για διάφορες ενδεικτικές οικονομικές δραστηριότητες εξηγώντας τις σχέσεις μεταξύ τους. Καθορίζεται έτσι ένα κατάλληλο 'ποιοτικό πλαίσιο-στόχος' (Quality Frame - QF) για την μετατροπή αυτών όλων αυτών των ταξινομήσεων που λειτουργούν ως των 'σχετικά πλαίσια' (source frames – SF) στο QF. Επιλέχθηκε η ονοματολογία για οικονομικές δραστηριότητες NACE Rev1 [Eurostat, 1993] ως πλαίσιο-στόχος και καθορίστηκαν οι αναγκαίες συνθήκες για την απεικόνιση των σχετικών πλαισίων (SF) σε αυτήν. Στα πλαίσια του καθορισμού αυτών των συνθηκών χρειάστηκε να μοντελοποιηθούν τα στοιχεία από τα οποία αποτελείται μία ταξινόμηση ώστε να είναι εμφανή τα επί μέρους στοιχεία που συμβάλλουν στη δυνατότητα

απεικόνισης. Η μέθοδος αναπτύσσεται λεπτομερώς στην παράγραφο 2.2 (Δες επίσης [Papageorgiou et.al., 2001d]).

ii) *Ασυνέχειες στο χώρο (breaks in space):* Αντιμετώπιση προβλημάτων που παρατηρούνται όταν καλούμαστε να συγκρίνουμε δεδομένα από διαφορετικές χώρες εξαιτίας μεθοδολογικών διαφορών στη δειγματοληπτική έρευνα.
Για την αντιμετώπιση του προβλήματος αυτού δημιουργήθηκε μια ταξινόμηση των ασυνεχειών που κωδικοποιεί την αιτία της μεθοδολογίας που προκαλεί την ασυνέχεια. Η ταξινόμηση έχει δύο επίπεδα: το πρώτο περιλαμβάνει επτά ευρείες κατηγορίες που είναι χρήσιμες για να χαρακτηρίσουμε τις ασυνέχειες σύμφωνα με τις αιτίες (δες παράγραφο 2.3.2, Πίνακα 1) και ένα δεύτερο πιο λεπτομερές επίπεδο περιλαμβάνοντας τις υποκατηγορίες που αντιστοιχούν στις ιδιαίτερες πτυχές της μεθοδολογίας (δες παράγραφο 2.3.2, Πίνακα 2) και θα χρησιμοποιηθούν ως πεδίο δοκιμής της εναρμόνισης [Papageorgiou et.al, 2001c].
Για τις ασυνέχειες που ταξινομήθηκαν προτάθηκαν συγκεκριμένοι συντελεστές (coefficients) που καταδεικνύουν πόσο σοβαρή ασυνέχεια παρουσιάζεται σε μια σειρά και επίσης πώς οι ασυνέχειες αυτές διανέμονται σύμφωνα με την αιτία. Αυτή η γνώση μπορεί να βοηθήσει το χρήστη να αποφασίσει εάν μια ιδιαίτερη μέθοδος ανάλυσης χρονικής σειράς είναι κατάλληλη. Στην περίπτωση ενός πίνακα που παρουσιάζει δεδομένα πολλών χωρών, τα αποτελέσματα περιλαμβάνουν έναν συντελεστή που προσδιορίζει πόσος διαφορετικές είναι μεθοδολογίες που ακολουθούνται σε δύο χώρες. Επομένως για έναν πίνακα με Ν στοιχεία, η πληροφορία που θα αφορά τις χώρες θα παράγει μια συμμετρική μήτρα, τα στοιχεία της οποίας προσδιορίζουν πόσες μεθοδολογίες για κάθε ζευγάρι των πηγών διαφέρουν η μια από την άλλη. Αυτός είναι ένας γρήγορος τρόπος να αξιολογηθεί εάν οι συγκρίσεις μπορούν να γίνουν και πόσο ασφαλές είναι για ένα χρήστη να εξάγει συμπεράσματα (Δες επίσης [Papageorgiou et.al, 2001c]). Οι ανωτέρω ασυνέχειες παρουσιάζονται διαγραμματικά στο σχήμα:



2.2 Ασυνέχειες σε χρονοσειρές (breaks in time series) και εναρμόνιση

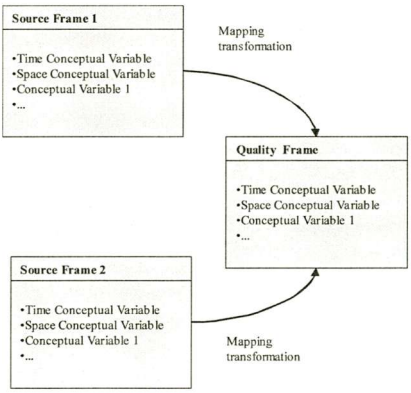
Ένα κλασσικό είδος ασυνεχειών ανάμεσα στα εξαγόμενα στατιστικά δεδομένα είναι η περίπτωση όταν *«δεδομένα που έχουν συλλεχθεί σε μια συγκεκριμένη χρονική περίοδο δεν είναι πλήρως συγκρίσιμα με αντίστοιχα δεδομένα προηγούμενων ή επόμενων χρονικών περιόδων»*. Στην περίπτωση αυτή λέμε ότι έχουμε μία *ασυνέχεια χρονοσειράς (break in time series)*. Τέτοιες περιπτώσεις συναντώνται κυρίως εξαιτίας της αλλαγής χρήσης κάποιων ταξινομήσεων (ονοματολογιών) και την αντικατάστασή τους με κάποιες σχετικές αλλά όχι πανομοιότυπες.

Για να επιτύχουμε την αποκατάσταση τέτοιου είδους ασυνέχειας πρέπει να εφαρμόσουμε κάποιους μετασχηματισμούς των υπαρχόντων δεδομένων (κωδικοποιημένα υπό την παλαιά ταξινόμηση) στην πιο πρόσφατη. Αυτή η διαδικασία καλείται ***μετασχηματισμός απεικόνισης*** (*mapping transformation*) και μπορεί να πραγματοποιηθεί μόνο εάν τα κατάλληλα μεταδεδομένα μοντελοποιηθούν, αναλυθούν και οι σχέσεις εξάρτησής τους περιγραφούν λεπτομερώς.

Στη συνέχεια θα αναπτύξουμε τις βασικότερες αιτίες ασυνεχειών σε χρονοσειρές και τα βήματα για την αντιμετώπισή τους.

2.2.1 Επιλογή πλαισίου αναφοράς και απεικόνισης

Η επίτευξη της εναρμόνισης των ταξινομήσεων, των στατιστικών μεταβλητών αλλά και των μονάδων μέτρησης των υπό εξέταση μεγεθών γίνεται εφικτή με την υιοθέτηση ενός βασικού πλαισίου αναφοράς (μιας συγκεκριμένης και ευρέως αποδεκτής για παράδειγμα ταξινόμησης).



Στην παρούσα διατριβή αυτό το πλαίσιο θα αναφέρεται στο εξής ως «Quality Frame (QF)» και θα αποτελεί τη ζητούμενη ταξινόμηση στην οποία θα πρέπει να αντιστοιχίσουμε όλες τις παλαιότερες. Οποιοδήποτε άλλο πλαίσιο (ταξινόμηση) από τις διαφορετικές πηγές της στατιστικής πληροφορίας - το οποίο θα αναφέρουμε ως Source Frame (SF) στο εξής – θα πρέπει να το απεικονίσουμε και να εναρμονίσουμε τα δεδομένα του με το πλαίσιο αναφοράς QF. Η ύπαρξη και ο καθορισμός των τμημάτων του QF, μας επιτρέπει την παγκόσμια, εναρμονισμένη ανάλυση της στατιστικής πληροφορίας. Επομένως, ο χρήστης δεν θα χρειάζεται απαραίτητα την πληροφορία για τη συνολική διαδικασία συλλογής, αποθήκευσης, φιλτραρίσματος, κλπ, των αρχικών δεδομένων [Papageorgiou et.al, 1999a].

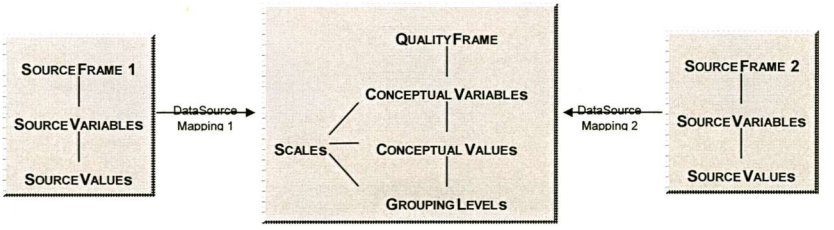
Ένα QF θεωρείται ως ένας πίνακας δεδομένων μαζί με τα αντίστοιχα μεταδεδομένα και πρέπει να κατασκευαστεί με τέτοιο τρόπο ώστε να λάβει υπόψη τις ασυμφωνίες των μεθοδολογιών και τις αδυναμίες εναρμόνισης συγκεκριμένων στατιστικών

μεταβλητών. Επίσης, είναι αναγκαίο να εξετάσει όλες τις υπάρχουσες ταξινομήσεις[2] πριν καταλήξει στην υιοθέτηση της βέλτιστης. Επιπρόσθετα, το επιλεγέν QF πρέπει συχνά να αναπροσαρμόζεται για να καλύπτει τις συνεχώς μεταβαλλόμενες ανάγκες των ερευνητών και των χρηστών. Παρόλ' αυτά, η αρχική του προσέγγιση πρέπει να είναι αρκετά «ελαστική» ώστε να μη χρειάζεται επαναδημιουργία του, αλλά απλά κάποιες προσαρμογές στο συνεχώς μεταβαλλόμενο στατιστικό περιβάλλον.

Το QUALITYFRAME (QF) αποτελεί το σημείο αναφοράς για την ομογενοποίηση των στατιστικών δεδομένων και την εναρμόνισή τους. Πρέπει να είναι ένας αρκετά γενικός πίνακας για να δέχεται όλα τα χαρακτηριστικά και τους διαφορετικούς τρόπους μέτρησης αυτών με ένα τρόπο που να επιτρέπει την εναρμόνιση όλων των πηγών δεδομένων.

Το SOURCEFRAME (SF), ο πίνακας δηλαδή των παρατηρήσεων, είναι μία ελαφρά διαφοροποιημένη μορφή του προηγούμενου, που πιθανώς χρησιμοποιείται από διαφορετικές χώρες.

Το Διάγραμμα 2 παρουσιάζει μία πρώτης μορφής αντιστοιχία/απεικόνιση (Source Mapping) μεταξύ ενός ή περισσοτέρων SFs στο προσχεδιασμένο QF.



Διάγραμμα 2

Κάθε QF και SF αποτελείται από μεταβλητές – *Conceptual variables και Source Variables αντίστοιχα* – και ορίζεται συγκεκριμένη αντιστοιχία μεταξύ των μεταβλητών των δύο Frames. Η αντιστοιχία αυτή πρέπει να έχει οριστεί με τέτοιο τρόπο ώστε η κάλυψη των διαφορετικών στοιχείων αν μπορεί να εκφράζεται με ακρίβεια.

Κρίνεται απαραίτητο ότι για κάθε QF και SF πρέπει να έχουν οριστεί τουλάχιστον τρεις μεταβλητές, από τις οποίες οι δύο πρέπει οπωσδήποτε να εκφράζουν το χρόνο (time) και το χώρο (space) [Papageorgiou et.al, 2001d].

Το σύνολο όλων των πιθανών τιμών μίας μέτρησης για μία συγκεκριμένη μεταβλητή ονομάζεται **κλίμακα (Scale).** Είναι το πρωταρχικό εργαλείο για μέτρηση ενός κοινωνικού, φυσικού, οικονομικού φαινομένου. Δίνουν πεδία τιμών για γενική χρήση. Για παράδειγμα, όταν δημιουργήσουμε μία Scale αρίθμησης, δεν είναι εφικτό να καταλάβουμε αν η μέτρηση αφορά άτομα, ώρες, αντικείμενα. Για να μπορέσουμε να

---

[2] Η έννοιες 'ταξινόμηση' και 'ονοματολογία', παρόλο που ενέχουν σημαντικές διαφορές όσον αφορά τη διάρθρωση, θα χρησιμοποιούνται με την ίδια έννοια.

αντιληφθούμε το αντικείμενο της μέτρησης, πρέπει να την αντιστοιχίσουμε στην υπό εξέταση παρατήρηση (μεταβλητή) ή σε μία μονάδα μέτρησης. Μία κλίμακα μπορεί να είναι:

- Εννοιολογική (categorical): όταν αναφέρεται στην παρατήρηση ενός φαινομένου (πχ. τύπου: discrete, spatial, ordinal, nominal)
- Αριθμητική (Numeric): όταν ορίζονται από μία μονάδα μέτρησης (πχ. τύπου: temporal, ordinal, cardinal, monetary, ratio, interval).

Επίσης, ανάλογα με τον τύπο των τιμών τους χωρίζονται και σε: nominal, ordinal και metric scales.

Οι ονοματολογίες και οι ταξινομήσεις αποτελούν ειδικές περιπτώσεις των nominal scales, οι οποίες ενσωματώνουν στις βασικές τους τιμές (Source και Conceptual values αντίστοιχα) μία ιεραρχία από **Επίπεδα ομαδοποίησης (Grouping Levels)**. Τα επίπεδα ομαδοποίησης δίνουν την ακρίβεια της μέτρησης στην κλίμακα που χρησιμοποιείται. Στην περίπτωση των αριθμητικών κλιμάκων, οι διαφορετικές ακρίβειες μέτρησης είναι δυνατό να δοθούν με απλές φόρμουλες μετατροπής (πχ. από χιλιόμετρα σε μέτρα με απλό πολλαπλασιασμό με 1000). Σε διαφορετικές περιπτώσεις θα χρειαστούν ειδικές φόρμουλες συσχετίσεων.

Η ενδεδειγμένη ονοματολογία-στόχος που θα επιλεγεί σαν βάση σύγκρισης θα πρέπει, μεταξύ άλλων, να έχει τα εξής χαρακτηριστικά:

- Να ακολουθεί όσο το δυνατόν μία γνωστή και ευρέως χρησιμοποιούμενη ονοματολογία, ώστε να διευκολύνει την εναρμόνιση των τρεχόντων δεδομένων.
- Να επιτρέπει την απόλυτη συγκέντρωση οικονομικών δεδομένων χρονοσειρών καθώς και την περαιτέρω ανάλυσή τους σε πιο λεπτομερή επίπεδα. Ένας τρόπος επίτευξης αυτής της ιδιότητας είναι η υιοθέτηση μιας ταξινόμησης με πολλά επίπεδα ανάλυσης (κεφάλαια, υποκεφάλαια, υπό-υποκεφάλαια, τουλάχιστον δηλαδή τρίτου επιπέδου). Τα επίπεδα ανάλυσης θα πρέπει να σταματούν όταν ο βαθμός συγκέντρωσης πέφτει πολύ χαμηλά (για λόγους όπως: εμπιστευτικότητα, κόστος έρευνας, κλπ.)
- Να μπορεί να εναρμονίζεται πλήρως με ιστορικά δεδομένα

Στην παρούσα διατριβή μελετήθηκε η δυνατότητα εναρμόνισης ταξινομήσεων οι οποίες χρησιμοποιούνται για την κωδικοποίηση δεδομένων από διαφορετικές περιοχές εφαρμογής (εργασίας, εκπαίδευσης, οικονομικής δραστηριότητας, κλπ). Στόχος ήταν να μπορέσουμε να απεικονίσουμε όλες αυτές τις ταξινομήσεις άμεσα ή έμμεσα σε μία αποδεχτή και ευρέως χρησιμοποιούμενη ονοματολογία. Επιλέξαμε την ονοματολογία για οικονομικές δραστηριότητες NACE Rev1 (Nomenclature des activités économiques) [Eurostat, 1993] ως το QF πάνω στο οποίο θα απεικονίσουμε όλα τα άλλα SFs και δημιουργήθηκε έτσι ένα διάγραμμα απεικονίσεων 30 Ευρωπαϊκών ταξινομήσεων, διακρίνοντας τέσσερις κατηγορίες δυνατών απεικονίσεων/μετασχηματισμών.

2.2.2 Συσχετίσεις ταξινομήσεων

Στη διαδικασία της εναρμόνισης, οι ταξινομήσεις μπορούν να περιγραφούν ως εξής [Hoffman, (1999), Papageorgiou et.al, (20001d)]:

- **Ταξινομήσεις αναφοράς (reference classifications):** Στην κατηγορία ανήκουν οι ταξινομήσεις που έχουν γίνει επισήμως αποδεκτές και συνιστώνται ως πρότυπα για τη δημιουργία άλλων σχετικών ονοματολογιών ή την αναθεώρηση των υπαρχόντων. Οι διεθνείς ταξινομήσεις (International Statistical Classifications) αποτελούν αναφορά όταν έχουν δημιουργηθεί μετά από διεθνή επίσημη συμφωνία από ένα διακρατικό φορέα όπως για παράδειγμα United Nations Statistical Commission, World Trade Organisation, International Monetary Fund, UNESCO, or International Labour Organisation, κλπ, ανάλογα με την περιοχή εφαρμογής της ταξινόμησης. Τέτοιες ταξινομήσεις είναι οι: Harmonised Commodity Description and Coding System (HS) [Eurostat, 1999], The International Standard Industrial Classification of All Economic Activities (ISIC) [United Nations, 2000], the Central Product Classification (CPC) [Eurostat, 1999], the International Standard Classification in Education (ISCED) [United Nations, 2000] and the International Standard Classification of Occupations (ISCO) [Eurostat, (1999)], [United Nations, (2000)].
- **Παραγόμενες ταξινομήσεις (derived classifications):** στηρίζονται στις ταξινομήσεις αναφοράς, όπως για παράδειγμα η *NACE* η οποία στηρίζεται στον *ISIC*.
- **Σχετιζόμενες/συναφείς ταξινομήσεις (related classifications):** θεωρούνται οι ταξινομήσεις εκείνες που παρέχουν ένα σύνολο οργανωμένων κατηγοριών για τις ίδιες μεταβλητές με την ταξινόμηση αναφοράς, αλλά για το οποίο οι κατηγορίες μπορούν μόνο μερικώς να αναφερθούν σε εκείνοι που καθορίζονται στις ταξινομήσεις αναφοράς, ή αυτός μπορεί μόνο να συνδεθεί στην ταξινόμηση αναφοράς σε συγκεκριμένα επίπεδα δομής. Παράδειγμα αποτελεί η Γαλλική ταξινόμηση NAF [INSEE, 1999] συναφής με τη NACE.

Η εναρμόνιση των υπαρχουσών στατιστικών ταξινομήσεων απαιτεί μια διαδικασία της απεικόνισης των διαφορετικών ταξινομήσεων και των στατιστικών προτύπων σε ένα κοινό πλαίσιο. Αυτό περιλαμβάνει τη χρήση των κοινών εννοιών και της ορολογίας, όπως και την καθιέρωση των συντονισμένων και συμφωνηθέντων πινάκων της σχέσης μεταξύ των κατηγοριών των διαφορετικών ταξινομήσεων, ή μέσω του προσδιορισμού των κοινών λεπτομερών δομικών μονάδων για αυτές τις κατηγορίες. Στην περίπτωση όπου οι διαφορετικές ταξινομήσεις καλύπτουν την ίδια μεταβλητή, η εναρμόνιση απαιτεί μια σαφή κατανόηση της βάσης για και τη φύση των διαφορών, όπως και εάν και το πώς αυτοί αντιστοιχούν στις διαφορετικές ανάγκες των χρηστών.

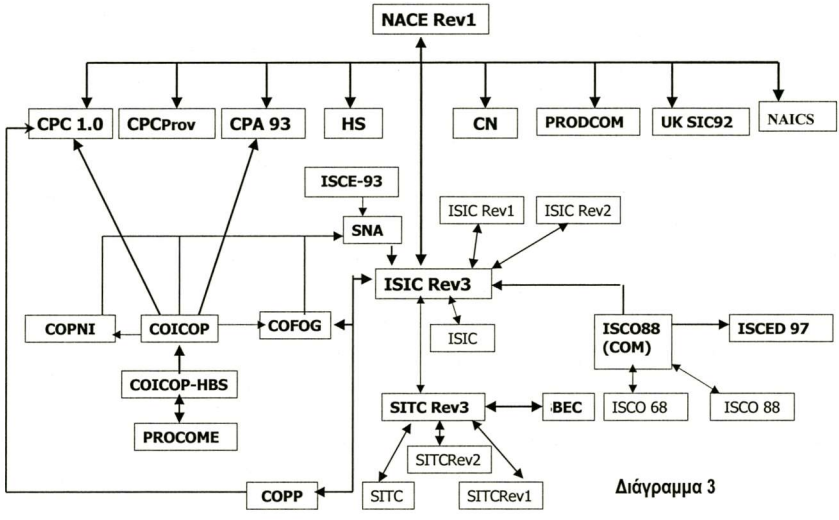
Προκειμένου να προταθεί μια γενική και ευρέως χρησιμοποιούμενη ονοματολογία για να επιτρέψει τις μετατροπές άλλων ονοματολογιών σε αυτή, τα ακόλουθα κριτήρια έχουν προταθεί [Papageorgiou et.al, 2001d]:

- Να χρησιμοποιηθεί μια διεθνώς αποδεκτή ονοματολογία

- Να υιοθετηθεί η πιο πρόσφατη έκδοση
- Να είναι αρκετά ευρεία ώστε να μετατρέψει άλλες υπάρχουσες ονοματολογίες σε αυτή.

Η ονοματολογία NACE Rev1 προτείνεται ως αυτή στην οποία (άμεσα ή μέσω της ταξινόμησης ISIC) διάφορες διεθνείς και εθνικές ταξινομήσεις μπορούν να απεικονιστούν. Η NACE Rev1 προτιμήθηκε από την ISIC, δεδομένου ότι χρησιμοποιείται από τις χώρες της ΕΕ και την EUROSTAT.

Το Διάγραμμα 3 επεξηγεί τις πιθανές σχέσεις της πλειοψηφίας των διεθνών ευρωπαϊκών ταξινομήσεων από διαφορετικούς τομείς εφαρμογής στη NACE Rev1.



Διάγραμμα 3

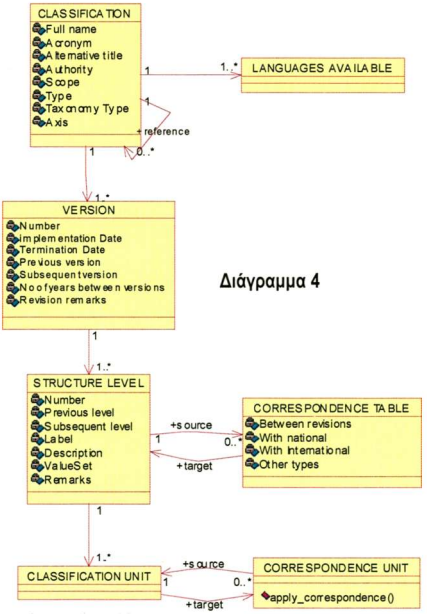
Οι ταξινομήσεις των οικονομικών δραστηριοτήτων (NACE, NAICS, ISIC [Eurostat,1999], και οι αναθεωρήσεις τους), το σύστημα των εθνικών απολογισμών (System of National Accounts (SNA) [OECD, 1998] και των λειτουργικών ταξινομήσεών του (COICOP, COPNI, COFOG και COPP [OECD, 1998]), ταξινομήσεις ερευνών οικιακών προϋπολογισμών (Household Budget Survey - HBS) (COICOP- HBS [OECD 1998] και η προκάτοχός της PROCOME), ταξινομήσεις των προϊόντων και των υπηρεσιών (CPC και οι αναθεωρήσεις αυτής), CPA, HS, SITC και οι αναθεωρήσεις τους BEC, CN, [Eurostat,1999] Prodcom, UK SIC92), ταξινομήσεις για την επαγγελματική εκπαίδευση (ISCED'97) [UN, 2000], Ταξινομήσεις των επαγγελμάτων (ISCO-88 και ρυθμισμένα ISCO-88COM [Eurostat,1999],[UN, 2000]) και ταξινομήσεις που περιγράφουν τη θέση της απασχόλησης (ISCED- 93) [UN, 2000], απεικονίζονται στη NACE Rev1. Περιγραφές και βασικές αρχές ταξινομήσεων παρατίθενται και στο [Hoffmann, 1999].

Αξίζει να αναφερθεί ότι ονοματολογίες όπως οι CPC, CPA, ISIC, CN, HS, PRODCOME list, NAICS και SIC μπορούν άμεσα να απεικονιστούν στη NACE Rev1 με τους απλούς μετασχηματισμούς απεικόνισης, ενώ, άλλες όπως COFOG, SITC, COPP, SNA και ISCO-88 μπορούν να μετασχηματιστούν μέσω του ISIC, ο όποιος συμφωνεί με τη NACE Rev1 σε διψήφιο επίπεδο. Εντούτοις, ταξινομήσεις όπως οι ISCED, COICOP, COPNI και οι σχετικές τους, έχουν τη μερική συμφωνία με τη NACE Rev1 και με το ISIC, και η δυνατότητα απεικόνισής τους απαιτεί τις προσαρμοσμένες και περίπλοκες διαδικασίες μετασχηματισμών που πρέπει να εξεταστούν σε σχέση με τις συγκεκριμένες κοινές μεταβλητές κάθε ταξινόμησης.

2.2.3 Μοντελοποίηση ταξινόμησης

Όπως έχει προαναφερθεί, η κατηγορία μετασχηματισμών ανάμεσα στις διαφορετικές ονοματολογίες που υιοθετούνται από τις διάφορες χώρες ή με τις αντιστοιχίσεις των αναθεωρήσεων της ίδιας ονοματολογίας καλείται *απεικόνιση ή μετασχηματισμός απεικόνισης (mapping transformation).* Αυτοί οι μετασχηματισμοί περιλαμβάνουν όλους εκείνους που απαιτούνται προκειμένου να μετατραπούν τα στοιχεία που συλλέγονται κάτω από τις προηγούμενες ταξινομήσεις σε αυτές που χρησιμοποιούνται σήμερα. Ο καθορισμός τέτοιων λειτουργιών καταδεικνύει ότι είναι δυνατό μέσω των μετασχηματισμών απεικόνισης να μετατραπούν τα υπάρχοντα στοιχεία σε μια διαφορετική μορφή.

Για να επιτύχουμε αυτό τον στόχο, η Ταξινόμηση (Classification) πρέπει να περιγραφεί περαιτέρω από τα συγκεκριμένα χαρακτηριστικά (μεταβλητές) και τις αλληλεξαρτήσεις τους που θα επιτρέψουν τους μετασχηματισμούς μεταξύ τους. Σημειώνουμε ότι οι ταξινομήσεις μπορούν να θεωρηθούν ως μετα-μεταδεδομένα και, ως εκ τούτου, τα μεταδεδομένα τους πρέπει να εξεταστούν.



Διάγραμμα 4

Στο Διάγραμμα 4 περιγράφονται τα μεταδεδομένα που πρέπει να ληφθούν υπόψη στο μοντέλο υπό την κλάση «ταξινόμηση» («Classification» στο μοντέλο του Κεφαλαίου 4). Αποτελεί ένα συμπληρωματικό υπο-μοντέλο για την συμπεριφορά της Ταξινόμησης στην περίπτωση εναρμόνισης.

Στο διάγραμμα παρατηρούμε ότι η κλάση ΤΑΞΙΝΟΜΗΣΗ (classification) περιγράφεται εξετάζοντας το ΠΛΗΡΕΣ ΟΝΟΜΑ το ΑΚΡΩΝΥΜΟ και τα

ΕΝΑΛΛΑΚΤΙΚΑ ΟΝΟΜΑΤΑ, καθώς επίσης και το ΣΚΟΠΟ της, (π.χ. ταξινόμηση προϊόντων), τον ΤΥΠΟ της (αριθμητικός, κείμενο ή μικτός), τον ΑΞΟΝΑ (χρονικός, χωρικός, κλπ.) και τον ΤΑΞΟΝΟΜΙΚΟ ΤΥΠΟ της (ενός επιπέδου, δύο, τριών, κλπ).

Κάθε ΤΑΞΙΝΟΜΗΣΗ μπορεί να είχε μεταφραστεί σε μια ή περισσότερες ΓΛΩΣΣΕΣ (Languages Available). Επιπλέον, μια ΤΑΞΙΝΟΜΗΣΗ μπορεί να έχει περισσότερες από μια ΕΚΔΟΣΕΙΣ (Version). Για κάθε ΕΚΔΟΣΗ πρέπει να ξέρουμε τουλάχιστον την ημερομηνία που εισήχθη, τον αριθμό ετών μεταξύ των αναθεωρήσεων, πότε σταμάτησε η χρήση της και ποιος ήταν ο σκοπός της αναθεώρησής της. Επιπλέον, διάφορα ΕΠΙΠΕΔΑ ΔΟΜΗΣ (Structure Level) εξετάζονται καθώς επίσης και τα επιτρεπόμενα ΣΥΝΟΛΑ ΤΙΜΩΝ (ValueSet).

Στην περίπτωση που έχουν εφαρμοστεί μετασχηματισμοί απεικόνισης προκειμένου να αποκατασταθεί η σχετική ασυνέχεια προκύπτει ένας ΠΙΝΑΚΑΣ ΑΝΤΙΣΤΟΙΧΙΣΗΣ (Correspondence Table) μεταξύ των δύο υπό εξέταση ΤΑΞΙΝΟΜΗΣΕΩΝ. Τέτοιοι μετασχηματισμοί μπορούν να εφαρμοστούν μεταξύ των ΕΚΔΟΣΕΩΝ της ίδιας ΤΑΞΙΝΟΜΗΣΗΣ ή των διαφορετικών ταξινομήσεων του ίδιου ή διαφορετικού τομέα εφαρμογής, κάτι το οποίο είναι δυσκολότερο.

Εν πάσει περιπτώσει, οι πιθανές απεικονίσεις πρέπει να εξεταστούν σύμφωνα με κάθε ΜΟΝΑΔΑ τής υπό μετατροπή ταξινόμησης (Source Classification Unit) και της ταξινόμησης-στόχο (Target Classification Unit). Η ΜΟΝΑΔΑ ΑΝΤΙΣΤΟΙΧΙΣΗΣ (Corresponding Unit) εκφράζει τη σχέση μεταξύ ενός στοιχείου της υπό μετατροπή ταξινόμησης και ενός αντίστοιχου στοιχείου της ταξινόμησης-στόχος.

2.2.4 Κατηγορίες δυνατών μετασχηματισμών απεικόνισης/ ταξινομήσεων

Όταν καλούμαστε να εφαρμόσουμε μετασχηματισμούς απεικόνισης τουλάχιστον οι ακόλουθες περιπτώσεις πρέπει να εξεταστούν:

- η εισαγωγή μιας νέας διεθνούς ονοματολογίας στον ίδιο τομέα της εφαρμογής
- μια αναθεώρηση της υπάρχουσας
- η εισαγωγή μιας διεθνούς ονοματολογίας σε έναν διαφορετικό τομέα εφαρμογής
- η εισαγωγή εθνικών ονοματολογιών

Θεωρώντας τη NACE Rev 1 ως ονοματολογία-στόχο, οι κατηγορίες των δυνατών απεικονίσεων/μετασχηματισμών που μπορούμε να συναντήσουμε ταξινομούνται σε τέσσερις βασικές ομάδες σύμφωνα με τη δομή της ονοματολογίας που πρέπει να μετατραπεί:

Ομάδα 1: περιλαμβάνει τις μετατροπές των ονοματολογιών που παράγονται με την υιοθέτηση της δομής και των κατηγοριών της ταξινόμησης NACE Rev 1, και έπειτα ενδεχομένως μετατρέπονται ανάλογα με τις ανάγκες μιας χώρας παρέχοντας κάποιες πρόσθετες λεπτομέρειες (όπως NAF [INSEE, 1999] ή μια συνάθροιση κατηγοριών.

Ομάδα 2: περιλαμβάνει τους μετασχηματισμούς των ονοματολογιών που παρέχουν ένα σύνολο οργανωμένων κατηγοριών για τις ίδιες μεταβλητές με τη NACE Rev1, αλλά

οι κατηγορίες τους μπορούν μόνο μερικώς να αναφερθούν σε εκείνες που καθορίζει η NACE Rev1, (π.χ NAICS [Eurostat, 1999]).

Ομάδα 3: περιλαμβάνει τους μετασχηματισμούς των αναθεωρήσεων (revisions) των ήδη υπαρχουσών ονοματολογιών.

Ομάδα 4: αφορά τους μετασχηματισμούς απεικόνισης των ονοματολογιών στις οποίες οι στατιστικές μονάδες είναι διαφορετικές από εκείνες της NACE Rev1 (π.χ η περίπτωση της CPC που είναι για προϊόντα ενώ η NACE Rev1 για οικονομικές δραστηριότητες).

Σε τέτοιες περιπτώσεις τα ακόλουθα βήματα πρέπει να εξεταστούν:

- Επιλογή των βασικών μεταβλητών της ταξινόμησης για τις οποίες πρέπει να ισχύσουν οι μετασχηματισμοί
- Προσδιορισμός των βασικών στατιστικών μονάδων της ταξινόμησης για την οποία η βασική μεταβλητή (-ες) μπορεί να περιγραφεί. Ως βασικές στατιστικές μονάδες θεωρούμε τις μονάδες που παρατηρούνται και αναφέρονται μόνο σε μία κατηγορία της ταξινόμησης και δεν σχετίζονται ούτε αναφέρονται σε άλλη στατιστική μονάδα.
- Προσδιορισμός των κανόνων για να συνδεθούν οι διαφορετικές στατιστικές μονάδες. Είναι απαραίτητο να υπάρξουν οι κανόνες για το μετασχηματισμό των υπό εξέταση στατιστικών μονάδες στην αντίστοιχη βασική στατιστική μονάδα της NACE Rev1.

Επιπλέον, πρέπει να τονιστεί ότι οι μετασχηματισμοί ισχύουν όχι μόνο για τα υπάρχοντα δεδομένα αλλά και για τα αντίστοιχα μεταδεδομένα τους [Papageorgiou et.al., 2000a] δεδομένου ότι τα μεταδεδομένα παρέχουν όλες τις απαραίτητες πληροφορίες για την σαφή κατανόηση των δεδομένων. Συνεπώς, προκειμένου να επιτραπούν οι διαδικασίες απεικόνισης, το αντικείμενο ταξινόμησης πρέπει να περιγραφεί από τα συγκεκριμένα χαρακτηριστικά (μεταβλητές) που θα επιτρέψουν τους μετασχηματισμούς μεταξύ τους και επομένως, τα αντίστοιχα αντικείμενα μεταδεδομένων πρέπει να ληφθούν υπόψη [Neuchatel Group, 2000], [Papageorgiou et.al., 2001d].

2.3 Ασυνέχειες στο χώρο (breaks in space) και εναρμόνιση

2.3.1 Εντοπισμός του προβλήματος

Οι ασυνέχειες εμφανίζονται συχνά στις χρονοσειρές και περιλαμβάνουν τις αλλαγές στα πρότυπα και τις μεθόδους που έχουν επιπτώσεις στη συγκρισιμότητα στοιχείων κατά τη διάρκεια του χρόνου, δεδομένου ότι καθιστούν τα στοιχεία πριν και μετά από την αλλαγή όχι πλήρως συγκρίσιμα.

Συνεπώς, η απόκτηση πληροφορίας για τις ασυνέχειες είναι ένα αρκετά σημαντικό κομμάτι των στατιστικών μεταδεδομένων και περιλαμβάνει πληροφορίες για τις υπάρχουσες ασυνέχειες στις χρονοσειρές που δείχνουν την αιτία της ασυνέχειας καθώς και το χρόνο που εμφανίστηκαν. Τέτοια πληροφορία μπορεί να καταστήσει το χρήστη στοιχείων προσεκτικό για οποιαδήποτε λανθασμένα συμπεράσματα μπορεί να έχουν εξαχθεί λόγω της ασυνέχειας.

Μερικές ασυνέχειες, εντούτοις, μπορούν να δημιουργήσουν περισσότερα προβλήματα από άλλες, συνεπώς πιστεύουμε ότι πρέπει να υπάρχει κάποια ένδειξη της σοβαρότητας της ασυνέχειας για την οικονομία ή τη χάραξη κάποιας πολιτικής.

Στην παρούσα διατριβή εξετάζονται και κωδικοποιούνται σε μία ταξινόμηση οι μεθοδολογικές ιδιαιτερότητες που μπορούν να προξενήσουν μία ασυνέχεια. Η προτεινόμενη ταξινόμηση έχει δύο επίπεδα: το πρώτο περιλαμβάνει επτά ευρείες κατηγορίες που είναι χρήσιμες όταν θέλουμε να χαρακτηρίσουμε την ασυνέχεια σε σχέση με την αιτία δημιουργίας της. Το δεύτερο πιο λεπτομερές επίπεδο περιλαμβάνει τις υποκατηγορίες των μεθοδολογικών προβλημάτων που παρουσιάζονται και θα χρησιμοποιηθούν για την εναρμόνιση. Για τις ασυνέχειες χρονοσειράς προτείνονται παράγοντες στάθμισης που δείχνουν το μέγεθος ασυνέχειας της σειράς καθώς και πώς οι ασυνέχειες κατανέμονται με βάση την αιτία δημιουργίας τους (βλ. και [Papageorgiou et.al, 2001c]).

Αυτή η πληροφορία μπορεί να βοηθήσει το χρήστη να αποφασίσει εάν μια ιδιαίτερη μέθοδος ανάλυσης χρονολογικής σειράς είναι κατάλληλη ή όχι.

Στην περίπτωση που θέλουμε να εξετάσουμε ασυνέχειες στο χώρο για περισσότερες από δύο χώρες ταυτόχρονα, απαιτείται η κατασκευή ενός πίνακα που εξετάζει ταυτόχρονα πολλές χώρες και τα αποτελέσματα περιλαμβάνουν έναν συντελεστή που προσδιορίζει πόσο διαφορετικές είναι οι μεθοδολογίες που ακολουθούνται ανά δύο χώρες. Επομένως για να εξεταστούν στοιχεία από Ν χώρες θα παραχθεί ένας συμμετρικός πίνακας τα στοιχεία του οποίου προσδιορίζουν πόσες μεθοδολογίες διαφέρουν η μια από την άλλη για κάθε ζευγάρι από τις Ν χώρες. Αυτός είναι ένας γρήγορος τρόπος να αξιολογηθεί εάν οι συγκρίσεις είναι εφικτό να γίνουν και πόσο ασφαλές είναι να βγάζουμε συμπεράσματα από ένα συγκεκριμένο σύνολο δεδομένων χωρίς να χρειάζεται να συνοδεύουμε την ανάλυσή μας με υπεράριθμες σελίδες στατιστικής μεθοδολογίας.

