| | |
|---|---|
| **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.2.3

# Μηχανισμοί επερώτησης και ανάκτησης πληροφορίας

## Σεπτέμβριος 2015

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ & ΘΡΗΣΚΕΥΜΑΤΩΝ, ΠΟΛΙΤΙΣΜΟΥ & ΑΘΛΗΤΙΣΜΟΥ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

| | | | | | | |
|---|---|---|---|---|---|---|
| **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 2.3** | Μηχανισμοί επερώτησης και ανάκτησης πληροφορίας | | | | | |
| **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 2.3 στοχεύει στον ορισμό μεθόδων και μηχανισμών επερώτησης και ανάκτησης της πληροφορίας ενός υπερχώρου. Ο μηχανισμός επερωτήσεων αρχικά αποτιμά τις αναζητήσεις του χρήστη στην κάθε πηγή ξεχωριστά με βάση ένα αρχικό σύνολο αντιστοιχήσεων, ενώ θα επιτρέπει τη σταδιακή προσαρμογή του σε πιο εξειδικευμένες και σύνθετες μορφές επερωτήσεων ανάλογα με την πλοήγηση του χρήστη. | | | | | |
| **Παραδοτέο** | Π.2.3 Μηχανισμοί επερώτησης και ανάκτησης πληροφορίας | | | | | |
| **Στόχος στο Τ.Δ.** | Τεχνική αναφορά που θα περιλαμβάνει τουλάχιστον 2 δημοσιεύσεις. | | | | | |
| **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.2.3 περιλαμβάνει τα αποτελέσματα της υποδράσης ΥΔ2.3. Στην ενότητα 1 παρουσιάζουμε το γενικότερο πλαίσιο του προβλήματος. Στην ενότητα 2 προτείνεται ένα framework για την απάντηση ερωτημάτων συντομότερης διαδρομής σε γράφους κοινωνικών δικτύων με τη χρήση σχεσιακού συστήματος βάσεων δεδομένων. Στην ενότητα 3 παρουσιάζεται ένα εργαλείο για την ιεραρχική χαρτογράφηση και οπτική εξερεύνηση διασυνδεδεμένων δεδομένων, ενώ στην ενότητα 4 ερευνάται η αποδοτική πλοήγηση και οπτικοποίηση πολύ μεγάλων RDF γράφων. Στην ενότητα 5 παρουσιάζεται μια πρώτη προσέγγιση στο πρόβλημα της διαφοροποίησης αποτελεσμάτων αναζήτησης σε RDF δεδομένα. Στις ενότητες 6 και 7 αναφερόμαστε σε αποτελέσματα σχετικά με την αναζήτηση σε RDF δεδομένα.

## 1.1 Πλαίσιο έρευνας

Ο σκοπός της Υποδράσης 2.3 είναι να ορίσει σε λογικό και σε αρχιτεκτονικό επίπεδο τη διαδικασία ανάκτησης πληροφορίας από ετερογενείς πηγές και να περιγράψει τους μηχανισμούς εφαρμογής επερωτήσεων σε περιβάλλοντα υπερχώρων, λαμβάνοντας υπόψη τις ιδιαιτερότητες που υπάρχουν στην ετερογένεια και αβεβαιότητα των δεδομένων και των αντιστοιχήσεων μεταξύ τους. Κεντρικό στοιχείο της προσέγγισης μας είναι η δημιουργία ενός μηχανισμού που θα υποστηρίζει πολλαπλές μορφές αναζήτησης και ανάκτησης πληροφορίας, από απλές αναζητήσεις μέσω λέξεων κλειδιών μέχρι σύνθετες επερωτήσεις πάνω σε δεδομένα και συσχετίσεις δεδομένων, ανάλογα με τις δυνατότητες και τα είδη αναζήτησης που θα παρέχονται από την κάθε πηγή δεδομένων.

