



ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ
ΗΛΕΚΤΡΟΝΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ
«Σύνδεση μιας μουσειακής συλλογής με τα Wikidata»

Του φοιτητή:
Έκκαρου Παπαδόπουλου
Αρ. Μητρώου: 2021206

Επιβλέπων
Χαράλαμπος Μπράτσας
Επίκουρος Καθηγητής

28 Αυγούστου, 2026

Τίτλος Δ.Ε.: Σύνδεση μιας μουσειακής συλλογής με τα Wikidata

Κωδικός Δ.Ε. 26163

Ονοματεπώνυμο φοιτητή Ίκαρος Παπαδόπουλος

Ονοματεπώνυμο εισηγητή Χαράλαμπος Μπράτσας

Ημερομηνία ανάληψης Δ.Ε. 11-3-2026

Ημερομηνία περάτωσης Δ.Ε. 25-8-2026

Βεβαιώνω ότι είμαι ο συγγραφέας αυτής της εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, έχω καταγράψει τις όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών, εικόνων και κειμένου, είτε αυτές αναφέρονται ακριβώς είτε παραφρασμένες. Επιπλέον, βεβαιώνω ότι αυτή η εργασία προετοιμάστηκε από εμένα προσωπικά, ειδικά ως διπλωματική εργασία, στο Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του ΔΙ.ΠΑ.Ε.

Η παρούσα εργασία αποτελεί πνευματική ιδιοκτησία του φοιτητή Ίκαρου Παπαδόπουλου που την εκπόνησε. Στο πλαίσιο της πολιτικής ανοικτής πρόσβασης, ο συγγραφέας/δημιουργός εκχωρεί στο Διεθνές Πανεπιστήμιο της Ελλάδος άδεια χρήσης του δικαιώματος αναπαραγωγής, δανεισμού, παρουσίασης στο κοινό και ψηφιακής διάχυσης της εργασίας διεθνώς, σε ηλεκτρονική μορφή και σε οποιοδήποτε μέσο, για διδακτικούς και ερευνητικούς σκοπούς, άνευ ανταλλάγματος. Η ανοικτή πρόσβαση στο πλήρες κείμενο της εργασίας δεν σημαίνει καθ' οιονδήποτε τρόπο παραχώρηση δικαιωμάτων διανοητικής ιδιοκτησίας του συγγραφέα/δημιουργού, ούτε επιτρέπει την αναπαραγωγή, αναδημοσίευση, αντιγραφή, πώληση, εμπορική χρήση, διανομή, έκδοση, μεταφόρτωση, ανάρτηση, μετάφραση ή τροποποίηση με οποιονδήποτε τρόπο, τμηματικά ή περιληπτικά, χωρίς τη ρητή προηγούμενη έγγραφη συναίνεση του συγγραφέα/δημιουργού.

Η έγκριση της διπλωματικής εργασίας από το Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του Διεθνούς Πανεπιστημίου της Ελλάδος δεν υποδηλώνει απαραίτητα και αποδοχή των απόψεων του συγγραφέα εκ μέρους του Τμήματος.

Στην οικογένεια και τους φίλους μου

Πρόλογος

Η παρούσα διπλωματική εργασία εκπονήθηκε στο πλαίσιο του προγράμματος σπουδών του Τμήματος Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του Διεθνούς Πανεπιστημίου της Ελλάδος. Αντικείμενό της είναι η διασύνδεση μιας μουσειακής ψηφιακής συλλογής με τα Wikidata, θέμα που εντάσσεται στο πεδίο των Linked Open Data και της ψηφιακής πολιτιστικής κληρονομιάς.

Το δείγμα που μελετήθηκε προέρχεται από το τμήμα Ελληνικής και Ρωμαϊκής Τέχνης του Metropolitan Museum of Art και αποτελείται από θραύσματα αττικών αγγείων της αρχαϊκής και πρώιμης κλασικής περιόδου. Η επιλογή αυτή προσδίδει στην εργασία διπλό χαρακτήρα: είναι ταυτόχρονα τεχνική μελέτη διασύνδεσης δεδομένων και συμβολή στην ψηφιακή ορατότητα της ελληνικής πολιτιστικής κληρονομιάς.

Από τις πρώτες κιόλας διαλέξεις του μαθήματος του Σημασιολογικού Ιστού, απέκτησα ενδιαφέρον ώστε να μελετήσω και να προσπαθήσω να κατανοήσω τον τρόπο με τον οποίο τα δεδομένα κατηγοριοποιούνται στον παγκόσμιο ιστό, πως από απλές λέξεις αποκτούν σημασία και οντολογικό ενδιαφέρον. Η παρούσα διπλωματική εργασία με βοήθησε σε αυτήν μου την αναζήτηση και η προσθήκη στοιχείων που αφορούν τον αρχαίο πολιτισμό, κατάφερε να προσδώσει μία πιο ευχάριστη νότα.

Περίληψη

Η ψηφιοποίηση των πολιτιστικών συλλογών έχει δημιουργήσει σημαντικούς όγκους δομημένων δεδομένων, οι οποίοι όμως παραμένουν συχνά απομονωμένοι και δεν αξιοποιούν τις δυνατότητες διασύνδεσης του οικοσυστήματος Linked Open Data. Η παρούσα εργασία αναπτύσσει και εφαρμόζει μια συστηματική, αναπαραγώγιμη μεθοδολογία για τη διασύνδεση μιας μουσειακής ψηφιακής συλλογής με τα Wikidata. Ως αντικείμενο μελέτης επιλέχθηκε δείγμα 501 αντικειμένων από το τμήμα Ελληνικής και Ρωμαϊκής Τέχνης του Metropolitan Museum of Art, το οποίο αποτελείται κατά 99,2% από θραύσματα αττικών αγγείων.

Η μεθοδολογία οργανώνεται σε τέσσερα στάδια: επιλογή και προετοιμασία του δείγματος μέσω OpenRefine, συμφιλίωση δημιουργών και υλικών με τα αντίστοιχα Wikidata items, εμπλουτισμό με ιδιότητες που δεν υπήρχαν στην πηγή, και μοντελοποίηση σχήματος RDF με εξαγωγή εντολών QuickStatements. Κάθε δήλωση τεκμηριώνεται με αναφορά πηγής P248 προς το Metropolitan Museum of Art.

Η αυτόματη συμφιλίωση απέδωσε 175 αντιστοιχίσεις δημιουργών, ποσοστό 34,9%. Στοχευμένη χειροκίνητη επιμέλεια τριών μόνο ονομάτων, επιλεγμένων με κριτήριο τη συχνότητα, ανέβασε το σύνολο σε 389, ποσοστό 77,6%. Η συμφιλίωση των υλικών κάλυψε 499 εγγραφές και ο εμπλουτισμός απέδωσε 291 τιμές εθνικότητας.

Η αξιοποίηση της στήλης Artist Wikidata URL του ίδιου του dataset ως χρυσού προτύπου επέτρεψε αντικειμενική αξιολόγηση, σπάνια διαθέσιμη σε αντίστοιχες εργασίες. Οι μετρικές ήταν: ακρίβεια 100%, ανάκληση 89,7%, F1-score 94,6%. Η τέλεια ακρίβεια, παρά τις 214 χειροκίνητες παρεμβάσεις, επικυρώνει τα κριτήρια απόφασης που εφαρμόστηκαν.

Η εργασία τεκμηριώνει τρία ευρήματα με ευρύτερη σημασία. Οι περιορισμοί τύπου στη συμφιλίωση αποδείχθηκαν προβληματικοί σε δύο ανεξάρτητες περιπτώσεις. Η αποτυχία αντιστοίχισης του πολυπληθέστερου δημιουργού οφειλόταν σε ψευδώς αρνητικό και όχι σε απουσία δεδομένων. Και η ιδιότητα P27 αποδείχθηκε σημασιολογικά ακατάλληλη για πρόσωπα της αρχαιότητας.

Λέξεις-κλειδιά: Linked Open Data, Wikidata, μουσειακές συλλογές, μεταδεδομένα, OpenRefine, Metropolitan Museum of Art, αττική αγγειογραφία, ελληνική αρχαιότητα, QuickStatements, GLAM

«Linking a Museum Collection with Wikidata»

«Ikaros Papadopoulos»

Abstract

The digitisation of cultural collections has produced significant volumes of structured data that often remain isolated and fail to exploit the connectivity offered by the Linked Open Data ecosystem. This thesis develops and applies a systematic, reproducible methodology for linking a museum digital collection with Wikidata. The case study is a sample of 501 objects from the Greek and Roman Art department of the Metropolitan Museum of Art, consisting almost entirely of Attic vase fragments.

The methodology comprises four stages: sample selection and preparation in OpenRefine, reconciliation of creators and materials against Wikidata items, enrichment with properties absent from the source, and RDF schema modelling with export of QuickStatements commands. Every statement carries a P248 source reference to the Metropolitan Museum of Art.

Automatic reconciliation matched 175 creator records, or 34.9%. Targeted manual curation of only three names, selected by frequency, raised the total to 389, or 77.6%. Material reconciliation covered 499 records and enrichment yielded 291 citizenship values.

Using the dataset's own Artist Wikidata URL column as a gold standard enabled objective evaluation, rarely available in comparable studies. The metrics were: precision 100%, recall 89.7%, F1-score 94.6%. Perfect precision despite 214 manual interventions validates the decision criteria applied.

The thesis documents three findings of broader significance. Type constraints in reconciliation proved problematic in two independent cases. The failure to match the most prolific creator was a false negative rather than a data gap. And the P27 property proved semantically unsuitable for figures of antiquity.

Keywords: Linked Open Data, Wikidata, museum collections, metadata, OpenRefine, Metropolitan Museum of Art, Attic vase painting, Greek antiquity, QuickStatements, GLAM

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον επιβλέπων καθηγητή μου Δρ. Χαράλαμπο Μπράτσα για την καθοδήγηση που έλαβα κατά τη διάρκεια της συγγραφής, καθώς και τα μέλη της ομάδας του για τις εύστοχες επισημάνσεις και την ευρύτερη βοήθεια.

Ευρετήριο Περιεχομένων

Πρόλογος.....	4
Περίληψη	5
Abstract	7
Ευχαριστίες	8
Ευρετήριο Περιεχομένων.....	9
Ευρετήριο Σχημάτων	14
Ευρετήριο Πινάκων.....	15
Εισαγωγή.....	16
Το πλαίσιο της εργασίας	16
Το αντικείμενο μελέτης.....	16
Σκοπός και στόχοι.....	16
Μεθοδολογική προσέγγιση	17
Παραδοτέα	17
Δομή της εργασίας	17
Κεφάλαιο 1ο: Θεωρητικό Πλαίσιο	19
1.1 Εισαγωγή.....	19
1.2 Semantic Web και Linked Open Data.....	19
1.2.1 Η ιδέα του Semantic Web: από τον Ιστό εγγράφων στον Ιστό δεδομένων	19
1.2.2 RDF: το θεμέλιο της αναπαράστασης γνώσης	20
1.2.3 Οντολογίες και γλώσσες σχήματος: RDFS και OWL	20
1.2.4 Το μοντέλο 5 αστεριών και οι αρχές LOD	21
1.2.5 SPARQL: η γλώσσα ερωτημάτων του Semantic Web	21
1.2.6 Το οικοσύστημα LOD και τα GLAM ιδρύματα	22
1.3 Wikidata: Δομή, Λειτουργία και Εργαλεία.....	22
1.3.1 Ιστορικό και γενική περιγραφή.....	22
1.3.2 Βασική δομή δεδομένων	23
1.3.3 Qualifiers, References και η αρχή της επαληθευσιμότητας.....	24
1.3.4 Εργαλεία του οικοσυστήματος Wikidata	24
1.4 Πρότυπα Μεταδομένων για Ψηφιακές Πολιτιστικές Συλλογές.....	25
1.4.1 Η σημασία των μεταδεδομένων στον χώρο των GLAM.....	25
1.4.2 Dublin Core: απλότητα και ευρεία υιοθέτηση	25
1.4.3 CIDOC-CRM: πλούσια σημασιολογική αναπαράσταση	26
1.4.4 Άλλα σχετικά πρότυπα.....	27

