«DARIAH-ΑΤΤΙΚΗ Ανάπτυξη της ελληνικής ερευνητικής υποδομής για τις ανθρωπιστικές επιστήμες ΔΥΑΣ»

ΑΚΑΔΗΜΙΑ ΑΘΗΝΩΝ	ΕΡΕΥΝΗΤΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΑΚΟ ΙΝΣΤΙΤΟΥΤΟ
ΕΚ ΑΘΗΝΑ	ΣΥΣΤΗΜΑΤΩΝ ΕΠΙΚΟΙΝΩΝΙΩΝ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ
ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ	ΑΝΩΤΑΤΗ ΣΧΟΛΗ ΚΑΛΩΝ ΤΕΧΝΩΝ
	ΙΔΡΥΜΑ ΤΕΧΝΟΛΟΓΙΑΣ ΚΑΙ ΕΡΕΥΝΑΣ

ΥΠΟΕΡΓΟ 4	Δράσεις ανάπτυξης της εθνικής υποδομής του ΕΠΙΣΕΥ
Παραδοτέο	ΠΑ.1.4.ΕΠΙΣΕΥ.1 Συστήματα αντιστοίχισης μεταδεδομένων, συλλογής και διαλειτουργικής διαχείρισης του περιεχομένου
Έκδοση 1.0-Τελική 31/12/2014	**Έγγραφο:**

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Υπουργείο Παιδείας & Θρησκευμάτων
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
ΥΠΟΥΡΓΕΙΟ ΑΝΑΠΤΥΞΗΣ & ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑΣ
ΕΠΑΝ II
ΕΣΠΑ 2007-2013
η περιφέρεια στο επίκεντρο της ανάπτυξης
Με τη συγχρηματοδότηση της Ελλάδας και του Ευρωπαϊκού Ταμείου Περιφερειακής Ανάπτυξης της Ευρωπαϊκής Ένωσης στο πλαίσιο του Ε.Π. Ανταγωνιστικότητα & Επιχειρηματικότητα και των Π.Ε.Π. Αττικής, Π.Ε.Π. Μακεδονίας Θράκης, Π.Ε.Π. Κρήτης & Νήσων Αιγαίου, Π.Ε.Π. Θεσσαλίας-Στερεάς Ελλάδας-Ηπείρου

Ταυτότητα Εγγράφου

ΥΠΟΕΡΓΟ	4. Δράσεις ανάπτυξης της εθνικής υποδομής του ΕΠΙΣΕΥ
Ενότητα εργασιών	1. ΚΟΙΝΩΝΙΑ ΨΗΦΙΑΚΩΝ ΠΟΡΩΝ
ΔΡΑΣΗ	1.4. Μητρώο λογισμικών υπηρεσιών
ΠΑΡΑΔΟΤΕΟ	ΠΑ.1.4.ΕΠΙΣΕΥ.1 Συστήματα αντιστοίχισης μεταδεδομένων, συλλογής και διαλειτουργικής διαχείρισης του περιεχομένου
Εργασία	Συγγραφή εγγράφου
Ομάδα Έργου	Στέφανος Κόλλιας Γεώργιος Μαρανδιανός Φοίβος Μυλωνάς Αλέξανδρος Χορταράς
Χαρακτηρισμός	Εσωτερικό Ομάδας Έργου

Ιστορικό Αλλαγών

Έκδ.	Ημ/μηνία	Αιτιολογία	Σύνοψη Αλλαγών	Επισπεύδοντες
1.0	31/12/2014	Τελική	Δημιουργία και Σύνταξη	Ομάδα Έργου

Πίνακας Περιεχομένων

Εισαγωγή

Το παραδοτέο αυτό περιγράφει τα βασικά σχήματα μεταδεδομένων και τις οντολογίες που χρησιμοποιούνται για την απεικόνιση των δεδομένων των παρόχων περιεχομένου, καθώς και το κοινό μοντέλο που χρησιμοποιείται στο πλαίσιο της Europeana και στο οποίο αντιστοιχίζονται όλα τα ετερογενή σχήματα μεταδεδομένων των παρόχων περιεχομένου.
Το παρόν Παραδοτέο είναι το πρώτο της Δράσης για Μελέτη και Δημιουργία «Μητρώου Λογισμικών Υπηρεσιών» που υλοποιεί το Ε.Π.Ι.Σ.Ε.Υ. στο πλαίσιο του έργου DARIAH-Αττική.

Σκοπός

Το παρόν Παραδοτέο επικεντρώνεται στη συλλογή και διαχείριση των μεταδεδομένων, που περιγράφουν πολιτιστικά τεκμήρια από τους παρόχους περιεχομένου, με διαλειτουργικό τρόπο, με στόχο την παροχή προηγμένων υπηρεσιών σημασιολογικής αναζήτησης και απάντησης στα επερωτήματα των χρηστών.

Ομάδες Στόχοι

Το παραδοτέο απευθύνεται τόσο στα μέλη των φορέων του έργου DARIAH-ΑΤΤΙΚΗ, και στους ερευνητές των φορέων αυτών που απασχολούνται στο πρόγραμμα, όσο και στους φορείς-χρήστες και στους ερευνητές που θα χρησιμοποιήσουν τα αποτελέσματα του έργου.

1. Πρότυπα Μεταδεδομένων

Παραδοσιακά, οι πρωτοβουλίες μεταδεδομένων στον πολιτισμικό τομέα εστιάζονται σε τρεις βασικούς τομείς: μουσεία, αρχεία και βιβλιοθήκες. Στο παρελθόν, κάθε τομέας (ή τμήμα αυτού) ανέπτυσσε ίδια πρότυπα, όπου παρουσιαζόταν το φαινόμενο πρότυπα ενός τομέα να μην χρησιμοποιούνται από άλλο τομέα (ακόμα και συναφούς δραστηριότητας). Στην πραγματικότητα, μέχρι πρόσφατα, η συνεργασία μεταξύ φορέων αποτελούσε την εξαίρεση παρά τον κανόνα. Οι βιβλιοθήκες ήταν οι πρώτοι πολιτισμικοί οργανισμοί που ανέπτυξαν συνεπή πρότυπα προκειμένου να υποστηρίξουν την απαίτηση ανταλλαγής εγγραφών καταλόγων μεταξύ συστημάτων. Οι υπόλοιποι τομείς διακρίνονταν από μικρότερης εντάσεως ανάγκη για διαμοιρασμό πληροφορίας. Με την έλευση όμως των τεχνολογιών επικοινωνίας και το όραμα της Κοινωνίας της Πληροφορίας, έχουν αναπτυχθεί διάφορα πρότυπα, που απευθύνονται σε όλες τις ανάγκες και το περιεχόμενο των πολιτισμικών οργανισμών.

Η πληροφορία για ένα ψηφιακό ή φυσικό αντικείμενο (δηλαδή τα μεταδεδομένα) πρέπει να εκφραστεί με ένα συγκεκριμένο τρόπο (δηλαδή ένα πρότυπο). Η υιοθέτηση ενός προτύπου μεταδεδομένων εξασφαλίζει ότι οι πόροι που έχουν περιγραφεί βάσει του συγκεκριμένου προτύπου θα είναι αναζητήσιμοι, ανεξάρτητα από τις ιδιαιτερότητες του παροχέα τους ή του οργανισμού που τους δημιούργησε. Ένα κοινό πρότυπο επιτρέπει επίσης σε συστήματα να μεταφέρουν τα χαρακτηριστικά των ψηφιακών πόρων σε άλλες ηλεκτρονικές εφαρμογές ή εργαλεία αναζήτησης. Επομένως, ένα πρότυπο επιτρέπει την αποτελεσματική κοινωνία της πληροφορίας από μία εφαρμογή ή σύστημα αναζήτησης σε άλλο, συντελώντας με αυτόν τον τρόπο στην επίτευξη διαλειτουργικότητας.

Θα εξετάσουμε σε αυτό το σημείο 4 βασικές δομές μεταδεδομένων. Οι δομές μεταδεδομένων δεν μπορούν να θεωρηθούν ως οντολογίες, δεδομένου ότι οι ορισμοί των στοιχείων που απαρτίζουν ένα πλαίσιο μεταδεδομένων στερούνται κάθε τυπικής προσέγγισης για τη σύλληψη ενός εννοιολογικού πεδίου. Για παράδειγμα, το πεδίο «*publisher*» του προτύπου Dublin Core, που θα εξετάσουμε παρακάτω, μπορεί να ερμηνευτεί ποικιλοτρόπως. Ενώ οι έννοιες μίας οντολογίας έχουν νόημα εκτός του περιβάλλοντος μίας δομής δεδομένων, τα στοιχεία ενός πεδίου μεταδεδομένων έχουν νόημα μόνο σε μία συγκεκριμένη ιεραρχία στοιχείων.

1.1 Dublin Core

Η πρωτοβουλία για το πρότυπο μεταδεδομένων **Dublin Core**[1] (Dublin Core Metadata Initiative) αποτελεί ένα φόρουμ αφιερωμένο στην ανάπτυξη online διαλειτουργικών προτύπων μεταδεδομένων, που υποστηρίζουν ένα ευρύ φάσμα σκοπών και επιχειρησιακών μοντέλων. Η πρωτοβουλία ανέπτυξε το Dublin Core Σύνολο Στοιχείων Μεταδεδομένων (Dublin Core Metadata Element Set) το 1995/1996, με στόχο την υποστήριξη απλής ανακάλυψης πόρων για ψηφιακές συλλογές σε διαφορετικά θεματικά πεδία. Αρχικό στόχο του προτύπου Dublin Core [DC04] αποτέλεσε η γεφύρωση του χάσματος ανάμεσα σε πρακτικές που ακολουθούνται σε αρχεία, βιβλιοθήκες και μουσεία, μειώνοντας τον αριθμό των στοιχείων για την περιγραφή ενός πολιτισμικού αντικειμένου σε 15. Το Dublin Core αποτελεί την πρώτη προσπάθεια ανάπτυξης ενός κοινού συνόλου στοιχείων που θα μπορούσαν να χρησιμοποιηθούν με συνέπεια για να περιγράψουν δικτυακούς πληροφοριακούς πόρους. Η χρησιμότητα και η επιτυχία του Dublin Core έγκειται στην απλότητα και την ευελιξία του: είναι αρκετά πλούσιο για να υποστηρίξει την αποτελεσματική αναζήτηση βάσει πεδίων, αλλά είναι επίσης απλό, ώστε να μην χρειάζεται εμπειρία ειδικών ή εκτενή χειρωνακτική προσπάθεια για τη δημιουργία εγγραφών. Το έτος 2000, το βασικό σύνολο στοιχείων επεκτάθηκε με την προσθήκη των Dublin Core Qualifiers για χρήση εντός τοπικών εφαρμογών ή συγκεκριμένων πεδίων. Το σύνολο αυτών των qualifiers δίνει στους οργανισμούς τη δυνατότητα να περιγράψουν επιπλέον το πολιτισμικό αντικείμενο.

Ειδικότερα, το Dublin Core ορίζει ένα σύνολο δεκαπέντε (15) βασικών στοιχείων, τα οποία είναι ευρέως χρήσιμα για την ανακάλυψη και ανάκτηση πόρων από διαφορετικά συστήματα. Τα στοιχεία μεταδεδομένων χωρίζονται σε τρείς ομάδες που κατά προσέγγιση υποδεικνύουν την κατηγορία ή τον σκοπό της πληροφορίας που περιγράφουν: (1) στοιχεία σχετικά κυρίως με το περιεχόμενο του πόρου: Title, Subject, Description, Source, Language, Relation, Coverage (2) στοιχεία σχετικά κυρίως με την πηγή ως πνευματική ιδιοκτησία: Creator, Publisher, Contributor, Rights και (3) στοιχεία σχετικά κυρίως με το στιγμιότυπο της πηγής: Date, Type, Format, Identifier. Κάθε στοιχείο είναι προαιρετικό και επαναλαμβανόμενο. Τα στοιχεία αυτά μπορούν να χρησιμοποιηθούν για να

[1] http://dublincore.org/

προστεθούν μεταδεδομένα σε HTML αρχεία[2] (χρησιμοποιώντας την επικεφαλίδα <meta>), αλλά μπορούν επίσης να χρησιμοποιηθούν σε άλλα περιεχόμενα για να δημιουργήσουν βασικά μεταδεδομένα για ένα ευρύ φάσμα ψηφιακών πόρων. Επίσης, εκτός από τα 14 αυτά στοιχεία το Dublin Core διαθέτει και ένα σύνολο όρων (DCTERMS) και τύπων (DCMI TYPE) για τη περιγραφή των μεταδεδομένων. Το σύνολο όρων DCTERMS ορίστηκε μετά το ορισμό των 15 βασικών στοιχείων του DC, δημιουργώντας τις ίδιες 15 σχέσεις αλλά με αυστηρώς ορισμένα πλέον τα πεδία ορισμού και τιμών. Το DCTERMS είναι εμπλουτισμένο με ένα πλήθος άλλων σχέσεων για καλύτερη περιγραφή του υλικού. Συνολικά περιέχει 40 ακόμα σχέσεις, μερίκες από αυτές είναι: η "Abstract" που ορίζεται ως η περίληψη του πόρου, η "Access Rights", που περιέχει πληροφορίες σχετικές με την πρόσβαση στον πόρο, η «Has Part" που χρησιμοποιείται για να περιγράψει άλλους πόρους που περιέχονται με φυσικό ή λογικό τρόπο στον περιγραφόμενο πόρο. Το σύνολο όρων DCMI TYPE περιέχει μια λίστα όρων (Collection, Dataset, Event, Image, Interactive Resource, Service, Software, Sound, Text**)** που μπορεί να χρησιμοποιηθεί για να περιγράψει με λεπτομέρεια το περιεχόμεντο του όρου Type. Το συγκεκριμένο πρότυπο, περιλαμβάνει δύο επίπεδα, το Simple Dublin Core και το Qualified Dublin Core. Το Simple Dublin Core χρησιμοποιεί 15 στοιχεία για την περιγραφή των τεκμηρίων, ενώ το Qualified Dublin Core χρησιμοποιεί τρία επιπλέον στοιχεία (Audience, Provenance, RightsHolder), ενώ ταυτόχρονα δίνει την δυνατότητα εισαγωγής προσδιοριστών (qualifiers), οι οποίοι βοηθούν στον καθορισμό της σημασιολογίας των στοιχείων με στόχο την ακριβέστερη αναζήτηση των ψηφιακών πόρων.

Το Dublin Core δεν επικεντρώνεται σε λεπτομερή διαχειριστικά ή τεχνικά μεταδεδομένα και για το λόγο αυτό είναι κατάλληλο για την αναζήτηση πόρων, παρά για την εσωτερική διαχείριση πόρων. Επιπλέον, εφόσον στόχος του είναι να αποτελέσει ένα απλό και ευρέως εφαρμόσιμο πρότυπο μεταδεδομένων για μια ποικιλία πόρων, το Dublin Core δεν παρέχει δομές για λεπτομερώς δομημένα μεταδεδομένα για συγκεκριμένους τύπους εγγράφων, όπως γίνεται στα πρότυπα TEI[3] και EAD, που θα εξετάσουμε παρακάτω. Για παράδειγμα, αν ένας χρήστης

[2] Παραδείγματα της HTML κωδικοποίησης του Dublin Core παρέχονται στο RFC 2731.

[3] Ο οργανισμός TEI (Text Encoding Initiative), που χρηματοδοτείται από τους οργανισμούς Association for Computers and the Humanities, Association for Computational Linguistics και Association for Literacy and Linguistic Computing,

αναζητεί πληροφορία μέσω διαδικτύου μπορεί να χρησιμοποιήσει το περιορισμένο λεξιλόγιο του Dublin Core για να αποκομίσει βασική πληροφορία σε μία γλώσσα που κατανοεί. Η πλήρης πρόσβαση στον πολιτισμό και τις υπηρεσίες του εξακολουθεί να απαιτεί γνώση του τοπικού λεξιλογίου και του περιβάλλοντος, αλλά ένα σύνολο απλών πληροφοριών κωδικοποιημένων στο Dublin Core μπορεί να επιστήσει την προσοχή του χρήστη σε μία ξένη πληροφοριακή πύλη, που σε άλλη περίπτωση θα διέφευγε της προσοχής του. Το Dublin Core μπορεί να αποθηκευτεί και να μεταφερθεί με ένα σύνολο τρόπων, συμπεριλαμβανομένων των HTML, XML, XML/RDF και σχεσιακών βάσεων δεδομένων.

Παρά την κριτική που έχει δεχτεί ότι αποτελεί περισσότερο ένα κακοσχεδιασμένο πρότυπο για την περιγραφή αντικειμένων πολιτισμικής κληρονομιάς, το Dublin Core χρησιμοποιείται όλο και περισσότερο και έχει υιοθετηθεί από ένα σύνολο κυβερνητικών οργανισμών, όπως εκείνους της Δανίας, του Ηνωμένου Βασιλείου και της Αυστραλίας. Συν τοις άλλοις, η Κοινοπραξία CIMI (CIMI Consortium) έχει εκτενώς ελέγξει το πρότυπο εντός του μουσειακού χώρου και έχει δημοσιεύσει τα αποτελέσματα [4] . Το Dublin Core ήδη υφίσταται σε περισσότερες από 20 μεταφράσεις, έχει υιοθετηθεί από τον οργανισμό CEN/ISSS (European Committee for Standardization / Information Society Standardization System) και τεκμηριώνεται σε δύο Internet RFC (Requests for Comments). Επίσης, έχει εγκριθεί ως αμερικανικό εθνικό πρότυπο (ANSI[5]/NISO[6] Z39.85), χρησιμοποιείται από περισσότερες από επτά κυβερνήσεις για την προώθηση της ανακάλυψης κυβερνητικής πληροφορίας σε ηλεκτρονική μορφή και έχει υιοθετηθεί από έναν

αναπτύσσει οδηγίες για την κωδικοποίηση και την ανταλλαγή μηχανικά αναγνώσιμων κειμένων στον τομέα των ανθρωπιστικών επιστημών. Έχει δύο βασικούς στόχους: (1) να καθορίσει μία κοινή μορφή ανταλλαγής και (2) να παράσχει ένα σύνολο προτάσεων για την κωδικοποίηση υλικού με τη μορφή κειμένου. Οι οδηγίες αυτές χρησιμοποιούνται όχι μόνο για να κωδικοποιήσουν έγγραφα, αλλά και αρχειακό υλικό. Το πρότυπο TEI, που εξυπηρετεί αυτούς τους στόχους, χρησιμοποιεί τη γλώσσα SGML, πρόγονο της XML.

