Έργο:	**«ΘΑΛΗΣ: Ενίσχυση της Διεπιστημονικής ή και Διιδρυματικής έρευνας και καινοτομίας με δυνατότητα προσέλκυσης ερευνητών υψηλού επιπέδου από το εξωτερικό μέσω της διενέργειας βασικής και εφαρμοσμένης έρευνας αριστείας»**
Τίτλος Υποέργου:	**«ΕΙΚΟΣ»: Θεωρητική και αλγοριθμική θεμελίωση για Προσωποκεντρικά Συνεργατικά Πληροφοριακά Συστήματα**

Παραδοτέο Π.2.1

Υπερχώρος και διαχείριση μοντέλων

Σεπτέμβριος 2015

Δράση 2	**Ολοκλήρωση παραδοσιακών και μη δεδομένων, πλοήγηση και αναζήτηση**				
Ομάδα	Ερ. Ομάδα 2	**Έναρξη**	01/06/2012	**Λήξη**	30/11/2015

Συντονιστής ΕΟ2	Τ. Σελλής (ΙΠΣΥ - ΕΚ «Αθηνά» & RMIT)	
Υποδράση: ΥΔ 2.1	Υπερχώρος και διαχείριση μοντέλων	
Συμμετέχοντες	*Μέλη ΚΕΟ*	Τ. Σελλής (ΙΠΣΥ - ΕΚ «Αθηνά» & RMIT), D. Pfoser (ΙΠΣΥ - ΕΚ «Αθηνά»), Β. Βασσάλος (ΟΠΑ), Γ. Κούτρικα (Μετακαλούμενη - IBM Almaden), Θ. Δαλαμάγκας (ΙΠΣΥ - ΕΚ «Αθηνά»),
	Μέλη ΟΕΣ	Γ.Παπαδάκης (ΙΠΣΥ - ΕΚ «Αθηνά»), Κ. Μακρυνιώτη (ΟΠΑ), Γ. Παπαστεφανάτος (ΙΠΣΥ - ΕΚ «Αθηνά»), M. Reczko (Ε.ΚΕ.Β.Ε. Α. Φλέμινγκ)
Σύντομη Περιγραφή	Η Υποδράση 2.1 στοχεύει να προσδιορίσει ένα εκφραστικό και γενικό μοντέλο υπερχώρων δεδομένων, ικανό να αναπαραστήσει με ενιαίο και ομοιόμορφο τρόπο την ετερογένεια των πηγών δεδομένων που απαρτίζουν ένα υπερχώρο δεδομένων. Για τον ορισμό του μοντέλου και των πράξεων, που θα υποστηρίζει, θα υιοθετήσουμε τεχνικές από την ερευνητική περιοχή της διαχείρισης μοντέλων.	
Παραδοτέο	Π.2.1 Υπερχώρος και διαχείριση μοντέλων	
Στόχος στο Τ.Δ.	Τεχνική αναφορά που θα περιλαμβάνει τουλάχιστον 1 δημοσίευση.	
Επίτευξη στόχου	100%	

Περιεχόμενα

1 Εισαγωγή

Ο βασικός στόχος του έργου ΕΙΚΟΣ είναι να προσφέρει τη μεθοδολογία, τη θεωρητική θεμελίωση, τις αλγοριθμικές τεχνικές και την αρχιτεκτονική του λογισμικού που απαιτείται ώστε τα πληροφοριακά συστήματα να μπορούν να προσφέρουν στους χρήστες αφενός την δυνατότητα εξατομίκευσης της παρεχόμενης πληροφορίας και αφετέρου τη δυνατότητα χρήσης ενσωματωμένων ετερογενών δεδομένων, ενδεχομένως διαφορετικής προέλευσης, με διαφανή τρόπο.

Στα πλαίσια του έργου, η Δράση 2 «Ολοκλήρωση Παραδοσιακών και Μη Δεδομένων, Πλοήγηση και Αναζήτηση» παρέχει τεχνικές για τον ορισμό υπερχώρων δεδομένων και την αξιοποίηση παραδοσιακών και μη δεδομένων σε τέτοια περιβάλλοντα. Η Δράση οργανώνεται σε τρεις θεμελιώδεις δράσεις, εκ των οποίων η πρώτη αφορά στον ορισμό του εννοιολογικού μοντέλου αναπαράστασης υπερχώρων, η δεύτερη την περιγραφή του μηχανισμού ενσωμάτωσης νέων πηγών σε έναν υπερχώρο και την έυρεση αντιστοιχήσεων μεταξύ ετερογενών πηγών δεδομένων και η τρίτη την αρχιτεκτονική και τους μηχανισμούς που θα πρέπει να διαθέτει ένα σύστημα υποστήριξης υπερχώρων δεδομένων για την εφαρμογή επερωτήσεων και την ανάκτηση πληροφορίας από αυτό.