1.5 Βιβλιογραφική Ανασκόπηση	27
1.5.1 Η αναδυόμενη σημασία των Wikidata για τα GLAM ιδρύματα.....	27
1.5.2 Σουηδικά μουσεία: μια τεκμηριωμένη εμπειρία	28
1.5.3 Το Metropolitan Museum of Art και τα ανοιχτά δεδομένα	28
1.5.4 Αξιολόγηση ανοιχτών πρακτικών μουσείων τέχνης	28
1.5.5 Ιταλικά GLAM ιδρύματα: μεγάλης κλίμακας διασύνδεση	29
1.5.6 Θέση της παρούσας εργασίας στη βιβλιογραφία	29
Κεφάλαιο 2ο: Η Ψηφιακή Συλλογή του Metropolitan Museum of Art.....	30
2.1 Εισαγωγή.....	30
2.2 Το Metropolitan Museum of Art.....	30
2.2.1 Ιστορικό και συλλογή.....	30
2.2.2 Η Open Access Initiative	30
2.3 Δομή του Dataset	31
2.3.1 Γενική επισκόπηση	31
2.3.2 Η κατανομή ανά τμήμα	31
2.3.3 Το κρίσιμο πεδίο Artist Wikidata URL	32
2.4 Επιλογή του Δείγματος	33
2.4.1 Το σκεπτικό της θεματικής εστίασης.....	33
2.4.2 Το εύρημα της ανωνυμίας.....	33
2.4.3 Η διαδικασία φιλτραρίσματος.....	34
2.5 Χαρακτηριστικά του Δείγματος.....	35
2.5.1 Σύνθεση και τυπολογία	35
2.5.2 Η ασυμμετρία στην κατανομή των δημιουργών	36
2.5.3 Η συλλογή von Bothmer	36
2.6 Η Αττική Αγγειογραφία ως Αντικείμενο Τεκμηρίωσης	37
2.6.1 Η σπανιότητα των υπογραφών.....	37
2.6.2 Η παράδοση Beazley και τα συμβατικά ονόματα	37
2.6.3 Οντολογικά ερωτήματα για τη διασύνδεση	37
2.7 Το Mapping προς τα Wikidata.....	38
Κεφάλαιο 3ο: Μεθοδολογία Διασύνδεσης με τα Wikidata	39
3.1 Εισαγωγή.....	39
3.1.1 Το περιβάλλον εργασίας	40
3.2 Στάδιο 1: Επιλογή και Προετοιμασία του Δείγματος	41
3.2.1 Εισαγωγή του dataset	41
3.2.2 Πρώτο φίλτρο: θεματική επιλογή	42

3.2.3 Δεύτερο φίλτρο: αποκλεισμός ανώνυμων έργων.....	43
3.2.4 Μετατροπή του Object ID σε αριθμητικό τύπο	44
3.2.5 Τρίτο φίλτρο: αριθμητικός περιορισμός	44
3.3 Στάδιο 2: Συμφιλίωση Οντοτήτων	45
3.3.1 Η έννοια και ο μηχανισμός	45
3.3.2 Αυτόματη συμφιλίωση των δημιουργών	46
3.3.3 Στρατηγική χειροκίνητης επιμέλειας	47
3.3.4 Η εκτέλεση της χειροκίνητης επιμέλειας	48
3.3.5 Οι περιπτώσεις όπου η αντιστοίχιση δεν έγινε	49
3.3.6 Το τελικό αποτέλεσμα.....	49
3.3.7 Δημιουργία της στήλης Artist_Wikidata_ID	50
3.4 Στάδιο 3: Συμφιλίωση Υλικών και Εμπλουτισμός	50
3.4.1 Η αποτυχία της πρώτης απόπειρας	50
3.4.2 Η δεύτερη απόπειρα	51
3.4.3 Μαζική αντιστοίχιση των υλικών	51
3.4.4 Ανάκτηση της εθνικότητας	52
3.5 Στάδιο 4: Μοντελοποίηση Σχήματος και Εξαγωγή	53
3.5.1 Η αρχιτεκτονική του Wikibase Schema.....	53
3.5.2 Ο υπό συνθήκη ορισμός του υποκειμένου	53
3.5.3 Οι δηλώσεις, οι ιδιότητες ταυτοποίησης και οι αναφορές	53
3.5.4 Ο υπό συνθήκη ορισμός της ιδιότητας P31	55
3.5.5 Η εξαγωγή και το σφάλμα της έκδοσης.....	56
3.5.6 Η δομή του παραγόμενου αρχείου	56
3.6 Δεοντολογικές Παρατηρήσεις.....	57
3.7 Η Τεκμηρίωση της Αναπαραγωγιμότητας.....	57
Κεφάλαιο 4ο: Αποτελέσματα και Αξιολόγηση.....	59
4.1 Εισαγωγή.....	59
4.2 Τα Παραδοτέα.....	59
4.2.1 Επισκόπηση.....	59
4.2.2 Το εμπλουτισμένο dataset.....	59
4.2.3 Το αρχείο μεθοδολογίας.....	60
4.2.4 Το αρχείο QuickStatements και η επαλήθευσή του	60
4.3 Αξιολόγηση της Συμφιλίωσης των Δημιουργών	61
4.3.1 Η τριμερής ανάλυση της αυτόματης φάσης.....	61
4.3.2 Η επίδραση της χειροκίνητης επιμέλειας.....	61

4.3.3 Το φαινόμενο Makron.....	62
4.3.4 Οι τεκμηριωμένες αρνήσεις.....	62
4.4 Αντικειμενική Αξιολόγηση έναντι Χρυσού Προτύπου	63
4.4.1 Το θεσμικό σύνολο αναφοράς	63
4.4.2 Οι μετρικές.....	63
4.4.3 Ερμηνεία των αποτελεσμάτων.....	64
4.4.4 Ανάκληση σε επίπεδο ονομάτων	64
4.4.5 Τα όρια του συνόλου αναφοράς.....	65
4.4.6 Τα δεκατέσσερα επιπλέον ευρήματα	65
4.5 Αξιολόγηση της Συμφιλίωσης των Υλικών.....	65
4.6 Η Επίδραση των Περιορισμών Τύπου: Ελεγχόμενο Πείραμα.....	66
4.6.1 Το ερευνητικό ερώτημα	66
4.6.2 Ο σχεδιασμός	66
4.6.3 Τα αποτελέσματα	67
4.6.4 Ερμηνεία	67
4.6.5 Συμπεράσματα του πειράματος.....	67
4.7 Αξιολόγηση του Εμπλουτισμού.....	68
4.7.1 Ποσοτικά αποτελέσματα.....	68
4.7.2 Η σημασιολογική ανεπάρκεια της ιδιότητας P27	68
4.8 Αξιολόγηση της Διαλειτουργικότητας.....	69
4.8.1 Η μετάβαση στο μοντέλο 5 αστεριών.....	69
4.8.2 Πρακτικές συνέπειες	70
4.9 Συνολική Αποτίμηση - Συζήτηση	70
Κεφάλαιο 5ο: Συμπεράσματα και Προτάσεις για Μελλοντική Έρευνα.....	71
5.1 Εισαγωγή.....	71
5.2 Απαντήσεις στα Ερευνητικά Ερωτήματα	71
5.2.1 Είναι εφικτή η διασύνδεση με ανοιχτά εργαλεία;.....	71
5.2.2 Ποια εμπόδια εμφανίζονται και πώς αντιμετωπίζονται;.....	71
5.2.3 Ποια είναι η αξία της διασύνδεσης;.....	72
5.3 Συστηματοποίηση των Συμπερασμάτων.....	72
5.3.1 Μεθοδολογικά συμπεράσματα.....	72
5.3.2 Εμπειρικά συμπεράσματα	72
5.3.3 Συμπεράσματα για τις αρχαιολογικές συλλογές.....	73
5.3.4 Το εύρος ισχύος των ευρημάτων	73
5.3.5 Η αναγκαιότητα της χειροκίνητης επιμέλειας	74

5.3.6 Συμβολή στη βιβλιογραφία.....	74
5.4 Περιορισμοί της Εργασίας.....	75
5.4.1 Περιορισμοί δείγματος.....	75
5.4.2 Περιορισμοί του mapping.....	75
5.4.3 Περιορισμοί της αξιολόγησης.....	75
5.4.4 Ο περιορισμός της μη εκτέλεσης.....	75
5.5 Προτάσεις για Μελλοντική Έρευνα.....	76
5.5.1 Βελτιστοποίηση των παραμέτρων συμφιλίωσης.....	76
5.5.2 Επέκταση του mapping.....	76
5.5.3 Εφαρμογή σε ελληνικές μουσειακές συλλογές.....	76
5.5.4 Ανάπτυξη τυποποιημένης διαδικασίας αξιολόγησης.....	77
5.5.5 Αξιοποίηση μεγάλων γλωσσικών μοντέλων.....	77
5.5.6 Μελέτη της αξιοσημείωτης καταχώρισης αρχαιολογικού υλικού.....	78
5.5.7 Εκπαιδευτικές εφαρμογές.....	78
5.6 Τελικά Σχόλια.....	78
Αναφορές.....	79
Βιβλιογραφία.....	81
Semantic Web και Linked Open Data.....	81
Wikidata.....	81
Πρότυπα μεταδεδομένων.....	81
GLAM και ψηφιακές συλλογές.....	81
Αρχαία ελληνική κεραμική.....	82
Εργαλεία και τεχνική τεκμηρίωση.....	82
Παράρτημα Α: Το Αρχείο Μεθοδολογίας.....	83
A.1 Επισκόπηση των κύριων ενεργειών.....	83
A.2 Ενδεικτικό απόσπασμα.....	84
Παράρτημα Β: Δείγμα του Αρχείου QuickStatements.....	85
B.1 Δημιουργία νέων items.....	85
B.2 Εμπλουτισμός υπάρχοντος item.....	85
Παράρτημα Γ: Πίνακας Αντιστοιχισμένων Δημιουργών.....	87

Ευρετήριο Σχημάτων

Σχήμα 2.1 Τα δέκα μεγαλύτερα τμήματα της συλλογής.....	32
Σχήμα 2.2 Η κυριαρχία της ανωνυμίας στην αρχαιολογική συλλογή	34
Σχήμα 3.1 Η συνολική ροή εργασιών της μεθοδολογίας.....	40
Σχήμα 3.2 Η οθόνη εισαγωγής CSV στο OpenRefine.....	42
Σχήμα 3.3 Το Text Facet με τα δεκαεννέα τμήματα.....	42
Σχήμα 3.4 Το αποτέλεσμα του πρώτου φίλτρου.....	43
Σχήμα 3.5 Το δεύτερο φίλτρο με invert.....	43
Σχήμα 3.6 Το τρίτο φίλτρο - Custom Text Facet.....	45
Σχήμα 3.7 Η ρύθμιση συμφιλίωσης με τύπο Human (Q5)	46
Σχήμα 3.8 Τα facets judgment και best candidate's score.....	47
Σχήμα 3.9 Η κατανομή των μη αντιστοιχισμένων ονομάτων.....	48
Σχήμα 3.10 Η διεπαφή Bulk Match	49
Σχήμα 3.11 Η αποτυχία της πρώτης απόπειρας - 499 errors	50
Σχήμα 3.12 Η δεύτερη απόπειρα χωρίς περιορισμό τύπου.....	51
Σχήμα 3.13 Η ανάκτηση της ιδιότητας P27	52
Σχήμα 3.14 Το Wikibase Schema	54
Σχήμα 3.15 Η προεπισκόπηση των δηλώσεων	55
Σχήμα 3.16 Δείγμα του αρχείου QuickStatements	57
Σχήμα 3.17 Το ιστορικό των ενεργειών στην καρτέλα Undo/Redo	58
Σχήμα 4.1 Ανάλυση αποτελεσμάτων αυτόματης συμφιλίωσης.....	61
Σχήμα 4.2 Η επίδραση της χειροκίνητης επιμέλειας	62
Σχήμα 4.3 Η ομοιογένεια των υλικών.....	66
Σχήμα 4.4 Οι ανακτηθείσες τιμές εθνικότητας.....	68

Ευρετήριο Πινάκων

Πίνακας 1.1 Το μοντέλο των 5 αστεριών Linked Open Data	21
Πίνακας 1.2 Wikidata properties που χρησιμοποιούνται στην εργασία	23
Πίνακας 1.3 Σύγκριση Dublin Core, CIDOC-CRM και Wikidata properties	26
Πίνακας 2.1 Η κατανομή των εγγραφών στα δεκαεννέα τμήματα	31
Πίνακας 2.2 Η απόδοση δημιουργού στο τμήμα Greek and Roman Art.....	33
Πίνακας 2.3 Η διαδοχική εφαρμογή των τριών φίλτρων.....	34
Πίνακας 2.4 Βασικά χαρακτηριστικά του δείγματος.....	35
Πίνακας 2.5 Οι συχνότεροι τύποι αντικειμένων	35
Πίνακας 2.6 Οι δώδεκα δημιουργοί με τα περισσότερα έργα	36
Πίνακας 2.7 Το mapping πεδίων dataset προς Wikidata properties	38
Πίνακας 3.1 Οι κύριες ενέργειες της μεθοδολογίας.....	39
Πίνακας 3.2 Η διαδοχική εφαρμογή των τριών φίλτρων.....	45
Πίνακας 3.3 Τριμερής ανάλυση των αποτελεσμάτων	46
Πίνακας 3.4 Τα δέκα συχνότερα μη αντιστοιχισμένα ονόματα.....	47
Πίνακας 3.5 Οι τρεις ενέργειες χειροκίνητης επιμέλειας.....	48
Πίνακας 3.6 Η επίδραση της χειροκίνητης επιμέλειας	50
Πίνακας 3.7 Αποτελέσματα συμφιλίωσης της στήλης Medium.....	51
Πίνακας 3.8 Οι τιμές της ιδιότητας P27.....	52
Πίνακας 4.1 Τα τρία παραδοτέα της εργασίας.....	59
Πίνακας 4.2 Πληρότητα των νέων στηλών.....	59
Πίνακας 4.3 Επαλήθευση της δομής του QuickStatements.....	60
Πίνακας 4.4 Τριμερής ανάλυση αυτόματης συμφιλίωσης.....	61
Πίνακας 4.5 Σύγκριση πριν και μετά τη χειροκίνητη επιμέλεια.....	61
Πίνακας 4.6 Σύγκριση με το θεσμικό σύνολο αναφοράς	63
Πίνακας 4.7 Μετρικές απόδοσης	64
Πίνακας 4.8 Εξέλιξη των μετρικών	64
Πίνακας 4.9 Κατανομή και αντιστοίχιση των υλικών	65
Πίνακας 4.10 Σύγκριση συμφιλίωσης με και χωρίς περιορισμό τύπου.....	67
Πίνακας 4.11 Οι τιμές της ιδιότητας P27.....	68
Πίνακας 4.12 Αξιολόγηση βάσει του μοντέλου 5 αστεριών.....	69
Πίνακας 5.1 Προτεινόμενες επεκτάσεις του mapping	76
Πίνακας Α.1 Οι κύριες ενέργειες με τις παραμέτρους τους.....	83
Πίνακας Γ.1 Οι 63 αντιστοιχισμένοι δημιουργοί.....	87

Εισαγωγή

Το πλαίσιο της εργασίας

Η ψηφιοποίηση των πολιτιστικών συλλογών αποτελεί μία από τις σημαντικότερες εξελίξεις στον χώρο της πολιτιστικής κληρονομιάς τις τελευταίες δεκαετίες. Μουσεία, αρχεία και βιβλιοθήκες παγκοσμίως έχουν επενδύσει σημαντικούς πόρους στη ψηφιοποίηση, δημιουργώντας τεράστιους όγκους δομημένων δεδομένων που αφορούν έργα τέχνης, δημιουργούς, ιστορικές περιόδους και γεωγραφικές θέσεις.