[4] CIMI Guide to Best Practice: Dublin Core (http://www.cimi.org/public_docs/meta_bestprac_v1_1_210400.pdf)

[5] **ANSI** = American National Standards Institute. Ο ANSI αποτελεί ένα οργανισμό που διευκολύνει την ανάπτυξη αμερικανικών εθνικών προτύπων, επιτυγχάνοντας τη συμφωνία μεταξύ των ενδιαφερόμενων μερών (http://www.ansi.org/).

[6] **NISO** = National Information Standards Organization. Ο NISO είναι ένας πιστοποιημένος ANSI οργανισμός που αναπτύσσει πρότυπα ειδικά για τους τομείς των βιβλιοθηκών, των εκδόσεων και των υπηρεσιών πληροφορικής (http://www.niso.org/).

αριθμό παγκόσμιων οργανισμών, όπως ο Παγκόσμιος Οργανισμός Υγείας (World Health Organization-WHO).

Αν και το τρέχον πρότυπο μεταδεδομένων Dublin Core αποτελεί έναν καλό συμβιβασμό για την επίτευξη διαλειτουργικότητας, διαθέτει κάποιες αδυναμίες, η κυριότερη από τις οποίες είναι η απώλεια των συμφραζομένων (loss of context), που ισοδυναμεί με απώλεια πληροφορίας. Για το λόγο αυτό, απαιτείται η χρήση προτύπων με μεγαλύτερη εκφραστικότητα, η οποία θα προσφέρει λεπτομέρεια στα μεταδεδομένα και κατ' επέκταση καλύτερη ποιότητα αναζήτησης χωρίς απώλεια της διαλειτουργικότητας. Άλλωστε, το Dublin Core δεν έχει ως στόχο την αντικατάσταση άλλων προτύπων μεταδεδομένων, αλλά στοχεύει στη συνύπαρξη –ακόμα και σε επίπεδο περιγραφής του ίδιου πόρου – με πρότυπα μεταδεδομένων που προσφέρουν μεγαλύτερη σημασιολογία. Επιπλέον, πλουσιότερα σημασιολογικά σχήματα μπορούν να απεικονιστούν στο Dublin Core για την εξαγωγή δεδομένων ή την ταυτόχρονη αναζήτηση σε πολλαπλά συστήματα. Αντιστοίχως, απλές εγγραφές Dublin Core μπορούν να χρησιμοποιηθούν ως εναρκτήριο σημείο για τη δημιουργία συνθετότερων περιγραφών.

1.2. EAD – Encoded Archival Description

Η ανάπτυξη του **EAD** ξεκίνησε το 1993 από ένα ερευνητικό πρόγραμμα του πανεπιστημιακού οργανισμού University of California, Berkeley. Το EAD Document Type Definition[7] (DTD) αποτελεί ένα πρότυπο για την κωδικοποίηση αρχειακών ευρημάτων χρησιμοποιώντας τις γλώσσες SGML[8] ή/και XML, το οποίο

[7] Η έκδοση 2002 του EAD DTD έχει σχεδιαστεί για να λειτουργεί ταυτόχρονα ως SGML και XML DTD και συμμορφώνεται με όλες τις προδιαγραφές των SGML/XML (ISO 8879). Έχει επίσης δοκιμαστεί εκτενώς από μία ποικιλία υπαρχόντων SGML/XML λογισμικών. Το EAD DTD καθορίζει τους κανόνες για τον υπομνηματισμό ενός βοηθήματος ευρημάτων, ορίζοντας τα βασικά στοιχεία (π.χ abbr, edition, family name, label, language κτλ), τα υποστοιχεία και τα γνωρίσματα στοιχείων.

[8] Η γλώσσα υπομνηματισμού SGML (Standard Generalized Markup Language) (http://www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=16387) χρησιμοποιεί ορισμένες από το χρήστη ετικέτες για να υπομνηματίσει το περιεχόμενο ενός ψηφιακού εγγράφου. Οι SGML ετικέτες είναι λέξεις που περιέχονται εντός των χαρακτήρων «<» και «>», οι οποίοι περιβάλλουν το περιεχόμενο. Η σημασιολογία αυτών των ετικετών ορίζεται είτε στο ίδιο το

συντηρείται από τον οργανισμό Network Development and MARC Standards Office of the Library of Congress (LC) σε συνεργασία με την Ένωση Αμερικανών Αρχειοφυλάκων (Society of American Archivists). Το EAD είναι ένα πρότυπο δομής δεδομένων, που σημαίνει ότι ορίζει μόνο την ύπαρξη και ακολουθία των στοιχείων, αφήνοντας την επιλογή του περιεχομένου στον οργανισμό αρχειοθέτησης.

Ο στόχος που επιτυγχάνεται με τη χρήση του EAD είναι ότι καθιστά τους αρχειακούς πόρους από διάφορα ιδρύματα ευκολότερα προσπελάσιμους από τους χρήστες. Το EAD έχει επίσης ενθαρρύνει την κοινότητα των αρχείων στη συμφωνία επί προτύπων δομών δεδομένων. Το EAD αποτελεί ένα διεθνές πρότυπο που χρησιμοποιείται από αρχεία και βιβλιοθήκες για να κωδικοποιήσουν τις περιγραφές χειρογράφων και αρχείων, παρέχοντας ένα τυπικό τρόπο για την κωδικοποίηση δεδομένων σε ένα βοήθημα ευρημάτων (finding aid) για μεγαλύτερη προσβασιμότητα. Αυτά είναι οδηγοί συλλογών ή κατάλογοι απογραφής που αποκαλύπτουν την προέλευση της συλλογής, τον τρόπο με τον οποίο είναι οργανωμένη και τα αντικείμενα που περιέχει. Ειδικότερα, τα βοηθήματα ευρημάτων είναι λεπτομερείς οδηγοί που περιγράφουν συλλογές από αδημοσίευτα προσωπικά έγγραφα, οργανωτικά αρχεία και φωτογραφίες, βοηθώντας τους ερευνητές να αναγνωρίσουν και να εντοπίσουν κιβώτια ή φακέλους που τους ενδιαφέρουν για μελέτη. Επίσης, παρέχουν πληροφορία για τον οργανισμό, τα άτομα ή την οικογένεια που δημιούργησε τα έγγραφα ή τις φωτογραφίες, μία σύνοψη της συλλογής και της διαρρύθμισής της και ένα λεπτομερή κατάλογο περιεχομένων. Με τον τρόπο αυτό, τα βοηθήματα ευρημάτων λειτουργούν ως εργαλεία για την αρχειακή περιγραφή, της ενέργειας δηλαδή αναγνώρισης και καταγραφής του υλικού μίας συλλογής ή μίας ομάδας αρχείων. Η διαδικασία αρχειοθέτησης υλικού συνήθως αφορά την ταξινόμηση του υλικού σε σειρές-κατηγορίες, π.χ αλληλογραφία, θεματικές σειρές, αποκόμματα ή αποκόμματα εφημερίδων και την στη συνέχεια ταξινόμηση του υλικού σε μία

έγγραφο είτε στο πρόγραμμα μεταγλώττισης. Η λέξη «creator», για παράδειγμα, μπορεί να οριστεί ως μία ετικέτα, η οποία όταν χρησιμοποιείται έχει τη μορφή <CREATOR>Salvator Dali</CREATOR>. Βάσει σύμβασης, οι ετικέτες λειτουργούν όπως οι παρενθέσεις, περιβάλλοντας κείμενο. Η ετικέτα που προηγείται (<CREATOR>) δηλώνει την κατηγορία, ενώ η ετικέτα που έπεται (</CREATOR>) σηματοδοτεί το περιεχόμενο – τιμή του πεδίου περιγραφής.

ιεραρχία, την οποία αναπαριστά το EAD μέσω «φωλιάσματος» ή αναδρομής στοιχείων.

1.3. CIDOC CRM – CIDOC Conceptual Reference Model

Το **CIDOC CRM** (CIDOC Εννοιολογικό Μοντέλο Αναφοράς) αποτελεί ένα οντοκεντρικό σημασιολογικό μοντέλο, το οποίο αναπτύχθηκε από ειδική Ομάδα Εργασίας του ICOM/CIDOC. Από τον Οκτώβριο 2006, το CIDOC CRM αποτελεί πρότυπο ISO (ISO 21127), γεγονός που επιτεύχθηκε με τη συνεργασία της ομάδας CIDOC CRM SIG και της επιτροπής ISO/TC46/SC4/WG9. Το CIDOC Conceptual Reference Model αποτελεί μια τυπική οντολογία με στόχο την καταγραφή εννοιών που βρίσκονται στους ποικίλους τύπους δεδομένων που χρησιμοποιούνται για την μουσειακή τεκμηρίωση και για την πολιτισμική κληρονομιά. Πρωτεύων ρόλος του CIDOC CRM είναι να αποτελέσει μία βάση για τη διαμεσολάβηση πολιτισμικής πληροφορίας, παρέχοντας με τον τρόπο αυτό τη σημασιολογική «κόλλα» που απαιτείται για τον μετασχηματισμό των τρεχουσών διεσπαρμένων, τοπικών πληροφοριακών πηγών σε μία συνδεδεμένη και πολύτιμη καθολική πηγή. Ειδικότερα, το CIDOC CRM ορίζει και περιορίζεται από την υποκείμενη σημασιολογία σχημάτων βάσεων δεδομένων και δομών εγγράφων που χρησιμοποιούνται στην πολιτισμική κληρονομιά και στη μουσειακή τεκμηρίωση με όρους μιας τυπικής οντολογίας, ενώ δεν στοχεύει στην πρόταση του υλικού που πρέπει να τεκμηριωθεί από τους πολιτιστικούς οργανισμούς. Αντιθέτως, εξηγεί τη λογική του υλικού που στην ουσία τεκμηριώνεται, συντελώντας με αυτό τον τρόπο στη σημασιολογική διαλειτουργικότητα.

Το CIDOC CRM επιτρέπει την κοινή αναπαράσταση δεδομένων που συλλέγονται κάτω από διαφορετικές απόψεις, στόχους και τύπους και δίνει με αυτό τον τρόπο την δυνατότητα αμοιβαίας μετατροπής και ολοκλήρωσης. Αυτή η ασυνήθιστη γενική χρήση επιτυγχάνεται με δύο τρόπους: (α) το εννοιολογικό μοντέλο του CIDOC συνιστά ένα επεκτάσιμο δίκτυο από συσχετιζόμενα επίπεδα αφαίρεσης και (β) είναι οργανωμένο σε ένα σύνολο κατηγοριών που διέπονται από βασικές διακριτές σχέσεις. Ειδικότερα, το CRM απαρτίζεται από 81 κλάσεις και 132 μοναδικές ιδιότητες. Το CIDOC CRM μπορεί να υλοποιηθεί με οποιοδήποτε σχεσιακό ή οντοκεντρικό σχήμα και τα CRM στιγμιότυπα μπορούν επίσης να κωδικοποιηθούν με τη χρήση της RDF, της XML, της DAML+OIL, της OWL και άλλων γλωσσών.

Εν γένει, η οργάνωση του CRM παρέχει ένα αξιόπιστο πλαίσιο για τον σχεδιασμό πολιτισμικών πληροφοριακών συστημάτων και γενικά την αναπαράσταση της υλικής πολιτισμικής κληρονομιάς. Επί του παρόντος, το CIDOC CRM είναι η μόνη διεθνώς αναγνωρισμένη λύση για τη σημασιολογική διασύνδεση των διαφόρων μορφών πολιτισμικής πληροφορίας και για το λόγο αυτό προτείνεται ως βάση για την επίτευξη σημασιολογικής συμβατότητας.

Είναι σαφές ότι τα πρότυπα μεταδεδομένων που περιγράφηκαν παραπάνω έχουν αρκετά κοινά χαρακτηριστικά και εν μέρει συναφείς στόχους και χρησιμοποιούνται από διαφορετικούς φορείς για τον χαρακτηρισμό παρόμοιου είδους μεταδεδομένων. Έτσι, προκειμένου να επιτευχθεί διαλειτουργικότητα είναι απαραίτητο να υπολογιστούν οι σχέσεις και αντιστοιχίσεις μεταξύ των διαφόρων εννοιών που ορίζουν, με βάση αλγορίθμους αυτόματης και ημιαυτόματης στοίχισης σχημάτων μεταδεδομένων. Με τον τρόπο αυτό θα υπάρχει η δυνατότητα ενοποιημένης διαχείρισης ολόκληρου του περιεχομένου του συστήματος, ανεξαρτήτως του τρόπου σημασιολογικής περιγραφής που έχουν επιλέξει οι επιμέρους φορείς που προσφέρουν το πολιτισμικό υλικό. Στην επόμενη ενότητα δίνεται μια περιγραφή του κοινού μοντέλου που χρησιμοποιείται στο πλαίσιο της Europeana, στο οποίο αντιστοιχίζονται τα σχήματα μεταδεδομένων των διαφόρων φορέων με στόχο της επίτευξη διαλειτουργικότητας.

1.4. EBU Core

Το EBU Core (http://tech.ebu.ch/lang/en/MetadataEbuCore) είναι ένα σύνολο μεταδεδομένων το οποίο αποτελεί ένα βασικό σχήμα για την περιγραφή των δομικών και τεχνικών χαρακτηριστικών ραδιοφωνικού και τηλεοπτικού περιεχομένου. Στόχος του είναι να καλύψει τις πληροφορίες που αφορούν την δημιουργία, τη διαχείριση και την διατήρηση του οπτικοακουστικού υλικού. Το EBU Core μπορεί να διευκολύνει την ανταλλαγή προγραμμάτων μεταξύ διαφόρων παραγωγών οπτικοακουστικού υλικού και μπορεί επίσης να χρησιμοποιηθεί για την περιγραφή περιεχομένου που πρόκειται να διανεμηθεί μέσω τηλεοπτικής εκπομπής, διαδικτύου, κινητών συσκευών ή συνδυασμούς αυτών. Το EBU Core έχει χρησιμοποιηθεί για την περιγραφή του υλικού στα πλαίσια του προγράμματος EUScreen.

Ο βασικός πυρήνας του EBU Core είναι το Dublin Core για τα μέσα επικοινωνίας. Το Dublin Core χρησιμοποιείται ως βασικός πυρήνας περιγραφής μεταδεδομένων από διάφορες βιβλιοθήκες και μουσεία σε διάφορα προγράμματα διαχείρισης πολιτισμικής κληρονομιάς. Ωστόσο το EBU Core ενδείκνυται για την περιγραφή της πρόσβασης σε τέτοιου είδους οπτικοακουστικό περιεχόμενο. Το EBU Core ενσωματώνει τις τελευταίες εξελίξεις που αφορούν τον Σημασιολογικό Ιστό και το Linked Open Data, και είναι διαθέσιμο με τη μορφή οντολογίας RDF.

Μεταξύ των πεδίων που ορίζει το EBU Core περιλαμβάνονται για παράδειγμα τα πεδία Title (το βασικό όνομα που χαρακτηρίζει ένα αντικείμενο), Creator (ο δημιουργός του αντικειμένου, φορέας ή πρόσωπο), Description (μια σύντομη κειμενική περιγραφή του αντικειμένου), Publisher (ο διανομέας του αντικειμένου), DateCreated (η ημερομηνία παραγωγής), Format (η φυσική μορφή του αντικειμένου), Language (οι γλώσσες του ακουστικού και κειμενικού υλικού), Location (οι περιοχές που σχετίζονται με το αντικείμενο).

Είναι φανερό το EBU Core προσφέρει πλούσια δυνατότητα περιγραφής των μεταδεδομένων που αφορούν τα χαρακτηριστικά του εκάστοτε αντικειμένου, ενώ μέσω πεδίων όπως το Description επιτρέπει την συμπερίληψη πλούσιας κειμενικής πληροφορίας από την επεξεργασία της οποίας, πιθανόν να μπορούν να εξαχθούν επιπλέον πληροφορίες για το αντικείμενο (π.χ. πρόσωπα ή περιοχές που εμφανίζονται σε ένα video, κτλ).

2. Διαλειτουργικότητα με την Ευρωπαϊκή Ψηφιακή Βιβλιοθήκη (EUROPEANA)

Οι συλλογές πολιτισμικού περιεχομένου δεικτοδοτούνται κυρίως με βάση ισχυρά αλλά ετερογενή πρότυπα μεταδεδομένων. Για παράδειγμα, το Dublin Core χρησιμοποιείται για απλές ευρέσεις, το SPECTRUM για πληροφορίες πλούσιων συλλογών, το AMICO για εικόνες μουσείων τέχνης, το MARC για βιβλιογραφικές εγγραφές, το IMS για εκπαιδευτικό υλικό. Αυτή η κατάσταση παρεμποδίζει σημαντικά το συνδυασμό και το άνοιγμα προς το ευρύ κοινό τέτοιων συλλογών. Η επίτευξη σημασιολογικής διαλειτουργικότητας μπορεί να αποτελέσει τη λύση για την αντιμετώπιση αυτού του προβλήματος.

Ο μεγάλος αριθμός των σχημάτων μεταδεδομένων σχεδιάστηκε με βάση τις απαιτήσεις συγκεκριμένων ομάδων χρηστών, προσανατολισμένων χρηστών, τύπων υλικού, θεματικών περιοχών, ιδιαίτερων αναγκών συγκεκριμένων έργων κλπ. Ενώ τα σχήματα μεταδεδομένων παρέχουν κοινά λεξιλόγια για την περιγραφή μιας συγκεκριμένης περιοχής ενδιαφέροντος, ανακύπτουν πολλά προβλήματα κάθε φορά που δημιουργούνται μεγάλες ψηφιακές βιβλιοθήκες ή χώροι αποθήκευσης με μεταδεδομένα τα οποία είχαν ετοιμαστεί σύμφωνα με διαφορετικά σχήματα. Έγιναν πολλές προσπάθειες για τον ορισμό της έννοιας της διαλειτουργικότητας. Μερικά παραδείγματα είναι και τα ακόλουθα:

- "Διαλειτουργικότητα είναι η δυνατότητα πολλαπλών συστημάτων που βασίζονται σε διαφορετικές πλατφόρμες υλικού και λογισμικού, δομών δεδομένων και διαπροσωπειών να ανταλλάσουν δεδομένα με ελάχιστη απώλεια περιεχομένου και λειτουργικότητας"

- "Διαλειτουργικότητα είναι η δυνατότητα δύο ή περισσοτέρων συστημάτων ή τμημάτων να ανταλλάσουν πληροφορίες και να χρησιμοποιούν τις ανταλλασσόμενες πληροφορίες χωρίς ιδιαίτερη προσπάθεια"

- "Διαλειτουργικότητα: Η συμβατότητα δύο ή περισσοτέρων συστημάτων έτσι ώστε να μπορούν να ανταλλάσσουν πληροφορίες και δεδομένα και να μπορούν να χρησιμοποιούν τις ανταλλασσόμενες πληροφορίες και δεδομένα χωρίς ιδιαίτερους χειρισμούς".