2.3.2 Μεθοδολογία που εφαρμόστηκε - Κατηγοριοποίηση ασυνεχειών και συγκρισιμότητα

Υποθέτουμε ότι η υπηρεσία που είναι υπεύθυνη για τη συλλογή στοιχείων έχει περιγράψει και καταγράψει ακριβώς τις αλλαγές στις διάφορες στατιστικές μεθοδολογίες. Έχουμε υιοθετήσει μια ταξινόμηση δύο επιπέδων βασισμένη στον πρότυπο κατάλογο μεταδεδομένων του ΟΟΣΑ (OECD) [OECD,1997]. Το πρώτο επίπεδο αποτελείται από τις ευρείες κατηγορίες ασυνεχειών/προβλημάτων που εμφανίζονται στον Πίνακα 1 (βλ. και [Papageorgiou et.al, 2001c]).

Πίνακας 1: Πρώτο επίπεδο κατηγοριοποίησης ασυνεχειών

Κωδ	Είδος	Περιγραφή
1	Πηγή	Χαρακτηριστικά της πηγής όπως πχ. περιοδικότητα.

2	Έννοιες και κάλυψη	Ορισμοί, περίοδος αναφοράς, στατιστικός πληθυσμός / εξαιρέσεις, γεωγραφική κάλυψη.
3	Πρότυπα	Ακολουθούμενα πρότυπα, ταξινομήσεις, αποκλίσεις από τα διεθνή πρότυπα.
4	Συλλογή δεδομένων	Μέθοδος συλλογής, Κατάλογος εγγραφών, κλπ
5	Χειρισμός δεδομένων	Συναθροίσεις, ρυθμιστικές μέθοδοι, κλπ
6	Ποιότητα	Σφάλμα δειγματοληψίας, σφάλμα μέτρησης, και ποσοστό μη απόκρισης
7	Άλλα	Όσα δεν μπορούν να κατηγοριοποιηθούν παραπάνω

Για τις ασυνέχειες στις χρονοσειρές ταξινομούμε τις πληροφορίες που παρέχονται από τους οργανισμούς συλλογής πληροφορίας. Εντούτοις, για να προσδιορίσουμε τις ασυνέχειες στο χώρο χρησιμοποιούμε το δεύτερο επίπεδο της ταξινόμησης σε έναν πίνακα ελέγχου (Πίνακας 2) και οι πληροφορίες για κάθε υποκατηγορία συγκρίνονται για κάθε είδος ασυνέχειας.

Μετά από έρευνα εντοπίστηκε ότι μερικά είδη ασυνεχειών στο χρόνο δεν μπορούν να δημιουργήσουν ασυνέχειες στο χώρο (όπως τη γεωγραφική κάλυψη) και αντίστροφα (όπως ο χρόνος αναφοράς). Στον Πίνακα 2 με «+» σημειώσαμε τις περιπτώσεις οι οποίες μπορούν να προκαλέσουν ασυνέχεια στο χώρο ή στο χώρο αντίστοιχα και με « - » τις περιπτώσεις που μπορεί να δημιουργηθεί σχετικό σφάλμα.

Πίνακας 2: Ανάλυση ασυνεχειών σε δεύτερο επίπεδο

Code	LEVEL I name	Sub-Code	Time	Space	LEVEL II name
1	Source (Πηγή)	11	+	+	Type of source (Τύπος πηγής)
		12	+	-	Periodicity (Περιοδικότητα)
		13	-	+	Reference period (Περίοδος αναφοράς)
		14	+	+	Unit of measurement (Μονάδα μέτρησης)
2	Concepts & coverage (Έννοιες & κάλυψη)	21	+	+	Indicator Definition (Καθορισμός δεικτών)
		22	+	-	Geographical Coverage (Γεωγραφική κάλυψη)
		23	+	+	Classification Coverage (Κάλυψη ταξινόμησης)
		24	+	+	Statistical population (Στατιστικός πληθυσμός)
		25	+	+	Population Exclusions (Αποκλεισμοί πληθυσμού)
3	Standards (Πρότυπα)	31	+	+	Standards framework (Πλαίσιο προτύπων)
		32	+	+	Departure from framework (εκτός πλαισίου)

		33	+	+	Classifications (Ταξινομήσεις)
4	Data collection (Συλλογή δεδομένων)	41	+	+	Reporting unit (Μονάδα αναφοράς)
		42	+	+	Reporting method (Μέθοδος αναφοράς)
		43	+	-	Master list (Κύριος κατάλογος)
		44	+	+	Sample size (Μέγεθος δείγματος)
5	Data Manipulation (Χειρισμός δεδομένων)	51	+	+	Aggregation method (Μέθοδος συνάθροισης)
		52	+	+	Grossing-up method
		53	+	+	Seasonal adjustment (Εποχιακή ρύθμιση)
6	Quality (Ποιότητα)	61	+	+	Sampling error (Λάθος δειγματοληψίας)
		62	+	+	Non-response rate (Ποσοστό μη ανταπόκρισης)
7	Other (Άλλο)	70	+	+	Other (Άλλο)

I) Επίπεδο σημαντικότητας της ασυνέχειας

Προφανώς, τα αποτελέσματα των μεθοδολογικών διαφορών μπορούν να είναι ευρέως φάσματος. Κάποια ίσως δεν παράγουν ουσιαστική ασυνέχεια ενώ άλλα (όπως οι μεταβολές στους ορισμούς) να έχουν σημαντικές συνέπειες σχετικά με τη συγκρισιμότητα στοιχείων. Επομένως πρέπει να εισαγάγουμε έναν παράγοντα στάθμισης w της σοβαρότητας (severity) για κάθε ασυνέχεια i που υπέβαλε η πηγή (χώρα) j, ο οποίος παίρνει τις τιμές για παράδειγμα:

$$w_{ij} = \begin{cases} 0 & no\ break \\ 0.5 & a\ break\ of\ low\ severity \\ 1.0 & a\ break\ of\ medium\ severity \\ 1.5 & a\ break\ of\ high\ severity \end{cases}$$

Θεωρούμε ότι θα ήταν σημαντικό κάθε χώρα που υποβάλλει έκθεση ασυνεχειών να δίνει και κάποια ένδειξη για τη δριμύτητα της ασυνέχειας (και μερικά είδη) που μπορεί να μας επιτρέψει να επιλέγουμε την κατάλληλη στάθμιση. Όταν αυτό δεν είναι διαθέσιμο και μία ασυνέχεια απλά αναφέρεται, επιλέγουμε τον παράγοντα στάθμισης βασισμένο στο είδος της ασυνέχειας (βλ. και [Papageorgiou et.al, 2001c])..

Επειδή δεν δίνονται όμως αυτά τα στοιχεία, και επειδή ο βασικός στόχος είναι να εξαχθούν αποτελέσματα που να μπορούν να χρησιμοποιηθούν άμεσα, επιλέξαμε για την περαιτέρω μελέτη να απλουστεύσουμε τη συνάρτηση καταγράφοντας απλά την ύπαρξη ή όχι ασυνέχειας ως εξής:

$$w_{ij} = \begin{cases} 0 & no\ break \\ 1 & break \end{cases}$$

II) Χαρακτηρισμός ασυνεχειών κατά είδος και συχνότητα εμφάνισης

Θα χαρακτηρίσουμε στη συνέχεια τις ασυνέχειες σύμφωνα με το είδος και την χώρα και θα εξετάσουμε την επίδραση των ασυνεχειών στις συγκρίσεις των χωρών με την εξαγωγή κατάλληλων συντελεστών.

Για κάθε αριθμό διαφορών στις χρονοσειρές ένας συντελεστής r_j που περιγράφει τη συχνότητα εμφάνισης των ασυνεχειών είναι:

$$r_j = \frac{\sum_{i=1}^{L_j} w_{ij}}{T-1} \qquad (3)$$

όπου L_j είναι το σύνολο των μεθοδολογικών διαφορών που ανέφερε η j χώρα, T είναι το μήκος της χρονοσειράς και $T\text{-}1$ είναι ο αριθμός των ζευγών από τις διαδοχικές περιπτώσεις ερευνών που μπορούν να συγκριθούν.

Εν συνεχεία, θα χαρακτηρίσουμε τις ασυνέχειες σύμφωνα με το είδος. Υποθέτουμε ότι ο υπεύθυνος οργανισμός για τη συλλογή δεδομένων και εξαγωγή αποτελεσμάτων σε κάθε χώρα έχει περιγράψει με ακρίβεια τις αλλαγές στη στατιστική μεθοδολογία. Για έναν πίνακα των στοιχείων με μεταβλητές το χρόνο και τη διάσταση χρησιμοποιούμε το πρώτο επίπεδο της ταξινόμησης ασυνεχειών του Πίνακα 1

Συμβολίζουμε με k τις επτά κατηγορίες ασυνεχειών του Πίνακα 1 (επίπεδο Ι), δηλ. k=1,2,…,7

Σύμφωνα με τον Πίνακα 2, κάθε μία από τις επτά αυτές κατηγορίες την αναλύσαμε σε υποκατηγορίες. Συνεπώς, έστω ότι κάθε κατηγορία του επιπέδου ΙΙ μπορεί να περιλαμβάνει D_k υποκατηγορίες επιπέδου ΙΙ.

Επομένως ορίζουμε τον δείκτη p_k ως:

$$p_k = \frac{\sum_{j=1}^{N}\sum_{i=1}^{D_k} w_{ij}}{\sum_{j=1}^{N}\sum_{i=1}^{D} w_{ij}} \qquad (4)$$

όπου $D = \sum D_k$ το σύνολο των κατηγοριών επιπέδου ΙΙ και Ν το σύνολο των υπό εξέταση χωρών.

Αντίστοιχος δείκτης μπορεί να υπολογιστεί ανάλογα πλέον με τη χώρα j που εξετάζουμε ως εξής:

$$p_j = \frac{\sum_{i=1}^{D} w_{ij}}{\sum_{j=1}^{N}\sum_{i=1}^{D} w_{ij}} \qquad (5)$$

ΙΙΙ) Χαρακτηρισμός ασυνεχειών σε πίνακα με στοιχεία από πολλές χώρες (ή πηγές)

Έστω D ένας επιλεγμένος αριθμός μεθοδολογιών υπό εξέταση. Για κάθε μία μεθοδολογία σκοπεύουμε να συγκρίνουμε τις σχετικές πρακτικές για κάθε ζεύγος από τις N εξεταζόμενες χώρες και να μετρήσουμε το επίπεδο εναρμόνισης (βλ. και [Papageorgiou et.al, 2001c])..

Δηλαδή θα εξετάζουμε το συνδυασμό $\binom{N}{2}$ ζευγών χωρών των συγκρίσεων για κάθε μεθοδολογία.

Ένας δείκτης που δείχνει πόσο διαφορετικές είναι οι μεθοδολογίες που χρησιμοποιούνται από δύο χώρες i, j ορίζεται ως εξής:

$$p_{ij} = \frac{\sum_{m=1}^{D} w_{ijm}}{D} \tag{6}$$

Εάν ορισμένα μεταδεδομένα πλήθους n λείπουν (δεν παρέχονται) από μια από τις δύο πηγές, προφανώς η σύγκριση δεν μπορεί να γίνει. Γι αυτές τις περιπτώσεις μεθοδολογιών τα αντίστοιχα βάρη θα είναι μηδενικά και αξιολογούμε τις περιπτώσεις D-n. Ένας γενικός συντελεστής που αποκαλύπτει πόσες ασυνέχειες μπορούν να παρατηρηθούν σε ολόκληρο τον πίνακα δίνεται από το μέσο όρο των ανωτέρω συντελεστών ως εξής:

$$p = \frac{\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} p_{ij}}{\binom{N}{2}} \tag{7}$$

2.4 Συνιστώσες ποιότητας στατιστικής πληροφορίας και διερεύνηση κριτηρίων αξιολόγησης της ποιότητας αποτελεσμάτων

Αρκετές Εθνικές Στατιστικές Υπηρεσίες παγκοσμίως έχουν δημιουργήσει και δημοσιεύσει πρακτικές και κριτήρια διασφάλισης της ποιότητας για ιδία χρήση. Παρατηρείται σύγκλιση απόψεων αλλά και διαφοροποιήσεις, κυρίως στον αριθμό των κριτηρίων που χρησιμοποιούνται αλλά και στις περιπτώσεις τις οποίες εξετάζουν. Ενδεικτικά αναφέρεται ότι η Στατιστική Υπηρεσία του Καναδά χρησιμοποιεί 6 κριτήρια ποιότητας [Statistics Canada (1998),] και η Στατιστική Υπηρεσία της Ολλανδίας δέκα [De Vries (1999).]. Διεθνείς Οργανισμοί όπως ο ΟΟΣΑ (OECD) και ο IMF έχουν τα δικά τους κριτήρια [Carson (2000)], [Eurostat, (2000d)]. Η Eurostat από την άλλη πλευρά έχει τη δική της προσέγγιση, υιοθετώντας επτά κριτήρια αξιολόγησης της ποιότητας [Eurostat, (2000b), (2000c), (2000d)] (δες επίσης [Depoutot & Arrondel, 1998], [Depoutot et. al, 1998]).

Δεδομένου ότι μια στατιστική υπηρεσία ασχολείται με θέματα που έχουν σχέση με τις σύγχρονες ανάγκες της κοινωνίας (σχετικότητα της πληροφορίας – relevance) και χρησιμοποιεί τις κατάλληλες έννοιες, τίθεται το ερώτημα αν τα αποτελέσματα των ερευνών αυτών διαθέτουν επαρκή ακρίβεια (accuracy) αλλά και αν η πληροφορία

διατίθεται έγκαιρα. Οι ακριβείς μετρήσεις έχουν συχνά απαγορευτικό κόστος και ορισμένες φορές είναι πολύ δύσκολο να πραγματοποιηθούν οπότε υπάρχει ενδιαφέρον για το αν έχει επιτευχθεί ένα αποδεκτό περιθώριο σφάλματος. Από την άλλη πλευρά, ακριβείς πληροφορίες σε ενδιαφέροντα θέματα δε θα είναι χρήσιμες στους χρήστες αν δε δημοσιευτούν έγκαιρα ώστε οι φορείς δημόσιας διοίκησης να πάρουν κάποιες αποφάσεις βασισμένοι σε αυτά. Η επικαιρότητα (timeliness) λοιπόν της στατιστικής πληροφορίας μπορεί να είναι μεγάλης σημασίας για βασικά μηνιαία οικονομικά μεγέθη αλλά να έχει μικρότερη σημασία για αργά μεταβαλλόμενα φαινόμενα.

Παράλληλα, όπως ήδη έχει ειπωθεί σε προηγούμενες ενότητες, οι χρήστες ορισμένες φορές είναι δυνατό να έχουν στη διάθεσή τους διαφορετικά σύνολα στατιστικών πληροφοριών, τα οποία έχουν παραχθεί *από διαφορετικές πηγές* και *σε διαφορετικές χρονικές στιγμές*. Η χρήση των αποτελεσμάτων μιας έρευνας διευκολύνεται αν μπορούν να συνδυαστούν με τα αποτελέσματα που έχουν προκύψει από άλλα σύνολα δεδομένων. Κάτι τέτοιο επιτυγχάνεται μέσω της χρήσης κοινών ή έστω συγκρίσιμων εννοιών και μεθοδολογιών. Ο βαθμός λοιπόν στον οποίο η στατιστική πληροφορία μπορεί να χρησιμοποιηθεί σε ευρύτερα πλαίσια και να κάνει χρήση τυποποιημένων εννοιών, μεταβλητών, ταξινομήσεων και μεθόδων αναφέρεται ως συμβιβαστότητα/συμβατότητα (coherence) της πληροφορίας ή της έρευνας. Τέλος, όπως έχει αναφερθεί ήδη στους μετασχηματισμούς, ενδιαφέρον παρουσιάζει η συγκρισιμότητα (comparability) των διαφόρων ερευνών, η οποία αναφέρεται στο κατά πόσο είναι δυνατό να συγκριθούν τα αποτελέσματα ερευνών που έχουν διεξαχθεί σε διαφορετικές χρονικές περιόδους ή από διαφορετικούς δημόσιους φορείς [Elvers, 1998].

Προκειμένου επίσης να γίνει σωστή χρήση των αποτελεσμάτων μιας έρευνας από τους χρήστες θα πρέπει αυτοί να γνωρίζουν το αντικείμενο της έρευνας και τις ιδιότητες των παρεχομένων πληροφοριών. Αυτό σημαίνει ότι ο δημόσιος φορέας θα πρέπει να παρέχει περιγραφές των εννοιών, των μεταβλητών και των ταξινομήσεων που έχουν χρησιμοποιηθεί κατά τη διενέργεια της έρευνας όπως επίσης και των μεθόδων συλλογής και επεξεργασίας των δεδομένων και των μεθόδων εκτίμησης, δηλαδή να παρέχει τα κατάλληλα ***μεταδεδομένα*** όπως έχουμε τονίσει και σε άλλα σημεία αυτής της διατριβής. Κρίνοντάς την από την οπτική γωνία της αξιολόγησης και διασφάλισης της ποιότητας των αποτελεσμάτων, η ιδιότητα αυτή της στατιστικής πληροφορίας αναφέρεται ως δυνατότητα ερμηνείας (interpretability) ή ευκρίνεια (clarity).

Πρέπει επίσης να σημειωθεί ότι οι περισσότερες από τις προαναφερθείσες ιδιότητες της στατιστικής πληροφορίας, δεν είναι προφανείς στους χρήστες χωρίς την παροχή βοηθητικών πληροφοριών από τη στατιστική υπηρεσία. Η ακρίβεια δεν είναι δυνατό να διαπιστωθεί κοιτάζοντας απλώς τους αριθμούς και η στατιστική υπηρεσία, η οποία διαθέτει πρόσβαση στα μικροδεδομένα και έχει γνώση και της μεθοδολογίας που χρησιμοποιήθηκε, θα πρέπει να δώσει κάποια μέτρα για την ακρίβεια. Η σχετικότητα επίσης της πληροφορίας φαίνεται να μη μπορεί να διαπιστωθεί χωρίς την παροχή πληροφοριών για τις έννοιες, τις ταξινομήσεις και τις μεθόδους που χρησιμοποιήθηκαν. Μόνο η επικαιρότητα και η δυνατότητα πρόσβασης είναι άμεσα παρατηρήσιμες από

τους χρήστες των αποτελεσμάτων. Επιπλέον η σχετικότητα, η δυνατότητα πρόσβασης και η συμβατότητα πρέπει συνήθως να θεωρούνται για σύνολα ερευνών και όχι για κάθε έρευνα χωριστά. Η σχετικότητα μιας πληροφορίας εξαρτάται από το τι άλλο είναι διαθέσιμο και χρειάζεται επομένως πρόσβαση σε ένα ολόκληρο πρόγραμμα ερευνών. Εξ' ορισμού το ίδιο ισχύει και για τη συμβατότητα ενώ αναφορικά με τη δυνατότητα πρόσβασης, τα περισσότερα στατιστικά προϊόντα δημοσιεύονται μέσω ενός κοινού συστήματος δημοσίευσης για ολόκληρη τη στατιστική υπηρεσία. Από την άλλη μεριά η ακρίβεια, η επικαιρότητα και η δυνατότητα ερμηνείας μπορούν να θεωρηθούν ως ιδιότητες κάθε στατιστικού αποτελέσματος χωριστά, ακόμα και αν κάθε έρευνα κάνει χρήση εργαλείων ή προσεγγίσεων που είναι κοινά μεταξύ διαφόρων προγραμμάτων.

Στη συνέχεια θα επικεντρωθούμε στην ευκρίνεια (clarity), η οποία σχετίζεται έμμεσα με τα προσφερόμενα μεταδεδομένα, καθώς και τη συγκρισιμότητα (comparability) και συμβατότητα (coherence) της πληροφορίας, οι οποίες επηρεάζονται άμεσα από τους προαναφερθέντες μετασχηματισμούς. Θα θεωρήσουμε τη δυνατότητα χρήσης των δεικτών τόσο σε εθνικό επίπεδο (χώρας) όσο και Ευρωπαϊκό επίπεδο (Eurostat).

2.4.1 Ευκρίνεια

Γενικά η ευκρίνεια είναι δύσκολο να εκτιμηθεί και συνδέεται κυρίως με την ποιότητα των μεταδεδομένων. Από αρκετές υπηρεσίες έχει προταθεί ως δείκτης για αυτή τη συνιστώσα της ποιότητας ο αριθμός ή το ποσοστό των δημοσιευμένων αποτελεσμάτων τα οποία συνοδεύονται από πλήρη μεταδεδομένα, σύμφωνα με ένα τυπικό σχέδιο παροχής μεταδεδομένων. Ένας προτεινόμενος δείκτης επίσης για την ευκρίνεια, ο οποίος όμως χρειάζεται περαιτέρω μελέτη, είναι ο λόγος των διαθέσιμων μεταδεδομένων προς τα απαιτούμενα μεταδεδομένα. Παρέχει αυτός ο δείκτης πληροφορίες για το βαθμό στον οποίο τα μεταδεδομένα είναι διαθέσιμα στους χρήστες και αναφέρεται στην πληρότητα των μεταδεδομένων για ένα συγκεκριμένο θέμα. Για τον υπολογισμό του δείκτη αθροίζεται η διαθεσιμότητα των μεταδεδομένων για τις διάφορες κατηγορίες (π.χ. 2 = πλήρως, 1 = μερικώς, 0 = μη διαθέσιμα) που περιλαμβάνονται σε έναν τυπικό (standard) πίνακα των απαιτούμενων μεταδεδομένων και στη συνέχεια γίνεται διαίρεση με το άθροισμα των απαιτούμενων μεταδεδομένων. Δηλαδή:

$$\text{Λόγος πληρότητας μεταδεδομένων} = \frac{\sum_{j=1}^{M} \text{διαθέσιμα μεταδεδομένα}}{\sum_{j=1}^{M} \text{σχετικά (απαιτούμενα) μεταδεδομένα}}$$

Όπου (j), j=1,2,...,M είναι οι διάφορες κατηγορίες μεταδεδομένων (βλ. το μοντέλο μεταδεδομένων που αναπτύσσεται στο κεφάλαιο 4 για τις διάφορες κατηγορίες μεταδεδομένων.

2.4.2 Συγκρισιμότητα (comparability)

Αυτή η συνιστώσα της ποιότητας έχει ως στόχο τη μέτρηση της επίδρασης των διαφορών στις στατιστικές έννοιες, τους ορισμούς και τις διαδικασίες μέτρησης όταν συγκρίνονται στοιχεία μεταξύ διαφορετικών γεωγραφικών περιοχών, τομέων μελέτης, ή χρονικών περιόδων. Η συγκρισιμότητα έχει έννοια όταν αναφέρεται σε έρευνες οι οποίες εκτιμούν ένα παρόμοιο χαρακτηριστικό και εκφράζει το βαθμό στον οποίο οι διαφορές που υπάρχουν οδηγούν σε διαφορές μεταξύ των πραγματικών τιμών ενός στατιστικού χαρακτηριστικού. Περιπτώσεις για τις οποίες παρατηρείται συχνά ενδιαφέρον για συγκρισιμότητα έχουν να κάνουν με γεωγραφικές περιοχές π.χ. συγκρίσεις μεταξύ χωρών, μεταξύ ενός κράτους-μέλους και της Ε.Ε. ή μεταξύ της Ε.Ε. και των Η.Π.Α. κτλ, με διάφορους τομείς π.χ. συγκρίσεις εισοδήματος για διαφορετικούς τύπους νοικοκυριών, ή με διαφορετικές χρονικές περιόδους π.χ. σύγκριση μεταξύ μιας περιόδου αναφοράς με προηγούμενες περιόδους αναφοράς για τους ίδιους τομείς μελέτης.

Υπάρχουν δύο κυρίως πηγές οι οποίες έχουν ως αποτέλεσμα την έλλειψη συγκρισιμότητας και οι οποίες πηγές είναι οι *εννοιολογικές διαφορές* και *οι διαφορές στη διαδικασία μέτρησης*. Στις εννοιολογικές διαφορές περιλαμβάνονται διαφορές στους ορισμούς των υπό μελέτη πληθυσμών ή διαφορετικές ταξινομήσεις που είναι πιθανό να εφαρμόζονται στις διάφορες χώρες και οι οποίες αρκετές φορές οφείλονται σε διαφορετικές παραδόσεις, νομοθεσίες ή πραγματικότητες ενώ οι διαφορές στη διαδικασία μέτρησης σχετίζονται με την ακρίβεια των εκτιμήσεων.

Αν περιοριστούμε ενδεικτικά στις βασικότερες πηγές μη συγκρισιμότητας, η ακόλουθη λίστα δημιουργήθηκε για να μας βοηθήσει στην εξαγωγή δεικτών συγκρισιμότητας.

1. Έννοιες

1.1 Στατιστικά χαρακτηριστικά (μια αριθμητική τιμή π.χ. συνολικός τζίρος)
1.2 Στατιστικά μέτρα (π.χ. μέσοι, ολικά, δείκτες)
1.3 Στατιστική μονάδα (π.χ. δειγματοληπτική μονάδα)
1.4 Πληθυσμός που ενδιαφέρει την έρευνα (ο «ιδανικός» πληθυσμός)
1.5 Πληθυσμός-στόχος (ο δειγματοληπτούμενος πληθυσμός)
1.6 Χρόνος αναφοράς (σημείο αναφοράς ή περίοδος αναφοράς)
1.7 Τομείς μελέτης (διαφορετικές υποκατηγορίες του πληθυσμού)
1.8 Ταξινομήσεις (π.χ. NACE Αναθ.1)
1.9 Άλλες εννοιολογικές απόψεις

2. Μέτρηση

2.1 Δειγματοληπτικές διαδικασίες
2.2 Συλλογή δεδομένων (μέθοδοι και εργαλεία)
2.3 Επεξεργασία δεδομένων (editing και imputation)
2.4 Προσαρμογές (adjustments) και εκτίμηση (σταθμίσεις κτλ)

2.5 Άλλα θέματα σχετικά με τις μετρήσεις για συγκεκριμένους τομείς μελέτης

Οι δείκτες που δημιουργήθηκαν έχουν να κάνουν κυρίως με τα αίτια που δυσκολεύουν τη δυνατότητα σύγκρισης και βασίζονται σε μεταδεδομένα. Στα πλαίσια της Ε.Ε. λοιπόν, με βάση την προηγούμενη λίστα, είναι δυνατό να αναπτυχθεί ένας δείκτης σχετικά με τη συγκρισιμότητα στοιχείων που αφορούν διαφορετικές γεωγραφικές περιοχές. Μέσω αυτού του δείκτη, κάθε χώρα-μέλος παρέχει στη Eurostat ένα σύνθετο μέτρο της γεωγραφικής συγκρισιμότητας για τις βασικές μεταβλητές των διαφόρων ερευνών που διεξάγει η στατιστική υπηρεσία της. Βέβαια η Eurostat θα πρέπει να ορίσει ποια είναι τα στατιστικά στοιχεία για τον υπολογισμό των οποίων οι υπηρεσίες των χωρών-μελών είναι δυνατό να ακολουθήσουν ένα ευρωπαϊκό πρότυπο. Επιπλέον οι στατιστικές υπηρεσίες χρειάζεται να γνωρίζουν ποιες είναι οι έννοιες και οι διαδικασίες μέτρησης στις οποίες διαφέρουν από τα ευρωπαϊκά πρότυπα. Προκειμένου λοιπόν να εκτιμηθεί η γεωγραφική συγκρισιμότητα, θα πρέπει τα κράτη-μέλη να εξετάζουν τα ίδια στατιστικά στοιχεία, τα ίδια στατιστικά χαρακτηριστικά (αριθμός και τύπος) και επιπλέον θα πρέπει να θεωρούνται οι ίδιες πηγές για την έλλειψη γεωγραφικής συγκρισιμότητας π.χ. οι πηγές της προηγούμενης λίστας.

Έστω λοιπόν ότι με (i) συμβολίζεται η χώρα που παρέχει στοιχεία για μια έρευνα, $i = 1,2,\ldots,N$ και με (j) συμβολίζεται η πηγή της έλλειψης της συγκρισιμότητας, $j = 1,2,\ldots,M$.

Ορίζονται επίσης οι μεταβλητές D_{ij} ως

$$D_{ij} = \begin{cases} 1, \text{ αν αναφέρεται διαφορά από το πρότυπο για την i – χώρα σχετικά με την j-πηγή} \\ 0, \text{ αν δεν αναφέρεται διαφορά από το πρότυπο για την i – χώρα σχετικά με την j-πηγή} \end{cases}$$

Επομένως για μια συγκεκριμένη χρονική περίοδο είναι δυνατό να παραχθεί ένας πίνακας σαν τον ακόλουθο για τις διαφορές από τα ευρωπαϊκά πρότυπα σύμφωνα με κάθε χώρα και κάθε πηγή και κάτι τέτοιο μπορεί να γίνει για τις βασικές έρευνες.

Πηγή (j)							
		1	2	…	j	…	M
	1	D_{11}	D_{12}		D_{1j}		D_{1M}
	2	D_{21}	D_{22}		D_{2j}		D_{2M}
Χώρα (i)	…						
	i	D_{i1}	D_{i2}		D_{ij}		D_{iM}
	…						
	N	D_{N1}	D_{N2}		D_{Nj}		D_{NM}

Ο αριθμός και το ποσοστό των διαφορών για μια βασική μεταβλητή μπορεί να υπολογιστεί από τους τύπους:

$$\sum_{i=1}^{N}\sum_{j=1}^{M} D_{ij} \quad \text{και} \quad \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} D_{ij}}{NM} \text{ , αντίστοιχα.}$$

Προφανώς αν $D_{i\,.} = \sum_{j=1}^{M} D_{ij} \neq 0$, τότε για τη χώρα (i) υπάρχει τουλάχιστον μια πηγή η οποία διαφέρει από το ευρωπαϊκό πρότυπο και μεγάλες τιμές του Di . υποδηλώνουν πιθανότητα ότι δεν είναι δυνατή η γεωγραφική σύγκριση.

Παρόμοια αν $D_{.\,j} = \sum_{i=1}^{N} D_{ij} \neq 0$ τότε για την πηγή (j) υπάρχει τουλάχιστον μία χώρα η οποία παρουσιάζει διαφορές από τα ευρωπαϊκά πρότυπα και σχετικά μεγάλες τιμές του D.j δηλώνουν πιθανή έλλειψη δυνατότητας σύγκρισης μεταξύ τομέων.

Οι διάφορες λεπτομέρειες, όπως για παράδειγμα ποιες ακριβώς είναι οι διαφορές που υπάρχουν, καλό είναι να παρέχονται από τις διάφορες χώρες ενώ ο υπολογισμός του δείκτη καλό είναι να γίνεται από την Eurostat. Επίσης η Eurostat είναι δυνατό να υπολογίσει το μέσο όρο του δείκτη από τις διάφορες χώρες και στη συνέχεια να υπολογίσει την απόκλιση κάθε χώρας από το μέσο όρο. Μαζί με τα μεταδεδομένα που αναφέρθηκαν προηγουμένως (ευρωπαϊκά πρότυπα), είναι χρήσιμο για την ερμηνεία του δείκτη να παρέχονται και περιγραφές των διαφορών καθώς και οι λόγοι για τους οποίους υπάρχουν αυτές οι διαφορές.

2.4.3 Συμβιβαστότητα, συνοχή (Coherence)

Η συμβιβαστότητα της στατιστικής πληροφορίας αναφέρεται στο βαθμό στον οποίο αυτή μπορεί να συνδυαστεί με άλλες στατιστικές πληροφορίες μέσα σε ένα γενικότερο πλαίσιο εργασίας . Εκφράζει δηλαδή την καταλληλότητα των στατιστικών αποτελεσμάτων να συνδυαστούν αξιόπιστα με διαφορετικούς τρόπους και για ποικίλες χρήσεις . Κάτι τέτοιο φαίνεται πως επιτυγχάνεται με τη χρήση τυπικών ταξινομήσεων ,

ορισμών , πληθυσμών – στόχων και κοινής μεθοδολογίας μεταξύ διαφόρων ερευνών που έχουν παρόμοια θέματα . Η συνοχή μάλιστα των στατιστικών στοιχείων εστιάζεται κυρίως στην κοινή χρήση στοιχείων που παράγονται για διαφορετικούς πρωταρχικούς σκοπούς .

Όταν υπάρχουν παρόμοια στατιστικά στοιχεία από διαφορετικές πηγές, τότε πρέπει αυτά να αναγνωρίζονται και πιθανές διαφορές, αν είναι δυνατό, να ποσοτικοποιούνται. Διαφορές μεταξύ δύο συνόλων στοιχείων που έχουν προκύψει από διαφορετικές έρευνες ίσως οφείλονται σε διαφορές στη διαδικασία συλλογής δεδομένων ή στις δειγματοληπτικές μονάδες με αποτέλεσμα να προκύπτουν διαφορετικές εκτιμήσεις . Υπάρχουν βέβαια πολλά θέματα για τα οποία είναι δυνατό να εξεταστεί η συνοχή αλλά στη συνέχεια θα εξεταστεί η συνοχή μεταξύ της πρωτογενούς και της κύριας δευτερογενούς χρήσης δεδομένων [Elvers, 1998]. Πολύ συχνά μάλιστα τα δεδομένα μιας έρευνας ενός στατιστικού οργανισμού χρησιμοποιούνται σύμφωνα με την κύρια δευτερογενή τους χρήση και σε άλλες έρευνες. Αφού λοιπόν καθοριστεί η πιο σημαντική δευτερογενής χρήση για ένα συγκεκριμένο σύνολο στοιχείων, είναι δυνατό να οριστούν οι επιθυμητές ιδιότητες για τη δευτερογενή χρήση υπό μορφή παραγόντων όπως στατιστικές μονάδες, πληθυσμός, μεταβλητές, περίοδοι αναφοράς, ταξινόμηση και στη συνέχεια να αποφασιστεί αν το σύνολο αυτό ικανοποιεί ή όχι τις απαιτήσεις για κάθε σχετικό παράγοντα .

Τα σύνολα των στοιχείων λοιπόν μπορούν να ταξινομηθούν σε πίνακες σύμφωνα με την κύρια δευτερογενή χρήση τους u , u=1,2,...,S όπως περιγράφεται στον παρακάτω πίνακα . Στη στήλη του παράγοντα (k) μπαίνει η τιμή 1 αν το προϊόν ικανοποιεί τις απαιτήσεις για δευτερογενή χρήση σχετικά με αυτόν τον παράγοντα ενώ διαφορετικά μπαίνει η τιμή 0 . Είναι για παράδειγμα:

Δευτερογενής Χρήση No	Σύνολο Στοιχείων No	Παράγων k_{11}	Παράγων k_{12}	Παράγων k_{13}	Παράγων ...	Παράγων ...	Παράγων K_{1n1}
1	1	0	1	1			1
1	2	1	0	0			1
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
1	N_1	0	0	1			0
Δευτερογενής Χρήση No		Παράγοντας k_{21}	Παράγοντας k_{22}	Παράγοντας k_{23}	Παρά-γοντας .	Παρά-γοντας .	Παράγοντας k_{2n2}
2	1	1	0	0			1
2	2	0	1	1			1
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
2	N_2	0	0	1			0
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
Δευτερογενής Χρήση No		Παράγοντας k_{s1}	Παράγοντας k_{s2}	Παράγοντας k_{s3}	Παρά-γοντας .	Παρά-γοντας .	Παράγοντας k_{sns}
S	1	0	1	1			1
S	2	0	0	1			0
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
S	N_s	1	0	0			1

Επομένως το ποσοστό των συνόλων των στοιχείων με την ίδια δευτερογενή χρήση u που ικανοποιούν τις απαιτήσεις σχετικά με τον παράγοντα k_{u1} είναι:

$$P_{u,ku1} = \frac{\sum_{j=1}^{Nu} I_{ku1,\,j}}{N_u}$$

όπου N_u είναι ο αριθμός των συνόλων των στοιχείων με την ίδια κύρια δευτερογενή χρήση u, k_{u1} είναι ο πρώτος παράγοντας για τη δευτερογενή χρήση u και

$$I_{ku1,j} = \begin{cases} 1, \text{ αν το j σύνολο στοιχείων ικανοποιεί τις απαιτήσεις του ku1} \\ 0, \text{ διαφορετικά} \end{cases}$$

Επίσης αν

$$I_{.,j} = \begin{cases} 1, \text{ αν το j σύνολο στοιχείων ικανοποιεί τις απαιτήσεις όλων των nu παραγόντων για τη δευτερογενή χρήση u} \\ 0, \text{ διαφορετικά} \end{cases}$$

τότε είναι δυνατό να υπολογιστεί το ποσοστό του συνόλου των στοιχείων με την ίδια δευτερογενή χρήση u που ικανοποιεί τις απαιτήσεις για όλους τους n_u σχετικούς παράγοντες, ως:

$$P_{u,.} = \frac{\sum_{j=1}^{Nu} I_{.,\,j}}{N_u}$$

Ένας γενικός δείκτης μπορεί να υπολογιστεί ως:

$$P_{.,.} = \frac{\sum_{u=1}^{S}\sum_{j=1}^{Nu} I_{.,\,j}}{\sum_{u=1}^{S} N_u} = \frac{\sum_{u=1}^{S} N_u P_{u,.}}{\sum_{u=1}^{S} N_u}$$

που είναι το ποσοστό των συνόλων των στοιχείων που ικανοποιούν τις απαιτήσεις για την κύρια δευτερογενή χρήση, όποια και αν είναι αυτή.

Οι προηγούμενοι δείκτες είναι δυνατό να υπολογιστούν τόσο σε επίπεδο Eurostat όσο και σε εθνικό επίπεδο και παρέχουν πληροφορίες για τις ενέργειες που πρέπει να γίνουν ώστε να αυξηθεί η συμβιβαστότητα. Χρειάζεται όμως καθορισμός των απαιτήσεων για τις δευτερογενείς χρήσεις. Η διαφορά, τέλος, που υπάρχει με τη συγκρισιμότητα είναι ότι οι δείκτες για τη συνοχή θεωρούν τις δευτερογενείς χρήσεις των δεδομένων ενώ στην περίπτωση της συγκρισιμότητας υπάρχει ενδιαφέρον για την κύρια χρήση των στατιστικών αποτελεσμάτων.

ΚΕΦΑΛΑΙΟ 3

ΚΑΤΑΣΚΕΥΗ ΣΤΑΤΙΣΤΙΚΩΝ ΜΟΝΤΕΛΩΝ ΜΕΤΑΔΕΔΟΜΕΝΩΝ - ΔΙΑΔΙΚΑΣΙΕΣ ΚΑΙ ΒΗΜΑΤΑ ΓΙΑ ΤΗ ΔΗΜΙΟΥΡΓΙΑ ΤΟΥΣ

3.1 Εισαγωγή

Ένα σημαντικό χαρακτηριστικό γνώρισμα των σύγχρονων συστημάτων επεξεργασίας στατιστικής πληροφορίας είναι η εκτεταμένη χρήση μεταδεδομένων όπως επισημαίνεται, μεταξύ των άλλων στις εργασίες των [Grossmann, 1999], [Papageorgiou et al, 2000a], [Sundgren, 1996]. Δυστυχώς, τα περισσότερα από τα στατιστικά συστήματα επεξεργασίας πληροφοριών μεταχειρίζονται τα μεταδεδομένα απλά ως κείμενο (free-text), χωρίς να αντιλαμβάνονται τη συνεισφορά των μεταδεδομένων στον εμπλουτισμό της πληροφορίας.