Οι εφαρμογές και χρήστες που προσπελαύνουν περιβάλλοντα υπερχώρων δεδομένων χρειάζονται μια ευρύτερη προσέγγιση στη σύνταξη και εφαρμογή επερωτήσεων σε σχέση με τα παραδοσιακά συστήματα βάσεων δεδομένων με στόχο την αποκόμιση σωστών και πλούσιων σημασιολογικά απαντήσεων από πολλαπλές πηγές δεδομένων. Ιδιαίτερα το διαδίκτυο και ο τεράστιος όγκος δεδομένων που παράγεται, οργανώνεται και ανταλλάσσεται καθημερινά μέσα σε αυτό ανάγουν την αναζήτηση της πληροφορίας σε πρωταρχική δραστηριότητα των χρηστών. Η αναζήτηση της πληροφορίας επεκτείνεται πέρα από την παραδοσιακή εφαρμογή επερωτήσεων σε δομημένες βάσεις δεδομένων, με πολλές άλλες μορφές: από την αναζήτηση με βάση μια λέξη κλειδί σε μια μηχανή αναζήτησης, στη χρήση υπηρεσιών διαδικτύου για την εύρεση και αγορά προϊόντων (π.χ., εύρεση πτήσεων), στην εφαρμογή σύνθετων κριτηρίων για την αναζήτηση αρχείων μέσα σε επιχειρησιακά δίκτυα και στη πλοήγηση μέσα σε δίκτυα ομοτίμων. Συνεπώς, μια από τις βασικές προκλήσεις σχετικά με την εφαρμογή επερωτήσεων σε υπερχώρους δεδομένων είναι η δυνατότητα των χρηστών να αναζητούν και να ρωτάνε με πολλαπλούς τρόπους. Για την ακρίβεια, η διάκριση μεταξύ αναζήτησης πληροφορίας και εφαρμογής επερωτήσεων σε δεδομένα θα πρέπει να ελαχιστοποιηθεί. Προς την κατεύθυνση αυτή, στοχεύουμε να μελετήσουμε μηχανισμούς σταδιακής αναζήτησης σε υπερχώρους δεδομένων. Οι χρήστες αρχικά θα πρέπει να αναζητούν με τον απλούστερο τρόπο (π.χ., μέσω λέξεων κλειδιών) και στη συνέχεια θα κατευθύνονται σε πιο εξειδικευμένες διεπαφές αναζήτησης, συσχέτισης

πληροφορίας και εφαρμογής επερωτήσεων ανάλογα με τη πληροφορία και τη λειτουργικότητα που είναι διαθέσιμη από κάθε πηγή δεδομένων. Η διαδικασία αυτή θα πρέπει να υποστηρίζεται αδιαφανώς και το σύστημα θα πρέπει να υποδεικνύει σταδιακά τους επόμενους πιθανούς τρόπους αναζήτησης ή εναλλακτικές μορφές αναζήτησης που μπορεί να χρησιμοποιήσει ο χρήστης. Επιπλέον, οι μηχανισμοί αυτοί θα λαμβάνουν υπόψη την ασάφεια που εμπεριέχεται στην πληροφορία, δίνοντας τη δυνατότητα για ανάκτηση προσεγγιστικών απαντήσεων.

## 1.2 Κίνητρα της έρευνας και κεντρική ιδέα

*Στο πλαίσιο της Υποδράσης 2.3 αναπτύξαμε μεθόδους επερώτησης και ανάκτησης πληροφορίας ενός υπερχώρου.*

