2.02

ΣΗΜΑΣΙΟΛΟΓΙΚΗ ΟΛΟΚΛΗΡΩΣΗ ΔΕΔΟΜΕΝΩΝ ΜΕ ΤΗ ΧΡΗΣΗ ΟΝΤΟΛΟΓΙΩΝ

Λίνα Μπουντούρη, Μανόλης Γεργατσούλης

Ιόνιο Πανεπιστήμιο, Τμήμα Αρχειονομίας–ΒιβλιοθηκονομίαςΠαλαιά Ανάκτορα, Πλ. Ελευθερίας, 491 00, Κέρκυρα
boudouri@ionio.gr, manolis@ionio.gr

Περίληψη: Η δημιουργία και η ταχεία εξέλιξη του Διαδικτύου οδήγησε στη διάθεση ποικίλλων δεδομένων και στην παροχή πρόσβασης σε μεγάλο αριθμό χρηστών. Ανάλογη με την εξέλιξη του Διαδικτύου είναι και η αύξηση του περιεχομένου που διατίθεται δικτυακά, το οποίο είναι συχνά ετερογενές, κυρίως επειδή προορίζεται για να καλύψει, είτε ευρείες ανάγκες πλήθους χρηστών, είτε εξειδικευμένες ανάγκες κοινοτήτων χρηστών. Σε κάθε περίπτωση, ο στόχος της αποτελεσματικής αναζήτησης και ανάκτησης δεδομένων που προέρχονται από ετερογενείς πηγές είναι δύσκολο να επιτευχθεί.

Με γνώμονα την επίτευξη του παραπάνω στόχου, το επιστημονικό πεδίο της Ολοκλήρωσης Δεδομένων (Data Integration) μελετά τη δημιουργία συστημάτων τα οποία παρέχουν δυνατότητες αναζήτησης και ανάκτησης από συλλογές αυτόνομων και ετερογενών πηγών δεδομένων σαν αυτές να αποτελούν ένα ενιαίο σύνολο. Με άλλα λόγια, μέσα από ένα σύστημα ολοκλήρωσης δεδομένων δίνεται η δυνατότητα στο χρήστη με ένα μόνο ερώτημα να ανακτήσει αποτελέσματα από διαφορετικές πηγές.

Τα τελευταία χρόνια, στα πλαίσια της δημιουργίας του Σημασιολογικού Ιστού (Semantic Web) και της ανάπτυξης συστημάτων ολοκλήρωσης δεδομένων, έχει δοθεί ιδιαίτερη έμφαση στην αντιμετώπιση θεμάτων Σημασιολογικής (ή Εννοιολογικής) Ολοκλήρωσης των δεδομένων (Semantic Integration). Η σημασιολογική ολοκλήρωση δεδομένων καλείται να αντιμετωπίσει προβλήματα σημασιολογικής ετερογένειας (semantic heterogeneity), τα οποία προκύπτουν από τη χρήση διαφορετικών όρων για την αποτύπωση της ίδιας έννοιας, τόσο σε επίπεδο σχημάτων κωδικοποίησης δεδομένων (schema level), όσο και στο επίπεδο των ίδιων των δεδομένων (data level).

Για την επίτευξη της σημασιολογικής ολοκλήρωσης, σημαντικός θεωρείται ο ρόλος των τεχνολογιών του σημασιολογικού ιστού, όπως είναι οι οντολογίες. Οι οντολογίες εκφράζουν αναλυτικούς προσδιορισμούς εννοιών και μπορούν, κατά συνέπεια, να λειτουργήσουν ως φορέας διαλειτουργικότητας της πληροφορίας ανάμεσα σε διάφορα πληροφοριακά συστήματα.

Η παρούσα εργασία περιλαμβάνει μία επισκόπηση του προβλήματος της σημασιολογικής ολοκλήρωσης δεδομένων δίνοντας έμφαση στις τεχνικές ολοκλήρωσης δεδομένων με χρήση οντολογιών. Στη συνέχεια, η εργασία παρουσιάζει ανοικτά ερευνητικά προβλήματα του συγκεκριμένου επιστημονικού πεδίου και προσεγγίσεις της ερευνητικής μας ομάδας για την επίλυσή τους.

Λέξεις κλειδιά: Ολοκλήρωση δεδομένων, σημασιολογική ολοκλήρωση, οντολογίες.

ONTOLOGY – BASED SEMANTIC INTEGRATION

2.02

Lina Bountouri, Manolis Gergatsoulis

Department of Archives and Library Science, Ionian University
Plateia Eleftherias, Palea Anaktora, 491 00, Corfu, Greece
boudouri@ionio.gr, manolis@ionio.gr

Abstract: The creation and the advance development of the Internet have led to the dissemination of various data sources and the enhanced access of users. Analogous to the development of the Internet is the growing number of the content that is disseminated over the web. This content is –most of the times– heterogeneous due to the fact that it is aiming to fulfil extremely "wide" or "sophisticated" informational needs of communities of users. In all cases, the main goal of efficient search and retrieval that come from heterogeneous data sources is hard to achieve.

In order to cope with this issue, the scientific field of Data Integration studies the design of information systems, which provide searching and retrieving of data from collection of autonomous and heterogeneous data sources as if these sources constitute –as a whole– a single data source. In other words, via a data integration system, the user can execute only one query and retrieve results from different data sources.

Nowadays, due to the development of the Semantic Web and data integration systems, there has been a great interest to solve Semantic Integration issues. The field of semantic integration is oriented to solve semantic heterogeneity problems, which emerge as a result of using different terms in order to state the same concept, both in the schema level and in the data level.

The role of semantic web technologies, such as ontologies, is of crucial importance so as to semantically integrate data. Given that ontologies are defined as an explicit specification of a conceptualization, they can be used as a tool to promote interoperability among different information systems.

The particular research effort presents an overview of the semantic integration problem, emphasizing in data integration technologies using ontologies. Moreover, our research presents open issues of the particular field and approaches of our team to meet them.