Αυτή η παθητική χρήση των μεταδεδομένων αγνοεί την ύπαρξη των μεταδεδομένων που χρησιμοποιούνται για μετασχηματισμούς [Kent, & Schuerhoff, 1997], [Papageorgiou et.al, 2000b], μειώνοντας κατά συνέπεια τα πλεονεκτήματα της μεταπληροφορίας. Το πρόβλημα αυτό μπορεί να αντιμετωπιστεί με τη δημιουργία ενός μοντέλου μεταδεδομένων λαμβάνοντας υπόψη τις ιδιαιτερότητες και τις ανάγκες των φορέων δημόσιας διοίκησης και των στατιστικών υπηρεσιών.

Στο κεφάλαιο αυτό εξετάζονται βήμα προς βήμα οι απαραίτητες διαδικασίες και επιλογές προκειμένου να δημιουργήσουμε ένα μοντέλο μεταδεδομένων χρήσιμο στους φορείς δημόσιας διοίκησης, το οποίο θα συμβάλει στην αυτοματοποίηση των διαδικασιών εξαγωγής ποιοτικών στατιστικών αποτελεσμάτων. Οι διαδικασίες αυτές λαμβάνονται υπόψη στο κεφάλαιο 4 στο οποίο περιγράφεται το προτεινόμενο μοντέλο μεταδεδομένων

3.2 Διαδικασίες και βήματα για τη δημιουργία ενός μοντέλου μεταδεδομένων

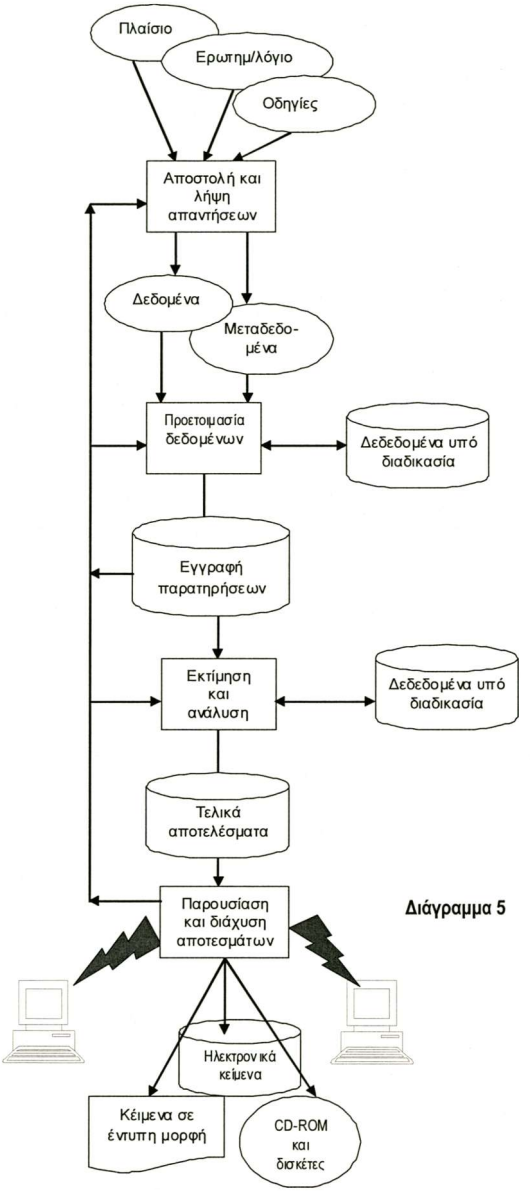
Κατά τη απόφαση ενσωμάτωσης μοντέλου μεταδεδομένων στα πληροφοριακά συστήματα των φορέων δημόσιας διοίκησης, οι παρακάτω προϋποθέσεις/διαδικασίες πρέπει να εφαρμοστούν:

- Εξέταση της ροής δεδομένων και μεταδεδομένων
- Συστηματική μελέτη του σχήματος δεδομένων των βάσεων ορισμένων ενδεικτικών φορέων
- Συγκεκριμενοποίηση των βασικών ιδιοτήτων/κριτηρίων που πρέπει να πληρεί το μοντέλο μεταδεδομένων
- Επιλογή κατηγοριών μεταδεδομένων (ποιά μεταδεδομένα θα μοντελοποιηθούν)
- Επιλογή τεχνικής μοντελοποίησης
- Επιλογή γλώσσας μοντελοποίησης

Τις διαδικασίες αυτές εξετάζουμε στη συνέχεια:

3.2.1 Τυπική διαδικασία ροής δεδομένων και μεταδεδομένων σε έναν στατιστικό οργανισμό

Στο Διάγραμμα 5 παρουσιάζεται μία απλοποιημένη διαδικασία ροής στατιστικής πληροφορίας και μεταπληροφορίας σε έναν οργανισμό, όπως για παράδειγμα σε μια Στατιστική Υπηρεσία.



Διάγραμμα 5

Ένα χαρακτηριστικό πλαίσιο είναι ένας κατάλογος επιχειρήσεων προς παρατήρηση και των επαφών με τα αντίστοιχα άτομα.

Τα παραδείγματα των μεταδεδομένων που συνοδεύουν τα (μικρο)δεδομένα είναι: ερωτήσεις που υποβάλλονται, διευκρινήσεις, ορισμοί, είδος δειγματοληψίας, κλπ.

Έπειτα έχουμε κωδικοποίηση και ανάλυση των δεδομένων και των μεταδεδομένων.

Ο κατάλογος και τα δεδομένα από άλλες πηγές, συμπεριλαμβανομένων των δεδομένων από τις προηγούμενες επαναλήψεις της ίδιας έρευνας, μπορούν να χρησιμοποιηθούν. Ο τελικός κατάλογος παρατήρησης μιας έρευνας περιέχει μικροδεδομένα και συνοδευτικά μεταδεδομένα, που οργανώνονται με έναν τυποποιημένο τρόπο προκειμένου να διευκολυνθούν οι επόμενες διαδικασίες.

Ακολουθεί υπολογισμός των κατ' εκτίμηση τιμών των στατιστικών χαρακτηριστικών για τον πληθυσμό και τα υποσύνολα πληθυσμού, εκτέλεση των στατιστικών αναλύσεων και τελικές πολυδιάστατες στατιστικές και συνοδευτικά μεταδεδομένα, που οργανώνονται με έναν τυποποιημένο τρόπο. Τα τελικά στατιστικά προϊόντα αποτελούνται από τα δεδομένα και τα μεταδεδομένα στην τελική βάση δεδομένων στατιστικών.

Τα προϊόντα μπορούν να αποθηκευτούν ως ηλεκτρονικά έγγραφα, και διαδίδονται μέσω των διαφορετικών μέσων.

3.2.2 Συστηματική μελέτη του σχήματος δεδομένων (data schema) των βάσεων ορισμένων ενδεικτικών φορέων

Πριν από τη δημιουργία του μοντέλου μεταδεδομένων πρέπει να μελετηθεί το σχήμα δεδομένων (data schema) της βάσης που χρησιμοποιούν οι φορείς δημόσιας διοίκησης.

Σε λογικό επίπεδο ένα σχήμα δεδομένων υπό μορφή συσχετίσεων-οντοτήτων θα αναπτυσσόταν για κάθε συλλογή δεδομένων. Αυτό θα χρησιμεύσει ως μια βάση για την περιγραφή των απαραίτητων σχέσεων μεταξύ των δεδομένων και θα διαμορφώσει τα μεταδεδομένα της βάσης. Αυτά τα μεταδεδομένα θα μεταφραστούν έπειτα στο πραγματικό σχήμα βάσεων δεδομένων ενός συστήματος.

Για το λόγο αυτό μελετήθηκε το σχήμα δεδομένων ορισμένων Ελληνικών φορέων δημόσιας διοίκησης. Πιο συγκεκριμένα, μελετήθηκαν τα λογικά σχήματα των εξής Ελληνικών οργανισμών:

- Γενική Γραμματεία Πληροφοριακών Συστημάτων (ΓΓΠΣ), Υπουργείου Οικονομίας και Οικονομικών, για τις δραστηριότητές τους σχετικά με Διασυνοριακές εμπορικές συναλλαγές
- Οργανισμός Επαγγελματικής Εκπαίδευσης και Κατάρτισης (ΟΕΕΚ), Υπουργείο Εθνικής Παιδείας και Θρησκευμάτων, σχετικά με την εκπαίδευση στα ΙΕΚ

- Εθνική Στατιστική Υπηρεσία Ελλάδος (ΕΣΥΕ) για τη συσχέτιση επαγγέλματος και οικονομικής δραστηριότητας

Τα επόμενα διαγράμματα παρατίθενται στα αγγλικά ώστε να μπορούν να χρησιμοποιηθούν στο σχεδιασμό μιας βάσης δεδομένων και στη συσχέτισή τους με αυτά άλλων χωρών και διεθνών οργανισμών.

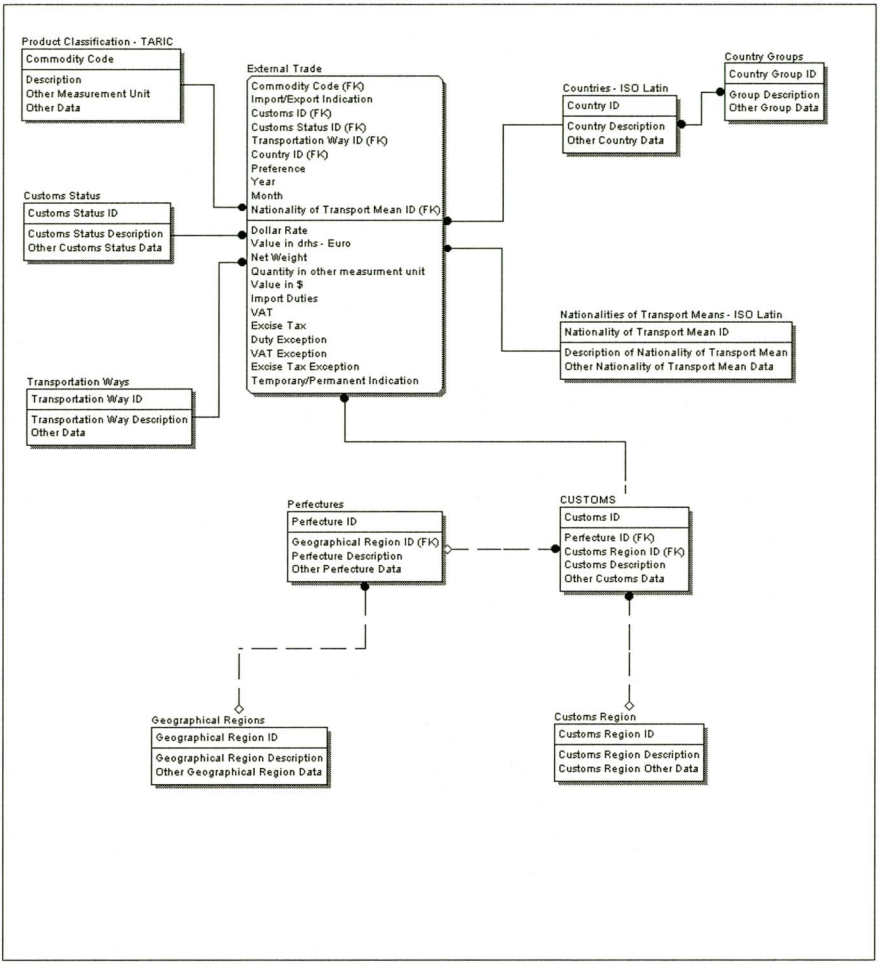
1. Διασυνοριακές εμπορικές συναλλαγές από ΓΓΠΣ

Οι βασικές **μεταβλητές** όπως απεικονίζονται στο διάγραμμα 6 είναι οι κάτωθι:

- Προϊόν (commodity) κατηγοριοποιημένο με κωδικούς **ταξινόμησης** TARIC και χρησιμοποιείται μία **μονάδα μέτρησης** (measurement unit)
- Τελωνειακή θέση (customs status)
- Τρόπος μεταφορών (transportation way)
- Προέλευση (υπηκοότητα) των μέσων μεταφοράς (nationality of transport mean)
- Χώρα (country)
- Τελωνειακό διαμέρισμα (customs prefecture)
- Περιοχή (region), η οποία όμως πρέπει να σχετίζεται και με τη γεωγραφικό διαμέρισμα αλλά και με το τελωνειακό διαμέρισμα.

Επιπρόσθετα, άλλες εξαρτημένες μεταβλητές είναι απαραίτητες αλλά και **ισοτιμίες** δηλαδή τίθεται η ανάγκη για **μετατροπές μιάς μονάδας μέτρησης σε άλλη.**

Σε κάθε περίπτωση υπάρχουν **επιπρόσθετες πληροφορίες** ως σημασιολογικά και μεταδεδομένα τεκμηρίωσης, αναγκαία για την κατανόηση των βασικών μεταβλητών.



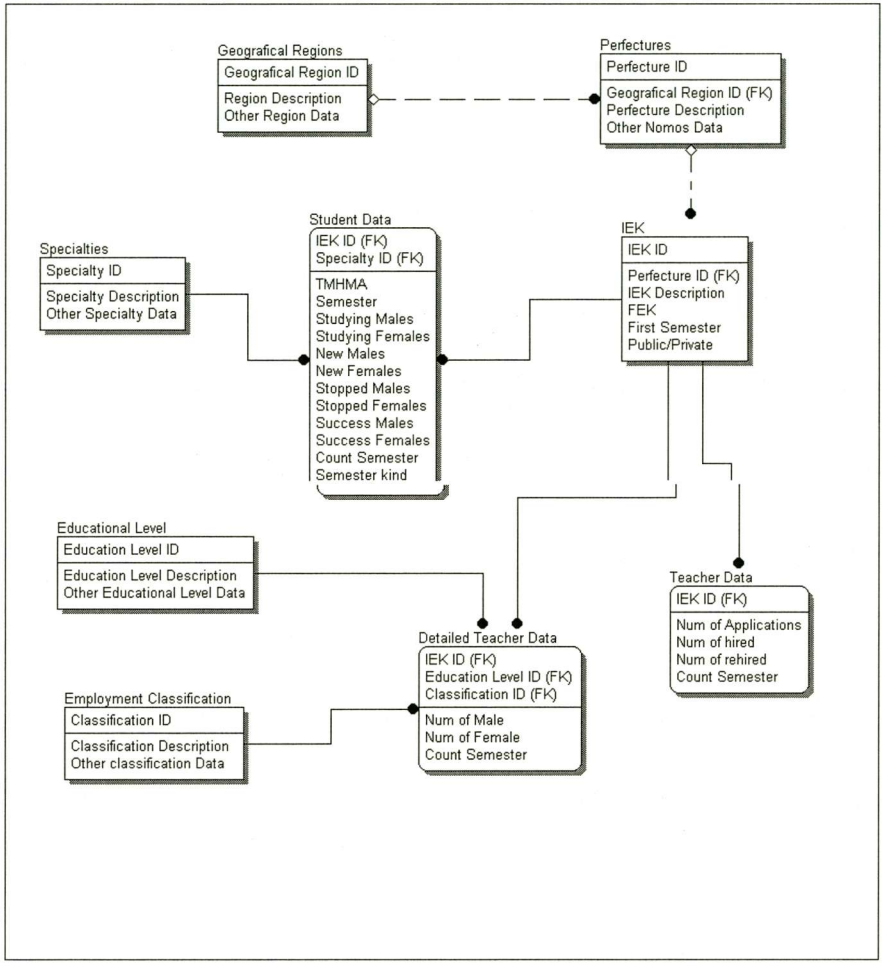
Διάγραμμα 6: Λογικό σχήμα για διασυνοριακές εμπορικές συναλλαγές από ΓΓΠΣ

2. Λογικό διάγραμμα για επαγγελματική εκπαίδευση από ΟΕΕΚ

Οι μεταβλητές που κρίνονται βασικές κατά το διάγραμμα 7 είναι

- Γεωγραφική περιοχή (geographical region) και μετράται σε διαμέρισμα, νομό, κλπ, δηλαδή με διαφορετική μονάδα μέτρησης όπου χρειάζεται.
- Όνομα ΙΕΚ
- Περίοδος φοίτησης
- Ιδιότητα (κρατικό ή ιδιωτικό)
- Επίπεδο κατάρτισης
- Ειδικότητα που μετράται με κάποια **ταξινόμηση**
- Επάγγελμα που μετράται με κάποια **ταξινόμηση**

- Φύλο
- Τμήμα
- άλλες επιμέρους **μεταβλητές**.



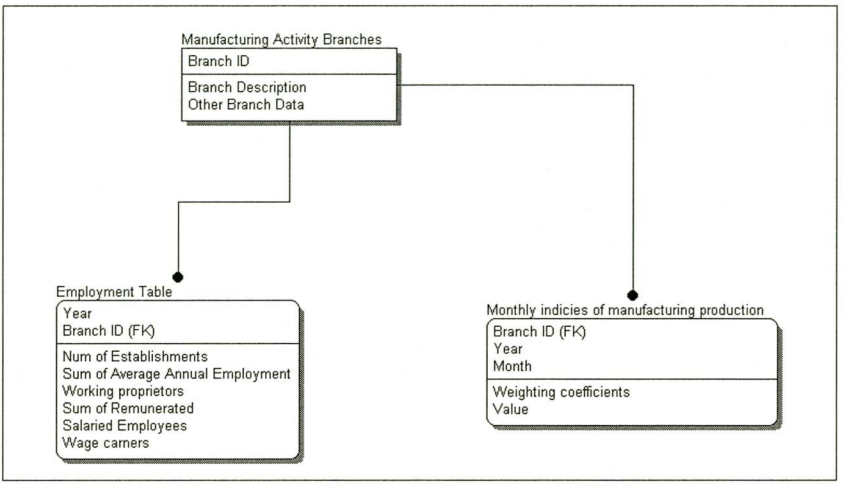
Διάγραμμα 7: Λογικό διάγραμμα για Επαγγελματική εκπαίδευση από ΟΕΕΚ

3. Λογικό σχήμα δεδομένων από ΕΣΥΕ

Παρατηρούμε απ' το διάγραμμα 8 ότι οι βασικές **μεταβλητές** είναι οι ακόλουθες τρεις:

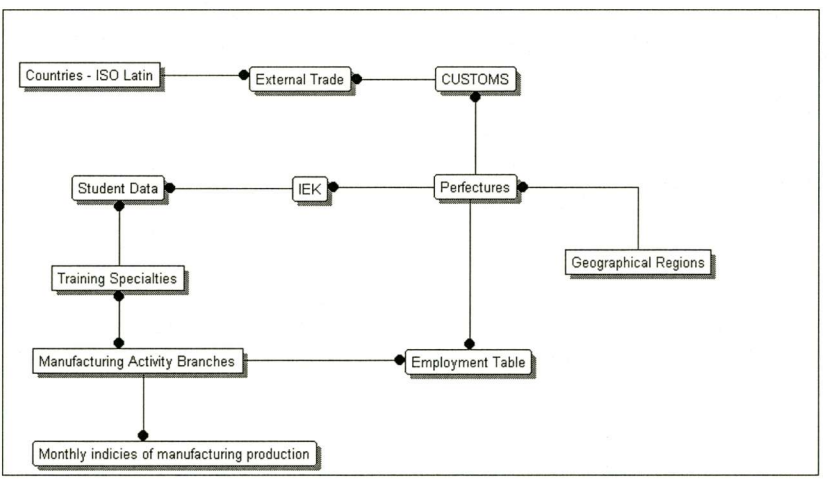
- Κλάδος Βιομηχανικής δραστηριότητας (manufacturing activity branch) κατά μία **ταξινόμηση.**
- Κατηγορία επαγγέλματος κατά μία **ταξινόμηση**
- Χρονική περίοδος (μετρημένη σε μήνες ή χρόνια ανάλογα με τις εξαρτημένες μεταβλητές)

Προκύπτουν επίσης ορισμένοι **δείκτες** (indices) με τη χρήση ορισμένων σταθμισμένων συντελεστών (weighting **coefficients**) και τέλος αποθηκεύεται η τιμή των δεικτών (**value**).



Διάγραμμα 8: Λογικό σχήμα δεδομένων από ΕΣΥΕ

Στη συνέχεια πραγματοποιήθηκε μία σύνθεση των λογικών σχημάτων των δεδομένων που παρέχονται από ΓΓΠΣ, ΟΕΕΚ και ΕΣΥΕ για να καταλήξουμε στις βασικές μεταβλητές και αντιπροσωπεύεται στο διάγραμμα 9. Πρέπει να τονιστεί ότι και τα τρία (ΓΓΠΣ, ΟΕΕΚ και ΕΣΥΕ) λογικά σχήματα δεδομένων απεικονίζονται σε ένα ενιαίο φυσικό σχήμα δεδομένων της βάσης δεδομένων ενός πληροφοριακού συστήματος.



Διάγραμμα 9: Σύνθεση λογικών σχημάτων των δεδομένων από ΓΓΠΣ, ΟΕΕΚ και ΕΣΥΕ

Αυτή η διαδικασία είχε ως στόχο να καταλήξουμε σε ένα λογικό σχήμα για το οποίο να εξετάσουμε τι απαιτήσεις έχει κάθε μεταβλητή για τον καθορισμό της. Τα επί μέρους διαγράμματα μας βοήθησαν επίσης να αντιληφθούμε πώς θέλει κάθε οργανισμός να αποθηκεύονται τα δεδομένα του και να παρουσιάζονται τα αποτελέσματα και τι μετασχηματισμοί είναι αναγκαίοι.

Η *διαδικασία* όμως εξαγωγής των αποτελεσμάτων που ακολουθεί ακόμα δεν ήταν ξεκάθαρα αποτυπωμένη στο λογικό διάγραμμα. Χρησιμοποιήθηκε σχετική βιβλιογραφία [Sundgren, 1996, 1999, 2000, 2004], [Scotney et al, 2002], [Olenski, 1996], [OECD, 1999], [Neuchatel group, 2000], [Karge, 1998], [Hatzopoulos et.al, 1998] από όπου έγινε η αποτύπωση της υπάρχουσας κατάστασης σε διεθνείς οργανισμούς αλλά και βάση μελετών από Ευρωπαϊκά Ερευνητικά προγράμματα στο θέμα αυτό όπως [IPIS consortium, 2001], [Grossmann, 1998], [Denk & Froeschl, 2000]. Επίσης, εξετάστηκαν οι απαιτήσεις των πληροφοριακών συστημάτων [Lenz & Shoshani, 1997], [Pedersen et al, 2002], [Agosta, 2000], [Shoshani, 2003].

Στο επόμενο κεφάλαιο παρουσιάζεται το μοντέλο μεταδεδομένων στο οποίο τελικά καταλήξαμε ότι αντιπροσωπεύει καλύτερα τις ανάγκες τόσο των δημοσίων φορέων της χώρας μας όσο και συναφείς διεθνείς οργανισμούς.

3.2.3. Συγκεκριμενοποίηση των βασικών ιδιοτήτων/κριτηρίων που πρέπει να πληρεί το μοντέλο μεταδεδομένων

Η έρευνα στους οργανισμούς και στις απαιτήσεις των φορέων δημόσιας διοίκησης καθώς και οι τεχνολογικές δυνατότητες που παρέχονται από αυτούς, μας έκαναν να καταλήξουμε σε ορισμένες προδιαγραφές για προϋποθέσεις που πρέπει να πληρεί το μοντέλο των μεταδεδομένων ως προς τη λειτουργικότητά του. Ορισμένες από αυτές είναι οι κάτωθι:

- Να ικανοποιεί κάποια κριτήρια ποιότητας σχεδίασης και περιεχομένου
- Το περιεχόμενο (μεταδεδομένα) να είναι διαθέσιμο από τους οργανισμούς
- Το περιεχόμενο (μεταδεδομένα) να είναι κατάλληλο για τους φορείς δημόσιας διοίκησης
- Να έχει δυνατότητα να εκτελούνται οι απαιτούμενοι μετασχηματισμοί
- Να έχει δυνατότητα να εκτελούνται συναθροίσεις (aggregations) και πολυεπίπεδες αναλύσεις.
- Να μπορεί να επεκταθεί όπου και όποτε χρειάζεται
- Να είναι σχετικά γενικό ώστε να έχει ευρεία εφαρμογή
- Να ακολουθεί το σχήμα δεδομένων του οργανισμού
- Να είναι σύμφωνο ως προς τα πρότυπα χωρών αλλά και διεθνών οργανισμών
- Να επιτρέπει τη σύγκριση των πληροφοριών μεταξύ χωρών και διαφορετικών χρονικών περιόδων

- Να μην είναι «ένα ακόμα μοντέλο» αλλά να περιλαμβάνει όλες τις διαδικασίες συλλογής, ανάλυσης και διάχυσης της πληροφορίας ώστε να χρησιμοποιείται αυτόνομα από κάθε σχετικό οργανισμό.

Όσον αφορά τις βασικές ιδιότητες που πρέπει να πληρούν τα υπό μοντελοποίηση μεταδεδομένα, οι κυριότερες που επιλέξαμε ως πιο αντιπροσωπευτικές για την περίπτωσή μας είναι οι ακόλουθες:

- **Πληρότητα:** Τα στοιχεία δεδομένων και μεταδεδομένων πρέπει να είναι όσο το δυνατόν πληρέστερα, δηλαδή να κρατώνται όσο το δυνατόν περισσότερα μεταδεδομένα, αλλά αποφεύγοντας να κρατούνται περισσότερες από μία φορές ή να μοντελοποιούνται μεταδεδομένα τα οποία δίνουν την ίδια πληροφορία αλλά υπό διαφορετικό όνομα.
- **Διαθεσιμότητα:** Το περιεχόμενο δεδομένων και μεταδεδομένων πρέπει να επιλεχτεί πρώτιστα από τις πηγές πληροφοριών για τις οποίες οι προμηθευτές στοιχείων μπορούν να εγγυηθούν ότι θα είναι διαθέσιμα και πλήρη [Froeschl & Grossmann, 2000].
- **Ενσωμάτωση:** Τα στοιχεία και τα μεταδεδομένα που θα αποθηκευτούν και θα διαμορφωθούν πρέπει να είναι σύμφωνα με τις διεθνώς απαραίτητες πληροφορίες για την ανάπτυξη δεικτών και προτύπων και να έχουν δυνατότητα σύγκρισης με άλλα στατιστικά μοντέλα μεταδεδομένων που αναπτύσσονται από Στατιστικές Υπηρεσίες ή άλλους φορείς δημόσιας διοίκησης, ή είναι προϊόντα ερευνητικών προγραμμάτων.

3.2.4 Επιλογή κατηγοριών μεταδεδομένων που θα μοντελοποιηθούν

Τα σύγχρονα συστήματα διαχείρισης βάσεων δεδομένων (Database Management Systems), προσφέρουν ποικίλα χρήσιμα χαρακτηριστικά γνωρίσματα όπως ο έλεγχος έκδοσης και διαχείρισης της δημοσίευσης στοιχείων [Date, 1990]. Στην περίπτωση των στατιστικών πληροφοριών, αυτά τα χαρακτηριστικά γνωρίσματα είναι τόσο σημαντικά που σήμερα, ουσιαστικά κάθε στατιστική υπηρεσία χρησιμοποιεί κάποια βάση δεδομένων.

Εντούτοις, προκειμένου να χρησιμοποιηθούν αυτά τα συστήματα, πρέπει να συλλεχθούν και αποθηκευτούν τα δεδομένα και τα μεταδεδομένα χρησιμοποιώντας ένα πρότυπο στοιχείων, δηλ. μια συλλογή των λογικών δομών και των χειριστών. Οι δομές διευκρινίζουν ποια στοιχεία και μεταδεδομένα είναι αναγκαία ενώ οι μετασχηματισμοί/διαδικασίες διευκρινίζουν κάθε έγκυρο χειρισμό αυτών των πληροφοριών [Layzell & Loukopoulos, 1989]. Κατά συνέπεια, πριν από τη χρησιμοποίηση ενός προγράμματος διαχείρισης βάσεων δεδομένων για την αποθήκευση των στατιστικών στοιχείων, πρέπει να αποφασίσουμε σχετικά με ποια δεδομένα και μεταδεδομένα πρέπει να μοντελοποιηθούν. Επιπλέον, πρέπει να λάβουμε υπόψη ότι τα μεταδεδομένα διαμορφώνονται με τέτοιο τρόπο, έτσι ώστε οι

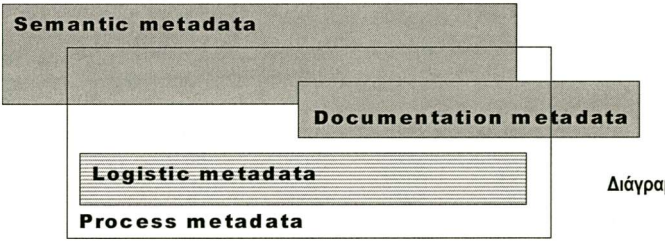
υπολογιστές μπορούν να καταλάβουν την έννοια των στοιχείων και να μας βοηθήσουν ενεργά στους χειρισμούς τους [Olenski, 1996].

Η απόφαση για το ποια μεταδεδομένα πρέπει να μοντελοποιηθούν είναι ένα δύσκολο έργο και έχει περαιτέρω αναλυθεί από [Froeschl (1997), (1999)], [Denk & Froeschl, 2000], [Grossmann et.al, 1998], [Pentaris & Vardaki 1998], [Papageorgiou et.al, 20001a, 2001b], [Vardaki 2004]. Καταλήξαμε ότι:

- δεν πρέπει να λάβουμε υπόψη μας εκείνα τα μεταδεδομένα τα οποία δε θα χρησιμοποιηθούν πουθενά ούτε θα προσδώσουν κάποιο όφελος στον αναλυτή ή τον τελικό χρήστη αλλά απλά θα υπερφορτώσουν το μοντέλο. Ο λόγος είναι ότι αυξάνοντας τον αριθμό των μεταδεδομένων που περικλείονται στο μοντέλο αυξάνεται και η πολυπλοκότητα χρησιμοποίησής του από τους χρήστες αλλά και τα πληροφοριακά συστήματα.
- θα πρέπει να λάβουμε υπόψη μας τις διαφορετικές κατηγορίες χρηστών οι οποίοι ενδιαφέρονται για διαφορετικά επίπεδα ανάλυσης και πληροφορίας, οπότε πρέπει να τους παρέχουμε όλη τη ζητούμενη πληροφορία και μεταπληροφορία. Σημειώνουμε απλά ότι εάν αποτύχουμε να περιλάβουμε ικανοποιητικό αριθμό και ποιότητα μεταδεδομένων, η όλη διαδικασία κινδυνεύει να αποβεί άχρηστη για τους τελικούς αποδέκτες των στατιστικών αποτελεσμάτων.

Στην παρούσα διατριβή μετά από συστηματική μελέτη της βιβλιογραφίας καταλήξαμε ότι τα μεταδεδομένα που θα μοντελοποιηθούν πρέπει να καλύπτουν όλο το φάσμα των βημάτων μιας δειγματοληπτικής έρευνας, από τον καθορισμό του δείγματος έως και τη δημοσίευση των αποτελεσμάτων, οπότε να ληφθεί υπόψη τόσο η επιθυμητή διάταξη των δεδομένων όσο και το μέσο επεξεργασίας και διάχυσης αποτελεσμάτων.

Δυστυχώς, δεν υπάρχει κανένας τρόπος να γνωρίζουμε με ακρίβεια τις μελλοντικές ανάγκες των διάφορων χρηστών. Εντούτοις, είναι απαραίτητο να παρασχεθούν τέσσερις σημαντικοί τύποι μεταπληροφορίας (δες επίσης [Vardaki, 2004a]) οι οποίοι αντιπροσωπεύουν αλληλοκαλυπτόμενες και σχετικές μεταξύ τους κατηγορίες, όπως προκύπτει από το διάγραμμα 10 που δείχνει τη συσχέτισή τους:



Διάγραμμα

Ο πρώτος τύπος είναι τα *σημασιολογικά μεταδεδομένα* (semantic metadata) τα οποία διευκρινίζουν τα μεταδεδομένα που χρησιμοποιούνται σε κάθε διαδικασία, όπως για παράδειγμα 'στατιστικός πληθυσμός', 'δειγματοληψία', 'ταξινόμηση', κλπ. Για να μπορέσουμε όμως να κατανοήσουμε την έννοια των παραπάνω μεταδεδομένων είναι

απαραίτητα τα *μεταδεδομένα τεκμηρίωσης* (documentation metadata). Αυτά είναι κομμάτια μεταπληροφορίας που παρέχουν την τεκμηρίωση (σημασία) ενός πίνακα. Παραδείγματα των μεταδεδομένων τεκμηρίωσης είναι ορισμοί όρων (π.χ. τι είναι "Έρευνα και Τεχνολογική ανάπτυξη"), των 'ετικετών' (π.χ. "Euro"), των καθορισμών του 'στατιστικού πληθυσμού', τεχνικών 'δειγματοληψίας', των 'ταξινομήσεων', κλπ. Ο τρίτος τύπος μεταπληροφορίας είναι τα *λογιστικά μεταδεδομένα* (logistic metadata). Αυτές οι πληροφορίες περιγράφουν πού βρίσκονται τα στοιχεία (π.χ. μια διεύθυνση Διαδικτύου) ή πώς μπορούμε να τα ανακτήσουμε (π.χ μια ερώτηση SQL). Ο τελευταίος τύπος μεταπληροφορίας είναι τα *μεταδεδομένα διαδικασίας* (process metadata). Αυτές οι πληροφορίες χρησιμοποιούνται για το χειρισμό των δεδομένων και καλούνται και διαδικασίες/μετασχηματισμοί (operations) στην παρούσα διατριβή (εξετάζονται ιδιατέρως στο κεφάλαιο 4).

Σημειώνουμε ότι, δεδομένου ότι τα μεταδεδομένα υποδεικνύουν την έννοια των στοιχείων, εάν τροποποιούμε ένα σύνολο δεδομένων πρέπει επίσης να αλλάξουμε τα συνοδευτικά μεταδεδομένα τους. Επομένως, πρέπει πάντα ταυτόχρονα να χειριζόμαστε τα δεδομένα και τα μεταδεδομένα για να διασφαλίσουμε την ποιότητα των στατιστικών αποτελεσμάτων [Papageorgiou *et al.*, 1999].

3.2.5 Επιλογή Τεχνικής μοντελοποίησης

Σχετικά με την επιλογή τεχνικής μοντελοποίησης, μία τυπική διαδικασία για την αποθήκευση μεταδεδομένων είναι με τη χρήση ενός σχεσιακού μοντέλου (Entity-Relationship (E-R) Model) [Chen, 1976]. Οι Layzell και Loucopoulos (1989) εξέτασαν τα βήματα που πρέπει να ακολουθηθούν πριν τη δημιουργία ενός τέτοιου μοντέλου, τα οποία σχετίζονται με την επιλογή μιας συγκεκριμένης, κατάλληλης μεθοδολογίας για δομημένη παρουσίαση των μεταδεδομένων.

Τα τελευταία χρόνια δύο μεθοδολογίες είναι οι επικρατέστερες: Η πρώτη είναι το Entity–Relationship Model (E-R Model) (μοντέλο οντοτήτων-σχέσεων) και η δεύτερη ακολουθεί το Object-Oriented Paradigm (O-O Model) [OMG,2002].

Η τεχνολογία του E-R μοντέλου προτάθηκε το 1976 (Chen, 1976) ως ένα εννοιολογικό μοντέλο δεδομένων που μετατρέπει τον 'πραγματικό κόσμο' σε οντότητες και σχέσεις μεταξύ τους. Χρησιμοποιήθηκε ευρέως τα τελευταία χρόνια επειδή αποτελεί τη βάση για τη μοντελοποίηση σε μία σχεσιακή βάση δεδομένων, όντας χρησιμοποιούμενη από πολλά πληροφοριακά συστήματα. Η βασικές έννοιες ήταν: οντότητα (entity), ιδιότητα/χαρακτηριστικό (attribute) and σχέση (relationship).

Η O-O μέθοδος στηρίζεται στην έννοια του αντικειμένου (object) – το οποίο προσδιορίζεται από κλάσεις (class) και χαρακτηριστικά/ιδιότητες κάθε κλάσης - και διέπεται και από σχέσεις μεταξύ των αντικειμένων, όπως και το E-R. Επιπρόσθετα όμως έχει και τη δυνατότητα εφαρμογής συγκεκριμένων μετασχηματισμών (operations/ transformations) πάνω στις κλάσεις βάσει των ιδιοτήτων τους, οι οποίοι καθιστούν την O-O μέθοδο κατάλληλη για την κατανόηση του εννοιολογικού μοντέλου από τα πληροφοριακά συστήματα [Papazoglou, et.al, 2000].

Βασιζόμενοι σε αυτό το πλεονέκτημα, χρησιμοποιήσαμε στην παρούσα διατριβή την Ο-Ο μέθοδο για τη δημιουργία των μοντέλων των μεταδεδομένων

3.2.6 Επιλογή Γλώσσας μοντελοποίησης

Η γλώσσα μοντελοποίησης επίσης παίζει ιδιαίτερο ρόλο στην πληρότητα και ευελιξία του μοντέλου. Η χρήση της UML (Universal Modeling Language) [OMG, 2002] έχει το πλεονέκτημα ότι το δημιουργημένο μοντέλο είναι πιο ευέλικτο, οπότε μπορούμε να προσθέτουμε ή να μετατρέπουμε μεταδεδομένα ανάλογα με τις ανάγκες μας και παρουσιάζει ξεκάθαρες και πολλαπλές σχέσεις μεταξύ των οντοτήτων. Επιπρόσθετα, οι δυνατοί μετασχηματισμοί που εφαρμόζονται στα διάφορα μεταδεδομένα μπορούν να απεικονιστούν στο ίδιο μοντέλο. Η Rational Rose και η Argo είναι δύο πακέτα μοντελοποίησης σε UML.

Ασφαλώς, η χρήση XML (eXtensible Markup Language)στις μέρες μας έχει ωθήσει προς μία στροφή στην εξαγωγή των μοντέλων από UML σε XML, ως το τελικό σχήμα που δέχεται ένα σύστημα για να επεξεργαστεί τα μεταδεδομένα και τα δεδομένα. Παρόλα αυτά, στο εννοιολογικό επίπεδο, πριν την ενσωμάτωση του μοντέλου στο πληροφοριακό σύστημα, η χρήση της UML ενδείκνυται για τη μοντελοποίηση των μεταδεδομένων, τόσο στον κατασκευαστή, επειδή είναι πιο εκφραστική και λειτουργική αλλά και στον τελικό χρήστη κυρίως επειδή το αποτέλεσμα είναι πιο περιγραφικό και παρουσιάζει καθαρά τις διαδικασίες, τους μετασχηματισμούς και τις σχέσεις.