Από τους παραπάνω ορισμούς μπορούμε να δούμε ότι η διαλειτουργικότητα σχετίζεται με το να έχουμε μια κοινή κατανόηση της σημασίας της πληροφορίας η οποία προέρχεται από ετερογενείς πηγές.

2.1 Σενάρια Διαλειτουργικότητας

Το κύριο ζήτημα χαρακτηρισμού της διαλειτουργικότητας ως σημασιολογικής έχει να κάνει με την κοινή αυτόματη ερμηνεία της σημασίας των πληροφοριών που ανταλλάσσονται, δηλαδή με τη δυνατότητα αυτόματης επεξεργασίας των πληροφοριών κατά τρόπο κατανοητό από τη μηχανή. Το πρώτο βήμα επίτευξης κάποιου αρχικού επιπέδου κοινής κατανόησης αποτελεί η γλώσσα αναπαράστασης η οποία ανταλλάσσει την τυπική σημασιολογία της πληροφορίας. Στη συνέχεια, τα συστήματα που κατανοούν αυτή τη σημασιολογία (εργαλεία συμπερασματολογίας, μηχανές οντολογικής αναζήτησης κλπ.) μπορούν να επεξεργαστούν την πληροφορία και να παρέχουν διαδικτυακές υπηρεσίες όπως είναι η αναζήτηση, η ανάκτηση κλπ. Οι τεχνολογίες του Σημασιολογικού Ιστού παρέχουν στο χρήστη ένα τυπικό πλαίσιο αναπαράστασης και επεξεργασίας διαφορετικών επιπέδων σημασιολογίας (πρότυπα W3C όπως τα RDF, OWL, SKOS, επεξεργασία οντολογιών, εργαλεία συμπερασματολογίας και απεικόνισης κλπ).

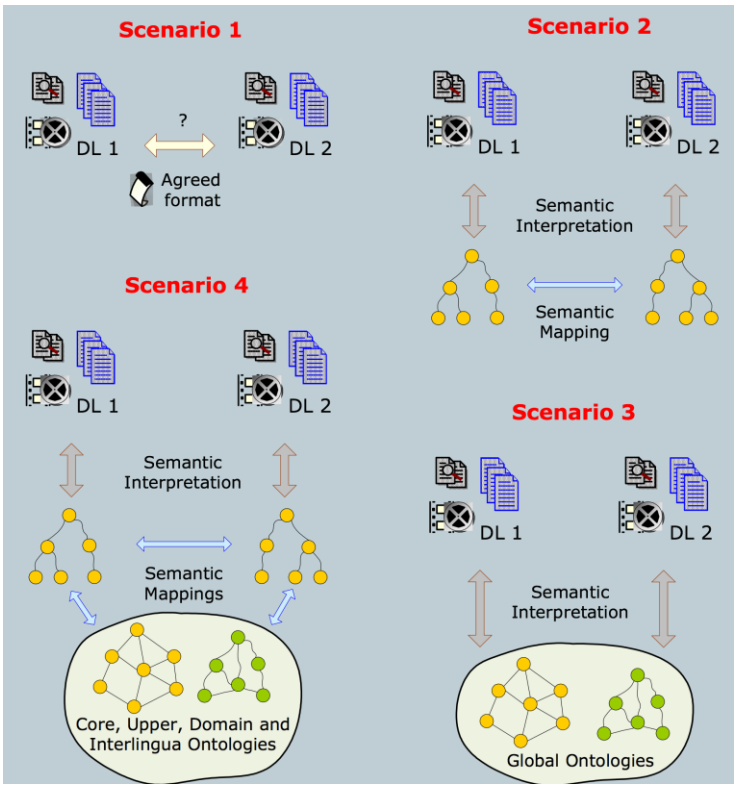
Στην Εικόνα 1 παρουσιάζονται μερικοί διαφορετικοί τρόποι επίτευξης σημασιολογικής διαλειτουργικότητας στο πλαίσιο του Σημασιολογικού Ιστού. Στο Σενάριο 1, διαφορετικές ψηφιακές βιβλιοθήκες ανταλλάσσουν πληροφορία με προκαθορισμένη και προσυμφωνημένη μορφοποίηση (η οποία πιθανώς βασίζεται σε πρότυπα). Στην περίπτωση που αυτή η πληροφορία παρέχεται σε XML ή παρόμοιες γλώσσες (βλέπε http://www.w3.org/XML/), δεν αναπαριστώνται τυπικές σημασιολογίες και κατά συνέπεια ο μόνος τρόπος να εκφραστεί η σημασία της πληροφορίας είναι μέσω αυστηρής συμφωνίας σχετικά με την μορφοποίηση ανταλλαγής. Είναι σημαντικό να προσέξουμε ότι σε αυτή την περίπτωση η σημασία δεν είναι κατανοητή από τη μηχανή, κάτι το οποίο σημαίνει ότι το Σενάριο 1 δεν υπάγεται τυπικά σε κανέναν ορισμό σημασιολογικής διαλειτουργικότητας. Εν' τούτοις, εφόσον από πλευράς λειτουργικότητας τα συστήματα μπορούν να συνεργαστούν κατ' αυτό τον τρόπο και επίσης οι γλώσσες τύπου XML παρέχουν δομική πληροφορία και μια περιορισμένη μορφή σημασίας κατανοητής από τον άνθρωπο (με βάση τα ονόματα που χρησιμοποιούνται),

μερικές φορές οι χρήστες εσφαλμένα νομίζουν ότι το Σενάριο 1 προσφέρει σημασιολογική ερμηνεία.

Το Σενάριο 2 βασίζεται στην υιοθέτηση γλωσσών αναπαράστασης τυπικής γνώσης παρέχοντας σημασιολογική ερμηνεία της πληροφορίας κατανοήσιμη από τη μηχανή, με βάση το μοναδικό τρόπο με τον οποίο οι δικτυακοί πόροι αναγνωρίζονται μέσω των Unified Recourse Identifiers (URIs, http://www.w3.org/TR/uri-clarification/). Κάθε ψηφιακή βιβλιοθήκη παρέχει μια τυπική περιγραφή των μεταδεδομένων της σε οντολογική μορφή, υιοθετώντας τις πιο κατάλληλες τυπικές γλώσσες αναπαράστασης σύμφωνα με το επίπεδο σημασιολογίας που πρέπει να εκφραστεί (λεπτομέρειες παρέχονται στην επόμενη ενότητα). Οπότε, με χρήση ενός πλαισίου απεικόνισης που διατηρεί τη σημασιολογία (αυτόματα ή μη αυτόματα), μπορεί να γίνει εκμετάλλευση της σημασιολογίας (ή τουλάχιστον κάποιων τμημάτων της) από άλλες παρόμοιες ψηφιακές βιβλιοθήκες. Παρά το γεγονός ότι η ιδέα αυτού του σεναρίου είναι απλή και ξεκάθαρη, δυστυχώς υπάρχουν πολλοί περιορισμοί στις περισσότερες περιπτώσεις, οι οποίες ανακύπτουν λόγω των ποικίλων τρόπων που χρησιμοποιούν οι ψηφιακές βιβλιοθήκες για την οργάνωση της πληροφορίας, των δυσκολιών που αντιμετωπίζουν οι χρήστες στην κατασκευή απομονωμένης οντολογίας και της ανικανότητας επίλυσης της ενδογενούς πολυπλοκότητας από εργαλεία αυτόματης απεικόνισης και συμπερασματολογίας. Μια πρώτη λύση στα παραπάνω προβλήματα υιοθετείται από το Σενάριο 3, όπου χρησιμοποιούνται καθολικές οντολογίες ως πλαίσιο αναφοράς, παρέχοντας κοινή κατανόηση της σημασιολογίας. Σε αυτή την περίπτωση οι ψηφιακές βιβλιοθήκες δεν κατασκευάζουν καμία οντολογία. Χρησιμοποιούν καθολικές οντολογίες προκειμένου να εξηγούν τα μεταδεδομένα τους. Το κύριο πρόβλημα αυτής της προσέγγισης είναι το γεγονός ότι όλες οι ψηφιακές βιβλιοθήκες θα πρέπει να χρησιμοποιούν τις ίδιες καθολικές οντολογίες, εισάγοντας κατ' αυτό τον τρόπο τα μειονεκτήματα της πρακτικής αμοιβαίας συμφωνίας (Σενάριο 1). Θα μπορούσαμε εδώ να ισχυριστούμε ότι το Σενάριο 3 αποτελεί τη "σημασιολογική έκδοση" του Σεναρίου 1, κληρονομώντας πολλά από τα μειονεκτήματά του.

Το Σενάριο 4 επιχειρεί να ικανοποιήσει με συνέπεια το όραμα της Σημασιολογικού Ιστού εισάγοντας όλες τις τεχνολογικές δυνατότητες. Διαφορετικές ψηφιακές βιβλιοθήκες χρησιμοποιούν τις δικές τους οντολογίες για να δηλώνουν τον ιδιαίτερο τρόπο οργάνωσης του αρχείου τους και να ορίζουν απεικονίσεις (αυτόματα ή μη αυτόματα) με παρόμοιες οντολογίες άλλων ψηφιακών

βιβλιοθηκών και με καθολικές οντολογίες που περιγράφουν γενική γνώση, την περιοχή εφαρμογής της αρχειακής συλλογής, γνώση σχετική με ανταλλαγή πληροφοριών κλπ. Αυτή η προσέγγιση έχει πολλά πλεονεκτήματα, ενώ το κύριο μειονέκτημά της παραμένει η ανικανότητα των σημασιολογικών τεχνολογιών να υποστηρίζουν ένα ικανοποιητικό επίπεδο αυτοματισμού, ιδιαίτερα στην περίπτωση εξαιρετικά εκφραστικών σημασιολογιών.



Εικόνα 1: Σενάρια Σημασιολογικής Διαλειτουργικότητας

2.2 Τεχνολογίες Σημασιολογικού Ιστού

Η W3C έχει ορίσει (ή σκοπεύει να ορίσει) διάφορες γλώσσες τυπικής αναπαράστασης γνώσης, στο πλαίσιο των συστάσεων όπως το Resource Description Framework (RDF, http://www.w3.org/RDF/), το Simple Knowledge Organization System (SKOS, http://www.w3.org/2004/02/skos/), την Web Ontology Language (OWL, http://www.w3.org/TR/owl-features/) κλπ. Κάθε γλώσσα προσφέρει διαφορετικές δυνατότητες αναπαράστασης με βάση το τυπικό συντακτικό της και τη σημασιολογία της. Κατά συνέπεια κάθε γλώσσα είναι κατάλληλη για διαφορετικούς τύπους ανταλλαγής πληροφορίας που βασίζεται

στην πολυπλοκότητά της. Η RDF έχει θεωρηθεί ως ένα γενικό πλαίσιο περιγραφής διαδικτυακών πόρων με βάση την τριπλέτα υποκείμενο-κατηγόρημα-αντικείμενο, όπου το υποκείμενο υποδηλώνει κάποιο πόρο, το κατηγόρημα κάποια συγκεκριμένη πλευρά (π.χ. κάποια ιδιότητα) του πόρου και το αντικείμενο μια σχέση του πόρου με κάποια προδιαγραφή (π.χ. την τιμή της ιδιότητας). Με βάση κάποιο απλό μοντέλο δεδομένων, η RDF είναι πράγματι κατάλληλη να παρέχει την απλή σημασιολογία διαδικτυακών πόρων κατά τυπικό και κατανοήσιμο από τη μηχανή τρόπο. Απ' την άλλη, το απλό συντακτικό περιορίζει την καταλληλότητά της να αναπαριστά σύνθετες δομές πληροφορίας οι οποίες απαιτούν μεγαλύτερη εκφραστικότητα τόσο από συντακτικής όσο και από σημασιολογικής πλευράς. Για παράδειγμα, η RDF δεν έχει κανένα εσωτερικό τρόπο να ξεχωρίζει διαφορετικούς τύπους πόρων (όπως είναι τα σχήματα, τα δεδομένα, οι έννοιες, οι ρόλοι, οι ιδιότητες, οι ιεραρχίες, οι ταξονομίες κλπ). Επιπρόσθετα, δεν διαθέτει τυπικό τρόπο για να εκφράζει την ισοδυναμία (equivalence), την υπαλληλία (subsumption) κλπ. οπότε παρέχει ανεπαρκείς δυνατότητες συμπερασματολογίας. Προκειμένου να επιλυθούν τα παραπάνω προβλήματα, ορίστηκε από την W3C η πιο εκφραστική Web Ontology Language (OWL), παρέχοντας μια υψηλότερου επιπέδου μηχανική ερμηνεία των διαδικτυακών πόρων. Η εκφραστικότητα της OWL βρίσκεται κοντά στις πλήρεις προδιαγραφές της λογικής πρώτης τάξης (first-order logic) και παρέχει σημαντικές δυνατότητες σε ευφυείς δικτυακούς πράκτορες οι οποίοι επιχειρούν να δημιουργήσουν υποδομή ευέλικτης διαμοίρασης και επαναχρησιμοποίησης πληροφορίας. Εν' τούτοις η εκφραστικότητα και η ερμηνευτική ισχύς της OWL δεν είναι πάντα απαραίτητες. Σε μερικές εφαρμογές ψηφιακών βιβλιοθηκών, μερικές φορές είναι αρκετό να μπορούν να αναπαριστώνται σημασιολογικά απλές ταξονομίες, θησαυροί, ταξινομήσεις κλπ. Όσον αφορά αυτές τις περιπτώσεις η απλότητα μπορεί να προηγείται της εκφραστικότητας, παρά το γεγονός ότι κάτι τέτοιο οδηγεί σε μείωση της συμπερασματολογικής ισχύος. Παρόμοιοι λόγοι παρακίνησαν την W3C να ακολουθήσει τη διαδικασία προτυποποίησης μιας οικογένειας τυπικών γλωσσών που είναι κατάλληλες για την αναπαράσταση του παραπάνω τύπου λεξιλογίων ελεγχόμενης δομής. Το SKOS επιχειρεί να συμβάλλει στη μείωση του χάσματος ανάμεσα στη χαμηλή ισχύ αναπαράστασης της RDF (ως γλώσσας οντολογιών) και στην πολυπλοκότητα της OWL.

Ο ορισμός της διαδικασίας σημασιολογικής απεικόνισης είναι προαπαιτούμενος σχεδόν για όλα τα σενάρια σημασιολογικής διαλειτουργικότητας που παρουσιάζονται στην Εικόνα 1. Οι σημασιολογικές απεικονίσεις παρέχουν ένα

κοινό στρώμα από το οποίο θα μπορούσαν να παρατηρηθούν διαφορετικές οντολογίες και διαφορετικές ψηφιακές βιβλιοθήκες να ανταλλάξουν πληροφορία με σημασιολογικά εύρωστο τρόπο. Προφανώς, μπορούν να οριστούν διαφορετικά είδη απεικονίσεων, τα οποία να υποστηρίζουν διαφορετικούς τύπους διαλειτουργικότητας. Για παράδειγμα οι τεχνικές αυτόματης ευθυγράμμισης οντολογιών (automatic ontology alignment) είναι ιδανικές για το ταίριασμα των οντολογιών που περιγράφουν τα μεταδεδομένα των ψηφιακών βιβλιοθηκών στο Σενάριο 2, ενώ στο Σενάριο 3 θα μπορούσαν να οριστούν πιο πολύπλοκες απεικονίσεις ανάμεσα στις καθολικές οντολογίες μέσω κάποιας μη αυτόματης διαδικασίας. Μια πιο τυπική προσέγγιση εξέτασης του επιπέδου και τύπου της διαδικασίας σημασιολογικής απεικόνισης βασίζεται στα ακόλουθα θέματα:

- Στο επίπεδο συμφωνίας ανάμεσα σε διαφορετικές ψηφιακές βιβλιοθήκες (υπάρχουν συμφωνημένες σχέσεις ανάμεσα σε πρότυπα; τι γίνεται με τις διαφορετικές γλώσσες; κλπ.)

- Ο χρόνος κατά τον οποίο εκτελείται η διαδικασία σημασιολογικής απεικόνισης (οι απεικονίσεις ορίζονται κατά τη διάρκεια της αλληλεπίδρασης ή off-line;)

- Ο τρόπος λειτουργίας (αυτόματος, ημι-αυτόματος, μη αυτόματος ;)

- Ο πράκτορας ή συγκεκριμένο άτομο που ορίζει τις απεικονίσεις (είναι ο διαδικτυακός πράκτορας; ο μηχανικός οντολογιών; ο δημιουργός του πράκτορα;)

- Το επίπεδο της σημασιολογικής απεικόνισης (διατηρεί τη δομή; διατηρεί τη σημασιολογία; διατηρεί τη λογική;)

- Ο τύπος των οντολογιών που απεικονίζονται (ποια είναι η γλώσσα αναπαράστασης;)

- Το επίπεδο ομοιογένειας των διαφορετικών οντολογιών προς απεικόνιση (ποιο είναι το σημασιολογικό επίπεδο των οντολογιών που χρησιμοποιούνται; αναπαριστώνται σε OWL;)

Όλα τα παραπάνω θέματα καθορίζουν τις κατάλληλες θεωρίες και τεχνολογίες που πρέπει να χρησιμοποιούνται για τον καθορισμό και την αναπαράσταση των σημασιολογικών απεικονίσεων. Από τεχνολογικής άποψης υπάρχουν πολλές

επιλογές, συνεπώς το πεδίο παραμένει αρκετά ανοικτό. Σε κάθε περίπτωση, τα πιο σημαντικά θέματα που πρέπει να διευκρινίζονται είναι η γλώσσα αναπαράστασης σημασιολογικής απεικόνισης και η μέθοδος καθορισμού σημασιολογικής απεικόνισης. Το πρώτο θέμα έχει εξεταστεί εκτενώς και έχουν προταθεί διάφορα πλαίσια αναπαράστασης, τα οποία καλύπτουν τα διαφορετικά σημασιολογικά επίπεδα και αρχιτεκτονικές της διαλειτουργικότητας (για παράδειγμα η OWL θα μπορούσε να είναι μια ισχυρή γλώσσα αναπαράστασης σε ορισμένες περιπτώσεις, ενώ η εκφραστικότητα των κανόνων θα μπορούσε να είναι πιο κατάλληλη σε άλλες περιπτώσεις). Το δεύτερο θέμα συνήθως προσεγγίζεται μέσω διαφορετικών κατευθύνσεων για τον υπολογισμό ομοιοτήτων και κατά συνέπεια για τον ορισμό απεικονίσεων (όπως είναι η συντακτική, η εξωτερική και η σημασιολογική). Όλες οι παραπάνω τεχνολογίες σχηματίζουν μια πλούσια τεχνολογική υποδομή πάνω στην οποία θα μπορούσε να στηριχθεί η σημασιολογική απεικόνιση.