Παρά την ψηφιακή τους μορφή, πολλές από αυτές τις συλλογές παραμένουν ουσιαστικά απομονωμένες. Τα δεδομένα τους αποθηκεύονται σε τοπικές βάσεις με ιδιαίτερα σχήματα μεταδεδομένων, χρησιμοποιούν εσωτερικά αναγνωριστικά χωρίς νόημα εκτός του ιδρύματος, και δεν αξιοποιούν τις δυνατότητες διασύνδεσης που παρέχει το οικοσύστημα του Semantic Web. Ένας ερευνητής που μελετά έναν καλλιτέχνη πρέπει να αναζητήσει χειροκίνητα σε δεκάδες διαφορετικά συστήματα, αντί να υποβάλει ένα ενιαίο ερώτημα.

Τα Wikidata, ως ανοικτή και συνεργατική βάση γνώσης του Ιδρύματος Wikimedia, προσφέρουν λύση. Παρέχουν πρότυπο για την καταγραφή και διασύνδεση πολιτιστικών δεδομένων με διεθνώς αναγνωρισμένο τρόπο, χρησιμοποιώντας παγκόσμια μοναδικά αναγνωριστικά και ανοικτή άδεια χρήσης. Η διασύνδεση μιας μουσειακής συλλογής με τα Wikidata μετατρέπει τις τοπικές εγγραφές σε κόμβους ενός παγκόσμιου γράφου γνώσης.

Το αντικείμενο μελέτης

Ως αντικείμενο μελέτης επιλέχθηκε το ανοιχτό dataset του Metropolitan Museum of Art, το οποίο περιλαμβάνει 484.956 εγγραφές κατανεμημένες σε δεκαεννέα τμήματα και διατίθεται υπό άδεια Creative Commons Zero.

Η εργασία εστίασε περαιτέρω στο τμήμα Ελληνικής και Ρωμαϊκής Τέχνης, το οποίο περιλαμβάνει 33.726 εγγραφές. Η επιλογή αυτή προσδίδει διπλό ενδιαφέρον. Αφενός εξετάζει ελληνική πολιτιστική κληρονομιά που φυλάσσεται σε ξένο ίδρυμα, θέμα με ιδιαίτερη σημασία για τον ελληνικό ακαδημαϊκό χώρο. Αφετέρου θέτει τη μεθοδολογία αντιμετώπιση με τις ιδιαιτερότητες των αρχαιολογικών συλλογών: συμβατικά ονόματα δημιουργών, συντριπτική κυριαρχία ανώνυμων αντικειμένων, και αρχαίες πολιτικές οντότητες που δεν αντιστοιχούν σε σύγχρονα κράτη.

Η διαδοχική εφαρμογή τριών φίλτρων οδήγησε στο τελικό δείγμα των 501 εγγραφών, το οποίο αποτελείται κατά 99,2% από θραύσματα αττικών αγγείων της αρχαϊκής και πρώιμης κλασικής περιόδου, με κοινή προέλευση από τη δωρεά Dietrich von Bothmer του 2011.

Σκοπός και στόχοι

Κύριος σκοπός της εργασίας είναι η μελέτη και εφαρμογή μιας μεθοδολογίας για τη διασύνδεση μουσειακής ψηφιακής συλλογής με τα Wikidata, με στόχο τον εμπλουτισμό των μεταδεδομένων, τη βελτίωση της διαλειτουργικότητας και την ενσωμάτωση της συλλογής στο οικοσύστημα Linked Open Data.

Οι επιμέρους στόχοι διατυπώνονται ως εξής:

Πρώτος στόχος. Ανάλυση των αρχών και των τεχνολογιών των Linked Open Data και διερεύνηση της δομής των Wikidata, ώστε να τεθεί το απαραίτητο θεωρητικό υπόβαθρο.

Δεύτερος στόχος. Εξέταση και χαρτογράφηση των μεταδεδομένων της συλλογής προς τις αντίστοιχες ιδιότητες των Wikidata.

Τρίτος στόχος. Ανάπτυξη και εφαρμογή αναπαραγωγίμης διαδικασίας διασύνδεσης των αντικειμένων με τα αντίστοιχα στοιχεία της βάσης γνώσης.

Τέταρτος στόχος. Αντικειμενική αξιολόγηση της ποιότητας των αποτελεσμάτων, της διαλειτουργικότητας και του εμπλουτισμού που επιτεύχθηκε.

Μεθοδολογική προσέγγιση

Η εργασία ακολουθεί συστηματική προσέγγιση πέντε φάσεων. Η πρώτη αφορά βιβλιογραφική ανασκόπηση σχετικά με ψηφιακές μουσειακές συλλογές, πρότυπα μεταδεδομένων όπως το Dublin Core και το CIDOC-CRM, και τεχνολογίες Linked Open Data. Η δεύτερη αφορά την ανάλυση της ψηφιακής συλλογής και τη δομή των μεταδεδομένων της.

Η τρίτη φάση αφορά τη χαρτογράφηση των μεταδεδομένων προς τις ιδιότητες των Wikidata, με στόχο τη σημασιολογικά ορθή αντιστοίχιση. Η τέταρτη αφορά την υλοποίηση της διασύνδεσης με τα εργαλεία OpenRefine, Wikidata Reconciliation Service και QuickStatements. Η πέμπτη αφορά την αξιολόγηση των αποτελεσμάτων ως προς την ακρίβεια, την πληρότητα και τη δυνατότητα επαναχρησιμοποίησης.

Ιδιαίτερο χαρακτηριστικό της προσέγγισης είναι η αξιοποίηση της στήλης Artist Wikidata URL, την οποία περιλαμβάνει το ίδιο το dataset, ως χρυσού προτύπου για αντικειμενική μέτρηση της ποιότητας. Η δυνατότητα αυτή, σπάνια διαθέσιμη σε αντίστοιχες εργασίες, επέτρεψε τον υπολογισμό των καθιερωμένων μετρικών precision, recall και F1-score.

Παραδοτέα

Η εργασία παράγει τρία συμπληρωματικά παραδοτέα.

Το πρώτο είναι το εμπλουτισμένο dataset σε μορφή CSV και XLSX, με 501 εγγραφές και 57 στήλες. Περιλαμβάνει τρεις νέες στήλες: τα Q-numbers των δημιουργών, τα Q-numbers των υλικών, και την ιδιότητα country of citizenship που ανακτήθηκε αυτόματα από τα Wikidata.

Το δεύτερο είναι το αρχείο μεθοδολογίας σε μορφή JSON, το οποίο περιέχει την πλήρη καταγραφή των ενεργειών που εκτελέστηκαν και διασφαλίζει την αναπαραγωγιμότητα της έρευνας. Παρατίθεται στο Παράρτημα Α.

Το τρίτο είναι το αρχείο εντολών QuickStatements σε σύνταξη V1, με 3.871 γραμμές, έτοιμο για εισαγωγή στα Wikidata. Δείγμα του παρατίθεται στο Παράρτημα Β.

Δομή της εργασίας

Η εργασία οργανώνεται σε πέντε κεφάλαια.

Το πρώτο κεφάλαιο θέτει το θεωρητικό πλαίσιο. Αναλύει τις αρχές του Semantic Web και των Linked Open Data, συμπεριλαμβανομένου του μοντέλου RDF, της γλώσσας SPARQL και του μοντέλου 5 αστεριών. Παρουσιάζει τη δομή και τα εργαλεία των Wikidata, εξετάζει τα πρότυπα Dublin Core και CIDOC-CRM, και ολοκληρώνεται με βιβλιογραφική ανασκόπηση διεθνών εμπειριών.

Το δεύτερο κεφάλαιο παρουσιάζει τη μουσειακή συλλογή. Περιγράφει το Metropolitan Museum of Art και την Open Access Initiative, αναλύει τη δομή του dataset, τεκμηριώνει την επιλογή του δείγματος, και εξετάζει την αττική αγγειογραφία ως αντικείμενο τεκμηρίωσης με έμφαση στην παράδοση Beazley.

Το τρίτο κεφάλαιο αποτελεί τον πυρήνα της εργασίας και παρουσιάζει βήμα προς βήμα την τετρασταδιακή μεθοδολογία: επιλογή δείγματος, συμφιλίωση οντοτήτων με αυτόματα και χειροκίνητη φάση, εμπλουτισμό, και μοντελοποίηση σχήματος με εξαγωγή εντολών.

Το τέταρτο κεφάλαιο παρουσιάζει και αξιολογεί τα αποτελέσματα. Περιλαμβάνει την επαλήθευση των παραδοτέων, την τριμερή ανάλυση της συμφιλίωσης, την αντικειμενική μέτρηση έναντι χρυσού προτύπου, και την αποτίμηση της διαλειτουργικότητας.

Το πέμπτο κεφάλαιο συνοψίζει τα συμπεράσματα, απαντά στα ερευνητικά ερωτήματα, συζητά κριτικά τους περιορισμούς, και παρουσιάζει επτά προτάσεις για μελλοντική έρευνα με έμφαση στην εφαρμογή σε ελληνικές μουσειακές συλλογές.

Η εργασία ολοκληρώνεται με τις ενότητες Αναφορών και Βιβλιογραφίας και με τρία παραρτήματα που περιλαμβάνουν το αρχείο μεθοδολογίας, δείγμα των εντολών QuickStatements και τον πλήρη πίνακα των αντιστοιχισμένων δημιουργών.

Κεφάλαιο 1ο: Θεωρητικό Πλαίσιο

1.1 Εισαγωγή

Το παρόν κεφάλαιο θέτει το θεωρητικό υπόβαθρο πάνω στο οποίο στηρίζεται η παρούσα εργασία. Η διασύνδεση ψηφιακών πολιτιστικών συλλογών με ανοιχτές βάσεις γνώσης δεν αποτελεί απλώς μια τεχνική διαδικασία, αλλά εντάσσεται σε ένα ευρύτερο οικοσύστημα αρχών, τεχνολογιών και προτύπων που διαμορφώνουν τον σύγχρονο ψηφιακό πολιτισμό. Η κατανόηση αυτού του οικοσυστήματος απαιτεί εξοικείωση με τέσσερις αλληλένδετες θεματικές ενότητες: τις αρχές του Semantic Web και των Linked Open Data, τη δομή και τη λειτουργία των Wikidata ως κεντρικού κόμβου γνώσης, τα διεθνή πρότυπα μεταδεδομένων για πολιτιστικές συλλογές, και τη σχετική ερευνητική βιβλιογραφία που τεκμηριώνει τη διεθνή τάση διασύνδεσης GLAM ιδρυμάτων με τα Wikidata.

Η δομή του κεφαλαίου ακολουθεί μια λογική πρόοδο από το γενικό στο ειδικό: ξεκινά από τις θεμελιώδεις έννοιες του Semantic Web που αποτελούν το τεχνολογικό υπόβαθρο κάθε LOD εφαρμογής, προχωρεί στη λεπτομερή παρουσίαση των Wikidata ως συγκεκριμένης πλατφόρμας υλοποίησης, εξετάζει τα πρότυπα μεταδεδομένων που χρησιμοποιούνται στον χώρο των μουσείων, και ολοκληρώνεται με βιβλιογραφική ανασκόπηση που αναδεικνύει τη θέση της παρούσας εργασίας στο διεθνές ερευνητικό πεδίο. Κάθε ενότητα οικοδομεί πάνω στην προηγούμενη, δημιουργώντας ένα συνεκτικό πλαίσιο που θα αξιοποιηθεί άμεσα στα επόμενα κεφάλαια.

Η επιλογή των συγκεκριμένων θεωρητικών περιοχών δεν είναι τυχαία. Το Semantic Web και τα LOD αποτελούν τη γλώσσα με την οποία εκφράζεται η διασύνδεση· τα Wikidata είναι η πλατφόρμα στην οποία πραγματοποιείται· τα πρότυπα μεταδεδομένων είναι η γέφυρα που συνδέει τα δεδομένα της συλλογής με τις απαιτήσεις της πλατφόρμας· και η βιβλιογραφική ανασκόπηση παρέχει τις εμπειρίες άλλων που έχουν ακολουθήσει παρόμοια μονοπάτια. Μαζί, αυτά τα τέσσερα στοιχεία αποτελούν τον θεωρητικό εξοπλισμό που χρειάζεται ο αναγνώστης για να κατανοήσει πλήρως τη μεθοδολογία και τα αποτελέσματα των επόμενων κεφαλαίων.

