Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

ΕΡΕΥΝΗΤΙΚΟ ΚΕΝΤΡΟ ΚΑΙΝΟΤΟΜΙΑΣ ΣΤΙΣ ΤΕΧΝΟΛΟΓΙΕΣ ΤΗΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΤΩΝ ΕΠΙΚΟΙΝΩΝΙΩΝ & ΤΗΣ ΓΝΩΣΗΣ «ΑΘΗΝΑ»

ΙΝΣΤΙΤΟΥΤΟ ΠΛΗΡΟΦΟΡΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

LODGOV

Διακυβέρνηση Δεδομένων στην εποχή του Ιστού Δεδομένων: δημιουργία, διαχείριση, διατηρησιμότητα, κοινοχρησία και προστασία πόρων στον Ιστό.

ΔΡΑΣΗ «ΑΡΙΣΤΕΙΑ»

Παραδοτέο 5.3: Τεχνική Αναφορά

ΔΡΑΣΗ «ΑΡΙΣΤΕΙΑ»
Παραδοτέο 5.3: Τεχνική Αναφορά

Τεχνική Αναφορά Βιώσιμων Ανοιχτών Διασυνδεδεμένων Δεδομένων

Γκίρτζου Κ., Χατζόπουλος Σ., Τσιτσίγκος Δ., Καραγκούνη Δ., Σαρτζετάκης Στ., Σταύρακας Ι., Ιωαννίδης Ι. και Δαλαμάγκας Θ.

Table of Contents

1 Εισαγωγή

Ο Ιστός Δεδομένων έχει αλλάξει δραματικά τον τρόπο που δημιουργούμε, μοιραζόμαστε, διασυνδέουμε και καταναλώνουμε μεγάλο όγκο δεδομένων. Όλο και περισσότερα εταιρικά, κυβερνητικά και δεδομένα χρηστών ξεφεύγουν από τα στενά όρια της ιδιωτικής διαχείρισης, δημοσιεύονται, και γίνονται διαθέσιμα για πιθανούς καταναλωτές δεδομένων, είτε είναι εφαρμογές/υπηρεσίες,, είτε μεμονωμένους χρήστες, είτε και ολόκληρες κοινότητες. Ο Ιστός Δεδομένων επεκτείνει την τρέχουσα υποδομή του Ιστού σε ένα ευρύτερο χώρο δεδομένων που περιλαμβάνει διασυνδεδεμένα δεδομένα από διαφορετικά domains.

Το μοντέλο των Διασυνδεδεμένων Δεδομένων (Linked Data – LD) είναι η πιο συνηθισμένη πρακτική για δημοσίευση, διαμοιρασμό και διαχείριση πληροφοριών στον Ιστό Δεδομένων, και προσφέρει ένα καινούργιο τρόπο ενοποίησης δεδομένων και διαλειτουργικότητας. Σύμφωνα με τη βασική ιδέα των Διασυνδεδεμένων Δεδομένων, οι πόροι (resources) που δημοσιεύονται στον Ιστό μπορούν να προσδιοριστούν μοναδικά από ένα URI (Uniform Resource Identifier), και οι σύνδεσμοι (typed links) μεταξύ των URIs χρησιμοποιούνται για να συνδέσουν εννοιολογικά τους πόρους αυτούς. Το σύννεφο Ανοιχτών Διασυνδεδεμένων Δεδομένων – ΑΔΔ (Linked Open Data - LOD) δημιουργείται χρησιμοποιώντας υπάρχοντα URIs αντί να δημιουργούμε νέα, και δείχνοντας από το ένα σύνολο δεδομένων στο άλλο αναφερόμενοι σε αυτά [Bizer2009].

Τα Διασυνδεδεμένα Δεδομένα κυρίως υλοποιούνται με το Resource Description Framework (RDF). Μια αναπαράσταση RDF είναι ένα σύνολο από δηλώσεις (statements) που χαρακτηρίζουν πόρους, γνωστές και ως τριπλέτες, δηλαδή εκφράσεις της μορφής *subject predicate object*. Το subject αναφέρεται στον πόρο που περιγράφεται. Στην πραγματικότητα, το subject είναι μια αναφορά URI σε αυτόν τον πόρο που τον προσδιορίζει μοναδικά. Predicates είναι συνήθως όροι από υπάρχοντα λεξικά ή οντολογίες και προσδιορίζονται επίσης από URIs. Τέλος, το object είτε μια literal τιμή είτε ένα άλλο URI το οποίο αναφέρεται σε έναν άλλο πόρο RDF. Ένα σύνολο από τριπλέτες RDF μπορεί να αναπαρασταθεί από έναν κατευθυνόμενο γράφο με labels, γνωστό και ως γράφο δεδομένων RDF. Ωστόσο στην πραγματικότητα, οι τριπλέτες RDF αποθηκεύονται σε σχεσιακές βάσεις δεδομένων, σε εγγενή triple/quad stores ή σε DBMS γράφων [Faye2012; Bizer2008].

Μια από της βασικές προκλήσεις όταν δημιουργούμε και δημοσιεύουμε δεδομένα στον Ιστό Δεδομένων ως Διασυνδεδεμένα Δεδομένα είναι να παράξουμε διατηρήσιμους (sustainable) χώρους ΑΔΔ. Η διατηρησιμότητα (sustainability) των χώρων ΑΔΔ είναι σημαντική καθώς διασφαλίζει ότι τα δεδομένα θα συνεχίσουν να υφίστανται ανεξάρτητα από το μέσο αποθήκευσης ή την εξέλιξη των ίδιων των δεδομένων. Επιπλέον, εκτός από τη βοήθεια στη διαθεσιμότητα των δεδομένων, η διατηρησιμότητα βοηθάει επίσης στην κατανάλωση και επαναχρησιμοποίηση των ΑΔΔ.

Στο παραδοτέο 5.1 του έργου LODGOV, μελετήσαμε εκτενώς τη διαθέσιμη βιβλιογραφία και συλλέξαμε τις καλύτερες προτεινόμενες πρακτικές για το σχεδιασμό διατηρήσιμων χώρων ΑΔΔ. Σε αυτή την αναφορά, εξετάζουμε αν και πώς δύο σύνολα ΑΔΔ που σχεδιάζονται και δημοσιεύονται στο LODGOV ενσωματώνουν τις ορθές πρακτικές και τι προκλήσεις συναντήσαμε στο σχεδιασμό τους.

Το υπόλοιπο του παραδοτέου είναι οργανωμένο ως εξής: το Κεφάλαιο 2 παρουσιάζει τα δύο Ανοιχτά Διασυνδεδεμένα Σύνολα Δεδομένων που προέρχονται από δύο διαφορετικά πεδία εφαρμογών και χρησιμοποιούνται σε αυτή την αναφορά ως case studies για το πρόβλημα της διατηρησιμότητας. Το Κεφάλαιο 3 παρουσιάζει πως τα δύο σύνολα δεδομένων ικανοποιούν τις προτεινόμενες πρακτικές για τη δημιουργία διατηρήσιμων χώρων ΑΔΔ. Τέλος, το Κεφάλαιο 4 κλείνει το παραδοτέο.