2.3 Συμβατότητα με την Europeana

Μία από τις σημαντικότερες δραστηριότητες της Ευρωπαϊκής Ένωσης τα τελευταία χρόνια ήταν η δημιουργία της Ευρωπαϊκής Ψηφιακής Βιβλιοθήκης (Europeana) που στοχεύει στην εύκολη πρόσβαση και παρουσίαση της ψηφιακής πολιτιστικής κληρονομιάς (βιβλία, φιλμ, χάρτες, φωτογραφίες, μουσική, βίντεο κ.α.). Η Europeana περιλαμβάνει ήδη άνω των 33.000.000 τεκμηρίων, η δε συσσώρευση και νέων τεκμηρίων συνεχίζεται και στη νέα περίοδο 2015-2020. Η μεγαλύτερη τεχνολογική πρόκληση ήταν η διασφάλιση της συντακτικής και σημασιολογικής διαλειτουργικότητας. Ο μεγάλος αριθμός των προτύπων αναπαράστασης μεταδεδομένων καθώς και τα διάφορα είδη ψηφιοποιημένου πολιτιστικού υλικού δημιουργούν ένα άκρως ετερογενές περιβάλλον. Οι τεχνολογίες του Σημασιολογικού Ιστού έχουν αποδείξει ότι μπορούν να προσφέρουν το μαθηματικό υπόβαθρο και τα κατάλληλα πρότυπα αναπαράστασης μεταδεδομένων για τη διασφάλιση της σημασιολογικής διαλειτουργικότητας.

Τα μεταδεδομένα της εφαρμογής είναι διαθέσιμα στη Europeana με τη χρήση του προτύπου OAI-PMH (Open Archives Initiative for Metadata Harvesting) που λειτουργεί ακριβώς προς αυτή την κατεύθυνση. Δημιουργεί ένα πλαίσιο διαλειτουργικότητας το οποίο επιτρέπει τη συγκέντρωση των μεταδεδομένων των ψηφιακών συλλογών και διευκολύνει την αποτελεσματική διάχυση του ψηφιακού περιεχομένου.

Μία πολύ σημαντική λειτουργία της προτεινόμενης υπηρεσίας είναι η εύρεση και εξόρυξη μεταδεδομένων από εξωτερικές πηγές του παγκοσμίου ιστού με αυτόματα εργαλεία με σκοπό τον εμπλουτισμό των μεταδεδομένων της εφαρμογής.

2.4 Το Σχήμα Μεταδεδομένων της Europeana

Ο μεγάλος αριθμός των προτύπων αναπαράστασης μεταδεδομένων καθώς και τα διάφορα είδη ψηφιακού πολιτιστικού υλικού που συσσωρεύει η Europeana δημιουργούν ένα άκρως ετερογενές περιβάλλον. Η μεγαλύτερη τεχνολογική πρόκληση που έχει να αντιμετωπίσει η Europeana είναι η διασφάλιση της συντακτικής, σημασιολογικής και πολυγλωσσικής διαλειτουργικότητας. Για το σκοπό αυτό, δημιουργήθηκε ένα σχήμα μεταδεδομένων που θα διασφαλίζει τη διαλειτουργικότητα μεταξύ των διαφορετικών σχημάτων μεταδεδομένων που χρησιμοποιούν οι πολιτιστικοί οργανισμοί ανά την Ευρώπη. Το σχήμα αυτό είναι βασισμένο στο ευρέως διαδεδομένο και καθιερωμένο πρότυπο Dublin Core (DC). Πιο συγκεκριμένα, δημιουργήθηκε ένα application profile του DC με προσθήκες πεδίων για την προβολή της ψηφιακής αναπαράστασης των φυσικών αντικειμένων. Είναι σημαντικό να τονίσουμε ότι η Europeana συλλέγει μόνο τα μεταδεδομένα που συνοδεύονται από online ψηφιακές αναπαραστάσεις των αντικειμένων. Λόγω της υπάρχουσας Ευρωπαϊκής νομοθεσίας για τα πνευματικά δικαιώματα των ψηφιακών αντικειμένων, η προβολή τους γίνεται μόνο από τους φορείς που έχουν τα δικαιώματα. Η Europeana προβάλει μόνο τα μεταδεδομένα και υποδεικνύει τον ιστοχώρο (του εκάστοτε πολιτιστικού φορέα) όπου μπορεί ο τελικός χρήστης να δει το ψηφιακό αντικείμενο.

Η επιλογή του σχήματος μεταδεδομένων της Europeana έγινε με βάση τις ανάγκες των βιβλιοθηκών που ήταν οι πρώτες που συνεισέφεραν περιεχόμενο στην πύλη και όχι με βάση τις ανάγκες του συνόλου των Ευρωπαϊκών πολιτιστικών οργανισμών (μουσεία, αρχεία, οπτικοακουστικά αρχεία). Πολύ γρήγορα δημιουργήθηκε η ανάγκη της προσαρμογής-επέκτασης του σχήματος μεταδεδομένων για την κάλυψη όλων των αναγκών και ιδιαίτερα σε σχέση με την αναπαράσταση των μουσειακών ψηφιακών αντικειμένων. Η ανάγκη για την υιοθέτηση ενός πιο περιεκτικού προτύπου μεταδεδομένων οδήγησε την

Europeana στη δημιουργία μίας ομάδας ειδικών για τον καθορισμό ενός κατάλληλου προτύπου. Αυτό το πρότυπο θα πρέπει να καλύπτει τις ανάγκες του συνόλου των πολιτιστικών οργανισμών καθώς και την ανάγκη χρήσης αυτών των μεταδεδομένων από εφαρμογές ηλεκτρονικής μάθησης.

2.5 Διαλειτουργικότητα και Διαχείριση Περιεχομένου

Μια σημαντική παράμετρος της ανάπτυξης του Εθνικού και Ευρωπαϊκού πολιτιστικού πεδίου είναι η διαλειτουργικότητα (interoperability) του ψηφιακού περιεχομένου που δημιουργείται και διατίθεται οργανωμένα από τους φορείς – κατόχους του περιεχομένου. Η διαλειτουργικότητα είναι ένα από τα σημαντικότερα θέματα που απασχολεί τους δημιουργούς και παρόχους του ψηφιακού περιεχομένου κατά τη δημιουργία, αποθήκευση και διαχείρισή του, διότι αποτελεί βασική προϋπόθεση για την προβολή και προσβασιμότητα σε αυτό.

Η **διαλειτουργικότητα** ορίζεται ως *η ικανότητα μεταφοράς και χρησιμοποίησης της πληροφορίας με ένα ομοιογενή και αποτελεσματικό τρόπο μεταξύ διαφόρων οργανισμών σε επίπεδο συστημάτων πληροφορικής* [ΠΔΗΔ].

Ένα από τα σημαντικότερα προβλήματα για την επίτευξη διαλειτουργικότητας είναι η ποικιλία και η διαφορετικότητα των υπαρχόντων συστημάτων διαχείρισης. Η πληροφορία που υπάρχει σε πολιτιστικούς οργανισμούς έχει συλλεχθεί και περιγραφεί με διαφορετικούς τρόπους, είτε λόγω της φύσης του υλικού είτε λόγω της φύσης των δεδομένων που συλλέγονται από τις μουσειακές συλλογές. Οι βιβλιοθηκάριοι έχουν χρησιμοποιήσει ένα πρότυπο δεδομένων, οι αρχειοθέτες ένα άλλο, οι έφοροι των μουσείων ένα άλλο και πιθανώς κάθε διαφορετικό τμήμα του πολιτισμικού οργανισμού έχει υιοθετήσει άλλο πρότυπο. Για μία ποικιλία λόγων – διαθεσιμότητα του λογισμικού, διαφοροποιήσεις στα πρότυπα, ανάγκες των συλλογών – η ετερογένεια εμποδίζει την ομαλή πρόσβαση, διαχείριση και αξιοποίηση του πολιτισμικού αποθέματος.

Η επίτευξη της διαλειτουργικότητας περιλαμβάνει δύο επί μέρους στόχους: την *τεχνική – συντακτική* και τη *σημασιολογική* διαλειτουργικότητα. Η **τεχνική-συντακτική διαλειτουργικότητα** αφορά τεχνικά θέματα που άπτονται της σύνδεσης υπολογιστικών συστημάτων με στόχο την ανταλλαγή πληροφορίας ή τη χρήση λειτουργικότητας από άλλα συστήματα. Η **σημασιολογική διαλειτουργικότητα** αναφέρεται στην εξασφάλιση ότι είναι κατανοητή η ακριβής

σημασιολογία της ανταλλασσόμενης πληροφορίας από κάθε άτομο ή εφαρμογή που γίνεται δέκτης αυτής.

Υπάρχουν διάφορα θέματα και επίπεδα διαλειτουργικότητας, που έχουν σχέση με την δυνατότητα των χρηστών να έχουν 'κοινού τύπου' πρόσβαση στο περιεχόμενο.

Ένα θέμα αφορά στον ορισμό της μονάδας πρόσβασης, ιδίως στην περίπτωση περιεχομένου που δημιουργείται σε ψηφιακή μορφή (born digital). Στο αναλογικό περιεχόμενο, τέτοιο πρόβλημα δεν υφίσταται, γιατί οι βασικές μονάδες είναι, για παράδειγμα, τα βιβλία ή τα άρθρα, με αναφορά στον αριθμό της σελίδας όπου βρίσκεται η πληροφορία. Αντίστοιχα είναι στην περίπτωση αρχείων (archives), τα records και τα files, τα εκθέματα στην περίπτωση των μουσείων, οι ταινίες στην περίπτωση των οπτικοακουστικών αρχείων, ή τα δοκίμια και τα τραγούδια στην περίπτωση της μουσικής.

Η πρόσβαση σε οργανωμένο ψηφιακό περιεχόμενο, στο πλαίσιο των Ευρωπαϊκών εξελίξεων (Europeana), γίνεται στην τωρινή μορφή σε επίπεδο 'πλήρους' ψηφιακού τεκμηρίου, δηλαδή αντίστοιχου με τα ανωτέρω αναφερθέντα αναλογικά τεκμήρια. Μακροπρόθεσμα, για το 2010-2011, προβλέπεται αύξηση της διακριτότητας σε επίπεδο μοντέλου, ορίζοντας δομές -εσωτερικές- του περιεχομένου και επιτρέποντας την πρόσβαση σε αυτές.

Ένα άλλο θέμα αφορά στα μεταδεδομένα. Το πρόβλημα εδώ συνίσταται στην ανάγκη για αναζήτηση και ανάκτηση περιεχομένου ανάμεσα σε διαφορετικές ψηφιακές συλλογές, τόσο σε επίπεδο του συνόλου των συλλογών, όσο και σε επίπεδο των αντικειμένων (objects) των συλλογών. Η βασική προσέγγιση στο πρόβλημα είναι η χρησιμοποίηση κοινών – προτύπων – μορφών μεταδεδομένων, και η αντιστοίχιση των διαφορετικών μορφών μεταδεδομένων μεταξύ τους.

Η πλέον σημαντική μορφή μεταδεδομένων από τεχνικής απόψεως είναι τα 'περιγραφικά (descriptive)' μεταδεδομένα. Τα μεταδεδομένα αυτά μπορεί να είναι τεχνικής, δομικής και διαχειριστικής φύσης, και θα αντιμετωπίζονται χωριστά το ένα από το άλλο.

Στο ανωτέρω πλαίσιο στην Europeana χρησιμοποιείται ένα Dublin Core Application Profile, δηλαδή προφίλ εφαρμογών βασισμένο στο πρότυπο μεταδεδομένων Dublin Core, για κάθε ειδικό τομέα περιεχομένου και σε επίπεδο αντικειμένου, ώστε να είναι δυνατή η αναζήτηση στον τομέα αυτό. Η δημιουργία αυτή είναι βασισμένη σε υπάρχοντα πρότυπα μεταδεδομένων, όπως το MARC για

τις βιβλιοθήκες (TEL AP), το EAD για τα αρχεία (Archives AP), το SPECTRUM/CIDOC για τα μουσεία (Museum AP), τα Scholarly Publications για ηλεκτρονικές εκτυπώσεις (ePrints AP) και άλλα, βασισμένα σε ελεγχόμενα λεξιλόγια (controlled vocabularies).

Οι διάφοροι φορείς που θέλουν να δημιουργήσουν ψηφιακό περιεχόμενο το οποίο θα είναι προσβάσιμο ηλεκτρονικά, πρέπει να αντιστοιχίζουν τα τοπικά τους μεταδεδομένα σε ένα από τα ανωτέρω Dublin Core Application Profiles. Απαιτείται η δημιουργία Application Profiles σε τομείς όπου σήμερα δεν υ*πάρχουν.*

Σε επίπεδο μεταδεδομένων αναζήτησης για ένα συγκεκριμένο μεγάλο τομέα ψηφιακού περιεχομένου – όπως είναι η Ευρωπαϊκή Ψηφιακή Βιβλιοθήκη, η οποία αφορά, προφανώς, και στο αναπτυσσόμενο από ελληνικούς φορείς ψηφιακό πολιτιστικό περιεχόμενο – θα πρέπει να δημιουργηθούν βάσεις (καταγραφής) μεταδεδομένων, όπως και κατάλογοι όρων, με ορισμούς όλων των όρων των μεταδεδομένων – των ιδιοτήτων και των σχημάτων κωδικοποίησής τους. Τα μη περιγραφικά μεταδεδομένα πρέπει επίσης να αναφέρονται σε τεχνικά θέματα, όπως είναι τα file formats. Οι τεχνικές οδηγίες του Δικτύου Minerva (www.minervaeurope.org) και του έργου PLANETS (www.planets-project.eu) μπορούν να αποτελέσουν την βάση για αντιμετώπιση των θεμάτων αυτών.

Η αναζήτηση με βάση περιγραφικά μεταδεδομένα μεταξύ διαφορετικών φορέων, προτύπων και γλωσσών (σε Ευρωπαϊκό επίπεδο) θα βοηθηθεί σημαντικά από την χρησιμοποίηση των υπαρχουσών και των συνεχώς εξελισσόμενων τεχνικών σημασιολογικής διαλειτουργικότητας, που βασίζεται στις τεχνολογίες και γλώσσες αναπαράστασης του Σημασιολογικού Ιστού (Semantic Web).

Οποιαδήποτε λειτουργικότητα αναφέρεται στο ψηφιακό περιεχόμενο, απαιτεί την (κοινή) αυτόματη κατανόηση της έννοιας της ανταλλασσόμενης πληροφορίας και επομένως την δυνατότητα αναπαράστασης και επεξεργασίας της πληροφορίας με ένα τρόπο που είναι κατανοητός από μηχανές.

Το πρώτο βήμα στην επίτευξη του στόχου αυτού είναι η χρησιμοποίηση μιας γλώσσας αναπαράστασης η οποία ανταλλάσσει την σημασιολογία των αντικειμένων. Με μια τέτοια γλώσσα, συστήματα τα οποία αναλύουν το περιεχόμενο με συμπερασματολογικά εργαλεία και οντολογίες (reasoning tools, ontology querying engines), θα μπορούν να παρέχουν ηλεκτρονικές υπηρεσίες, όπως αποτελεσματική – ενοποιημένη - αναζήτηση και ανάκληση περιεχομένου. Οι τεχνολογίες του Σημασιολογικού Ιστού παρέχουν το πλαίσιο για την αναπαράσταση και επεξεργασία διαφορετικών επιπέδων σημασιολογίας. Η

σημασιολογική διαλειτουργικότητα θα αφορά στο περιεχόμενο σε εννοιολογικό επίπεδο, στο επίπεδο της διαπροσωπείας του χρήστη (user interface) και στην αυτόματη επεξεργασία του περιεχομένου. Η ανάπτυξη και χρησιμοποίηση των γλωσσών RDF, OWL, οντολογιών και πλαισίων όπως το CIDOC CRM αποτελούν τη βάση για την ανάπτυξη αυτή.

Στο πλαίσιο αυτό, υπάρχοντα μεταδεδομένα, ελεγχόμενα λεξιλόγια πρέπει να μετατραπούν σε μορφή αναγνωρίσιμη από τη μηχανή, συνεισφέροντας στη δημιουργία ενός επιπέδου, ικανού να δεχτεί σημασιολογικές μεθόδους επερωτήσεων. Η μορφή αυτή μπορεί να είναι το πρότυπο SKOS σε πολλές περιπτώσεις, ή και οι RDF και OWL σε άλλες Το κέρδος από την σημασιολογική διαλειτουργικότητα μπορεί να επιδειχτεί μέσω της μετατροπής πολλών ελεγχομένων λεξιλογίων και θησαυρών όρων στην μορφή SKOS και στη χρήση τους για εννοιολογική πρόσβαση στο περιεχόμενο.

3. Πηγές Γνώσης για Σημασιολογικό Εμπλουτισμό

3.1 DBPedia

Η DBpedia είναι μια από τις πλουσιότερες οντολογίες γενικού περιεχομένου. Στόχος της είναι η εξαγωγή δομημένου περιεχομένου από τις πληροφορίες που περιέχει η διαδικτυακή εγκυκλοπαίδεια Wikipedia. Η δομημένη αυτή πληροφορία διατίθεται στον Παγκόσμιο Ιστό, και επιτρέπει στους χρήστες να θέτουν ερωτήματα σχετικά με τις έννοιες και τις ιδιότητες των πηγών της Wikipedia, συμπεριλαμβανομένων των συνδέσεων προς άλλα σύνολα δεδομένων (που αποτελούν μέρος του Linked Data). Για την αναπαράσταση των στιγμιοτύπων η DBpedia χρησιμοποιεί το πρότυπο RDF, ενώ η δομή της οντολογίας διατίθεται στη μορφή OWL.