Η εφαρμογή επερωτήσεων και η ανάκτηση πληροφορίας σε παραδοσιακά συστήματα ολοκλήρωσης δεδομένων έχει αρκετούς περιορισμούς ως προς τη δυνατότητα χειρισμού ετερογενούς εξελισσόμενης πληροφορίας. Απαιτεί την ύπαρξη ενός καθολικού σχήματος το οποίο θα ενσωματώνει όλη τη σημασιολογία και τα μεταδεδομένα των επιμέρους σχημάτων, ένα διαμεσολαβητή που θα διατηρεί το σύνολο των αντιστοιχήσεων μεταξύ των σχημάτων και κυριότερα μια συγκεκριμένη γλώσσα επερωτήσεων που καθορίζεται από το μοντέλο δεδομένων του καθολικού σχήματος. Μια τέτοια αρχιτεκτονική δυστυχώς δεν μπορεί να ανταποκριθεί στις ιδιαιτερότητες των υπερχώρων δεδομένων, όπου τόσο η δημιουργία αντιστοιχήσεων γίνεται με δυναμικό τρόπο, όσο και η ετερογένεια των δεδομένων και των μεταδεδομένων της κάθε πηγής επιβάλλει διαφορετικό τρόπο αναζήτησης κάθε φορά. Στα πλαίσια αυτά, παρουσιάζουμε μεθόδους και συστήματα για την επερώτηση και ανάκτηση πληροφορίας κυρίως RDF δεδομένων σε διάφορα σενάρια, όπου το RDF επιτελεί ως ένα μοντέλο ενσωμάτωσης ετερογενών δεδομένων. Διαχειριζόμαστε ερωτήματα συντομότερης διαδρομής σε γράφους μεγάλης κλίμακας, τα οποία μπορούν να βρουν εφαρμογή σε διάφορα ερωτήματα σε Διασυνδεδεμένα Δεδομένα. Επιπλέον, μελετάμε λύσεις για την αναζήτηση RDF δεδομένων με λέξεις κλειδιά, καθώς και το πρόβλημα της διαφοροποίησης αποτελεσμάτων αναζήτησης σε RDF δεδομένα. Ένα σημαντικό πρόβλημα σχετικό με την ανάκτηση δεδομένων, είναι η οπτικοποίηση δεδομένων με τη

μορφή γράφου. Με τον τρόπο αυτό παρέχεται η δυνατότητα οπτικής αναζήτησης και πλοήγησης σε υπερχώρους δεδομένων.

# 2 COLD. Επανεξέταση ετικετών hub σε βάση δεδομένων για γράφους μεγάλης κλίμακας

Ο υπολογισμός της συντομότερης διαδρομής σε γράφους είναι ένα από τα πλέον γνωστά προβλήματα στην θεωρία αλγορίθμων. Μια καινούρια σχετικά ερευνητική περιοχή που έχει προσελκύσει το ενδιαφέρον των ερευνητών είναι η χρήση παραδοσιακών συστημάτων σχεσιακών βάσεων δεδομένων σε συνδυασμό με αλγόριθμους συντομότερους διαδρομής, ώστε να μπορούν να απαντηθούν ερωτήματα συντομότερης διαδρομής σε γράφους μεγάλης κλίμακας. Σε αυτήν την κατεύθυνση, η παρούσα εργασία προτείνει ένα νέο, πρωτοποριακό και βελτιστοποιημένο framework που απαντά ερωτήματα συντομότερης διαδρομής σε γράφους κοινωνικών δικτύων και η οποία υλοποιήθηκε εξολοκλήρου σε ένα σύστημα διαχείρισης σχεσιακών βάσεων δεδομένων ανοιχτού λογισμικού. Το προτεινόμενο SQL framework με το όνομα COLD (COmpressed Labels on the Database) μπορεί και απαντά πλήθος ερωτημάτων συντομότερης διαδρομής (κόμβος-προς-κόμβο, ένα-προς-πολλούς, k-πλησιέστερων γειτόνων και Reverse k-πλησιέστερων γειτόνων) τα οποία δεν μπορούσαν να απαντηθούν από τις υπάρχουσες μεθόδους και συνεπώς μπορεί να χρησιμοποιηθεί σε πλήθος πρακτικών εφαρμογών σε γράφους μεγάλης κλίμακας. Η εκτεταμένη πειραματική μας αξιολόγηση έδειξε πως το προτεινόμενο COLD framework ξεπερνά σε απόδοση όλες τις προηγούμενες μεθόδους (συμπεριλαμβανομένων εξειδικευμένων βάσεων δεδομένων για γράφους) σε ταχύτητα απόκρισης ερωτημάτων, ενώ απαιτεί σημαντικά μικρότερο χώρο αποθήκευσης από τις προηγούμενες λύσεις.