2 Σύνολα Δεδομένων

Οι προτεινόμενες τεχνικές του LODGOV αξιολογούνται με δύο Σύνολα Ανοιχτών Διασυνδεδεμένων Δεδομένων που προέρχονται από δύο διαφορετικά πεδία εφαρμογών. Το πρώτο σύνολο δεδομένων περιλαμβάνει πληροφορίες σχετικές με προϊόντα βιομάζας και λέγεται AI4B[1], ενώ το δεύτερο είναι ένα σύνολο δεδομένων για μόρια microRNA και λέγεται DIANA[2]. Ο Πίνακας 1 δείχνει ποσοτικά χαρακτηριστικά των δύο συνόλων δεδομένων.

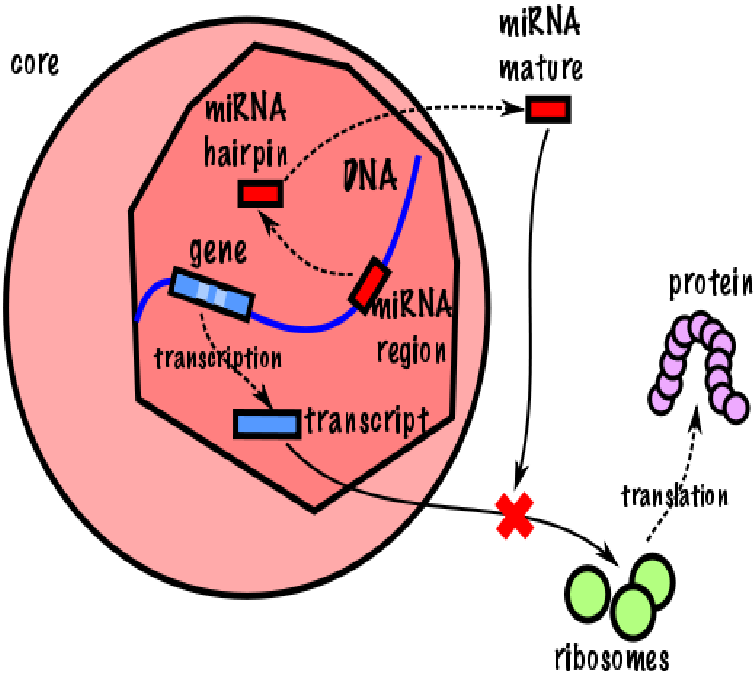
Πίνακας 1: Συγκεντρωτικά στατιστικά για τα ΑΔΔ AI4B και DIANA

Dataset	# Triplets	# Classes	# Properties	# Unique String Values
AI4B	$2,7*10^6$	15	148	6.350
DIANA	$4,6*10^9$	16	76	613.408

Το σύνολο δεδομένων AI4B είναι ένα ανοιχτό σύνολο δεδομένων προϊόντων βιομάζας σχεδιασμένο να καλύψει τις ανάγκες της βιομηχανίας ανανεώσιμων πηγών ενέργειας και της αγοράς της. Με το όρο βιομάζα αναφερόμαστε σε κάθε υλικό που προέρχεται από ζωντανούς οργανισμούς. Πιο συγκεκριμένα, η βιομάζα για ενεργειακούς σκοπούς περιλαμβάνει κάθε υλικό που μπορεί να χρησιμοποιηθεί για την παραγωγή στερεών, υγρών ή/και αέριων καυσίμων. Στην πράξη, υπάρχουν δύο ήδη βιομάζας: πρώτον τα υπολείμματα (κάθε είδος από υπολείμματα φυτών και ζώων καθώς και το οργανικό μέρος των αστικών αποβλήτων) και δεύτερον η βιομάζα που παράγεται από ειδικά ενεργειακά φυτά. Το σύνολο δεδομένων AI4B μοντελοποιεί και τα δύο αυτά προϊόντα βιομάζας, τις τεχνολογίες που μετατρέπουν τα προϊόντα βιομάζας σε ενέργεια καθώς επίσης και τις ανάγκες της αγοράς όπως προμηθευτές, πελάτες, καθώς και την τρέχουσα προσφορά και ζήτηση. Σκοπός του είναι να βοηθήσει στη δημιουργία ενός βιομηχανικού δικτύου που να μπορεί να επεξεργάζεται απόβλητα και εμπορικά υλικά, ενέργεια και νερό ώστε να πετύχει οικονομικό, περιβαλλοντικό και κοινωνικό όφελος. Οι πληροφορίες για το σύνολο δεδομένων AI4B έχουν συλλεχθεί από το Ερευνητικό Κέντρο «Αθηνά» σε συνεργασία με εταίρους από τη βιομηχανία ανανεώσιμων πηγών ενέργειας στην Ελλάδα.

Το δεύτερο σύνολο δεδομένων, DIANA, προέρχεται από τις επιστήμες της ζωής και είναι ένα ανοιχτό σύνολο δεδομένων που περιέχει πληροφορίες για βιολογικές οντότητες από τον κόσμο των microRNA. MicroRNAs (ή miRNAs) είναι μικρά μόρια RNA που δένουν σε μετάγραφα (transcripts) αγγελιοφόρων RNA (mRNA), που ονομάζονται στόχοι (targets), και ελέγχουν (down-regulate) τον μεταβολισμό τους. Τα miRNA στόχοι είναι σημαντική πληροφορία καθώς μπορούν να χρησιμοποιηθούν σε θεραπείες. Για παράδειγμα, η γνώση αυτή μπορεί να χρησιμοποιηθεί ώστε να μειωθεί ο μεταβολισμός των γονιδίων εισάγοντας τεχνητά miRNAs στα κύτταρα. Η γένεση των microRNAs, όπως φαίνεται στην Εικόνα 1, αποτελείται από δύο βήματα. Στο πρώτο βήμα, ένα μετάγραφο του microRNA γονιδίου δημιουργεί το πρόδρομο (precursor) miRNA (γνωστό και ως hairpin miRNA). Στο δεύτερο βήμα, όπου γίνεται η φάση της ωρίμανσης, το πρόδρομο μετατρέπεται σε ώριμο miRNA. Το ώριμο miRNA είναι το ενεργό μόριο που στοχεύει σε μετάγραφα mRNA, ρυθμίζοντας έτσι τον μεταβολισμό ενός γονιδίου.