Ένας αριθμός μοντέλων έχει τα τελευταία χρόνια καταγραφεί στη βιβλιογραφία, κυρίως ως προϊόν σχετικών ερευνητικών προγραμμάτων (projects) που έχουν χρηματοδοτηθεί από την Ευρωπαϊκή Ένωση. Τα μεταδεδομένα που έχουν χρησιμοποιηθεί είχαν σχέση με συγκεκριμένες λειτουργίες και ανάγκες που το συγκεκριμένο project είχε να καλύψει και να δώσει λύσεις. Ενδεικτικά αναφέρονται τα μοντέλα μεταδεδομένων των projects IDARESA (Grossmann et al. (1998)], ADDSIA [Hatzopoulos et al. (1998)], MISSION (http://www.epros.ed.ac.uk/mission), METANET (http://www.epros.ed.ac.uk/metanet), IQML (http://www.epros.ed.ac.uk/iqml), COSMOS (http://www.epros.ed.ac.uk/cosmos) και IPIS (http://www.instore.gr/ipis).

Παρόλα αυτά, κάθε ένα από αυτά τα μοντέλα έχουν δημιουργηθεί για συγκεκριμένες ανάγκες του κύκλου μιάς δειγματοληπτικής διαδικασίας, για παράδειγμα μόνο για τη δημιουργία ερωτηματολογίων (IQML), τη διάχυση αποτελεσμάτων και εναρμόνισή τους (IPIS), τη συλλογή και διαμόρφωση της πληροφορίας, κλπ. Χρειάζεται λοιπόν ένα ολοκληρωμένο μοντέλο να αποτυπώνει τη σειρά των διαδικασιών από τη συλλογή των στοιχείων μέχρι τη διάχυση των στατιστικών αποτελεσμάτων, το οποίο να μπορεί να χρησιμοποιηθεί από τους περισσότερους φορείς δημόσιας διοίκησης και διεθνείς οργανισμούς.

Συνεπώς στην παρούσα διατριβή, χρησιμοποιήθηκε η UML (Rational Rose 2000 Enterprise Edition) για τη δημιουργία του μοντέλου μεταδεδομένων για να εκφράσουμε τις κλάσεις, τις διαδικασίες, τους μετασχηματισμούς και τις σχέσεις που διέπουν τη συνολική διαδικασία δειγματοληπτικής έρευνας.

ΚΕΦΑΛΑΙΟ 4

ΠΡΟΤΕΙΝΟΜΕΝΟ ΣΤΑΤΙΣΤΙΚΟ ΜΟΝΤΕΛΟ ΜΕΤΑΔΕΔΟΜΕΝΩΝ ΓΙΑ ΤΗ ΣΥΛΛΟΓΗ, ΕΠΕΞΕΡΓΑΣΙΑ ΚΑΙ ΔΙΑΧΥΣΗ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ / ΜΕΤΑΠΛΗΡΟΦΟΡΙΑΣ

4.1 Εισαγωγή

Ένα ολοκληρωμένο, σημασιολογικά πλούσιο στατιστικό μοντέλο μεταδεδομένων σχεδιάζεται για να καλύψει τα σημαντικότερα στάδια της επεξεργασίας στατιστικών πληροφοριών (συλλογή και ανάλυση δεδομένων συμπεριλαμβανομένης της εναρμόνισης, της επεξεργασίας των δεδομένων και των μεταδεδομένων τους και της διάδοσης/ διάχυσης των αποτελεσμάτων), το οποίο μπορεί να ελαχιστοποιήσει την πολυπλοκότητα των δεδομένων που αποθηκεύονται καθώς και των προβλημάτων συμβατότητας μεταξύ των Στατιστικών Πληροφοριακών Συστημάτων (ΣΠΣ) (Statistical Information Systems).

Η σημασιολογία του μοντέλου αναλύεται, περιγράφοντας κάθε μέρος της στατιστικής επεξεργασίας. Επιπλέον, καθορίζονται τα λογιστικά μεταδεδομένα (logistic metadata) για τη θέση και το σχήμα των στοιχείων.

Προκειμένου να διασφαλίσουμε την ποιότητα στοιχείων από τον κίνδυνο ενός λανθασμένου ή ατελούς συνδυασμού δεδομένων-μεταδεδομένων, εισάγουμε ένα σύνολο μετασχηματισμών/διαδικασιών (operations/transformations) και εξετάζουμε τις διάφορες ιδιότητες τους. Οι μετασχηματισμοί είναι αλγόριθμοι που καθορίζονται πέρα από την κοινή περιοχή των στοιχείων και των μεταδεδομένων. Μερικοί από αυτούς είναι η προσθήκη και η αφαίρεση μιας μεταβλητής, η προσθήκη και η επιλογή των στοιχείων, και ο μετασχηματισμός ομαδοποίησης. Παρουσιάζονται πάνω στις κλάσεις στις οποίες εφαρμόζονται και εξετάζονται οι ιδιότητές τους.

Επιπλέον, συζητάμε πώς το προτεινόμενο πλαίσιο μπορεί να διευκολύνει την είσοδο και την ανάλυση πληροφοριών σε ένα ΣΠΣ. Τέλος, καταδεικνύουμε σε μια περιπτωσιολογική μελέτη πώς το προτεινόμενο μοντέλο μεταδεδομένων μπορεί να εφαρμοστεί για την παραγωγή νέων οικονομικών δεικτών με ταυτόχρονη χρήση μεταδεδομένων.

4.2 Δομή του προτεινόμενου Στατιστικού Μοντέλου Μεταδεδομένων

Στη συνέχεια, προτείνουμε ένα στατιστικό μοντέλο δεδομένων/μεταδεδομένων το οποίο εξετάζει τα κύρια στάδια της επεξεργασίας στατιστικών στοιχείων, δηλαδή: συλλογή δεδομένων, επεξεργασία, ανάλυση, εναρμόνιση και διάχυση αποτελεσμάτων. Ένα τέτοιο ολοκληρωμένο μοντέλο θα επιτύχει την ελαχιστοποίηση των προβλημάτων πολυπλοκότητας και συμβατότητας των διάφορων υποσυστημάτων ενός πληροφοριακού συστήματος, βασισμένων σε μια κοινή βάση αποθήκευσης μεταδεδομένων για τη συνολική επεξεργασία πληροφοριών.

Θεωρούμε τα σημασιολογικά και τα μεταδεδομένα τεκμηρίωσης και, προκειμένου να διασφαλίσουμε την ποιότητα των δεδομένων από τυχόν λανθασμένους συνδυασμούς δεδομένων/μεταδεδομένων, εισάγουμε ορισμένα μεταδεδομένα διαδικασίας τα οποία καθορίζονται βάσει του ρόλου τους στο μοντέλο και των ιδιοτήτων τους. Τέλος, τα λογιστικά μεταδεδομένα λαμβάνονται υπόψη για τη θέση των δεδομένων μέσα στη βάση (δες επίσης [Muller et al, 1999], [Papageorgiou et al, 2001a, 2001b]).

Το μοντέλο σχεδιάστηκε με τη χρήση μιας Unified Modeling Language (UML) [OMG, 2002] ώστε να διασφαλιστεί η ευελιξία του σε μελλοντικές πιθανές ανάγκες προσθήκης μεταδεδομένων καθώς επίσης και για την καλύτερη κατανόηση της διεπίπεδης αναπαράστασης των μεταδεδομένων.

Το πρώτο επίπεδο στο μοντέλο αποτελεί η '*κλάση*' η οποία αναπαριστάται με ένα κουτάκι στη UML και αποτελεί ένα βασικό σημασιολογικό ή λογιστικό μεταδεδομένο το οποίο χρειάζεται και άλλα μεταδεδομένα για να περιγραφεί πλήρως. Τα επί μέρους αυτά μεταδεδομένα είναι τα χαρακτηριστικά/ιδιότητες της κάθε κλάσης και αποτελούν το δεύτερο επίπεδο μοντελοποίησης.

Οι μετασχηματισμοί (operators/transformations) περικλείονται σε κάθε κλάση όπου εφαρμόζονται σε ένα παράλληλο επίπεδο με τα χαρακτηριστικά/ιδιότητες της κλάσης μια που και οι μετασχηματισμοί αποτελούν και αυτοί ιδιότητες της κλάσης.

Προκειμένου να παρουσιαστούν και να εξεταστούν τα μεταδεδομένα που περιλαμβάνονται στο μοντέλο μας, παρουσιάζουμε το μοντέλο κατά τμήματα, ανάλογα με τη διαδικασία που περιγράφει, δηλαδή: i) συλλογή και ανάλυση δεδομένων (κύριο μέρος) συμπεριλαμβανομένης της εναρμόνισης και επεξεργασίας δεδομένων ii) και διαδικασίες παραγωγής αποτελεσμάτων και iii) διάχυση αποτελεσμάτων. Για κάθε περίπτωση παρουσιάζονται και τα μεταδεδομένα διαδικασίας και οι μετασχηματισμοί που εφαρμόζονται στις αντίστοιχες κλάσεις, καθώς επίσης και τα λογιστικά μεταδεδομένα. Επιπλέον απεικονίζεται η δυνατότητα διασφάλισης της ποιότητας κάθε σταδίου της διαδικασίας.

Το μοντέλο απεικονίζεται στην αγγλική γλώσσα επειδή αυτή υποστηρίζεται από τη Rational rose UML που χρησιμοποιήθηκε αλλά και για να είναι δυνατή η χρήση του από διεθνείς οργανισμούς.

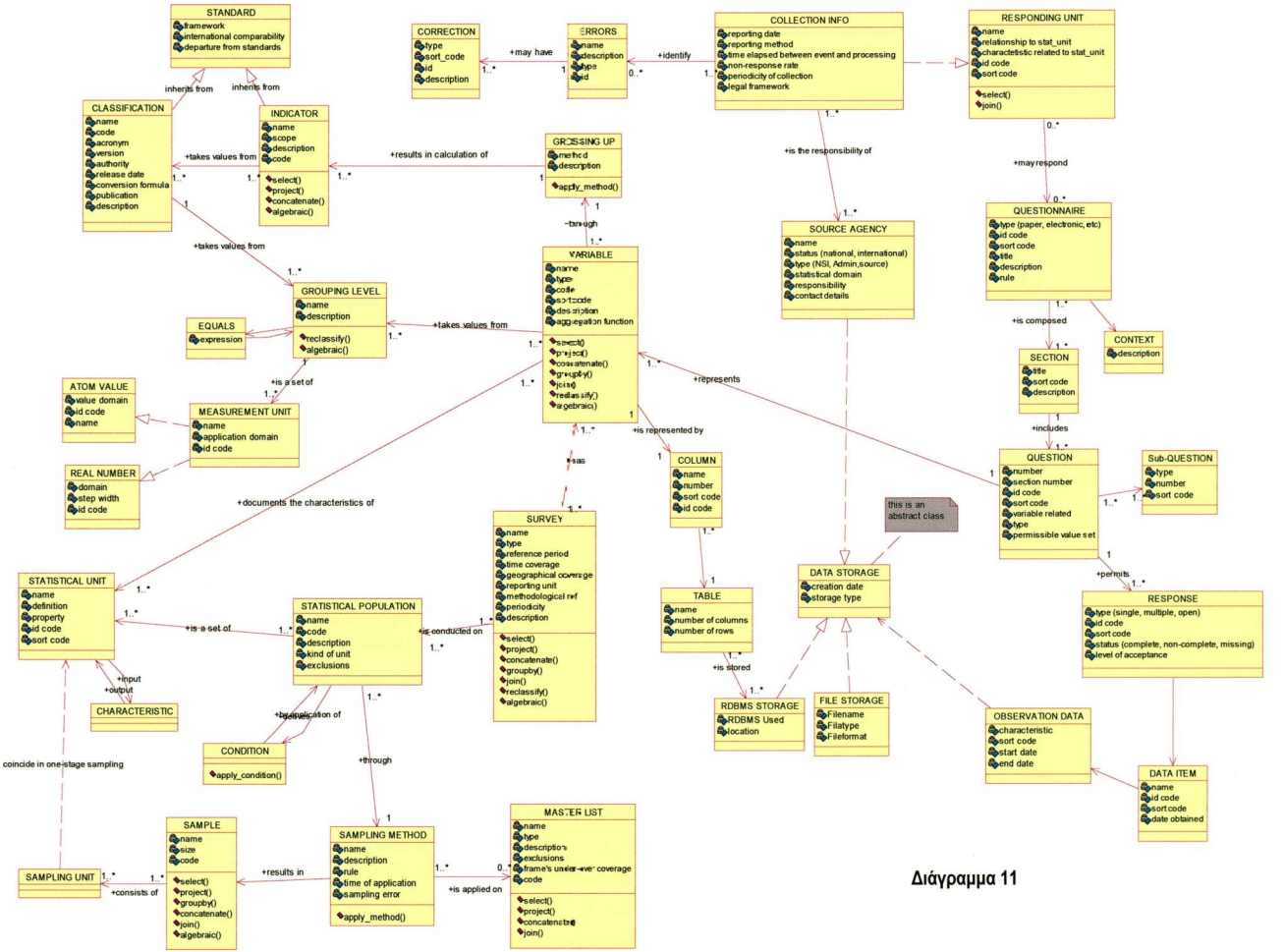
4.2.1 Τμήμα συλλογής και ανάλυσης δεδομένων του μοντέλου

Στο Διάγραμμα 11 παρουσιάζεται το κύριο μέρος του μοντέλου. Αποτελείται από τα στάδια συλλογής, επεξεργασίας και ανάλυσης δεδομένων καθώς και μεταδεδομένα που συμβάλλουν σε περιπτώσεις εναρμόνισης.

Η φάση συλλογής δεδομένων αποτελείται από το πλαίσιο ερωτηματολογίων, τη διαδικασία συλλογής απαντήσεων και έπειτα την αποθήκευση του συνόλου των αποκτηθέντων δεδομένων. Η διαδικασία αρχίζει με την ΣΥΛΛΟΓΗ ΠΛΗΡΟΦΟΡΙΩΝ (COLLECTION INFO). Ένας ή περισσότεροι ΟΡΓΑΝΙΣΜΟΙ (SOURCE AGENCIES) είναι αρμόδιοι για τη συλλογή και τη σύνταξη των στατιστικών. Σε όλους τους τύπους

ΕΡΕΥΝΩΝ (SURVEY) (απογραφή, δειγματοληπτική έρευνα ή απευθείας από πηγή δημόσιου φορέα) οι πληροφορίες συλλέγονται μέσω ενός ΕΡΩΤΗΜΑΤΟΛΟΓΙΟΥ (QUESTIONNAIRE) που πρέπει να περιγραφεί λεπτομερώς, προκειμένου να είναι σε θέση να κωδικοποιηθούν και να ταξινομηθούν οι ερωτήσεις και οι σχετικές απαντήσεις επαρκώς σε ένα ΣΠΣ. Οι ΕΝΟΤΗΤΕΣ (SECTIONS) του ΕΡΩΤΗΜΑΤΟΛΟΓΙΟΥ περιγράφονται καθώς επίσης και κάθε ΕΡΩΤΗΣΗ (QUESTION) και υπο—ερώτηση (SUB-QUESTION). Κάθε ΕΡΩΤΗΣΗ αντιπροσωπεύει μια ή περισσότερες ΜΕΤΑΒΛΗΤΕΣ (VARIABLES) του SURVEY και εξετάζεται σύμφωνα με τον τύπο της και το επιτρεπόμενο σύνολο τιμών προκειμένου να προχωρήσουμε στην ΚΩΔΙΚΟΠΟΙΗΣΗ (CODING) και EDITING (βλ. παράγραφο 4.2.2.1), καθώς επίσης και τη δυνατότητα χρήση της κατάλληλης ΜΟΝΑΔΑΣ ΜΕΤΡΗΣΗΣ (MEASUREMENT UNIT). Αυτοί που απαντούν (RESPONDING UNITS) σε κάθε ΕΡΩΤΗΜΑΤΟΛΟΓΙΟ δίνουν μια ή περισσότερες ΑΠΑΝΤΗΣΕΙΣ (RESPONSES) σε κάθε ΕΡΩΤΗΣΗ (ανάλογα με τον τύπο ερώτησης). Όλα τα αποκτηθέντα ΣΤΟΙΧΕΙΑ (DATA ITEMS) αποθηκεύονται ως ΣΤΟΙΧΕΙΑ ΠΑΡΑΤΗΡΗΣΗΣ (OBSERVATIONS DATA) για την επεξεργασία και ανάλυση. Επιπλέον, κατά τη διάρκεια της συλλογής και της επεξεργασίας εντοπίζονται τυχόν ΛΑΘΗ (ERRORS) τα οποία μπορεί να επιδέχονται διόρθωση (CORRECTED) (βλ. επίσης τις παραγράφους 4.2.2.2 και 4.2.3 για περισσότερες λεπτομέρειες) Επίσης, η ΕΡΕΥΝΑ που πραγματοποιείται μπορεί να έχει ένα σύνολο ΜΕΤΑΒΛΗΤΩΝ που αντιπροσωπεύει το σύνολο ή ένα μέρος των ΕΡΩΤΗΣΕΩΝ που περιλαμβάνονται στο ΕΡΩΤΗΜΑΤΟΛΟΓΙΟ. Αυτή η ΕΡΕΥΝΑ συσχετίζεται με έναν ΣΤΑΤΙΣΤΙΚΟ ΠΛΗΘΥΣΜΟ (STATISTICAL POPULATION) που καθορίζεται από ένα σύνολο μονάδων ενδιαφέροντος, οι οποίες καλούνται ΣΤΑΤΙΣΤΙΚΕΣ ΜΟΝΑΔΕΣ (STATISTICAL UNITS). Αυτό το σύνολο μονάδων είτε περιγράφεται χρησιμοποιώντας έναν ΚΑΤΑΛΟΓΟ (MASTER LIST) είτε είναι ένα υποσύνολο ενός μεγαλύτερου ΣΤΑΤΙΣΤΙΚΟΥ ΠΛΗΘΥΣΜΟΥ. Στη δεύτερη περίπτωση, οι δύο σχετικοί ΣΤΑΤΙΣΤΙΚΟΙ ΠΛΗΘΥΣΜΟΙ συσχετίζονται ο ένας με τον άλλον μέσω ενός CONDITION.

Στην περίπτωση μιας δειγματοληπτικής ΕΡΕΥΝΑΣ, μια ΔΕΙΓΜΑΤΟΛΗΠΤΙΚΗ ΜΕΘΟΔΟΣ (SAMPLING METHOD) εφαρμόζεται στο ΣΤΑΤΙΣΤΙΚΟ ΠΛΗΘΥΣΜΟ για να παραγάγει ένα μικρότερο, αλλά ο αρκετά αντιπροσωπευτικό, σύνολο μονάδων που ονομάζουμε ΔΕΙΓΜΑ (SAMPLE), το οποίο αποτελείται από ΜΟΝΑΔΕΣ (SAMPLING UNITS).



Διάγραμμα 11

Κάθε ΜΕΤΑΒΛΗΤΗ της ΕΡΕΥΝΑΣ παίρνει τις τιμές από ένα ΕΠΙΠΕΔΟ ΟΜΑΔΟΠΟΙΗΣΗΣ (GROUPING LEVEL), που αποτελείται από ΜΟΝΑΔΕΣ ΜΕΤΡΗΣΗΣ (MEASUREMENT UNITS). Όταν αναφερόμαστε σε κατηγορικές μεταβλητές, μια ΜΟΝΑΔΑ ΜΕΤΡΗΣΗΣ είναι ένα ATOM VALUE και όταν αναφερόμαστε σε αριθμητικές μεταβλητές, είναι ένας ΠΡΑΓΜΑΤΙΚΟΣ ΑΡΙΘΜΌΣ (REAL NUMBER). Κάθε ΜΟΝΑΔΑ ΜΕΤΡΗΣΗΣ συσχετίζεται με άλλες ΜΟΝΑΔΕΣ ΜΕΤΡΗΣΗΣ μέσω της κλάσης "EQUALS", που διευκρινίζει το είδος αυτής της σχέσης, π.χ. 'περιέχει', 'ισοδύναμος με', κλπ. Ένα διαταγμένο σύνολο ΕΠΙΠΕΔΩΝ ΟΜΑΔΟΠΟΙΗΣΗΣ είναι μια ΤΑΞΙΝΟΜΗΣΗ (CLASSIFICATION). Μια ή περισσότερες ΜΕΤΑΒΛΗΤΕΣ χρησιμοποιούνται για να υπολογίσουν έναν ΔΕΙΚΤΗ (INDICATOR). Τέλος, οι ΤΑΞΙΝΟΜΗΣΕΙΣ και οι ΔΕΙΚΤΕΣ μπορούν να συμμορφωθούν με ορισμένα ΠΡΟΤΥΠΑ (STANDARDS).

Η ύπαρξη των δυνατοτήτων μετατροπής ΜΟΝΑΔΩΝ ΜΕΤΡΗΣΗΣ (μέσω της κλάσης "EQUALS",) και των ΤΑΞΙΝΟΜΗΣΕΩΝ στο μοντέλο μεταδεδομένων συμβάλλει στη διαδικασία εναρμόνισης και δίνει τη δυνατότητα ενσωμάτωσης αυτοματοποιημένων μετασχηματισμών σε ένα ΣΠΣ. Πρέπει επίσης να αναφερθεί ότι η χρήση ενός μοντέλου μεταδεδομένων είναι καθαυτή ένα μεγάλο βήμα προς την εναρμόνιση μεταδεδομένων.

4.2.2 Διαδικασία επεξεργασίας δεδομένων

4.2.2.1 Περιγραφή διαδικασιών

Μετά τη συλλογή των δεδομένων οι ακόλουθες διαδικασίες πρέπει να ακολουθηθούν.

Εισαγωγή δεδομένων (DATA ENTRY): Πρόκειται για εκείνη τη φάση της έρευνας κατά την οποία η πληροφορία η οποία έχει καταγραφεί σε ένα ερωτηματολόγιο κατά τη συλλογή των δεδομένων, μετατρέπεται σε μια μορφή η οποία ερμηνεύεται εύκολα από έναν ηλεκτρονικό υπολογιστή.

EDITTING: Το *statistical data editing* είναι η διαδικασία κατά την οποία εξετάζεται αν τα στατιστικά δεδομένα περιέχουν σφάλματα και τα σφάλματα εκείνα τα οποία ανακαλύπτονται διορθώνονται στη συνέχεια. Είναι δηλαδή μια διαδικασία για την ανακάλυψη και την προσαρμογή (adjusting) μεμονωμένων σφαλμάτων τα οποία προκύπτουν κατά τη συλλογή και την εισαγωγή των δεδομένων ή ακόμα και για την εξέταση δεδομένων για τα οποία υπάρχει αμφιβολία ως προς την εγκυρότητά τους. Πρόκειται λοιπόν για μια απαραίτητη διαδικασία κατά τη διεξαγωγή μιας έρευνας καθώς τα σφάλματα στα δεδομένα της έρευνας είναι πιθανό να επηρεάσουν αρνητικά τις εκτιμήσεις, να προκαλέσουν δυσκολίες στην περαιτέρω επεξεργασία της στατιστικής πληροφορίας και φυσικά να μειώσουν την εμπιστοσύνη των χρηστών στα αποτελέσματα της έρευνας.

Γενικά, το editing ξεκινά με τη συγκεκριμενοποίηση ενός συνόλου κανόνων οι οποίοι καλούνται edits ή edit rules. Πρόκειται για λογικές συνθήκες ή περιορισμούς στις τιμές των δεδομένων που πρέπει να ικανοποιούνται για να θεωρηθούν τα δεδομένα έγκυρα. Είναι δηλαδή έλεγχοι προκειμένου να αναγνωριστούν λανθασμένα ή «ύποπτα» δεδομένα. Για παράδειγμα, πιθανά edits σε μια κοινωνικο-οικονομική έρευνα είναι ότι στην ηλικία δεν πρέπει να υπάρχει κενό, ότι η ηλικία πρέπει να είναι ένας ακέραιος

αριθμός μεταξύ του 0 και του 120, ότι αν η ηλικία είναι μικρότερη των 16 τότε η οικογενειακή κατάσταση θα πρέπει να είναι ανύπαντρος κ.τ.λ. Αυτά είναι παραδείγματα microedits, κανόνων δηλαδή που σχετίζονται με τις πληροφορίες μεμονωμένων ερωτώμενων. Οι κανόνες ανακάλυψης εκτρόπων παρατηρήσεων (outliers) είναι επίσης edits με την ευρεία έννοια του όρου καθώς ανακαλύπτουν όχι απαραίτητα εσφαλμένες αλλά «ύποπτες» τιμές, οι οποίες βρίσκονται εκτός ενός εύρους το οποίο καθορίζεται από το σύνολο των δεδομένων ή από προηγούμενη εμπειρία και γνώση. Τέτοια edits, τα οποία συνδέουν δεδομένα διαφόρων αποκρινόμενων, καλούνται macroedits.

Το editing καταναλώνει ένα σημαντικό ποσοστό του συνολικού προϋπολογισμού μιας έρευνας. Έχει εκτιμηθεί ότι το κόστος αυτής της διαδικασίας φθάνει το 20% του προϋπολογισμού σε έρευνες νοικοκυριών ενώ σε έρευνες που αφορούν επιχειρήσεις φθάνει το 40%.

Οι περισσότεροι τύποι των edit rules προκύπτουν από μια προκαθορισμένη θεώρηση για το πώς πρέπει να συμπεριφέρεται ο πληθυσμός. Κάτι τέτοιο αναφέρεται ως μοντέλο editing και σε ορισμένες περιπτώσεις αυτό το μοντέλο είναι αρκετά απλό. Περισσότερες πληροφορίες σε [De Jong, 2003], [De Waal & Quere, 2003], [De Waal & Pannekoek, 2004], [Nordbotten, 2000], [Petrakos et.al, 2004].

Κωδικοποίηση (CODING): Είναι μια διαδικασία κατά την οποία ακατέργαστα (raw) δεδομένα μιας έρευνας, συνήθως υπό μορφή απαντήσεων σε ερωτήσεις ανοικτού τύπου, ταξινομούνται και μετασχηματίζονται σε μία μορφή η οποία μπορεί να χρησιμοποιηθεί στα τελικά στάδια της έρευνας, όπως στην εκτίμηση, την παρουσίαση και την ανάλυση των αποτελεσμάτων. Σκοπός λοιπόν της κωδικοποίησης είναι να ταξινομήσει τις πληροφορίες που περιέχουν τα δεδομένα σε κατηγορίες, οι οποίες μπορούν να χρησιμοποιηθούν στην ανάλυση των δεδομένων.

Μια συνηθισμένη διαδικασία κωδικοποίησης περιλαμβάνει δύο στάδια. Κατά το πρώτο στάδιο ο ερευνητής πρέπει να δημιουργήσει ένα πλαίσιο ταξινόμησης ή κωδικοποίησης, πρέπει δηλαδή να επιλέξει και να ορίσει τις κατηγορίες οι οποίες θα χρησιμοποιηθούν για την κωδικοποίηση και να αντιστοιχίσει σε κάθε κατηγορία έναν αριθμό, έναν κωδικό. Αυτό το πλαίσιο είναι γνωστό ως ονοματολογία/ταξινόμηση (nomenclature/ classification) ή λεξικό (dictionary) ή λίστα κωδικών και περιλαμβάνει ένα σύνολο οδηγιών για τον τρόπο με τον οποίο θα γίνει η κωδικοποίηση. Περιέχει επίσης ορισμούς και περιγραφές των διαφόρων κατηγοριών και αποδίδει σε κάθε κατηγορία ένα συγκεκριμένο αριθμό. Κατά το δεύτερο τώρα στάδιο ταξινομούνται οι γραπτές ή προφορικές απαντήσεις των ερωτώμενων στις προκαθορισμένες κατηγορίες, εργασία η οποία μπορεί να εκτελεστεί με διάφορους τρόπους όπως θα διαπιστωθεί στη συνέχεια. Πρόκειται λοιπόν για μια σύνθετη διαδικασία και στις μέρες μας οι περισσότερες στατιστικές υπηρεσίες διενεργούν έρευνες ώστε να βελτιωθούν τα δύο στάδια της κωδικοποίησης και αναζητούν κυρίως νέους τρόπους για την ταξινόμηση των απαντήσεων στις διάφορες κατηγορίες. Όσο για τις τεχνικές κωδικοποίησης, αυτές ποικίλλουν ανάλογα με τον οργανισμό. Οι Michiels and Hacking (2004) παρουσιάζουν τρεις τεχνικές κωδικοποίησης που χρησιμοποιεί η Statistics Nethelands και τις συγκρίνουν σε περιπτώσεις ταξινόμησης επαγγελμάτων, μορφωτικού επιπέδου και οικονομικής δραστηριότητας. Στη χώρα μας η ΕΣΥΕ χρησιμοποιεί manual coding

προκειμένου να κωδικοποιήσει τις απαντήσεις στις διάφορες έρευνες, με αποτέλεσμα να απαιτείται μεγάλο χρονικό διάστημα για την ολοκλήρωση της διαδικασίας της κωδικοποίησης και επομένως και της επεξεργασίας των δεδομένων.

IMPUTATION: (απόδοση τιμών) είναι η διαδικασία που χρησιμοποιείται για την αντικατάσταση τιμών που είναι ελλείπουσες ή μη έγκυρες ή μη συμβατές με τις άλλες τιμές και εντοπίστηκαν μέσω της διαδικασίας του editing. Πραγματοποιείται συνήθως με την αλλαγή ορισμένων από τις απαντήσεις ή με την απόδοση κάποιων τιμών σε αυτές έτσι ώστε να διασφαλιστεί η υψηλή ποιότητα των εκτιμήσεων και να δημιουργηθεί ένα σύνολο δεδομένων το οποίο θα φαίνεται λογικό και θα διακρίνεται από εσωτερική συμβατότητα. Βέβαια αρκετά από αυτά τα προβλήματα θα μπορούσαν να είχαν λυθεί νωρίτερα, σε προηγούμενα στάδια της έρευνας , μέσω εκ νέου επαφής με τον ερωτώμενο ή μέσω manual διόρθωσης των ερωτηματολογίων. Όμως γενικά κάτι τέτοιο δεν είναι πάντοτε δυνατό είτε λόγω προβλημάτων ενόχλησης του ερωτώμενου (response burden) είτε λόγω κόστους είτε λόγω χρονικών περιορισμών ώστε να εκδοθούν εγκαίρως τα αποτελέσματα.

Η στρατηγική του imputation έχει ως στόχο την ελάττωση της μεροληψίας των εκτιμήσεων της έρευνας που προκύπτει λόγω των missing values. Επίσης επιτρέπει τη διεξαγωγή της έρευνας σαν να υπήρχε πλήρες σύνολο δεδομένων και διευκολύνει την ανάλυση και την παρουσίαση των αποτελεσμάτων. Όμως η εφαρμογή αυτής της στρατηγικής δεν οδηγεί υποχρεωτικά σε εκτιμήσεις οι οποίες είναι λιγότερο μεροληπτικές από εκείνες οι οποίες θα προέκυπταν από το μη πλήρες σύνολο δεδομένων.

Μία μέθοδος imputation είναι η λογική απόδοση τιμών (deductive or logical imputation) η οποία χρησιμοποιείται σε εκείνες τις περιπτώσεις στις οποίες η τιμή που λείπει είναι δυνατό να συμπληρωθεί με μία τιμή η οποία προκύπτει με βάση κάποιους λογικούς συλλογισμούς. Συχνά επίσης χρησιμοποιείται και η μέθοδος της αντικατάστασης από τη μέση τιμή (mean imputation overall). Πρόκειται για μία απλή στην εφαρμογή της μέθοδο η οποία σε κάθε missing value για τη μεταβλητή y αποδίδει το μέσο όρο όπως αυτός υπολογίζεται από το σύνολο των αποκρινόμενων για τη μεταβλητή αυτή (δες και [Laaksonen, 2002], [De Waal, 2000]).

Σε αρκετές στατιστικές υπηρεσίες υπάρχει διαθέσιμο software για τη διαδικασία του imputation. Ανάμεσά τους και η Statistics Canada η οποία έχει αναπτύξει δύο συστήματα, το GENESIS (GENEralised SImulation System) και το SEVANI (System for Estimation of VAriance due to Nonresponse and Imputation) [Beaumont et.al (2003)].

4.2.2.2 Περιγραφή του τμήματος του μοντέλου για την επεξεργασία δεδομένων

Σε όλες τις προαναφερθείσες διαδικασίες, η εφαρμογή ορισμένων *μοντέλων* μπορεί να εξασφαλίσει τη συμβατότητα με άλλες πηγές και την υψηλή ποιότητα των στατιστικών αποτελεσμάτων [Olenski, 1996], [Papageorgiou et.al. 1999b], [Froeschl & Grossmann, 2001].

Στο Διάγραμμα 12 εξετάζεται το τμήμα του μοντέλου που απεικονίζει τη σειρά διαδικασιών. Γίνεται όλο και περισσότερο κοινή πρακτική να ενσωματωθούν μερικές

από τις διαδικασίες προετοιμασίας δεδομένων μαζί με τη διαδικασία συλλογής δεδομένων, κάτι που βελτιώνει τη δυνατότητα να ανιχνευθούν και να διορθωθούν τα λάθη σε ένα αρχικό στάδιο, π.χ. κατά τη διάρκεια της ολοκλήρωσης ερωτηματολογίων στο web.

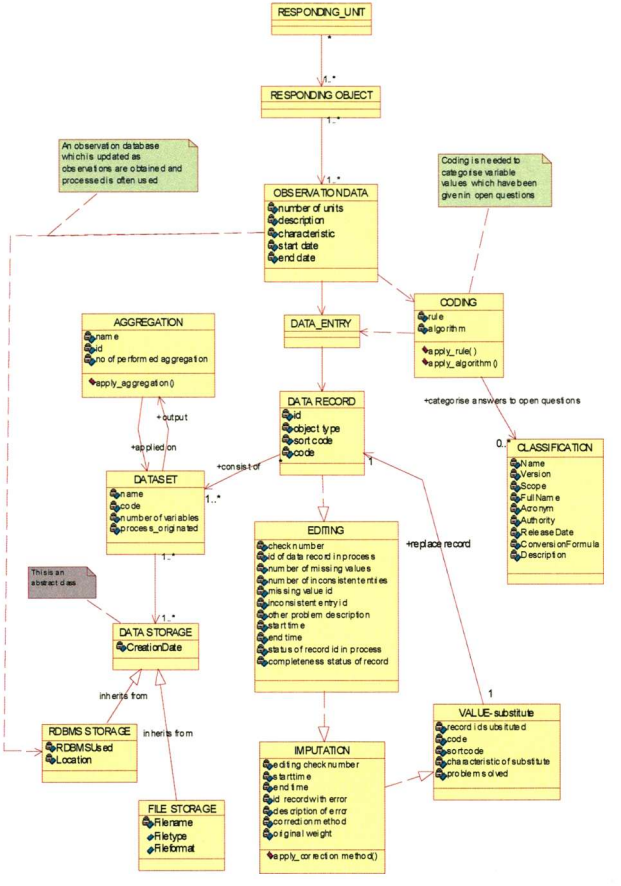
Σε αυτό το διάγραμμα, τα ΔΕΔΟΜΕΝΑ ΠΑΡΑΤΗΡΗΣΗΣ (OBSERVATION DATA) είτε εισάγονται στο ΣΠΣ του φορέα είτε, εάν έχουν υπάρχουν ανοικτές ερωτήσεις (open questions), οι αντίστοιχες απαντήσεις πρέπει να ΚΩΔΙΚΟΠΟΙΗΘΟΥΝ (CODING) σύμφωνα με τους προκαθορισμένους ΚΑΝΟΝΕΣ (RULES) και ΑΛΓΟΡΙΘΜΟΥΣ (ALGORITHMS), σε διάφορες προκαθορισμένες κατηγορίες, π.χ. τις κατηγορίες που καθορίζονται από μια τυποποιημένη ΤΑΞΙΝΟΜΗΣΗ (CLASSIFICATION). Φυσικά, κατά τη διάρκεια των διαφορετικών σταδίων συλλογής, τα ΔΕΔΟΜΕΝΑ ΠΑΡΑΤΗΡΗΣΗΣ πρέπει να αποθηκευτούν. Επομένως, μια βάση δεδομένων διοικούμενη από ένα σύστημα διαχείρισης βάσεων δεδομένων πρέπει να εξεταστεί, και, ως εκ τούτου, τα ΣΤΟΙΧΕΙΑ ΠΑΡΑΤΗΡΗΣΗΣ συσχετίζονται με ένα τέτοιο σχεσιακό σύστημα διαχείρισης RDBMS (Relation Data Management System). Η μεταφορά των στοιχείων από το έντυπο ερωτηματολόγιο στον υπολογιστή στοιχεία αναφέρεται ως ΕΙΣΑΓΩΓΗ ΔΕΔΟΜΕΝΩΝ (DATA ENTRY) Τα προκύπτοντα ΑΡΧΕΙΑ ΔΕΔΟΜΕΝΩΝ (DATA RECORDS) ελέγχονται για την πληρότητα και τα πιθανά λάθη.

Κατόπιν, η ανίχνευση και η διόρθωση των λαθών των απαντήσεων του ερωτηματολογίου είναι ένα κρίσιμο στάδιο, δεδομένου ότι αυτό είναι πιθανώς η κύρια πηγή λαθών μέτρησης. Το DATA EDITING είναι η εφαρμογή των ελέγχων που προσδιορίζουν τις ελλείπουσες, άκυρες ή ασυμβίβαστες καταχωρήσεις.

Κάθε ΑΡΧΕΙΟ ΔΕΔΟΜΕΝΩΝ ελέγχεται και τα προβλήματα εμφανίστηκαν από τα προαναφερθέντα λάθη προσδιορίζονται και διορθώνονται, εφόσον ενδείκνυται, στη διαδικασία ΚΩΔΙΚΟΠΟΙΗΣΗΣ (CODING).

Σε αυτό το στάδιο, μερικές από τις απαντήσεις ή τα ελλείποντα αρχεία δεδομένων αντικαθίστανται από τις τιμές υποκατάστατων (VALUE SUBSTITUTE) που αντικαθιστούν τις προβληματικές τιμές. Πρέπει επίσης να αναφερθεί ότι τα λάθη μπορούν επίσης να εμφανιστούν στο στάδιο ΕΙΣΑΓΩΓΗΣ ΔΕΔΟΜΕΝΩΝ καθώς και κατά τη διάρκεια της ΚΩΔΙΚΟΠΟΙΗΣΗΣ.

Τέλος, ένα σύνολο δεδομένων (DATASET) δημιουργείται κάθε φορά που πραγματοποιούνται αλλαγές ή προστίθεται πληροφορία. Μια διαδικασία συνάθροισης (AGGREGATION) εφαρμόζεται έπειτα στο τελικό ΣΥΝΟΛΟ ΔΕΔΟΜΕΝΩΝ (που είναι ταξινομημένο σε πίνακες) προκειμένου να παραχθούν οι ΔΕΙΚΤΕΣ (βλ. παράγραφο 4.2.1 για την περιγραφή αυτής της διαδικασίας).



Διάγραμμα 12 : Διαδικασία επεξεργασίας δεδομένων

4.2.3 Τμήμα μοντέλου για τη διαδικασία διάχυσης αποτελεσμάτων

Το Διάγραμμα 13 απεικονίζει τη διαδικασία διάχυσης αποτελεσμάτων του υπό εξέταση μοντέλου μεταδεδομένων.

