Ανάπτυξη της ελληνικής ερευνητικής Υποδομής

για τις Ανθρωπιστικές Επιστήμες ΔΥΑΣ

MIS 441245

ΠΑΡΑΔΟΤΕΟ ΠΑ 2.8 ΙΤΕ 1

Οδηγίες συντήρησης αλληλοαναφορών

Σεπτέμβριος 2015

ΙΔΡΥΜΑ ΤΕΧΝΟΛΟΓΙΑΣ ΚΑΙ ΕΡΕΥΝΑΣ (ΙΤΕ)

ΥΠΟΕΡΓΟ: «ΔΡΑΣΕΙΣ ΑΝΑΠΤΥΞΗΣ ΤΗΣ ΔΥΑΣ»

ΕΠΙΧΕΙΡΗΣΙΑΚΑ ΠΡΟΓΡΑΜΜΑΤΑ : **«ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ & ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ» ΚΑΙ «ΠΕΡΙΦΕΡΕΙΩΝ ΣΕ ΜΕΤΑΒΑΣΗ»**

ΕΘΝΙΚΟ ΣΤΡΑΤΗΓΙΚΟ ΠΛΑΙΣΙΟ ΑΝΑΦΟΡΑΣ

ΕΣΠΑ 2007-2013

η περιφέρεια στο επίκεντρο της ανάπτυξης

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης - Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης (ΕΤΠΑ), στο πλαίσιο του Ε.Π. Ανταγωνιστικότητα και Επιχειρηματικότητα (ΕΠΑΝ ΙΙ) και των Π.Ε.Π. Αττικής, Π.Ε.Π. Μακεδονίας - Θράκης

ΠΑΡΑΔΟΤΕΟ ΠΑ 2.8 ITE 1

Έγγραφο:	**DARIAH KRHTH ΠΑ 2.8 ITE 1**
Τίτλος παραδοτέου	**ΟΔΗΓΙΕΣ ΣΥΝΤΗΡΗΣΗΣ ΑΛΛΗΛΟΑΝΑΦΟΡΩΝ**
Ενότητα Εργασίας:	**Διαχείριση περιεχομένου, αποθετήρια**
Υπεύθυνος Φορέας/μονάδα:	**ΙΤΕ – ΙΠ**
Μονάδα	**Εργαστήριο Πληροφοριακών Συστημάτων** **Κέντρο Πολιτισμικής Πληροφορικής**
Άλλοι συμμετέχοντες:	
Συγγραφείς:	**Κωνσταντίνα Κονσολάκη, Μαρία Δασκαλάκη, Μάρτιν Ντέρ, Αθηνά Κριτσωτάκη**
Κατάσταση Εγγράφου:	**Α' έκδοση**
Ημερομηνία πρώτης έκδοσης	Σεπτέμβριος 2015
Ημερομηνία τελευταίας επικαιροποίησης	
Σχετικοί σύνδεσμοι	

ΠΕΡΙΕΧΟΜΕΝΑ

1 Περίληψη

Στην ψηφιακή εποχή που διανύουμε και στην ταχύτατη ανάπτυξη του διαδικυακού κόσμου της πληροφορίας, υπάρχει πληθώρα αλληλοαναφορών και αλληλοσυσχετίσεων στα δεδομένα. Διαφορετική γνώση και διαφορετικές περιγραφές πολύ συχνά αναφέρονται στα ίδια στιγμιότυπα. Η διαχείριση και ταυτοποίηση αυτών παραμένει ένα άλυτο πρόβλημα στην πληροφορική.

Πολλοί πάροχοι πληροφοριών επιθυμούν πλέον να διαθέσουν τη γνώση που έχουν μαζέψει. Συχνά διαφορετικά αποθετήρια κρατούν πληροφορία που αναφέρεται στην ίδια οντότητα. Άρα υπάρχει ανάγκη για μία μεθοδολογίας διασύνδεσης όλων αυτών των δεδομένων.

Ο όρος «co-reference» όπως προσδιορίζεται στον τομέα της γλωσσολογίας, ορίζει την κατάσταση στην οποία διαφορετικοί όροι χρησιμοποιούνται για να αναφερθούν στο ίδιο αναφερόμενο. Στο σημασιολογικό δίκτυο (πληροφορική) ο όρος ορίζει μία κατάσταση όπου διαφορετικά URI χρησιμοποιούνται για να περιγράψουν τον ίδιο μη πληροφοριακό πόρο.

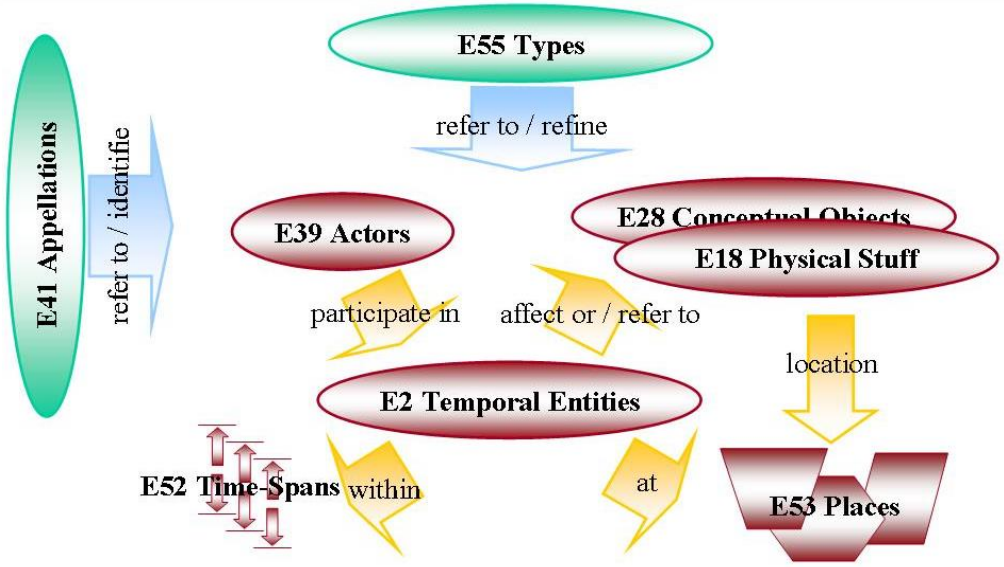
Με τον όρο «co-reference» εννοούμε την αναγνώριση συγκεκριμένων παραπομπών/αναφορών σε πρόσωπα, αντικείμενα, γεγονότα κ.λ.π. ανά διαφορετικά έγγραφα, στα οποία ο χρήστης-αναγνώστης ανακαλύπτει ότι «αναφέρονται στο ίδιο πράγμα». Ο μηχανισμός αυτός βασίζεται πάνω σε συσχετίσεις που βρίσκουμε σε ψηφιακά αντικείμενα ή μεταδεδομένα, σε προϋπάρχουσα γνώση.