Τα αποτελέσματά μας δημοσιεύθηκαν στο 14th International Symposium on Spatial and Temporal Databases (SSTD2015) [EfEP15].

# 3 rdf:SynopsViz - Ένα framework για την οπτική εξερεύνηση και ανάλυση ιεραρχιών διασυνδεδεμένων δεδομένων

Ο στόχος της οπτικοποίησης δεδομένων είναι να παρέχει διαισθητικούς και φιλικούς τρόπους για την ανάκτηση και εξερεύνηση μεγάλου όγκου πληροφορίας, ειδικά για τους μη ειδικούς χρήστες. Οι πρόσφατες τεχνολογίες των διασυνδεδεμένων δεδομένων (linked Data) και ευρύτερα το διαδίκτυο δεδομένων (Web of Data) επιτρέπουν τη δημοσίευση δομημένης πληροφορίας στο διαδίκτυο. Η ευρεία υιοθέτηση και χρήση αυτών των τεχνολογιών έχει οδηγήσει στη διαθεσιμότητα ενός τεράστιου όγκου ανοιχτών δεδομένων από πολλές διαφορετικές πηγές. Ωστόσο, ο όγκος και η ετερογένεια των διαθέσιμων πληροφοριών καθιστούν δύσκολο για τους ανθρώπους να εξερευνήσουν και να αναλύσουν τα δεδομένα αυτά με μη αυτόματο τρόπο και να πλοηγηθούν σε αυτά. Σε αυτή την εργασία, παρουσιάζουμε το rdf: SynopsViz, ένα εργαλείο για την ιεραρχική χαρτογράφηση και την οπτική εξερεύνηση ανοιχτών διασυνδεδεμένων δεδομένων - Linked Open Data (LOD). Η ιεραρχική εξερεύνηση βασίζεται στην αφαίρεση και ομαδοποίηση των οντοτήτων δεδομένων σε ένα ιεραρχικό μοντέλο. Τα διάφορα επίπεδα που απαρτίζουν την ιεραρχία του μοντέλου σχετίζονται μεταξύ τους με βάση τις τιμές των ιδιοτήτων των οντοτήτων που περιέχονται σε κάθε ομάδα. Το προτεινόμενο ιεραρχικό μοντέλο επιτρέπει στο χρήστη να πλοηγείται σε μεγάλους όγκους δεδομένων (αριθμητικών και ημερομηνίας), μειώνοντας την υπερχείλιση πληροφορίας που χαρακτηρίζει συνήθως εφαρμογές οπτικοποίησης δεδομένων. Επίσης, παρέχει αποτελεσματική άντληση πληροφοριών και περιλήψεων και στατιστικών στοιχείων για κάθε ομάδα.

Τα αποτελέσματά μας δημοσιεύθηκαν στο άρθρο [BiSP14] που παρουσιάστηκε στο 11th European Semantic Web Symposium (ESWC 2014) Satellite Events και συγκεκριμένα στο Demo Track.