1http://web.imis.athena-innovation.gr/projects/lodgov/ai4b
2http://web.imis.athena-innovation.gr/projects/lodgov/diana



Εικόνα 1: Η γένεση ενός microRNA

Όλη η πληροφορία σχετικά με τα miRNAs έχει συλλεχθεί από το Ερευνητικό Κέντρο «Αθηνά» σε συνεργασία με το Πανεπιστήμιο Θεσσαλίας και το Ερευνητικό Κέντρο Βιοϊατρικών Επιστημών «Α. Φλέμινγκ». Μια πληθώρα βάσεων δεδομένων ανοιχτής πρόσβασης έχουν χρησιμοποιηθεί ως πόροι: η βάση δεδομένων miRBase (έως και την έκδοση 18)[3] που παρέχει τη βασική πληροφορία σχετικά με τα microRNAs, τόσο για την hairpin όσο και για την mature μορφή τους, η βάση δεδομένων ENSEMBL (έκδοση 69)[4] που παρέχει πληροφορία σχετική με γονίδια, ενώ η βάση δεδομένων μονοπατιών KEGG[5] που παρέχει πληροφορία σχετική με αλληλεπιδράσεις μορίων και των δικτύων αντίδρασης. Τα κύρια χαρακτηριστικά της βάσης δεδομένων DIANA είναι τα ακόλουθα (α) ο όγκος της, όπως φαίνεται και από τα στατιστικά του Πίνακας 1, και (β) το γεγονός ότι τα δεδομένα αυτά εξελίσσονται στο χρόνο. Πιο συγκεκριμένα αλλαγές παρατηρούνται και στις δύο μορφές του microRNA, το hairpin miRNA και το mature miRNA. Οι αλλαγές που έχουν παρατηρηθεί εξετάζοντας τα αρχεία μεταδεδομένων που παρέχονται σε κάθε έκδοση της βάσης δεδομένων miRBase φαίνονται λεπτομερώς στον Πίνακας 2. Από αριστερά προς τα δεξιά φαίνεται το είδος της αλλαγής, μια περιγραφή της αλλαγής και αν παρατηρείται σε hairpin ή ώριμα miRNAs.

Πίνακας 2: Σύνοψη των αλλαγών που παρατηρούνται στη βάση δεδομένων miRBase για hairpin και ώριμα miRNAs

Change Type	Description	Hairpin	Mature
NEW	Insertion of a miRNA entry in miRBase.	✓	✓
NAME	A miRNA changes the value of its name.	✓	✓

3http://www.mirbase.org/
4http://www.ensembl.org/index.html
5http://www.genome.jp/kegg/pathway.html

SEQUENCE	A miRNA changes the value of its sequence.	✓	✓
NAME/SEQUENCE	A miRNA changes both the value of is name and its sequence.	✓	✓
DELETE	Deletion of a miRNA entry in miRBase.	✓	✓
ADD HAIRPIN PARENT	Creation of the relationship among a pair of hairpin and a mature miRNA.	✓	✓
REMOVE HAIRPIN PARENT	Deletion of the relationship among a pair of hairpin and a mature miRNA.	✓	✓

3 Διατηρησιμότητα ΑΔΔ

Το πρώτο βήμα ως προς τη διατηρησιμότητα του χώρου των ΑΔΔ σχετίζεται με τα χαρακτηριστικά των URIs. Το δεύτερο βήμα περιλαμβάνει τα λεξικά που χρησιμοποιήθηκαν για τη μοντελοποίηση και την περιγραφή των δεδομένων ενώ το τελευταίο αφορά τη σύνδεση με άλλα σύνολα δεδομένων προκειμένου να υποβοηθηθεί η συνδεσιμότητα, η εξερεύνηση και ο σχηματισμός του Σύννεφου Διασυνδεδεμένων Δεδομένων.