Το εξαιρετικό εύρος του περιεχομένου της Wikipedia, η οποία αποτελεί μια εξαιρετικά πλούσια και συνεχώς επεκτεινόμενη εγκυκλοπαίδεια, καθιστά την DBpedia μια εξαιρετική πηγή πληροφορίας, για την αποδοτική χρησιμοποίηση της οποίας πρέπει να ληφθεί υπόψη ο πολύ μεγάλος όγκος της και η σχετικά χαμηλή σημασιολογική εκφραστικότητά της (καθώς χρησιμοποιεί το πρότυπο RDF αντί της OWL). Είναι χαρακτηριστικό ότι το σύνολο των δεδομένων της DBpedia περιγράφει περισσότερα από 3.64 εκατομμύριο αντικείμενα, 1.83 εκατομμύριο από τα οποία είναι ενταγμένα σε μια καλά δομημένη οντολογία, που περιλαμβάνει 416,000 πρόσωπα, 526,000 περιοχές, 106,000 μουσικά άλμπουμ, 60,000 ταινίες, 17,500 βινετοπαιχνίδια, 169,000 οργανισμούς, 183,000 είδη ζωής και 5,400 ασθένειες. Το σύνολο δεδομένων της DBpedia περιέχει περιγραφές μέχρι και σε 97 διαφορετικές γλώσσες, 2,724,000 συνδέσμους προς εικόνες 6,300,000 συνδέσμους προς εξωτερικές ιστοσελίδες, 6,200,000 συνδέσμους προς άλλα σύνολα δεδομένων τύπου RDF.

3.2 Wordnet

Για την καλύτερη επεξεργασία των κειμενικών περιγραφών είναι συνήθως χρήσιμη η χρήση ενός είδους θησαυρού που επιτρέπει την εξαγωγή εννοιών από τα κείμενα. Το WordNet είναι μια τέτοια μεγάλη λεξικογραφική βάση δεδομένων της αγγλικής γλώσσας.

Τα ουσιαστικά, τα ρήματα, τα επίθετα και τα επιρρήματα της γλώσσας είναι ομαδοποιημένα σε σύνολα συνωνύμων, καθένα από τα οποία εκφράζει μια διαφορετική έννοια. Τα σύνολα αυτά διασυνδέονται μεταξύ τους μέσω εννοιακών, σημασιολογικών και λεξιλογικών σχέσεων. Το WordNet μοιάζει με θησαυρό καθώς ομαδοποιεί τις λέξεις με βάση την σημασία τους. Ωστόσο, διαφέρει σε αρκετά σημεία από έναν παραδοσιακό θησαυρό. Πρώτον το WordNet δεν διασυνδέει μόνο απλά λέξεις, αλλά σημασίες λέξεων. Αυτό βοηθάει διευκρίνιση των διαφορών μεταξύ λέξεων που βρίσκονται κοντά η μία στην άλλη. Δεύτερον, το WordNet ονοματίζει τις σημασιολογικές σχέσεις μεταξύ των διαφόρων λέξεων, ενώ η ομαδοποίηση των λέξεων σε έναν θησαυρό δεν ακολουθεί κάποιο συγκεκριμένο πρότυπο πέραν της σημασιολογικής ομοιότητας.

Η βασική σχέση μεταξύ των λέξεων στο WordNet είναι η συνωνυμία, όπως π.χ. μεταξύ των λέξεων car και automobile. Τα συνώνυμα, δηλαδή οι λέξεις που σημαίνουν το ίδιο πράγμα και μπορούν να χρησιμοποιηθούν η μία στη θέση της άλλης ομαδοποιούνται σε μη διατεταγμένα σύνολα που καλούνται synset. Καθένα από τα 117.000 synset του WordNet συνδέεται με τα άλλα synset μέσω ενός μικρού πλήθους «εννοιακών σχέσεων». Οι λέξεις που έχουν πολλές διαφορετικές μεταξύ τους σημασίες περιέχονται σε περισσότερα του ενός synset. Η πιο συνήθης σχέση μεταξύ των synset είναι η σχέση υπαγωγής (ή σχέση ISA). Η σχέση αυτή συνδέει τα πιο γενικά synset όπως furniture με πιο εξειδικευμένα synset όπως bed. Όλες οι ιεραρχίες ουσιαστικών καταλήγουν σε έναν ριζικό κόμβο και όλες οι σχέσεις υπαγωγής είναι μεταβατικές. Πέραν της υπαγωγής μοντελοποιείται και η σχέση της μερονυμίας, δηλαδή της σχέσης μέρους προς όλο, όπως μεταξύ των synset chair και back. Τα ρήματα ταξινομούνται επίσης σε ιεραρχίες από synset και τα ρήματα που βρίσκονται στα κατώτερα επίπεδα της ιεραρχίας εκφράζουν ολοένα και πιο εξειδικευμένες πράξεις, όπως π.χ. στην ακολουθία communicate-talk-whisper. Η οργάνωση των επιθέτων γίνεται με βάση τα αντώνυμα. Τα ζεύγη άμεσων αντωνύμων όπως young-old εκφράζουν μια ισχυρή σημασιολογική σχέση. Πέραν αυτών ορίζονται όμως και ασθενέστερες σχέσεις.

4. Εργαλεία αντιστοίχισης

Σκοπός του παρόντος κεφαλαίου είναι η παρουσίαση των διαθεσίμων συστημάτων και εργαλείων που εξυπηρετούν την ευθυγράμμιση των οντολογιών και την αντιστοίχιση των σχημάτων μεταδεδομένων.

4.1 AgreementMakerLight

Το AgreementMakerLight (AML) [1] είναι ένα σύστημα ευθυγράμμισης οντολογιών, αυτοματοποιημένο και επεκτάσιμο που αναπτύχθηκε κυρίως για τον κλάδο των βιοεπιστημών όπου χρησιμοποιούνται οντολογίες πολύ μεγάλου μεγέθους.

Το σύστημα διαθέτει μία διεπαφή για το χρήστη η οποία προκειμένου να είναι εύχρηστη και επεκτάσιμη απεικονίζει μόνο τη γειτονιά μιας ευθυγράμμισης την κάθε στιγμή ενώ παρέχει διάφορες επιλογές πλοήγησης στην ευθυγράμμιση.

Το AML Περιέχει τρεις βασικές μονάδες, τη μονάδα φόρτωσης της οντολογίας, τη μονάδα αντιστοίχισης και τη μονάδα επιλογής των αντιστοιχίσεων και επιδιόρθωσης. Η μονάδα φόρτωσης είναι υπεύθυνη για την ανάγνωση των οντολογιών και τη σάρωση της πληροφορίας τους με σκοπό την οργάνωσή της στις δομές δεδομένων της οντολογίας του συστήματος. Η πιο σημαντική δομή για υλοποίηση της αντιστοίχισης είναι ο πίνακας Lexicon που περιλαμβάνει τα ονόματα των τάξεων και τα συνώνυμα μιας οντολογίας και χρησιμοποιεί ένα σύστημα κατάταξης για να αποδίδει βάρη στις αντιστοιχίσεις. Η μονάδα ταιριάσματος περιέχει αλγορίθμους ταιριάσματος του AML ή προσαρμογείς. Οι προσαρμογείς διακρίνονται σε πρωτεύοντες και δευτερεύοντες. Οι πρωτογενείς είναι γραμμικής πολυπλοκότητας και εφαρμόζονται συνολικά σε όλα τα προβλήματα και οι δεύτεροι είναι πολυωνυμικής πολυπλοκότητας και μπορούν να εφαρμοστούν μόνο τοπικά σε μεγάλα σε προβλήματα. Η χρήση της διαθέσιμης γνώσης στους πρωτεύοντες προσαρμογείς είναι ένα βασικό χαρακτηριστικό του AML το οποίο περιλαμβάνει ένα καινοτόμο αλγόριθμο επιλογής της γνώσης υποβάθρου. Η μονάδα επιλογής και επιδιόρθωσης των αντιστοιχίσεων εξασφαλίζει ότι η τελική ευθυγράμμιση έχει την επιθυμητή πληθυκότητα και ότι είναι συνεπής. Ο προσεγγιστικός αλγόριθμος επιδιόρθωσης της ευθυγράμμισης του AML εφαρμόζει ένα βήμα βελτιστοποίησης που εντοπίζει το ελάχιστο σύνολο των κλάσεων που απαιτείται να εξεταστούν ως προς τη συνέπεια μειώνοντας έτσι το μέγεθος του προβλήματος επιδιόρθωσης.

Το AML σχεδιάστηκε με βασικό στόχο τη δυνατότητα επέκτασης του συστήματος, ενώ διατηρεί χαμηλή πολυπλοκότητα. Το σύστημα εκτελεί τη διαδικασία ευθυγράμμισης σε ένα προσωπικό υπολογιστή σε λιγότερο από ένα λεπτό για μεσαίου μεγέθους προβλήματα (μέχρι 10.000 τάξεις ανά οντολογία) έως το πολύ 20 λεπτά για πολύ μεγάλα προβλήματα (έως και 100.000 τάξεις ανά οντολογία). Το AML έχει πετύχει αξιόλογες επιδόσεις σε περιπτώσεις οντολογιών από το πεδίο των βιοεπιστημών στην Πρωτοβουλία Αξιολόγησης της Ευθυγράμμισης των Οντολογιών (Ontology Alignment Evaluation Initiative, OAEI 2013), ενώ χρησιμοποιείται σε πολλές άλλες εφαρμογές.

Πρόκειται για ένα σύστημα ανοιχτού κώδικα που διατίθεται σε μορφή εκτελέσιμου (.jar) και σα διεπαφή του Eclipse (https://github.com/AgreementMakerLight).

4.2 AOT / AOTL

Τα AOT και AOTL [2] είναι δύο συστήματα ευθυγράμμισης οντολογιών. Το σύστημα AOT χρησιμοποιεί διάφορετικούς αλγόριθμους ευθυγράμμισης που βασίζονται στην ορολογία προκειμένου να υπολογίσει τις ομοιότητες και χρησιμοποιεί ένα τοπικό φίλτρο προκειμένου να επιλεγούν οι καλύτερες αντιστοιχίσεις. Το AOTL συνδυάζει τις ομοιότητες που υπολογίζονται από τους διάφορους αλγόριθμους ταιριάσματος συμβολοσειράς (σε επίπεδο ορολογίας) χωρίς τοπικό φίλτρο. Στη συνέχεια οι ομοιότητες που έχουν προκύψει συνδυάζονται με όσες προκύπτουν σε γλωσσικό επίπεδο και υπολογίζονται με χρήση του εξωτερικού λεξικού WordNet. Έπειτα, το σύστημα συνδυάζει τις διάφορες ομοιότητες δίνοντας προτεραιότητα στη γλωσσολογική αντιστοίχιση.

Στόχος του συστήματος AOT είναι να μελετηθούν πάνω του διάφορα φίλτρα και συναρτήσεις άθροισης προκειμένου να βελτιωθούν μελλοντικά τα συστήματα ευθυγράμμισης. Στο AOTL χρησιμοποιείται το WordNet για την επιλογή των σημασιολογικών αντιστοιχίσεων με απώτερο στόχο το σύστημα να ανακαλύπτει νέες σημασιολογικές αντιστοιχίες παρά να υπολογίζει την βέλτιστη.

4.3 BioMixer

Το εργαλείο BioMixer [3] είναι μια διαδικτυακή εφαρμογή για την οπτικοποίηση οντολογιών. Υποστηρίζει την συνεργατική απεικόνιση οντολογιών που αφορά στην κοινή χρήση των οπτικών αναπαραστάσεων των δεδομένων από περισσότερους από έναν χρήστη, με στόχο την από κοινού επεξεργασία τους.

Η συνεργατική ανάπτυξη οντολογιών υποστηρίζεται στους διάφορους κειμενογράφους οντολογιών όπως WebProtégé2 (http://webprotege.stanford.edu) οι οποίοι δίνουν τη δυνατότητα στους χρήστες να επεξεργάζονται, να συζητούν ή να εμπλουτίζουν- σημειώνουν τις οντολογίες μέσω διαδικτύου. Η συνεργατική οπτικοποίηση προσφέρει πλεονεκτήματα όπως η δυνατότητα πρόσβασης στις οντολογίες από περισσότερους χρήστες και όχι μόνο από τους ειδικούς , ενώ ταυτόχρονα εκμεταλλευέται τη δυναμική του διαδικτύου για να πυροδοτήσει τη συζήτηση και τη συμμετοχή μεγάλων ομάδων χρηστών. Η τρέχουσα έκδοση του BioMixer μπορεί να χαρακτηριστεί ως ένα εργαλείο ασύγχρονης συνεργασίας για οπτικοποίηση των οντολογιών, δηλαδή συνεργασίας που λαμβάνει χώρα σε διαφορετικές τοποθεσίες και σε διαφορετικές χρονικές στιγμές (π.χ. με χρήση του WebProtégé). Ωστόσο, προγραμματίζεται η υποστήριξη της σύγχρονης συνεργασίας σε μελλοντικές εκδόσεις.

Η διαδικτυακή διεπαφή του BioMixer επιτρέπει την πρόσβαση στο σύστημα μέσω του περιηγητή διαδικτύου χωρίς να απαιτείται η εγκατάσταση επιπλέον λογισμικού. Μέσω της διεπαφής οι χρήστες έχουν τη δυνατότητα να χρησιμοποιούν κοινό περιβάλλον εργασίας οπτικοποίησης και να δημοσιεύουν τις απεικονίσεις εισάγοντας τις σε εξωτερικές ιστοσελίδες. Μπορούν ακόμα να αλληλεπιδρούν μεταξύ τους στέλνοντας ένα υπάρχον περιβάλλον εργασίας οπτικοποίησης στους υπόλοιπους συνεργάτες μέσω ηλεκτρονικού ταχυδρομείου και προσθέτοντας σημειώσεις στις οπτικοποιήσεις προκειμένου να εκκινήσουν συζητήσεις. Μέσω της δεπαφής του το σύστημα υποστηρίζει χρήστες με διαφορετικά γνωστικά υπόβαθρα και προτιμήσεις παρέχοντας τη δυνατότητα πολλαπλών όψεων που πραγματοποιούν διαφορετικές λειτουργίες.

Μέσω του εργαλείου ο χρήστης έχει τη δυνατότητα να αναζητήσει μία έννοια μεταξύ των διαθέσιμων οντολογιών, να προβάλλει όποιες από αυτές θέλει σε μορφή γράφου, να προσθέσει σημειώσεις, να επιλέξει μια έννοια-κόμβο και να προβάλει τις σχετικές έννοιες και τις αντιστοιχίσεις, να οπτικοποιήσει διαφορετικά είδη σχέσεων. Στο μέλλον πρόκειται να συμπεριληφθεί και η οπτικοποίηση

διαφορετικών ειδών ευθυγράμμισης (ακριβής, σχετική κ.α.). Οι οπτικοποιήσεις μπορούν να γίνουν σε διαφορετικές διατάξεις (όπως δέντρου, πλέγματος κ.α.) οι οποίες είναι κατάλληλες για διαφορετικές εργασίες και επιτρέπουν την επεκτασιμότητα όταν αυξάνονται οι κόμβοι.

Το BioMixer είναι ένα σύστημα ανοικτού κώδικα και διατίθεται στο http://github.com/thechiselgroup/biomixer.

4.4 OPTIMA

Το Optima [4] είναι ένα εργαλείο ευθυγράμμισης οντολογιών το οποίο αναγνωρίζει αυτόματα και ευθυγραμμίζει συσχετιζόμενες έννοιες και ρόλους μεταξύ οντολογιών. Διαθέτει μια διεπαφή χρήστη που διευκολύνει την οπτικοποίηση και την ανάλυση των οντολογιών σε N3, RDF και OWL καθώς και των αποτελεσμάτων της ευθυγράμμισης.

Το Optima βασίζεται κυρίως στα σχήματα των οντολογιών για να παράγει τις αντιστοιχίσεις και χρησιμοποιεί τα στιγμιότυπα όταν υπάρχουν, προκειμένου να βελτιώσει τα αποτελέσματα. Βρίσκει αντιστοιχίσεις τύπου "πολλά προς ένα" επιτρέποντας έτσι, σε πολλές έννοιες μιας οντολογίας να αντιστοιχιστούν σε μία έννοια της άλλης. Για τον υπολογισμό της ευθυγράμμισης χρησιμοποιεί έναν αλγόριθμο που υπολογίζει τη δομική και την λεξιλογική ομοιότητα μεταξύ των σχημάτων προκειμένου να υπολογίσει την πιθανότητα μιας αντιστοίχισης.

Πρόκειται μια εφαρμογή βασισμένη σε Java η οποία μπορεί να εκτελεστεί τοπικά είτε σε Windows ή Linux πλατφόρμες, είτε απευθείας από το διαδίκτυο με χρήση του Java Web Start. Χρησιμοποιεί τη βιβλιοθήκη Jena για τη σάρωση των οντολογιών που είναι γραμμένες σε N3, RDF και OWL. Η διεπαφή της βασίζεται στο Welkin, ένα πρόγραμμα περιήγησης οντολογιών ανοικτού κώδικα που επιτρέπει στο χρήστη να φορτώνει και να βλέπει τις οντολογίες. Για εύκολη περιήγηση διαθέτει πολλές διαφορετικές διατάξεις των γραφημάτων των οντολογιών (όπως δέντρου, κύκλου, τυχαίων σημείων) ενώ φιλτράρει τις μη σχετικές έννοιες για να την απλούστευση της προβολής των αποτελεσμάτων. Ανάλογα με το μέγεθος και την πολυπλοκότητα των οντολογιών η διαδικασία της ευθυγράμμισης, η πρόοδος της οποίας είναι εμφανής στο χρήστη, μπορεί να κυμαίνεται από μερικά δευτερόλεπτα έως ώρες. Τα αποτελέσματα της ευθυγράμμισης αποθηκεύονται σε μορφή XML. Τέλος, σημειώνουμε ότι δίνεται η

δυνατότητα στους χρήστες να εισάγουν χειροκίνητα κάποιες αρχικές αντιστοιχίσεις των κόμβων για τη διευκόλυνση της διαδικασίας.

Οδηγίες χρήσης και demo για το εργαλείο διατίθενται στην σελίδα http://lsdis.cs.uga.edu/projects/sensormap/.