Η αύξηση των πληροφοριακών πόρων σε μορφή RDF οδήγησε ταυτόχρονα σε μία αύξηση του πλήθους των URI που χρησιμοποιούνται για την ταυτοποίηση διαφορετικών οντοτήτων. Αυτή η πολλαπλότητα των URI δημιουργεί το πρόβλημα του co-reference, όπου διαφορετικά URI χρησιμοποιούνται για να περιγράψουν την ίδια οντότητα. Σε ένα ευρύ σημασιολογικό δίκτυο υπάρχει η ανάγκη να συνδεθεί γνώση από διαφορετικούς παρόχους πληροφορίας.

Μια ενδεικτική περίπτωση coreference εμφανίζεται στις ψηφιακές βιβλιοθήκες, όταν π.χ η ταυτότητα του συγγραφέα μιας δημοσίευσης χρειάζεται να αποσαφηνιστεί. Υπάρχουν δηλ. πολλοί συγγραφείς που μοιράζονται το ίδιο όνομα, και το πρόβλημα γίνεται πιο σύνθετο με την χρήση των αρχικών του ονόματος , ή διαφορετικών μορφών ονομασίας. Οι λύσεις για τέτοιας φύσης πρόβλημα ποικίλουν και συνήθως βασίζονται σε διαδικτυακές αναζητήσεις που καθορίζουν εαν ο συγγραφέας είναι ο ίδιος με κάποιον άλλον.

Στο παραδοτέο αυτό περιγραφεται η υπάρχουσα κατάσταση και προτείνεται ένα μοντέλο δημιουργίας υποθέσεων ταυτότητας - αναφοράς. Το μοντέλο αυτό βασίζεται στο CIDOC CRM [Crofts et.al, 2005].

Το CIDOC CRM είναι μία οντολογία που προτείνει μία δομή που βασίζεται στην τεκμηρίωση γεγονότων και διαδικασιών. Ορίζει συσχετίσεις μεταξύ εννοιών/οντοτήτων και όχι ορολογία. Αποτελεί ένα γενικό σχήμα που μοντελοποιεί έννοιες συμμετοχής/παρουσίας, μέρους-όλου, αναφοράς και ταξινόμησης σε θεμελιώδεις έννοιες που συνδέουν πρόσωπα, αντικείμενα, έννοιες, χρόνο και τόπο (εικ.1). Το CIDOC CRM δίνει έμφαση στην περιγραφή της πραγματικής γνώσης δημιουργώντας έτσι ένα πρώτο επίπεδο αφαίρεσης (μοντέλο). Επίσης δίδει τη δυνατότητα για προσθήκη πρόσθετων επιπέδων αφαίρεσης (μεταμοντέλων).



Εικόνα 1: Το CIDOC CRM : οι πιο γενικές έννοιες και συσχετίσεις

Η φιλοσοφία του CIDOC CRM (ISO 21127) βασίζεται σε τρείς παραδοχές:

- Η σχέση μεταξύ οντοτήτων και αναγνωριστικών (identifiers) που χρησιμοποιούνται για την αναφορά οντοτήτων, η αμφισημία της αναφοράς, είναι μια ιστορική πραγματικότητα η οποία πρέπει να περιγραφεί παρά να επιλυθεί με τεχνητό τρόπο εκ των προτέρων. Το CIDOC CRM (ISO 21127) διακρίνει κόμβους που αντιστοιχούν σε υπαρκτές οντότητες και σε κόμβους που αναπαριστούν ονόματα / αναγνωριστικά.
- Οι κατηγορίες και τα ταξινομικά συστήματα δεν αποτελούν μόνο μέσα περιγραφής της πραγματικότητας αλλά είναι τεκμήρια μια ιστορικής πραγματικότητας ανθρώπινης

επινόησης. Παρομοίως, η τεκμηρίωση περιγράφεται όπως ακριβώς και το υπό τεκμηρίωση περιεχόμενο.

- Ένας χαρακτηριστικός τρόπος να αναλυθεί και να περιγραφεί το παρελθόν είναι ο διαχωρισμός του σε διακριτά γεγονότα. Το υπό τεκμηρίωση παρελθόν μπορεί να περιγραφεί σαν γεγονότα που εμπλέκουν οντότητες (υλικές και φυσικές οντότητες, όπως άνθρωποι, ζώα ή πράγματα ή εννοιολογικές οντότητες όπως ιδέες, έννοιες, προϊόντα της φαντασίας ή κοινά ονόματα). Οι υλικές ή άυλες οντότητες είναι παρόντες στα γεγονότα είτε μέσω του φυσικού τους φορέα είτε σαν έννοιες.

2 Υπάρχουσα κατάσταση

Κάθε αποθετήριο έχει το δικό του σχήμα ονομασίας για συγγραφείς και δημοσιεύσεις - η επικάλυψη μεταξύ κάθε αποθετηρίου είναι σημαντική. Προκειμένου να υπάρξει ένας σημασιολογικός Ιστός που παρέχει την εξελιξιμότητα για να αντιμετωπίσει τέτοιες ασυνέπειες πρέπει να αντιμετωπιστεί το ζήτημα του coreference. Στις βάσεις δεδομένων το πρόβλημα του coreference ορίζεται ως σύνδεσμος καταλόγων. Η ανάγκη για το σύνδεσμο καταλόγων προκύπτει όταν πρέπει να ενωθούν ή να συγχωνευθούν οι κατάλογοι ή τα αρχεία από διαφορετικές βάσεις δεδομένων. Κάθε βάση δεδομένων θα μπορούσε να έχει τα διπλά αρχεία του ίδιου προσώπου ή του πράγματος που, σε περίπτωση συγχωνεύσης, θα καθιστούσαν τα δεδομένα μη συμβατά.

Οι λύσεις ποικίλουν, από μαθηματικές αναλύσεις και αλγορίθμους που ψάχνουν equivalences, έως εκκαθαρίσεις δεδομένων (data cleaning), όμως το πρόβλημα παραμένει. Η επικάλυψη και ο διπλασιασμός της πληροφορίας είναι ένα φαινόμενο συχνό μετά από ενοποίηση βάσεων δεδομένων και αυτό που συνήθως επιχειρείται είναι να αφαιρεθούν αυτές οι διπλοεγγραφές που αφορούν το ίδιο πράγμα. Οι data cleaning μεθοδολογίες συγκρίνουν τις ιδιότητες των πραγμάτων σε διαφορετικές πηγές και εκτιμούν την πιθανότητα να εννοούν το ίδιο πράγμα. Δυστυχώς διαφορετικές πηγές συνήθως αποδίδουν διαφορετικές ιδιότητες για τα ίδια πράγματα, καθιστώντας τη μέθοδο αναξιόπιστη. Επίσης, τέτοιες μεθοδολογίες προϋποθέτουν προϋπάρχουσα κοινή γνώση σχετική με τις αξίες των ιδιοτήτων που συγκρίνονται, γνώση που δύσκολα αποκτάται.