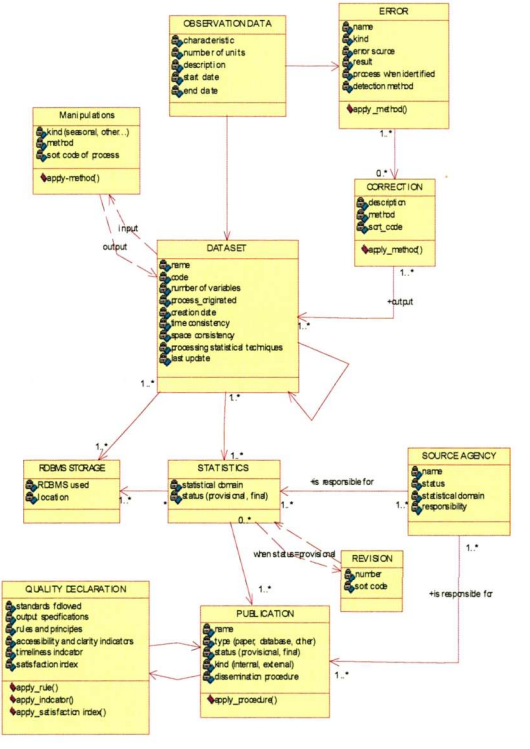
Μετά από την εφαρμογή της σειράς διαδικασιών που διευκρινίζονται στο διάγραμμα 12, διαμορφώνεται το ΣΥΝΟΛΟ ΔΕΔΟΜΕΝΩΝ (DATASET). Στη συνέχεια, διάφοροι ΧΕΙΡΙΣΜΟΙ (MANIPULATIONS) εκτελούνται προκειμένου να μετασχηματιστούν τα ΣΥΝΟΛΑ ΔΕΔΟΜΕΝΩΝ στη ζητούμενη μορφή.

Η ειδική λειτουργία της συνάθροισης (aggregation) για την ανάπτυξη των δεικτών έχει εξηγηθεί στο διάγραμμα 12. Άλλοι ΧΕΙΡΙΣΜΟΙ αφορούν seasonal adjustments,

weights, ή άλλων διαδικασιών στο σύνολο δεδομένων για να παρουσιάσουμε τα στατιστικά αποτελέσματα κατά τα ζητούμενα κριτήρια.

Επιπλέον, τα ΛΑΘΗ (ERRORS) που ανιχνεύονται σύμφωνα με την ΠΗΓΗ (SOURCE) τους, ΤΟ ΕΙΔΟΣ (TYPE), και το αποτέλεσμά τους εξετάζονται επίσης μαζί με το ΣΤΑΔΙΟ ΔΙΑΔΙΚΑΣΙΑΣ (PROCESS STAGE) της ανίχνευσής τους και της ΔΙΟΡΘΩΣΗΣ τους (CORRECTION).

Τα ΣΥΝΟΛΑ ΔΕΔΟΜΕΝΩΝ καθώς επίσης και των τελικών (ή προσωρινών) ΣΤΑΤΙΣΤΙΚΩΝ ΑΠΟΤΕΛΕΣΜΑΤΩΝ (STATISTICS) αποθηκεύονται και ρυθμίζονται από RDBMSs. Η αποθήκευση εκτελείται από το ίδιο SOURCE AGENCY ή ένα εναλλακτικό από αυτό που είναι αρμόδιο για τη ΔΗΜΟΣΙΕΥΣΗ (PUBLICATION) των στατιστικών αποτελεσμάτων. Η ΔΗΜΟΣΙΕΥΣΗ μπορεί να είναι είτε σε μορφή εγγράφου είτε υπό μορφή βάσης δεδομένων.



Διάγραμμα 13: Διαδικασία διάχυσης αποτελεσμάτων

Η ποιότητα αυτών των δημοσιευμένων αποτελεσμάτων εξετάζεται επίσης (QUALITY DECLARATION) από την άποψη της επικαιρότητας, της δυνατότητας πρόσβασης και της σαφήνειας της δημοσίευσης, λαμβάνοντας υπόψη την ικανοποίηση χρηστών.

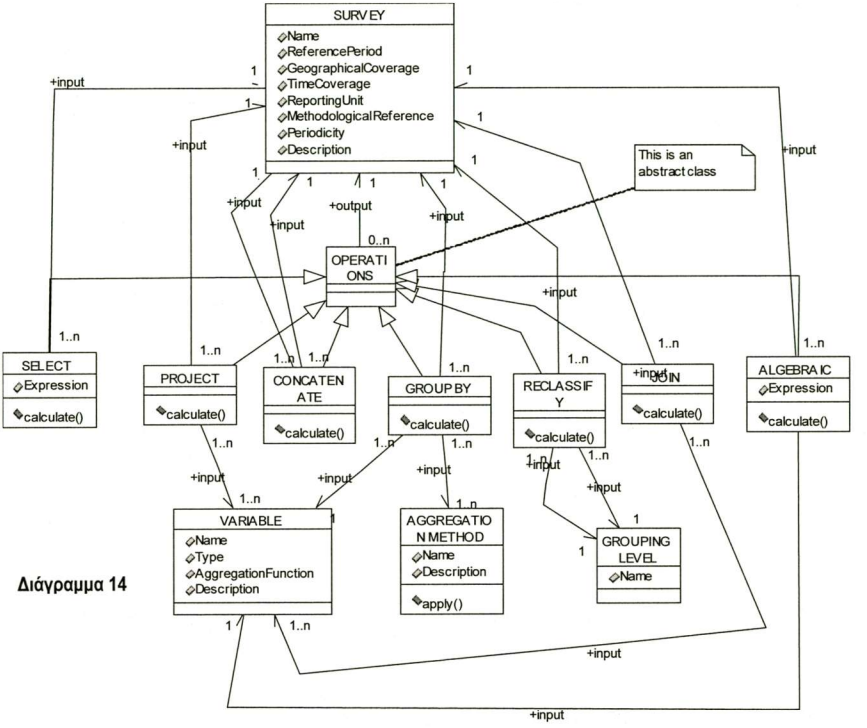
4.3 Μετασχηματισμοί/διαδικασίες (transformations/operations)

Στη διατριβή χρησιμοποιούμε τους όρους "μετασχηματισμός (transformation)" ή/και "διαδικασία (operation)" για να χαρακτηρίσουμε ένα τύπο χειρισμού δεδομένων και μεταδεδομένων. Δηλαδή ένας μετασχηματισμός είναι πραγματικά ένα αλγόριθμος που καθορίζεται πέρα από την κοινή περιοχή των στοιχείων και των μεταδεδομένων. Οι μετασχηματισμοί αναμένονται για να είναι ένα σημαντικό χαρακτηριστικό γνώρισμα της προσεχούς παραγωγής των συστημάτων στατιστικών πληροφοριών (Papageorgiou *et al.*, 2000a, 2000b). Μια εφαρμογή που εξοπλίζεται με τους μετασχηματισμούς μπορεί να εμποδίσει τις πιθανά εσφαλμένες διαδικασίες χρηστών. Με τη χρησιμοποίηση σχετικών μεταδεδομένων, το σύστημα μπορεί να συναγάγει εάν ένας χειρισμός των στοιχείων ισχύει ή όχι. Παραδείγματος χάριν, η συγχώνευση στοιχείων που συλλέγονται κάτω από τις διαφορετικές ονοματολογίες δεν είναι ένας έγκυρος χειρισμός, γιατί πρέπει πρώτα να τα εναρμονίσουμε.

Σημειώνουμε ότι, εάν ένα σύστημα δεν χρησιμοποιεί τους μετασχηματισμούς δεν θα είναι σε θέση να διευκρινίσει την έννοια των μεταδεδομένων που ακολουθούν ένα χειρισμό και επομένως, δεν θα προστατεύσει το χρήστη από τα λάθη στους επόμενους χειρισμούς. Μια πρόσθετη πιθανή χρήση των μετασχηματισμών είναι στην καθοδήγηση του χρήστη κατά τη στατιστική επεξεργασία. Σε αυτήν την περίπτωση, ο χρήστης περιγράφει απλά τον πίνακα που χρειάζεται χρησιμοποιώντας μεταδεδομένα και το σύστημα υπολογίζει αυτόματα το ζητούμενο πίνακα από τα διαθέσιμα στοιχεία. Αυτό μπορεί να έχει πολλές εφαρμογές, όπως π.χ. στη δημιουργία των σύνθετων εκθέσεων αποτελεσμάτων μέσω του Internet.

Ορισμοί και ιδιότητες των μετασχηματισμών του προτεινόμενου μοντέλου

Παρουσιάζεται ένα λογικό διάγραμμα (διάγραμμα 14) των μετασχηματισμών/ διαδικασιών όπως αυτοί επιδρούν στις διάφορες κλάσεις του μοντέλου. Εκφράζονται στην 'αφηρημένη' εκδοχή τους με την κλάση Operators και διακρίνονται στις επτά υποκατηγορίες ανάλογα με την κλάση του μοντέλου όπου εφαρμόζονται.



Διάγραμμα 14

1. Μετασχηματισμός Επιλογής (Selection)

Ο μετασχηματισμός επιλογής *(σcondition)* σχετίζεται με αυτόν της γραμμικής άλγεβρας. Το αποτέλεσμα της εφαρμογής αυτού του μετασχηματισμού είναι ένας νέος πίνακας που διατηρεί μόνο ένα υποσύνολο των αρχικών δεδομένων, ικανοποιώντας το κριτήριο επιλογής. Από τον τρόπο που καθορίσαμε το μετασχηματισμό επιλογής σε έναν πίνακα T εύκολα καταλήγουμε ότι:

$$\sigma condition1(\sigma condition2(T)) = \sigma condition2(\sigma condition1(T)) = \sigma condition1 \wedge condition2(T)$$

Είναι επίσης απαραίτητο να υπολογιστούν εκ νέου όλοι οι αλλαγμένοι δείκτες (Indicators) του πίνακα με βάση το νέο σύνολο μεταβλητών. Εντούτοις, αυτό είναι εφικτό μόνο εάν τα αρχικά μικροδεδομένα είναι διαθέσιμα ή εάν η λειτουργία συνάθροισης (aggregation) των αλλαγμένων δεικτών επιτρέπει τον εκ νέου υπολογισμό των νέων τιμών από τους υπάρχοντες δείκτες (πχ. MAX(), MIN(), SUM(), αλλά όχι AVERAGE())[3].

---

[3] Πρέπει να σημειωθεί ότι αυτός ο μετασχηματισμός είναι πάντα μεταθετικός υπό τον όρο ότι τα μικροδεδομένα είναι διαθέσιμα. Εάν αυτό δεν ισχύει, έπειτα οι αλλαγμένοι δείκτες θα πρέπει να υπολογιστούν εκ νέου χρησιμοποιώντας μόνο τους διαθέσιμους, ήδη υπάρχοντες δείκτες. Σε αυτήν την

Η έκφραση όταν εφαρμόζεται ο μετασχηματισμός επιλογής σε ένα survey είναι:

```
SURVEY2=SELECT.CALCULATE(SURVEY1,CONDITION)
```

Παραδείγματος χάριν, εάν ένας χρήστης θέλει να δει τα δεδομένα της ανεργίας (unemployment), τα οποία αναφέρονται στις γυναίκες (female), η ανωτέρω έκφραση θα ήταν:

```
Unemployment1=SELECT.CALCULATE(Unemployment, sex="female")
```

Όπου οι μεταβλητές sex και unemployment είναι μεταβλητές του survey και unemployment1 είναι το αποτέλεσμα της εφαρμογής του μετασχηματισμού επιλογής.

2. Προβολής (Projection)

Ο μετασχηματισμός projection (pvariables), όπως ο μετασχηματισμός επιλογής, πολύ μοιάζει με αυτός από τη γραμμική άλγεβρα. Το αποτέλεσμα της εφαρμογής αυτού του μετασχηματισμού είναι ένας νέος πίνακας που περιλαμβάνει μόνο ένα υποσύνολο των μεταβλητών και των δεικτών των αρχικών δεδομένων.

Εάν το σύνολο μεταβλητών ομαδοποίησης (Varsgroup) οποιουδήποτε δείκτη του πίνακα *T* δεν είναι ένα υποσύνολο Vars', κατόπιν αυτός ο δείκτης θα πρέπει να υπολογιστεί εκ νέου, χρησιμοποιώντας ένα νέο Varsgroup' = Varsgroup ∩Vars'.

Από τον τρόπο που καθορίσαμε το μετασχηματισμό projection σε έναν πίνακα *T*, εύκολα καταλήγουμε ότι:

$$p_{Vars1}(p_{Vars2}(T)) = p_{Vars2}(p_{Vars1}(T)) = p_{Vars1 \cap Vars2}(T).$$

Η έκφραση όταν εφαρμόζεται ο μετασχηματισμός projection σε ένα survey είναι:

```
SURVEY2=PROJECT.CALCULATE(SURVEY1,VARIABLE1,VARIABLE2,...)
```

Παραδείγματος χάριν, εάν ένας χρήστης θέλει να δει δύο από τις τρεις μεταβλητές (ηλικία, φύλο, υπηκοότητα) σχετικά με την ανεργία, η ανωτέρω έκφραση θα ήταν:

```
Unemployment2 =PROJECT.CALCULATE(Unemployment, Age, Nationality)
```

Όπου οι μεταβλητές sex, unemployment και nationality είναι μεταβλητές του survey και unemployment2 είναι το αποτέλεσμα της εφαρμογής του μετασχηματισμού projection.

3. Επισύναψης (Concatenation)

Το αποτέλεσμα της εφαρμογής του μετασχηματισμού concatenation (τ) σε δύο πίνακες (T1, T2) είναι ένας νέος πίνακας (T3) συνδυάζοντας τα στοιχεία των αρχικών πινάκων. Από τον τρόπο καθορίσαμε το μετασχηματισμό αλληλουχίας καταλήγουμε στο συμπέρασμα ότι:

$$\tau(T1, T2) = \tau(T2, T1) \quad \text{και}$$

περίπτωση, ο μετασχηματισμός, ανάλογα με τη χρησιμοποιημένη μέθοδο συνάθροισης, μπορεί να μην είναι μεταθετικός.

$$\tau(\tau(T1, T2), T3) = \tau(T1, \tau(T2, T3))$$

Αυτός ο μετασχηματισμός έχει τις ακόλουθες προϋποθέσεις προκειμένου να ισχύσει:

- Και οι δύο πίνακες T1, T2 πρέπει να έχουν τα "ίδια" σύνολα μεταβλητών που μετρούνται με την ίδια μονάδα μέτρησης διαφορετικά, ο προκύπτων πίνακας θα έχει πολλές μηδενικές/άγνωστες τιμές.
- Και οι δύο πίνακες T1, T2 πρέπει να αναφερθούν στο ίδιο είδος μονάδων, δηλ. δεν μπορούμε να συνδέσουμε δύο πίνακες με πληροφορίες για διαφορετικές μονάδες μέτρησης όπως π.χ. άτομα και αγαθά
- Η τομή των δύο στατιστικών πληθυσμών (statistical populations) Sp1, Sp2 πρέπει να είναι το κενό, ώστε να διασφαλιστεί ότι ο νέος πίνακας δεν περιέχει διπλές εγγραφές που μπορεί να οδηγήσουν σε εσφαλμένα αποτελέσματα.
  Δηλαδή $Sp1 \cap Sp2 = \varnothing$

Η έκφραση που αντιπροσωπεύει τη χρήση του μετασχηματισμού:

SURVEY3=CONCATENATE.CALCULATE(SURVEY1,SURVEY2)

Παραδείγματος χάριν, εάν ένας χρήστης θέλει να δει τα στοιχεία δύο surveys με την επισύναψη των δύο σχετικών πινάκων, π.χ. την ανεργία στην Ελλάδα και ανεργία στη Γαλλία, υπό τον όρο ότι οι προϋποθέσεις ισχύουν, η ανωτέρω έκφραση θα ήταν:

```
Unemployment3=CONCATENATE.CALCULATE (Unemployment in
Greece, Unemployment in France)
```

Όπου Unemployment3 είναι το αποτέλεσμα της εφαρμογής του Concatenation.

4. Μετασχηματισμός Ομαδοποίησης (GroupBy)

Ο μετασχηματισμός GroupBy χρησιμοποιείται για τη δημιουργία των νέων δεικτών από τα υπάρχοντα δεδομένα. Ο νέος δείκτης υπολογίζεται μετά από να ομαδοποιήσει τις τιμές ενός πίνακα σύμφωνα με τις διακριτές τιμές μιας ή περισσότερων υπαρχουσών μεταβλητών που καλούνται μεταβλητές ομαδοποίησης (Varsgroup). Ο παραγόμενος δείκτης υπολογίζεται με την εφαρμογή μιας λειτουργίας συνάθροισης (AgrF) επί των τιμών ενός υποσυνόλου (Varsind) των μεταβλητών του πίνακα.

Η μόνη προϋπόθεση της εφαρμογής ενός μετασχηματισμού GroupBy είναι ότι εάν ο πίνακας έχει ήδη έναν δείκτη, τότε οι μεταβλητές ομαδοποίησης του μετασχηματισμού πρέπει να είναι οι ίδιες με τις μεταβλητές ομαδοποίησης των υπαρχόντων δεικτών.

Εκφράζεται ως:

SURVEY2=GROUPBY.CALCULATE(SURVEY1,VARIABLE1,VARIABLE2,
AGGR.METHOD1,...)

Παραδείγματος χάριν, εάν θέλουμε να αθροίσουμε τα δεδομένα της ανεργίας, σύμφωνα με τη μεταβλητή υπηκοότητα, και να μετρήσει τον αριθμό των ατόμων κάθε υπηκοότητας, η ανωτέρω έκφραση θα ήταν:

```
Unemployment4=GROUPBY.CALCULATE(Unemployment, Nationality,
                    Age, Count(Person))
```

όπου Unemployment4 είναι ο παραγόμενος πίνακας.

5. Επαναταξινόμησης (Reclassification)

Ο μετασχηματισμός επαναταξινόμησης μετατρέπει τις τιμές μιας μεταβλητής ή ενός δείκτη από ένα επίπεδο ομαδοποίησης (Grouping Level) σε ένα διαφορετικό. Ο μετασχηματισμός επαναταξινόμησης δεν θέτει κανέναν περιορισμό στο Grouping Level-στόχο. Επομένως, τα αποτελέσματα μιας επαναταξινόμησης μπορούν να οδηγήσουν σε πίνακες με missing values, ή με τις ανακριβείς τιμές, ειδικά όταν η το προϋπάρχον Grouping Level δεν είναι πλήρως μετατρέψιμο στο νέο (πχ, μετατροπή της ταξινόμησης NABS στην ταξινόμηση του OECD).

Μια χαρακτηριστική περίπτωση είναι όταν και τα δύο Grouping Levels ανήκουν στην ίδια ταξινόμηση, όπως όταν μετατρέπουμε από το επίπεδο υποκεφαλαίων NABS ($2^{ο}$ επίπεδο ομαδοποίησης ψηφίων) στο επίπεδο κεφαλαίων ($1^{ο}$ επίπεδο ομαδοποίησης ψηφίων). Σε αυτήν την περίπτωση, η επαναταξινόμηση δεν έχει επιπτώσεις στην ποιότητα των στοιχείων.

Οι νέες τιμές του προκύπτοντος πίνακα υπολογίζονται χρησιμοποιώντας τις πληροφορίες που περιλαμβάνονται στις περιπτώσεις της κλάσης Equals, η οποία "συνδέει" δύο (ή περισσότερες) μονάδες μέτρησης οι οποίες ανήκουν σε διαφορετικά Grouping Levels.

Η έκφραση είναι:

SURVEY2=RECLASSIFY.CALCULATE(SURVEY1,VARIABLE,
GROUPING_LEVEL1,GROUPING_LEVEL2)

Εάν για παράδειγμα ένας χρήστης θέλει να μετατρέψει τις τιμές της μεταβλητής 'Ηλικία' από το Grouping Level A, το οποίο αποτελείται από τις τιμές 15-24, 25-49, 50-64, στο Grouping Level B, που αποτελείται από τις τιμές 15-49, 50-64, τότε θα είχαμε:

```
Unemployment5=RECLASSIFY.CALCULATE(Unemployment, Age, A, B)
```

όπου Unemployment5 είναι το αποτέλεσμα της επαναταξινόμησης. Προφανώς, προκειμένου να εκπληρωθεί αυτή η μετατροπή, η σχέση μεταξύ των τιμών των δύο Grouping Levels πρέπει να καθοριστεί. Δηλαδή κάπως πρέπει να διευκρινιστεί ότι οι τιμές 15-24 και 25-49 του A είναι ίσες με την αξία 15-49 του B. Επίσης, πρέπει να καθοριστεί ότι η αξία 50-64 είναι η ίδια και στα δύο επίπεδα ομαδοποίησης.

Παρατήρηση 4.3.1:

Ο μετασχηματισμός επαναταξινόμησης από ένα Grouping Level (GL) σε άλλο (GL1, GL2) μπορεί να έχει επιπτώσεις στην ποιότητα των παραχθέντων στατιστικών δεδομένων για δύο λόγους. Ο πρώτος συσχετίζεται με το γεγονός ότι η απεικόνιση μεταξύ GL1 και GL2 μπορεί να είναι ελλιπής. Σε αυτήν την περίπτωση, ο προκύπτων πίνακας θα περιέχει μηδενικές/άγνωστες τιμές. Ο δεύτερος λόγος συσχετίζεται με το γεγονός ότι ο αρχικός πίνακας μπορεί να έχει μερικές missing values.

6. Μετασχηματισμός ένωσης (Join)

Ο μετασχηματισμός αυτός έχει πολλά κοινά στοιχεία με της γραμμικής άλγεβρας. Εφαρμόζεται σε δύο πίνακες που έχουν μια ή περισσότερες κοινές μεταβλητές (join variables). Το αποτέλεσμα είναι ένας νέος πίνακας που έχει όλες τις μεταβλητές και τους δείκτες και των δύο πινάκων (προφανώς, οι κοινές μεταβλητές και τους δείκτες συμπεριλαμβάνονται μιά φορά).

Δεδομένου ότι το αποτέλεσμα του μετασχηματισμού είναι ένας νέος πίνακας, όλοι οι δείκτες αυτού του πίνακα πρέπει να έχουν τις ίδιες μεταβλητές ομαδοποίησης (Varsgroup). Συνεπώς, ο μετασχηματισμός join έχει τις ακόλουθες προϋποθέσεις:

- Οι δύο πίνακες πρέπει να έχουν τους ίδιους στατιστικούς πληθυσμούς και δείγμα
- Οι δείκτες των 'ενωμένων' πινάκων πρέπει να έχουν τις ίδιες μεταβλητές ομαδοποίησης (Varsgroup).
- Εάν οι μεταβλητές ομαδοποίησης Varsgroup δεν είναι το κενό σύνολο (δηλ. οι ενωμένοι πίνακες έχουν δείκτες) έπειτα οι 'ενωμένες' μεταβλητές πρέπει να είναι ίσες με Varsgroup.

Η έκφραση είναι:

SURVEY3=JOIN.CALCULATE(SURVEY1,SURVEY2,COMMON_VARIABLE1,..)

Παραδείγματος χάριν, εάν ένας χρήστης θέλει να ενώσει τα στοιχεία της ανεργίας με εκείνοι του επιπέδου εκπαίδευσης των ανέργων, σύμφωνα με την κοινή μεταβλητή 'επάγγελμα', θα έχουμε:

```
Unemployment6=JOIN.CALCULATE(Unemployment, Level of
                Education, Occupation)
```

όπου Unemployment6 ο νέος πίνακας.

7. Αλγεβρικοί μετασχηματισμοί (algebraic)

Η τελευταία κατηγορία μετασχηματισμών είναι οι αλγεβρικοί μετασχηματισμοί. Είναι γενικοί μετασχηματισμοί που χρησιμοποιούνται για όλους τους μαθηματικούς μετασχηματισμούς (π.χ. προσθήκες, πολλαπλασιασμοί, κλπ) που εφαρμόζονται συχνά

σε έναν πίνακα. Ο αλγεβρικός μετασχηματισμός είναι ο μοναδικός ο οποίος απαιτεί την ταυτόχρονη διευκρίνιση μεταδεδομένων τεκμηρίωσης καθώς και σημασιολογικά μεταδεδομένα από το χρήστη. Ο λόγος είναι ότι δεν υπάρχει κανένας αυτοματοποιημένος τρόπος να γίνει κατανοητή από τα ΣΠΣ η σημασιολογική έννοια ενός τέτοιου μετασχηματισμού.

Η έκφραση είναι:

SURVEY2=ALGEBRAIC.CALCULATE(SURVEY1,VARIABLE,EXPRESSION)

Παραδείγματος χάριν, εάν ένας χρήστης θέλει να πολλαπλασιάσει τις τιμές της μεταβλητής A της ανεργίας επί 2, θα έχει:

```
Unemployment 7 =ALGEBRAIC.CALCULATE(Unemployment, A, (x2))
```

Όπου Unemployment 7 είναι το αποτέλεσμα μετά τον πολλαπλασιασμό.

Παρατήρηση 4.3.2:

Μελετώντας προσεκτικά τους μετασχηματισμούς που καθορίστηκαν, εύκολα παρατηρούμε ότι όλοι έχουν την ιδιότητα της κλειστότητας. Δηλαδή το αποτέλεσμα της εφαρμογής ενός μετασχηματισμού στα δεδομένα του survey είναι ένα νέο survey. Μια τέτοια ιδιότητα βεβαιώνει ότι μπορούμε να εφαρμόσουμε τις διαδικασίες σε ένα survey κατ' επανάληψη, κάτι ιδιαίτερα χρήσιμο αφού μας δίνει την ευκαιρία να παράγουμε νέους δείκτες.

4.4 Εφαρμογή

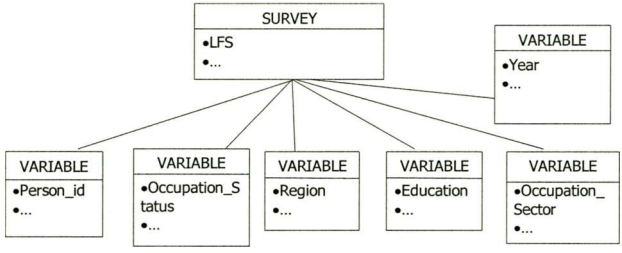
Η εφαρμογή αυτή έχει δύο στόχους: να δείξει πώς το προτεινόμενο μοντέλο μπορεί να χρησιμοποιηθεί για την παραγωγή νέων οικονομικών δεικτών βάσει μίας δειγματοληπτικής έρευνας και επίσης να εξετάσει τη συμβολή των προτεινόμενων μετασχηματισμών σε αυτή τη δραστηριότητα.

Έστω ότι θέλουμε να δημιουργήσουμε το δείκτη

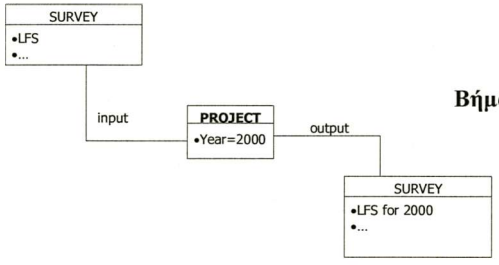
R = ποσοστό ανεργίας (Unemployment rate)

ο οποίος ορίζεται ως **Αριθμός ανέργων προς τον αριθμό συνολικού εργατικού δυναμικού** (Number of unemployed persons divided by the Number of total Labour Force).

Έστω η έρευνα για το εργατικό δυναμικό (Labor Force Survey –LFS) και τις μεταβλητές Άτομο (Person_id), Επαγγελματική κατάσταση (Occupation_Status), Περιοχή (Region), Εκπαίδευση (Education), Τομέας απασχόλησης (Occupation_sector) και Έτος (Year), όπως φαίνεται διαγραμματικά:

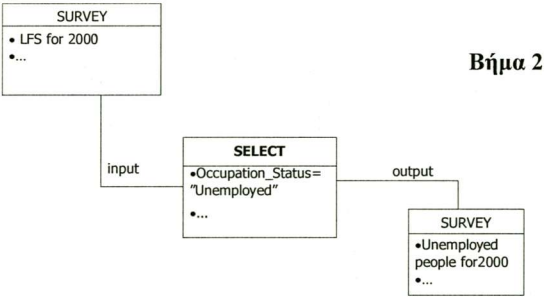


Βήμα 1: Επιλέγουμε τις μεταβλητές που μας ενδιαφέρουν για την εξαγωγή του δείκτη. Έστω για παράδειγμα ότι μας ενδιαφέρει η χρονιά 2000, οπότε εφαρμόζουμε τον μετασχηματισμό *Προβολής (Projection)* στη μεταβλητή Year πάνω στο αρχικό survey οπότε προκύπτει LFS for 2000.



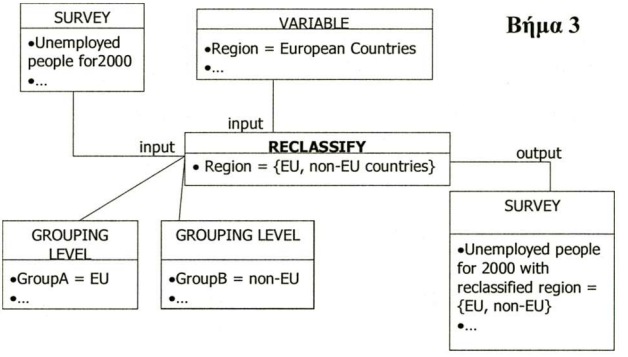
Βήμα 1

Βήμα 2: Κατόπιν μας ενδιαφέρει να απομονώσουμε από τη μεταβλητή Επαγγελματική κατάσταση (Occupation_Status) την τιμή «άνεργος». Έτσι, επιλέγουμε εφαρμόζοντας τον μετασχηματισμό *Επιλογής (selection)* την τιμή Occupation_Status = Unemployed πάνω στο αρχικό survey LFS for 2000, οπότε προκύπτει ένα νέο survey για τους ανέργους του 2000 (Unemployed people for 2000)



Βήμα 2

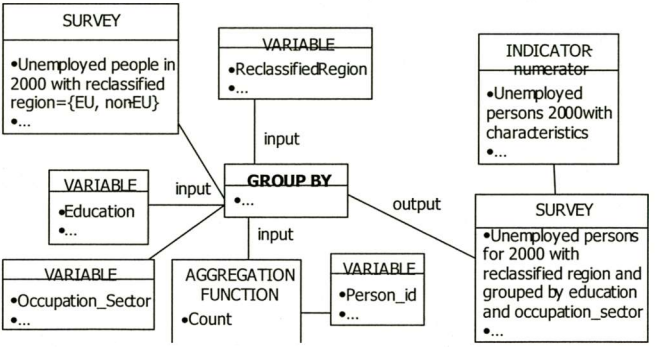
Βήμα 3: Εάν οι τιμές μιας μεταβλητής δεν βρίσκονται στο επιθυμητό επίπεδο ομαδοποίησης εφαρμόζουμε μία *Επαναταξινόμηση (Reclassify)* στη μεταβλητή αυτή. Για παράδειγμα αν η μεταβλητή Περιοχή (Region) = {Ευρωπαϊκές χώρες} και εμείς θέλουμε επιμέρους διαμέριση σε EU και non-EU, τότε στο survey Unemployed people for 2000 εφαρμόζουμε την επαναταξινόμηση σε πιο λεπτομερές επίπεδο ομαδοποίησης και τότε το νέο survey θα εξετάζεται ως προς region={EU, non-EU}



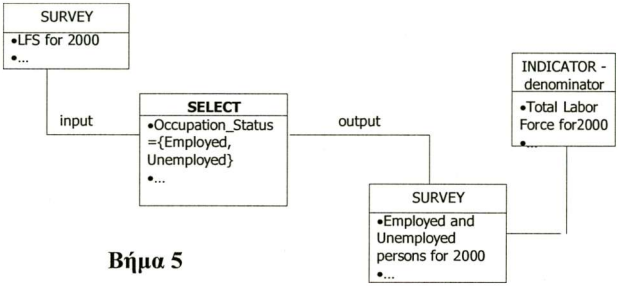
Βήμα 3

Βήμα 4: Εφαρμόζουμε απαραίτητες *ομαδοποιήσεις (group by)* και *αθροιστικές συναρτήσεις (aggregation functions)* για να βρούμε τον αριθμητή του δείκτη R (indicator numerator).

Βήμα 4

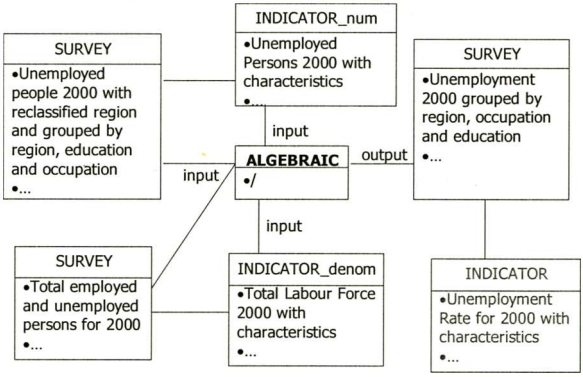


Βήμα 5: Για να βρούμε τον παρονομαστή του δείκτη (indicator denominator) R, εφαρμόζουμε τον μετασχηματισμό *Επιλογής (select)* στη μεταβλητή "Occupation_status" και από αυτή επιλέγουμε όλες τιμές {employed, unemployed}, οπότε προκύπτει ο παρονομαστής ως Συνολικό εργατικό δυναμικό το 2000 (Total Labor Force for 2000).



Βήμα 5

Βήμα 6: Τέλος, εφαρμόζουμε τον Αλγεβρικό μετασχηματισμό (algebraic): διαίρεση (/) ανάμεσα στον αριθμητή του βήματος 4 και του παρονομαστή του βήματος 5 και προκύπτει ο ζητούμενος δείκτης (indicator) R



Βήμα 6

ΚΕΦΑΛΑΙΟ 5

ΕΝΣΩΜΑΤΩΣΗ ΕΝΟΣ ΜΟΝΤΕΛΟΥ ΜΕΤΑΔΕΔΟΜΕΝΩΝ ΣΕ ΜΙΑ ΒΑΣΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΧΡΗΣΗ ΤΟΥ ΑΠΟ ΤΑ ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ

Οι κανόνες που ακολουθούνται προκειμένου να μετατραπεί το μοντέλο μεταδεδομένων σε ένα σχεσιακό σχήμα βάσεων δεδομένων είναι οι εξής:

i) Κάθε κλάση γίνεται ένας χωριστός πίνακας και οι ιδιότητές της είναι οι τμήματα του πίνακα

ii) Κάθε σχέση μία-προς-πολλά μεταξύ της κλάσης Α και της κλάσης Β 'μεταφράζεται' με την προσθήκη ενός 'foreign key' στην κλάση Β. Ένα παράδειγμα είναι η σχέση μεταξύ της κλάσης «Statistical Population» και της κλάσης «Survey»

iii) Κάθε πολλαπλή σχέση μεταξύ της κατηγορίας Α και της κατηγορίας Β "μεταφράζεται" με την προσθήκη ενός νέου βοηθητικού πίνακα. Ένα παράδειγμα είναι η σχέση μεταξύ της κλάσης «Adjustments» και της κλάσης «Survey».

Μπορούμε κατ' αυτόν τον τρόπο να απεικονίσουμε τις κλάσεις του μοντέλου ως εξής (ορισμένες κλάσεις δίνονται ως παράδειγμα):

SURVEY(id, name, description, referPeriod, geoRef, timeCover, reportUnit, methodRef, periodicity, statPopId, sampleName, quality_id, collection_id)

statPopId: foreign key to STATISTICAL_POPULATION.id

quality_id: foreign key to DATA_QUALITY.id

collection_id: foreign key to COLLECTION_INFO.id

STATISTICAL_POPULATION(id, name, description, kindOfUnit)

CONDITION(name, expression, genStatPop, specStatPop)

genStatPop: foreign key to STATISTICAL_POPULATION.id

specStatPop: foreign key to STATISTICAL_POPULATION.id

SAMPLING_METHOD(id, name, description, initStatPop, sample, master_list)

initStatPop: foreign key to STATISTICAL_POPULATION.id

VARIABLE(id, name, type, description, gr_level)

gr_level: foreign key to GR_LEVEL.id

CHARACTERISTIC(id, name, description)

SUR_VARS_CHARS(survey_id, var_id, charact_id) (auxiliary table)

INDICATOR(id, name, description, scope, aggr_function, gross_id)

gross_id: foreign key to GROSSUP_METHOD.id

CLASSIFICATION(id, name, version, pr_version, scope, description, fullname, acronym, authority, releaseDate, publication)

GR_LEVEL(id, name, classification)

classification: foreign key to CLASSIFICATION.id

MEASUREMENT_UNIT(id, name, type, start, end, step, value, kind, gr_level)

gr_level: foreign key to GR_LEVEL.id

EQUALS(meas_unit1, expression, meas_unit2)

meas_unit1, meas_unit2: foreign keys to MEASURE_UNIT.id

DATA_STORAGE(id, type, RDBMSused, location, filename, filetype, fileformat, location, path)

DBMS_DB(RDBMS, database, table, column) (auxiliary table)

column: foreign key to VARIABLE.id

DATA_QUALITY(id, timeliness, prelimEstimates, MissingData, revPolicy, corrobEvid, estimMethod)

ERRORS(id, kind, description, correction_descr)

QUAL_ERROR(qual_id, error_id) (auxiliary table)

qual_id: foreign key to DATA_QUALITY.id

error_id: foreign key to ERRORS.id

ADJ_SURVEY(adj_id, survey_id) (auxiliary table)

adj_id: foreign key to ADJUSTMENTS.id

survey_id: foreign key to SURVEY.id

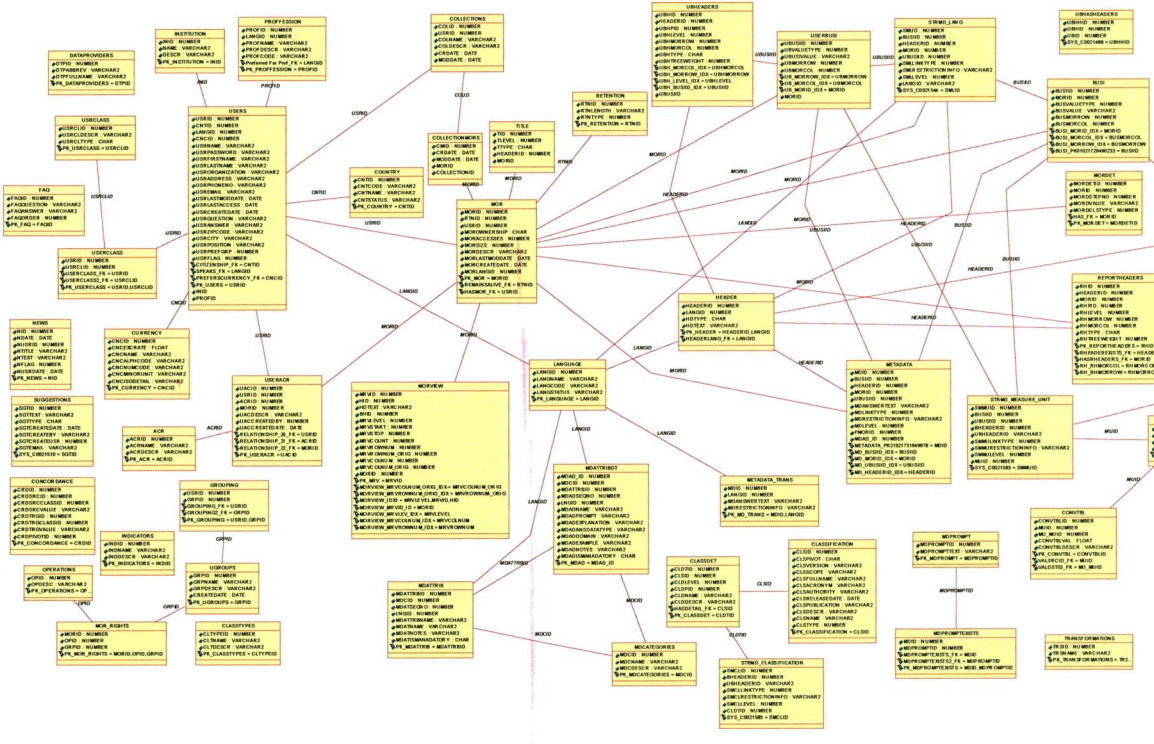
COLLECTION_INFO(id, descr_quest, reportDate, reportMethod, nonRespRate, timeLapse)

ADMIN_SOURCES(id, name, updateProc, access_control, processTech, spaceConsistency, timeConsistency, systemQuality, legalBase, restrictions)

SOURCE_AGENCY(id, name, domain, kind, beginDateAvail, endDateAvail)

kind: takes either the value "for collection" or "for compilation"

Το σχήμα της σχεσιακής βάσης που μπορούμε να δημιουργήσουμε για το μοντέλο μεταδεδομένων που αναπτύχθηκε στο προηγούμενο κεφάλαιο παρουσιάζεται στο διάγραμμα 15 και μπορεί να χρησιμοποιηθεί από κάθε ΣΠΣ.