# 4 Προς μια κλιμακούμενη οπτική εξερεύνηση πολύ μεγάλων RDF γράφων

Σε αυτή την εργασία, περιγράφουμε συνοπτικά την ανάπτυξη μιας πλατφόρμας η οποία βασίζεται σε σκληρό δίσκο, και προσφέρει αποδοτική πλοήγηση και

οπτικοποίηση πολύ μεγάλων γράφων. Σε αντίθεση με τις υφιστάμενες εργασίες, προτείνουμε μια γενική πλατφόρμα, που ονομάζεται graphVizdb, η οποία προσφέρει δυνατότητες κλιμακούμενης οπτικοποίησης γράφων οι οποίες δεν εξαρτώνται από τα χαρακτηριστικά του εκάστοτε γράφου. Η αποτελεσματικότητα της προτεινόμενης πλατφόρμας βασίζεται σε μια νέα τεχνική για την ευρετηρίαση και την αποθήκευση του γράφου. Η βασική ιδέα είναι ότι στη φάση προεπεξεργασίας, ο γράφος απεικονίζεται, χρησιμοποιώντας οποιαδήποτε από τις υπάρχουσες βιβλιοθήκες απεικόνισής γράφων. Στην συνέχεια, οι συντεταγμένες που έχουν ανατεθεί στους κόμβους του γράφου δεικτοδοτούνται κάνοντας χρήση χωρικών ευρετηρίων (R-tree), και αποθηκεύονται σε μια βάση δεδομένων. Κατά τον χρόνο εκτέλεσης, και ενώ ο χρήστης πλοηγείται πάνω από τον γράφο, συγκεκριμένα τμήματα του γράφου ανακτώνται με βάση τις συντεταγμένες τους.

Τα αποτελέσματά μας δημοσιεύθηκαν στο άρθρο [Bi++15] που παρουσιάστηκε στο 12th Extended Semantic Web Conference (ESWC2015) και συγκεκριμένα στο Poster Track.

# 5 RDivF: Διαφοροποιώντας την αναζήτηση με λέξεις κλειδιά σε RDF γράφους

Σε αυτήν την εργασία παρουσιάζουμε μία πρώτη προσέγγιση του προβλήματος της διαφοροποίησης (diversification) αποτελεσμάτων αναζήτησης σε σημασιολογικά RDF δεδομένα. Το πρόβλημα έγκειται στην ανάκτηση ενός συνόλου από k σημασιολογικά αποτελέσματα αναζήτησης με λέξεις κλειδιά, τα οποία θα παρουσιάζουν τη μεγαλύτερη δυνατή ετερογένεια, τόσο από άποψη περιεχομένου, όσο και από άποψη δομής και συσχετίσεων, μιας και αναφερόμαστε σε δομημένα δεδομένα RDF που μπορεί να ακολουθούν συγκεκριμένο σχήμα και που συσχετίζονται μεταξύ τους με τη βοήθεια ιδιοτήτων.

Αν και η διαφοροποίηση αποτελεσμάτων αναζήτησης σε αδόμητα δεδομένα (π.χ. ιστοσελίδες) είναι ευθύ πρόβλημα, αφού στόχος είναι η συλλογή αποτελεσμάτων ετερογενών ως προς το περιεχόμενο, δεν ισχύει το ίδιο για δομημένα δεδομένα. Η λύση που προτείνουμε είναι ένα πλαίσιο που θα

συνδυάζει διάφορες πτυχές των σημασιολογικών δεδομένων (περιεχόμενο, σχήμα, δομή, συσχετίσεις οντοτήτων) και θα παράγει ετερογενή σύνολα από γράφους - αποτελέσματα, οι οποίοι θα αντιστοιχούν μεν στις λέξεις κλειδιά του αντίστοιχου ερωτήματος, αλλά θα διαφοροποιούν τις οντότητες - κόμβους, τις ιδιότητες - ακμές που τους συνδέουν και τις κλάσεις της οντολογίας που τους χαρακτηρίζουν.

Τα αποτελέσματά μας δημοσιεύθηκαν στο άρθρο [Bi++13] που παρουσιάστηκε στο 17th International Conference on Theory and Practive of Digital Libraries (TPDL 2013).