Το παρόν Παραδοτέο Π.2.1 περιλαμβάνει τα αποτελέσματα της υποδράσης ΥΔ2.1. Στην ενότητα 1 παρουσιάζουμε το γενικότερο πλαίσιο του προβλήματος. Στην ενότητα 2 παρουσιάζουμε μία μελέτη περίπτωσης για τη δημοσίευση στατιστικών δεδοδμένων ως Ανοικτά Διασυνδεδεμένα Δεδομένα. Στην ενότητα 3 παρουσιάζουμε αποτελέσματα σχετικά με την ολοκλήρωση Διασυνδεδεμένων Δεδομένων.

1.1 Πλαίσιο έρευνας

Ο σκοπός της δράσης 2.1 είναι να παρέχει μεθόδους μοντελοποίησης υπερχώρων βασισμένες στις τεχνικές που εφαρμόζονται στη διαχείριση μοντέλων. Αρχικά θα μελετήσουμε τρόπους αντιστοίχησης και αναπαράστασης υπερχώρων ως αφαιρετικών μοντέλων δεδομένων και στη συνέχεια θα προτείνουμε ένα σύνολο από πράξεις διαχείρισης και αντιστοίχησης μοντέλων

με στόχο την ολοκλήρωση σχημάτων και ετερογενών δεδομένων που προέρχονται από τις διάφορες πηγές.

Με τον όρο υπερχώρος δεδομένων (dataspace) αναφερόμαστε σε ένα σύνολο από χαλαρά συνδεδεμένους χώρους πληροφορίας (information container). Ένας χώρος πληροφορίας είναι ένας πόρος που περιέχει πληροφορία και μπορεί να αναφερθεί μέσω ενός αναγνωριστικού που είναι μοναδικό για τον υπερχώρο. Παραδείγματα τέτοιων πόρων είναι οι βάσεις δεδομένων, οι σχέσεις βάσεων δεδομένων, οι εγγραφές βάσεων δεδομένων, αρχεία, εγγραφές μέσα σε αρχεία, ροές δεδομένων, εγγραφές ροών δεδομένων, έγγραφα, τμήματα κειμένων, χάρτες, τροχιές, κλπ. Ένα σύνολο από χώρους πληροφορίας είναι χαλαρά συνδεδεμένο όταν ο υπερχώρος περιέχει μία αντιστοίχηση από τον ένα χώρο στον άλλο. Σε αυτό το πλαίσιο, μία αντιστοίχηση αποτελεί δήλωση συσχέτισης που μπορεί να υποδηλώνει part-of σχέση, σχέση συμμετοχής συνόλου, προέλευση, σημασιολογική ομοιότητα, αντιστοίχηση σχημάτων, ή οποιαδήποτε άλλη μορφή διασύνδεσης.

Οι υπερχώροι δεδομένων αποτελούν μια διευρυνόμενη τα τελευταία χρόνια προσέγγιση στην ποικιλομορφία των δεδομένων που παράγονται σήμερα από ένα πλήθος ετερογενών πηγών, όπως είναι εταιρικές βάσεις δεδομένων, ροές δεδομένων, γεωγραφικά δεδομένα, XML έγγραφα και σελίδες στο web, ημερολόγια sites, κ.α., και στην ολοκληρωμένη διαχείρισή τους ανεξάρτητα από την προέλευση και τη δομή που αυτά έχουν. Ένας από τους κύριους στόχους των υπερχώρων δεδομένων είναι οι χρήστες να έχουν τη δυνατότητα να εξερευνούν και να χειρίζονται δεδομένα χωρίς να απαιτείται η προηγούμενη γνώση των πηγών, της δομής και τους είδους των δεδομένων που είναι διαθέσιμα. Προσφέρουν στους χρήστες χρήσιμες απαντήσεις ακόμα και όταν ελάχιστη προσπάθεια από αυτούς έχει δαπανηθεί στην κατανόηση και συσχέτιση των δεδομένων. Από την άλλη πλευρά, η διαχείριση μοντέλων προσφέρει μεθόδους και τεχνικές για τη διαχείριση μεταδεδομένων και ειδικότερα για τη διαχείριση σχημάτων και αντιστοιχήσεων μεταξύ τους. Τα προβλήματα στα οποία επικεντρώνονται οι τεχνικές αυτές σχετίζονται με το ταίριασμα και την ολοκλήρωση σχημάτων (schema matching and integration), το μετασχηματισμό μοντέλων και την αναπαράσταση αντιστοιχήσεων μεταξύ μοντέλων (model mappings). Στη παρούσα Υποδράση, θα χρησιμοποιήσουμε τις