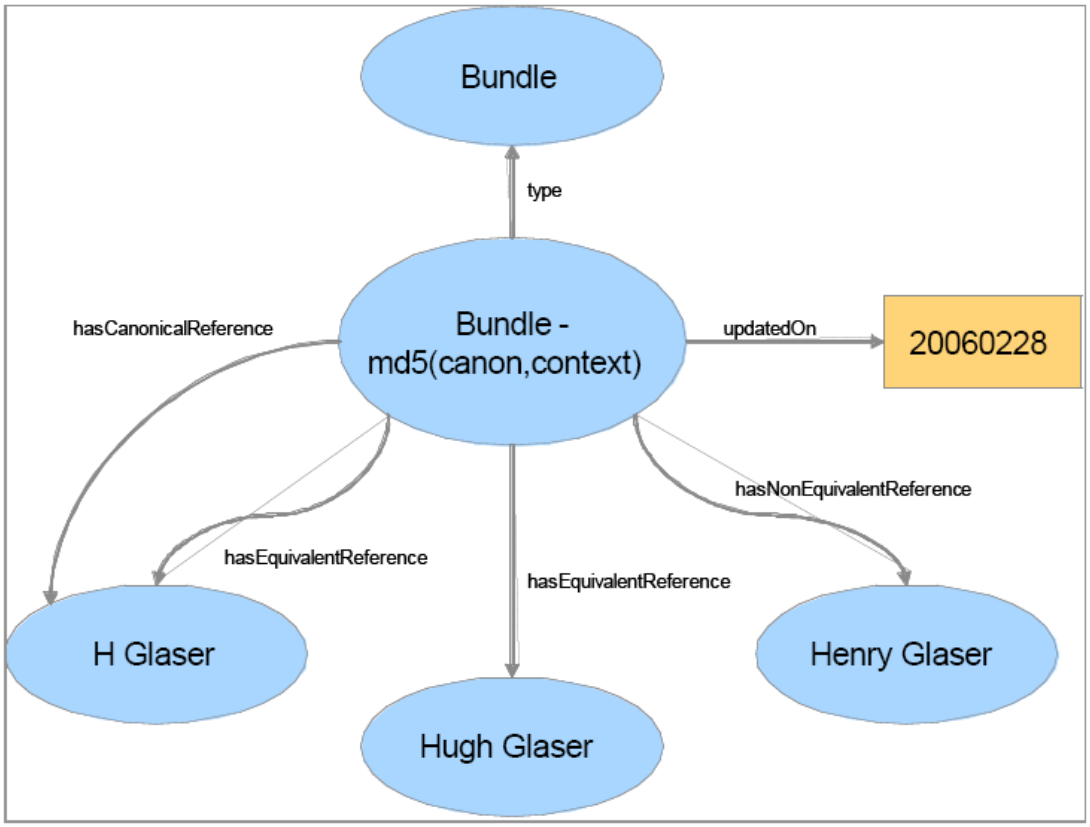
Αλλοι πάλι σπαταλούν χρόνο στο να φτιάχνουν authority files/λίστες ονομάτων με σκοπό να τα χρησιμοποιούν ως κωδικούς τοπικά Αυτές οι μέθοδοι είναι ανακριβείς και μη συστηματικές και δεν αποδίδουν ιδιότητες στις οντότητες με σκοπό να βοηθήσουν τελικά στην ταυτοποίησή/ εύρεση της ταυτότητάς τους.

Όλες αυτές οι μέθοδοι προσφέρουν μονομερείς λύσεις. Αντίθετα χρειάζεται ένα θεωρητικό πλαίσιο που αντιμετωπίζει τη σημασία του coreference με έναν ενιαίο τρόπο και κατά συνέπεια η υλοποίηση αυτού.

Η τυπική μέθοδος διαχείρισης της πληθώρας URIs είναι η χρήση της *owl:sameAs* που τα συνδέει. Η σημασιολογία του owl:sameAs εννοεί ότι όλα τα URIs που συνδέονται με αυτήν την ιδιότητα έχουν την ίδια ταυτότητα , πράγμα που σημαίνει ότι το θέμα (subject) και το αντικείμενο (object) θα πρέπει να είναι ο ίδιος πόρος. Το κύριο μειονέκτημα σε αυτό είναι ότι τα δύο URIs γίνονται αδιάκριτα ακόμη κι όταν αναφέρονται σε διαφορετικές οντότητες σύμφωνα με τα συμφραζόμενα στα οποία χρησιμοποιούνται. Επίσης μία λάθος ισοδυναμία μπορεί εν συνεχεία να προκαλέσει άλλες λάθος ισοδυναμίες που θα αναφέρονται.

Το CRS (Consistent Reference Service) δημιουργήθηκε για να διαχειριστεί το πρόβλημα του co reference ανάμεσα σε εκατομμύρια URIs. Για να το κάνει αυτό χρησιμοποιεί την έννοια *bundle* που ομαδοποιεί τα URIs που αναφέρονται στον ίδιο πόρο.

Το CRS μπορεί να χρησιμοποιεί διαφορετικούς αλγόριθμους για να αναγνωρίσει ισοδύναμους πόρους. Κάθε URI σε ένα αποθετήριο έχει το δικό του bundle στο CRS. Όταν βρεθεί η ισοδυναμία τα bundles ενώνονται και δημιουργούν ένα νέο bundle. Με αυτόν τον μηχανισμό το CRS λειτουργεί σαν βάση γνώσης για ένα συγκεκριμένο URI. Ουσιαστικά το CRS στα πλαίσια του bundle (δομές μεταδεδομένων που περιλαβάνουν ένα σύνολο αναφορών που θεωρητικά αναφέρονται ή δεν αναφέρονται στον ίδιο πόρο υπό συγκεκριμένες συνθήκες) επιτρέπει στους χρήστες να συγκρίνουν χειροκίνητα τα έγγραφα που αναφέρονται στον ίδιο πόρο/πληροφορία. Ομως η περιγραφή αυτής της πληροφορίας είναι περιορισμένη και συσχετίζει τα URLS των εγγράφων με αυτήν – δεν καθορίζει μέρη των εγγράφων που αναφέρονται στον συγκεκριμένο πόρο και δημιουργεί σύγχυση/παρερμηνείες ως προς την ταυτότητά του.