Keywords: Data integration, semantic integration, ontologies

1. ΕΙΣΑΓΩΓΗ

Ένας συνεχώς αυξανόμενος αριθμός δεδομένων διατίθενται μέσω του διαδικτύου λόγω της ευρείας εξάπλωσης και χρήσης του. Ο μεγάλος αριθμός δεδομένων προκύπτει ως ένα βαθμό από την προσπάθεια να καλυφθούν οι πληροφοριακές ανάγκες ποικίλων ομάδων χρηστών. Με γνώμονα την κάλυψη των συγκεκριμένων αναγκών έχουν δημιουργηθεί πληροφοριακά συστήματα, τα οποία χαρακτηρίζονται από μεγάλο βαθμό αυτονομίας σε ποικίλα επίπεδα, όπως είναι οι διαφορετικές δυνατότητες αναζήτησης–ανάκτησης δεδομένων σε κάθε ένα από αυτά.

Η αυτονομία στο σχεδιασμό των πληροφοριακών συστημάτων οδηγεί στην εμφάνιση ετερογένειας σε τέσσερα διαφορετικά επίπεδα:

- **ετερογένεια συστημάτων (system heterogeneity)**: Προκύπτει από τη χρήση διαφορετικών πλατφόρμων υλικού, λειτουργικών συστημάτων, πρωτοκόλλων δικτύου κτλ.
- **ετερογένεια στη σύνταξη (syntactic heterogeneity)**: Διαφορές σε κωδικοποίηση, γλώσσες επερωτήσεων, πρωτόκολλα επικοινωνίας, "formats" δεδομένων κτλ.
- **ετερογένεια σχημάτων (schematic heterogeneity)**: Αποτέλεσμα της χρήσης διαφορετικών μοντέλων δεδομένων, δομών δεδομένων και σχημάτων κωδικοποίησής τους ανάμεσα σε πηγές.
- **σημασιολογική ετερογένεια (semantic heterogeneity)**: Παράγεται από τις σημασιολογικές αντιθέσεις, οι οποίες προκύπτουν όταν η σημασία των δεδομένων μπορεί να εκφραστεί με διαφορετικούς τρόπους και με ποικίλες ερμηνείες.

Αποτέλεσμα των όσων προαναφέρθηκαν είναι η ενιαία αναζήτηση και ανάκτηση δεδομένων να γίνεται εξαιρετικά δύσκολη από τους χρήστες, οι οποίοι δεν αναγνωρίζουν τις διαφορές και τα επίπεδα ετερογένειας των πληροφοριακών συστημάτων. Αντίθετα, ενδιαφέρονται για "ενοποιημένες διαδρομές" αναζήτησης και ανάκτησης σε ποικίλες πηγές με στόχο να καλύψουν τις πληροφοριακές τους ανάγκες.

Τα *συστήματα ολοκλήρωσης δεδομένων* (data integration systems) έρχονται να καλύψουν τις συγκεκριμένες ανάγκες, παρέχοντας στο χρήστη πρόσβαση σε συλλογή αυτόνομων πηγών, σαν αυτές να αποτελούν ως σύνολο μία πηγή δεδομένων (Koffina, Serfiotis και Christophides 2004). Το επιστημονικό πεδίο της Ολοκλήρωσης Δεδομένων ασχολείται με το πρόβλημα του συνδυασμού δεδομένων που προέρχονται από διαφορετικές πηγές, διαφορετικά συστήματα, περιγράφονται και μοντελοποιούνται από διαφορετικά σχήματα, και την ενοποίησή τους σε ένα ομογενοποιημένο σύνολο αποτελεσμάτων, διευκολύνοντας ταυτόχρονα την πρόσβαση του χρήστη (Lenzerini 2002, Halevy 2001, Ullman 2000).

- Σε γενικές γραμμές, στα συστήματα ολοκλήρωσης δεδομένων υπάρχει ένα *σχήμα διαμεσολάβησης* (mediated schema) ή αλλιώς *καθολικό σχήμα* (global schema), στο οποίο ο χρήστης υποβάλλει το ερώτημα. Το σχήμα διαμεσολάβησης δεν περιλαμβάνει δεδομένα από τις τοπικές πηγές αποθηκευμένα σε αυτό· αποτελεί επί της ουσίας έναν

εικονικό πυρήνα έκφρασης των δεδομένων που αυτές διαθέτουν. Οι σχέσεις ανάμεσα στο καθολικό σχήμα και στις *τοπικές πηγές* (local sources) εκφράζονται μέσα από *όψεις* (views). Υπάρχουν δύο τρόποι έκφρασης των σχέσεων ανάμεσα στο καθολικό σχήμα και τις τοπικές πηγές. Στην πρώτη προσέγγιση το καθολικό σχήμα εκφράζεται ως μία όψη των τοπικών σχημάτων οπότε και οι σχέσεις ανάμεσα στα δύο είναι άμεσες (Global–As–View (GAV)) (Cali κ.α. 2002). Στη δεύτερη προσέγγιση το καθολικό σχήμα εκφράζεται ανεξάρτητα από τις τοπικές πηγές και οι σχέσεις ανάμεσα τους καθιερώνονται εκφράζοντας κάθε τοπική πηγή ως όψη του καθολικού σχήματος (Local–As–View (LAV)) (Duschka, Genesereth και Levy 2000).

- Τα βασικά θέματα που πρέπει να αντιμετωπιστούν σε ένα σύστημα ολοκλήρωσης δεδομένων είναι: α) ο τρόπος που ορίζουμε τις σημασιολογικά σχετικές πληροφορίες των πηγών με το καθολικό σχήμα, και β) η διαδικασία εκτέλεσης ενός ερωτήματος και ο μετασχηματισμός του σε επιμέρους ερωτήματα προς κάθε τοπική πηγή.

- Για να αντιμετωπιστούν τα προβλήματα σημασιολογικής ετερογένειας ανάμεσα στα τοπικά δεδομένα και το καθολικό σχήμα, ορίζονται *κανόνες συσχέτισης* (mapping rules). Οι κανόνες συσχέτισης ορίζουν μία σχέση ανάμεσα σε δύο μοντέλα· το πρώτο είναι το *πηγαίο σχήμα* (source schema), π.χ. μία τοπική πηγή δεδομένων, και το δεύτερο είναι το *τελικό σχήμα* (target schema), π.χ. το σχήμα διαμεσολάβησης. Πιο αναλυτικά, διατηρούν τη σημασιολογική σχέση ανάμεσα στις πιθανές ερμηνείες των πεδίων των δύο μοντέλων. Για παράδειγμα, μία σημασιολογική σχέση ανάμεσα σε δύο μοντέλα είναι η συσχέτιση των πεδίων ενός σχήματος μεταδεδομένων, π.χ. UNIMARC (IFLA 2000), με ένα διαφορετικό σχήμα μεταδεδομένων, π.χ. Encoded Archival Description (Library of Congress 2006).