3.1 Χαρακτηριστικά των URIs

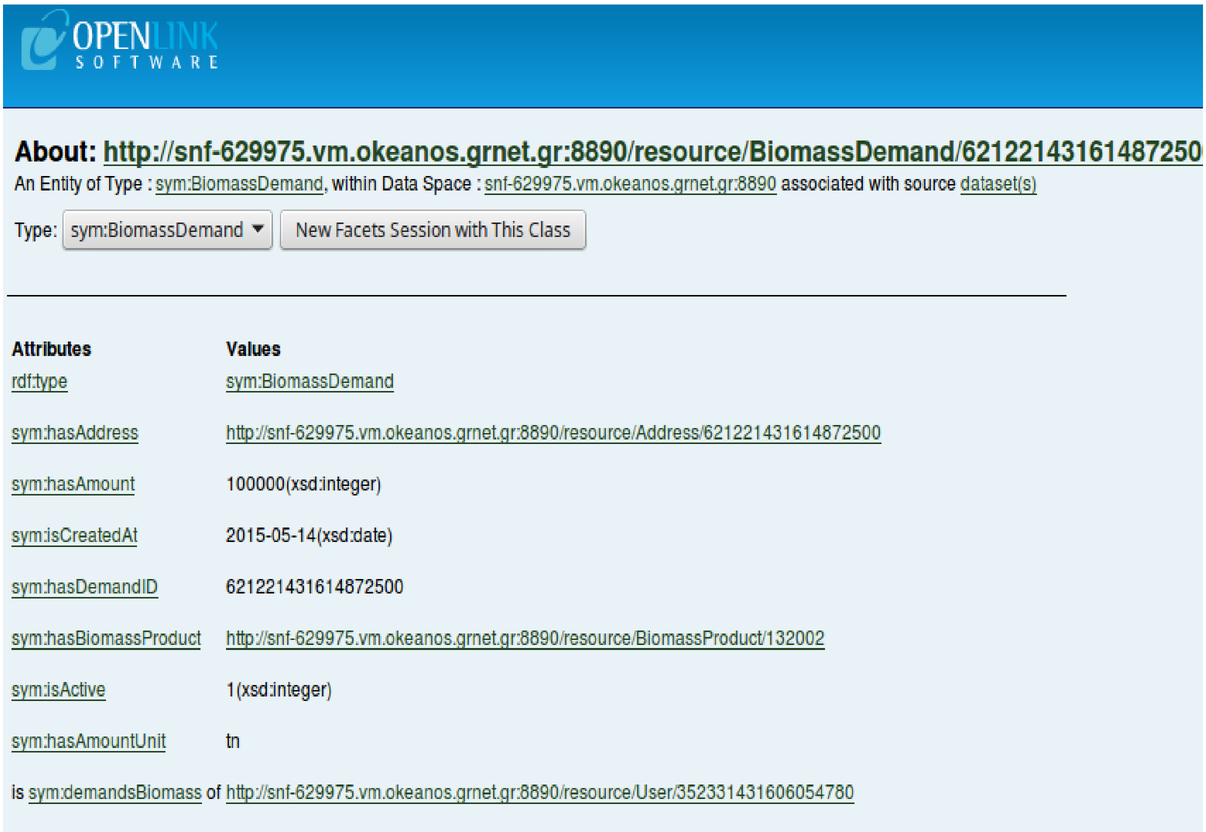
3.1.1 Ανακατευθυνόμενα URIs

Τα URIs είναι θεμελιώδη στο μοντέλο Διασυνδεδεμένων Δεδομένων, καθώς τόσο οι πόροι (resources) που μοντελοποιούνται σε ένα σύνολο δεδομένων, τόσο και τα κατηγορήματα (predicates) που περιγράφουν τα δεδομένα προσδιορίζονται μοναδικά από μια αναφορά URI. Το πιο σημαντικό χαρακτηριστικό που πρέπει να πληροί ένα URI είναι η ανακατεύθυνση. Το ανακατευθυνόμενο URI είναι ένας μηχανισμός ανάκτησης πόρων που χρησιμοποιεί κάποιο πρωτόκολλο του internet, όπως το HTTP, ώστε να αποκτήσει ένα αντίγραφο ή μια αναπαράσταση του πόρου που προσδιορίζει. Στα πλαίσια του παραδοσιακού ιστού HTML, ένα URI αναφέρεται σε μια ιστοσελίδα και όταν ανακατευθυνόμαστε σε αυτό ο web server επιστρέφει ένα αντίγραφο αυτής της σελίδας. Στην περίπτωση των Διασυνδεδεμένων Δεδομένων, η αναπαράσταση έχει τη μορφή ενός αρχείου που περιγράφει τον πόρο που προσδιορίζει URI. Το αρχείο αυτό είναι συνήθως σε μορφή HTML ή XML.

Τόσο το σύνολο δεδομένων AI4B όσο και το σύνολο δεδομένων DIANA είναι αποθηκευμένα σε Virtuoso[6], ένα υβριδικό σύστημα που περιλαμβάνει ένα σχεσιακό σύστημα βάσεων δεδομένων και μια web server εφαρμογή που προσφέρει SQL, XML και RDF διαχείριση των δεδομένων. Η χρήση του Virtuoso εξασφαλίζει ότι όλοι οι αποθηκευμένοι σε αυτό πόροι (resources) θα έχουν ανακατευθυνόμενα URIs που θα αντιπροσωπεύονται από έγγραφα HTML. Παραδείγματα από ανακατευθυνόμενα URIs από τα δικά μας σύνολα δεδομένα και τις HTML σελίδες που αντιπροσωπεύουν φαίνονται στην Εικόνα 2. Η Εικόνα 2 δείχνει την HTML σελίδα που παίρνουμε όταν ανακατευθυνόμαστε στο URI `.../BiomassDemand/621221431614872500` το οποίο αναπαριστά τη ζήτηση ενός προϊόντος βιομάζας του AI4B. Αντίστοιχα, στην Εικόνα 3 φαίνεται η HTML σελίδα που

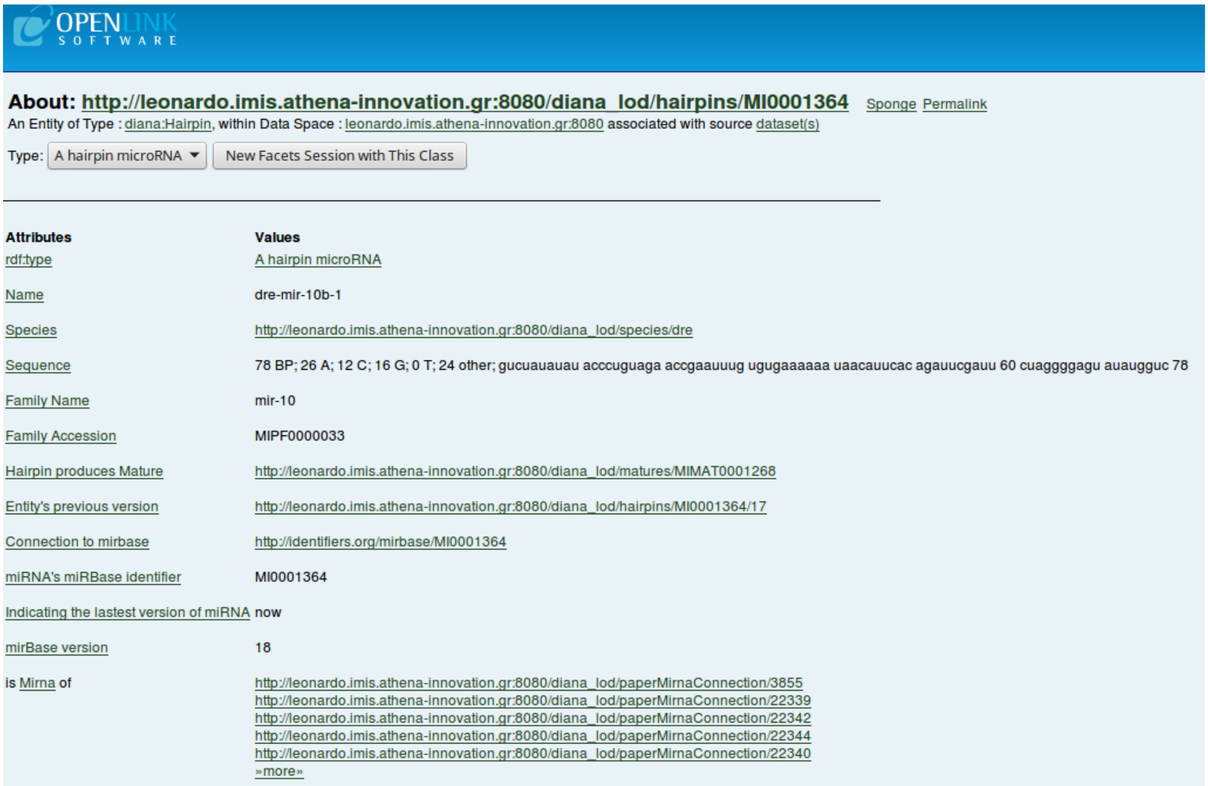
6virtuoso.openlinksw.com

παίρνουμε όταν ανακατευθυνόμαστε στο URI `…/hairpins/MI0001364` που αντιπροσωπεύει έναν hairpin miRNA πόρο από το σύνολο δεδομένων DIANA. Τα ανακατευθυνόμενα URIs χρησιμοποιούνται επίσης για σημασιολογικές συνδέσεις, όπως τα κατηγορήματα. Για να περιγράψουμε τα σύνολα δεδομένων AI4B και DIANA, σχεδιάσαμε δύο λεξικά για τις ανάγκες του καθενός, το λεξικό *ai4b* και το *diana* αντίστοιχα.