Εικόνα 2: Παραδειγμα γράφου με χρήση bundle

3 Μοντέλο Δημιουργίας Υποθέσεων Ταυτότητας – Αναφοράς

Στα πλαίσια μίας αναζήτησης λύσης/μοντέλου, ορίζουμε «co-reference» τον σχολιασμό μίας συγκεκριμένης γνώσης για μία αναφερόμενη οντότητα σε διαφορετικά έγγραφα, που ο χρήστης εντοπίζει και αναγνωρίζει ως όμοιο ή διαφορετικό ή πιθανόν όμοιο με.

Η πρόκληση είναι να διακρίνει κανείς τις διαφοροποιήσεις στη γνώση που εκφράζεται από διαφορετικούς χρήστες που περιγράφουν και ταυτίζουν οντότητες – επίσης, εξίσου σημαντικό είναι ενα πληροφοριακό σύστημα να υποστηρίζει την συνεργασία και ενοποίηση των διαφορετικών απόψεων και περιγραφών χρηστών.

Η ενοποίηση δεδομένων κάτω από ένα κοινό σχήμα, γίνεται μέσω ενός δικτύου γνώσης για το οποίο απαιτούνται:

1) εκτίμηση ποιο από τα δεδομένα που προέρχονται από διαφορετικές πηγές αναφέρεται στο ίδιο πράγμα- αυτή είναι η σχέση «co-reference»

2) διατήρηση της σχέσης «co-reference» και διάθεσή της σε διαφορετικές πηγές στο δίκτυο. Η αναγνώριση της σχέσης «co-reference» είναι η πιο σημαντική λειτουργία για την ολοκλήρωση της πληροφορίας.

Η προτεινομένη μεθοδολογία στοχεύει στην αντιμετώπιση τυπικών καταστάσεων που εμφανίζονται σε εργασιακές ομάδες που συντηρούν τεράστιους καταλόγους δεδομένων με δισεκατομμύρια κωδικών να αποδίδονται σε ανθρώπους, πράγματα, μέρη κ.λ.π.

Στις περιπτώσεις αυτές απαιτείται ένα μοντέλο με απόψεις δραστών σχετικά με το αν μιλάνε για το ίδιο αντικείμενο ή όχι. Επίσης είναι αναγκαίο το μοντέλο να καθιστά δυνατή τη διάκριση της διαφορετικής γνώσης που διαθέτουν οι δράστες όταν περιγράφουν με διαφορετικό τρόπο τα πράγματα, καθώς επίσης να εντοπίζει και τα βοηθητικά γνωρίσματα ταυτοποίησης αυτών. Τα πράγματα αυτά μπορεί να είναι αντικείμενα, τόποι, πρόσωπα, συμβάντα ή τα πάντα στα οποία αναφερόμαστε. Όλα αυτά τα πράγματα μπορεί να αντιστοιχηθούν στη γενική οντότητα E1 CRM Entity του CIDOC – CRM.

Η κύρια έννοια επομένως στο μοντέλο είναι η οντότητα στην οποία αναφερόμαστε. Αυτή η οποία συγκεντρώνει κάποια χαρακτηριστικά, ώστε να μπορούμε να την εντοπίσουμε και να την αναγνωρίσουμε, και γενικά είναι η πιθανή οντότητα για την οποία κάποιος μπορεί να εκφράζει γνώση, ερμηνεία κ.λ.π., επομένως είναι πιθανόν να αναφερόμαστε σε ο,τιδήποτε, όπως: E1 CRM Entity: Thing , Place, Person, Event του CIDOC - CRM. Τα αναγνωριστικά γνωρίσματα για την κάθε οντότητα είναι διαφορετικά ανάλογα με το είδος της, δηλ. π.χ άλλα είναι τα γνωρίσματα που χρησιμοποιούνται συνδυαστικά για την αναγνώριση/ταύτιση ενός ανθρώπου και άλλα αυτά που χρησιμοποιούνται π.χ για την αναγνώριση ενός τόπου.

Για τον **άνθρωπο (CIDOC- CRM: E21 Person),** συγκεκριμένα, οι απαραίτητες πληροφορίες για την ταύτισή του είναι:

- Πολλά ή εναλλακτικά ονόματα που φέρει
- Πλήθος κωδικών που του έχουν αποδώσει authority files
- Ημερομηνία γέννησης
- Ημερομηνία θανάτου
- Ρόλος/θέση που κατέχει
- Χρονικό διάστημα έναρξης-λήξης αυτής του της θέσης
- Σχέσεις με άλλους ανθρώπους

Για τον **τόπο (CIDOC- CRM: E53 Place),** ένας συνδυασμός γνωρισμάτων που χρησιμοποιείται για την αναγνώρισή του, θα μπορούσε να περιλαμβάνει:

- Πολλές ή εναλλακτικές ονομασίες
- Πλήθος κωδικών που του έχουν αποδώσει authority files
- Γεωγραφικό μήκος-γεωγραφικό πλάτος
- Γεωπολιτική διαίρεση
- Χρονικό διάστημα ισχύος της γεωπολιτικής διαίρεσης

Για το **γεγονός (CIDOC- CRM: E5 Event),** τα χαρακτηριστικά εκείνα που μπορούν να χρησιμοποιηθούν για ταύτιση είναι αρκετά δύσκολο να οριστούν, μια και γενικά είναι δύσκολο μία τέτοια εννοια να προσδιοριστεί μοναδικά από κάποια γνωρίσματα. Προτείνουμε έναν συνδυασμό ιδιοτήτων , όπως:

- Πολλές ή εναλλακτικές ονομασίες
- Πλήθος κωδικών που του έχουν αποδώσει authority files
- Χρόνος συμβάντος
- Τόπος συμβάντος

Για το **αντικείμενο (CIDOC- CRM: E70 Thing),** ένας συνδυασμός χαρακτηριστικών που συμβάλλουν στην αναγνώρισή του, είναι:

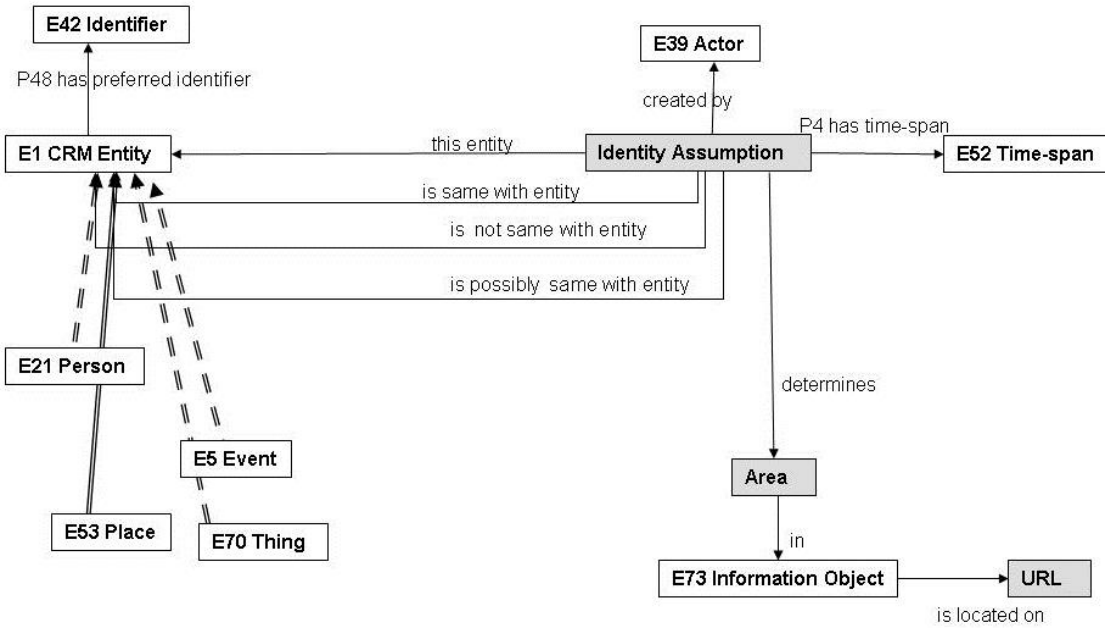
- Πολλές ή εναλλακτικές ονομασίες, εφόσον έχει
- Μία ταξινόμησή του/ το είδος του ή ένας χαρακτηρισμός που του αποδίδεται
- Πλήθος κωδικών που του έχουν αποδώσει τα authority files
- Διαστάσεις
- Περίοδο δημιουργίας/κατασκευής
- Δημιουργός

Για τις παραπάνω οντότητες ο χρήστης κάνει υποθέσεις ταυτότητας και με αυτόν τον τρόπο αναφέρεται σε αυτές: συγκεκριμένα:

3.1 Οι Υποθέσεις Ταυτότητας

Οι υποθέσεις ταυτότητας (Identity Assumption) είναι διαδικασίες κατά τις οποίες ο χρήστης εκφράζει εκτιμήσεις για οντότητες αναφερόμενες που συγκρίνει με σκοπό να προσδιορίσει την ταυτότητά τους, βάσει της προυπάρχουσας γνώσης. Οι υποθέσεις ταυτότητας εκφράζουν τα παρακάτω είδη ισοδυναμίας:

1) Όμοιες (ομοιότητα δύο αναφερόμενων οντοτήτων) (is same with entity)
2) Μη όμοιες (διαφορά δύο οντοτήτων) (is not same with entity)
3) Πιθανόν όμοιες (σε περίπτωση που ο χρήστης δεν είναι σίγουρος σχετικά με την ταυτότητά τους) (is possibly same with entity)



Εικόνα 3: προτεινόμενο σχήμα co-reference

Ο χρήστης μπορεί να εφαρμόσει μόνο ένα είδος από αυτές τις υποθέσεις ταυτότητας για δύο συγκεκριμένες οντότητες. Επίσης, όταν ο χρήστης έχει εκτιμήσει στη βάση γνώσης ότι δύο αναφερόμενες οντότητες είναι ίδιες, τότε αυτές οι δύο οντότητες συνενώνονται και εκλαμβάνονται ως μία. Αντίστοιχα με τη διαδικασία της συνένωσης, υπάρχει και η διαδικασία της διάκρισης για την αντίθετη περίπτωση.

Κατά την υπόθεση υπάρχει ένας χρήστης που περιγράφει και προσπαθεί να ταυτίσει μία συγκεκριμένη οντότητα βάσει της συγκεκριμένης γνώσης που έχει για αυτήν. Παράλληλα σε συγκεκριμένες περιοχές πληροφοριακών κειμένων προσδιορίζει και εντοπιζει αυτήν την οντότητα.

Δηλαδή, μία οντότητα μπορεί να έχει πολλές αναφορές σε διαφορετικά έγγραφα – μία λύση για τον προσδιορισμό της αναφοράς είναι να περιγραφεί η περιοχή στο έγγραφο (E73 Information Object) όπου αναφέρεται (Area).

Η εκτίμηση της ταυτότητας είναι μία σχετική διαδικασία και καταγράφεται ως προσδιοριζόμενη από τη γνώση κάθε χρήστη – δεδομένου αυτού, λοιπόν, δεν αποδίδουμε απόλυτες ομοιότητες/συγκρίσεις μεταξύ πραγμάτων.

Το συγκεκριμένο μοντέλο αναπαριστά τον εντοπισμό και ταύτιση αναφερόμενων οντοτήτων σε διαφορετικά έγγραφα και καθιστά δυνατή την περιγραφή τους βάσει της γνώσης που διαθέτουν οι χρήστες.

4 Αναφορές

Crofts, N., Doerr, M., Gill, T., Stead, S., Stiff, M. (2005): Definition of the CIDOC Conceptual Reference Model, ISO

Doerr, M. (2009), “Ontologies for Cultural Heritage”, in: Second Edition of: Steffen Staab, Rudi Studer (eds): Handbook on Ontologies, Springer

Glaser,H., Jaffri,A. and Millard,I. (2009) Managing Co-reference on the Semantic Web. In, *WWW2009 Workshop: Linked Data on the Web (LDOW2009), Madrid, Spain*

Jaffri,A., Glaser, H. and Millard, I. (2007) URI Identity Management for Semantic Web Data Integration and Linkage. In, *3rd International Workshop On Scalable Semantic Web Knowledge Base Systems, Vilamoura, Algarve, Portugal, 25 - 30 Nov 2007.* Springer.

Meghini C., Spyratos, N., Doerr, M., (2007). Sharing co-reference knowledge for data integration, *Second DELOS Conference on Digital Libraries. Working Notes (Tirrenia, Pisa, Italy, 5-7 Dicembre 2007). Proceedings, DELOS Network of Excellence in Digital Libraries, 2007.*

Meghini, C., Doerr, M., & Spyratos, N. (2009). Managing Co-reference Knowledge for Data Integration*./Proceeding of the 2009 conference on Information Modelling and Knowledge Bases XX/, (pp. 224-244). Amsterdam, The Netherlands, The Netherlands: IOS Press (978-1-58603-957-8),*

Melessanakis, V. (2011). Design and development of a platform for the management and collaborative identification of co-reference on digital resources.