- Αναφορικά με τη διαδικασία της ερώτησης, ένα ερώτημα υποβάλλεται ως προς το καθολικό σχήμα και στη συνέχεια μετασχηματίζεται σε ερωτήματα στις επιμέρους πηγές (μετασχηματισμός ερωτήματος). Στη διαδικασία αυτή αξιοποιούνται οι κανόνες συσχέτισης, οι οποίοι όσο πιο πολύπλοκοι είναι στην έκφρασή τους τόσο πιο δύσκολη γίνεται η διαδικασία του μετασχηματισμού των ερωτημάτων. Επιπλέον ζήτημα στη διαδικασία της ερώτησης είναι η μετατροπή του ερωτήματος από μία γλώσσα επερώτησης σε μία άλλη, σε περίπτωση που οι τοπικές πηγές είναι σε διαφορετική σύνταξη και κωδικοποίηση σε σχέση με το καθολικό σχήμα. Το ζήτημα της εκτέλεσης ερωτημάτων είναι εκτός των θεμάτων της συγκεκριμένης εργασίας.

2. ΣΗΜΑΣΙΟΛΟΓΙΚΗ ΟΛΟΚΛΗΡΩΣΗ ΔΕΔΟΜΕΝΩΝ

- Η παραδοσιακή Ολοκλήρωση Δεδομένων αποτελεί μία ανοιχτή ερευνητική περιοχή στο πεδίο των Βάσεων Δεδομένων τα τελευταία χρόνια. Πρόσφατα, το ερευνητικό ενδιαφέρον μεταφέρεται από την ολοκλήρωση δεδομένων στη σημασιολογική ολοκλήρωση δεδομένων σε ποικίλες ερευνητικές περιοχές, όπως είναι η *διαχείριση της πληροφορίας* (information management) και οι *οντολογίες* (ontologies). Η συγκεκριμένη τάση επηρεάζεται άμεσα από τη πρόθεση "μετασχηματισμού" του παγκόσμιου

ιστού σε πλέον πλούσιες σημασιολογικές φόρμες με τη χρήση τεχνολογιών του *Σημασιολογικού Ιστού* (Semantic Web).

- Η σημασιολογική ολοκλήρωση δεδομένων καλείται να αντιμετωπίσει προβλήματα *σημασιολογικής ετερογένειας* (semantic heterogeneity) τα οποία προκύπτουν από τη χρήση διαφορετικών όρων για την αποτύπωση της ίδιας έννοιας, τόσο σε *επίπεδο σχημάτων κωδικοποίησης δεδομένων* (schema level), όσο και στο *επίπεδο των δεδομένων* (data level) (Doan, Noy και Halevy 2004).
- Στο επίπεδο του σχήματος, τα δεδομένα μπορούν να διαφέρουν στις σχέσεις, στα ονόματα πεδίων και ιδιοτήτων, στο επίπεδο λεπτομέρειας και στην κάλυψη ενός συγκεκριμένου τομέα. Στο επίπεδο των δεδομένων, συναντούνται διαφορετικές παρουσιάσεις της ίδιας έννοιας, όπου οι διαφορές αυτές αντιμετωπίζονται με ποικίλες μεθόδους, όπως η δημιουργία συνδέσμων μεταξύ τους (record linkage).

2.1. Οντολογίες

Μία οντολογία ορίζεται ως "*μια τυπική* (formal), *κατηγορηματική* (explicit) προδιαγραφή μιας *διαμοιρασμένης* (shared) *εννοιολογικής αναπαράστασης* (conceptualization)" (Gruber 1993).

Οι οντολογίες δημιουργήθηκαν για να λειτουργήσουν ως ένας μηχανισμός διαλειτουργικότητας ανάμεσα σε ανθρώπους, φορείς και συστήματα. Εφόσον μπορούν να αναπαραστήσουν εννοιολογικά έναν τομέα, μπορούν να αξιοποιηθούν αποτελώντας μία "ομπρελά" όρων και σημασιών που εκφράζουν την ίδια έννοια. Με βάση αυτή τους την ιδιότητα, λειτουργούν ως φορέας επικοινωνίας μεταξύ διαφορετικών πληροφοριακών συστημάτων, με διαφορετικούς χρήστες, παρέχοντας μία κοινή βάση ανάμεσα σε αυτά, αναπαριστώντας και αναλύοντας τις οντότητες που περιγράφουν τα δεδομένα τους.

2.2. Ο ρόλος των οντολογιών στη σημασιολογική ολοκλήρωση δεδομένων

Σύμφωνα με τους (Cruz, Xiao και Hsu 2005) η σημασιολογική ολοκλήρωση δεδομένων είναι η διαδικασία της χρήσης εννοιολογικών αναπαραστάσεων των δεδομένων και των σχέσεών τους με στόχο την εξάλειψη της ετερογένειας. Εφόσον οι οντολογίες επιτρέπουν την πολύπλοκη έκφραση εννοιών και των σχέσεών τους, ο ρόλος τους στη σημασιολογική ολοκλήρωση δεδομένων είναι ιδιαίτερα ενεργός.

Στο συγκεκριμένο άρθρο, αναφερόμαστε στην αξιοποίηση των οντολογιών ως σχήματα διαμεσολάβησης για την ολοκλήρωση δεδομένων σε επίπεδο σχήματος. Μία οντολογία επιλέγεται ως καθολικό σχήμα διότι μπορεί να ορίσει πολύπλοκες σημασιολογικές σχέσεις ενός θεματικού χώρου. Παράλληλα, είναι διατυπωμένη σε αυστηρό μαθηματικό φορμαλισμό, ο οποίος επιτρέπει την εξαγωγή συμπερασμάτων, για παράδειγμα τον ορισμό επιπρόσθετων σχέσεων ανάμεσα στις έννοιες. Στο πλαίσιο αυτό έχουν αναπτυχθεί από ερευνητικές ομάδες συγκεκριμένες οντολογίες οι οποίες ορίζουν έννοιες και σχέσεις διαφόρων τομέων, π.χ. πολιτιστική κληρονομιά, υπολογιστική γλωσσολογία και γνωστική επιστήμη, οικονομικά.