1.2 Semantic Web και Linked Open Data

1.2.1 Η ιδέα του *Semantic Web*: από τον *Ιστό εγγράφων* στον *Ιστό δεδομένων*

Ο Παγκόσμιος Ιστός (World Wide Web), από τη δημιουργία του στις αρχές της δεκαετίας του 1990, σχεδιάστηκε ως ένα σύστημα διασύνδεσης εγγράφων μέσω υπερσυνδέσμων. Τα HTML έγγραφα που αποτελούν τον ιστό είναι κατεξοχήν κατανοητά από ανθρώπους: ένας χρήστης μπορεί να τα διαβάσει, να τα ερμηνεύσει και να εξαγάγει από αυτά χρήσιμες πληροφορίες. Αντίθετα, μια μηχανή - ένας υπολογιστής ή ένα πρόγραμμα - μπορεί να «δει» το κείμενο και τους υπερσυνδέσμους, αλλά δεν μπορεί να κατανοήσει τη σημασία τους. Δεν μπορεί να ξέρει, για παράδειγμα, ότι η λέξη «Αθήνα» σε ένα κείμενο αναφέρεται στην πρωτεύουσα της Ελλάδας, ή ότι ο «Παρθενώνας» είναι κτίσμα που βρίσκεται σε αυτήν [1].

Αυτό το θεμελιώδες πρόβλημα - η απουσία μηχαναγνώσιμης σημασίας στον Ιστό - ήταν αυτό που κινητοποίησε τον Tim Berners-Lee να οραματιστεί το Semantic Web. Σε ένα άρθρο-σταθμό που δημοσιεύθηκε στο Scientific American το 2001 μαζί με τους James Hendler και Ora Lassila, πρότεινε μια επέκταση του Ιστού στην οποία η πληροφορία δεν θα ήταν απλώς αναγνώσιμη από ανθρώπους, αλλά θα είχε καλά ορισμένη σημασία που θα μπορούσαν να επεξεργαστούν και οι μηχανές. Το όραμα

αυτό περιγράφηκε ως ένας «ιστός δεδομένων» (Web of Data) παράλληλος με τον υπάρχοντα «ιστό εγγράφων» (Web of Documents) [1].

Η πρακτική υλοποίηση αυτού του οράματος απαιτεί δύο βασικά συστατικά. Πρώτον, ένα τυποποιημένο τρόπο αναπαράστασης πληροφορίας που να είναι κατανοητός από μηχανές - αυτός είναι ο ρόλος του RDF (Resource Description Framework). Δεύτερον, έναν μηχανισμό μοναδικής ταυτοποίησης κάθε οντότητας στον Ιστό - αυτός είναι ο ρόλος των URI (Uniform Resource Identifiers). Μαζί, αυτά τα δύο συστατικά επιτρέπουν τη δημιουργία ενός δικτύου δεδομένων όπου κάθε πληροφορία είναι τυποποιημένη, μοναδικά αναγνωρίσιμη και διασυνδεδεμένη με σχετικές πληροφορίες [2].

1.2.2 RDF: το θεμέλιο της αναπαράστασης γνώσης

Το Resource Description Framework (RDF) αποτελεί το θεμελιώδες μοντέλο δεδομένων του Semantic Web. Αναπτύχθηκε από την κοινοπραξία W3C (World Wide Web Consortium) και αποτελεί επίσημη W3C Recommendation από το 1999, με σημαντικές αναθεωρήσεις το 2004 (RDF 1.0) και το 2014 (RDF 1.1). Η κεντρική ιδέα του RDF είναι απλή αλλά ισχυρή: κάθε πληροφορία εκφράζεται ως μια τριπλέτα (triple) της μορφής υποκείμενο - κατηγορημα - αντικείμενο (subject - predicate - object) [2].

Αυτή η δομή τριπλέτας μοιάζει με μια απλή πρόταση στη φυσική γλώσσα. Η πρόταση «Ο Βαν Γκογκ γεννήθηκε στη Ζούντερτ» μεταφράζεται στην τριπλέτα (Vincent_van_Gogh, placeOfBirth, Zundert). Η πρόταση «Ο Βαν Γκογκ ζωγράφισε τους Ηλιάνθους» μεταφράζεται στην τριπλέτα (Vincent_van_Gogh, created, Sunflowers). Η πρόταση «Οι Ηλιάνθοι βρίσκονται στην Εθνική Πινακοθήκη του Λονδίνου» μεταφράζεται στην τριπλέτα (Sunflowers, currentLocation, NationalGallery). Αυτές οι τριπλέτες μπορούν να συνδυαστούν για να σχηματίσουν ένα πλέγμα (γράφο) γνώσης που αναπαριστά σύνθετες σχέσεις μεταξύ οντοτήτων [2].

Κάθε στοιχείο μιας τριπλέτας αναγνωρίζεται μέσω ενός μοναδικού URI. Αυτό σημαίνει ότι δεν υπάρχει σύγχυση: το «Vincent van Gogh» ως ζωγράφος είναι διαφορετική οντότητα από το «Vincent van Gogh» ως τοπωνύμιο ή ως album μουσικής. Κάθε οντότητα έχει έναν και μόνον URI που τη χαρακτηρίζει μοναδικά σε παγκόσμιο επίπεδο. Στα Wikidata, για παράδειγμα, ο Vincent van Gogh αναγνωρίζεται από τον URI <http://www.wikidata.org/entity/Q5582>, ο οποίος είναι αποκλειστικά δικός του σε ολόκληρο τον Ιστό.

Το RDF υποστηρίζει διάφορες σειριακές μορφές (serialization formats) για την αποθήκευση και μεταφορά δεδομένων: RDF/XML (η αρχική και πιο διαδεδομένη), Turtle (πιο ευανάγνωστη από ανθρώπους), N-Triples (μια γραμμή ανά τριπλέτα), JSON-LD (συμβατή με JSON) και άλλες. Η επιλογή της μορφής εξαρτάται από τη χρήση: για αποθήκευση και επεξεργασία από μηχανές συνήθως προτιμάται το RDF/XML ή το Turtle, ενώ για ανταλλαγή δεδομένων μέσω web APIs το JSON-LD είναι πιο πρακτικό.

1.2.3 Οντολογίες και γλώσσες σχήματος: RDFS και OWL

Το RDF από μόνο του επιτρέπει την αναπαράσταση γεγονότων, αλλά δεν παρέχει τρόπο να εκφραστεί η σημασία των σχέσεων και των κατηγοριών. Για αυτόν τον σκοπό αναπτύχθηκαν δύο επιπλέον γλώσσες: το RDF Schema (RDFS) και η Web Ontology Language (OWL), επίσης W3C Recommendations.

Το RDFS επεκτείνει το RDF επιτρέποντας τον ορισμό κλάσεων (classes) και ιδιοτήτων (properties). Μια κλάση είναι μια κατηγορία οντοτήτων - για παράδειγμα η κλάση «Ζωγράφος» ή «Έργο Τέχνης». Μια ιδιότητα ορίζει μια σχέση μεταξύ οντοτήτων - για παράδειγμα η ιδιότητα «δημιούργησε» που

συνδέει έναν Ζωγράφο με ένα Έργο Τέχνης. Το RDFS επιτρέπει επίσης τον ορισμό ιεραρχιών κλάσεων (υπερκλάση - υποκλάση), που αντικατοπτρίζουν τη γνωσιολογική δομή του κόσμου.

Η OWL είναι μια πιο εκφραστική γλώσσα που επιτρέπει τον ορισμό σύνθετων λογικών σχέσεων μεταξύ κλάσεων και ιδιοτήτων. Με την OWL μπορεί κανείς να εκφράσει, για παράδειγμα, ότι δύο κλάσεις είναι αμοιβαία αποκλειστικές, ότι μια ιδιότητα είναι μεταβατική (αν $A \rightarrow B$ και $B \rightarrow C$ τότε $A \rightarrow C$), ή ότι μια ιδιότητα έχει μοναδική τιμή. Αυτές οι εκφράσεις επιτρέπουν σε λογικούς συλλογιστές (reasoners) να εξάγουν νέα συμπεράσματα από υπάρχουσες πληροφορίες - μια δυνατότητα που ανοίγει τον δρόμο για έξυπνες εφαρμογές που «σκέφτονται» με βάση γνώση.

1.2.4 Το μοντέλο 5 αστεριών και οι αρχές LOD

Τα Linked Open Data (LOD) αποτελούν την πρακτική εφαρμογή των αρχών του Semantic Web για τη δημοσίευση δεδομένων στον Ιστό. Ο Tim Berners-Lee πρότεινε ένα πενταεπίπεδο μοντέλο αξιολόγησης (5-star model) για τη μέτρηση της ωριμότητας των ανοιχτών δεδομένων [11]. Ο πίνακας που ακολουθεί παρουσιάζει αναλυτικά τα πέντε επίπεδα:

Πίνακας 1.1: Το μοντέλο 5 αστεριών Linked Open Data

Επίπεδο	Περιγραφή	Παράδειγμα	Περιορισμός
★	Ανοιχτά δεδομένα με οποιαδήποτε μορφή	Σκαναρισμένο PDF	Μη μηχαναγνώσιμο
★★	Δομημένα και μηχαναγνώσιμα	Αρχείο Excel	Ιδιοκτησιακή μορφή
★★★	Μη ιδιοκτησιακή μορφή	CSV, JSON	Χωρίς URIs
★★★★	URIs για αναγνώριση πόρων	RDF/XML	Χωρίς διασύνδεση
★★★★★	Διασύνδεση με άλλα LOD datasets	Wikidata, DBpedia	-

Το dataset του Metropolitan Museum of Art που αξιοποιείται στην παρούσα εργασία βρίσκεται αρχικά στο επίπεδο των τριών αστεριών: είναι δημόσια διαθέσιμο υπό άδεια CC0 σε μορφή CSV. Μέσω της μεθοδολογίας που αναπτύσσεται στην παρούσα εργασία, το dataset αναβαθμίζεται στο επίπεδο των πέντε αστεριών: οι οντότητές του αποκτούν URI-based αναγνωριστικά (Q-numbers) και διασυνδέονται με το παγκόσμιο οικοσύστημα LOD μέσω των Wikidata. Αυτή η μετάβαση από τρία σε πέντε αστέρια αποτελεί στην ουσία τον κεντρικό στόχο της εργασίας.

Πέρα από το μοντέλο 5 αστεριών, τα Linked Open Data διέπονται από τέσσερις βασικές αρχές που διατύπωσε ο Berners-Lee: χρήση URIs ως ονομάτων για πράγματα, χρήση HTTP URIs ώστε να μπορεί κανείς να τα αναζητήσει, παροχή χρήσιμων πληροφοριών μέσω αυτών των URIs χρησιμοποιώντας RDF/SPARQL, και σύνδεση με άλλα URIs ώστε να μπορούν οι χρήστες να ανακαλύψουν περισσότερα. Αυτές οι αρχές διασφαλίζουν ότι τα LOD δεν είναι απλώς ανοιχτά, αλλά πραγματικά διασυνδεδεμένα και αξιοποιήσιμα.

1.2.5 SPARQL: η γλώσσα ερωτημάτων του Semantic Web

Για την αναζήτηση και ανάκτηση δεδομένων από γράφους RDF χρησιμοποιείται η γλώσσα ερωτημάτων SPARQL (SPARQL Protocol and RDF Query Language). Αποτελεί W3C Recommendation από το 2008 (SPARQL 1.0) με σημαντική επέκταση το 2013 (SPARQL 1.1). Η SPARQL επιτρέπει την υποβολή πολύπλοκων ερωτημάτων σε RDF datasets, με σύνταξη που θυμίζει SQL αλλά προσαρμοσμένη στη δομή των γράφων [2].

Ένα βασικό SPARQL ερώτημα αποτελείται από τρία βασικά μέρη: το SELECT (τι θέλουμε να ανακτήσουμε), το WHERE (ποια μοτίβα τριπλέτων πρέπει να ισχύουν) και προαιρετικά πρόσθετες ρήτρες όπως FILTER (φίλτρα τιμών), ORDER BY (ταξινόμηση), LIMIT (περιορισμός αριθμού αποτελεσμάτων) και GROUP BY (ομαδοποίηση). Η SPARQL 1.1 εισήγαγε επίσης ερωτήματα τύπου ASK (επιστρέφει αληθές/ψευδές), CONSTRUCT (επιστρέφει νέο RDF γράφο), DESCRIBE (επιστρέφει πληροφορίες για μια οντότητα) και UPDATE (για τροποποίηση δεδομένων).

Στο πλαίσιο της παρούσας εργασίας, η SPARQL αξιοποιείται έμμεσα μέσω του Wikidata Query Service. Το SPARQL endpoint των Wikidata (query.wikidata.org) επιτρέπει την ανάκτηση πληροφοριών για τις οντότητες που συμφιλώνονται. Για παράδειγμα, ένα ερώτημα που ανακτά την εθνικότητα (P27) όλων των καλλιτεχνών που εμφανίζονται στη συλλογή θα είχε τη μορφή: `SELECT ?artist ?citizenship WHERE { ?artist wdt:P27 ?citizenship }`. Αυτή η λειτουργία αξιοποιείται αυτόματα από το OpenRefine κατά τη φάση του εμπλουτισμού δεδομένων.