Διάγραμμα 15

ΚΕΦΑΛΑΙΟ 6

ΕΠΕΚΤΑΣΕΙΣ ΣΤΑΤΙΣΤΙΚΩΝ ΜΟΝΤΕΛΩΝ ΜΕΤΑΔΕΔΟΜΕΝΩΝ ΣΕ ΟΜΑΔΟΠΟΙΗΜΕΝΑ ΔΕΔΟΜΕΝΑ

6.1 Εισαγωγή

Οι κλασσικές μέθοδοι ανάλυσης στατιστικών δεδομένων αναπτύχθηκαν και εφαρμόστηκαν με επιτυχία σε περιπτώσεις κατά τις οποίες τα υπό εξέταση δεδομένα είτε αναφέρονταν σε ένα μόνο άτομο/αντικείμενο, είτε κάθε μία από τις μεταβλητές που τα χαρακτήριζαν αναφερόταν σε μία και μόνη παρατήρηση.

Στην περίπτωση που αναφερόμαστε σε πιο σύνθετους τύπους δεδομένων, τα συμβολικά δεδομένα (symbolic data), τα οποία εξετάζει η μέθοδος ανάλυσης συμβολικών δεδομένων (symbolic data analysis method). Στη μέθοδο αυτή, τα υπάρχοντα δεδομένα ομαδοποιούνται σε «υπερ-δεδομένα», τα «συμβολικά δεδομένα», τα οποία ομαδοποιημένα υπό κοινά χαρακτηριστικά παράγουν τα συμβολικά αντικείμενα (Symbolic Objects (SO)). Η βασική χρησιμότητά τους είναι ότι επιτρέπουν τον χειρισμό μεγάλου όγκου δεδομένων με δομημένο τρόπο.

Αποτέλεσμα αυτής της συγκέντρωσης πληροφορίας καθώς και της ανάγκης επεξεργασίας της, ήταν η επέκταση των κλασσικών μεθόδων ανάλυσης δεδομένων σε μεθόδους ανάλυσης συμβολικών δεδομένων. Ένας κώδικας περιγραφής των μεταβλητών οι οποίες χαρακτηρίζουν τα συμβολικά δεδομένα αναπτύχθηκε στη βιβλιογραφία (βλ. [Billard & Diday, 2003], [Bock & Diday, 2000], [Diday, 1991], [Diday et.al., 1995], [Hebrail, 1996], [Stephan et.al, 1997]) όπου αναλύθηκαν τα είδη και οι ιδιότητες των μεταβλητών, οι σχέσεις μεταξύ των τιμών κάθε μεταβλητής αλλά και μεταξύ των ιδίων των μεταβλητών, ώστε να είναι δυνατή η δημιουργία των συμβολικών δεδομένων. Αρκετά διεθνή συνέδρια (όπως π.χ. OSDA (Conference on Ordinal and Symbolic Data Analysis), KESDA – (Conference on Knowledge Extraction and Symbolic Data Analysis), etc) και επιστημονικά περιοδικά (όπως JSL (Journal of Symbolic Logic), JSDA (Journal of Symbolic Data Analysis), etc) είναι αφιερωμένα στη συμβολική ανάλυση.

Τα συμβολικά δεδομένα δεν χρησιμοποιούνται μόνο για να ομαδοποιούν τα τεράστια σύνολα δεδομένων αλλά ταυτόχρονα οδηγούν σε περισσότερο σύνθετους πίνακες δεδομένων (Symbolic Data Tables (SDT)), διευκολύνοντας τον χειρισμό των μεγάλων αριθμών δεδομένων που συλλέγονται. Για να μπορεί όμως να πραγματοποιηθεί ταυτόχρονα και ο χειρισμός των σχετικών εννοιών, ένα μοντέλο μεταδεδομένων, ειδικά σχεδιασμένο για τα συμβολικά δεδομένα, είναι απαραίτητο.

Το μοντέλο μεταδεδομένων που αναπτύχθηκε στην προηγούμενη ενότητα καλύπτει αυτή την περίπτωση και δημιουργεί, την ίδια στιγμή, τις προϋποθέσεις για την εξέταση πολλαπλών παρατηρήσεων και την ανάλυση των αντίστοιχων δεδομένων. Παρόλα αυτά η κλασσικές μέθοδοι ανάλυσης που ακολουθήθηκαν δεν καλύπτουν πλήρως την ανάγκη επεξεργασίας και ανάλυσης μεγάλου αριθμού πολύπλοκων δεδομένων τα οποία είχαν συλλεχθεί και αποθηκευτεί σε βάσεις δεδομένων

Στατιστικών Υπηρεσιών, Οργανισμών Δημόσιας Διοίκησης και Ερευνητικών Ινστιτούτων [Diday, 1991], [Diday et al., 1995].

Η παρούσα διατριβή επεκτείνει την κλασσική ανάλυση των δεδομένων, όπως αυτή αναπτύχθηκε στις προηγούμενες ενότητες, με τη δημιουργία ενός μοντέλου μεταδεδομένων το οποίο περιέχει την απαραίτητη μεταπληροφορία για τα Symbolic Objects (SO) και τη δημιουργία των αντίστοιχων πινάκων (Symbolic Data Tables (SDT)), με απώτερο στόχο την ακριβή εφαρμογή των στατιστικών μεθόδων συμβολικής ανάλυσης (Symbolic Data Analysis methods).

Σε αυτό το κεφάλαιο αναπτύσσουμε ένα στατιστικό μοντέλο μεταδεδομένων το οποίο λαμβάνει υπόψη τις ιδιομορφίες των συμβολικών δεδομένων σχετικά με τα απαιτούμενα μεταδεδομένα. Το μοντέλο αυτό διατηρεί τα θεμελιώδη στοιχεία μεταδεδομένων του βασικού μοντέλου που αναπτύχθηκε και αναλύθηκε στο προηγούμενο κεφάλαιο και επιπλέον αναπροσαρμόσθηκε για να διατηρεί τη μεταπληροφορία όχι μόνο των κλασσικών (αρχικών) δεδομένων (μεταβλητές ερευνών, στατιστικές μονάδες, πληθυσμός πλαισίων, κλπ) αλλά και τα συμβολικά δεδομένα.

Πιο συγκεκριμένα, το αναμορφωμένο μοντέλο διατηρεί τα μεταδεδομένα για τα κύρια στάδια των διαδικασιών της κλασσικής ανάλυσης στοιχείων, και είναι εμπλουτισμένο με τις συμβολικές διαδικασίες ανάλυσης στοιχείων. Το μοντέλο αυτό περιέχει πληροφορία τόσο για τα κλασσικά (αρχικά) δεδομένα (μεταβλητές, στατιστικό πληθυσμό, άτομα/αντικείμενα, κλπ) όσο και για τα συμβολικά δεδομένα. Περιέχει δηλαδή μεταπληροφορία τόσο για τις κύριες φάσεις της κλασσικής ανάλυσης, όσο και για της συμβολικής, διατηρώντας έτσι την ιστορικότητα των δεδομένων και των διαδικασιών από τις αρχικές μεταβλητές και τα πρωτότυπα άτομα/αντικείμενα έως τη δημιουργία των Symbolic Objects και Symbolic Data Tables, καθώς και την εφαρμογή στατιστικών μεθόδων επί αυτών.

Είναι επίσης σημαντικό ότι το παρόν μοντέλο, διατηρεί την ιστορία της επεξεργασίας δεδομένων, όχι μόνο των κλασσικών δεδομένων όπως το μοντέλο του κεφαλαίου 4, αλλά και, σύμφωνα τώρα με τις νέες απαιτήσεις, την ιστορικότητα των διαδικασιών από την αρχική δημιουργία μεταβλητών μέχρι τη δημιουργία SDT και την εφαρμογή μεθόδων καθώς και την απεικόνιση των τελικών αποτελεσμάτων.

Τα στάδια τα οποία ακολουθήθηκαν μέχρι την υλοποίηση του μοντέλου μεταδεδομένων και περιγράφονται σε αυτή την ενότητα είναι τα εξής:

- Μελετήθηκαν οι σχέσεις που διέπουν τα συμβολικά δεδομένα και τα στοιχεία που πρέπει να κρατούνται καθ' όλη τη διάρκεια της διαδικασίας ώστε να είναι διαθέσιμη η ιστορικότητα των διαδικασιών για τον τελικό χρήστη.
- Το υπάρχον μοντέλο μεταδεδομένων αναμορφώθηκε να για καλύψει τις απαιτήσεις των συμβολικών δεδομένων δίνοντας έμφαση στην επεξεργασία των αναγκαίων σχέσεων προκειμένου να γίνει εφικτός ο χειρισμός και η ανάλυση των συμβολικών δεδομένων μέσα από τα αρχικά δεδομένα, που προέκυψαν από τη μελέτη της βιβλιογραφίας.

- Τέλος, ένα παράδειγμα παρατίθεται χρησιμοποιώντας δεδομένα από τη βιβλιογραφία [Bock & Diday, 2000, (p.58)] για να δειχθεί η χρησιμότητα του μοντέλου αλλά και η δυνατότητα να βελτιωθεί η γραφική αναπαράσταση των συμβολικών δεδομένων μέσα από κατάλληλα πακέτα που αναπτύχθηκαν.

6.2 Ορισμοί και σπουδαιότητα των συμβολικών δεδομένων και διαφορές με τα κλασσικά δεδομένα

Διεθνείς οργανισμοί, κυβερνήσεις και γενικότερα χρήστες και επεξεργαστές δεδομένων συλλέγουν και αποθηκεύουν προς επεξεργασία τεράστιο αριθμό στατιστικών δεδομένων. Η ανάγκη κατηγοριοποίησης και ομογενοποίησης του συνόλου αυτών των δεδομένων σε νέες «στατιστικές μονάδες» (statistical units) καθώς και της επεξεργασίας αυτών, αποτελεί σήμερα ένα θέμα ιδιαίτερης σημασίας. Σημαντικά ποσά έχουν δαπανηθεί τα τελευταία χρόνια για την ικανοποίηση των παραπάνω αναγκών με στόχο τη διαρκή βελτίωση της ανάλυσης πολύπλοκων μορφών στατιστικής πληροφορίας. Έμφαση δίνεται στη μείωση του αριθμού των δεδομένων που χρειάζεται να επεξεργαστούμε με τη μικρότερη όμως δυνατή απώλεια πληροφορίας [Papageorgiou & Vardaki, (2004)].

6.2.1 Ορισμοί[4]

Οντότητα/άτομο (Individual)

Μία οντότητα/άτομο (individual) αποτελεί τη μονάδα του στατιστικού πληθυσμού που ερευνήθηκε. Η κάθε οντότητα έχει κάποια χαρακτηριστικά που μετρώνται με συγκεκριμένες αρχικές μεταβλητές (original variables).

Συμβολικό αντικείμενο (Symbolic object)

Έστω **Ω** ένα σύνολο παρατηρηθέντων οντοτήτων (ατόμων/αντικειμένων) και **dωi** είναι η περιγραφή του **ωi** ατόμου/αντικειμένου υπό εξέταση.

Εάν **D** είναι ένα σύνολο περιγραφών **ω** αντικειμένων, τότε **dωi** $\in$ **D**

Έστω **R** είναι η σχέση (relation) μεταξύ των περιγραφών **di**, η οποία ορίζεται στο **D**

Έστω επίσης ότι **L** είναι το σύνολο των συγκρίσεων δύο περιγραφών **d** και **d'** μέσω της **R**, δηλαδή [**d' R d**] ϵ **L**

Τότε, αν ***a*** μία απεικόνιση : **Ω → L** τα συμβολικά αντικείμενα είναι τριπλέτες

$$s=(a, R, d)$$

που περιγράφουν τα άτομα *d* (από ένα σύνολο περιγραφών D), τη σχέση *R* μεταξύ των περιγραφών και μιας απεικόνισης *a* από το σύνολο ατόμων Ω στο L.

---

[4] Στις επόμενες ενότητες οι όροι θα αναφέρονται κυρίως με την αγγλική τους ορολογία και συντομογραφία επειδή θα χρησιμοποιηθούν κατ' αυτόν τον τρόπο και στο μοντέλο μεταδεδομένων που είναι στα αγγλικά.

Συμβολική μεταβλητή (Symbolic Variable - SVar)

Μία συμβολική μεταβλητή (symbolic variable) είναι μία μεταβλητή που μετράει τα χαρακτηριστικά των συμβολικών αντικειμένων και αναφέρεται σε μία ομάδα οντοτήτων (group of individuals).

Συμβολικός Πίνακας (Symbolic Data Table -SDT)

Είναι ο πίνακας που έχει ως γραμμές τα συμβολικά αντικείμενα και ως στήλες τις συμβολικές μεταβλητές

6.2.2 Σπουδαιότητα συμβολικής ανάλυσης

Όπως έχει προαναφερθεί, η μέθοδος ανάλυσης συμβολικών δεδομένων αναπτύχθηκε από την ανάγκη επεξεργασίας και ανάλυσης μεγάλου αριθμού πολύπλοκων δεδομένων τα οποία είχαν συλλεχθεί και αποθηκευτεί σε βάσεις δεδομένων Στατιστικών Υπηρεσιών και άλλων οργανισμών. Συμβολικά δεδομένα μπορούν να προκύψουν από κάθε πηγή, στοχεύοντας στη συγκέντρωση και κατηγοριοποίηση τεράστιων συνόλων πινακοποιημένων δεδομένων. Προκύπτουν από την κατανομή πιθανότητας, τα ποσοστά ή το εύρος κάθε τυχαίας μεταβλητής η οποία σχετίζεται με κάθε κελί παρόμοιου πίνακα. Προκύπτουν επίσης από Σχεσιακές Βάσεις Δεδομένων (κατά τη συγχώνευση ορισμένων σχέσεων ή συγκεντρώνοντας και κατηγοριοποιώντας απαντήσεις σε συγκεκριμένες ερωτήσεις), από την Ανάλυση δεδομένων (παραγοντική ανάλυση, αθροιστικές μέθοδοι, δίκτυα, κλπ) των συμβολικών αντικειμένων που παίρνουν δύο μόνο τιμές («σωστό, λάθος», «ναι, όχι», κλπ) [Diday 1991], [Diday et al 1995].

Επιπρόσθετα, η σπουδαιότητα των συμβολικών δεδομένων είναι ότι δίνουν τη δυνατότητα να εξεταστούν περιπτώσεις και παρατηρήσεις οι οποίες δεν στηρίζονται σε ακριβή και πειραματικά δεδομένα, αλλά ενέχουν κάποια έλλειψη ακρίβειας ή αμφιβολία κατά την καταγραφή των τιμών μιας μεταβλητής ή ακόμη και ορισμένα μη υπάρχοντα στοιχεία (missing values). Η πρακτική σημασία της χρήσης συμβολικών δεδομένων ενισχύεται από το γεγονός ότι η δημιουργία και παρουσίασή τους είναι αρκετά εύκολη αν ακολουθηθεί ένας συγκεκριμένος αριθμός βημάτων:

- Τα individuals πρέπει να χωριστούν σε ομάδες ή «οικογένειες» (groups or classes of individuals) ανάλογα με κάποιο κοινό χαρακτηριστικό το οποίο καθορίζει η τυχαία μεταβλητή.
- Διακρίνουμε τις ιδιότητες κάθε individual και τις αναπαριστούμε με συγκεκριμένες μεταβλητές οι οποίες μπορεί να είναι οποιουδήποτε τύπου (ποσοτικές, ποιοτικές, ταξινομικές, κλπ).
- Εξετάζουμε ποιες μεταβλητές είναι ανεξάρτητες και ποιες εξαρτημένες.
- Διακρίνουμε το είδος της εξάρτησης στις μεταβλητές (ιεραρχική, λογική, στοχαστική, κλπ).

- Δημιουργία νέων πινάκων με συμβολικά πλέον δεδομένα (Symbolic Data Tables, (SDT)).

Εφ' όσον τα δεδομένα προς επεξεργασία τοποθετηθούν σωστά (έχοντας ακολουθήσει τα αναγκαία βήματα) σε ανάλογο πρόγραμμα στον υπολογιστή, τα αποτελέσματα επιδέχονται περαιτέρω ανάλυση ανάλογα με τις ανάγκες κάθε οργανισμού και υπηρεσίας. Η ανάλυσή τους τόσο σε μορφή γραπτού κειμένου όσο και σε διδιάστατη ή τρισδιάστατη γραφική παράσταση [Noirhomme, 2002], [Noirhomme & Rouad, 1997], [Verde et.al, 2003] είναι δυνατό να αξιοποιηθεί από τους ενδιαφερόμενους χρήστες και επεξεργαστές δεδομένων με απώτερο στόχο τη δυνατότητα σύγκρισης των δεδομένων ανάμεσα στις διάφορες χώρες ή/και αντίστοιχες υπηρεσίες.

Για να είναι όμως δυνατή η σύγκριση των αποτελεσμάτων είναι απαραίτητο οι χρήστες και αναλυτές της στατιστικής πληροφορίας να αντιλαμβάνονται με τον ίδιο τρόπο τη μέθοδο δημιουργίας των συμβολικών δεδομένων. Αν μάλιστα αναλογιστούμε ότι οποιαδήποτε μορφή συγχώνευσης και κατηγοριοποίησης δεδομένων περιέχει κάποιες 'κρυμμένες' πληροφορίες ή πιθανώς κάποιες εντελώς χαμένες πληροφορίες, η ανάγκη για καλύτερη κατανόηση της ακριβούς έννοιας κάθε δημιουργημένου symbolic object, προϋποθέτει την ύπαρξη και ταυτόχρονη χρήση των αντίστοιχων μεταδεδομένων (metadata).

6.2.3 Διαφορές στη χρήση μεταδεδομένων στην κλασσική και συμβολική ανάλυση πινακοποιημένων δεδομένων.

Όπως αναλύθηκε και στο κεφάλαιο 4, στην κλασσική ανάλυση πινάκων, τα μεταδεδομένα προσφέρουν πληροφορία για τα υπό εξέταση individuals (το σύνολο Ω) και τις μεταβλητές, δηλαδή ουσιαστικά την απεικόνιση του Ω στο σύνολο μεταβλητών Y_i. Το σύνολο των individuals που εξετάζουμε, συνήθως περιγράφεται κατά την εξέταση των διαφόρων φάσεων μιας έρευνας, δηλαδή κατά το σχεδιασμό της έρευνας, τα στάδια υλοποίησης και τέλος τη συλλογή δεδομένων καθαυτή. Για τα στοιχεία αυτών των φάσεων, τα αντίστοιχα μεταδεδομένα είναι ουσιαστικά αυτά που εξετάσαμε στο κεφάλαιο 4 και συσχετίζονται λειτουργικά βάση του μοντέλου μεταδεδομένων που αναλύθηκε σε εκείνη την ενότητα.

Στην περίπτωση των συμβολικών δεδομένων, ο πίνακας (symbolic data table) μοιάζει σαν ένα κλασσικό πίνακα με την έννοια όμως ότι περιέχει γραμμές που αντιστοιχούν σε συμβολικά αντικείμενα, στήλες που αντιστοιχούν σε συμβολικές μεταβλητές (symbolic variables) και κάθε κελί αναφέρεται σε ένα σύνολο αντικειμένων (class/group of individuals). Σε ένα τέτοιο πίνακα, τα individuals μπορούν να συνενώνονται σε διαφορετικά σύνολα αντικειμένων χρησιμοποιώντας τις τιμές ενός ή περισσοτέρων αρχικών μεταβλητών και να σχηματίζουν τελικά πολλαπλά σύνολα αντικειμένων ή ακόμα και ενώσεις από σύνολα αντικειμένων.

Τα μεταδεδομένα που χρησιμοποιούνταν για την περιγραφή των αντικειμένων και των αρχικών μεταβλητών καθώς και των διάφορων φάσεων της έρευνας αναφέρονται μόνο στα αρχικά δεδομένα (κλασσικά δεδομένα) και δεν διαθέτουν επαρκείς

πληροφορίες για τα σύνολα αλλά και τις ενώσεις των συνόλων των αντικειμένων ούτε για τις αντίστοιχες μεταβλητές στις οποίες αναφέρονται. Και ασφαλώς, δεν επαρκούν για να περιγράψουν **πώς προέκυψαν** αυτά τα σύνολα αντικειμένων, **ποια ήταν τα κοινά και μη κοινά τους χαρακτηριστικά** καθώς και φυσικά όλες **τις διαδικασίες που διέπουν αυτές τις μεταβολές στον πίνακα**.

Για να επιτύχουμε την καταγραφή όλων των παραπάνω διαδικασιών πρέπει να περάσουμε από τον κλασσικά τρόπο σκέψης και αντιμετώπισης ενός πίνακα στη συμβολική ανάλυση. Πρέπει δηλαδή να ληφθούν υπόψη, εκτός όλων των μεταδεδομένων της κλασσικής ανάλυσης και τα κάτωθι:

- Τα μεταδεδομένα που περιγράφουν τη διαδικασία δημιουργίας κάθε symbolic object, λαμβάνοντας υπόψη τον μετασχηματισμό που εφαρμόστηκε στην κάθε αρχική μεταβλητή, το είδος των υποθέσεων και τις νέες επιτρεπόμενες τιμές των μεταβλητών.
- Στην περίπτωση ενώσεων symbolic objects η σειρά των διαδικασιών είναι απαραίτητο να είναι γνωστή στον τελικό χρήστη

Μία βασική διαφορά με τα μεταδεδομένα που εξετάζονται στην κλασσική ανάλυση είναι ότι η μεταπληροφορία για τα σύνολα των individuals είναι απαραίτητη για την κατανόηση των συμβολικών δεδομένων, ενώ στην κλασσική ανάλυση οι τελικοί χρήστες ενδιαφέρονται συνήθως για το σύνολο των αντικειμένων και όχι για τα χαρακτηριστικά του καθενός ξεχωριστά (πληροφορία η οποία άλλωστε συνήθως δεν είναι διαθέσιμη λόγω εμπιστευτικότητας των δεδομένων). Στην περίπτωση όμως συνόλων individuals είναι απαραίτητη η περιγραφή τους και η γνώση για τα κοινά και μη κοινά στοιχεία τους και τη διαδικασία της σύνθεσής τους.

Σημειώνεται, ότι η χρήση τους για τον έλεγχο ποιότητας των αποτελεσμάτων ενός survey περιγράφεται και στο [Marcelo, 2002].

6.2.4 Χρήση συμβολικών δεδομένων

Οι περιπτώσεις στις οποίες κατεξοχήν χρησιμοποιείται η μέθοδος των συμβολικών δεδομένων είναι εκείνες κατά τις οποίες καλούμαστε να αναλύσουμε ομογενείς κλάσεις (κατηγορίες) individuals (αναφέρονται στη βιβλιογραφία και ως «υπερ-άτομα», «συγκεντρωτικά αντικείμενα» ή «δεύτερης τάξης αντικείμενα»), [Bock & Diday, 2000], [Diday 1991], [Diday et al, 1995], [Stephan et.al, 1997].

Ένα χαρακτηριστικό παράδειγμα προέρχεται από τις χώρες στις οποίες η κάθε μία αντιστοιχεί σε ένα σύνολο οργανισμών οι οποίοι χρηματοδοτούνται για έρευνα και τεχνολογική ανάπτυξη (ΕΤΑ).

Στον τομέα της Έρευνας και Τεχνολογικής Ανάπτυξης (ΕΤΑ) η οποία διενεργείται στον κλάδο των Επιχειρήσεων, μπορούμε να εφαρμόσουμε τη μεθοδολογία ανάλυσης συμβολικών δεδομένων με στόχο να δημιουργήσουμε συγκρίσιμους δείκτες σχετικά με τους πόρους οι οποίοι διατίθενται για ΕΤΑ, με δυνατότητα να χρησιμοποιηθούν για σύγκριση των τάσεων στις χώρες της Ευρωπαϊκής Ένωσης. Με τον τρόπο αυτό θα καταστεί εφικτή η σύγκριση των ανωτέρω διατιθέμενων σε κάθε χώρα ποσών,

εξαλείφοντας κατ' αυτό τον τρόπο τυχόν παρερμηνείες ή λανθασμένες συγκρίσεις. Όλες οι έννοιες και τα αρχικά δεδομένα είναι διαθέσιμα σε επίσημες εκδόσεις της Ευρωπαϊκής Στατιστικής Υπηρεσίας (Eurostat). Στην Ελλάδα απαγορεύεται από τη νομοθεσία η δημοσίευση τέτοιων στοιχείων για κάθε επιχείρηση και είναι διαθέσιμα μόνο τα συγκεντρωτικά στοιχεία τα οποία δημοσιεύονται σε εκδόσεις της Ευρωπαϊκής Ένωσης. Στο παράδειγμά αυτό, ως βασική μεταβλητή κατηγοριοποίησης σε ομάδες (ή κλάσεις) μπορούμε να θεωρήσουμε είτε την ταξινόμηση NACE (σε αναλογία με τα πορίσματα του κεφ.2) η οποία κατηγοριοποιεί τις επιχειρήσεις κατά κύρια δραστηριότητα, είτε τον ισολογισμό, το μέγεθος της επιχείρησης, τον αριθμό του προσωπικού, κλπ. και να εξετάσουμε όλες τις εξαρτημένες μεταβλητές και τους κανόνες εξάρτησης αυτών.

6.3 Επέκταση μοντέλου μεταδεδομένων για συμβολικά δεδομένα

Βάσει των απαιτήσεων της δημιουργίας των Symbolic Variables (Svars), των Symbolic Objects (SO) και των Symbolic Data Tables (SDT), συγκεκριμένες κατηγορίες μεταδεδομένων λήφθηκαν υπόψη, όχι μόνο για την περιγραφή της διαδικασίας δημιουργίας των SO και SDT αλλά και για την περαιτέρω χρήση τους και ανάλυση με την εφαρμογή των μεθόδων συμβολικής ανάλυσης. Δύο βασικές κατηγορίες (και υποκατηγορίες αυτών) θεωρήθηκαν απαραίτητες και αναλύονται περαιτέρω:

Μεταδεδομένα για τα αρχικά δεδομένα (Original Data)

Σε αυτή την κατηγορία ανήκουν τα μεταδεδομένα που περιγράφουν τα αρχικά δεδομένα που χρησιμοποιήθηκαν για τη δημιουργία των υπό εξέταση SO και SDT:

i) **Μεταδεδομένα της έρευνας (Survey metadata):** είναι τα μεταδεδομένα εκείνα που περιγράφουν πώς πραγματοποιήθηκε η έρευνα σε όλα της τα στάδια. Συγκεκριμένα, περιγράφουν το στατιστικό πληθυσμό (statistical population), την περιγραφή των χαρακτηριστικών των individuals που συμμετείχαν στην έρευνα, τη δειγματοληπτική μέθοδο (sampling method) που εφαρμόστηκε, κλπ. Εν γένει, αυτή η κατηγορία μεταδεδομένων αναφέρεται σε όλη τη διαδικασία συλλογής δεδομένων χωρίς να συγκεκριμενοποιεί κάθε μεταβλητή ή individual. Αυτό σημαίνει ότι αν παρουσιάσουμε τα δεδομένα μιάς έρευνας σε ένα πίνακα, όπου κάθε στήλη παριστάνει μία μεταβλητή και κάθε γραμμή ένα individual, τότε τα survey metadata αναφέρονται στο σύνολο του πίνακα.

ii) **Μεταδεδομένα για τη μεταβλητή (Variable metadata):** αναφέρονται σε μία συγκεκριμένη μεταβλητή που μετράται σε μία έρευνα. Για παράδειγμα, αν η «απασχόληση» είναι μία από τις ερωτήσεις του ερωτηματολογίου της έρευνας, δε μπορούμε να συγκρίνουμε αντίστοιχα δεδομένα από διαφορετικές έρευνες αν δεν γνωρίζουμε ακριβώς τον ορισμό της «απασχόλησης» καθώς και τις πληροφορίες που τη συνοδεύουν, όπως την ταξινόμηση των διαφόρων ειδών απασχόλησης, κλπ. Συνεπώς, αυτή η κατηγορία μεταδεδομένων αναφέρεται στην συγκεκριμένη μεταβλητή που αντιπροσωπεύεται από μια συγκεκριμένη στήλη του πίνακα.

Ένα σχετικό και πιο εκτενές μοντέλο μεταδεδομένων για τα κλασσικά δεδομένα έχει αναλυθεί από τους [Papageorgiou et al, 2001a, 2001b] και ένα ακόμα πιό αναλυτικό μοντέλο για όλες τις φάσεις συλλογής, ανάλυσης και διάχυσης της στατιστικής μεταπληροφορίας δημιουργήθηκε για την παρούσα διατριβή και αναλύθηκε στο κεφάλαιο 4.

Συμβολικά μεταδεδομένα (Symbolic Metadata)

Σε αυτή την κατηγορία περιλαμβάνονται τα εξής:

i) Μεταδεδομένα για symbolic object:

Τα μεταδεδομένα αυτά περιγράφουν τη διαδικασία δημιουργίας ενός SO καθορίζοντας τις symbolic variables, τους τελεστές (operators) που εφαρμόστηκαν σε αυτές τις μεταβλητές (Average, sum etc.), τις αναγκαίες προϋποθέσεις και τις αντίστοιχες τιμές των μεταβλητών (upper, lower limits, thresholds κλπ.). Μία σημαντική διαφορά με τα κλασσικά δεδομένα, όπως προαναφέρθηκε, είναι ότι η μεταπληροφορία για τα ομαδοποιημένα πλέον individuals σε classes of individuals είναι πολύ σημαντική πληροφορία, ενώ στα κλασσικά δεδομένα η αντίστοιχη πληροφορία για τα individuals δεν είναι σημαντική ή δεν προσφέρεται για λόγους εμπιστευτικότητας. Με το να έχει κανείς την περιγραφή και τη διαδικασία δημιουργίας των classes of individuals του δίνεται η ευκαιρία για περαιτέρω ανάλυση και νέες ομαδοποιήσεις με βάση άλλο κοινό χαρακτηριστικό. Ενδιαφέροντα στοιχεία μεταπληροφορίας είναι ο αριθμός των individuals που συνενώθηκαν για τη δημιουργία του SO, ο αριθμός των individuals του πληθυσμού και του δείγματος από όπου προέκυψαν.

Επομένως, τα μεταδεδομένα που πρέπει οπωσδήποτε να ληφθούν υπόψη σε αυτή την κατηγορία και τα οποία αναφέρονται με την αγγλική τους ορολογία επειδή έτσι συμμετέχουν στο μοντέλο, είναι τα εξής:

- *name,*
- *id.,*
- *code,*
- *number of individuals participating,*
- *condition of aggregation (type, threshold and restrictions),*
- *description (d),*
- *the mapping association (a),*
- *the Relation (R),*
- *the size,*
- *the SDT row (όπου το SO περιγράφεται)*
- *operations (transformations) που μπορούν να εφαρμοστούν για τη δημιουργία νέων SO*

ii) **Μεταδεδομένα για Symbolic Variables**:

Οι SVars παράγονται από το μετασχηματισμό που οδηγεί στη δημιουργία κάθε SO. Σε κάθε SO αναφέρεται ένα σύνολο individuals (class/group of individuals) αντί για

ένα μόνο αντικείμενο και κάθε SO σχετίζεται με ένα σύνολο τιμών, αντί για μία μοναδική τιμή. Ένας μετασχηματισμός σε αυτό το σύνολο τιμών δημιουργεί την symbolic variable. Συνεπώς, τα αντίστοιχα μεταδεδομένα πρέπει να περιγράφουν τη διαδικασία που αυτές οι symbolic variables δημιουργήθηκαν από τα αρχικά δεδομένα.

Επιπρόσθετα, οι SVars πρέπει να ονομαστούν, να προσδιοριστεί το πεδίο ορισμού και το σύνολο τιμών τους και το είδος τους [Bock & Diday, 2000]. Πιό συγκεκριμένα, για κάθε κατηγορία των symbolic variables πρέπει να οριστούν τα κάτωθι:

a. Το εύρος
b. Η ταξινόμηση από την οποία παίρνει τις τιμές της
c. Σύνολο ορισμού και πεδίο τιμών
d. Η στήλη του συμβολικού πίνακα (SDT column) όπου παρουσιάζεται η μεταβλητή αυτή

iii) Μεταδεδομένα για το Symbolic Data Table:

Όπως έχει προαναφερθεί, ένα Symbolic Data Table μοιάζει με ένα πίνακα της κλασσικής ανάλυσης αλλά οι γραμμές του αναφέρονται σε SOs, κάθε κελί περιγράφει groups of individuals και οι στήλες του σε symbolic variables.

Για το SDT πρέπει να κρατηθεί κάθε πληροφορία σχετικά με τη δημιουργία του καθώς και με όλες τις πράξεις και μετασχηματισμούς (operations) που μπορούν να εφαρμοστούν πάνω στις γραμμές και στήλες του. Οποιαδήποτε διαφοροποίηση δε του SDT δημιουργεί ένα νέο SDT, συνεπώς κάθε στάδιο μετατροπής πρέπει να είναι εφικτό να ανακληθεί και να είναι εφικτό στο χρήστη να αναζητήσει το αρχικό SDT. Συνεπώς, πρέπει να είναι διαθέσιμες οι περιγραφές των operators που εφαρμόστηκαν και το αποτέλεσμα που έδωσαν σε κάθε στάδιο εφαρμογής τους.

6.4 Περιγραφή του μοντέλου μεταδεδομένων

Όπως έχει αναλυθεί στο Κεφάλαιο 3, η ενδεδειγμένη τεχνική μοντελοποίησης για το παρόν πρόβλημα είναι η **Object-Oriented (O-O)** οπότε και το αναμορφωμένο μοντέλο, δημιουργήθηκε με εργαλείο που ακολουθεί O-O τεχνολογία, τη UML Rational Rose 2000 Enterprise Edition.

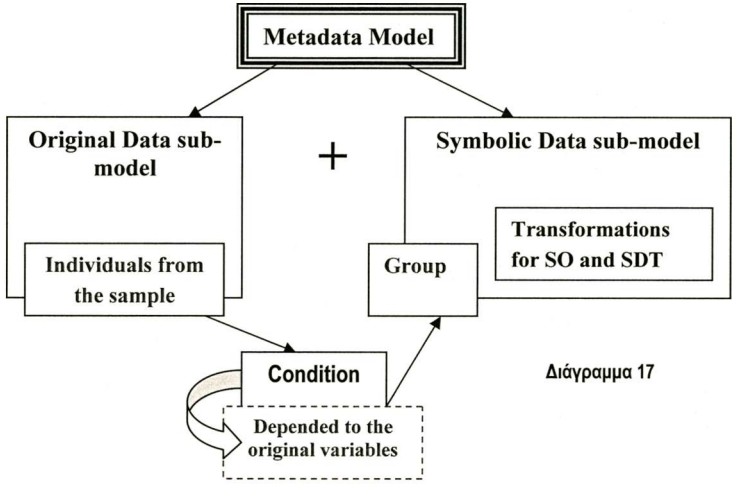
Ασφαλώς, όπως και στο μοντέλο του κεφαλαίου 4, το τροποποιημένο αυτό μοντέλο πραγματοποιήθηκε χρησιμοποιώντας αγγλική ορολογία που υποστηρίζεται από το συγκεκριμένο UML εργαλείο. Επιπρόσθετα, επειδή το συγκεκριμένο UML μοντέλο έχει ως στόχο να μπορεί να χρησιμοποιηθεί από πληροφοριακά συστήματα και να δύναται να περιγραφεί σε XML (μέσω ενός xmi file που δημιουργείται αυτόματα από τη Rational Rose), θα ήταν αδύνατη οποιαδήποτε περαιτέρω χρήση του (ή απλά δημοσίευσή του) αν είχε δημιουργηθεί στην ελληνική γλώσσα.

Για την ευκολότερη μελέτη του μοντέλου, χωρίστηκε σε δύο μέρη σύμφωνα με την παραπάνω ανάλυση των κατηγοριών των μεταδεδομένων:

- Το τμήμα του μοντέλου μεταδεδομένων για τα αρχικά δεδομένα (Original data (survey, original variables))
- Το τμήμα του μοντέλου μεταδεδομένων για τα symbolic data (SOs, SDT, Symbolic Variables (Svars)).