Σε ένα σενάριο ολοκλήρωσης δεδομένων, υπάρχουν τρεις προσεγγίσεις σχετικές με το ρόλο των οντολογιών (Wache κ.α. 2001):

- **προσέγγιση μονής οντολογίας** (single ontology approach): Μία καθολική οντολογία (global ontology) παρέχει ένα διαμοιρασμένο λεξιλόγιο (shared vocabulary) για τον ορισμό των σημασιών αυτόνομων πηγών δεδομένων οι οποίες σχετίζονται με αυτήν. Η καθολική οντολογία περιγράφει ένα συγκεκριμένο τομέα, οπότε η συγκεκριμένη προσέγγιση εφαρμόζεται ιδιαίτερα σε πηγές δεδομένων που παρουσιάζουν διαφορετικές "όψεις" του ίδιου τομέα. Μία οντολογία που μπορεί να λειτουργήσει κατά αυτόν τον τρόπο είναι η CIDOC CRM για τον τομέα της πολιτισμικής κληρονομιάς (International Council of Museums 2006).
- **προσέγγιση πολλαπλών οντολογιών** (multiple ontology approach): Στην προσέγγιση αυτή κάθε τοπικό σύστημα δεδομένων περιγράφεται από μία ξεχωριστή *τοπική οντολογία* (local ontology). Η απουσία μίας καθολικής οντολογίας διευκολύνει την αυτόνομη ανάπτυξη τοπικών οντολογιών, οι οποίες εκφράζουν αναλυτικά και με συνέπεια τις έννοιες και τις σχέσεις κάθε τοπικού συστήματος δεδομένων. Εν τούτοις, το θετικό αυτό χαρακτηριστικό αποτελεί παράλληλα και ανοικτό ερευνητικό πρόβλημα κυρίως αναφορικά με τον *ορισμό των κανόνων συσχέτισης μεταξύ των οντολογιών* (ontology mapping) (Kalfoglou και Schorlemmer 2005).
- **προσέγγιση υβριδικής οντολογίας** (hybrid ontology approach): Η συγκεκριμένη προσέγγιση συνδυάζει χαρακτηριστικά από τις δύο προαναφερθείσες προσεγγίσεις. Κάθε τοπική πηγή περιγράφεται από μία ξεχωριστή τοπική οντολογία της οποίας η δημιουργία έχει βασιστεί είτε στις πρωτογενείς έννοιες μίας καθολικής οντολογίας είτε στη μετατροπή της τοπικής πηγής σε οντολογία. Το γεγονός αυτό διευκολύνει τη συσχέτιση των τοπικών οντολογιών και κατά συνέπεια των τοπικών πηγών δεδομένων.

Η πλειοψηφία των τοπικών πηγών δεν είναι δομημένη σύμφωνα με τις γλώσσες οντολογιών, αλλά σε XML μορφή ή σε σχεσιακό μοντέλο (βάσεις δεδομένων). Ως συνέπεια, δεν είναι δυνατή η άμεση συσχέτισή τους με τις κλάσεις και τις ιδιότητες της καθολικής οντολογίας και απαιτούνται διαδικασίες μετατροπής.

Τα τελευταία χρόνια, ένας διαρκώς αυξανόμενος όγκος δεδομένων και μεταδεδομένων δομείται και διατίθεται σύμφωνα με τη γλώσσα κωδικοποίησης XML. Στόχος αυτής της επιλογής αποτελεί η ευκολότερη έκφραση, διάθεση, μεταφορά και αποθήκευση των δεδομένων. Παράλληλα, η χρήση της XML ως γλώσσας έκφρασης δεδομένων λύνει προβλήματα ετερογένειας στη κωδικοποίηση δεδομένων (ετερογένεια στη σύνταξη). Στη διεθνή ερευνητική βιβλιογραφία παρατηρείται ιδιαίτερη ενασχόληση με τη δημιουργία κανόνων συσχέτισης ανάμεσα σε XML τοπικές πηγές και γλώσσες οντολογιών, όπως είναι η RDF και η OWL, με στόχο τη σημασιολογική ολοκλήρωσή τους (Lehti και Fankhauser 2004, Cruz, Xiao και Hsu 2004, Amann κ.α. 2001). Σχετικές αναφορές γίνονται στο Κεφάλαιο 5. Για την έκφραση των κανόνων αυτών χρησιμοποιούνται πλούσιες σημασιολογικά γλώσσες, π.χ. οι γλώσσες οντολογιών, καθώς και γλώσσες που προέρχονται από τις βάσεις δεδομένων όπως είναι η Datalog και η Frame Logic (F—Logic).

Σε γενικές γραμμές μπορούμε να πούμε ότι υπάρχουν δύο διαφορετικές προσεγγίσεις για τη δημιουργία κανόνων συσχέτισης XML δεδομένων με γλώσσες οντολογιών.

Η πρώτη προσέγγιση αντιμετωπίζει τους κανόνες συσχέτισης με *"διαδικαστικό"* (operational) τρόπο (Stuckenschmidt και Uschold 2005). Συγκεκριμένα, οι κανόνες συσχέτισης στην προσέγγιση αυτή περιγράφουν τη διαδικασία μετατροπής των XML δεδομένων στις κλάσεις και τις ιδιότητες μίας οντολογίας. Για παράδειγμα, τα "complexType" στοιχεία της XML μετατρέπονται σε RDF resources και τα "simpleType" στοιχεία και γνωρίσματα σε RDF properties (Cruz, Xiao και Hsu 2004).

Η δεύτερη προσέγγιση αντιμετωπίζει τους κανόνες συσχέτισης με *"δηλωτικό"* (declarative) τρόπο. Πιο αναλυτικά, οι κανόνες συσχέτισης στην προσέγγιση αυτή περιγράφουν τη (σημασιολογική) σχέση που πρέπει να υπάρχει μεταξύ των XML δεδομένων και των κλάσεων και ιδιοτήτων μίας οντολογίας, χωρίς να παρέχουν κατά ανάγκη και κάποια διαδικασία μετασχηματισμού. Η δεύτερη προσέγγιση φαίνεται να είναι πλέον αξιοποιήσιμη σε χώρους που έχουν ήδη αναπτυχθεί εννοιολογικά μοντέλα για το σημασιολογικό τους ορισμό.