4.5 LogMap

Το LogMap [5] είναι ένα επεκτάσιμο σύστημα ευθυγράμμισης οντολογιών με ενσωματωμένες δυνατότητες συλλογιστικής και διάγνωσης. 'Εχει τη δυνατότητα να διαχειριστεί σημασιολογικά πλούσιες οντολογίες που περιλαμβάνουν δεκάδες ή εκατοντάδες χιλιάδες κλάσεις. Από την πειραματική αξιολόγηση προκύπτει ότι μπορεί να κατασκευάσει την ευθυγράμιση για τις μεγαλύτερες βιοιατρικές οντολογίες όπως οι NCI (http://ncit.nci.nih.gov), FMA (http://sig.biostr.washington.edu/projects/fm/) και SNOMED CT (http://www.ihtsdo.org/snomed-ct). Λαμβάνει υπόψη τη σημασιολογία των OWL οντολογιών και περιλαμβάνει αλγορίθμους για την ανίχνευση και επιδιόρθωση της μη ικανοποιησιμότητας των κλάσεων και της ασυνέπειας, κατά τη διάρκεια εκτέλεσης του προγράμματος. Σαν αποτέλεσμα, παράγει ένα σύνολο αντιστοιχίσεων το οποίο σε συνδυασμό με τις αρχικές οντολογίες είναι συνεπές και δεν περιέχει μη ικανοποιήσιμες κλάσεις.

Το LogMap δημιουργεί ευρετήρια με τις ετικέτες των κλάσεων για κάθε οντολογία εισόδου. Συγκεκριμένα, αποδίδει ετικέτες στις κλάσεις της οντολογίας καθώς και για τις λεξιλογικές διαφορές τους και επιτρέπει τη δυνατότητα εμπλουτισμού των ευρετηρίων με χρήση ενός εξωτερικού λεξικού (π.χ., WordNet ή UMLS- lexicon). Το σύστημα υπολογίζει την επεκταμένη ιεραρχία των οντολογιών εισόδου με χρήση είτε απλών ευριστικών συναρτήσεων είτε με συλλογιστή περιγραφικών λογικών. Στη συνέχεια, κατασκευάζει ένα σύνολο αντιστοιχίσεων λαμβάνοντας υπόψη τις ετικέτες της κάθε οντολογίας εισόδου με βάση το οποίο θα υπολογιστούν νέες αντιστοιχίσεις. Ο πυρήνας του LogMap είναι μια επαναληπτική διαδικασία επιδιόρθωσης των μη ικανοποίησιμων κλάσεων που προκύπτουν από την ευθυγράμμιση των οντολογιών, και εύρεσης νέων αντιστοιχίσεων. Οι νέες αντιστοιχίσεις παράγονται εξερευνώντας επαναληπτικά τις οντολογίες εισόδου ξεκινώντας από τις αναφορές και χρησιμοποιώντας την εκτεταμένη ιεραρχία κλάσεων των οντολογιών. Για να ανακαλύψει νέες χαρτογραφήσεις το LogMap διατηρεί δύο σύνολα σημασιολογικά σχετικών κλάσεων για κάθε αρχική αντιστοίχιση. Τα σύνολα επεκτείνονται παράλληλα χρησιμοποιώντας την ιεραρχία

των κλάσεων. Οι νέες αντιστοιχίσεις υπολογίζονται αντιστοιχώντας τις κλάσεις των σχετικών συνόλων χρησιμοποιώντας μια συνάρτηση ομοιότητας για κάθε ζεύγος γραμματοσειρών. Η επαναληπτική διαδικασία της συνεχίζεται μέχρι να μην είναι δυνατή η επέκταση του συνόλου των αντιστοιχίσεων χωρίς τη δημιουργία ασυνέπειας στην οντολογία. Τέλος, πραγματοποιείται μια αρκετά ακριβής εκτίμηση του μέρους επικάλυψης των οντολογιών, που μπορεί να χρησιμοποιηθεί μετέπειτα από τους χρήστες για χειροκίνητη ευθυγράμμιση των οντολογιών καθώς οι περισσότερες πιθανές αντιστοιχίσεις που δεν έχουν εξαχθεί αυτόματα είναι πιθανό να ανήκουν σε αυτό.

Το LogMap δέχεται τις ίδιες μορφές οντολογιών όπως και το OWL API, δηλαδή RDF/XML, OWL/XML, OWL Functional, OBO, KRSS, and Turtle (N3). Είναι ένα εργαλείο ανοικτού κώδικα και διατίθεται υπό την άδεια GNU Lesser General Public License 3.0. Είναι διαθέσιμο σαν project του Google Code (https://code.google.com/p/logmap-matcher/) ή σαν πακέτο της πλατφόρμας SEALS (https://code.google.com/p/logmap-matcher/downloads/list). Μπορεί να χρησιμοποιηθεί από τη γραμμή εντολών μέσω της αυτόνομης διανομής του ή απευθείας από την διαδικτυακή διεπαφή του (http://csu6325.cs.ox.ac.uk/). Επίσης μπορεί εύκολα να ενσωματωθεί σε άλλες εφαρμογές java.

4.6 RiMOM-IM

Πρόκειται για ένα σύστημα [6] που έχει αναπτυχθεί για την ευθυγράμμιση βάσεων γνώσης μεγάλης κλίμακας. Η βασική ιδέα είναι η πλήρης αξιοποίηση της υπάρχουσας διαθέσιμης πληροφορίας ευθυγράμμισης για τη βελτίωση της αποδοτικότητας και την αποφυγή διάδοσης σφαλμάτων. Στο πλαίσιο αυτό αξιοποιούνται διαθέσιμες μέθοδοι ευθυγράμμισης στιγμιοτύπων ενώ εφαρμόζεται μια μέθοδος συνάθροισης που βελτιώνει την ευθυγράμμιση στιγμιοτύπων βάσει της ομοιότητας.

Το σύστημα χρησιμοποιεί δύο βασικές τεχνικές για την ενίσχυση της διαδικασίας ευθυγράμμισης σε βάσεις γνώσης μεγάλης κλίμακας, το μπλοκάρισμα και την επαναληπτική ευθυγράμμιση. Στην τεχνική μπλοκαρίσματος δημιουργείται ένα ευρετήριο για τα στιγμιότυπα στις δύο βάσεις γνώσεις ξεχωριστά και στη συνέχεια επιλέγονται τα στιγμιότυπα με τα ίδια κλειδιά ως υποψήφια ζεύγη. Στην επαναληπτική ευθυγράμμιση οι αντιστοιχίσεις εντοπίζονται σε διαφορετικές επαναλήψεις, στην καθε επανάληψη καθορίζεται ένας μικρός αριθμός αντιστοιχίσεων ο οποίος χρησιμοποιείται για τον καθορισμό των επόμενων.

Η τελευταία έκδοση του RiMOM-IM είναι διαθέσιμη στη σελίδα http://keg.cs.tsinghua.edu.cn/project/RiMOM/.

4.7 The RSDL workbench

Το RSDL workbench [7] είναι μια συλλογή εργαλείων που αναπτύχθηκε ως μέρος μιας πλατφόρμας σύνθεσης υπηρεσιών για τις αγορές υπηρεσιών και παρέχει εργαλεία για τον καθορισμό της δομής και λειτουργικότητας υπηρεσιών που βασίζονται στη γλώσσα Rich Service Description Language (RSDL). Επιτρέπει την αντιστοίχιση των αιτήσεων παροχής υπηρεσιών και των προσφορών με βάση τα μοντέλα δεδομένων τους, τις λειτουργίες τους, και πρωτόκολλα. Πραγματοποιεί την αντιστοίχιση των μοντέλων δεδομένων αξιοποιώντας διαθέσιμες οντολογίες προκειμένου να εντοπιστούν τα συσχετιζόμενα αντικείμενα των μοντέλων δεδομένων.

Το σύστημα RSDLWB σχεδιάστηκε για να αντιστοιχίζει UML μοντέλα χρησιμοποιώντας τεχνικές αντιστοίχισης βασισμένες σε κανονικοποίηση, τεκμηρίωση, ομοιότητες τύπων, σε γλωσσικές πηγές και οντολογίες. Επιπλέον, το RSDLWB αντιστοιχίζει σχέσεις δεδομένων καθώς και σχέσεις αντικειμένων. Επειδή το σύστημα σχεδιάστηκε για να αντιστοιχίζει UML μοντέλα αρχικά δεν υποστήριζε την διαχείριση OWL οντολογιών. Για το λόγο αυτό έχει εισαχθεί ένα στρώμα αφαίρεσης προκειμένου να καταστήσει δυνατή την ευθυγράμμιση οντολογιών σε OWL. Τα χαρακτηριστικά της γλώσσας OWL που αναγνωρίζει το σύστημα είναι οι ετικέτες των κλάσεων, οι σχέσεις δεδομένων καθώς και οι σχέσεις αντικειμένων. Το σύστημα ευθυγράμμισης κατασκευάζει αντιστοιχίσεις τύπου "μία προς μία". Η σύγχρονη έκδοση του συστήματος βασίζεται σε μεγάλο βαθμό στις ετικέτες και δεν υποστηρίζει οντολογίες μεγάλης κλίμακας.

Το σύστημα έχει υλοποιηθεί σε μορφή plug-in του Eclipse και υλοποιεί τη διεπαφή EMF Compare (http://www.eclipse.org/emf/compare/).

4.8 XMap++

Το XMap++ [8] είναι ένα εργαλείο ευθυγράμμισης οντολογιών σε OWL. Χρησιμοποιεί μια σειρά διαφορετικών μέτρων ομοιότητας που εμπίπτουν σε γλωσσολογικές, δομικές και άλλες κατηγορίες για να αναγνωρίσει την σημασιολογία που αναπαριστούν οι οντολογίες. Αυτά τα μέτρα ομοιότητας που οδηγούν στην ακριβέστερη ευθυγράμμιση αναπαριστάνται με διανύσματα με βάρη

από τα οποία προκύπτει ένα συνολικό μέτρο ομοιότητας. Το εργαλείο αποτελεί επέκταση δύο προηγουμένων εκδόσεων, των XMapGen και XMapSig, τα οποία ξεπερνά στον χρόνο επεξεργασίας μεγάλων οντολογιών καθώς και στη δυνατότητα αντιμετώπισης διαφορετικών φυσικών γλωσσών.

Το σύστημα αποτελείται από τρία βασικά υποσυστήματα που πραγματοποιούν αντιστοίχιση συμβολοσειρών με βάση γλωσσολογικούς κανόνες που εφαρμόζονται στις έννοιες της οντολογίας, δομική αντιστοίχιση με βάση τις σχέσεις των γειτονικών κόμβων της, ενώ γίνεται και γλωσσική αντιστοίχιση με βάση το WordNet η οποία ενισχύει τις άλλες δύο και ξεκαθαρίζει αμφισημείες. Επίσης, πραγματοποιείται μετάφραση των όρων με χρήση της διαδικτυακής υπηρεσίας Bing Translator. Όλες οι οντότητες μιας οντολογίας ελέγχονται σε ομοιότητα με κάθε μια από τις οντότητες της δεύτερης οντολογίας ενώ παράγονται πίνακες ομοιότητας για κάθε ζεύγος οντοτήτων. Οι πίνακες που προκύπτουν από τα υποσυστήματα ενοποιούνται σε ένα μοναδικό πίνακα μέσω διαδικασιών συνάθροισης, επιλογής και συνδυασμού. Τελικά με βάση ένα προκαθορισμένο κατώφλι τα αποτελέσματα για τις τιμές ομοιότητας φιλτράρονται και επιλέγονται "μία-προς-μία" αντιστοιχίσεις.

Το σύστημα XMap++ παρουσιάζει πολύ καλή ακρίβεια στα αποτελέσματα, ενώ η ευστοχία του είναι φθίνουσα ανάλογα με το μέγεθος της οντολογίας. Αυτό αποδίδεται στους περιορισμούς του WordNet για εξειδικευμένους όρους. Σε οντολογίες όπου απαιτείται μετάφραση πολλών όρων, ο χρόνος επεξεργασίας αυξάνεται λόγω της καθυστέρησης από την μετάφραση μέσω διαδικτύου. Σε ελέγχους που συνδυάζουν ευθυγράμμιση και απάντηση ερωτημάτων, η εφαρμογή ήταν ανάμεσα σε λίγες που κατάφεραν να δώσουν απάντηση σε όλα τα ερωτήματα (διαγωνισμός OAEI 2014).

Το Xmap++ που αναπτύσσεται από την ομάδα LabGED του πανεπιστημίου Badji Mokhtar της Αλγερίας και είναι διαθέσιμο μέσω της σελίδας http://www.labged.net/index.php?rubrique=mapage38.

4.9 YAM++

Το YAM++ [9] είναι ένα επεκτάσιμο σύστημα ευθυγράμμισης οντολογιών. Η τελευταία έκδοσή του έχει διάφορες αναβαθμίσεις για ταχύτερη επεξεργασία, μειώνοντας την πολυπλοκότητα των αλγορίθμων που χρησιμοποιούσε η προηγούμενη έκδοση και υιοθετώντας μέθοδο αποθήκευσης δεδομένων στο δίσκο κατά την λειτουργία. Το εργαλείο εξάγει δεδομένα για την δομή, το περιεχόμενο,

τη σημασιολογία, την ορολογία και τις επεκτάσεις των οντοτήτων μιας οντολογίας με την κάθε κατηγορία να ελέγχεται από ένα αντίστοιχο υποσύστημα το οποίο ανακαλύπτει τον μέγιστο αριθμό πιθανών αντιστοιχίσεων. Για την καλύτερη ευθυγράμμιση έχουν εισαχθεί διαφορετικά σχήματα ευθυγράμμισης ώστε να εντοπίζονται περισσότερες ασυνέπειες. Η τελική βέλτιστη ευθυγράμμιση γίνεται με έναν προσεγγιστικό αλγόριθμο βελτιωμένης ταχύτητας.

Η εφαρμογή επεξεργάζεται τις οντολογίες στην είσοδο με την ανοιχτή βιβλιοθήκη OWLAPI (http://owlapi.sourceforge.net) και τους συλλογιστές ELK^2 (https://www.cs.ox.ac.uk/isg/tools/ELK/) και Pellet (https://github.com/complexible/pellet). Στη συνέχεια κωδικοποιεί σε ευρετήρια πληροφορίες για τις ανασημάνσεις, τη δομή και το περιεχόμενο των οντοτήτων, μεταφράζοντας λέξεις με χρήση της πλατφόρμας Bing Translator (https://www.bing.com/translator/). Έπειτα οι υποψήφιες αντιστοιχίσεις ταξινομούνται στην μηχανή αναζήτησης Lucene (https://lucene.apache.org), πραγματοποιούνται αυτόματες αναζητήσεις σε κάθε οντότητα και επιλέγονται μόνο τα πρώτα αποτελέσματα. Με τον τρόπο αυτό επιτυγχάνεται η μείωση του όγκου επεξεργασίας στα επόμενα βήματα. Έπειτα, πραγματοποιείται μια γρήγορη αναζήτηση για οντότητες με σχεδόν όμοιες περιγραφές που διαφέρουν σε πολύ λίγες λέξεις. Οι οντότητες που απομένουν συγκρίνονται μεταξύ τους για ομοιότητες με διάφορα υποσυστήματα που βασίζονται στην ορολογία, τα στιγμιότυπα και το περιεχόμενο. Μετά από αυτές τις συγκρίσεις πραγματοποιούνται επιπρόσθετες με βάση ομοιότητες στη δομή της οντολογίας και τη σημασιολογία (με ελέγχους για αντιφάσεις). Οι συγκεκριμένες μέθοδοι είναι αποτελεσματικές στην περίπτωση μικρών οντολογιών με λιγότερες από 1000 οντότητες.

Σημειώνουμε ότι στο διαγωνισμό ευθυγράμμισης οντολογιών OAEI 2013 το YAM++ είχε πρώτη θέση σε αρκετές κατηγορίες, και έδειξε καλύτερα αποτελέσματα από την προηγούμενη έκδοσή του. Είναι διαθέσιμο στην ιστοσελίδα http://www.lirmm.fr/yam-plus-plus/.

4.10 AgreementMaker

Το AgreementMaker [10] είναι ένα εργαλείο ευθυγράμμισης που αναπτύχθηκε για χρήση από εξειδικευμένους χρήστες στο εκάστοτε πεδίο αντικειμένου του σχήματος της βάσης ή της οντολογίας. Η αρχιτεκτονική του περιλαμβάνει πολλά διαφορετικά υποσυστήματα που χαρακτηρίζονται από παραμέτρους όπως το περιεχόμενο που πρόκειται να αντιστοιχίσουν, δηλαδή εννοιολογικό ή δομικό, την ανάγκη συμμετοχής του χρήστη ή όχι, τον περιορισμό στο σχήμα ή και στα στιγμιότυπα, και τον τύπο χρήσης, δηλαδή ανεξάρτητα ή σε συνδυασμό με άλλα υποσυστήματα. Η εφαρμογή συμπεριλαμβάνει δείκτες αποδοτικότητας όπως ακρίβεια, ευστοχία και μετρήσεις χρόνου.

Ο χρήστης ελέγχει τις παραμέτρους του συστήματος από ένα γραφικό περιβάλλον από όπου μπορεί να επιλέξει από μια λίστα τα υποσυστήματα ευθυγράμμισης που θα χρησιμοποιηθούν και τις αντίστοιχες παραμέτρους τους. Η εφαρμογή ακολουθεί μια αντικειμενοστρεφή σχεδίαση ούτως ώστε να εξυπηρετεί καλύτερα αυτό το σκοπό. Τα υποσυστήματα ευθυγράμμισης χωρίζονται σε τρία επίπεδα όπου το πρώτο δημιουργεί πίνακες ομοιοτήτων ενώ το δεύτερο και τρίτο βελτιώνουν τα αποτελέσματα του πρώτου. Οι μέθοδοι που μπορεί να επιλέξει ο χρήστης περιλαμβάνουν τις συνήθεις που απαντώνται και στα υπόλοιπα συστήματα αλλά και ορισμένες που αναπτύχθηκαν ειδικά για αυτή την εφαρμογή. Οι είσοδοι του προγράμματος μπορούν να είναι οντολογίες σε XML, RDFS, OWL ή σχήματα N3 και εμφανίζονται σε μορφή δέντρου στο γραφικό περιβάλλον. Οι αντιστοιχίσεις στην έξοδο μπορούν να είναι και αυτές σε διαφορετική μορφή όπως πχ XML ή Excel.

Τα αποτελέσματα της ευθυγράμμισης μπορούν να εκτιμηθούν από τις μετρικές, ή την οπτική παρουσίασή τους. Στην περίπτωση που είναι διαθέσιμη κάποια ευθυγράμμιση αναφοράς, το αποτέλεσμα μπορεί να ελεγχθεί και από αυτόματη σύγκριση από την οποία μπορεί να γίνει ρύθμιση των παραμέτρων των μεθόδων ευθυγράμμισης.

Το εργαλείο αναπτύχθηκε στο πανεπιστήμιο του Σικάγο και είναι διαθέσιμο στη σελίδα http://agreementmaker.org/.