1.2.6 Το οικοσύστημα LOD και τα GLAM ιδρύματα

Στον χώρο των GLAM ιδρυμάτων, η υιοθέτηση των Linked Open Data έχει αναδειχθεί ως στρατηγική προτεραιότητα. Η διεθνής πλατφόρμα Europeana, που συγκεντρώνει ψηφιακά αντικείμενα από πάνω από 4.000 ευρωπαϊκά πολιτιστικά ιδρύματα, αποτελεί το μεγαλύτερο παράδειγμα LOD στον χώρο της πολιτιστικής κληρονομιάς. Το Europeana Data Model (EDM), που χρησιμοποιεί η πλατφόρμα, βασίζεται στις αρχές του RDF και του Semantic Web, επιτρέποντας τη διαλειτουργική ανταλλαγή δεδομένων μεταξύ εκατοντάδων ιδρυμάτων [4].

Σε εθνικό επίπεδο, πολλές εθνικές βιβλιοθήκες και μουσεία έχουν υιοθετήσει LOD πολιτικές για τη δημοσίευση των συλλογών τους. Η Βιβλιοθήκη του Κογκρέσου των ΗΠΑ, η Βρετανική Βιβλιοθήκη, τα Εθνικά Μουσεία της Σουηδίας και πολλά άλλα ιδρύματα έχουν δημοσιεύσει authority data και μεταδεδομένα συλλογών ως LOD. Αυτή η τάση επιταχύνεται με την ανάπτυξη εργαλείων που καθιστούν τη δημοσίευση LOD προσβάσιμη σε ιδρύματα χωρίς εξειδικευμένο τεχνικό προσωπικό - ένα από τα οποία είναι το OpenRefine που χρησιμοποιείται στην παρούσα εργασία.

Τα οφέλη της υιοθέτησης LOD από τα GLAM ιδρύματα είναι πολλαπλά και τεκμηριωμένα. Πρώτον, βελτιώνεται η ανακαλυψιμότητα (discoverability): τα δεδομένα γίνονται ορατά σε μηχανές αναζήτησης και εφαρμογές που «κατανοούν» RDF, αυξάνοντας την πιθανότητα οι χρήστες να βρουν, ό,τι ψάχνουν. Δεύτερον, ενισχύεται η διαλειτουργικότητα (interoperability): ιδρύματα με διαφορετικά συστήματα μπορούν να ανταλλάξουν δεδομένα χωρίς απώλεια σημασίας. Τρίτον, αυξάνεται η δυνατότητα επαναχρησιμοποίησης (reusability): ερευνητές, εκπαιδευτικοί και εφαρμογές τρίτων μπορούν να αξιοποιήσουν τα δεδομένα για δικούς τους σκοπούς. Τέλος, επιτρέπεται ο αυτόματος εμπλουτισμός τοπικών δεδομένων με πληροφορίες από εξωτερικές πηγές, όπως ακριβώς συμβαίνει στην παρούσα εργασία [4].

1.3 Wikidata: Δομή, Λειτουργία και Εργαλεία

1.3.1 Ιστορικό και γενική περιγραφή

Τα Wikidata είναι μια ανοιχτή, συνεργατική, πολύγλωσση βάση γνώσης που ιδρύθηκε από το Ίδρυμα Wikimedia (Wikimedia Foundation) τον Οκτώβριο του 2012, με αρχικό χρηματοδότη τον Paul Allen (Microsoft). Η ανάπτυξή τους υπήρξε εντυπωσιακά ταχεία: ήδη τον Μάρτιο του 2013, τα Wikidata άρχισαν να χρησιμοποιούνται ως πηγή δεδομένων για τη Wikipedia σε διάφορες γλώσσες, και το 2014 η αγγλική Wikipedia υιοθέτησε τα Wikidata ως κύρια πηγή για infoboxes. Σήμερα, τα Wikidata

περιλαμβάνουν εκατοντάδες εκατομμύρια statements και αποτελούν έναν από τους κεντρικότερους κόμβους του παγκόσμιου οικοσυστήματος Linked Open Data [4].

Η φιλοσοφία των Wikidata βασίζεται στη συνεργατικότητα: όποιος μπορεί να δημιουργήσει λογαριασμό και να συνεισφέρει δεδομένα. Ωστόσο, σε αντίθεση με τη Wikipedia όπου κάθε επεξεργασία είναι ορατή ως κείμενο, στα Wikidata κάθε συνεισφορά αφορά δομημένα δεδομένα: προσθήκη ή τροποποίηση statements, δημιουργία νέων items, εισαγωγή αναφορών πηγής. Αυτή η δομημένη φύση καθιστά τις συνεισφορές πιο ελέγξιμες και τις τυχόν ανακρίβειες πιο εύκολα εντοπίσιμες.

Ένα από τα πιο σημαντικά χαρακτηριστικά των Wikidata είναι η πολυγλωσσία. Κάθε item μπορεί να έχει ετικέτες (labels), περιγραφές (descriptions) και ψευδώνυμα (aliases) σε εκατοντάδες γλώσσες. Αυτό σημαίνει ότι ο ίδιος Vincent van Gogh μπορεί να αναζητηθεί στα αγγλικά, στα ελληνικά (Βίνσεντ βαν Γκογκ), στα γαλλικά (Vincent Van Gogh) ή στα ολλανδικά (Vincent van Gogh), και το σύστημα θα αναγνωρίσει ότι πρόκειται για το ίδιο item (Q5582). Αυτή η γλωσσική ουδετερότητα αποτελεί θεμελιώδες πλεονέκτημα για την αναπαράσταση πολιτιστικών δεδομένων που αφορούν διαφορετικές γλωσσικές κοινότητες [4].

1.3.2 Βασική δομή δεδομένων

Η βασική μονάδα δεδομένων στα Wikidata είναι το Item. Κάθε item αντιπροσωπεύει μια οντότητα του πραγματικού κόσμου - ένα πρόσωπο, έναν τόπο, ένα αντικείμενο, μια έννοια, ένα γεγονός - και αναγνωρίζεται από έναν μοναδικό αριθμητικό κωδικό με το πρόθεμα «Q» (Q-number). Τα Q-numbers αποδίδονται αυτόματα και σειριακά κατά τη δημιουργία κάθε νέου item. Ορισμένα χαρακτηριστικά παραδείγματα: Q5 αντιστοιχεί στην έννοια «άνθρωπος» (Human), Q762 στον Leonardo da Vinci, Q5582 στον Vincent van Gogh, Q160236 στο Metropolitan Museum of Art.

Οι Properties ορίζουν τις σχέσεις μεταξύ items ή τις τιμές που μπορεί να έχει ένα item, και αναγνωρίζονται με το πρόθεμα «P» (P-number). Κάθε property έχει έναν τύπο τιμής: μπορεί να δέχεται ως τιμή ένα άλλο item (item property), μια ημερομηνία (time property), ένα κείμενο (string property), έναν αριθμό (quantity property), μια συντεταγμένη (coordinate property) κ.ά. Για τους σκοπούς της παρούσας εργασίας, ιδιαίτερη σημασία έχουν οι properties που παρουσιάζονται στον παρακάτω πίνακα:

Πίνακας 1.2: Wikidata properties που χρησιμοποιούνται στην παρούσα εργασία

Κωδικός	Ονομασία	Τύπος τιμής	Χρήση στην εργασία
P170	creator (δημιουργός)	Item (Human)	Σύνδεση αντικειμένου - καλλιτέχνη
P186	material used (υλικό)	Item (Material)	Σύνδεση αντικειμένου - υλικού
P27	country of citizenship	Item (Country)	Εθνικότητα καλλιτέχνη
P248	stated in (αναφέρεται σε)	Item	Τεκμηρίωση πηγής (Met)
P571	inception (ημερομηνία δημιουργίας)	Χρονολογία	Έτος δημιουργίας έργου
P195	collection (συλλογή)	Item	Μουσείο που φυλάσσει το έργο
P217	inventory number (αρ. καταλόγου)	String	Αριθμός καταλόγου μουσείου
P31	instance of (αποτελεί)	Item	Κατηγορία αντικειμένου

Κάθε statement (δήλωση) συνδυάζει ένα item με μια property και μια τιμή. Για παράδειγμα, η δήλωση «Ο Βαν Γκογκ είναι ολλανδικής υπηκοότητας» αποδίδεται ως (Q5582, P27, Q55), όπου Q55 είναι το

item για τις Κάτω Χώρες. Ένα item μπορεί να έχει χιλιάδες statements που το περιγράφουν από διαφορετικές οπτικές γωνίες. Το σύνολο των items και των statements των Wikidata αποτελεί έναν τεράστιο γράφο γνώσης που μπορεί να αναρωτηθεί μέσω SPARQL ή να αξιοποιηθεί απευθείας από εφαρμογές μέσω API.

1.3.3 *Qualifiers, References και η αρχή της επαληθευσimότητας*

Ένα από τα πιο ενδιαφέροντα χαρακτηριστικά της αρχιτεκτονικής των Wikidata είναι η δυνατότητα προσθήκης Qualifiers (προσδιοριστών) και References (αναφορών) σε κάθε statement. Οι Qualifiers εξειδικεύουν τη δήλωση, προσθέτοντας περιστάσεις όπως χρόνο (π.χ. «ήταν Γάλλος πολίτης από το 1888»), τόπο, ρόλο, ή άλλα πλαίσια. Η χρήση qualifiers επιτρέπει την αναπαράσταση σύνθετων πραγματικοτήτων που δεν μπορούν να αποδοθούν με απλές τριπλέτες.

Οι References αποτελούν ίσως το πιο κρίσιμο στοιχείο της αρχιτεκτονικής των Wikidata από την άποψη της ακαδημαϊκής αξιοπιστίας. Κάθε statement μπορεί (και πρέπει) να συνοδεύεται από τεκμηρίωση της πηγής του: μια επαληθεύσιμη εξωτερική αναφορά που επιβεβαιώνει την ορθότητα της πληροφορίας. Χωρίς Reference, μια δήλωση μπορεί να επισημανθεί ως αμφισβητήσιμη ή να αφαιρεθεί από άλλους επεξεργαστές. Αυτή η αρχή - «no original research», όπως διατυπώνεται στην κοινότητα - διασφαλίζει ότι τα Wikidata παραμένουν μια βάση γνώσης με επαληθεύσιμο περιεχόμενο, και όχι ένα forum για αυθαίρετες ισχυρισμούς.

Στην παρούσα εργασία, η αρχή αυτή τηρείται συστηματικά: κάθε statement που εισάγεται στα Wikidata συνοδεύεται από Reference με property P248 (stated in) και τιμή Q160236 (Metropolitan Museum of Art). Αυτή η αναφορά τεκμηριώνει ότι η πληροφορία προέρχεται από το επίσημο ανοιχτό dataset του μουσείου, ένα αξιόπιστο και επαληθεύσιμο source. Η συστηματική τήρηση αυτής της πρακτικής δεν είναι απλώς τυπική απαίτηση, αλλά αντικατοπτρίζει τη δέσμευση της εργασίας στις αρχές της ανοιχτής και διαφανούς επιστήμης.

1.3.4 *Εργαλεία του οικοσυστήματος Wikidata*

Το οικοσύστημα των Wikidata περιλαμβάνει μια σειρά εργαλείων που υποστηρίζουν διαφορετικές ανάγκες - από την αναζήτηση δεδομένων έως τη μαζική εισαγωγή. Τα τρία εργαλεία που χρησιμοποιούνται στην παρούσα εργασία αναλύονται παρακάτω.

OpenRefine: Ένα ισχυρό εργαλείο ανοιχτού κώδικα για τον καθαρισμό, τη μετατροπή και τη διασύνδεση δεδομένων. Αναπτύχθηκε αρχικά ως Freebase Gridworks (2009), στη συνέχεια ως Google Refine (2010-2012) και τελικά έγινε ανοιχτού κώδικα ως OpenRefine από το 2012. Διαθέτει ενσωματωμένη λειτουργία Reconciliation που επιτρέπει τη live σύνδεση με εξωτερικά reconciliation services, συμπεριλαμβανομένου του Wikidata reconciliation service. Υποστηρίζει πλήθος μορφών δεδομένων (CSV, TSV, JSON, XML, Google Sheets) και παρέχει τη γλώσσα GREL (General Refine Expression Language) για εκφράσεις μετασχηματισμού [5].

Wikidata Query Service (WDQS): Ένα SPARQL 1.1 endpoint διαθέσιμο στη διεύθυνση query.wikidata.org. Επιτρέπει την υποβολή ερωτημάτων SPARQL σε πραγματικό χρόνο στη βάση δεδομένων των Wikidata. Παρέχει γραφικό περιβάλλον με αυτόματη συμπλήρωση κώδικα (autocomplete), παραδείγματα ερωτημάτων και δυνατότητα οπτικοποίησης αποτελεσμάτων σε χάρτες, γραφήματα χρόνου, δέντρα και άλλες μορφές. Χρησιμοποιείται στην παρούσα εργασία έμμεσα, κατά τη φάση του εμπλουτισμού δεδομένων.