3. ΣΗΜΑΣΙΟΛΟΓΙΚΗ ΔΙΑΛΕΙΤΟΥΡΓΙΚΟΤΗΤΑ ΜΕΤΑΔΕΔΟΜΕΝΩΝ

Το πρόβλημα της ενιαίας αναζήτησης σε ετερογενή δεδομένα προκύπτει ιδιαίτερα στο χώρο της επιστήμης της πληροφορίας και στους φορείς που διαχειρίζονται πληροφορία όπως τα αρχεία, οι βιβλιοθήκες και τα μουσεία. Οι συγκεκριμένοι φορείς διαθέτουν και αναπτύσσουν ποικίλες συλλογές με διαφορετικό υλικό το οποίο περιγράφεται με διαφορετικά σχήματα μεταδεδομένων —τα οποία συχνά είναι σε σύνταξη XML— ανάλογα με τις ανάγκες περιγραφής και τεκμηρίωσής του. Προκύπτει, κατ' επέκταση, το θέμα αναζήτησης και ανάκτησης ανάμεσα σε διαφορετικά σχήματα κωδικοποίησης. Για να επιτευχθεί η λειτουργία αυτή πρέπει είτε να μετατρέπεται ένα σχήμα μεταδεδομένων σε ένα άλλο (στο οποίο θα γίνεται η αναζήτηση) είτε όλα τα σχήματα να αντιστοιχούν σε ένα καθολικό σχήμα περιγραφής.

Αναφορικά με την επίτευξη της σημασιολογικής διαλειτουργικότητας των μεταδεδομένων στο επίπεδο σχημάτων, έχουν αναπτυχθεί εργαλεία μετατροπής από ένα σχήμα μεταδεδομένων σε ένα άλλο (conversion of metadata records). Για παράδειγμα, στη Βιβλιοθήκη του Κογκρέσου μετατρέπουν τις αρχειακές περιγραφές τους (EAD) (Library of Congress 2006) σε βιβλιογραφικές εγγραφές MARC 21 (Library of Congress 2005) για να εξασφαλίζουν επιπλέον πρόσβαση στις περιγραφές των αρχείων τους με τη χρήση του online βιβλιογραφικού καταλόγου. Εν τούτοις, στα πλαίσια αυτής της μετατροπής, η απεικόνιση της περιγραφής ενός αρχείου σε ένα βιβλιογραφικό πρότυπο δεν καλύπτει τις ανάγκες τεκμηρίωσης της αρχειακής περιγραφής· το γεγονός αυτό ισχύει για όλα τα σχήματα μεταδεδομένων, εφόσον κάθε ένα από αυτά έχει δημιουργηθεί με στόχο την τεκμηρίωση συγκεκριμένου τύπου υλικού.

Η σημασιολογική ολοκλήρωση δεδομένων με τη χρήση οντολογιών στο χώρο των μεταδεδομένων δίνει τη δυνατότητα πρόσβασης σε περιγραφές ποικίλου τύπου υλικού. Επιπρό-

σθετα, σε ένα σχετικό σενάριο ολοκλήρωσης μεταδεδομένων, μπορούν να αξιοποιηθούν οντολογικά μοντέλα που έχουν αναπτυχθεί στο χώρο της επιστήμης της πληροφορίας για αυτό το σκοπό, και τα οποία αποτελούν "ομπρελά" για σχήματα μεταδεδομένων και τοπικές πηγές, όπως είναι το εννοιολογικό μοντέλο Functional Requirements for Bibliographic Records (FRBR) (IFLA 1998) για βιβλιογραφικά δεδομένα, η οντολογία CIDOC Conceptual Reference Model (CRM) για την τεκμηρίωση πολιτιστικού περιεχομένου (αρχεία, βιβλιοθήκες, μουσεία) (International Council of Museums 2006) και η οντολογία ABC για ποικίλου τύπου μεταδεδομένα (Lagoze και Hunter 2001).

4. ΑΝΟΙΧΤΑ ΠΡΟΒΛΗΜΑΤΑ ΣΤΟΥΣ ΚΑΝΟΝΕΣ ΣΗΜΑΣΙΟΛΟΓΙΚΗΣ ΣΥΣΧΕΤΙΣΗΣ

Όπως προαναφέρθηκε, ένας κανόνας συσχέτισης ορίζει μία σημασιολογική σχέση ανάμεσα σε δύο πεδία. Στα πλαίσια ορισμού κανόνων συσχέτισης ανάμεσα σε XML δεδομένα και οντολογίες, έχει αντιμετωπιστεί μεγάλος αριθμός προβλημάτων. Εν τούτοις, είναι αρκετά δύσκολο μέσα από μία γλώσσα κανόνων να οριστούν όλες οι πιθανές σχέσεις, οι οποίες –κατά συνέπεια– ορίζονται μεμονωμένα με στόχο την κάλυψη συγκεκριμένων αναγκών ενός συστήματος ολοκλήρωσης δεδομένων.

Τα προβλήματα που προκύπτουν στο ορισμό κανόνων συσχέτισης ανάμεσα σε XML δεδομένα και πλουσιότερες σημασιολογικές φόρμες αφορούν σε μεγάλο βαθμό το ποιες σχέσεις πρέπει να καλυφθούν με έναν κανόνα. Οι κανόνες έχουν ως στόχο να ορίσουν σχέσεις ανάμεσα σε δύο πεδία, όμως συχνά οι σχέσεις που πρέπει να δηλωθούν γίνονται πιο πολύπλοκές προκειμένου να αντιμετωπιστούν προβλήματα σημασιολογικής ετερογένειας. Σε γενικές γραμμές, οι βασικές σημασιολογικές σχέσεις σε επίπεδο σχημάτων είναι οι παρακάτω:

- **συνωνυμία** (synonymy): Σχέση η οποία συνδέει πεδία με διαφορετικό όνομα τα οποία όμως αφορούν την ίδια έννοια. Για παράδειγμα, το στοιχείο ‹title› του XML σχήματος μεταδεδομένων Metadata Object Description Schema (MODS) (Library of Congress 2006) το οποίο αποδίδει τον τίτλο ενός τεκμηρίου, αναφέρεται στην ίδια έννοια με το γνώρισμα "Title of the Manifestation" της οντότητας "Manifestation" του εννοιολογικού μοντέλου Functional Requirements For Bibliographic Records (FRBR) (IFLA 1998).
- **ευρύτερος όρος** (hypernym (broader term)): Μία έννοια μπορεί να είναι περισσότερο γενική από μία άλλη. Πιο αναλυτικά, η σημασία μίας έννοιας εμπεριέχει τη σημασία μίας άλλης έννοιας, της οποίας είναι ευρύτερος όρος. Η αντίθετη σχέση είναι η σχέση "στενότερος όρος" (hyponym (narrower term)). Για παράδειγμα, το στοιχείο ‹Architectural_Drawing› ενός XML εγγράφου είναι στενότερος όρος της έννοιας "Drawing" μίας οντολογίας και το αντίστροφο.
- **holonymy**: Δηλώνει τη σχέση ανάμεσα στο σύνολο και στα μέρη του συνόλου. Η αντίστροφη σχέση είναι η σχέση του μέρους προς το σύνολό του και ονομάζεται "meronymy". Ενδεικτικά αναφέρουμε ότι το στοιχείο ‹PhD_Students› ενός XML εγγράφου αποτελεί υπό–μέρος της έννοιας "Postgraduate_Students" μίας οντολογίας.

- **1: κανένα** (one–to–null): Μία έννοια σε ένα XML έγγραφο δεν έχει σημασιολογικά ισοδύναμη έννοια στα πλαίσια της οντολογίας. Ενδεικτικά αναφέρεται ότι το επίπεδο περιγραφής σε μία XML αρχειακή περιγραφή (EAD) (Library of Congress 2006) το οποίο δηλώνεται μέσα από το γνώρισμα "level" δεν έχει σημασιολογικά αντίστοιχο πεδίο στο εννοιολογικό μοντέλο CIDOC CRM (International Council of Museums).

- **1: πολλά** (one–to–many):Ένα πεδίο από ένα XML έγγραφο αντιστοιχεί σε ένα πεδίο της οντολογίας και το αντίστροφο. Για παράδειγμα, τα στοιχεία ‹subject›/‹topic› και ‹subject›/‹name› του XML σχήματος μεταδεδομένων MODS αντιστοιχούν στο πεδίο Subject του σχήματος Dublin Core.

- **disjoint**: Σχέση ανάμεσα σε δύο έννοιες οι οποίες δεν έχουν κοινές πληροφορίες μεταξύ τους και είναι διακριτές. Για παράδειγμα, το στοιχείο ‹Undergraduate_Students› ενός XML εγγράφου δεν έχει καμία κοινή πληροφορία με την έννοια "Postgraduate _Students" μίας οντολογίας.

- **δομή–ιεραρχία**: Σχέσεις ιεραρχίας ανάμεσα σε πεδία. Στην XML οι πληροφορίες δομούνται με μία δενδροειδή ιεραρχική δομή, η οποία αρκετές φορές υποδηλώνει μία σημασιολογική σχέση, π.χ. σχέση μέρους–συνόλου. Στο XML σχήμα μεταδεδομένων EAD (Library of Congress 2006) αποτυπώνεται η ιεραρχία της δομής ενός αρχείου (αρχείο, υπό–αρχείο, σειρά, υπό–σειρά κτλ) μέσα από στοιχεία που εμπεριέχονται. Συγκεκριμένα, το στοιχείο ‹archdesc› στο οποίο περιγράφεται το αρχείο περιλαμβάνει υπό–στοιχεία (‹c01› – ‹c12›) τα οποία περιγράφουν τα υπό–μέρη του αρχείου.

- **δομή–σειρά**: Σχέσεις σειράς ανάμεσα σε πεδία. Σε ένα XML έγγραφο, όπως προαναφέραμε τα στοιχεία ακολουθούν μία ιεραρχία. Στα πλαίσια της ιεραρχίας αυτής, συχνά η σειρά των στοιχείων υποδηλώνει μία επιπλέον σημασία. Για παράδειγμα, σε ένα XML έγγραφο που περιγράφεται βιβλιογραφία, μέσα σε ένα στοιχείο ‹authors› εμπεριέχονται επιπλέον στοιχεία ‹author› για να καταγραφούν τα ονόματα των συγγραφέων ενός έργου. Το πρώτο στοιχείο ‹author› θα είναι αυτό που θα περιέχει το βασικά πνευματικά υπεύθυνο του έργου, οπότε η σειρά των στοιχείων δηλώνει μία επιπλέον σημασιολογία.

Επιπλέον πρόβλημα στον ορισμό κανόνων ανάμεσα στην XML και στις γλώσσες οντολογιών προκύπτει λόγω της διαφορετικής λογικής των μοντέλων τους. Ενδεικτικά αναφέρουμε ότι η XML είναι ένα δενδροειδές μοντέλο με ετικέτες στους κόμβους, οι οποίες επιπλέον είναι σε μία ιεραρχία. Η RDF είναι ένα μοντέλο γράφου με ετικέτες στους κόμβους και στις ακμές και στους οποίους κόμβους δεν υπάρχει ιεραρχία. Με άλλα λόγια, οι γλώσσες οντολογιών έχουν μία επίπεδη δομή σε αντίθεση με την XML όπου υπάρχει ιεραρχική δενδροειδής δομή.

Τέλος, αξίζει να σημειωθεί ότι στο πλαίσιο των μεταδεδομένων η σημασιολογική συσχέτιση γίνεται πιο πολύπλοκη λόγω ιδιαιτεροτήτων που παρουσιάζει κάθε ένα από αυτά (Pierre και LaPlant 1998). Θα αναφερθούμε ενδεικτικά σε μερικά σημεία τα οποία επηρεάζουν ουσιαστικά τη δημιουργία κανόνων συσχέτισης ανάμεσα στα XML μεταδεδομένα και στις οντολογίες .

- **Μεταδεδομένα ενός ή πολλών αντικειμένων** (single versus multiple objects metadata): Για παράδειγμα, μία εγγραφή μεταδεδομένων μπορεί να περιγράφει ένα σύνολο αντικειμένων που ανήκουν στο ίδιο επίπεδο (π.χ. πολλές μονογραφίες) και μία εγγραφή ένα μόνο αντικείμενο (π.χ. μία μονογραφία).
- Η αποτύπωση όλων των πληροφοριών που καταχωρούνται για την περιγραφή διαφόρων τύπων υλικού δεν μπορούν να αποδοθούν πλήρως σε ένα γενικό εννοιολογικό μοντέλο.
- Διάφορα XML μεταδεδομένα έχουν επιμέρους περιορισμούς όπως πεδία επαναλαμβανόμενα ή μη επαναλαμβανόμενα, υποχρεωτικά ή προαιρετικά, οι οποίοι να μην εκφράζονται στα πλαίσια μίας οντολογίας.