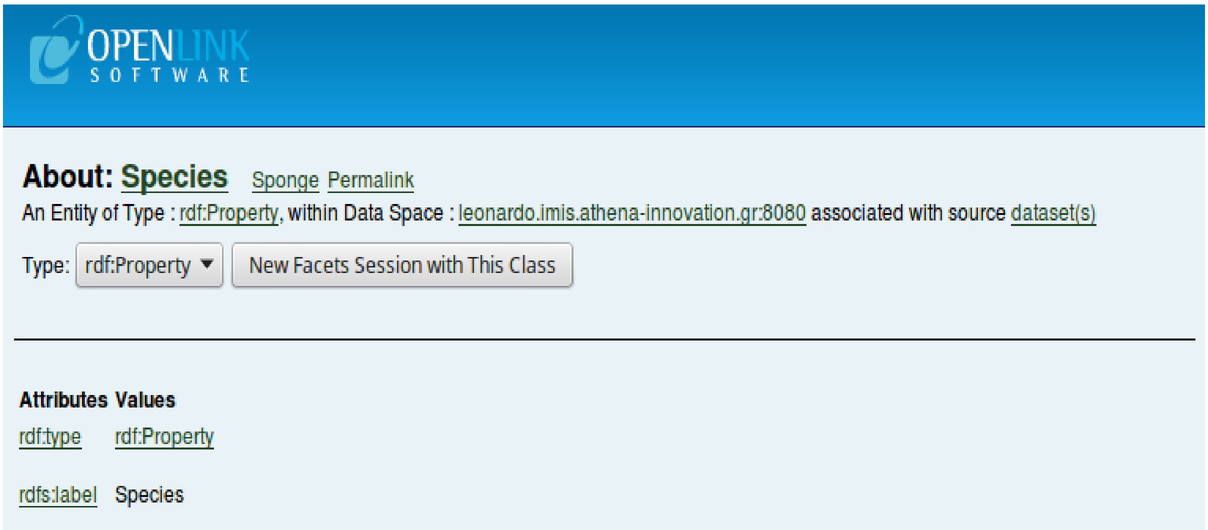


Εικόνα 2: Η HTML σελίδα από ανακατεύθυνση ενός πόρου ζήτησης βιομάζας από το σύνολο δεδομένων AI4B.

Όλες οι RDF κλάσεις καθώς και οι ιδιότητές τους αποθηκεύτηκαν σε βάση δεδομένων Virtuoso, επομένως κάθε URI επιστρέφει μια HTML σελίδα. Στην Εικόνα 4 φαίνεται ένα παράδειγμα HTML σελίδας μετά από ανακατεύθυνση στο κατηγόρημα `diana:species` του λεξικού που χρησιμοποιήθηκε για τη μοντελοποίηση του συνόλου δεδομένων DIANA. Πρόβλημα με την ανακατεύθυνση μπορεί να προκύψει με την επαναχρησιμοποίηση όρων που προέρχονται από ανοιχτά και δημόσια διαθέσιμα λεξικά (για περισσότερες πληροφορίες για την επαναχρησιμοποίηση λεξικών στο Κεφάλαιο 3.2). Το πρόβλημα δημιουργείται καθώς ο πάροχος των δεδομένων δεν έχει τον έλεγχο των ανακατευθυνόμενων όρων και βασίζεται στη σταθερότητα του παρόχου του λεξικού. Για να ελαχιστοποιήσουμε αυτό το ρίσκο και να επαναχρησιμοποιήσουμε λεξικά, καθώς συνίσταται στο σχεδιασμό Διασυνδεδεμένων Δεδομένων, χρησιμοποιήσαμε όρους που προέρχονται από γνωστούς και έμπιστους παρόχους λεξικών. Ο Πίνακας 3 δείχνει όλα τα λεξικά που χρησιμοποιήθηκαν για τα σύνολα δεδομένων AI4B και DIANA. Από αριστερά προς τα δεξιά φαίνονται το όνομα του λεξικού, το πρόθεμά του, η περιγραφή και το URI του.



Εικόνα 3: : Η HTML σελίδα από ανακατεύθυνση ενός πόρου hairpin miRNA από το σύνολο δεδομένων DIANA.



Εικόνα 4: Η HTML σελίδα που επιστρέφει η ανακατεύθυνση στο κατηγόρημα `diana:species` του λεξικού που σχεδιάστηκε για το σύνολο δεδομένων DIANA.

Πίνακας 3: Γνωστά και έμπιστα λεξικά που χρησιμοποιήθηκαν στα σύνολα δεδομένων AI4B και DIANA

Vocabulary Name	**Prefix**	**Description**	**URI**
RDF Syntax	`rdf:`	RDF Syntax provides the basic RDF vocabulary.	http://www.w3.org/1999/02/22-rdf-syntax-ns#
RDF Schema	`rdfs:`	RDF Schema provides a data-modelling vocabulary for RDF data.	http://www.w3.org/2000/01/rdf-schema#>
OWL	`owl:`	OWL describes the built-in classes and properties that together form the basis of the RDF/XML syntax of OWL 2.	http://www.w3.org/2002/07/owl#
PROV-O	`prov:`	PROV-O expresses provenance information modeled by the PROV Data Model.	http://www.w3.org/TR/prov-o/
FOAF	`foaf:`	FOAF vocabulary described people and linking them with information.	http://xmlns.com/foaf/spec/
Dublin Core	`dcterms:`	Dublin Core is a vocabulary that describes web resources and physical resources.	http://dublincore.org/documents/dcmi-terms/

3.1.2 Ιεραρχικά URIs

Η χρήση ιεραρχικών URIs συστήνεται όταν θέλουμε να χρησιμοποιήσουμε δεξαμενές δεδομένων (containers) ή όταν θέλουμε να εκφράσουμε μια φυσική ιεραρχία που υπάρχει στα δεδομένα μας, όπως για παράδειγμα η σχέση πατέρα-παιδιού. Μεταφέροντας αυτή την ιεραρχία και στη δομή των URIs τα κάνει πιο κατανοητά και οι χρήστες μπορούν να πλοηγούνται εύκολα προς τα πάνω στην ιεραρχία διαγράφοντας μέρη του URI.

Χρησιμοποιήσαμε ιεραρχικά URIs μόνο στο σχεδιασμό του συνόλου δεδομένων DIANA. Πιο συγκεκριμένα, ιεραρχικά URIs χρησιμοποιήθηκαν για την μοντελοποίηση των εξελισσόμενων οντοτήτων hairpin και ώριμων miRNAs. Τα URIs που δημιουργήθηκαν με το pattern *`http://{domain}/{concept}/{identifier}`* αναπαριστούν τη γενική αναπαράσταση μια συγκεκριμένης έννοιας και του μοναδικού αναγνωριστικού της. Χρησιμοποιώντας ένας τέτοια URIs, μπορούμε να αντλήσουμε την περιγραφή RDF για την τελευταία και την τρέχουσα έκδοση ενός πόρου RDF. Καθώς οι έννοιες που μας ενδιαφέρουν εξελίσσονται, προσθέτουμε timestamps στο τέλος των προηγούμενων URIs έτσι ώστε να δημιουργηθεί μια ιεραρχία. Οπότε τα URIs που δημιουργήθηκα με το ακόλουθο pattern `http://{domain}/{concept}/{identifier}/{timestamp}` αναπαριστούν συγκεκριμένες εκδόσεις μια έννοιας σε συνδυασμό με ένα μοναδικό αναγνωριστικό. Για παράδειγμα, το URI …`/hairpins/MI0001364` είναι η γενική αναπαράσταση της hairpin miRNA οντότητας με μοναδικό αναγνωριστικό το MI0001364 της miRBase βάσης δεδομένων και μπορεί να χρησιμοποιηθεί για την ανάκτηση της τελευταίας έκδοσης του hairpin miRNA, που είναι η miRBase έκδοση 18. Αντίθετα το URI …`/hairpins/MI0001364/12.0` αναπαριστά την ίδια hairpin miRNA οντότητα για την έκδοση miRBase 12.0.