QuickStatements: Ένα εργαλείο μαζικής επεξεργασίας Wikidata, αναπτυγμένο από εθελοντή της κοινότητας και φιλοξενούμενο στη διεύθυνση quickstatements.toolforge.org. Δέχεται αρχεία εντολών

σε απλή σύνταξη (V1 ή V2 syntax) και εκτελεί μαζικές προσθήκες, τροποποιήσεις ή διαγραφές statements στα Wikidata. Κάθε γραμμή του αρχείου V1 syntax αντιστοιχεί σε ένα statement προς εισαγωγή, με τη μορφή: [QItem] TAB [Property] TAB [Value] TAB [Reference Property] TAB [Reference Value]. Αποτελεί το τελικό βήμα της μεθοδολογίας που αναπτύσσεται στην παρούσα εργασία [5].

1.4 Πρότυπα Μεταδομένων για Ψηφιακές Πολιτιστικές Συλλογές

1.4.1 Η σημασία των μεταδομένων στον χώρο των GLAM

Ο όρος GLAM (Galleries, Libraries, Archives, Museums - Γκαλερί, Βιβλιοθήκες, Αρχεία, Μουσεία) αναφέρεται συλλογικά στα ιδρύματα που φυλάσσουν, διαχειρίζονται και παρέχουν πρόσβαση σε τεκμήρια πολιτιστικής και επιστημονικής κληρονομιάς. Η ψηφιοποίηση των συλλογών αυτών τις τελευταίες δεκαετίες δημιούργησε τεράστιους όγκους ψηφιακών δεδομένων που απαιτούν συστηματική τεκμηρίωση και οργάνωση. Τα μεταδεδομένα - δηλαδή τα δεδομένα που περιγράφουν άλλα δεδομένα - αποτελούν το θεμέλιο αυτής της τεκμηρίωσης [6].

Η ποιότητα και η δομή των μεταδεδομένων καθορίζουν σε μεγάλο βαθμό την αξία μιας ψηφιακής συλλογής. Πλούσια, δομημένα και διαλειτουργικά μεταδεδομένα επιτρέπουν: την αποτελεσματική αναζήτηση και ανάκτηση αντικειμένων, τη σύγκριση και ανάλυση αντικειμένων από διαφορετικές συλλογές, την αυτόματη επεξεργασία για στατιστική ανάλυση και έρευνα, τη διαλειτουργική ανταλλαγή δεδομένων μεταξύ ιδρυμάτων, και τη μακροχρόνια διατήρηση με σαφή τεκμηρίωση. Αντίθετα, ελλιπή ή αδόμητα μεταδεδομένα μειώνουν δραστικά τη χρησιμότητα ακόμα και της πιο εκτεταμένης ψηφιακής συλλογής.

Η διαλειτουργικότητα (interoperability) αποτελεί τον κεντρικό στόχο της τυποποίησης μεταδεδομένων στον χώρο των GLAM. Ένας χρήστης που αναζητά έργα ενός καλλιτέχνη θα ήθελε να μπορεί να εντοπίσει αντικείμενα σε πολλαπλές συλλογές ταυτόχρονα, ανεξάρτητα από το ίδρυμα που τα φυλάσσει ή το σύστημα που χρησιμοποιεί. Αυτό προϋποθέτει κοινά πρότυπα που εξασφαλίζουν ότι η ίδια έννοια (π.χ. «δημιουργός») εκφράζεται με τον ίδιο τρόπο σε διαφορετικά συστήματα. Τα πρότυπα Dublin Core και CIDOC-CRM αντιμετωπίζουν αυτήν την ανάγκη από διαφορετικές οπτικές γωνίες.

1.4.2 Dublin Core: απλότητα και ευρεία υιοθέτηση

Το Dublin Core (DC) είναι ένα σχήμα μεταδεδομένων που αναπτύχθηκε αρχικά το 1995 στο Dublin του Οχάιο, σε ένα workshop που συνδιοργάνωσαν το OCLC (Online Computer Library Center) και το NCSA (National Center for Supercomputing Applications). Η κεντρική φιλοσοφία του εκφράστηκε από τους ιδρυτές του με τη φράση «better a simple standard than no standard» - καλύτερα ένα απλό πρότυπο από κανένα πρότυπο. Αυτή η φιλοσοφία εξηγεί γιατί το Dublin Core έχει υιοθετηθεί ευρύτατα: αποτελεί σήμερα επίσημο πρότυπο ISO (ISO 15836-1:2017) και χρησιμοποιείται από χιλιάδες ιδρύματα παγκοσμίως [7].

Ο πυρήνας του Dublin Core αποτελείται από 15 στοιχεία: Title (τίτλος), Creator (δημιουργός), Subject (θέμα), Description (περιγραφή), Publisher (εκδότης), Contributor (συντελεστής), Date (ημερομηνία), Type (τύπος), Format (μορφή), Identifier (αναγνωριστικό), Source (πηγή), Language (γλώσσα), Relation (σχέση), Coverage (κάλυψη) και Rights (δικαιώματα). Κάθε ένα από αυτά τα 15 στοιχεία είναι προαιρετικό, επαναλαμβανόμενο και επεκτάσιμο. Αυτή η ευελιξία το καθιστά κατάλληλο για περιγραφή ποικίλων τύπων ψηφιακών πόρων - από βιβλία και άρθρα έως εικόνες, βίντεο και μουσειακά αντικείμενα.

Το Dublin Core Metadata Terms (DCMI Terms), που επεκτείνει το Simple Dublin Core, εισάγει πρόσθετες έννοιες και καθορίζει με μεγαλύτερη ακρίβεια τη σημασία των βασικών στοιχείων. Για παράδειγμα, το στοιχείο Date χωρίζεται σε πιο εξειδικευμένα στοιχεία όπως created (ημερομηνία δημιουργίας), modified (τελευταία τροποποίηση), issued (ημερομηνία δημοσίευσης) κ.ά. Αυτή η εξειδίκευση βελτιώνει την ακρίβεια της περιγραφής χωρίς να θυσιάζει την απλότητα του πυρήνα.

Για την παρούσα εργασία, το Dublin Core λειτουργεί ως εννοιολογική γέφυρα: τα πεδία του dataset του Met μπορούν να χαρτογραφηθούν (mapped) σε στοιχεία Dublin Core, και αυτά με τη σειρά τους σε Wikidata properties. Για παράδειγμα: Artist Display Name → DC:Creator → Wikidata P170, Medium → DC:Format → Wikidata P186, Object Date → DC:Date → Wikidata P571. Αυτή η διπλή χαρτογράφηση διασφαλίζει τη σημασιολογική συνέπεια της διαδικασίας.

1.4.3 CIDOC-CRM: πλούσια σημασιολογική αναπαράσταση

Το CIDOC Conceptual Reference Model (CIDOC-CRM) αποτελεί ένα ριζικά διαφορετικό πρότυπο από το Dublin Core: ενώ το Dublin Core είναι ένα απλό σχήμα μεταδεδομένων, το CIDOC-CRM είναι μια επίσημη οντολογία (formal ontology) σχεδιασμένη ειδικά για τον χώρο της πολιτιστικής κληρονομιάς. Αναπτύχθηκε από την Επιτροπή Τεκμηρίωσης του Διεθνούς Συμβουλίου Μουσείων (CIDOC - Committee for Documentation of ICOM) και αποτελεί επίσημο πρότυπο ISO 21127 από το 2006. Η τελευταία δημοσιευμένη έκδοση (7.1.3) κυκλοφόρησε το 2023 [8].

Η κεντρική φιλοσοφία του CIDOC-CRM είναι η γεγονотоκεντρική (event-centric) αναπαράσταση: αντί να περιγράφει απευθείας τις ιδιότητες ενός αντικειμένου (π.χ. «αυτό το έργο δημιούργησε ο Βαν Γκογκ»), περιγράφει τα γεγονότα στα οποία συμμετείχε (π.χ. «υπήρξε ένα γεγονός δημιουργίας, στο οποίο συμμετείχε ο Βαν Γκογκ ως δημιουργός, και το αποτέλεσμα ήταν αυτό το έργο»). Αυτή η προσέγγιση επιτρέπει την αναπαράσταση σύνθετων ιστορικών πραγματικοτήτων με πολλή μεγαλύτερη ακρίβεια από ένα απλό σχήμα μεταδεδομένων.

Οι κεντρικές κλάσεις του CIDOC-CRM οργανώνονται σε δύο βασικές κατηγορίες: Enduring Entities (μόνιμες οντότητες, με πρόθεμα E) και Temporal Entities (χρονικές οντότητες - γεγονότα). Ορισμένες από τις πιο σημαντικές κλάσεις για τον χώρο των μουσείων είναι: E22 Human-Made Object (τεχνητό αντικείμενο, π.χ. ένας πίνακας), E39 Actor (δράστης - οποιοδήποτε πρόσωπο ή οργανισμός), E52 Time-Span (χρονικό διάστημα), E53 Place (τόπος) και E65 Creation (γεγονός δημιουργίας ενός αντικειμένου). Οι ιδιότητες (properties) συνδέουν αυτές τις κλάσεις: η P14 carried out by (εκτελέστηκε από) συνδέει ένα γεγονός με τον δράστη, η P45 consists of (αποτελείται από) συνδέει ένα αντικείμενο με τα υλικά του.

Ο παρακάτω πίνακας παρουσιάζει συγκριτικά τα δύο πρότυπα και τη σχέση τους με τα Wikidata:

Πίνακας 1.3: Σύγκριση Dublin Core, CIDOC-CRM και Wikidata properties

Κριτήριο	Dublin Core	CIDOC-CRM	Wikidata
Τύπος	Σχήμα μεταδεδομένων	Επίσημη οντολογία	Βάση γνώσης (LOD)
Πολυπλοκότητα	Χαμηλή (15 στοιχεία)	Πολύ υψηλή (100+ κλάσεις)	Μέτρια (εκατ. properties)
Πρότυπο	ISO 15836	ISO 21127	CC0 (άδεια)
Χρήση	Γενική περιγραφή πόρων	Πλούσια τεκμηρίωση μουσείων	Παγκόσμια διασύνδεση

Αντιστοίχιση DC:Creator	DC:Creator	P14 carried out by	P170 creator
Αντιστοίχιση DC:Format	DC:Format	P45 consists of	P186 material used
Αντιστοίχιση DC:Date	DC:Date	E52 Time-Span	P571 inception

Η σύγκριση αυτή αναδεικνύει ότι τα τρία πρότυπα δεν είναι ανταγωνιστικά αλλά συμπληρωματικά. Το Dublin Core παρέχει το απλό, ευρέως αποδεκτό πλαίσιο· το CIDOC-CRM παρέχει τη σημασιολογική βαθύτητα· και τα Wikidata παρέχουν την παγκόσμια διασύνδεση. Στην παρούσα εργασία, το Dublin Core χρησιμεύει ως εννοιολογικός χάρτης για την ταυτοποίηση των κεντρικών πεδίων, το CIDOC-CRM παρέχει τη σημασιολογική βάση για την κατανόηση του mapping, και τα Wikidata αποτελούν τον στόχο της διασύνδεσης.

1.4.4 Άλλα σχετικά πρότυπα

Πέρα από το Dublin Core και το CIDOC-CRM, υπάρχουν αρκετά άλλα πρότυπα μεταδεδομένων που χρησιμοποιούνται στον χώρο των GLAM και είναι σχετικά με την παρούσα εργασία. Το Spectrum αποτελεί ένα πρότυπο διαχείρισης μουσειακών συλλογών ευρέως χρησιμοποιούμενο στο Ηνωμένο Βασίλειο. Το VRA Core (Visual Resources Association Core) είναι εξειδικευμένο για οπτικά υλικά. Το LIDO (Lightweight Information Describing Objects) είναι ένα XML-based σχήμα για ανταλλαγή μουσειακών δεδομένων, χρησιμοποιούμενο ευρέως από την Europeana.

Ιδιαίτερη μνεία αξίζει το Europeana Data Model (EDM), το οποίο αναπτύχθηκε ειδικά για τις ανάγκες της Europeana και βασίζεται στις αρχές του RDF και του Semantic Web. Το EDM επιτρέπει τη χαρτογράφηση από ποικίλα τοπικά σχήματα μεταδεδομένων σε ένα κοινό μοντέλο, διευκολύνοντας τη διαλειτουργική συγκέντρωση δεδομένων από εκατοντάδες ευρωπαϊκά ιδρύματα. Η σχέση μεταξύ EDM και Wikidata έχει αποτελέσει αντικείμενο ερευνητικής μελέτης, με αρκετές εργασίες να εξερευνούν τρόπους διασύνδεσης των δύο οικοσυστημάτων.

1.5 Βιβλιογραφική Ανασκόπηση

1.5.1 Η αναδυόμενη σημασία των Wikidata για τα GLAM ιδρύματα

Τα Wikidata έχουν αναδειχθεί ταχύτατα σε κεντρικό κόμβο του οικοσυστήματος LOD για τον χώρο των GLAM. Πρόσφατη συστηματική ανασκόπηση της βιβλιογραφίας (Candela et al., 2024), που εξέτασε δεκάδες εργασίες σχετικά με τη χρήση των Wikidata από GLAM ιδρύματα, κατέδειξε ότι η εν λόγω χρήση καλύπτει ένα ευρύ φάσμα σκοπών: δημοσίευση authority data, εμπλουτισμός μεταδεδομένων, δημιουργία ψηφιακών εκθέσεων, υποστήριξη ερευνητικών έργων και βελτίωση της ορατότητας των συλλογών [6]. Η ίδια έρευνα υπογράμμισε ότι διεθνείς πρωτοβουλίες όπως η International GLAM Labs Community, η Europeana και το πρόγραμμα Collections as Data συστήνουν ρητά τη χρήση των Wikidata.