5. ΣΧΕΤΙΚΕΣ ΕΡΓΑΣΙΕΣ

Στη διεθνή ερευνητική κοινότητα, υπάρχει μεγάλο ενδιαφέρον για τη σημασιολογική ολοκλήρωση δεδομένων με τη χρήση οντολογιών. Θα αναφερθούμε ενδεικτικά σε έργο ομάδων που ασχολούνται με την επίλυση του προβλήματος και που έχουν αντιμετωπίσει μεμονωμένα μερικά από τα προβλήματα που προαναφέρθηκαν.

Οι (Lakshmanan και Sadri 2003) στην εργασία τους, ορίζουν ότι όταν μία XML πηγή θέλει να συμμετάσχει σε ένα διαλειτουργικό σύνολο πηγών, πρέπει να συσχετίσει τα εννοιολογικά δεδομένα της με μία οντολογία. Σε αυτό το πλαίσιο, προτείνουν ένα μοντέλο συσχέτισης XML *διαδρομών* (XML paths) με μεταβλητές που δηλώνουν κλάσεις και ιδιότητες μίας συγκεκριμένης οντολογίας, αξιοποιώντας τη γλώσσα XPath.

Οι (Cruz, Xiao και Hsu 2004) στην GAV προσέγγισή τους, συσχετίζουν τα XML δεδομένα με RDF δεδομένα μετατρέποντας τα στοιχεία και τα γνωρίσματα των XML εγγράφων σε RDF κλάσεις και ιδιότητες. Ένα επιπλέον χαρακτηριστικό της προσέγγισής τους είναι η διατήρηση της ιεραρχικής δομής (δομή–ιεραρχία) των XML εγγράφων μέσα στην RDF οντολογία που δημιουργείται για το κάθε ένα από αυτά. Στη συνέχεια, οι επιμέρους οντολογίες ενοποιούνται σε μία καθολική οντολογία στην οποία τίθεται το ερώτημα. Σε μία επιπλέον εργασία τους δημιούργησαν μία γλώσσα για να δηλώσουν σημασιολογικούς κανόνες συσχέτισης ανάμεσα σε XML δεδομένα και RDF οντολογίες. Η γλώσσα αυτή βασίστηκε στην RDF, ονομάζεται RDFMS και αξιοποιείται για να δηλώσει σχέσεις και βαθμούς ισοδυναμίας (super, sub, equ) και λειτουργίες όπως πρόσθεση στοιχείων (and).

Οι (Amann κ.α. 2001) παρουσιάζουν μία προσέγγιση στην οποία στόχος είναι να εκμεταλλευτούν τη δομή των XML δεδομένων για να συσχετίσουν κομμάτια της πληροφορίας με οντολογίες. Για να ορίσουν τους κανόνες έχουν δημιουργήσει μία απλή γλώσσα συσχέτισης η οποία περιγράφει τις XML πηγές αξιοποιώντας το XPath και συνδέει τις XPath διαδρομές ενός XML εγγράφου με κλάσεις και ιδιότητες της οντολογίας. Η προσέγγισή τους είναι ανάλογη με των (Lakshmanan και Sadri 2003).

Οι (Lehti και Fankhauser 2004, Lehti και Fankhauser 2005) επιτυγχάνουν τη σημασιολογική ολοκλήρωση μετατρέποντας τα XML δεδομένα σε μία OWL καθολική οντολογία, ορίζο-

ντας κανόνες συσχέτισης οι οποίοι επίσης βασίζονται στην OWL. Με τη χρήση της συγκεκριμένης γλώσσας για την έκφραση κανόνων αντιμετωπίζουν προβλήματα συνωνυμίας και δομής–ιεραρχίας.

6. ΜΕΛΛΟΝΤΙΚΗ ΚΑΤΕΥΘΥΝΣΗ

Στην παρούσα εργασία παρουσιάστηκε το πλαίσιο της σημασιολογικής ολοκλήρωσης δεδομένων, ενώ δόθηκε ιδιαίτερη έμφαση στη διερεύνηση των προβλημάτων που αφορούν τους κανόνες συσχέτισης ανάμεσα σε ετερογενή XML δεδομένα και μεταδεδομένα και σε πλούσιες σημασιολογικά μορφές, όπως είναι οι γλώσσες οντολογιών. Η δική μας κατεύθυνση έχει προσδιοριστεί στην ανάπτυξη ενός μοντέλου για τη δημιουργία σημασιολογικών κανόνων συσχέτισης. Τα θέματα που θα εξεταστούν στη συνέχεια της εργασίας της ερευνητικής μας ομάδας στη συγκεκριμένη περιοχή είναι

- οι σημασιολογικές σχέσεις που θα αποδοθούν μέσα από το μοντέλο συσχέτισης, με ιδιαίτερη έμφαση στις σχέσεις που αφορούν XML μεταδεδομένα στο χώρο της επιστήμης της πληροφορίας,
- η επιλογή ή ο σχεδιασμός πλούσιας εκφραστικά γλώσσας για την απόδοση των σημασιολογικών σχέσεων, και
- η αξιοποίηση των σημασιολογικών αυτών σχέσεων στη διαδικασία μετασχηματισμού και απάντησης των ερωτημάτων.