5 ΠΑΡΑΡΤΗΜΑ

Στο κεφάλαιο περιγράφονται οι έννοιες και ιδιότητες του CIDOC CRM, που χρησιμοποιούνται, καθώς και νέες έννοιες και ιδιότητες που προτείνονται ως επέκταση αυτού.

Οι περιγραφές των κλάσεων και ιδιοτήτων αποδίδονται στα αγγλικά για τους παρακάτω λόγους:

(α) η τελευταία έκδοση του CIDOC CRM δεν έχει μεταφραστεί στα ελληνικά. Η μετάφραση στα ελληνικά δεν εμπίπτει στις εργασίες αυτού του παραδοτέου.

(β) οι προτεινόμενες προσθέσεις και διαφοροποιήσεις αποδίδονται επίσης στα αγγλικά για ομοιομορφία και για να συζητηθούν στις συναντήσεις της ειδικής ομάδας εργασίας του CIDOC CRM.

Για την παρουσίαση της οντολογίας υιοθετείται ο τρόπος περιγραφής των κλάσεων και ιδιοτήτων που χρησιμοποιεί το CIDOC CRM. Συγκεκριμένα:

Για τις κλάσεις δηλώνονται:

- Τα ονόματα των κλάσεων παρουσιάζονται σαν τίτλοι σε έντονη μορφή (bold)
- Η γραμμή “Subclass of:” δηλώνει την υπερκλάση της κατηγορίας από την οποία κληρονομεί τις ιδιότητες
- Η γραμμή “Superclass of:”είναι μια παραπομπή στις υποκατηγορίες αυτής της κατηγορίας;
- Η γραμμή “Scope note:” περιέχει τον κειμενικό ορισμό της έννοιας που η κατηγορία αντιπροσωπεύει
- Η γραμμή “Examples:” περιέχει κατάλογο παραδειγμάτων των περιπτώσεων αυτής της κατηγορίας
- Η γραμμή “Properties:”δηλώνει τον κατάλογο των ιδιοτήτων της κατηγορίας

Για τις ιδιότητες δηλώνονται:

Τα ονόματα των ιδιοτήτων παρουσιάζονται σαν τίτλοι σε έντονη μορφή (bold)

- Η γραμμή “Domain:” δηλώνει την κλάση για την οποία η ιδιότητα ορίζεται
- Η γραμμή “Range:” δηλώνει την κλάση στην οποία δείχνει η ιδιότητα ή αποδίδει αξία της ιδιότητας
- Η γραμμή “Subpropertyof:” είναι μια παραπομπή στις ιδιότητες στην οποία είναι υπο-ιδιότητα
- Η γραμμή “Superpropertyof:”είναι μια παραπομπή στις υποιδιότητες αυτής τηςιδιότητας

- Η γραμμή “Scopenote:” περιέχει τον κειμενικό ορισμό της έννοιας που η ιδιότητα αντιπροσωπεύει
- Η γραμμή “Examples:” περιέχει κατάλογο παραδειγμάτων των περιπτώσεων αυτής της ιδιότητας

5.1 Νέες κλάσεις

Identity Assumption

Subclass of: E7 Activity

Scope note: This class comprises a process that an Actor determines an assumption about the identity of two named entities.

Examples:

Properties:

this entity: E1 CRM Entity
is same with entity: E1 CRM Entity
is not same with entity: E1 CRM Entity
is possibly same with: E1 CRM Entity
determines: Area
created by: E39 Actor

Area

Subclass of: E73 Information Object

Scope note: This class comprises a part of the document that refers to the entity. It is the cited area.

Examples:

Properties:

in: E73 Information Object

URL

Subclass of:

Scope note: This class specifies where an identified resource is available in the web and it is used as a location for an information object.

Examples:

Properties:

5.2 Νέες ιδιότητες

this entity

Domain: Identity Assumption

Range: E1 CRM Entity

Subproperty of:

Quantification:

Scope note: This property declares that this entity is referred and associated with another entity.

Examples:

is same with entity

Domain: Identity Assumption

Range: E1 CRM Entity

Subproperty of:

Quantification:

Scope note: This property declares that this entity is identified as same with another entity. It is an identity type.

Examples:

is not same with entity

Domain: Identity Assumption

Range: E1 CRM Entity

Subproperty of:

Quantification:

Scope note: This property declares that this entity is identified as not same with another entity. It is an identity type.

Examples:

is possibly same with entity

Domain: Identity Assumption

Range: E1 CRM Entity

Subproperty of:

Quantification:

Scope note: This property declares that this entity is identified as possibly same with another entity. It is an identity type.

Examples:

determines

Domain: Identity Assumption

Range: Area

Subproperty of:

Quantification:

Scope note: This property identifies the specific part of the document (area) that is determined by an identity assumption process.

Examples:

in

Domain: Area

Range: E73 Information Object

Subproperty of:

Quantification:

Scope note: This property specifies that an area is included as a part in an information object.

Examples:

is located on

Domain: E73 Information Object

Range: URL

Subproperty of:

Quantification:

Scope note: This property records that an information object has a URL as a specific location used to identify it.

Examples:

created by

Domain: Identity Assumption

Range: E39 Actor

Subproperty of: E7 Activity: P14 carried out by (performed): E39 Actor

Quantification:

Scope note: This property records the active participation of an actor in an identity creation activity.

Examples:

5.3 Οι αναφερόμενες κλάσεις και ιδιότητες του CIDOC CRM

Μια και το προτεινόμενο μοντέλο χρησιμοποιεί τμήματα από το CIDOC – CRM ver. 6.0, η παρακάτω ενότητα περιλαμβάνει μία κατανοητή λίστα όλων των δομών που χρησιμοποιούνται μαζί με τους ορισμούς τους. Ο ολοκληρωμένος ορισμός του εννοιολογικού μοντέλου CIDOC CRM βρίσκεται στον επίσημο δικτυακό τόπο: http://www.cidoc-crm.org/official_release_cidoc.html.

5.3.1 **κλάσεις/έννοιες που χρησιμοποιούνται από το CIDOC CRM**

Η ενότητα που ακολουθεί περιλαμβάνει τους ορισμούς των κλάσεων του εννοιολογικού μοντέλου CIDOC CRM, που χρησιμοποιούνται στο προτεινόμενο μοντέλο. Οι ιδιότητες που περιλαμβάνονται σε αυτούς τους ορισμούς και χρησιμοποιούνται στο μοντέλο εμφανίζονται με έντονο χρώμα (bold).