# 6 Αναζήτηση με λέξεις κλειδιά σε RDF βασιζόμενη σε μετάφραση από λέξεις κλειδιά σε SPARQL

Στην εν λόγω εργασία παρουσιάζεται μία περίληψη παλαιότερων εργασιών μας στην αναζήτηση RDF δεδομένων με βάση λέξεις κλειδιά, υπερτονίζοντας τρέχουσες και μελλοντικές κατευθύνσεις. Πιο συγκεκριμένα, παρουσιάζεται η προσέγγισή μας για την αναζήτηση με βάση λέξεις κλειδιά σε δεδομένα γράφου, και συγκεκριμένα RDF. Η μέθοδός μας αντί να παρέχει απαντήσεις απευθείας από τον RDF γράφο, παράγει αυτόματα ένα σύνολο από υποψήφια SPARQL ερωτήματα, δηλαδή SPARQL ερωτήματα που προσπαθούν να περιγράψουν την απαιτούμενη πληροφορία για τον χρήστη όπως αυτή περιγράφεται από τις λέξεις κλειδιά. Επιπλέον, παρέχεται και μια περιγραφή σε φυσική γλώσσα των αντίστοιχων ερωτημάτων, βασιζόμενοι στο SPARQL2NL. Ένα πλήρως λειτουργικό πρωτότυπο σύστημα είναι διαθέσιμο στον σύνδεσμο http://snf-624527.vm.okeanos.grnet.gr:8080/KeywordSearchDiana/web/.

Τα αποτελέσματά μας δημοσιεύθηκαν στο άρθρο [GkPD15] που θα παρουσιαστεί στο 1st International Workshop on Novel Web Search Interfaces and Systems (NWSearch 2015).

# 7 Αναζήτηση και Εξερεύνηση RDF πόρων στην πλαρφόρμα LinkZoo

Η ιδέα του Ιστού των Δεδομένων, καθώς επεκτείνει την υπάρχουσα υποδομή του ιστού προς ένα ενιαίο χώρο δεδομένων που περιέχει και διασυνδέει

δεδομένα προερχόμενα από ένα μεγάλο και διαφορετικό πλήθος περιοχών, επιδρά δραματικά στον τρόπο με τον οποίο όχι μόνο δημιουργούμε και διασυνδέουμε μεγάλους όγκους πληροφορίας, αλλά και πως τους καταναλώνουμε. Τα Διασυνδεδεμένα Δεδομένα είναι η πιο ευρέως διαδεδομένη πρακτική για την διαχείριση, την δημοσίευση και την διάδοση πληροφορίας στον Ιστό των Δεδομένων, και η οποία προσφέρει έναν νέο τρόπο ενσωμάτωσης και διαλειτουργικότητας. Δύο είναι οι κεντρικές ιδέες των Διασυνδεδεμένων Δεδομένων: (α) ότι όλοι οι πόροι που δημοσιεύονται στον Ιστό αναγνωρίζονται μοναδικά από ένα URI και (β) συνδέσεις ανάμεσα σε πόρους έχουν σημασιολογική έννοια, όπως αυτή ορίζεται από το εκάστοτε λεξιλόγιο. Η επαναχρησιμοποίηση υπάρχοντων URIs και λεξιλογίων αντί για την δημιουργία νέων, καθώς και η σύνδεση ενός συνόλου δεδομένων με άλλα σύνολα δεδομένων δημιουργεί το Σύννεφο Ανοιχτών Διασυνδεδεμένων Δεδομένων.