Η επέκταση της χρήσης Wikidata από GLAM ιδρύματα δεν αποτελεί απλώς ακαδημαϊκή τάση, αλλά αντικατοπτρίζει πρακτικές ανάγκες: τα ιδρύματα αντιμετωπίζουν αυξανόμενη πίεση να καταστήσουν τις συλλογές τους πιο ορατές και προσβάσιμες στο διαδίκτυο, και τα Wikidata αποτελούν έναν αποτελεσματικό και οικονομικό τρόπο να το επιτύχουν. Επιπλέον, η άδεια CC0 των Wikidata εξαλείφει νομικά εμπόδια που συχνά συνδέονται με άλλες πλατφόρμες.

1.5.2 Σουηδικά μουσεία: μια τεκμηριωμένη εμπειρία

Μια από τις πιο ολοκληρωμένες και καλά τεκμηριωμένες εμπειρίες διασύνδεσης μουσειακών συλλογών με τα Wikidata προέρχεται από τη Σουηδία. Η Fagervig et al. (2023) παρουσίασε λεπτομερώς το έργο που υλοποίησαν το Nationalmuseum της Στοκχόλμης και τα Εθνικά Ιστορικά Μουσεία σε συνεργασία με το Wikimedia Sverige. Στόχος ήταν η δημιουργία authority records στα Wikidata για πρόσωπα (καλλιτέχνες, δωρητές, κ.ά.) που σχετίζονται με τις μουσειακές συλλογές, και η διασύνδεση των υπαρχόντων εγγραφών των βάσεων δεδομένων των μουσείων με τα αντίστοιχα Wikidata items [5].

Η μεθοδολογία που ακολουθήθηκε ήταν παρόμοια με αυτή της παρούσας εργασίας: χρήση OpenRefine για τη συμφιλίωση εγγραφών με τα Wikidata, με GREL για την προεπεξεργασία δεδομένων και QuickStatements για τη μαζική εισαγωγή. Ένα ιδιαίτερα ενδιαφέρον εύρημα ήταν η αποτελεσματικότητα του OpenRefine ακόμα και για χρήστες χωρίς τεχνικό υπόβαθρο: τα μέλη της ομάδας του μουσείου, αν και δεν ήταν προγραμματιστές, μπόρεσαν να εκμάθουν τις απαραίτητες δεξιότητες GREL και να ολοκληρώσουν επιτυχώς τη συμφιλίωση [5]. Αυτό επιβεβαιώνει ότι η μεθοδολογία διασύνδεσης με τα Wikidata είναι ρεαλιστική και για μικρότερα ιδρύματα χωρίς εξειδικευμένο τεχνικό προσωπικό.

Επιπλέον, η εν λόγω έρευνα επισήμανε τη σημασία της κοινοτικής διάστασης: η αλληλεπίδραση με την κοινότητα των Wikidata, η λήψη ανατροφοδότησης για τα εισαγόμενα δεδομένα και η συμμετοχή σε εργαστήρια (edit-a-thons) αποτέλεσαν σημαντική πηγή γνώσης και αφύλησης. Αυτή η κοινοτική διάσταση αποτελεί χαρακτηριστικό των Wikidata που απουσιάζει από άλλες πλατφόρμες LOD.

1.5.3 Το Metropolitan Museum of Art και τα ανοιχτά δεδομένα

Το dataset του Metropolitan Museum of Art, που αποτελεί το αντικείμενο μελέτης της παρούσας εργασίας, έχει χρησιμοποιηθεί σε πλήθος ερευνητικών εργασιών. Στιμλερ και Ρόλινσον (2019) ανέλυσαν τον αντίκτυπο που είχε στα Wikidata και τη Wikipedia η δημοσίευση ανοιχτών datasets από μουσεία τέχνης, συμπεριλαμβανομένου του Met. Τα αποτελέσματά τους έδειξαν ότι η δημοσίευση ανοιχτών μεταδεδομένων οδηγεί σε αυξημένες προβολές Wikipedia, εμπλουτισμό items Wikidata και βελτιωμένη αναπαράσταση των καλλιτεχνών της συλλογής [3].

Πιο πρόσφατη έρευνα (Nakasone et al., 2026) αξιοποίησε το OpenRefine και τα Wikidata για τη συμφιλίωση εγγραφών του Index of American Design (Smithsonian) με linked data entities. Ιδιαίτερο ενδιαφέρον παρουσίασε η αντιμετώπιση ονομάτων με ελλιπείς πληροφορίες ή με πολλά πιθανά matches - μια πρόκληση που είναι κοινή σε πολλά μουσειακά datasets, συμπεριλαμβανομένου αυτού του Met που χρησιμοποιείται στην παρούσα εργασία [11].

1.5.4 Αξιολόγηση ανοιχτών πρακτικών μουσείων τέχνης

Η πιο πρόσφατη και ολοκληρωμένη αξιολόγηση των πρακτικών ανοιχτών δεδομένων μουσείων τέχνης στα Wikidata προέρχεται από τους Candela et al. (2025). Η έρευνα αυτή εξέτασε συστηματικά τον τρόπο με τον οποίο διεθνώς αναγνωρισμένα μουσεία τέχνης - μεταξύ των οποίων το Met, το Rijksmuseum, το Louvre και η Tate Gallery - χρησιμοποιούν τα Wikidata για τη δημοσίευση και διασύνδεση των συλλογών τους [9].

Τα αποτελέσματα κατέδειξαν τρία κεντρικά ευρήματα. Πρώτον, η διασύνδεση με τα Wikidata βελτιώνει σημαντικά την ανακαλυψιμότητα, τη διαλειτουργικότητα και τη δυνατότητα επαναχρησιμοποίησης των δεδομένων. Δεύτερον, η ποιότητα των εισαγόμενων δεδομένων - και ειδικά η παρουσία αναφορών πηγής (references) - αποτελεί κρίσιμο παράγοντα για την αποδοχή τους από την κοινότητα. Τρίτον, παρά τα σαφή οφέλη, πολλά μουσεία εξακολουθούν να αξιοποιούν τα Wikidata μερικώς ή

αναποτελεσματικά, κυρίως λόγω έλλειψης γνώσης ή ανθρώπινων πόρων [9]. Αυτό το εύρημα υπογραμμίζει τη σημασία της ανάπτυξης προσβάσιμων μεθοδολογιών και εκπαιδευτικών εργαλείων - ακριβώς αυτό που επιδιώκει η παρούσα εργασία.

1.5.5 Ιταλικά GLAM ιδρύματα: μεγάλης κλίμακας διασύνδεση

Το έργο Empowering Italian GLAMs (2022–2024), που παρουσιάστηκε από τους Rossenova et al. (2026), αποτελεί το πιο εκτεταμένο παράδειγμα συστηματικής διασύνδεσης μουσειακών δεδομένων με τα Wikidata σε εθνικό επίπεδο. Το έργο δημιούργησε 1.304 νέα items στα Wikidata αντιστοιχώντας σε ιταλικά μουσεία, και εμπλούτισε πάνω από 6.800 υπάρχοντα items με δεδομένα από τις εθνικές στατιστικές έρευνες του ISTAT (Ιταλικό Εθνικό Ινστιτούτο Στατιστικής) [10].

Ιδιαίτερα ενδιαφέροντα ήταν η μεθοδολογία διασφάλισης ποιότητας (quality assurance) που εφαρμόστηκε: πριν από κάθε μαζική εισαγωγή δεδομένων, εκτελείτο συστηματικός έλεγχος μέσω OpenRefine για τον εντοπισμό και διόρθωση σφαλμάτων. Επιπλέον, η χειροκίνητη επιμέλεια πραγματοποιήθηκε περιοχή προς περιοχή, εξασφαλίζοντας ότι κάθε εγγραφή ελέγχθηκε από ειδικό με γνώση της τοπικής πραγματικότητας. Αυτή η προσέγγιση - συνδυασμός αυτόματης επεξεργασίας με στοχευμένη χειροκίνητη επιμέλεια - αποτελεί βέλτιστη πρακτική που θα μπορούσε να υιοθετηθεί και σε μελλοντικές ελληνικές εφαρμογές.

1.5.6 Θέση της παρούσας εργασίας στη βιβλιογραφία

Η σύγκριση με τη διεθνή βιβλιογραφία αναδεικνύει τη μοναδική θέση της παρούσας εργασίας. Σε αντίθεση με μεγάλα ευρωπαϊκά έργα όπως το Empowering Italian GLAMs, η παρούσα εργασία εστιάζει σε έναν μεμονωμένο ερευνητή που εφαρμόζει τη μεθοδολογία σε δείγμα ενός διεθνούς dataset. Αυτή η οπτική είναι εξίσου σημαντική: αποδεικνύει ότι η διασύνδεση με τα Wikidata είναι εφικτή ακόμα και χωρίς θεσμικούς πόρους ή εξειδικευμένη ομάδα.

Σε σχέση με τη σουηδική εμπειρία (Fagervig et al., 2023), η παρούσα εργασία παρουσιάζει παρόμοια μεθοδολογία αλλά διαφορετικό πλαίσιο: αντί για authority data εθνικών μουσείων, εξετάζει τη διασύνδεση δεδομένων ενός ανοιχτού dataset με τα Wikidata. Αυτή η διαφορά αναδεικνύει ένα πρόσθετο σενάριο χρήσης: τη διασύνδεση ήδη δημοσιευμένων ανοιχτών datasets - όχι μόνο εσωτερικών βάσεων δεδομένων ιδρυμάτων.

Τέλος, η εργασία αυτή αποτελεί μία από τις πρώτες ελληνόγλωσσες ακαδημαϊκές εργασίες που πραγματεύονται συστηματικά τη διασύνδεση μουσειακής συλλογής με τα Wikidata. Αυτό είναι σημαντικό διότι καλύπτει ένα κενό στην ελληνική βιβλιογραφία, και θέτει τις βάσεις για μελλοντική έρευνα που θα μπορούσε να εφαρμόσει παρόμοιες μεθοδολογίες σε ελληνικές μουσειακές συλλογές, συμβάλλοντας στην ενσωμάτωση της ελληνικής πολιτιστικής κληρονομιάς στο παγκόσμιο οικοσύστημα Linked Open Data.

Κεφάλαιο 2ο: Η Ψηφιακή Συλλογή του Metropolitan Museum of Art

2.1 Εισαγωγή

Το παρόν κεφάλαιο εστιάζει στη μουσειακή ψηφιακή συλλογή που αποτελεί το αντικείμενο μελέτης της εργασίας. Παρουσιάζεται αρχικά το Metropolitan Museum of Art και η πολιτική ανοιχτής πρόσβασης που ακολουθεί, στη συνέχεια αναλύεται η δομή του ανοιχτού dataset του, και τέλος τεκμηριώνεται η διαδικασία επιλογής του δείγματος εργασίας από το τμήμα Ελληνικής και Ρωμαϊκής Τέχνης.

Η επιλογή του συγκεκριμένου dataset στηρίχθηκε σε τρεις παράγοντες. Πρόκειται για ένα από τα πιο εκτεταμένα ανοιχτά μουσειακά datasets διεθνώς, με 484.956 εγγραφές. Διατίθεται υπό άδεια Creative Commons Zero, η οποία επιτρέπει την απρόσκοπτη επαναχρησιμοποίηση και είναι συμβατή με τις απαιτήσεις των Wikidata. Και χρησιμοποιείται ευρέως στη διεθνή έρευνα, γεγονός που επιτρέπει τη σύγκριση των αποτελεσμάτων με άλλες εργασίες.

Η περαιτέρω εστίαση στο τμήμα Ελληνικής και Ρωμαϊκής Τέχνης προσδίδει στην εργασία διπλό ενδιαφέρον. Αφενός αφορά ελληνική πολιτιστική κληρονομιά που φυλάσσεται σε ξένο ίδρυμα, θέμα με ιδιαίτερη σημασία για τον ελληνικό ακαδημαϊκό χώρο. Αφετέρου θέτει τη μεθοδολογία αντιμετώπισης με ιδιαιτερότητες που δεν εμφανίζονται σε συλλογές νεότερης τέχνης και δοκιμάζουν τα όρια των εργαλείων συμφιλίωσης.

2.2 Το Metropolitan Museum of Art

2.2.1 Ιστορικό και συλλογή

Το Metropolitan Museum of Art ιδρύθηκε το 1870 στη Νέα Υόρκη με πρωτοβουλία ομάδας Αμερικανών επιχειρηματιών, πολιτικών, νομικών και καλλιτεχνών που οραματίζονταν τη δημιουργία ενός εθνικού μουσείου τέχνης αντάξιου των μεγάλων ευρωπαϊκών ιδρυμάτων. Η πρώτη έκθεση άνοιξε για το κοινό το 1872 στο κτίριο Dodworth Studio, ενώ το μουσείο μεταφέρθηκε στη σημερινή του θέση στο Central Park το 1880.

Η μόνιμη συλλογή του αριθμεί σήμερα περισσότερα από 1,5 εκατομμύριο αντικείμενα, που εκτείνονται σε χρονικό ορίζοντα άνω των 5.000 ετών και προέρχονται από κάθε γεωγραφική περιοχή του πλανήτη. Το μουσείο λειτουργεί σε τρεις τοποθεσίες και προσελκύει ετησίως εκατομμύρια επισκέπτες, καταλαμβάνοντας σταθερά θέση μεταξύ των δέκα πιο επισκεπτόμενων μουσείων παγκοσμίως.