E1 CRM Entity

Superclass of: E2TemporalEntity

E52 Time-Span

E53 Place

E54 Dimension

E77 Persistent Item

E92 Spacetime Volume

Scope note: This class comprises all things in the universe of discourse of the CIDOC Conceptual Reference Model.

It is an abstract concept providing for three general properties:

1. Identification by name or appellation, and in particular by a preferred identifier
2. Classification by type, allowing further refinement of the specific subclass an instance belongs to
3. Attachment of free text for the expression of anything not captured by formal properties

With the exception of E59 Primitive Value, all other classes within the CRM are directly or indirectly specialisations of E1 CRM Entity.

Examples:

- the earthquake in Lisbon 1755 (E5)

Properties:

P1 is identified by (identifies): E41 Appellation

P2 has type (is type of): E55 Type

P3 has note: E62 String

(P3.1 has type: E55 Type)

P48 has preferred identifier (is preferred identifier of): E42 Identifier

P137 exemplifies (is exemplified by): E55 Type

(P137.1 in the taxonomic role: E55 Type)

E5 Event

Subclass of: E4 Period

Superclass of: E7 Activity

E63 Beginning of Existence

E64 End of Existence

Scope note: This class comprises changes of states in cultural, social or physical systems, regardless of scale, brought about by a series or group of coherent physical, cultural, technological or legal phenomena. Such changes of state will affect instances of E77 Persistent Item or its subclasses.

The distinction between an E5 Event and an E4 Period is partly a question of the scale of observation. Viewed at a coarse level of detail, an E5 Event is an 'instantaneous' change of state. At a fine level, the E5 Event can be analysed into its component phenomena within a space and time frame, and as such can be seen as an E4 Period. The reverse is not necessarily the case: not all instances of E4 Period give rise to a noteworthy change of state.

Examples:

- the birth of Cleopatra (E67)
- the destruction of Herculaneum by volcanic eruption in 79 AD (E6)
- World War II (E7)
- the Battle of Stalingrad (E7)
- the Yalta Conference (E7)
- my birthday celebration 28-6-1995 (E7)
- the falling of a tile from my roof last Sunday
- the CIDOC Conference 2003 (E7)

Properties:

P11 had participant (participated in): E39 Actor

P12 occurred in the presence of (was present at): E77 Persistent Item

E7 Activity

Subclass of: E5 Event

Superclass of: E8 Acquisition

E9 Move

E10 Transfer of Custody

E11 Modification

E13 Attribute Assignment

E65 Creation

E66 Formation

E85 Joining

E86 Leaving

E87 Curation Activity

Scope note: This class comprises actions intentionally carried out by instances of E39 Actor that result in changes of state in the cultural, social, or physical systems documented.

This notion includes complex, composite and long-lasting actions such as the building of a settlement or a war, as well as simple, short-lived actions such as the opening of a door.

Examples:

- the Battle of Stalingrad
- the Yalta Conference
- my birthday celebration 28-6-1995
- the writing of "Faust" by Goethe (E65)
- the formation of the Bauhaus 1919 (E66)
- calling the place identified by TGN '7017998' 'Quyunjig' by the people of Iraq
- Kira Weber working in glass art from 1984 to 1993
- Kira Weber working in oil and pastel painting from 1993

Properties:

P14 carried out by (performed): E39 Actor

(P14.1 in the role of: E55 Type)

P15 was influenced by (influenced): E1 CRM Entity

P16 used specific object (was used for): E70 Thing

(P16.1 mode of use: E55 Type)

P17 was motivated by (motivated): E1 CRM Entity

P19 was intended use of (was made for): E71 Man-Made Thing

(P19.1 mode of use: E55 Type)

P20 had specific purpose (was purpose of): E5 Event

P21 had general purpose (was purpose of): E55 Type

P32 used general technique (was technique of): E55 Type

P33 used specific technique (was used by): E29 Design or Procedure

P125 used object of type (was type of object used in): E55 Type

P134 continued (was continued by): E7 Activity

E21 Person

Subclass of: E20 Biological Object

E39 Actor

Scope note: This class comprises real persons who live or are assumed to have lived.

Legendary figures that may have existed, such as Ulysses and King Arthur, fall into this class if the documentation refers to them as historical figures. In cases where doubt exists as to whether several persons are in fact identical, multiple instances can be created and linked to indicate their relationship. The CRM does not propose a specific form to support reasoning about possible identity.

Examples:

- Tut-Ankh-Amun
- Nelson Mandela

Properties:

P152 has parent (is parent of): E21 Person

E39 Actor

Subclass of: E77 Persistent Item

Superclass of: **E21 Person**

E74 Group

Scope note: This class comprises people, either individually or in groups, who have the potential to perform intentional actions of kinds for which someone may be held responsible.

The CRM does not attempt to model the inadvertent actions of such actors. Individual people should be documented as instances of E21 Person, whereas groups should be documented as instances of either E74 Group or its subclass E40 Legal Body.

Examples:

- London and Continental Railways (E40)
- the Governor of the Bank of England in 1975 (E21)
- Sir Ian McKellan (E21)

Properties:

P74 has current or former residence (is current or former residence of): E53 Place

P75 possesses (is possessed by): E30 Right

P76 has contact point (provides access to): E51 Contact Point

P131 is identified by (identifies): E82 Actor Appellation

E42 Identifier

Subclass of: E41 Appellation

Scope note: This class comprises strings or codes assigned to instances of E1 CRM Entity in order to identify them uniquely and permanently within the context of one or more organisations. Such codes are often known as inventory numbers, registration codes, etc. and are typically composed of alphanumeric sequences. The class E42 Identifier is not normally used for machine-generated identifiers used for automated processing unless these are also used by human agents.

Examples:

- "MM.GE.195"
- "13.45.1976"
- "OXCMS: 1997.4.1"

- ISSN "0041-5278"
- ISRC "FIFIN8900116"
- Shelf mark "Res 8 P 10"
- "Guillaume de Machaut (1300?-1377)" [a controlled personal name heading that follows the French rules]

E52 Time-Span

Subclass of: E1 CRM Entity

Scope note: This class comprises abstract temporal extents, in the sense of Galilean physics, having a beginning, an end and a duration.

Time Span has no other semantic connotations. Time-Spans are used to define the temporal extent of instances of E4 Period, E5 Event and any other phenomena valid for a certain time. An E52 Time-Span may be identified by one or more instances of E49 Time Appellation.

Since our knowledge of history is imperfect, instances of E52 Time-Span can best be considered as approximations of the actual Time-Spans of temporal entities. The properties of E52 Time-Span are intended to allow these approximations to be expressed precisely. An extreme case of approximation, might, for example, define an E52 Time-Span having unknown beginning, end and duration. Used as a common E52 Time-Span for two events, it would nevertheless define them as being simultaneous, even if nothing else was known.