Στην παρούσα δουλειά, παρουσιάζουμε επιγραμματικά την πλατφόρμα του LinkZoo και των βασικών του μερών, και εμβαθύνουμε ιδιαίτερα στις λειτουργίες αναζήτησης που παρέχει για την εξερεύνηση και την ανάκτηση πόρων στον Ιστό των Δεδομένων. Το LinkZoo είναι μια διαδικτυακή πλατφόρμα για την συνεργατική διαχείριση, επεξεργασία, υποσημείωση και διάδοση πόρων του Ιστού των Δεδομένων, καθώς και της δημοσίευσής τους ως Διασυνδεδεμένα Δεδομένα. Η πλατφόρμα διαθέτει δυο βασικές λειτουργίες αναζήτησης: (α) μια διαδραστική αναζήτηση με λέξεις κλειδιά για την αναζήτηση πόρων που εμπεριέχονται μέσα στο LinkZoo, πόροι που είτε έχουν εισαχθεί από τον ίδιο τον χρήστη, είτε έχουν δοθεί στον χρήστη μέσω ενός άλλου χρήστη, και (β) μια αναζήτηση με λέξεις κλειδιά για την εξερεύνηση απομακρυσμένων RDF δεδομένων μέσω επερωτήσεων SPARQL. Η πρώτη λειτουργία αναζήτησης δημιουργεί επερωτήσεις σε φυσική γλώσσα με βάση τις λέξεις κλειδιά που δίνει ο χρήστης, καθώς και με χρήση της ταξινομίας τόσο των πόρων όσο και των ιδιοτήτων τους. Η δεύτερη λειτουργία αναζήτησης παράγει αυτόματα ένα σύνολο από υποψήφιες επερωτήσεις SPARQL δεδομένου ενός συνόλου λέξεων-κλειδιών που παρέχει ο χρήστης. Οι επερωτήσεις SPARQL που παράγονται προσπαθούν ουσιαστικά να συλλάβουν και να εκφράσουν ποια είναι πληροφορία που πιθανώς αναζητεί ο χρήστης. Με αυτόν τον δεύτερο

μηχανισμό, το LinkZoo επιτρέπει την εξερεύνηση απομακρυσμένων συνόλων δεδομένων. Επιπλέον, το LinkZoo προσφέρει την δυνατότητα ενσωμάτωσης μέσα στην πλατφόρμα για περαιτέρω επεξεργασία. Με την συνδυαστική χρήση των μηχανισμών αναζήτησης που προσφέρει το LinkZoo επιτρέπει στους χρήστες αρχικά να ανακαλύπτουν απομακρυσμένα σύνολα δεδομένων και κατόπιν να τα εξετάζουν σε βάθος χωρίς να χρειάζεται να γνωρίζουν πλήρως την δομή τους. Τέλος, εξετάζουμε την αποτελεσματικότητα των υπηρεσιών του LinkZoo και ιδιαίτερα των υπηρεσιών αναζήτησης μέσα από ένα πραγματικό σενάριο χρήσης με χρήστες που δουλεύουν με δεδομένα από τον χώρο των επιστημονικών Διασυνδεδεμένων Δεδομένων.

Τα αποτελέσματά μας δημοσιεύθηκαν στο άρθρο [Me++15] που παρουσιάστηκε στο 4th International Conference on Data Management Technologies and Applications (DATA 2015).

# 8 Ανακεφαλαίωση

Το παρόν παραδοτέο Π2.3 παρουσιάζει τα αποτελέσματα της υποδράσης ΥΔ2.3 του έργου ΕΙΚΟΣ. Ο στόχος της υποδράσης ΥΔ2.3 ήταν να αναπτύξουμε μεθόδους επερώτησης και ανάκτησης πληροφορίας ενός υπερχώρου.

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

1. Παρουσιάσαμε το σύστημα COLD που θα διαχειρίζεται πλήθος ερωτημάτων συντομότερης διαδρομής σε γράφους μεγάλης κλίμακας, το οποίο μπορεί να χρησιμοποιηθεί και να αξιοποιηθεί από πλήθος εφαρμογών και πληθώρα διαφορετικών τύπων επερωτήσεων σε διασυνδεμένα δεδομένα. Επιπλέον, καθώς η υλοποίηση του συγκεκριμένου framework έγινε σε ένα σύστημα διαχείρισης σχεσιακών βάσεων δεδομένων ανοιχτού λογισμικού, είναι δυνατή η κλιμάκωσή του και η χρήση του από μεγάλο αριθμό ταυτόχρονων χρηστών.
2. Παρουσιάσαμε ένα ιεραρχικό μοντέλο και ένα προγραμματιστικό εργαλείο για την οπτική αναζήτηση και πλοήγηση σε υπερχώρους δεδομένων, όπως είναι αυτοί που δημιουργούνται με τα ανοιχτά διασυνδεδεμένα δεδομένα.