3.1.3 Σχετικά URIs

Η χρήση των σχετικών URIs προσφέρει διάφορα πλεονεκτήματα στις πλατφόρμες Διασυνδεδεμένων Δεδομένων, όπως είναι (α) το μειωμένο μήκος σε σχέση με τα ολόκληρα URIs, (β) τόσο οι πόροι (resources) RDF όσο και οι ιδιότητες (properties) RDF γίνονται πιο εύκολο να απομνημονευτούν και να χρησιμοποιηθούν, (γ) διευκολύνεται η διαδικασία αυτόματης δημιουργίας των URIs και κυρίως η διαδικασία αναπαράστασης της εξέλιξής τους. Στη διαδικασία σχεδιασμού και ανάπτυξης των συνόλων δεδομένων AI4B και DIANA, κάνουμε χρήση σχετικών URIs για να επωφεληθούμε από τα ανωτέρω πλεονεκτήματα.

3.2 Λεξικά και Τύποι

Μια από τις βασικές αρχές των Διασυνδεδεμένων Δεδομένων είναι οι συνδέσεις μεταξύ των RDF οντοτήτων ή μεταξύ οντοτήτων και literal τιμών, συνδέσεις που υποδηλώνουν σημασιολογική συσχέτιση (semantic connotations). Οι σημασιολογικές συσχετίσεις δεν περιορίζονται μόνο στις συνδέσεις, αλλά χρησιμοποιούνται και για την περιγραφή RDF οντοτήτων και literal τιμών. Όλες αυτές οι συσχετίσεις – γνωστές και ως όροι (terms) – που χρησιμοποιούνται για τη μοντελοποίηση ενός συνόλου δεδομένων ως διασυνδεδεμένου μπορούν να προέλθουν από τον ορισμό ενός νέου λεξικού ή από την επαναχρησιμοποίηση ενός υπάρχοντος. Για να ενισχυθεί η διατηρησιμότητα χώρων ΑΔΔ σχετικά με τη μοντελοποίηση, οι ακόλουθες πρακτικές συχνά προτείνονται στη βιβλιογραφία (α) επαναχρησιμοποίηση λεξικών, (β) η χρήση του όρου `rdf:type` και (γ) η χρήση ενός XML σχήματος τύπων δεδομένων για τη ορισμό literal τιμών.

3.2.1 Επαναχρησιμοποίηση Λεξικών

Η επαναχρησιμοποίηση λεξικών είναι η πιο ευρέως προτεινόμενη πρακτική όταν δημοσιεύουμε δεδομένα ως Διασυνδεδεμένα Δεδομένα, καθώς έχει μια σειρά από πλεονεκτήματα. Αρχικά, ενισχύει τη συνδεσιμότητα και υποβοηθά την εξερεύνηση με υπάρχοντες μηχανισμούς. Δεύτερον, υποστηρίζει τη διαλειτουργικότητα (interoperability) των δεδομένων καθώς τα καθιερωμένα λεξικά έχουν γνωστές συσχετίσεις (connotations) και οι υπάρχοντες μηχανισμοί μπορούν να επεξεργαστούν τα δεδομένα χωρίς να αλλάξει το status τους. Τρίτον, προσθέτει αξιοπιστία στο σχήμα, προωθώντας το δεδομένα αυτά καθαυτά. Τελικά, η επαναχρησιμοποίηση λεξικών κάνει τα πράγματα πιο εύκολα καθώς δεν χρειάζεται να αναληφθεί δουλειά που έχει ήδη γίνει.

Εκτός από τα πλεονεκτήματα που περιγράψαμε παραπάνω, η επαναχρησιμοποίηση λεξικών έχει το ρίσκο των ανακατευθυνόμενων URIs (βλέπε Κεφάλαιο 3.1.1). Για να ελαχιστοποιήσουμε το ρίσκο αυτό, στο σχεδιασμό του AI4B και του DIANA χρησιμοποιήθηκαν όροι που προέρχονται από έγκυρα και άκρως αξιόπιστα λεξικά. Τα λεξικά αυτά φαίνονται στον Πίνακας 3. Ο Πίνακας 4 από τα αριστερά προς τα δεξιά δείχνει το όνομα του λεξικού, τους ακριβείς όρους που χρησιμοποιήθηκαν καθώς και την περιγραφή τους.

Πίνακας 4: Όροι από έγκυρα λεξικά που χρησιμοποιήθηκαν στα σύνολα δεδομένων AI4B και DIANA

Vocabulary Name	Terms	Description
RDF Syntax	`rdf:type`	A property stating that a resource is an instance of an RDF class.
	`rdf:Property`	A class declaring RDF properties. `rdf:Property` is an instance of `rdfs:Class`.
Dublin Core	`dcterms:description`	A property providing a description of the resource.
	`dcterms:title`	A property providing a name to the resource.
	`dcterms:identifier`	A property indicating an unambiguous reference to the resource within a given context.
	`dcterms:BibliographicResource`	A class declaring a bibliographic resource, such as a book, an article, or any other documentary resource.
	`dcterms:creator`	A property declaring the primarily responsible entity for making the resource.
	`dcterms:date`	A property declaring the point or period of time associated with an event in the lifecycle of the resource.
OWL	`owl:sameAs`	A property linking a resource to another resource indicating that two URI references actually refer to the same thing.
	`prov:wasAssociatedWith`	A property assigning responsibility to an agent for an activity, indicating that the agent had a role in the activity.
	`prov:wasAttributedTo`	A property ascribing of an entity to an agent.

PROV-O	`prov:Agent`	A class declaring that something bears some form of responsibility for an activity taking place, for the existence of an entity, or for another agent's activity.
	`prov:Entity`	A class declaring that the entity is a physical, digital, conceptual, or other kind of thing with some fixed aspects; entities may be real or imaginary.
	`prov:Activity`	A class declaring that an activity is something that occurs over a period of time and acts upon or with entities; it may include consuming, processing, transforming, modifying, relocating, using, or generating entities.
	`prov:used`	A property declaring the beginning of utilizing an entity by an activity. Before usage, the activity had not begun to utilize this entity and could not have been affected by the entity.
FOAF	`foaf:Document`	A class representing those things which are, broadly conceived, as documents.
	`foaf:homepage`	A property representing a homepage for something.
RDF Schema	`rdfs:label`	A property providing a human-readable version of a resource's name.
	`rdfs:subPropertyOf`	A property stating that all resources related by one property are also related by another. It declares hierarchy.
	`rdfs:subClassOf`	A property stating that all the instances of one class are instances of another.
		A class declaring RDF classes. `rdfs:Class` is an instance

	rdfs:Class	of rdfs:Class.
	rdfs:Literal	A class declaring literal values such as strings and integers. Property values such as textual strings are examples of RDF literals.