Automatic processing and querying of instances of E52 Time-Span is facilitated if data can be parsed into an E61 Time Primitive.

Examples:

- 1961
- From 12-17-1993 to 12-8-1996
- 14h30 – 16h22 4th July 1945
- 9.30 am 1.1.1999 to 2.00 pm 1.1.1999
- duration of the Ming Dynasty

Properties:

P78 is identified by (identifies): E49 Time Appellation

P79 beginning is qualified by: E62 String

P80 end is qualified by: E62 String

P81 ongoing throughout: E61 Time Primitive

P82 at some time within: E61 Time Primitive

P83 had at least duration (was minimum duration of): E54 Dimension

P84 had at most duration (was maximum duration of): E54 Dimension

P86 falls within (contains): E52 Time-Span

E53 Place

Subclass of: E1 CRM Entity

Scope note: This class comprises extents in space, in particular on the surface of the earth, in the pure sense of physics: independent from temporal phenomena and matter.

The instances of E53 Place are usually determined by reference to the position of "immobile" objects such as buildings, cities, mountains, rivers, or dedicated geodetic marks. A Place can be determined by combining a frame of reference and a location with respect to this frame. It may be identified by one or more instances of E44 Place Appellation.

It is sometimes argued that instances of E53 Place are best identified by global coordinates or absolute reference systems. However, relative references are often more relevant in the context of cultural documentation and tend to be more precise. In particular, we are often interested in position in relation to large, mobile objects, such as ships. For example, the Place at which Nelson died is known with reference to a large mobile object – H.M.S Victory. A resolution of this Place in terms of absolute coordinates would require knowledge of the movements of the vessel and the precise time of death, either of which may be revised, and the result would lack historical and cultural relevance.

Any object can serve as a frame of reference for E53 Place determination. The model foresees the notion of a "section" of an E19 Physical Object as a valid E53 Place determination.

Examples:

- the extent of the UK in the year 2003
- the position of the hallmark on the inside of my wedding ring
- the place referred to in the phrase: "Fish collected at three miles north of the confluence of the Arve and the Rhone"
- here -> <-

Properties:

P87 is identified by (identifies): E44 Place Appellation

P89 falls within (contains): E53 Place

P121 overlaps with: E53 Place

P122 borders with: E53 Place

P157 is at rest relative to (provides reference space for): E18 Physical Thing

E70 Thing

Subclass of: E77 Persistent Item

Superclass of: E71 Man-Made Thing

E72 Legal Object

Scope note: This general class comprises discrete, identifiable, instances of E77 Persistent Item that are documented as single units, that either consist of matter or depend on being carried by matter and are characterized by relative stability.

They may be intellectual products or physical things. They may for instance have a solid physical form, an electronic encoding, or they may be a logical concept or structure.

Examples:

- my photograph collection (E78)
- the bottle of milk in my refrigerator (E22)
- the plan of the Strassburger Muenster (E29)
- the thing on the top of Otto Hahn's desk (E19)
- the form of the no-smoking sign (E36)
- the cave of Dirou, Mani, Greece (E27)

Properties

P43 has dimension (is dimension of): E54 Dimension

P101 had as general use (was use of): E55 Type

P130 shows features of (features are also found on): E70 Thing

(P130.1 kind of similarity: E55 Type)

E73 Information Object

Subclass of: E89 Propositional Object

E90 Symbolic Object

Superclass of: E29 Design or Procedure

E31 Document

E33 Linguistic Object

E36 Visual Item

Scope note: This class comprises identifiable immaterial items, such as a poems, jokes, data sets, images, texts, multimedia objects, procedural prescriptions, computer program code, algorithm or mathematical formulae, that have an objectively recognizable structure and are documented as single units.

An E73 Information Object does not depend on a specific physical carrier, which can include human memory, and it can exist on one or more carriers simultaneously.

Instances of E73 Information Object of a linguistic nature should be declared as instances of the E33 Linguistic Object subclass. Instances of E73 Information Object of a documentary nature should be declared as instances of the E31 Document subclass. Conceptual items such as types and classes are not instances of E73 Information Object, nor are ideas without a reproducible expression.

Examples:

- image BM000038850.JPG from the Clayton Herbarium in London
- E. A. Poe's "The Raven"
- the movie "The Seven Samurai" by Akira Kurosawa
- the Maxwell Equations
- The Getty AAT as published as Linked Open Data, accessed 1/10/2014

5.3.2 Ιδιότητες που χρησιμοποιούνται από το CIDOC CRM

Η ενότητα που ακολουθεί περιλαμβάνει τους πλήρεις ορισμούς των ιδιοτήτων του εννοιολογικού μοντέλου CIDOC CRM, που χρησιμοποιούμε στο μοντέλο που προτείνουμε.

P4 has time-span (is time-span of)

Domain: E2 Temporal Entity

Range: E52 Time-Span

Quantification:many to one, necessary, dependent (1,1:1,n)

Scope note: This property describes the temporal confinement of an instance of an E2 Temporal Entity.

The related E52 Time-Span is understood as the real Time-Span during which the phenomena were active, which make up the temporal entity instance. It does not convey any other meaning than a positioning on the “time-line” of chronology. The Time-Span in turn is approximated by a set of dates (E61 Time Primitive). A temporal entity can have in reality only one Time-Span, but there may exist alternative opinions about it, which we would express by assigning multiple Time-Spans. Related temporal entities may share a Time-Span. Time-Spans may have completely unknown dates but other descriptions by which we can infer knowledge.

Examples:

the Yalta Conference (E7) *has time-span* Yalta Conference time-span (E52)

P48 has preferred identifier (is preferred identifier of)

Domain: E1 CRM Entity

Range: E42 Identifier

Subproperty of: E1 CRM Entity.P1 is identified by (identifies): E41 Appellation

Quantification:many to one (0,1:0,n)

Scope note: This property records the preferred E42 Identifier that was used to identify an instance of E1 CRM Entity at the time this property was recorded.

More than one preferred identifier may have been assigned to an item over time.

Use of this property requires an external mechanism for assigning temporal validity to the respective CRM instance.

P48 has preferred identifier (is preferred identifier of), is a shortcut for the path from E1 CRM Entity through *P140 assigned attribute to (was attributed by)*, E15 Identifier Assignment, *P37 assigned (was assigned by)* to E42 Identifier. The fact that an identifier is a preferred one for an organisation can be better expressed in a context independent form by assigning a suitable E55 Type to the respective instance of E15 Identifier Assignment using the *P2 has type* property.

Examples:

- the pair of Lederhosen donated by Dr Martin Doerr (E22) *has preferred identifier* “OXCMS:2001.1.32” (E42)