4.11 ++ Spicy

Το ++Spicy [11] είναι ένα εργαλείο ευθυγράμμισης σχήματος ανοιχτού κώδικα. Έχει σχεδιαστεί με σκοπό την εφαρμογή σε πραγματικά προβλήματα διαχείρισης, συγχώνευσης και εκκαθάρισης δεδομένων καθώς και διαδικασιών εξαγωγής, μετατροπής και φόρτωσης (ETL – Extract, Transform, Load) σε αποθήκες δεδομένων. Για αυτό το λόγο δίνεται βάρος στον επιτυχή και γρήγορο υπολογισμό λύσεων χωρίς να θυσιάζεται η βελτιστοποίηση τους.

Το σύστημα δέχεται στην είσοδό του τα δύο σχήματα (πηγής και στόχου) μαζί με τους περιορισμούς κλειδιών και ξένων κλειδιών, γενικά χαρακτηριστικά της ευθυγράμμισης και ένα ή περισσότερα στιγμιότυπα του σχήματος στόχου. Η έξοδος του συστήματος είναι μια εκτελέσιμη μετατροπή στο σχήμα πηγής που περιλαμβάνει την λύση που κατασκευάστηκε. Παράγεται επίσης ένα αρχείο κώδικα σε SQL ή XQuery που μπορεί να δοθεί σε μια εξωτερική μηχανή απάντησης ερωτημάτων. Η αλληλεπίδραση με το χρήστη γίνεται σε γραφικό περιβάλλον και υποστηρίζεται σχεσιακό σχήμα και σε μορφή XML. Η έξοδος του συτήματος περιλαμβάνει ελάχιστους πλεονασμούς εγγραφών και ενημερώνει το χρήστη για τυχόν ασυνέπειες.

Η απόδοση σε μεγάλους όγκους δεδομένων ή πολύπλοκα σχήματα είναι πολύ καλή με βάση το χρόνο καθώς και την ποιότητα αποτελεσμάτων με το εργαλείο να πετυχαίνει μία πολύ καλή ισορροπία σε αυτά τα δύο μεγέθη. Είναι διαθέσιμο στην σελίδα [11].

4.12 Alignment API

Το Alignment API [12] είναι μια διεπαφή για την ευθυγράμμιση οντολογιών. Το σύστημα έχει σχεδιαστεί με σκοπό την παροχή πλαισίου για την αντιστοίχηση και ευθυγράμμιση οντολογιών και θεμελίων για εργαλεία διαχείρισης εργασιών σύγκρισης, ερμηνείας και ανάλυσης. Σε σχέση με άλλα αντίστοιχα εργαλεία υπερτερεί στο ότι παρέχει μια αυτόνομη και φορητή βιβλιοθήκη.

Το εργαλείο διαθέτει τέσσερεις κύριες κλάσεις. Η κλάση των αντιστοιχίσεων είναι η βασική κλάση του API και αντιπροσωπεύει ένα σύνολο κελιών και μεταδεδομένων για την αντιστοίχηση, όπως στοιχεία για τις οντολογίες ή άλλα δεδομένα που συνδέονται με την ευθυγράμμιση. Τα κελιά είναι κλάσεις που γενικά αντιπροσωπεύουν μια αντιστοιχία μεταξύ δύο οντοτήτων, δηλαδή στοιχείων μιας

οντολογίας ή και άλλα πιο τεχνικά μεταδεδομένα. Η κλάση του δικτύου οντολογιών δέχεται ένα σύνολο οντολογιών και ένα σύνολο αντιστοιχίσεων και βοηθά στην ευκολία της διαχείρισης αυτών. Τέλος, η κλάση της σχέσης μεταξύ οντοτήτων. Οι κλάσεις αυτές δίνουν πρόσβαση στην πληροφορία των στιγμιοτύπων και σε μεθόδους επεξεργασίας της.

Εκτός από την αποθηκευτική δομή του, το σύστημα προσφέρει και μια δομή επεξεργασίας για εργασίες ευθυγράμμισης, με μια διεπαφή για όλους τους ταιριαστές, μια δεύτερη για εκτιμητές αποτελεσμάτων και μια τρίτη για χρήστες που πραγματοποιούν διαφορετικές μορφοποιήσεις στις αντιστοιχήσεις. Αυτές οι τρεις διεπαφές αντιπροσωπεύονται από αντίστοιχες κλάσεις.

Το API προσφέρει μια βασική εφαρμογή όλων των δομών για τις τυπικές εργασίες που προσφέρουν τα αντίστοιχα εργαλεία όπως η δημιουργία και επεξεργασία ευθυγραμμίσεων. Αυτή η εφαρμογή μπορεί να επεκταθεί ή να οριστεί μια καινούργια με τις δομές του API με πολλές διαφορετικές δυνατότητες όπως βιβλιοθήκες για επεξεργασία διάφορων μορφοποιήσεων όπως OWL, SKOS, HTML, RDF, XSLT, SWRL, ή βιβλιοθήκες με εκτιμητές και ταιριαστές. Επιπλέον, το API προσφέρει διεπαφές για συνεργασία με άλλα APIs επεξεργασίας οντολογιών με το πακέτο Ontowrap. Ακόμη, με το πακέτο Ontosim, προσφέρεται μία διεπαφή για τον υπολογισμό της ομοιότητας μεταξύ οντολογιών και στοιχείων τους καθώς και εφαρμογής διαφόρων διαθέσιμων μετρικών ή ανάπτυξής νέων, με έμφαση στην συγκέντρωση στοιχείων από πίνακες ομοιότητας. Τέλος, όλες οι λειτουργίες του API υποστηρίζουν την πρότυπη γλώσσα EDOAL (http://alignapi.gforge.inria.fr/edoal.html) που περιέχει ένα σύνολο κατασκευαστών και τελεστών ειδικά για την ευθυγράμμιση με δομή κοντά σε αυτή της OWL.

Το Alignment API αναπτύσσεται πάνω από δέκα χρόνια, από τα ερευνητικά εργαστήρια INRIA και LIG στη Γαλλία. Είναι διαθέσιμο μέσω της ιστοσελίδας http://alignapi.gforge.inria.fr/.

4.13 PARIS

Το σύστημα PARIS [13] επιτρέπει την αυτόματη ευθυγράμμιση οντολογιών σε RDF. Αντιστοιχίζει στιγμιότυπα, σχέσεις και κλάσεις οντολογιών. Η ευθυγράμμιση σε επίπεδο στιγμιοτύπου γίνεται σε συνδυασμό με την ευθυγράμμιση στο επίπεδο

σχήματος παρέχοντας έτσι μία συνολική αντιμετώπιση του προβλήματος της αντιστοίχισης.

Η βασική προσέγγιση του εργαλείου είναι πιθανοτική με την έννοια ότι υπολογίζεται κάποιος βαθμός αντιστοίχισης ανάλογα με πιθανοτικές εκτιμίσεις. Το σύστημα με αυτόν τον τρόπο πραγματοποιεί την ευθυγράμμιση χωρίς να είναι απαραίτητος ο καθορισμός παραμέτρων. Σύμφωνα με πειραματική αξιολόγηση το εργαλείο πετυχαίνει ακρίβεια 90% σε περιπτωσεις οντολογιών μεγάλης κλίμακας.

Το εργαλείο καθώς διατίθεται μέσω της ιστοσελίδας http://webdam.inria.fr/paris/.

4.14 Hertuda

Το Hertuda [14] είναι απλό εργαλείο ευθυγράμμισης οντολογιών που βασίζεται τη σύγκριση συμβολοσειρών και στο φιλτράρισμα των μη σχετικών αντιστοιχίσεων. Παρά την απλότητά του παρουσιάζει πολύ καλές επιδόσεις σε σύγκριση με τα υπάρχοντα εργαλεία αντιστοίχισης.

Το σύστημα αντιστοιχίζει αντικείμενα βασιζόμενο στη σύγκριση συμβολοσειρών και υποστηρίζει την εκφραστικότητα της OWL DL. Οι έννοιες, οι σχέσεις αντικειμένων και οι σχέσεις δεδομένων αντιμετωπίζονται ξεχωριστά. Για κάθε ένα από αυτά τα δομικά στοιχεία της οντολογίας ορίζονται τρια ξεχωριστά κατώφλια. Εάν ο βαθμός εμπιστοσύνης για μία σύγκριση είναι υψηλότερος από το σχετικό κατώφλι τότε η αντιστοίχιση λαμβάνεται υπόψιν στην έξοδο του συστήματος.

Διατίθεται μέσω της ιστοσελίδας http://www.ke.tu-darmstadt.de/resources/ontology-matching/hertuda.

4.15 Coma 3.0

Το Coma 3.0 είναι ένα εργαλείο αντιστοίχισης οντολογιών που στοχεύει στην εύρεση σημασιολογικών αντιστοιχίσεων μεταξύ των δομών μεταδεδομένων που περιγράφονται σε XML σχήματα, βάσεις δεδομένων ή οντολογίες.

Αποτελεί επέκταση των πρότυπων εργαλείων COMA και COMA++ επιτρέποντας την ευθυγράμμιση οντολογιών. Είναι κατάλληλο για την αντιμετώπιση προβλημάτων πραγματικού κόσμου. Η γραφική διεπαφή χρήστη παρέχει ένα σύνολο από

ευκολίες που επιτρέπουν στο χρήστη να επεμβαίνει στη διαδικασία αντιστοίχισης με ποικίλους τρόπους.

Το σύστημα επιτρέπει την αντιστοίχιση με βάση τρεις διαφορετικές προσεγγίσεις. Η πρώτη προσέγγιση πραγματοποιεί την αντιστοίχιση με βάση το περιεχόμενο και αφορά στις περιπτώσεις όπου αντιστοιχίζονται διαφορετικά σχήματα που περιγράφουν στοιχεία με κοινό περιεχόμενο. Η δεύτερη προσέγγιση εφαρμόζει την τεχνική "διαίρει και βασίλευε" και αφορά στις περιπτώσεις μεγάλων σχημάτων. Σκοπός της συγκεκριμένης προσέγγισης είναι η βελτίωση του χρόνου εκτέλεσης αλλά και της ποιότητας της αντιστοίχισης σε επίπεδο σχήματος. Τέλος, η προσέγγιση της επαναχρησιμοποίησης επιτρέπει την αξιοποίηση υπάρχουσας πληροφορίας αντιστοίχισης.

Το σύστημα είναι διαθέσιμο μέσω της ιστοσελίδας http://dbs.uni-leipzig.de/en/Research/coma_index.html.

4.16 Codi-Matcher

Το Codi (Combinatorial Optimization for Data Integration) [15] είναι ένα σύστημα πιθανολογικής και λογικής ευθυγράμμισης το οποίο παρέχει ένα πλαίσιο για την ευθυγράμμιση των ατόμων, των εννοιών και των ιδιοτήτων δύο ετερογενών οντολογιών. Το CODI συνδυάζει τις λεξιλογικές μετρήσεις ομοιότητας και τις λογικές πληροφορίες του σχήματος, προκειμένου να αντιμετωπίσει αποτελεσματικά την δημιουργία ασυνέπειας κατά τη διαδικασία ευθυγράμμισης. Οι ευθυγραμμίσεις υπολογίζονται μέσω της επίλυσης συνδυαστικών προβλημάτων βελτιστοποίησης. Επίσης μπορούν να αναγνωριστούν ζεύγη οντολογιών που ανήκουν σε διαφορετικές εκδόσεις της ίδιας οντολογίας.

Το CODI βασίζεται στο συντακτικό και τη σημασιολογία της λογικής Markov, που συνδυάζει τη λογική πρώτης τάξης και τους μη κατευθυνόμενους πιθανολογικούς γράφους, και μετασχηματίζει το πρόβλημα της ευθυγράμμισης σε ένα πρόβλημα μέγιστης εκ των υστέρων (maximum-a-posteriori) βελτιστοποίησης. Η λογική Markov είναι ικανή να εντοπίσει τόσο τα ισχυρά λογικά αξιώματα όσο και τα αρκετά αβέβαια σε σχέση με τις πιθανές αντιστοιχίσεις ορίζοντας μια κατανομή πιθανότητας των πιθανών ευθυγραμμίσεων.

Κατά την ευθυγράμμιση εφαρμόζονται αυστηροί περιορισμοί πληθικότητας που επιβάλουν την επιλογή μιας συναρτησιακής ευθυγράμμισης με πληθικότητα ένα-προς-ένα καθώς και περιορισμοί μείωσης της ασυνέπειας των ευθυγραμμίσεων σε

σχέση με την οντολογία, κατά τη διάρκεια της διαδικασίας υπολογισμού των ευθυγραμμίσεων (και όχι σε επόμενο στάδιο). Εφαρμόζονται επίσης περιορισμοί που εξασφαλίζουν ότι οι ισοδυναμίες εννοιών που προκύπτουν δε θα αλλάζουν τα δομικά στοιχεία της γνώσης, δηλαδή δεν θα καθιστούν ισοδύναμες έννοιες που οι θυγατρικές τους έννοιες (child concepts) στην ιεραρχία δεν θα είναι ισοδύναμες.

Όσον αφορά στον υπολογισμό της ευθυγράμμισης των στιγμιότυπων υλοποιήθηκε μια προσέγγιση προκειμένου να μην υπολογίζεται η λεξιλογική ομοιότητα για όλα τα διαθέσιμα ζευγάρια. Η προσέγγιση αυτή χρησιμοποιεί τις ιδιότητες των αντικειμένων προκειμένου να επιλέξει τα ζευγάρια για τα οποία πρέπει να υπολογιστεί η ομοιότητα. Αυτή θεωρεί ότι υπάρχει μία κοινή οντολογία-σώμα ορολογίας και δύο διαφορετικα σύνολα ισχυρισμών, επομένως θεωρεί ότι οι δύο οντολογίες έχουν ήδη ενσωματωθεί. Αρχικά υπολογίζονται οι ευθυγραμμίσεις αναφοράς, συγκρίνοντας ένα μικρό υποσύνολο των ατόμων μεταξύ τους (π.χ. όλα τα άτομα που αποδίδονται σε μια συγκεκριμένη έννοια, όπως Κινηματογράφος), υπολογίζοντας τις λεξιλογικές ομοιότητες τους και προσθέτοντας τα στις ευθυγραμμίσεις αναφοράς αν οι τιμές ομοιότητας τους ξεπερνούν κάποιο κατώφλι. Στη συνέχεια υπολογίζεται η λεξιλογική ομοιότητα για όλα τα άτομα που συνδέονται με ένα από τα άτομα του πρώτου ζεύγους ευθυγράμμισης και προσθέτονται στο τέλος της λίστας αναφορών αν οι τιμές ξεπερνούν κάποιο κατώφλι. Στη συνέχεια η διαδικασία επαναλαμβάνεται για όλες τις ευθυγραμμίσεις του αρχικού συνόλου.

Κατά τη διάρκεια της διαδικασίας ευθυγράμμισης γίνονται έλεγχοι συνέπειας ενώ παραλείπονται τα ζευγάρια που δεν έχουν καμία κοινή υπο-έννοια προκειμένου να μειωθεί ο αριθμός των συγκρίσεων. Η ιδέα αυτή επεκτείνεται με κάποια βήματα μετά την επεξεργασία. Προκειμένου να εντοπιστούν οι αντιστοιχίες οι οποίες δεν συνδέονται με μια ιδιότητα αντικειμένου, συγκρίνονται μεταξύ τους όλα τα υπόλοιπα άτομα που δεν έχουν εισαχθεί στις ευθυγραμμίσεις αναφοράς και αν η ομοιότητα τους ξεπερνάει το κατώφλι εισάγονται στο σύνολο. Στο τέλος χρησιμοποιείται ένας άπληστος αλγόριθμος για τον υπολογισμό μια ευθυγράμμισης ένα-προς-ένα. Οι τεχνικές αυτές μειώνουν σημαντικά το χρόνο εκτέλεσης όταν υπάρχει πολύ μεγάλος αριθμός στιγμιότυπων.

Το Codi διατίθεται δωρεάν υπό ακαδημαϊκή άδεια και είναι διαθέσιμο για λήψη μαζί με οδηγίες εγκατάστασης (για Linux και Windows) στο Google code (https://code.google.com/p/codi-matcher/).

4.17 MapOnto

Το MapOnto [16] είναι ένα ερευνητικό πρόγραμμα που στοχεύει στην εύρεση σημασιολογικών αντιστοιχίσεων μεταξύ διαφορετικών μοντέλων δεδομένων, για παράδειγμα σχήματα βάσεων δεδομένων, εννοιολογικά σχήματα και οντολογίες. Στα πλαίσια του έχουν αναπτυχθεί εργαλεία για την ανακάλυψη σημασιολογικών αντιστοιχίσεων είτε μεταξύ σχημάτων βάσεων δεδομένων και οντολογιών είτε μεταξύ δύο σχημάτων βάσεων δεδομένων.

Το εργαλείο ευθυγράμμισης είναι ένα πρότυπο ερευνητικό εργαλείο που λειτουργεί με διαδραστικό και ημι-αυτόματο τρόπο. Ξεκινώντας από ένα σύνολο αντιστοιχίσεων τύπου 'χαρακτηριστικό προς χαρακτηριστικό', το εργαλείο αναλύει τη σημασιολογία των δύο μοντέλων εισόδου και παράγει ένα σύνολο λογικών τύπων που αντιπροσωπεύουν τη σημασιολογική σχέση μεταξύ των δύο μοντέλων. Στη συνέχεια η λίστα με τους τύπους ταξινομείται βάζοντας τις πιο «λογικές» αντιστοιχίσεις στην κορυφή της. Τελικά ο χρήστης μπορεί να επιλέξει κάποια αντιστοίχιση από τη λίστα.

Το Maponto υποστηρίζεται μέσω μίας διασύνδεσης στην πλατφόρμα Protege. Προς το παρόν λειτουργεί μόνο με το Protege 3.0 και παλιότερες εκδόσεις του. Στην είσοδό του δέχεται αρχεία SQL DDL για σχεσιακές βάσεις και αρχεία OWL για οντολογίες. Το plugin MAPONTO έχει υλοποιηθεί ως μια διεπαφή χρήστη όπου οι χρήστες μπορούν να επιλέξουν σχήματα βάσεων δεδομένων και οντολογίες, να δημιουργήσουν και να επεξεργαστούν αντιστοιχίσεις, να παράγουν και να επεξεργαστούν τις υποψήφιες χαρτογραφήσεις και να παράγουν και να αποθηκεύσουν τις τελικές αντιστοιχίσεις σε αρχεία.