βασικές αρχές από τη διαχείριση μοντέλων για να υποστηρίζουμε διαδικασίες διαχείρισης υπερχώρων. Οι υπάρχουσες τεχνικές θα επεκταθούν και θα εμπλουτιστούν έτσι ώστε να αντιμετωπίσουμε την υψηλή ετερογένεια δεδομένων που συναντάται σε περιβάλλοντα υπερχώρων.

1.2 Κίνητρα της έρευνας και κεντρική ιδέα

Στο πλαίσιο της Υποδράσης Δ2.1 παρέχονται μέθοδοι μοντελοποίησης υπερχώρων δεδομένων, για την αναπαράσταση με ενιαίο και ομοιόμορφο τρόπο της ετερογένειας των πηγών δεδομένων που απαρτίζουν ένα υπερχώρο δεδομένων.

Το πρόβλημα που επιχειρεί να λύσει η συγκεκριμένη Υποδράση σχετίζεται με το ευρύτερο πρόβλημα μοντελοποίησης υπερχώρων δεδομένων και ολοκλήρωσης δεδομένων. Οι υπερχώροι αποτελούν μια νέα ερευνητική περιοχή στο χώρο της διαχείρισης δεδομένων και οι μέχρι τώρα προσεγγίσεις σκιαγραφούν λειτουργικές ιδιότητες που πρέπει να έχουν χωρίς να επεκτείνονται στην εισαγωγή συγκεκριμένων τρόπων μοντελοποίησης τους. Στη δράση αυτή θα ορίσουμε σε εννοιολογικό επίπεδο την έννοια του υπερχώρου, καθώς και τη δυνατότητα ενοποίησης πληροφοριών. Στην προτεινόμενη μέθοδο, η υλοποίηση των υπερχώρων βασίζεται στα Διασυνδεδεμένα Δεδομένα (Linked Data), τα οποία αποτελούν μια ευρέως διαδεδομένη πρακτική για τη δημοσίευση, ανταλλαγή και σύνδεση δεδομένων στον Ιστό. Με τη μέθοδο αυτή δημοσιεύονται δομημένα δεδομένα, τα οποία είναι συνδεδεμένα μεταξύ τους, επιτρέποντας έτσι την προηγμένη επεξεργασία και επερώτησή τους. Στα πλαίσια αυτά, το RDF είναι το μοντέλο δεδομένων που χρησιμοποιείται για τη μοντελοποίηση των υπερχώρων. Αποτελεί ένα γενικό μοντέλο δεδομένων, μέσω του οποίου δίνεται η δυνατότητα διαχείρισης της ετερογένειας στο σχήμα των δεδομένων με ενιαίο τρόπο, αντιστοιχίζοντας τα επιμέρους σχήματα σε αυτό.

2 Ολοκλήρωση Διασυνδεδεμένων Δεδομένων

Τα Διασυνδεδεμένα Δεδομένα (Linked Data) αποτελούν σημαντική πλευρά της εξέλιξης του Ιστού Δεδομένων (Web of Data). Με τον όρο αυτό περιγράφονται σύνολα δεδομένων που έχουν δομηθεί, δημοσιευθεί και συνδεθεί σύμφωνα με τους κανόνες που όρισε πρώτος ο Tim Berners-Lee. Στην παρούσα εργασία αναπτύσσεται μια web-based εφαρμογή που συλλέγει ανομοιογενή δεδομένα