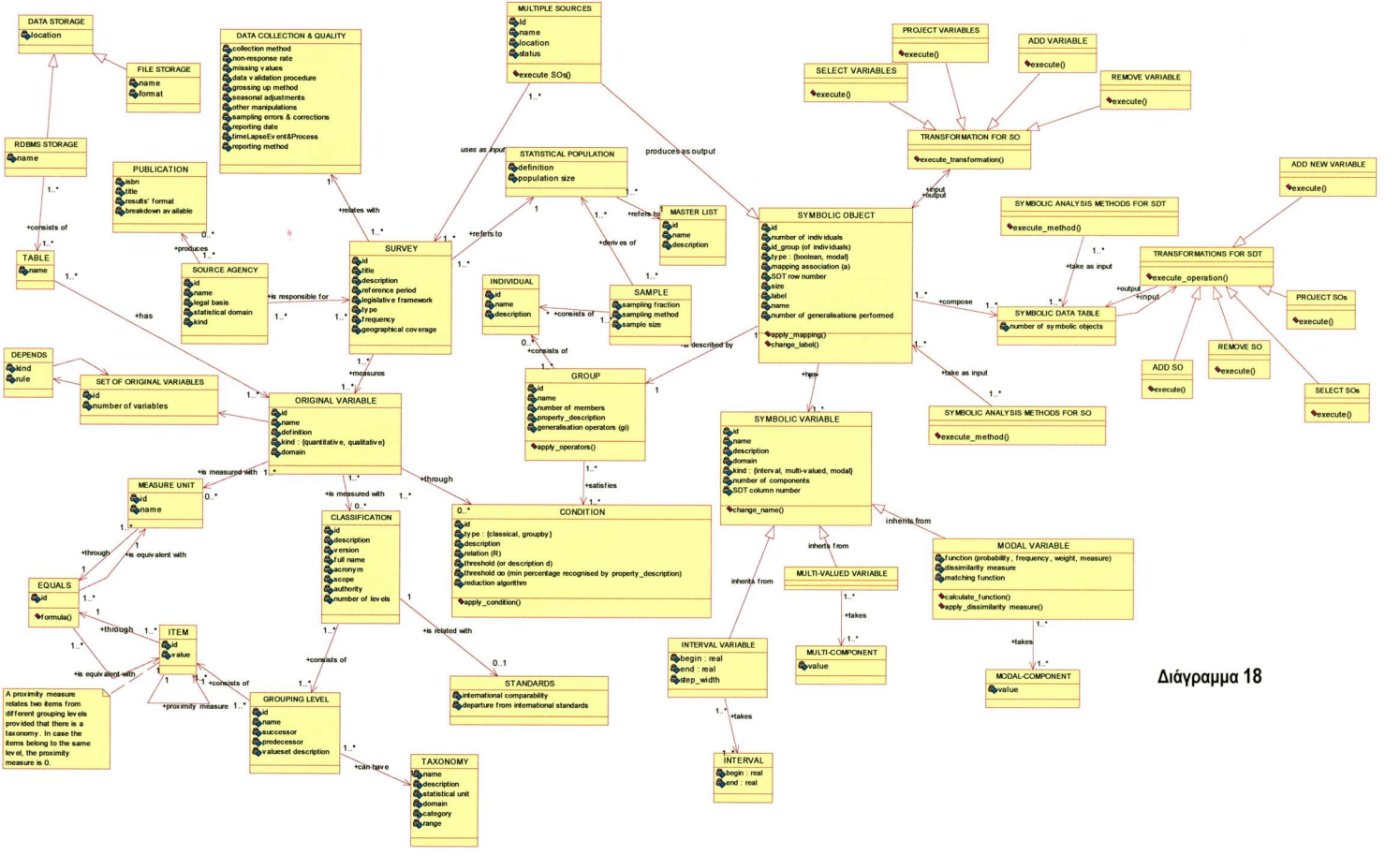
Αυτά τα δύο τμήματα του μοντέλου ενέχουν πληροφορία για διαφορετικά στάδια της ανάλυσης και την δημιουργίας των SO. Παρόλα αυτά είναι αλληλένδετα και η κύρια διαδικασία ένωσής πραγματοποιείται μέσω της κλάσης "GROUP", η οποία δίνει πληροφορία για τις ενώσεις των individuals που ανήκουν στο δείγμα (SAMPLE), και ικανοποιεί ένα συγκεκριμένο 'CONDITION' που σχετίζεται με τις ORIGINAL VARIABLES.

Μία απλουστευμένη διαγραμματική μορφή των δύο υπο-μοντέλων και της σύνδεσής τους παρουσιάζεται στο διάγραμμα 17.



Διάγραμμα 17

Το συνολικό μοντέλο παρουσιάζεται στο διάγραμμα 18

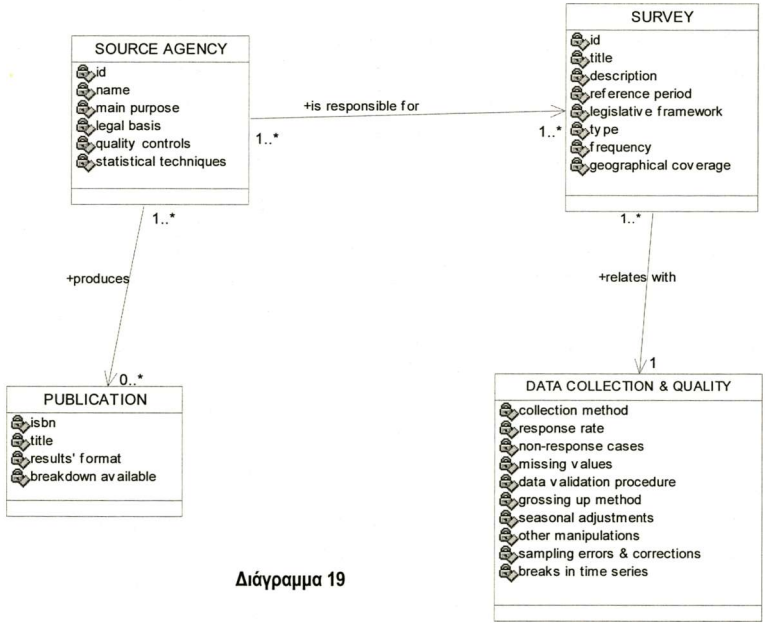


Διάγραμμα 18

Το μοντέλο κρατάει πληροφορία για τα εξής:

- Οργανισμούς υπεύθυνους για συλλογή και επεξεργασία πληροφορίας
- Πληροφορίες συλλογής δεδομένων
- Στατιστικούς πληθυσμούς
- Κλασσικές μεταβλητές και πρότυπα.
- Συμβολικά δεδομένα και συμβολικές μεταβλητές.
- Λογιστικά μεταδεδομένα.

Στο διάγραμμα 19 περιλαμβάνεται μέρος του μοντέλου με τις πιθανές πηγές δεδομένων, τις δημοσιεύσεις αποτελεσμάτων και τα μεταδεδομένα για τη συλλογή και την ποιότητα των συλλεχθέντων δεδομένων.



Διάγραμμα 19

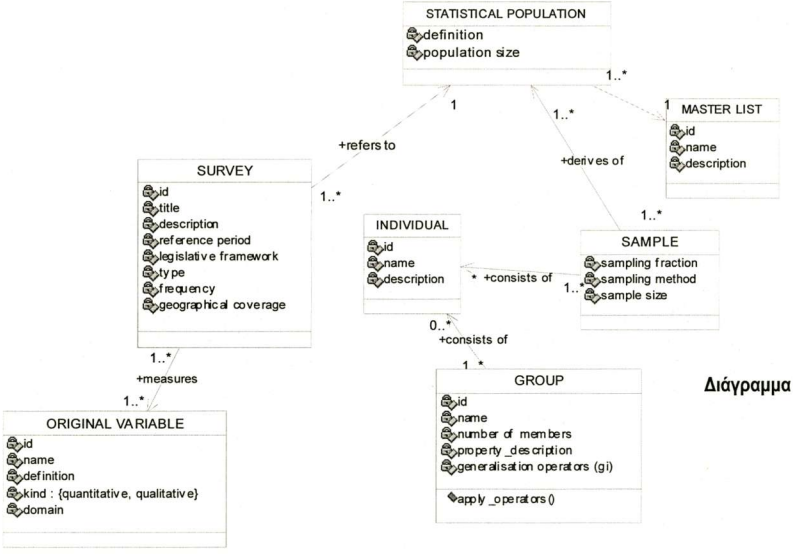
Συγκεκριμένα παρουσιάζει ότι ένα SURVEY έχει τουλάχιστον ένα SOURCE AGENCY (ένα στατιστικό οργανισμό), αρμόδιο για τη συλλογή ή τη διάταξη και ανάλυση των δεδομένων. Επιπλέον, είναι σημαντικό να κρατηθούν οι πληροφορίες για τη διαδικασία της συλλογής δεδομένων (COLLECTION INFO AND QUALITY) και οι πιθανές ασυνέχειες στις χρονοσειρές προκειμένου να μπορούν οι χρήστες να αξιολογούν την ποιότητα των στοιχείων. Ανάμεσα στις πληροφορίες που μοντελοποιούνται υπό την κλάση COLLECTION INFO AND QUALITY είναι η μέθοδος

συλλογής, το ποσοστό απόκρισης (π.χ. εάν το ποσοστό απάντησης μιας έρευνας είναι πολύ χαμηλό, είναι αβέβαιο εάν τα αποτελέσματα είναι αντιπροσωπευτικά), οι missing values, διάφορα σφάλματα και μέθοδοι αποκατάστασής τους, κλπ.

Επίσης, τα SOURCE AGENCIES δημοσιεύουν συνήθως τα αποτελέσματα μιας έρευνας (PUBLICATION), έτσι ώστε οι τελικοί χρήστες μπορούν να έχουν πρόσβαση σε αυτά.

Στο διάγραμμα 20 παρουσιάζεται μία ομάδα (GROUP) που περιγράφει τα άτομα (INDIVIDUALS) που συμμετέχουν στο SURVEY. Ο καθορισμός του συνόλου όλων των ατόμων που καταγράφεται σε μια έρευνα είναι ο στατιστικός πληθυσμός (STATISTICAL POPULATION) της έρευνας (SURVEY), ενώ ένας κύριος κατάλογος (MASTER LIST) είναι ένα σύνολο από όπου μπορούμε να εξάγουμε το δείγμα (SAMPLE). Παραδείγματος χάριν, θα μπορούσε να είναι ο κατάλογος όλων των επιχειρήσεων με προσωπικό περισσότερο από 100 άτομα.

Επιπλέον, τα άτομα διαμορφώνουν τις ομάδες (GROUP) σύμφωνα με την εφαρμογή ενός CONDITION σε ένα ή περισσότερα χαρακτηριστικά (CHARACTERISTICS) που αντιπροσωπεύονται από τις αρχικές μεταβλητές (ORIGINAL VARIABLES).



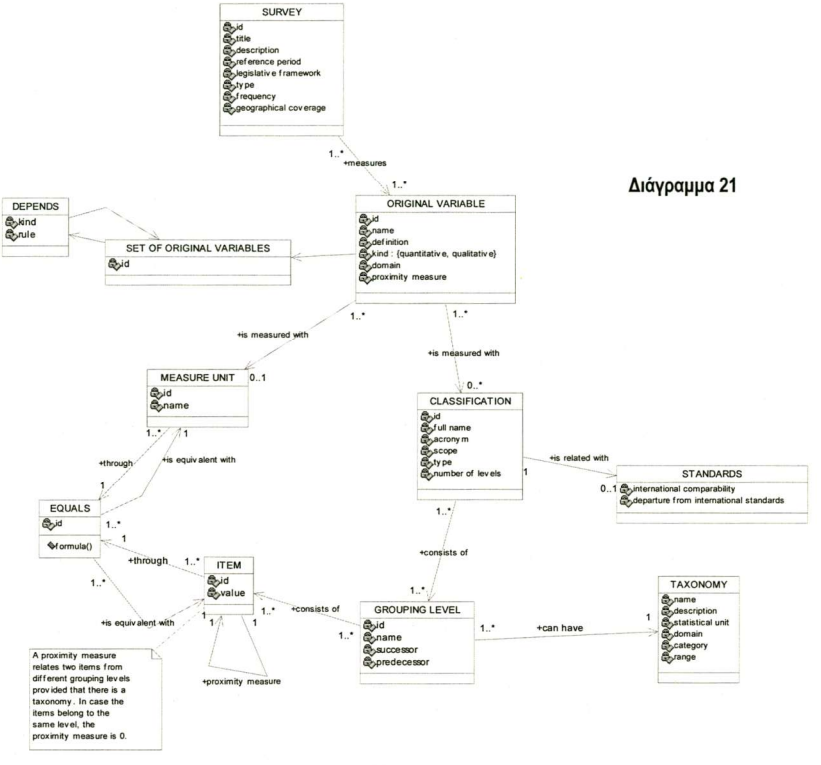
Διάγραμμα 20

Παρατήρηση: Η κατηγορία "GROUP" εμπλουτίζεται από το 'property_description' και generalisation operators(gi)' που χρησιμοποιούνται στον 'αλγόριθμο μείωσης SO'

που περιγράφεται από τους Bock και Diday (2000). Επίσης, στο διάγραμμα 22 εξετάζεται εκτενέστερα η κλάση "CONDITION" για περισσότερες πληροφορίες για τον τρόπο με τον οποίο ο αλγόριθμος εφαρμόζεται).

Στο διάγραμμα 21 παρατηρούμε πώς οι αρχικές μεταβλητές (ORIGINAL VARIABLES) συσχετίζονται με ένα SURVEY. Πραγματικά, ένα SURVEY μετρά μια ή περισσότερες ORIGINAL VARIABLES που αντιπροσωπεύουν ορισμένα χαρακτηριστικά (CHARACTERISTICS) των ατόμων (INDIVIDUALS) στο στατιστικό πληθυσμό (STATISTICAL POPULATION). Μια αρχική μεταβλητή μπορεί να είναι είτε ποσοτική (quantitative) είτε ποιοτική (qualitative) ανάλογα με τον τρόπο μέτρησής της. Στην πρώτη περίπτωση, οι τιμές μετριούνται σύμφωνα με μια μονάδα μέτρησης (MEASUREMENT UNIT), η οποία συσχετίζεται πιθανώς με μια άλλη μέσω της κλάσης EQUALS. Αυτό σημαίνει ότι εάν υπάρχει μια διευκρινισμένη σχέση μεταξύ δύο μονάδων μέτρησης, αυτή μπορεί να μετασχηματιστεί σε άλλη και αντίστροφα. Π.χ. η γιάρδα μετατρέπεται σε μέτρα σύμφωνα με έναν συγκεκριμένο τύπο (1yard = 0,91meters).

Στη δεύτερη περίπτωση, οι τιμές της αρχικής μεταβλητής μετριούνται με μια ταξινόμηση (CLASSIFICATION) που αποτελείται από επίπεδα ομαδοποίησης (GROUPING LEVELS), τα οποία με τη σειρά τους απαρτίζονται από τιμές (VALUES). Για παράδειγμα, η NUTS 1999 ταξινόμηση για τις γεωγραφικές περιοχές αποτελείται από τέσσερα επίπεδα. Το πρώτο επίπεδο (επίπεδο 0) περιέχει τα ονόματα όλες χώρα στην ΕΕ, έτσι τα στοιχεία του επιπέδου είναι: Ελλάδα Βέλγιο, Δανία, Γερμανία, Ισπανία, Γαλλία, Ιρλανδία, Λουξεμβούργο, Ολλανδία, Ιταλία, Πορτογαλία, Sverige, Suomi, Αυστρία και Ηνωμένο Βασίλειο.



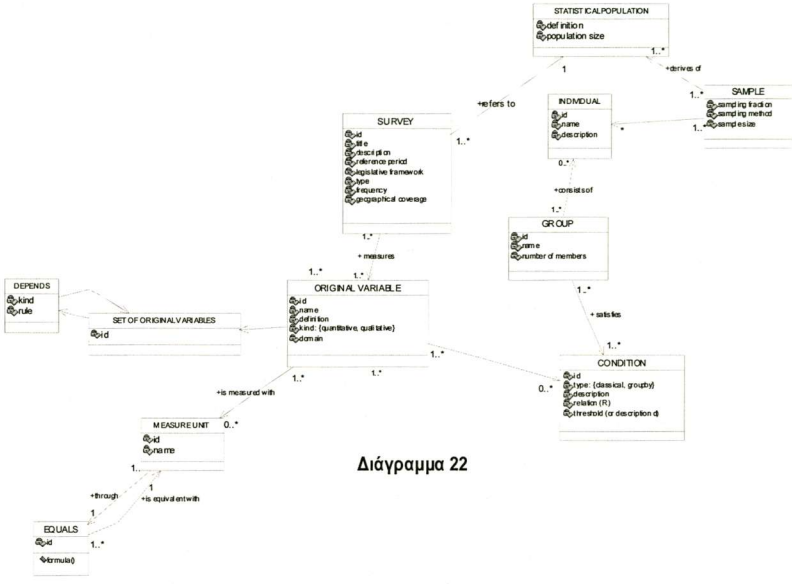
Διάγραμμα 21

Στο διάγραμμα 22 παρουσιάζεται η σχέση των μεταδεδομένων των INDIVIDUALS με τα μεταδεδομένα των συμβολικών αντικειμένων (SYMBOLIC OBJECTS) που αναφέρονται στις ομάδες (GROUPS) ατόμων/αντικειμένων.

Ιδιαίτερα, ένα GROUP of INDIVIDUALS που ικανοποιεί ένα σύνολο CONDITIONS στις ORIGINAL VARIABLES περιγράφει ένα SYMBOLIC OBJECT.

Παραδείγματος χάριν, υποθέτουμε ότι από το σύνολο αυτοκινήτων θέλουμε να δημιουργήσουμε ένα symbolic object που αναφέρεται στην ομάδα (group) 'limousines που παράγονται στην Ευρωπαϊκή Ένωση μετά το 1995'. Έτσι, τα αυτοκίνητα ομαδοποιούνται σύμφωνα με τον 'τύπο αυτοκινήτου' με επιπλέον μεταβλητές τον 'τόπο παραγωγής' και το 'έτος παραγωγής', λειτουργώντας ως CONDITIONS για τη δημιουργία των SYMBOLIC OBJECTS. Επίσης, τα δύο CONDITIONS εφαρμόζονται συνδετικά, δηλαδή ο *τελεστής* του συμβολικού αντικειμένου είναι ίσος με το λογικό ΚΑΙ ($\wedge$). Κατά συνέπεια, δημιουργούνται Ν συμβολικά αντικείμενα έχοντας τρεις νέες συμβολικές μεταβλητές (symbolic variables): ο τύπος αυτοκινήτων, ο χρόνος παραγωγής και οι περιοχή της παραγωγής. Οι τυχόν υπόλοιπες αρχικές μεταβλητών, π.χ. τιμή, ιπποδύναμη, κλπ, γίνονται

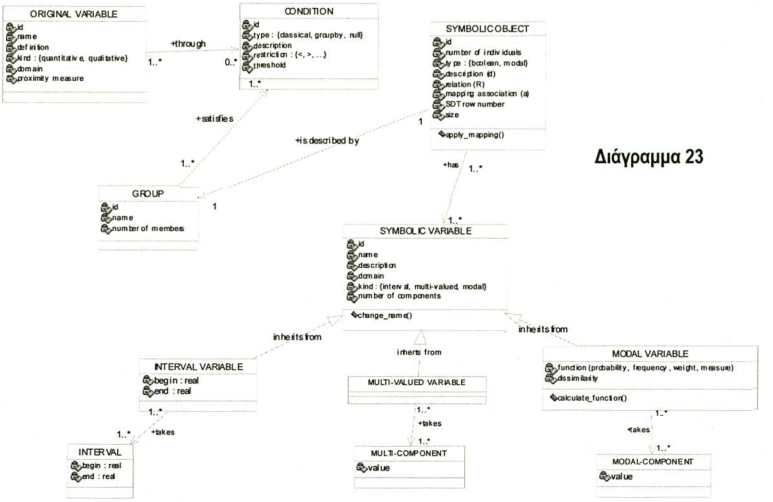
συμβολικές μεταβλητές χωρίς conditions (ή με "null condition"). Συνεπώς, ένα συμβολικό αντικείμενο συσχετίζεται με την ομάδα ατόμων που προέρχεται από και με τις μεταβλητές του.



Διάγραμμα 22

Ένα άλλο σημαντικό στοιχείο που έχει μοντελοποιηθεί είναι τόσο ο αριθμός σειρών (ROWS NUMBER) ενός SDT καθώς και η σειρά στην οποία παρουσιάζεται ένα SO στον SDT (column), που είναι απαραίτητες πληροφορίες για τους περαιτέρω χειρισμούς στο SDT.

Επιπλέον, στο διάγραμμα 23 μια συμβολική μεταβλητή είναι είτε μια μεταβλητή INTERVAL–οι τιμές που της είναι διαστήματα αριθμών ή διαταγμένων κατηγορικών τιμών - ή μια MULTI-VALUED μεταβλητή –οι τιμές της είναι σύνολα τιμών - ή μια MODAL VARIABLE, η οποία είναι πιό σύνθετη από άλλες και η τιμές της είναι ένα σύνολο ζευγών, όπου κάθε ζεύγος αποτελείται από μια τιμή που παρατηρείται στο συγκεκριμένο συμβολικό αντικείμενο.



Διάγραμμα 23

Τέλος, ορισμένες μέθοδοι που παρατηρούνται να εφαρμόζονται στη συμβολική ανάλυση μπορούν να μοντελοποιηθούν στο ίδιο μοντέλο, ώστε να είναι εφικτή η διατήρηση της ιστορικότητας των χειρισμών ενός χρήστη αλλά και να είναι δυνατή η μετατροπή όλων των SO και SDT που είναι αποθηκευμένα στη βάση δεδομένων. Αυτοί οι μετασχηματισμοί μπορούν αν εφαρμόζονται είτε στο SYMBOLIC OBJECT, είτε στη SYMBOLIC VARIABLE είτε στο SYMBOLIC DATA TABLE.

Στο διάγραμμα 24 παρουσιάζονται οι πιθανοί μετασχηματισμοί που επιλέξαμε να μοντελοποιήσουμε μετά από τη μελέτη των απαιτήσεων σύμφωνα με [Bock & Diday, 2000]

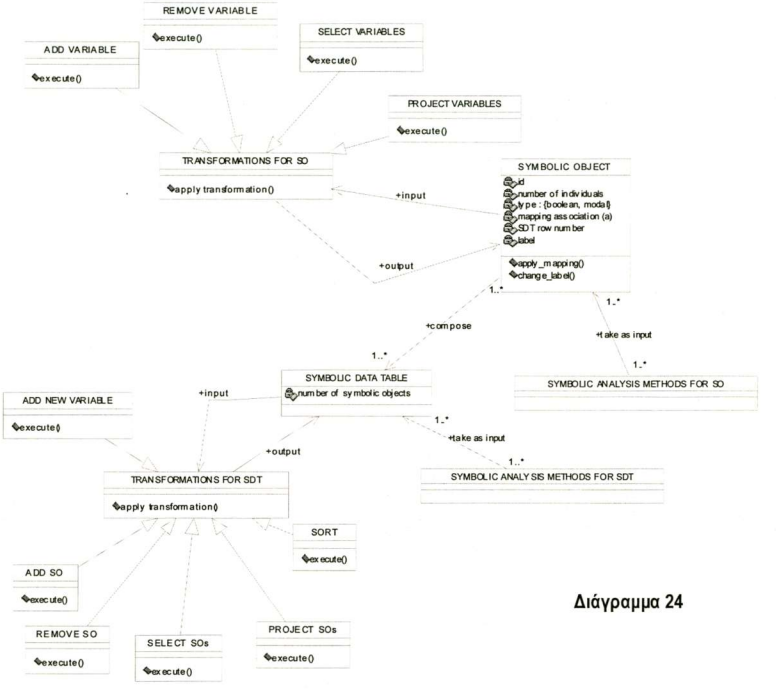
Μέθοδοι που εφαρμόζονται σε ένα SYMBOLIC OBJECT:

- Επιλογή (**SELECTION**) ενός SO για περαιτέρω επεξεργασία
- Προσθήκη (**ADDITION**) ή αφαίρεση (**REMOVAL**) μιας SVar. Σε κάθε τέτοια περίπτωση ένα νέο SO δημιουργείται
- **PROJECTION** ενός συνόλου από μία ή περισσότερες SVars. Σε κάθε τέτοια περίπτωση ένα νέο SO δημιουργείται
- Η δυνατότητα για **προσθήκη** ή **αφαίρεση** ορισμένων individuals από ένα SO είναι επίσης μία ενδιαφέρουσα ιδιότητα αλλά δεν αναπτύσσεται στη παρούσα διατριβή.

Μέθοδοι που εφαρμόζονται σε ένα SYMBOLIC DATA TABLE:

- Επιλογή (**SELECTION**) ενός ή περισσοτέρων SO από ένα SDT.
- Προσθήκη (**ADDITION**) ή Αφαίρεση (**REMOVAL**) ενός SO. Σε κάθε τέτοια περίπτωση ένα νέο SDT παράγεται.

- **PROJECTION** ενός ή περισσοτέρων SVars. Σε κάθε τέτοια περίπτωση ένα νέο SDT παράγεται.
- Αλλαγή σειράς (**SORTING**) των SYMBOLIC OBJECTS τα οποία περιλαμβάνονται σε ένα SYMBOLIC DATA TABLE. Σε κάθε τέτοια περίπτωση ένα νέο SDT παράγεται



Διάγραμμα 24

6.5 Εφαρμογές

Στην ενότητα αυτή παραθέτουμε δύο εφαρμογές του μοντέλου μεταδεδομένων στα συμβολικά δεδομένα. Στην *Εφαρμογή 1* εξηγείται βήμα προς βήμα πώς το προτεινόμενο μοντέλο μπορεί να χρησιμοποιηθεί στην αυτοματοποιημένη δημιουργία συμβολικών αντικειμένων, όταν γνωρίζουμε μόνο τις αρχικές μεταβλητές μιάς δειγματοληπτικής έρευνας και τον αριθμό των ατόμων/αντικειμένων.

Στη *δεύτερη εφαρμογή* περιγράφεται η δυνατότητα αυτοματοποιημένης παρουσίασης των μοντελοποιημένων μεταδεδομένων ταυτόχρονα με την εξαγωγή των συμβολικών πινάκων ή των γραφημάτων αναπαράστασης των συμβολικών αντικειμένων.

6.5.1 Εφαρμογή 1

Το παράδειγμα στηρίζεται στο (Bock and Diday, 2000, σελ.58, πίνακας 4.3).

Θεωρούμε ένα πίνακα κλασσικών μεταβλητών (Classical Table) ο οποίος περιέχει πληροφορία για 6 χώρες (Am, Be, Ci, Do, Fu, Ga), τον αριθμό των individuals, τη βιομηχανική δραστηριότητα (industrial activity) (οι κατηγορίες των βιομηχανιών παρουσιάζονται ως A, Co, C, E, I, En), το Εθνικό Ακαθάριστο Προϊόν (Gross National Product- GNP) και το Κόμμα (Political party) που κέρδισε τις εκλογές του 1995.

Όλες οι σχετικές πληροφορίες κρατώνται από τα attributes των κλάσεων SURVEY και ORIGINAL VARIABLES (**yi**) και απεικονίζονται στο διάγραμμα 25 Συγκεκριμένα έχουμε:

Number of Individuals=308

Classical Variables = 5 :

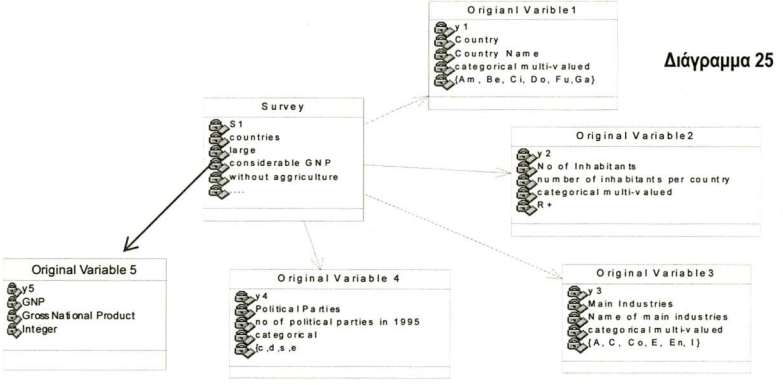
y1 =country,

y2=Number of Inhabitants,

y3=Main Industries,

y4=Political party in the 1995 elections

y5=GNP



Διάγραμμα 25

Αν ομαδοποιήσουμε κατά τη μεταβλητή y_1, τότε παράγεται ένας νέος πίνακας, πλέον ένας συμβολικός πίνακας (SDT) με τις εξής Symbolic Variables (Y_i):

Y_1=Number of Individuals,

Y_2=GNP,

Y_3=Main Industries,

Y_4=Political Party,

οι οποίες είναι ομαδοποιημένες έχοντας χρησιμοποιήσει έναν 'GroupBy' operator κατά χώρα (Country).

Παρατηρούμε ότι η πληροφορία για τις χώρες (από την αρχική μεταβλητή y_1 = country) «χάθηκε» από το συμβολικό πίνακα. Το μοντέλο όμως μεταδεδομένων κρατάει αυτή την πληροφορία, καθώς επίσης και τη διαδικασία που πραγματοποιήθηκε, ώστε ο χρήστης αν επιθυμεί να μπορεί να επιστρέψει στον αρχικό του πίνακα.

Έστω τώρα ότι θέλουμε να εξάγουμε συμβολικό αντικείμενο (SO) με πληροφορία:

«Όλες οι χώρες που έχουν πληθυσμό μεγαλύτερο από 30 εκατομμύρια με GNP μεταξύ 200 και 800 δις ευρώ και δεν έχουν βιομηχανία τύπου Α»

Στην περίπτωση αυτή λοιπόν, εξετάζουμε ποιες από τις μεταβλητές του διαγράμματος 25 σχετίζονται με το ζητούμενο. Βλέπουμε ότι πουθενά δεν μας ενδιαφέρει η πληροφορία για το "Political Party".

Επομένως, οι υπό εξέταση μεταβλητές που σχετίζονται με το προς παραγωγή SO δε σχετίζονται καθόλου με τη μεταβλητή Y_4, αλλά έχουν ως ζητούμενο για τις υπόλοιπες μεταβλητές τα εξής:

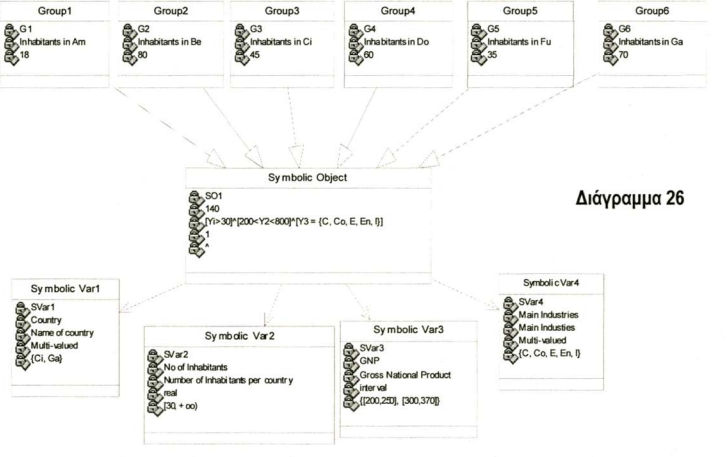
$Y_1 > 30$ (σε εκατομμ) και $200 < Y_2 < 800$ (σε δις) και $Y_3 = \{C, Co, E, En, I\}$

Η διαδικασία, όπως παρουσιάζεται στο διάγραμμα 26. Τα άτομα (individuals) χωρίζονται σε 6 Groups σύμφωνα με τις τιμές της αρχικής μεταβλητής

$$y_1 = country = \{Am, Be, Ci, Do, Fu, Ga\}$$

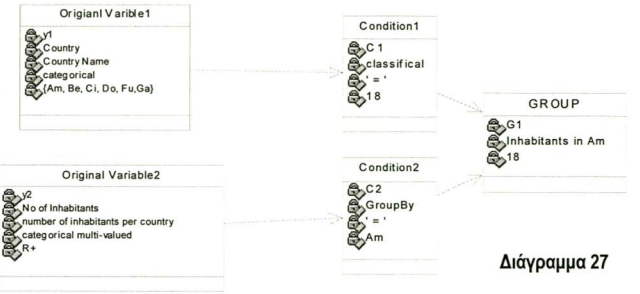
Αυτή η διαδικασία ομαδοποίησης δημιούργησε τα συμβολικά αντικείμενα (SO) και τις συμβολικές μεταβλητές (SVar) που παρουσιάζονται στο διάγραμμα 26

Σημείωση: Στο διάγραμμα 26 περιλαμβάνονται μόνο τα SO και SVar που έχουν νόημα σχετικό με τους περιορισμούς που εφαρμόστηκαν.



Διάγραμμα 26

Μπορούμε επίσης να δείξουμε πώς οι original variables σχετίζονται με το κάθε Group. Στο διάγραμμα 27 παρουσιάζεται αυτή η σχέση για το Group 1 του διαγράμματος 26 καθώς με τον ίδιο τρόπο δημιουργούνται και τα υπόλοιπα Group.



Διάγραμμα 27

6.5.2 Εφαρμογή 2

Έστω ένας πίνακας έχει τρεις Original Variables (οι οποίες αντιπροσωπεύουν τις στήλες του πίνακα) ως εξής:

y1= επίπεδο εκπαίδευσης (*educational level)*

y2 = επαγγελματική απασχόληση (*occupation*)

y3 = επαγγελματική κατάσταση *(status of employment)*

Οι μεταβλητές αυτές παίρνουν τιμές ως εξής:

y_1 = **Educational level** (κατά την ταξινόμηση ISCED):

1. Early childhood education
2. Primary education
3. Lower secondary education
4. Upper secondary education
5. Non-university higher education
6. University higher education
7. Graduate-professional higher education
8. Other

y_2 = **Occupation**

1. Legislators, senior officials and managers
2. Professionals
3. Technicians and associate professionals
4. Clerks
5. Service workers and shop and market sales workers

6. Skilled agricultural and fishery workers
7. Craft and related trades workers
8. Plant and machine operators and assemblers
9. Elementary occupations
10. Armed forces

y_3 = Status of employment:

1. Self employed w/out personnel
2. Self employed with personnel
3. Employee
4. Helper in family business

Έστω τώρα ότι η έρευνα πραγματοποιήθηκε σε 100 άτομα (individuals) βάση των μεταβλητών ομαδοποίησης Φύλο (sex) = { m, f } και της ηλικίας (age) η οποία παίρνει τιμές:

Age:

1. 15 – 25
2. 25 – 35
3. 35 – 45
4. 45 – 55
5. 55 – 65

Τότε, ***οκτώ symbolic objects*** δημιουργούνται χρησιμοποιώντας κατάλληλο σύστημα διαχείρισης συμβολικών δεδομένων (βλ. για παράδειγμα [Hebrail, 1996]) χρησιμοποιώντας τις μεταβλητές y_1, y_2, y_3 .

Όλα τα symbolic objects έχουν ομαδοποιηθεί κατά *sex* και *age*. Τα οκτώ symbolic objects παρουσιάζονται στον επόμενο πίνακα, ο οποίος έχει εξαχθεί από το κατάλληλο σύστημα.

	educ	occup	status
15-25/f	graduate (0.67), upper se (0.33)	Elementa (0.33), Professi (0.33), Technici (0.33)	employee (0.33), Helper i (0.33), self emp (0.33)
25-35/f	primary (1.00)	Clerks (1.00)	self emp (1.00)
35-45/f	nonunive (1.00)	Plant an (1.00)	self emp (1.00)
35-45/m	nonunive (0.33), primary (0.33), upper se (0.33)	Clerks (0.33), Service (0.33), Skilled (0.33)	Helper i (0.33), self emp (0.67)
45-55/f	graduate (0.17), lower se (0.33), nonunive (0.17), universi (0.17), upper se (0.17)	Plant an (0.17), Professi (0.17), Service (0.17), Skilled (0.33)	Helper i (0.17), self emp (0.33), self emp (0.50)
45-55/m	early ch (1.00)	Plant an (1.00)	employee (1.00)
55-65/f	lower se (0.50), primary (0.50)	Clerks (0.50), Plant an (0.50)	employee (1.00)
55-65/m	lower se (0.50), nonunive (0.50)	Clerks (0.50), Plant an (0.50)	employee (0.50), self emp (0.50)

Ο εξαγόμενος πίνακας δεν περιλαμβάνει κανένα μεταδεδομένο. Επομένως, η ερμηνεία του θα ήταν δύσκολη χωρίς τις αρχικές εξηγήσεις που δόθηκαν για τις εξεταζόμενες μεταβλητές και τους περιορισμούς που εφαρμόστηκαν.

Επίσης, έχει χαθεί η πληροφορία για τον αριθμό των ατόμων που συμμετείχαν στην έρευνα

Το μοντέλο μεταδεδομένων έχει κρατήσει όλες αυτές τις επιπλέον πληροφορίες που είναι απαραίτητες για την κατανόηση του πίνακα, ταυτόχρονα με τα δεδομένα. Αν το μοντέλο 'εμφυτευτεί' στη βάση συμβολικών δεδομένων του συστήματος που χρησιμοποιήθηκε, ο πίνακας θα μπορεί να παρουσιάσει τις επεξηγηματικές πληροφορίες για SO και SVars με ένα απλό κλικ στα αντίστοιχα κελιά του πίνακα όπως διευκρινίζονται στον ακόλουθο εμπλουτισμένο πίνακα.

SYMBOLIC OBJECT

Number of individuals: 100
OriginalVariables.classMembership:
Age, Sex
Operatortype: Aggregation
Condition type: numerical

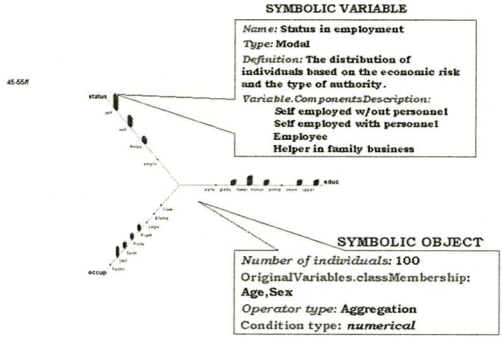
SYMBOLIC VARIABLE

Name: Status in employment
Type: Modal
Definition: The distribution of individuals based on the economic risk and the type of authority.
Variable.ComponentsDescription::
Self employed w/out personnel
Self employed with personnel
Employee
Helper in family business

	educ	occup	status
15-25/f	graduate (0.67), upper se (0.33)	Elements (0.33), Professi (0.33), Technici (0.33)	employee (0.33), Helper i (0.33), self emp (0.33)
25-35/f	primary (1.00)	Clerks (1.00)	self emp (1.00)
35-45/f	nonunive (1.00)	Plant an (1.00)	self emp (1.00)
35-45/m	nonunive (0.33), primary (0.33), upper se (0.33)	Clerks (0.33), Service (0.33), Skilled (0.33)	Helper i (0.33), self emp (0.67)
45-55/f	graduate (0.17), lower se (0.33), nonunive (0.17), universi (0.17), upper se (0.17)	Plant an (0.17), Professi (0.17), Service (0.17), Skilled (0.33)	Helper i (0.17), self emp (0.33), self emp (0.50)
45-55/m	early ch (1.00)	Plant an (1.00)	employee (1.00)
55-65/f	lower se (0.50), primary (0.50)	Clerks (0.50), Plant an (0.50)	employee (1.00)
55-65/m	lower se (0.50), nonunive (0.50)	Clerks (0.50), Plant an (0.50)	employee (0.50), self emp (0.50)

Σε αυτήν την περίπτωση, κάνοντας κλικ σε κάθε έναν από τους τίτλους στηλών, οι πληροφορίες για τη συγκεκριμένη αρχική μεταβλητή θα παρουσιαστούν (ορισμός, όνομα, τύπος, τιμές, κ.λπ..) σύμφωνα με τα χαρακτηριστικά / ιδιότητες της κλάσης "ORIGINAL VARIABLE" του μοντέλου μεταδεδομένων που προτείνεται στο κεφάλαιο αυτό. Ομοίως, κάνοντας κλικ στο πάνω αριστερό κουμπί, θα εμφανιστούν οι πληροφορίες για το συγκεκριμένο συμβολικό αντικείμενο σύμφωνα με τις ιδιότητες της κλάσης "SYMBOLIC OBJECT" του μοντέλου.