Ο πηγαίος κώδικας και το εκτελέσιμο του MapOnto είναι διαθέσιμα για λήψη στο sourceforge (http://sourceforge.net/projects/maponto/) ενώ οδηγίες εγκατάστασης και χρήσης μαζί με κάποια δεδομένα δοκιμής είναι διαθέσιμα στη σελίδα του (http://www.cs.toronto.edu/semanticweb/maponto/).

4.18 MatchIT

Το MatchIT [17] είναι ένα εργαλείο με σκοπό την αυτοματοποίηση και τη διευκόλυνση της χαρτογράφησης των σχημάτων διαφορετικών πηγών δεδομένων κάνοντας χρήση της σημασιολογίας των όρων που χρησιμοποιούνται στα σχήματα. Το MatchIT περιλαμβάνει την αγγλική οντολογία WordNet, αλλά προσφέρει στους

χρήστες τη δυνατότητα να εγκαταστήσουν και να χρησιμοποιήσουν τις δικές τους οντολογίες.

Πριν εκτελέσει τη διαδικασία της χαρτογράφησης το MatchIT χρησιμοποιεί διάφορες τεχνικές προκειμένου να δημιουργήσει ένα λεξιλόγιο των σχημάτων εξάγοντας μεμονωμένες λέξεις από τα ονόματα των όρων των σχημάτων και στη συνέχεια αντιστοιχίζει αυτές τις λέξεις σε σχετικές της εγκατεστημένης οντολογίας. Το λεξιλόγιο περιέχει όλες τις λέξεις που χρησιμοποιούνται στα σχήματα, καθώς και το νόημα τους όπως ορίζεται από την οντολογία. Αν χρειαστεί μπορεί να τροποποιηθεί ώστε να αναπαριστά με ακρίβεια τον τομέα των σχημάτων. Το λεξιλόγιο μπορεί να εξαχθεί σαν αρχείο OWL για χρήση σε άλλα εργαλεία.

Το MatchIT χρησιμοποιεί το λεξιλόγιο για τον υπολογισμό της σημασιολογικής ομοιότητας των υποψήφιων προς αντιστοίχιση όρων των σχημάτων. Μπορεί να αντιστοιχίσει αυτόματα διαφορετικά σχήματα προς ένα σχήμα στόχο. Οι υποψήφιες αντιστοιχίσεις κατατάσσονται και βαθμολογούνται με χρήση αλγορίθμων συμβολοσειρών, σημασιολογικών αλγορίθμων και προσαυξημένης ομοιότητας και τους αποδίδονται μεταδεδομένα. Τα λεξιλόγια που προκύπτουν και οι σημασίες των εννοιών τους μπορούν να τροποποιηθούν και οι αλγόριθμοι αντιστοίχισης μπορούν να ρυθμιστούν έτσι ώστε να μπορούν να παράγουν εναλλακτικές προτάσεις αντιστοίχισης.

Το MatchIT παρέχει διαγνωστικά στους χρήστες που επεξηγούν κάθε προτεινόμενη αντιστοίχιση διευκολύνοντας τους να κατανοήσουν τη σημασιολογία. Τους επιτρέπει επίσης να συνεργάζονται, να δημιουργούν και να συντηρούν διαφορετικές εργασίες αντιστοίχισης οι οποίες διατηρούν όλα τα αποτελέσματα, τις εισαγωγές και τις εξαγωγές, τις τροποποιήσεις του λεξιλογίου και τα λεξιλόγια που δημιουργηθήκαν. Επιτρέπει επίσης την καταγραφή, τις σημειώσεις και τα σχόλια που επιτρέπουν την παρακολούθηση της προόδου και την παρακολούθηση των εκκρεμοτήτων.

Διατίθεται είτε ως αυτόνομη εφαρμογή είτε ως διεπαφή στο Eclipse [17].

4.19 Falcon-AO

Το Falcon-AO [18] είναι ένα εργαλείο αυτόματης ευθυγράμμισης οντολογιών του διαδικτύου οι οποίες είναι εκφρασμένες σε RDF (S) και OWL.

Το Falcon-AO αποτελείται από πέντε δομικά στοιχεία. Την αποθήκη δεδομένων για την προσωρινή αποθήκευση τους κατά τη διαδικασία αντιστοίχισης, το χώρο μοντέλων (Model Pool) για τη διαχείριση των οντολογιών και τη δημιουργία των διαφορετικών μοντέλων για τους διαφορετικούς ευθυγραμμιστές, το σύνολο ευθυγράμμισης για τη δημιουργία και την αξιολόγηση των ευθυγραμμίσεων που προκύπτουν, τη βιβλιοθήκη των ευθυγραμμιστών του και τον κεντρικό ελεγκτή για τη σύνθεση των στρατηγικών ευθυγράμμισης και την εκτέλεση των λειτουργιών ταιριάσματος.

Η βιβλιοθήκη των ευθυγραμμιστών αποτελείται από τέσσερις διακριτούς ευθυγραμμιστές. Οι V-Doc και I-Sub είναι δύο ελαφρείς (light-weighted) γλωσσολογικοί ευθυγραμμιστές όπου ο πρώτος υπολογίζει τις ομοιότητες των εννοιών με βάση ένα σύνολο λέξεων που εξάγει από το όνομα, τις ετικέτες, τα σχόλια και τους γείτονες της κάθε μίας ενώ ο δεύτερος βασίζεται στις συμβολοσειρές. Ο GMO είναι ένας επαναληπτικός δομικός ευθυγραμμιστής που μετράει τη δομική ομοιότητα μεταξύ γραφημάτων RDF με βάση τις ομοιότητες που κληρονομούνται κατά την ευθυγράμμιση. Ο PBM χρησιμοποιεί την στρατηγική "διαίρει και βασίλευε" για να διαιρέσει σε τμήματα τις οντολογίες μεγάλης κλίμακας και να εντοπίσει τις αντιστοιχίες μεταξύ αυτών των τμημάτων με χρήση των υπολοίπων ευθυγραμμιστών.

Το Falcon-AO χρησιμοποιεί κάποιους κανόνες συντονισμού για να μειώσει την ετερογένεια μεταξύ των οντολογιών προς ευθυγράμμιση. Χρησιμοποιεί επίσης μια νέα προσέγγιση για την αντιστοίχιση "πολλά προς πολλά". Πρόκειται για μια προσέγγιση διάσπασης που λαμβάνει υπόψη τα γλωσσικά και τα δομικά χαρακτηριστικά των οντοτήτων και αξιολογεί τα αποτελέσματα της αντιστοίχισης. Επίσης χρησιμοποιεί μια προσέγγιση για την αντιστοίχιση οντολογιών με σχήματα βάσεων δεδομένων η οποία βρίσκει απλές αντιστοιχίσεις μεταξύ τους με χρήση εικονικών εγγράφων και ελέγχει την συνέπεια των αντιστοιχίσεων.

Το εργαλείο παρέχει μια γραφική διεπαφή χρήστη που επιτρέπει στους χρήστες να θέτουν παραμέτρους ταιριάσματος καθώς και να προβάλουν και να χειρίζονται τις

προκύπτουσες ευθυγραμμίσεις. Το Falcon-AO έχει υλοποιηθεί σε Java και είναι ένα λογισμικό ανοικτού κώδικα που διατίθεται υπό την άδεια Apache 2.0. (http://ws.nju.edu.cn/falcon-ao/res/falcon.zip).

4.20 S-Match

Το S-Match [19] είναι ένα εργαλείο σημασιολογικής αντιστοίχισης το οποίο μπορεί να πάρει σαν είσοδο δύο δεντροειδείς δομές όπως σχήματα βάσεων δεδομένων ή lightweight οντολογίες. Το εργαλείο παρέχει διάφορους αλγορίθμους σημασιολογικής αντιστοίχισης ενώ είναι επεκτάσιμο και υποστηρίζει την ανάπτυξη νέων αλγορίθμων.

Οι αλγορίθμοι σημασιολογικής αντιστοίχισης που υλοποιεί είναι ο αρχικός αλγόριθμος S-Match, ο αλγόριθμος ελάχιστης αντιστοίχισης και ο αλγόριθμος αντιστοίχισης που διατηρεί τη δομή. Ο αρχικός αλγόριθμος είναι γενικού σκοπού, προσαρμόσιμος και κατάλληλος για πολλές εφαρμογές. Ο ελάχιστος αλγόριθμος εκμεταλλεύεται την επιπλέον γνώση που υπάρχει κωδικοποιημένη στη δομή της εισόδου και είναι ικανός να παράγει την ελάχιστη ευθυγράμμιση, που παραλείπει τις περιττές αντιστοιχίσεις ανάλογα με κάποια κριτήρια και είναι κατάλληλη για χειροκίνητη αξιολόγηση, και τη μέγιστη ευθυγράμμιση η οποία περιέχει όλες τις πιθανές αντιστοιχίσεις και είναι κατάλληλη για εφαρμογές που δε λαμβάνουν υπόψη τη σημασιολογία. Ο αλγόριθμος διατήρησης της δομής είναι κατασκευασμένος για την ευθυγράμμιση δομικά όμοιων στοιχείων και είναι κατάλληλος για διεπαφές υπηρεσιών ευθυγράμμισης και σχήματα βάσεων δεδομένων. Αντιστοιχίζει τις εισόδους κάνοντας διάκριση μεταξύ των δομικών στοιχείων, όπως οι συναρτήσεις ή οι μεταβλητές.

Το S-Match παρέχει πρόσβαση είτε μέσω γραμμής εντολών είτε μέσω της γραφικής διεπαφής του η οποία διαθέτει διάφορες γραμμές εργαλείων για την επεξεργασία των αντιστοιχίσεων και τη διευκόλυνση της διαδικασίας ευθυγράμμισης.

Το S-Match είναι ένα λογισμικό ανοιχτού κώδικά το οποίο διανέμεται υπό την άδεια LGPL License. Είναι διαθέσιμο για λήψη στο SourceForge (http://sourceforge.net/projects/s-match/) και στο GitHub (https://github.com/s-match/s-match-core/wiki) όπου είναι διαθέσιμο και ένα εγχειρίδιο χρήσης.

4.21 CogZ

Το CogZ [20] είναι ένα εργαλείο για την υποστήριξη και ανθρώπινη καθοδήγηση της διαδικασίας ευθυγράμμισης οντολογιών. Το εργαλείο συμπεριλαμβάνεται στην τελευταία έκδοση του Protégé και επεκτείνει το υπάρχον εργαλείο χαρτογράφησης (και διεπαφή του Protégé) PROMPT.

Σκοπός του CogZ είναι η διευκόλυνση της αλληλεπίδρασης μεταξύ του χρήστη και του εργαλείου ευθυγράμμισης αντιμετωπίζοντας θέματα όπως η δυσκολία των χρηστών να θυμούνται τι έχουν εξετάσει και εκτελέσει, να κατανοήσουν την έξοδο του αλγορίθμου, να θυμούνται γιατί έκαναν μια πράξη, να αντιστρέφουν τις αποφάσεις τους και να διαθέτουν δικαιολογητικά στοιχεία που στηρίζουν τις αποφάσεις τους. Η αρχιτεκτονική του CogZ επιτρέπει στους ερευνητές να ενσωματώσουν εύκολα τους δικούς τους αλγορίθμους, διάφορα στοιχεία διεπαφών και μορφές αρχείων χαρτογράφηση επεκτείνοντας την διεπαφή PROMPT.

Το εργαλείο εισάγει απεικονίσεις για την υποστήριξη της δραστηριότητας του χρήστη. Στο συνδυασμό των εργαλείων PROMPT και COGZ χρησιμοποιούνται η αναπαράσταση δενδρικής δομής για την επισκόπηση της οντολογίας και των πιθανών αντιστοιχίσεων, που επιτρέπει την προβολή μεγάλης ποσότητας δεδομένων σε μικρή επιφάνεια της διεπαφής και επιτρέπουν την εύκολη κλιμάκωση της απεικόνισης για μεγάλες οντολογίες. Οι περιοχές με πολλές υποψήφιες αντιστοιχίσεις και οι περιοχές που έχουν ήδη αντιστοιχιστεί περιγράφονται με χρώματα διαφορετικής έντασης. Παρέχεται επίσης οπτικοποίηση σχετικά με τον αριθμό των υποψήφιων αντιστοιχίσεων, των εννοιών που έχουν ήδη χαρτογραφηθεί και των εννοιών χωρίς αντιστοιχίσεις σε κάθε κομμάτι της οντολογίας, δίνοντας μια γενική εικόνα του βαθμού ολοκλήρωσης της διαδικασίας. Υποστηρίζεται επίσης η σύγκριση των γειτονιών των εννοιών, δηλαδή των περιοχών που έχουν άμεσες δομικές σχέσεις με κάθε έννοια, στην οποία απεικονίζεται η δομική σύγκριση μεταξύ δύο εννοιών.

Η διεπαφή του COGZ παρέχει φίλτρα χαρτογράφησης που βασίζονται στις κατηγορίες πιθανών αντιστοιχίσεων του PROMPT (π.χ. ταίριασμα ονομάτων ή συνωνύμων), τα οποία μειώνουν τον αριθμό των αντιστοιχίσεων που προβάλλονται επιτρέποντας στο χρήστη να επικεντρωθεί σε συγκεκριμένους τύπους αντιστοιχίσεων. Δίνεται επίσης η δυνατότητα χρήσης ιεραρχικών φίλτρων για την προβολή αντιστοιχίσεων που ανήκουν σε συγκεκριμένες περιοχές της οντολογίας. Για την υποστήριξη της μνήμης του χρήστη κατά τη διάρκεια της διαδικασίας, του

δίνεται η δυνατότητα να σχολιάζει κάποιες υποψήφιες αντιστοιχίσεις σαν προσωρινές και να τις προβάλει στη συνέχεια μόνες τους ή μαζί με τις υπόλοιπες με διαφορετικό χρώμα, ενώ ο χρήστης ειδοποιείται όταν εκτελεί χαρτογράφηση με μια έννοια που έχει ήδη μια προσωρινή. Παρέχεται επίσης στο χρήστη η δυνατότητα σχολιασμού για τη δικαιολόγηση κάποιας αντιστοίχισης.

Η δεντρική οπτικοποίηση του COGZ υποστηρίζει το σημασιολογικό ζουμ το οποίο εμφανίζει τις περιπτώσεις πολλαπλής κληρονομικότητας και τη διαδραστική αναζήτηση η οποία φιλτράρει τα δέντρα της οντολογίας με βάση το ερώτημα του χρήστη. Τα δέντρα μπορούν επίσης να φιλτραριστούν ώστε να εμφανίζουν μόνο τους όρους με ή χωρίς αντιστοιχίσεις για πιο εύκολη επισκόπηση της διαδικασίας από το χρήστη. Τα παραπάνω διευκολύνουν την οπτικοποίηση και τη διαδικασία ευθυγράμμισης όταν αυξάνεται ο όγκος των αντιστοιχίσεων.

5. ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Faria D., Pesquita C., Santos E., Palmonari M., Cruz I.F., Couto F.M. The Agreement Maker Light Ontology Matching System. In: Meersman R., Panetto H., Dillon T.S., Eder J., Bellahsene Z., Ritter N., Leenheer P.D., Dou D., editors. On the Move to Meaningful Internet Systems: OTM 2013 Conferences—Confederated International Conferences. Vol. 8185. Lecture Notes in Computer Science; Springer; Berlin/Heidelberg, Germany: 2013. pp. 527–541.

[2] Abderrahmane Khiat, Moussa Benaissa. AOT/AOTL Results for OAEI 2014. In Proceedings of the 9th International Workshop on Ontology Matching co-located with the 13th International Semantic Web Conference, ISWC 2014.

[3] Elena Voyloshnikova, Bo Fu, Lars Grammel, Margaret-Anne Storey. BioMixer: Visualizing Mappings of Biomedical Ontologies. The Third International Conference on Biomedical Ontologies (ICBO 2012), 2012.

[4] R. Kolli, P. Doshi. Optima: Tool for ontology alignment with application to semantic reconciliation of sensor metadata for publication in sensormap. Semantic Computing, IEEE International Conference, 2008.

[5] https://www.cs.ox.ac.uk/isg/projects/LogMap/

[6] Z. Dragisic, K. Eckert, J. Euzenat, D. Faria, A. Ferrara, R. Granada, V. Ivanova, E. Jim{\'e}nez-Ruiz, A. Kempf, Andreas Oskar and P. Lambrix et. al. Results of the ontology alignment evaluation initiative 2014. In Proceedings of the 9th International Workshop on Ontology Matching Collocated with the 13th International Semantic Web Conference, ISWC 2014.

[7] S. Schwichtenberg, C. Gerth, Christian, G. Engels. RSDL Workbench Results for OAEI 2014.

[8] W. Djeddi, K. Eddine, T. Mohammed. XMAP: a novel structural approach for alignment of OWL-full ontologies. Machine and Web Intelligence (ICMWI), 2010 International Conference, pp. 368--373, IEEE, 2010.

[9] D. Ngo, Z. Bellahsene. YAM++: A multi-strategy based approach for ontology matching task. In Knowledge Engineering and Knowledge Management, pp. 421--425, Springer, 2012.

[10] I. Cruz, F. P. Antonelli, C. Stroe. AgreementMaker: efficient matching for large real-world schemas and ontologies. In Proceedings of the VLDB Endowment. Vol 2, pp. 1586--1589, VLDB Endowment, 2009.

[11] http://www.db.unibas.it/projects/spicy/

[12] J. David, J. Euzenat, F. Scharffe, C. T. Dos Santos. The alignment api 4.0. In Semantic web journal, Vol. 2, pp. 3—10, 2011.

[13] Fabian M. Suchanek, Serge Abiteboul, Pierre Senellart. PARIS: Probabilistic Alignment of Relations, Instances, and Relations. VLDB, 2012.

[14] Sven Hertling. Hertuda results for OEAI 2012. In Seventh International Workshop on Ontology Matching (OM 2012), 2012.

[15] J. Huber, T. Sztyler, J. Noessner, C. Meilicke. CODI: Combinatorial optimization for data integration--results for OAEI 2011. In Journal Ontology Matching, Vol. 134, 2011.

[16] R. Press. Ontology and database mapping: a survey of current implementations and future directions. Journal of Web Engineering, Vol. 7, pp.001—024, 2008.

[17] http://www.revelytix.com/?q=content/matchit

[18] W. Hu, Y. Qu. Falcon-AO: A practical ontology matching system. Web Semantics journal: Science, Services and Agents on the World Wide Web. Vol 6, pp. 237—239, Elsevier, 2008.

[19] F. Giunchiglia, P. Shvaiko, M. Yatskevich. S-Match: an algorithm and an implementation of semantic matching. ESWS, Vol. 3053, pp. 61—75, Springer, 2004.

[20] https://code.google.com/p/cogz/