από ποικίλες πηγές και τα εντάσσει σε χώρους δεδομένων. Η διαδικασία ολοκλήρωσης περιλαμβάνει τέσσερα κύρια βήματα: (α) τη συλλογή δεδομένων, (β) τον μετασχηματισμό σχήματος, (γ) την αναγνώριση όμοιων οντοτήτων, (δ) την αποθήκευση των δεδομένων και εκτέλεση ερωτημάτων SPARQL στο σχηματισμένο χώρο δεδομένων. Η εφαρμογή μετατρέπει τα δεδομένα που δεν βρίσκονται σε RDF μορφή σε RDF χρησιμοποιώντας έναν αλγόριθμο που αγνοεί το σημασιολογικό περιεχόμενο, μεταφράζει διαφορετικά σχήματα σε ένα ενιαίο τοπικό σχήμα και συνδέει όμοιες οντότητες. Βασίζεται σε ένα απλό περιβάλλον διεπαφής χρήστη και στα ανοικτού κώδικα εργαλεία R2R Framework και Silk Framework. Τα εισηγμένα δεδομένα σχηματίζουν διακριτά σύνολα δεδομένων που διαμορφώνουν ένα χώρο δεδομένων (dataspace) στον οποίο εφαρμόζονται SPARQL ερωτήματα. Η κύρια συνεισφορά της εφαρμογής είναι η δυνατότητα των χρηστών - με μια σειρά απλών βημάτων - να συνδυάζουν ετερογενή δεδομένα και να εξάγουν χρήσιμες πληροφορίες από αυτά.

Τα αποτελέσματά μας δημοσιεύτηκαν στη διπλωματική εργασία [Κανα13].

3 Δημοσιεύοντας δεδομένα απογραφής ως Ανοικτά Διασυνδεδεμένα Δεδομένα. Μελέτη περίπτωσης

Στην εν λόγω εργασία παρουσιάζεται μελέτη περίπτωσης για τη δημοσίευση στατιστικών δεδομένων ως Ανοικτά Διασυνδεδεμένα Δεδομένα (Linked Open Data). Τα στατιστικά δεδομένα διατηρούνται από τις στατιστικές υπηρεσίες και οργανισμούς, ενώ συγκεντρώνονται μέσω ερευνών ή από άλλες πηγές και αφορούν κυρίως την παρατήρηση κοινωνικοοικονομικών δεικτών. Σε αυτή την μελέτη περίπτωσης, παρουσιάζουμε τη δημοσίευση των προκαταρκτικών αποτελεσμάτων της απογραφής του πληθυσμού που διαμένει στην Ελλάδα, η οποία διεξήχθη το 2011, ως Ανοικτά Διασυνδεδεμένα Δεδομένα. Για τον σκοπό αυτό, έχουμε χρησιμοποιήσει το Data Cube λεξιλόγιο και το εργαλείο Google Refine για τη μοντελοποίηση και τη δημοσίευση των αποτελεσμάτων της απογραφής.

Τα αποτελέσματά μας δημοσιεύθηκαν στο άρθρο [PePD13] που παρουσιάστηκε στο 2nd International Workshop on Open Data (WOD 2013).

4 Ανακεφαλαίωση

Το παρόν παραδοτέο Π2.1 παρουσιάζει τα αποτελέσματα της υποδράσης ΥΔ2.1 του έργου ΕΙΚΟΣ. Ο στόχος της υποδράσης ΥΔ2.1 ήταν η παροχή μεθόδων μοντελοποίησης υπερχώρων δεδομένων, για την αναπαράσταση με ενιαίο και ομοιόμορφο τρόπο της ετερογένειας των πηγών δεδομένων που απαρτίζουν ένα υπερχώρο δεδομένων.

Στα πλαίσια της διερεύνησής μας, λοιπόν, επιτύχαμε να ανταποκριθούμε στο στόχο της υποδράσης με τους ακόλουθους τρόπους:

1. Παρουσιάσαμε μια εφαρμογή για την ολοκλήρωση δεδομένων υπό τη μορφή Διασυνδεδεμένων Δεδομένων, τα οποία αποτελούν μια διαδεδομένη πρακτική για την διαχείριση και δημοσίευση πληροφορίας στον Ιστό Δεδομένων, η οποία προσφέρει έναν νέο τρόπο ενσωμάτωσης και διαλειτουργικότητας.
2. Επιπλέον, παρουσιάσαμε πώς στατιστικά δεδομένα μπορούν να δημοσιευτούν ως Ανοικτά Διασυνδεδεμένα Δεδομένα.

Δημοσιεύσεις

[PePD13] Irene Petrou, George Papastefanatos, Theodore Dalamagas. Publishing Census as Linked Open Data. A Case Study. In Proceedings 2nd International Workshop on Open Data (WOD 2013), Paris, France, June 3, 2013.

[Κανα13] Ιωάννης Κανακάκης. Ολοκλήρωση Διασυνδεδεμένων Δεδομένων. Διπλωματική Εργασία, ΕΜΠ, Αθήνα, Δεκέμβριος 2013.

Παράρτημα