ΒΙΒΛΙΟΓΡΑΦΙΑ

- Amann, B., I. Fundulaki, M. Scholl, C. Beeri και A–M. Vercoustre 2001. Mapping XML fragments to community web ontologies. Εργασία στο *Proceedings of the Fourth International Workshop on the Web and Databases*, WebDB 2001, Santa Barbara, California, USA, May 24–25, 2001, (επ.) Giansalvatore Mecca, Jérôme Siméon, 97–102.
- Cali, A., D. Calvanese, G.D. Giacomo και M. Lenzerini 2002. Data integration under integrity constraints. Εργασία στο *Advanced Information Systems Engineering: 14th International Conference*, CAiSE 2002 Toronto, Canada, May 27–31, Lecture Notes In Computer Science Vol. 2348, 262–279. Heidelberg: Springer–Verlag.
- Cruz, I.F., H. Xiao και F. Hsu 2004. Peer–to–Peer semantic integration of XML and RDF data sources. Εργασία στο *Agents and Peer–to–Peer Computing: Third International Workshop*, AP2PC 2004, New York, NY, USA, July 19, 2004, Lecture Notes In Computer Science Vol. 3601, 108–119. Heidelberg: Springer–Verlag.
- Cruz, I. και H. Xiao 2005. The role of ontologies in data integration. *Journal of Engineering Intelligent Systems* 13, (4), **http://www.cs.uic.edu/~advis/publications/dataint/eis05j.pdf** (πρόσβαση στις 20 Ιουλίου 2006).
- Doan, A., N.F. Noy και A.Y. Halevy 2004. Introduction to the special issue on semantic integration. *SIGMOD Record* 33, (4): 11–13.
- Duschka, O.M., M.R. Genesereth και A.Y. Levy 2000. Recursive query plans for data integration. *Journal of Logic Programming* 43, (1): 49–73.
- Gruber, T. 1993. A translation approach to portable ontology specifications. *Knowledge Acquisition* 5, (2): 199–220.
- Halevy, A.Y. 2001. Answering queries using views: a survey. *Very Large Databases Journal* 10 (4): 270–294.
- IFLA 2000. UNIMARC *Manual: Bibliographic Format 1994*, **http://www.ifla.org/VI/3/p1996–1/sec–uni.htm** (πρόσβαση στις 5 Ιουλίου 2006)
- IFLA 1998. Functional Requirements for Bibliographic Records: Final Report. München: IFLA Section on Cataloguing, K.G. Saur, **http://www.ifla.org/VII/s13/frbr/frbr.pdf** (πρόσβαση στις 2 Ιουλίου 2006).
- International Council of Museums 2006. *CIDOC Conceptual Reference Model (CRM)*, **http://cidoc.ics.forth.gr** (πρόσβαση στις 2 Ιουλίου 2006)
- Kalfoglou, Y. και W.M. Schorlemmer 2005. Ontology mapping: the state of the art. Εργασία στο *Proceedings of the Semantic Interoperability and Integration*. Schloss Dagstuhl: IBFI, **http://drops.dagstuhl.de/opus/volltexte/2005/40/pdf/04391.KalfoglouYannis.Paper.40.pdf** (πρόσβαση στις 20 Ιουλίου 2006).
- Koffina, I., G. Serfiotis και V. Christophides 2004. *Foundations for information integration: a state of the art*. Technical report, DELOS Network Of Excellence on Digital Libraries.
- Lagoze, C. και J. Hunter 2001. The ABC ontology and model. *Journal of Digital*

Information 2, (2), **http://jodi.tamu.edu/Articles/v02/i02/Lagoze/** (πρόσβαση στις 20 Ιουλίου 2006).

- Lakshmanan, L.V.S. και F. Sadri 2003. Interoperability on XML data. Εργασία στο *Proceedings of The Semantic Web—ISWC 2003, Second International Semantic Web Conference*, Lecture Notes In Computer Science Vol. 2870, 146—163. Heidelberg: Springer—Verlag.
- Lehti, P. και P. Fankhauser 2004. XML data integration with owl: experiences and challenges. Εργασία στο *2004 Symposium on Applications and the Internet* (SAINT 2004), 26—30 January 2004, 160—70. χ.τ: IEEE Computer Society.
- Lehti, P. και P. Fankhauser 2005. SWQL—a query language for data integration based on OWL. Εργασία στο *On the Move to Meaningful Internet Systems 2005: OTM Workshops: OTM Confederated Internationl Workshops and Posters, AWeSOMe, CAMS, GADA, MIOS+INTEROP, ORM, PhDS, SeBGIS, SWWS, and WOSE 2005,* Agia Napa, Cyprus, October 31—November 4, 2005, επ. Robert Meersman, Zahir Tari, Pilar Herrero, Lecture Notes In Computer Science Vol. 3762, 926—935 Heidelberg: Springer—Verlag.
- Lenzerini, M. 2002. Data integration: a theoretical perspective. Εργασία στο *Proceedings of the 21st ACM SIGACT—SIGMOD—SIGART Symposium on Principles of Database Systems (PODS02)*, 233—46. New York: ACM.
- Library of Congress 2006. *Encoded Archival Description*, **http://www.loc.gov/ead/** (πρόσβαση 2 Ιουλίου 2006).
- Library of Congress 2005. *MARC Standards*, **http://www.loc.gov/marc/** (πρόσβαση 2 Ιουλίου 2006).
- Library of Congress 2006. *Metadata Object Description Schema (MODS)*, **http://www.loc.gov/standards/mods** (πρόσβαση 6 Ιουλίου 2006).
- Pierre, M. St. και W.P. Jr. LaPlant, 1998. Issues in crosswalking content metadata standards, **http://www.niso.org/press/whitepapers/crsswalk.html** (πρόσβαση στις 2 Μάη 2006)
- Stuckenschmidt, H. και M. Uschold 2005. Representation of semantic mappings. Εργασία στο *Proceedings of the Semantic Interoperability and Integration.* Schloss Dagstuhl: IBFI, **http://drops.dagstuhl.de/opus/volltexte/2005/53/pdf/04391.SWM7.ExtAbstract.53.pdf** (πρόσβαση στις 20 Ιουλίου 2006).
- Ullman, J.D. 2000. Information integration using logical views. *Theoretical Computer Science* 239, (2): 189—210.
- Wache, H., T. Vogele, U. Visser, H. Stuckenscmidt, G. Schuster, H. Neumann και S. Hubner 2001. Ontology—based integration of information: a survey of existing approaches. Εργασία στο *Proceedings of the Second Workshop on Ontology Learning OL'*2001 Seattle, USA, August 4 2001, **http://citeseer.ist.psu.edu/cache/papers/cs/27032/http:zSzzSzwww.cs.vu.nlzSz~heinerzSzpubliczSzois—2001.pdf/wache01ontologybased.pdf** (πρόσβαση στις 31 Ιουλίου 2006).

This research was partially co–funded by the European Social Fund (75%) and National Resources (25%) – Operational Program for Educational and Vocational Training (EPEAEK II) and particularly by the Research Program "PYTHAGORAS II".