3. Προτείναμε την πλατφόρμα graphVizdb που υποστηρίζει την αποδοτική οπτικοποίηση και ανάκτηση δεδομένων γράφων μέσω χωρικών συναρτήσεων.
4. Προτείναμε μία πρώτη προσέγγιση στο πρόβλημα της διαφοροποίησης των αποτελεσμάτων αναζήτησης σε RDF δεοδμένα, λαμβάνοντας υπόψη διάφορα χαρακτηριστικά τους, όπως το περιεχόμενο, το σχήμα, τη δομή, τις συσχετίσεις οντοτήτων.
5. Παρουσιάσαμε λύσεις για την αναζήτηση RDF δεδομένων με βάση λέξεις κλειδιά, καθώς και τρέχουσες και μελλοντικές κατευθύνσεις έρευνας.
6. Παρουσιάσαμε τη διαδικτυακή πλατφόρμα LinkZoo για την συνεργατική διαχείριση, επεξεργασία, υποσημείωση και διάδοση πόρων του Ιστού των Δεδομένων, καθώς και την δημοσίευσή τους ως Διασυνδεδεμένα Δεδομένα. Εστιάσαμε κυρίως στις λειτουργίες αναζήτησης που παρέχει για την εξερεύνηση και την ανάκτηση πόρων στον Ιστό των Δεδομένων.

## Δημοσιεύσεις


[EfEP15] Alexandros Efentakis, Christodoulos Efstathiades, Dieter Pfoser. COLD. Revisiting Hub Labels on the Database for Large-Scale Graphs. In Proceedings 14th International Symposium on Spatial and Temporal Databases (SSTD2015), Hong Kong, China, August 26-28, 2015.

[BiSP14] Nikos Bikakis, Melina Skourla, George Papastefanatos. rdf: SynopsViz - A Framework for Hierarchical Linked Data Visual Exploration and Analysis. In 11th European Semantic Web Symposium (ESWC 2014) Satellite Events, Anissaras, Crete, Greece, May 25-29, 2014.

[Bi++15] Nikos Bikakis, John Liagouris, Maria Krommyda, George Papastefanatos and Timos Sellis. Towards Scalable Visual Expliration of Very Large RDF Graphs. In 12th European Semantic Web Symposium (ESWC 2015), Portoroz, Slovenia, May 31 - June 4, 2015.

[Bi++13] Nikos Bikakis, Giorgos Giannopoulos, John Liagouris, Dimitrios

Skoutas, Theodore Dalamagas, Timos Sellis. RDivF: Diversifying Keyword Search on RDF Graphs. In Proceedings 17th International Conference on Theory and Practive of Digital Libraries (TPDL 2013), Valletta, Malta, September 22-26, 2013.

[GkPD15] Katerina Gkirtzou, George Papastefanatos, Theodore Dalamagas. RDF keyword search based on keywords-to-SPARQL translation. In 1st International Workshop on Novel Web Search Interfaces and Systems (NWSearch 2015), Melbourne, Australia, October 23, 2015.

[Me++15] Marios Meimaris, George Alexiou, Katerina Gkirtzou, George Papastefanatos, Theodore Dalamagas. RDF Resource Search and Exploration with LinkZoo. In Proceedings of 4th International Conference on Data Management Technologies and Applications (DATA 2015), Colmar, Alsace, France, 20-22 July 2015.

## Παράρτημα