3.2.2 Χρήση του rdf:type

Ο όρος rdf:type είναι ένας επαναχρησιμοποιήσιμος όρος που η χρήση του προτείνεται εκτενώς από τη βιβλιογραφία. Ο όρος αυτός έχει μεγάλη σημασία σε ένα σύνολο δεδομένων καθώς ορίζει την κλάση στην οποία ανήκει μια οντότητα. Επιπλέον βοηθάει στη δημιουργία του σχήματος που μοντελοποιεί τα δεδομένα. Τόσο στο AI4B όσο και στο DIANA, κάνουμε εκτενή χρήση του όρου αυτού, ορίζοντας για κάθε οντότητα τον τύπο της RDF κλάσης της. Ο Πίνακας 5 δείχνει τις πιο σημαντικές RDF κλάσεις του AI4B συνόλου ΑΔΔ και Πίνακας 6 για το DIANA αντίστοιχα. Να σημειωθεί ότι στη δεύτερη στήλη των πινάκων αυτών δίνουμε πληροφορίες για ιεραρχικές κλάσεις όπου υπάρχουν.

Πίνακας 5: Οι πιο σημαντικές RDF κλάσεις στο σύνολο δεδομένων AI4B

RDF class	**Description**
sym:BiomassProduct	A class representing a biomass, natural product, e.g. fruit residues.
sym:OtherMaterial	A class representing other materials, e.g. biogas.
sym:BiomassOffer	A class representing an offer of a biomass product.
sym:BiomassDemand	A class representing a demand of a biomass product.
sym:EnablingTechnology	A class representing a technology that converts a biomass product into another material.

Πίνακας 6: Οι πιο σημαντικές RDF κλάσεις στο σύνολο δεδομένων DIANA.

RDF class	**Subclass Of**	**Description**
diana:Hairpin	prov:Entity	A class representing a hairpin microRNA.
diana:Mature	prov:Entity	A class representing a mature microRNA.
diana:AddHairpinParent	prov:Activity	A class representing a new relationship between a hairpin microRNA and a mature microRNA.
diana:RemoveHairpinParent	prov:Activity	A class representing the end of a relationship between a

		hairpin microRNA and a mature microRNA.
`diana:Species`		A class representing a species.
`diana:Gene`		A class representing a gene.
`diana:Interaction`		A class representing an interaction between a mature miRNA and a gene's transcript.
`diana:Transcript`		A class representing the transcript of a gene.
`diana:ScientificDatabase`	`prov:Agent`	A class representing a scientific database.

3.2.3 Τύποι δεδομένων literal τιμών

Τα RDF literals αναπαριστούν τιμές όπως αλφαριθμητικά, νούμερα και ημερομηνίες. Ο ορισμός τύπων δεδομένων που συσχετίζουν οντότητες με literal τιμές είναι μια ακόμη συνήθης πρακτική που βοηθάει στη διατηρησιμότητα των χώρων ΑΔΔ. Η πρακτική αυτή βοηθάει επίσης την ερμηνεία της πληροφορίας που έχει μοντελοποιηθεί και την αναγνώριση ασυνεπειών που μπορεί να προκύψουν. Η γενίκευση των τύπων δεδομένων που χρησιμοποιείται στο RDF είναι σύμφωνη με το XML Schema [Peterson2012]. Κάθε τύπος δεδομένων που συμφωνεί με αυτή τη γενίκευση μπορεί να χρησιμοποιηθεί στο RDF, ακόμα και αν δεν ορίζεται στο XML Schema. Για όλους τους ανωτέρω λόγους, ορίζουμε όλες τις ιδιότητες που συνδέουν οντότητες με literal τιμές με ένα τύπο δεδομένων. Ο Πίνακας 7 δείχνει όλους τους διαφορετικούς τύπους δεδομένων που χρησιμοποιήθηκαν στο AI4B και στο DIANA μαζί με μια σύντομη περιγραφή τους. Να επισημάνουμε ότι το πρόθεμα `xsd:` αναφέρεται στο URI http://www.w3.org/2001/XMLSchema#.

Πίνακας 7: Οι τύποι δεδομένων που χρησιμοποιήθηκαν στα σύνολα δεδομένων AI4B και DIANA.

Data type	**Description**
`xsd:decimal`	It is used to specify a numeric value.
`xsd:string`	It is used for values that contains character strings.
`xsd:integer`	It is used to specify a numeric value without a fractional component.
`xsd:double`	It is used to specify a numeric value with a fractional component.
`xsd:gyear`	It is used to represent a specific calendar year. The letter g signifies "Gregorian."
`xsd:Boolean`	It is used to specify a true or false value.
`xsd:dateTime`	It is used to specify a date and a time. The dateTime is specified in the following form "YYYY-MM-DDThh:mm:ss" where YYYY indicates the year, MM indicates the month, DD indicates the day, T indicates the start of the required time section, hh indicates the hour,

	mm indicates the minute and ss indicates the second.
`xsd:date`	It is used to specify a date. The date is specified in the following form "YYYY-MM-DD" where YYYY indicates the year, MM indicates the month and DD indicates the day.

3.3 Συνδεσιμότητα

Μια από τις πιο θεμελιώδεις έννοιες των Διασυνδεδεμένων Δεδομένων είναι οι συνδέσεις ενός συνόλου δεδομένων με άλλα Διασυνδεδεμένα Δεδομένα που μπορεί να προέρχονται από ετερογενείς πόρους. Η συνδεσιμότητα επιτυγχάνεται με τη σύνδεση δύο RDF οντοτήτων με μια ιδιότητα (property). Η σημασιολογία των ιδιοτήτων αυτών μπορεί να οριστεί στο νέο λεξικό που δημιουργείται για το σχεδιασμό του συγκεκριμένου συνόλου δεδομένων ή μπορεί να προέλθει από την επαναχρησιμοποίηση λεξικών όπως περιγράψαμε στο Κεφάλαιο 3.2.1. Μια από τις ιδιότητες που χρησιμοποιούνται περισσότερο για τη σύνδεση δυο οντοτήτων είναι η `owl:sameAs`. Αποτελεί μια built-in OWL ιδιότητα που συνδέει δυο οντότητες και ορίζει ότι τα URIs αναφέρονται στο ίδιο αντικείμενο: τα αντικείμενα έχουν την ίδια «ταυτότητα» [Dean2004].