Αντίστοιχα, στην περίπτωση που παρουσιάζουμε γραφικά ένα συμβολικό αντικείμενο, για παράδειγμα με τη βοήθεια ενός Zoom Star (δες [Noirhomme, 2002]), τα μεταδεδομένα μπορούν να εμφανίζονται με ένα κλικ πάνω στο zoom star, τόσο για τα SO, όσο και για Svar, ή, αν θέλουμε και συγκεκριμένες πληροφορίες για το Survey (μεταδεδομένα για αρχικά δεδομένα) μπορούν επίσης να παρουσιάζονται σε ένα μικρό πίνακα κάτω από το zoom star.



Labor Force Survey, Greece, 1997 Survey Metadata

Relevant Unit:	Unit of Labor Force survey
Survey's frequency:	quarterly
Geographical coverage:	The whole country (Greece)
International comparability:	NACE REV.1, STEP-92 (ISCO-88), ISCED-97

ΚΕΦΑΛΑΙΟ 7

ΣΥΜΠΕΡΑΣΜΑΤΑ, ΠΡΟΤΑΣΕΙΣ ΚΑΙ ΠΡΟΟΠΤΙΚΕΣ ΜΕΛΛΟΝΤΙΚΗΣ ΈΡΕΥΝΑΣ

Αρκετοί φορείς δημόσιας διοίκησης και στατιστικές υπηρεσίες αντιμετωπίζουν δυσχέρειες από την έλλειψη τυποποίησης στις διαδικασίες και στην ανταλλαγή πληροφοριών μεταξύ σχετικών φορέων με τους οποίους συνεργάζονται. Επιπλέον, οι μέθοδοι συλλογής και επεξεργασίας πληροφοριών καθώς και το τεχνολογικό περιβάλλον (πληροφοριακά συστήματα και αντίστοιχες βάσεις δεδομένων) ποικίλλουν μεταξύ των οργανισμών και επιπλέον, η μεταπληροφορία αντιμετωπίζεται διαφορετικά, δημιουργώντας κατά συνέπεια προβλήματα στη συγκρισιμότητα των αποτελεσμάτων και τη συμβατότητα των διαδικασιών.

Τα αποτελέσματα της διατριβής μπορούν να επικεντρωθούν κυρίως στην προστιθέμενη αξία του μοντέλου μεταδεδομένων όσον αφορά την προτυποποίηση των διαδικασιών συλλογής, επεξεργασίας και διάχυσης της στατιστικής πληροφορίας, αποτελώντας έτσι ένα πολύτιμο εργαλείο για τους φορείς δημόσιας διοίκησης στην εξαγωγή ακριβέστερων, ποιοτικών στατιστικών αποτελεσμάτων και δεικτών οικονομικής πολιτικής, με χαμηλότερο κόστος και μειωμένο φόρτο εργασίας. Αποτελούν λοιπόν μία ολοκληρωμένη πρόταση προς τους εν λόγω φορείς και άλλους οργανισμούς επεξεργασίας στατιστικών δεδομένων για ποιοτικά και συγκρίσιμα αποτελέσματα.

Πιο αναλυτικά, η συμβολή του μοντέλου μεταδεδομένων στην προτυποποίηση (standardisation) πραγματοποιείται με δύο τρόπους: i) προτυποποίηση όσον αφορά τις διαδικασίες και την ανταλλαγή δεδομένων και ii) προτυποποίηση όσον αφορά το σχεδιασμό. Συγκεκριμένα:

7.1 Προτυποποίηση των διαδικασιών

Όσον αφορά την τυποποίηση των διαδικασιών και της ανταλλαγής δεδομένων, το προτεινόμενο μοντέλο μπορεί να θεωρηθεί ως ένα σημαντικό βήμα προς την τυποποίηση επειδή το μοντέλο:

- περιλαμβάνει περισσότερα από 150 μεταδεδομένα που επιλέγονται σύμφωνα με τις ανάγκες που εκφράζονται από τον ΟΟΣΑ (OECD) στον κατάλογο κύριων οικονομικών δεικτών (Main economic indicators [OECD, 1997]), τον IMF (SDDS standard), και ακολουθεί τους ορισμούς από τα Ηνωμένα Έθνη (UN terminology model) και Eurostat, επομένως γίνονται αποδεκτά ευρέως και χρησιμοποιούνται από τους εθνικούς και διεθνείς οργανισμούς.
- εξετάζει όλα τα κύρια στάδια της επεξεργασίας στατιστικών στοιχείων: συλλογή δεδομένων, επεξεργασία, ανάλυση και διάχυση αποτελεσμάτων.

- περιλαμβάνει μέρος της διαδικασίας εναρμόνισης που ελαχιστοποιεί τα προβλήματα συγκρισιμότητας κυρίως στον τομέα των μετασχηματισμών απεικόνισης των ταξινομήσεων και των μονάδων μέτρησης.
- κρατάει πληροφορία για την ποιότητα, παρέχοντας τη δυνατότητα να ελέγχεται η ποιότητα ταυτόχρονα με την επεξεργασία δεδομένων, με εφαρμογή ποιοτικών αλληλοσχετιζόμενων δεικτών (όπως παραδείγματος χάριν για να εξετάσουν τη σχέση μεταξύ της επικαιρότητας και της ακρίβειας, ή τη δυνατότητα πρόσβασης των πληροφοριών)
- περιλαμβάνει επτά μετασχηματισμούς/διαδικασίες (operations) που κρατούν την ιστορία επεξεργασίας κάθε χειρισμού συνόλου δεδομένων,
- επιτρέπει την παραγωγή νέων οικονομικών δεικτών, το οποίο είναι ένα χρήσιμο εργαλείο για τις Εθνικές Στατιστικές Υπηρεσίες και άλλους οργανισμούς δημόσιας διοίκησης και πολιτικής.

7.2 Προτυποποίηση σχεδιασμού

Το μοντέλο δημιουργήθηκε ακολουθώντας το Object Oriented paradigm με τη χρήση UML. Δεν υιοθετήσαμε το Entity-relational model (E-R) επειδή θέλαμε το μοντέλο να είναι ευέλικτο και προσαρμόσιμο στις μελλοντικές απαιτήσεις.

Το παρόν UML μοντέλο μπορεί στη συνέχεια να μετατραπεί σε ένα XML schema (eXtensible Markup Language), αν χρειαστεί. Εάν δημιουργούσαμε το μοντέλο σε XML από την αρχή δεν θα ήταν η ενδεδειγμένη λύση γιατί – από την πλευρά του χρήστη του μοντέλου, όπως οι διάφοροι φορείς στους οποίου απευθυνόμαστε - η UML υπερτερεί έναντι της XML επειδή:

- η παράσταση των διαδικασιών σε UML είναι σχηματικά κατανοητή χωρίς να είναι απαραίτητη η γνώση της γλώσσας μοντελοποίησης, κάτι που στην XML δεν είναι εφικτό. Οπότε, είναι ευκολότερο για τους χρήστες να κατανοήσουν τα μοντέλα που χρησιμοποιούν UML από να αντιμετωπίσουν όλες τις τεχνικές λεπτομέρειες ενός σχήματος XML.
- Αρκετοί οργανισμοί στην Ελλάδα δεν χρησιμοποιούν ακόμα XML αλλά ανταλλάσουν τις πληροφορίες και αποτελέσματα μέσω HTML ή απλά με ηλεκτρονικό ταχυδρομείο οπότε δεν έχουν επαρκή γνώση της XML.
- Η UML είναι μία πλούσια γλώσσα μοντελοποίησης επιτρέποντας δύο επίπεδα μοντελοποίησης (class – attribute), την προσθήκη operators και οι σχέσεις που διέπουν τις κλάσεις είναι εμφανής άμεσα.

Επιπλέον, ένα από τα κύρια πλεονεκτήματα μοντελοποίησης με UML είναι ότι το εννοιολογικό πρότυπο που αντιπροσωπεύεται από την UML παραμένει αμετάβλητο καθώς καθορίζονται οποιεσδήποτε νέες τεχνικές εφαρμογές. Το ίδιο εννοιολογικό πρότυπο μπορεί να χρησιμοποιηθεί για να παραγάγει διαφορετικά σχήματα XML και XMI (eXtensible Markup Interface). Ακόμα κι αν υιοθετηθεί μια νέα τεχνολογική

υποδομή, το αρχικό εννοιολογικό μοντέλο σε UML παραμένει το ίδιο, ή μπορεί να εμπλουτιστεί εύκολα.

Στην περίπτωση που η χρήση XML γίνει ευρέως χρησιμοποιούμενη από τους σχετικούς οργανισμούς, τότε προτείνουμε τα εξής βήματα ώστε να δημιουργηθεί ένα κατανοητό και εύχρηστο XML σχήμα:

- καθορισμός του εννοιολογικού μοντέλου σε UML.
- καθορισμός των κανόνων απεικόνισης από UML στο σχήμα XML.
- εφαρμογή αυτών των κανόνων από τα εργαλεία παραγωγής κώδικα.

7.3 Προτάσεις και προοπτικές μελλοντικής έρευνας

Στην ενότητα αυτή παρατηρούμε εν συντομία την τρέχουσα κατάσταση όπως αποτυπώνεται από την έρευνά μας για την ανάπτυξη του μοντέλου μεταδεδομένων και δίνονται κάποιες προτάσεις και συνιστώμενες ενέργειες:

Παρατήρηση 1: Οι ευρωπαϊκές χώρες είναι σε διαφορετικά στάδια τεχνολογικής προόδου. Παρατηρήθηκαν διαφορετικές εικόνες ως προς τη συλλογή δεδομένων στις Σκανδιναβικές χώρες, τα άλλα κράτη-μέλη της ΕΕ, τις χώρες που πρόσφατα έγιναν δεκτές καθώς και τις υπόλοιπες υπό ένταξη χώρες.

Πρόταση: Συστήνουμε την υιοθέτηση ενός πλαισίου XML/EDI (Electronic Data Intechange) για την ανταλλαγή των διαφορετικών τύπων στοιχείων τα οποία έχουν αποθηκευτεί σε ποικίλες βάσεις δεδομένων και αναλυθεί από διαφορετικά πληροφοριακά συστήματα. Συνιστούμε λοιπόν ως βάση το μοντέλο μεταδεδομένων σε UML που αναπτύχθηκε ως πλατφόρμα. Επιπλέον, θεωρούμε απαραίτητη μια ανάλυση απαιτήσεων, καθώς επίσης και η εξέταση των δυνατοτήτων κάθε νεοενταγμένου στην ΕΕ κράτους (καθώς και των υπό ένταξη).

Παρατήρηση 2: Η ετερογένεια στις μεθόδους και τις τεχνολογίες φαίνεται να επιδεινώνεται δεδομένης της διεύρυνσης της Ευρωπαϊκής Ένωσης. Η ποικιλία των πλαισίων νομοθεσίας και των μεθοδολογικών ιδιαιτεροτήτων μεταξύ των χωρών θα γίνει σύντομα αντιληπτή.

Πρόταση: Προκειμένου να έχουμε μια ακριβή άποψη σχετικά με την παρούσα κατάσταση όλων των ευρωπαϊκών χωρών από την άποψη της ποιότητας και συγκρισιμότητας των αποτελεσμάτων, μια έρευνα πρέπει να πραγματοποιηθεί στις ευρωπαϊκές χώρες και να περιλαμβάνει τις πληροφορίες για τυχόν ασυνέχειες στις χρονοσειρές, καθώς κάποια ένδειξη της δριμύτητάς της συνέπειας που είχε αυτή η ασυνέχεια. Στη συνέχεια, θα πρέπει να συνεχιστεί η έρευνα και η προσπάθεια για την εναρμόνιση των δεδομένων σε σχέση με τα αίτια των ασυνεχειών και τα προβλήματα που δημιούργησαν. Τα αποτελέσματα της παρούσης διατριβής σχετικά με την εναρμόνιση (κεφάλαιο 2) μπορούν αν χρησιμοποιηθούν αλλά τα νέα δεδομένα θα απαιτήσουν νέους μετασχηματισμούς και αξιολόγηση υπό διαφορετικό πρίσμα. Επιπρόσθετα, η σύγκριση των δεδομένων των νεοενταγμένων στην ΕΕ κρατών είναι

απαραίτητη για τη διαμόρφωση ευρωπαϊκής οικονομικής πολιτικής και την ανάπτυξη της διεθνούς συνεργασίας στην εποχή της παγκοσμιοποίησης.

Παρατήρηση 3: Έχουν γίνει προσπάθειες να μετρηθεί σε ευρωπαϊκό επίπεδο η ποιότητα των στατιστικών αποτελεσμάτων συγκρίνοντας τα αποτελέσματα που στέλνουν τα κράτη με διεθνή πρότυπα και κριτήρια, όπως βάσει του Special Data Dissemination Standard (SDDS) του IMF, ή των επτά κριτηρίων της Eurostat.

Πρόταση: Η έρευνα πάνω στους δείκτες διασφάλισης της ποιότητας όπως αυτά που δημιουργήθηκαν στο κεφάλαιο 2, παράγραφος 2.4 πρέπει να συνεχιστεί ώστε να εξαχθούν δείκτες ελέγχου της ποιότητας των αποτελεσμάτων, τόσο από την πλευρά του παραγωγού της πληροφορίας όσο και του χρήστη.

Παρατήρηση 4: Παρατηρείται έλλειψη κοινών προτύπων (standards)

Πρόταση: Χρειάζεται αρκετή έρευνα ακόμα για τη θέσπιση κοινών προτύπων στο ευρωπαϊκό στατιστικό σύστημα και είναι μια επιτακτική ανάγκη για τους διεθνείς οργανισμούς. Συνεπώς, στο παρόν και στο εγγύς μέλλον, τα αποτελέσματα από διαφορετικές πηγές θα παρουσιάζουν ασυνέχειες και πρέπει τουλάχιστον να βρεθεί ένας τρόπος να 'ποσοτικοποιηθεί' το πρόβλημα που δημιουργείται. Σήμερα η συνήθης πρακτική είναι να παρέχεται, μαζί με τη δημοσίευση των αποτελεσμάτων, και μια συνήθως ογκώδης μεθοδολογική δημοσίευση που περιέχει τα κατάλληλα μεταδεδομένα. Ο χρήστης που θέλει να κάνει μια σύγκριση πρέπει να διαβάσει αυτή τη μεθοδολογική δημοσίευση, να κάνει μια αξιολόγηση του επιπέδου εναρμόνισης και να χρησιμοποιήσει αναλόγως τα παρεχόμενα στοιχεία. Η υιοθέτηση ενός μοντέλου μεταδεδομένων όπως αυτό που δημιουργήθηκε στη διατριβή αυτή αποτελεί ένα βήμα προς την αναγκαία αυτή τυποποίηση. Ο εμπλουτισμός ενός τέτοιου μοντέλου με περισσότερους μετασχηματισμούς και στοιχεία ποιότητας θα μπορούσε να διευκολύνει τον χρήστη στην αξιολόγηση των στατιστικών αποτελεσμάτων.

Παρατήρηση 5: Το κόστος σε υλικοτεχνική υποδομή και σε ανθρώπινο δυναμικό του φορέα για τη συλλογή όλων των μεταδεδομένων που ενδεχομένως να ζητηθούν είναι τεράστια

Πρόταση: Η ολοκλήρωση ενός μοντέλου μεταδεδομένων όπως αυτό που δημιουργήθηκε θα ελαχιστοποιούσε το φορτίο των εργαζομένων, θα μείωνε τις δαπάνες και θα βοηθούσε στην κατανόηση και σωστή χρήση των στατιστικών μεταβλητών που χρησιμοποιούνται από δευτερεύουσες πηγές στοιχείων. Τα συλλεχθέντα μεταδεδομένα θα είναι τα ίδια με τα μεταδεδομένα που θα ζητηθούν από όλες τις υπηρεσίες. Δεδομένου ότι τα μεταδεδομένα θα υποβληθούν σε επεξεργασία ταυτόχρονα με τα στοιχεία, θα επιτευχθεί από την υπηρεσία η αυτόματη ανάκτηση των δεδομένων μαζί με τα μεταδεδομένα.

Εντούτοις, υπάρχουν ακόμα αρκετές προοπτικές για μελλοντική έρευνα:

Ποιότητα: Αυτή τη στιγμή τα μεταδεδομένα που αναφέρονται στις ποιοτικές πληροφορίες των αποτελεσμάτων και των διαδικασιών περιορίζονται στις διορθώσεις και την ακρίβεια λαθών κατά τη διάρκεια των διαδικασιών, καθώς επίσης και τη δυνατότητα πρόσβασης και τη σαφήνεια / επάρκεια των πληροφοριών. Αρκετές άλλες συνιστώσες της ποιότητας χρειάζεται να μελετηθούν καθώς και οι περιπτώσεις αλληλοεπίδρασης των διαφόρων συνιστωσών της ποιότητας (επέκταση αποτελεσμάτων κεφαλαίου 2, ενότητα 2.4).

Πρότυπα (standards): η ύπαρξη προτύπων είναι βασική για την ανταλλαγή των στοιχείων, εντούτοις, αυτή τη στιγμή δεν υπάρχουν γενικά αποδεκτά πρότυπα για τη συλλογή και ανταλλαγή δεδομένων. Υπάρχουν πολλά πρότυπα για διαφορετικούς λόγους, αλλά η ύπαρξη πολλών προτύπων σημαίνει ότι δεν υπάρχει κανένα γενικό αποδεκτό πρότυπο. Τα 'standards' που εξετάζονται στο μοντέλο αναφέρονται μόνο στις ταξινομήσεις και τους δείκτες, καθώς επίσης και την τυποποιημένη ανταλλαγή της μεταπληροφορίας. Εκτενέστερη έρευνα είναι αναγκαία.

Κοινά αποδεκτή ορολογία: λόγω της πολυπλοκότητας των μεθόδων και της υπάρχουσας εμπειρίας στις διαφορετικές χώρες, δεν υπάρχει μία πρότυπη ορολογία των μεταδεδομένων. Το μοντέλο της διατριβής ακολουθεί την ορολογία του OECD, και όπου χρειάστηκε της Eurostat και UN. Εντούτοις, παρόλες τις διαφορετικές ορολογίες που έχουν δημιουργηθεί από διεθνείς οργανισμούς μια ενοποιημένη προσέγγιση καθυστερεί ακόμα.

Χρήστες: οι χρήστες, ανάλογα με την κατηγορία τους, έχουν διαφορετική αντίληψη και απαιτήσεις σχετικά με τα μεταδεδομένα που χρειάζονται. Τα αιτήματά τους πρέπει να εξεταστούν περισσότερο από τους φορείς συλλογής δεδομένων και η σχετικότητα των πληροφοριών μπορεί να τυποποιηθεί σύμφωνα με τα είδη χρηστών.

Χρήση δευτερεύουσας πηγή στοιχείων: είναι βασικό να μειωθεί το κόστος συλλογής δεδομένων με τη χρησιμοποίηση των υπαρχουσών δευτερευουσών πηγών στοιχείων. Η έλλειψη ενοποιημένων μεταδεδομένων είναι το κύριο εμπόδιο για τη δευτεροβάθμια συλλογή δεδομένων, μια ανεπάρκεια που προσπαθήσαμε να ελαχιστοποιήσουμε με το αναπτυγμένο κοινό μοντέλο.

Εργαλεία/συστήματα: το προσωπικό των στατιστικών οργανισμών απαιτεί απλά και εύχρηστα εργαλεία/ συστήματα για να παραγάγουν μεταδεδομένα κατά τη διάρκεια της διαδικασίας συλλογής δεδομένων. Μία ανάλυση κόστους-χρησιμότητας είναι απαραίτητη για τις υπάρχουσες τεχνολογικές δομές.

Συμπερασματικά, τα αποτελέσματα της παρούσης διατριβής αποτελούν ένα βήμα προς την προτυποποίηση και την δομημένη παρουσίαση και χρήση των μεταδεδομένων, καθώς και την εναρμόνιση των αποτελεσμάτων. Η συνέχιση της έρευνας σε αυτούς τομείς καθώς και στην διασφάλιση της ποιότητας αποτελεί ιδιαίτερη πρόκληση για τους μελετητές της περιοχής και κρίνεται απαραίτητη για τους διεθνείς οργανισμούς και τους φορείς ανάλυση στατιστικής πληροφορίας.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] **Agosta L. (2000).** "The Essential Guide to Data Warehousing", *Prentice Hall*, UK.

[2] **Beaumont J.-F., Haziza D., Mitchell C. & Rancourt E. (2003),** "New tools at Statistics Canada to Measure and Evaluate the Impact of Nonresponse and Imputation", presented at the Research Conference of Federal Committee on Statistical Methodology.

[3] **Billard L. and Diday E. (2003)**, "From the Statistics of Data to the Statistics of knowledge: Symbolic Data Analysis", *Journal of the American Statistical Association (JASA),* 98(462), 470-487.

[4] **Bock H.-H and Diday E. (2000).** *"Analysis of Symbolic Data", Springer*, Heidelberg.

[5] **Carson S. (2000),** *Toward a Framework for Assessing Data Quality,* IMF, www.imf.org

[6] **Chen P. P. (1976).** "The Entity-Relationship model: Toward a unied view of data". *ACM Transactions on Database Systems*, 1(1):9-36.

[7] **Date, C.J. (1990).** "An Introduction to Database Systems", Volume I, 5th Ed., *Addison-Wesley*, ISBN 0-201-52878-9.

[8] **De Jong A. (2003).** "IMPECT: Recent Developments in Harmonized Processing and Selective Editing", *Conference of European Statisticians, Work Session on Statistical Data Editing,* Madrid. Available at: http://www.unece.org/stats/documents/2003/10/sde/wp.31.e.pdf

[9] **Denk M. & Froeschl K.A (2000).** „The IDARESA Data Mediation Architecture for Statistical Aggregates", *Research in Official Statistics (ROS),* Vol 3 No 1, pp. 7-38.

[10] **Denk M., Froeschl K.A & Grossmann W. (2002).** "Statistical Composites: A Transformation-bound Representation of Statistical Datasets", *Fourteenth International Conference on Scientific and Statistical Database Management, SSDBM'02.* Edinburgh, UK, 217-226, IEEE Computer Society.

[11] **Depoutot Raoul, Arondel Philippe and Linden, H. (1998)** "Quality Aspects of International Statistics" GSS Conference, London. Available at: http://europa.eu.int/en/comm/eurostat/research/quality/documents/international_quality.pdf

[12] **Depoutot Raoul and Arondel Philippe (1998).**"International Comparability and Quality of Statistics" *CAED97 International Conference on Comparability Analysis of Enterprise (micro) Data*, Bergamo, Italy. Available at: http://europa.eu.int/en/comm/eurostat/research/quality/documents/international_comparability.pdf

[13] **Deutsch P., Emtage A., Koster M. and Stumpf M. (1995).** "Publishing Information on the Internet with Anonymous FTP (IAFA Templates)", *IETF IAFA WG,* Internet Draft at: ftp://nic.merit.edu/documents/internet-drafts/draft-ietf-iiir-publishing-03.txt

[14] **De Vries, W. (1999).** "Ranking: right or wrong? Some problems in comparing national statistical offices and systems". *Netherlands Official Statistics*, Statistics Netherlands, Netherlands, 14, 4-6.

[15] **De Waal T. (2000).** "A brief overview of imputation methods applied at Statistics Netherlands", Netherlands Official Statistics 2000-3.

[16] **De Waal T. & Pannekoek J. (2004).** "Automatic Edit and Imputation for Business Surveys", *European Conference on Quality and Methodology in Official Statistics*, Mainz, Germany.

[17] **De Waal T. & Quere R. (2003).** "A Fast and Simple Algorithm for Automatic Editing of Mixed Data", *Journal of Official Statistics (JOS),* 19, pp. 383-402.

[18] **Diday E. (1991).** "Des Objets de l' analyse des donnees a ceux de l' analyse des connaissances", *Induction symbolique numerique a partir de donnees.* Diday E and Kodratoff Y. (eds.), Vol I. Cepadues, Toulouse, 9-76.

[19] **Diday E., Lechevallier Y., Opitz O., (1995).** "Ordinal and Symbolic Data Analysis". Conference on *Ordinal and Symbolic Data Analysis (OSDA 95),* Paris. Springer Verlag, Heidelberg-Berlin, 1996.

[20] **Elvers, E. (1998).** "Other Aspects of Quality: Comparability and Coherence", *Model quality report in business statistics*, Davies, P. and Smith, P. (eds), Eurostat, Final Deliverable of SUPCOM97 Lot.

[21] **EUROSTAT (1993).** NACE Rev1, *EEC Official Journal* L 83.

[22] **EUROSTAT (1999).** 'Inventory of International Statistical Classifications', *Statistical Office of the European Communities*, Luxembourg, ISBN 92-828-8204-7.

[23] **EUROSTAT (2000a),** "The metadata problem in a European Context, Workshop on Statistical Metadata", Luxembourg, Working paper No: 1.1.

[24] **Eurostat, (2000b).** "Definition of quality in statistics". Available at: http://www.forum.europa.eu.int/Public/irc/dsis/qis/library?l=/public&vm=detailed&sb=Title

[25] **Eurostat, (2000c).** "Standard quality report." Available in electronic form at: http://www.forum.europa.eu.int/Public/irc/dsis/qis/library?l=/public&vm=detailed&sb=Title.

[26] **Eurostat, (2000d).** "Glossary on quality in statistics", Doc. Eurostat/A4/Quality/00/ General/Glossary.

[27] **Froeschl, K.A. (1997).** "Metadata Management in Statistical Information Processing", Wien: *Springer*, ISBN 3-211-82987-3.

[28] **Froeschl, K.A. (1999).** "Metadata Management in Official Statistics – An IT-Based Methodology Approach", *Austrian Journal of Statistics,* Vol 28 No2, pp. 49–79

[29] **Froeschl K.A and Grossmann W.(2001).** "Deciding Statistical Data Quality", *NTTS-ETK Conference*, Vol 1, pp. 567-575, European Communities, ISBN 92-894-1176-7.

[30] **Froeschl K.A and Grossmann W., (2000).** "The Role of Metadata in Using Administrative Sources", *Research in Official Statistics (ROS),* 3(1), 65 – 82.

[31] **Froeschl K.A., Yamada T. & Kudrna R. (2002).** "Industrial Statistics Revisited: From Footnotes to Meta-Information Management", *Austrian Journal of Statistics,* 31(1), 9–34.

[32] **Ghosh, S.P. (1988).** "Statistics Metadata". *Encyclopedia of Statistical Sciences*, 8, 743-746, S. Kotz, N. L. Johnson & C. B. Read, Eds., John Wiley and Sons, New York.

[33] **Grossmann, W., Froeschl, K. & Walk, M. (1998)**. "The Idaresa Data Model – Final Version", IDARESA, The Milestone IV Package, Second Part.,

[34] **Grossmann, W. and Papageorgiou, H., (1997).** "Data and Metadata Representation of Highly Aggregated Economic Time-Series". 51st Session of the International Statistical Institute, Contributed Papers, 2, 485-486.

[35] **Grossmann, W.(1999).** "Metadata*", Encyclopedia of Statistical Sciences*, S. Kotz, Editor-in-Chief, John Wiley and Sons, New York, update Vol 3, pp. 811-815.

[36] **Hand D.J. (1993).** "Data, metadata, and information". *Statistical Journal of the United Nations*, 10, 143–151.

[37] **Hatzopoulos, M., Karali, I. & Viglas, E. (1998).** "Attacking Diversity in NSIs' Storage Infrastructure: the ADDSIA Approach". *NTTS '98*, Sorrento, Italy, 229-234,

[38] **Hebrail G. (1996)** "SODAS (Symbolic Official Data Analysis System)". *Conference of the International Federation of Classification Societies (IFCS'96),* Kobe, Japan. Springer Verlag.

[39] **Hoffmann,E. (1999).** "Standard Statistical Classifications: Basic Principles", *Statistical Commission*, 30th session, New York.

[40] **INSEE (1999).** "Nomenclatures d'activités et de produits françaises.NAF-CPF", *Les éditions des Journaux officiels*, 1402, 1- 73.

[41] **IPIS consortium (2001).** 'Detailed Design of the Pilot Prototype System', Deliverable D7.1.

[42] **Karge, R. (1998).** "'Integrated Metadata-Systems within Statistical Offices". Tenth International Conference on Scientific and Statistical Database Management *(SSDBM 98),* Capri, Italy, 216-219, IEEE Computer Society.

[43] **Kent, J-P. & Schuerhoff, M.(1997).** "Some Thoughts About a Metadata Management System", *Ninth International Conference on Scientific and Statistical Database Management (SSDBM 97)*, Olympia, Washington, 174-185, IEEE Computer Society.

[44] **Laaksonen S. (2002).** "Traditional and new techniques for imputation", *Statistics in Transition Journal of the Polish Statistical Association*, 5 (6), 1013-1035.

[45] **Layzell, P. and Loucopoulos, P. (1989).** "System Analysis and Development", *Chartwell-Bratt,* 3rd Edition, ISBN 0-86238-215-7.

[46] **Lenz, H.-J. & Shoshani, A. (1997).** "Summarizability in OLAP and Statistical Databases", *Ninth International Conference on Scientific and Statistical Database Management (SSDBM 97),* Olympia Washington, 132-143, IEEE Computer Society.

[47] **Malvestuto, F.M. (1993).** "A Universal-Scheme Approach to Statistical Databases Containing Homogeneous Summary Tables". *ACM Transactions on Database Systems*, 18, 678-708.

[48] **Marcelo C. (2002),** "Quality Control in the Portuguese Labor Force Survey", *e-Journal of Symbolic Data Analysis (JSDA),* Vol 0, 40-46

[49] **Marsh, C., Dale, A. & Skinner, C. (1994).** "Safe Data versus Safe Settings: Access to Microdata from the British Census". *International Statistical Review,* 62, 35-53.

[50] **Michiels J. & Hacking W. (2004).** "Computer assisted coding by interviewers", *European Conference on Quality and Methodology in Official Statistics*, Mainz, Germany.

[51] **Muller R., Stohr Th. and Rahm E.(1999).** "An Integrative and Uniform Model for Metadata Management in Data Warehousing Environments", *International Workshop on design and Management of data Warehouses (DMDW'99),* Heidelberg, Germany.

[52] **Neuchatel Group (2000).** "'The Neuchatel Terminology: Classification database object types and their attributes", *UN/ECE Work Session on Statistical Metadata*, Working Paper no 10, Geneva, Switzerland.

[53] **Noirhomme- Fraiture M. (2002).** "Visualisation of Large Datasets: The Zoom Star solution", e-Journal of Symbolic Data Analysis (JSDA), 0, 26-39

[54] **Noirhomme- Fraiture M. and Rouad M. (1997).** "Computer Graphics for Symbolic Objects", *ASMDA 97*, Naples.

[55] **Nordbotten S. (2000).** "Evaluating efficiency of statistical data editing: General framework", *Conference of European Statisticians*, Methodological material, United Nations, Geneva.

[56] **Organisation for Economic Cooperation and Development – OECD (1999).** "Metadata for International Comparisons and Adherence to International Standards", *UN/ECE Work Session on Statistical Metadata,* Working paper No 3.

[57] **OECD (1997).** "Main Economic Indicators, Sources and Methods, Labour and Wage Statistics",*OECD*, Statistics Directorate.

[58] **OECD (1998).** "Functional Classifications of the 1993 SNA, COICOP, COPNI, COFOG, COPP", OECD Statistics Directorate, *OECD Meeting of National Accounts Experts*, Paris.

[59] **Olenski J.(1996).** "Practical problems of implementing metadata standards in official statistics", *Eighth International Conference on Scientific and Statistical Database Management (SSDBM 96),* Stockholm, Sweden, 130-147, IEEE Computer Society.

[60] **OMG (2002).** "Unified Modeling Language" at http://www.omg.org/uml/

[61] **Ozsoyoglu, G., Matos, V. & Ozsoyoglu, Z. M.(1989).** "Query Processing Techniques in the Summary-Table-by-Example Database Query Language", *ACM Transactions on Database Systems*, 14, 526-573.

[62] **Papageorgiou H., Artikis G. and Vardaki M. (1999a).** "On updating highly aggregated economic time series", *Wiadomości Statystyczne,* 46(4), 4-11.

[63] **Papageorgiou, H. Vardaki, M. & Pentaris, F. (1999b).** "Quality of Statistical Metadata", *Research in Official Statistics (ROS),* 2(1), 45-57.

[64] **Papageorgiou, H., Vardaki, M. & Pentaris, F. (2000a).** "Data and Metadata Transformations", *Research in Official Statistics (ROS*), 3(2), 27-43.

[65] **Papageorgiou, H., Vardaki, M. & Pentaris, F. (2000b)** "Recent Advances on Metadata", *Computational Statistics,* 15(1), 89-97.

[66] **Papageorgiou, H., Pentaris, F., Theodorou E., Vardaki M. & Petrakos M. (2001a).** "Modelling Statistical Metadata", *Thirteenth International Conference on Scientific and Statistical Database Management (SSDBM 01)*, Virginia, USA, 25-35, IEEE Computer Society.

[67] **Papageorgiou, H., Pentaris, F., Theodorou E., Vardaki M. & Petrakos M. (2001b).** "A statistical metadata model for simultaneous manipulation of data and metadata", *Journal of Intelligent Information Systems (JIIS),* 17(2/3), 169-192.

[68] **Papageorgiou, H., Petrakos M, Vardaki M., Theodorou E. & Pentaris, F. (2001c)** "Metadata based Assessment of the level of fragmentation of Data

Series and Multisource Statistical Tables" *NTTS-ETK 2001 Conference*, Crete, Vol 1, pp.263- 272. Published by Eurostat.

[69] Papageorgiou, H., Vardaki M., Petrakos M, Theodorou E. & Pentaris, F. (2001d). "Harmonisation of Economic Classifications and related Transformations", *NTTS-ETK 2001 Conference*, Crete, 1, 345-354. Published by Eurostat.

[70] **Papageorgiou, H., Vardaki M., Theodorou E. and Pentaris, F. (2002).** "The use of Statistical Metadata Modelling and related transformations to assess the quality of statistical reports", Invited paper, *Joint UNECE/Eurostat Seminar on Integrated Statistical Information Systems and Related Matters (ISIS 2002),* Geneva, Switzerland. Available at www.unece.org/stats/documents/ces/sem.47/24.e.pdf

[71] **Papageorgiou, H. and Vardaki M., (2004).** "A Statistical Metadata Model for Symbolic Objects", to appear in the forthcoming book of *"Symbolic Data Analysis and the SODAS Software'*, Diday & Noirhomme Eds., Wiley.

[72] **Papazoglou M.P., Spaccapietra S. & Tari Z. (2000).** "Advances in Object-Oriented Data Modeling". Cambridge, USA: *The MIT Press,* ISBN 0-262-16189-3.

[73] **Pedersen D., Riis K. & Pedersen T.B. (2002).** "A powerful and SQL-compatible data model and query language for OLAP", *Thirteenth Australasian conference on Database technologies,* Melbourne, Victoria, Australia, 121-130.

[74] **Pentaris F. and Vardaki M. (1998).** "Statistical Metadata", *HERCMA 98,* ΑΣΟΕΕ, Αθήνα, 1, 1022-1026

[75] **Petrakos G., Conversano C., Farmakis G., Mola F., Siciliano R. & Stavropoulos P. (2004).** "New ways of specifying data edits", *Journal of the Royal Statistical Society,* A167, 1-26.

[76] **Poole J., Chang D., Tolbert D., Mellor D. (2002).** "Common warehouse metamodel - an introduction to the standard for data warehouse integration", *Wiley*, New York.

[77] **Scotney B., Dunne J., & McClean S.(2002).** "Statistical database modeling and compatibility for processing and publication in a distributed environment", *Research in Official Statistics (ROS),* 5(1), 5-18.

[78] **Shoshani Arie (2003).** "Multidimensionality in statistical, OLAP, and scientific databases". Multidimensional Databases: Problems and Solutions. Maurizio Rafanelli (ed.) Idea Group, 46-68.

[79] **Statistics Canada (1998).** "Quality Guidelines, v.3", http://www.statcan.ca

[80] **Stephan V., Hebrail G. and Lechevallier Y. (1997)**. "Improving Symbolic Descriptions of sets of individuals: the reduction of assertions". *Applied Stochastic Models and Data Analysis Conference,* Capri, Italy.

[81] **Stonebraker, M. (editor) (1994).** "Readings in Database Systems", Second Edition, *Morgan Kaufmann*, ISBN 1-55860-252-6.

[82] **Sundgren, B., (1996).** "Making Statistical Data More Available", *International Statistical Review,* 64, 23-38

[83] **Sundgren, B., (1999).** "Information Systems Architecture for National And International Statistical Offices Guidelines And Recommendations", *Conference Of European Statisticians, UNECE, Statistical Standards and Studies*, paper No. 51, UN, Geneva.

[84] **Sundgren B., (2000).** "The Swedish Statistical Metadata System", *ECE Workshop on Statistical Metadata,* Working Paper 1.6, Luxembourg.

[85] **Sundgren B. (2004).** "Documentation templates and metadata models at Statistics Sweden", *Metadata Working Group 2004,* Luxembourg.

[86] **United Nations (2000).** "International Economic and Social Classifications", Sales no E/CN.3/2000/17.

[87] **Vardaki M. & Papageorgiou H. (2004).** "An Integrated Metadata Model for Statistical Data Collection and Processing", *Sixteenth International Conference on Scientific and Statistical Database Management (SSDBM 04)*, Santorini, Greece, 363-372, IEEE Computer Society.

[88] **Vardaki M. (2004a).** "Statistical Metadata in Data Processing and Interchange". Invited paper for the Encyclopedia of Data Warehousing and Mining, John Wang (Ed), IDEA Group publishing, USA. To appear April 2005.

[89] **Vardaki M. (2004b).** "Metadata for Symbolic Objects". *e-Journal of Symbolic Data Analysis (JSDA)*, Vol 2(1), 1-8.

[90] **Verde R., Lechevallier Y. And Chavent M (2003),** "Symbolic clustering interpretation and visualization", e-Journal of Symbolic Data Analysis (JSDA),1, 24-31.

[91] **Westlake Andrew (1997).** "A simple Structure for Statistical Meta-Data", *Tenth International Conference on Scientific and Statistical Database Management (SSDBM),* Capri, Italy, 186-195, IEEE Computer Society.