Η υψηλή συνδεσιμότητα βοηθάει την ολοκλήρωση των δεδομένων, την ερμηνεία των δεδομένων καθώς και την εξερεύνησή τους. Η χρησιμοποίηση συνδέσμων είναι ένας φυσικός τρόπος να ενσωματωθούν δεδομένα από διάφορες πηγές. Επιπλέον, σύνδεσμοι όπως το `owl:sameAs`, βοηθούν την ερμηνεία των δεδομένων χωρίς το επιπρόσθετο κόστος της επανάληψης και της συντήρησης. Τέλος, υπάρχοντες μηχανισμοί εξερεύνησης δεδομένων μπορούν να ανακαλύψουν τα πρόσφατα δημοσιευμένα δεδομένα ακολουθώντας τις συνδέσεις. Η συνδεσιμότητα σχετίζεται συχνά με το ρίσκο της ανακατεύθυνσης (βλέπε Κεφάλαιο 3.1.1), καθώς η συντήρηση των ανακατευθύνομενων URIs βασίζεται στις συνδεδεμένες πηγές. Παρά το ρίσκο, η συνδεσιμότητα είναι μια πολύ σημαντική έννοια για τα Διασυνδεδεμένα Δεδομένα.

Στο σχεδιασμό μας, συνδέουμε τα σύνολα δεδομένων μας με επαναχρησιμοποίηση λεξικών όπως είδαμε εκτενώς στον Πίνακας 4. Επιπρόσθετα στο DIANA, χρησιμοποιήθηκε και το `owl:sameAs` όπως φαίνεται στην Εικόνα 5.

4 Σύνοψη

Σε αυτή την αναφορά, εξετάσαμε το σχεδιασμό του AI4B και DIANA ως διατηρήσιμους χώρους ΑΔΔ. Και στα δύο σύνολα δεδομένων εφαρμόσαμε ευρέως διαδεδομένες πρακτικές για την εξασφάλιση της διατηρησιμότητας, όπως η επαναχρησιμοποίηση λεξικών, οι τύποι δεδομένων και η σύνδεση με άλλα σύνολα δεδομένων. Επιπλέον, ελαχιστοποιήσαμε το ρίσκο της αδυναμίας ανακατεύθυνσης των URIs με τη χρήση γνωστών και έγκυρων λεξικών.

http://leonardo.imis.athena-innovation.gr:8080/diana_lod/species/sme	http://bio2rdf.org/taxonomy:79327
http://leonardo.imis.athena-innovation.gr:8080/diana_lod/species/sly	http://bio2rdf.org/taxonomy:4081
http://leonardo.imis.athena-innovation.gr:8080/diana_lod/species/ghr	http://bio2rdf.org/taxonomy:3635
http://leonardo.imis.athena-innovation.gr:8080/diana_lod/species/sof	http://bio2rdf.org/taxonomy:286192
http://leonardo.imis.athena-innovation.gr:8080/diana_lod/species/pbi	http://bio2rdf.org/taxonomy:61621
http://leonardo.imis.athena-innovation.gr:8080/diana_lod/species/ssy	http://bio2rdf.org/taxonomy:405039
http://leonardo.imis.athena-innovation.gr:8080/diana_lod/species/oar	http://bio2rdf.org/taxonomy:9940
http://leonardo.imis.athena-innovation.gr:8080/diana_lod/species/bol	http://bio2rdf.org/taxonomy:3712
http://leonardo.imis.athena-innovation.gr:8080/diana_lod/species/lmi	http://bio2rdf.org/taxonomy:7004
http://leonardo.imis.athena-innovation.gr:8080/diana_lod/species/bra	http://bio2rdf.org/taxonomy:51350
http://leonardo.imis.athena-innovation.gr:8080/diana_lod/species/lca	http://bio2rdf.org/taxonomy:9447
http://leonardo.imis.athena-innovation.gr:8080/diana_lod/species/lja	http://bio2rdf.org/taxonomy:34305
http://leonardo.imis.athena-innovation.gr:8080/diana_lod/species/hru	http://bio2rdf.org/taxonomy:6454
http://leonardo.imis.athena-innovation.gr:8080/diana_lod/species/sja	http://bio2rdf.org/taxonomy:6182
http://leonardo.imis.athena-innovation.gr:8080/diana_lod/species/cla	http://bio2rdf.org/taxonomy:6221
http://leonardo.imis.athena-innovation.gr:8080/diana_lod/species/bkv	http://bio2rdf.org/taxonomy:10629
http://leonardo.imis.athena-innovation.gr:8080/diana_lod/species/vun	http://bio2rdf.org/taxonomy:3917
http://leonardo.imis.athena-innovation.gr:8080/diana_lod/species/cpa	http://bio2rdf.org/taxonomy:3649
http://leonardo.imis.athena-innovation.gr:8080/diana_lod/genes/AC3.5	http://bio2rdf.org/ensembl:AC3.5
http://leonardo.imis.athena-innovation.gr:8080/diana_lod/genes/AH6.10	http://bio2rdf.org/ensembl:AH6.10
http://leonardo.imis.athena-innovation.gr:8080/diana_lod/transcripts/AH6.10	http://bio2rdf.org/ensembl:AH6.10
http://leonardo.imis.athena-innovation.gr:8080/diana_lod/genes/AH6.12	http://bio2rdf.org/ensembl:AH6.12
http://leonardo.imis.athena-innovation.gr:8080/diana_lod/transcripts/AH6.12	http://bio2rdf.org/ensembl:AH6.12
http://leonardo.imis.athena-innovation.gr:8080/diana_lod/genes/AH6.14	http://bio2rdf.org/ensembl:AH6.14
http://leonardo.imis.athena-innovation.gr:8080/diana_lod/transcripts/AH6.14	http://bio2rdf.org/ensembl:AH6.14
http://leonardo.imis.athena-innovation.gr:8080/diana_lod/genes/AH6.2	http://bio2rdf.org/ensembl:AH6.2
http://leonardo.imis.athena-innovation.gr:8080/diana_lod/transcripts/AH6.2	http://bio2rdf.org/ensembl:AH6.2

Εικόνα 5: Η χρήση του owl:sameAs στο DIANA.

Βιβλιογραφία

[Bizer2008] Bizer, Christian and Schultz, Andreas. “Benchmarking the performance of storage systems that expose SPARQL endpoints.” In Proceedings of the ISWC Workshop on Scalable Semantic Web Knowledgebase, (2008)

[Bizer2009] Bizer, Christian, Heath, Tom and Berners-Lee, Tim. "Linked Data - The Story So Far." International Journal on Semantic Web and Information Systems 5, no. 3 (2009): 1--22.

[Dean2004] Dean, Mike and Schreiber, Guus. “OWL Web Ontology Language Reference.” (2004), W3C Recommendation, URL: http://www.w3.org/TR/owl-ref/

[Faye2012] Faye, David, Cure, Olivier and Blin, Guillaume. “A survey of RDF storage approaches.” ARIMA Journal, vol. 15, (2012): 11—35.

[Peterson2012] Peterson, David, Gao, Sandy, Malhotra, Ashok, Sperberg-McQueen, Michael, Thompson, Henry, Biron, Paul V.et al. “W3C XML Schema Definition Language (XSD) 1.1 Part 2: Datatypes.” (2012), W3C Recommendation. URL: http://www.w3.org/TR/xmlschema11-2/