2.2.2 Η Open Access Initiative

Τον Φεβρουάριο του 2017 το Met ανακοίνωσε την Open Access Initiative, σύμφωνα με την οποία τα ψηφιακά αρχεία των έργων που βρίσκονται στον δημόσιο τομέα διατίθενται ελεύθερα για λήψη και επαναχρησιμοποίηση υπό την άδεια Creative Commons Zero. Η απόφαση σηματοδότησε στρατηγική μετατόπιση: το μουσείο εγκατέλειψε το μοντέλο των τελών αδειοδότησης εικόνων και υιοθέτησε τη λογική της ανοιχτής γνώσης ως δημόσιου αγαθού.

Η επιλογή της άδειας CC0 έχει καθοριστική σημασία για την παρούσα εργασία. Σε αντίθεση με άλλες άδειες Creative Commons που επιβάλλουν υποχρέωση αναφοράς ή περιορισμούς εμπορικής χρήσης, η CC0 συνιστά πλήρη παραίτηση από δικαιώματα. Τα Wikidata λειτουργούν επίσης υπό CC0 και δεν θα

μπορούσαν να δεχθούν δεδομένα με περιοριστικότερη άδεια, οπότε η συμβατότητα αυτή αποτελεί προϋπόθεση και όχι απλώς διευκόλυνση.

Παράλληλα με τις εικόνες, το Met δημοσίευσε στην πλατφόρμα GitHub το πλήρες dataset μεταδεδομένων της συλλογής του. Το αποθετήριο ενημερώνεται τακτικά και δέχεται αναφορές σφαλμάτων από την κοινότητα μέσω του μηχανισμού issues, πρακτική που καθιστά τη διαδικασία επιμέλειας διαφανή και συμμετοχική.

2.3 Δομή του Dataset

2.3.1 Γενική επισκόπηση

Το dataset διανέμεται σε μορφή CSV με κωδικοποίηση UTF-8-BOM. Στην έκδοση που χρησιμοποιήθηκε στην παρούσα εργασία περιλαμβάνει 484.956 εγγραφές και 54 στήλες μεταδεδομένων. Κάθε γραμμή αντιστοιχεί σε ένα αντικείμενο και αναγνωρίζεται μοναδικά από το πεδίο Object ID.

Οι 54 στήλες ομαδοποιούνται σε πέντε κατηγορίες. Τα πεδία ταυτοποίησης περιλαμβάνουν το Object ID, το Object Number και το Department. Τα πεδία περιγραφής αφορούν τον τίτλο, τον τύπο αντικειμένου, το υλικό, τις διαστάσεις και την κατηγοριοποίηση. Τα πεδία δημιουργού καλύπτουν το ονοματεπώνυμο, τον ρόλο, την εθνικότητα και τις χρονολογίες ζωής. Τα πεδία προέλευσης αφορούν τη χρονολόγηση, τον πολιτισμό, την περίοδο και τη γεωγραφική θέση. Τέλος, τα πεδία διαχείρισης περιλαμβάνουν το έτος απόκτησης, την αίθουσα έκθεσης και το καθεστώς δικαιωμάτων.

2.3.2 Η κατανομή ανά τμήμα

Η ανάλυση του πεδίου Department μέσω Text Facet στο OpenRefine αποκάλυψε ότι η συλλογή οργανώνεται σε δεκαεννέα τμήματα, με έντονα ανομοιόμορφη κατανομή. Το μεγαλύτερο τμήμα, τα Σχέδια και Χαρακτικά, συγκεντρώνει από μόνο του πάνω από το ένα τρίτο των εγγραφών.

Πίνακας 2.1: Η κατανομή των εγγραφών στα δεκαεννέα τμήματα του Met

Τμήμα	Εγγραφές	Ποσοστό
Drawings and Prints	172.630	35,6%
European Sculpture and Decorative Arts	43.051	8,9%
Photographs	37.459	7,7%
Asian Art	37.000	7,6%
Greek and Roman Art	33.726	7,0%
Costume Institute	31.652	6,5%
Egyptian Art	27.969	5,8%
The American Wing	18.532	3,8%
Islamic Art	15.573	3,2%
Modern and Contemporary Art	14.696	3,0%
Arms and Armor	13.623	2,8%
Arts of Africa, Oceania, and the Americas	12.367	2,6%

Medieval Art	7.142	1,5%
Ancient Near Eastern Art	6.223	1,3%
Musical Instruments	5.227	1,1%
European Paintings	2.626	0,5%
Robert Lehman Collection	2.586	0,5%
The Cloisters	2.340	0,5%
The Libraries	534	0,1%
Σύνολο	484.956	100%

OpenRefine Met_Museum_Wikidata_Thesis

127.0.0.1:3333/project?project=2167982886393

Open... Export... Help

Facet / Filter Undo / Redo 0/0 484,956 rows Extensions Wikibase

Refresh Reset all Remove all Show as: rows records Show: 5 10 25 50 100 500 1000 rows

Object Number Is Highlight Is Timeline Work Is Public Domain Object ID Gallery Number Department AccessionYear Object Name Title

1. 1979.486.1 False False False 1 The American Wing 1979 Coin One-dollar Liberty Head Coin

2. 1980.264.5 False False False 2 The American Wing 1980 Coin Ten-dollar Liberty Head Coin

3. 67.265.9 False False False 3 The American Wing 1967 Coin Two-and-a-Half Dollar Coin

4. 67.265.10 False False False 4 The American Wing 1967 Coin Two-and-a-Half Dollar Coin

5. 67.265.11 False False False 5 The American Wing 1967 Coin Two-and-a-Half Dollar Coin

6. 67.265.12 False False False 6 The American Wing 1967 Coin Two-and-a-Half Dollar Coin

7. 67.265.13 False False False 7 The American Wing 1967 Coin Two-and-a-Half Dollar Coin

8. 67.265.14 False False False 8 The American Wing 1967 Coin Two-and-a-Half Dollar Coin

9. 67.265.15 False False False 9 The American Wing 1967 Coin Two-and-a-Half Dollar Coin

10. 1979.486.3 False False False 10 The American Wing 1979 Coin Two-and-a-half-dollar Indian Head

11. 1979.486.2 False False False 11 The American Wing 1979 Coin Two-and-a-half-dollar Liberty Head

12. 1979.486.7 False False False 12 The American Wing 1979 Coin Twenty-dollar Liberty Head Coin

13. 1979.486.4 False False False 13 The American Wing 1979 Coin Five-dollar Indian Head Coin

14. 1979.486.5 False False False 14 The American Wing 1979 Coin Five-dollar Liberty Head Coin

15. 16.74.49 False False False 15 The American Wing 1916 Coin, 1/2 Real

16. 16.74.27 False False False 16 The American Wing 1916 Coin, 1/4 Peso

17. 16.74.28 False False False 17 The American Wing 1916 Coin, 1/4 Peso

18. 16.74.29 False False False 18 The American Wing 1916 Coin, 1/4 Peso

Σχήμα 2.1: Τα δέκα μεγαλύτερα τμήματα της συλλογής

Το τμήμα Greek and Roman Art κατέχει την πέμπτη θέση με 33.726 εγγραφές, ποσοστό 7,0% της συλλογής. Το μέγεθος αυτό είναι ιδανικό για τους σκοπούς της εργασίας: αρκετά μεγάλο ώστε να προσφέρει ποικιλία και στατιστική επάρκεια, αλλά όχι τόσο ώστε να καθιστά αδύνατη τη λεπτομερή εξέταση.

2.3.3 Το κρίσιμο πεδίο Artist Wikidata URL

Το dataset περιλαμβάνει μια στήλη που αποδείχθηκε καθοριστική για την αξιολόγηση της μεθοδολογίας: την Artist Wikidata URL. Πρόκειται για σύνδεσμο προς το Wikidata item κάθε δημιουργού, τον οποίο έχει καταχωρίσει το ίδιο το μουσείο ως μέρος της πολιτικής διασύνδεσης των authority records του.

Η ύπαρξη της στήλης θα μπορούσε εκ πρώτης όψεως να θεωρηθεί ότι υπονομεύει τη σκοπιμότητα της εργασίας. Δεν ισχύει, για τρεις λόγους. Πρώτον, η στήλη αφορά αποκλειστικά τους δημιουργούς· τα υλικά, οι πολιτισμοί, οι χρονολογήσεις και τα ίδια τα αντικείμενα παραμένουν αδιασύνδετα. Δεύτερον, η στήλη δεν είναι πλήρης, όπως θα φανεί στο δείγμα. Τρίτον, και σημαντικότερο, η ύπαρξη ανεξάρτητα

τεκμηριωμένων αντιστοιχίσεων παρέχει ένα θεσμικό σύνολο αναφοράς έναντι του οποίου μπορεί να αξιολογηθεί αντικειμενικά η ποιότητα της αυτόματης συμφιλίωσης.

Η δυνατότητα αυτή αξιοποιείται συστηματικά στο Κεφάλαιο 4 και αποτελεί μία από τις σημαντικότερες μεθοδολογικές συνεισφορές της εργασίας. Στην πλειονότητα των αντίστοιχων ερευνών, η αξιολόγηση της συμφιλίωσης παραμένει ποιοτική ή απαιτεί χειροκίνητη κατασκευή δείγματος ελέγχου. Εδώ το πρότυπο υπήρχε εξ αρχής μέσα στα δεδομένα.

2.4 Επιλογή του Δείγματος

2.4.1 Το σκεπτικό της θεματικής εστίασης

Η επιλογή του δείγματος έγινε με θεματικό και όχι με τυχαίο κριτήριο. Η απόφαση αυτή αξίζει αιτιολόγησης, καθώς η τυχαία δειγματοληψία θεωρείται συνήθως προτιμότερη σε στατιστικές μελέτες.

Ο πρώτος λόγος αφορά τη θεματική συνοχή. Ένα τυχαίο δείγμα 500 εγγραφών από τις 484.956 θα περιλάμβανε αιγυπτιακές σαρκοφάγους, ευρωπαϊκά χαρακτικά, ιαπωνικά υφάσματα και αμερικανικά έπιπλα - αντικείμενα χωρίς κοινά χαρακτηριστικά ως προς τα υλικά, τη χρονολόγηση ή την τεκμηρίωση. Η ανάλυση των αποτελεσμάτων θα ήταν αντίστοιχα διάχυτη.

Ο δεύτερος λόγος αφορά την ερευνητική σημασία για τον ελληνικό χώρο. Η διασύνδεση ελληνικής πολιτιστικής κληρονομιάς με μια παγκόσμια, πολύγλωσση βάση γνώσης έχει αξία που υπερβαίνει την τεχνική άσκηση: συμβάλλει στην ορατότητα και την ερευνητική προσβασιμότητα του ελληνικού πολιτισμού διεθνώς.

Ο τρίτος λόγος αφορά τη μεθοδολογική πρόκληση. Οι αρχαιολογικές συλλογές παρουσιάζουν ιδιαιτερότητες που δοκιμάζουν τα όρια των εργαλείων: συμβατικά ονόματα δημιουργών αντί για τεκμηριωμένα πρόσωπα, συντριπτική πλειονότητα ανώνυμων αντικειμένων, και αρχαίες πολιτικές οντότητες που δεν αντιστοιχούν σε σύγχρονα κράτη. Οι προκλήσεις αυτές συνιστούν πρωτότυπη συνεισφορά της εργασίας και δεν θα εμφανίζονταν σε άλλο δείγμα.

2.4.2 Το εύρημα της ανωνυμίας

Η εφαρμογή Text Facet στη στήλη Artist Display Name, περιορισμένη στο τμήμα Greek and Roman Art, απέδωσε ένα από τα σημαντικότερα ευρήματα της εργασίας. Από τις 33.726 εγγραφές, οι 31.581 δεν διαθέτουν καμία τιμή στο πεδίο του δημιουργού.

Πίνακας 2.2: Η απόδοση δημιουργού στο τμήμα Greek and Roman Art

Κατηγορία	Εγγραφές	Ποσοστό
Χωρίς επώνυμο δημιουργό	31.581	93,6%
Με επώνυμο δημιουργό	2.145	6,4%
Σύνολο τμήματος	33.726	100%



Σχήμα 2.2: Η κυριαρχία της ανωνυμίας στην αρχαιολογική συλλογή

Το ποσοστό 93,6% δεν συνιστά ελάττωμα της τεκμηρίωσης του μουσείου αλλά αντανακλά την ίδια τη φύση του αρχαιολογικού υλικού. Στην αρχαιότητα, η ταυτότητα του τεχνίτη σπανίως καταγραφόταν και ακόμη σπανιότερα διασώθηκε. Η υπογραφή σε αγγείο αποτελούσε εξαίρεση και όχι κανόνα.

Το εύρημα έχει άμεσες συνέπειες για τη μεθοδολογία. Η ιδιότητα P170 (creator), που αποτελεί τον έναν από τους δύο άξονες του mapping, είναι εφαρμόσιμη μόλις στο 6,4% της συλλογής. Για μια πληρέστερη διασύνδεση αρχαιολογικού υλικού θα χρειαζόνταν ιδιότητες όπως η P2596 (culture), η P571 (inception) ή η P189 (location of discovery), οι οποίες είναι διαθέσιμες για πολύ μεγαλύτερο ποσοστό αντικειμένων. Η παρατήρηση αυτή αναπτύσσεται περαιτέρω στο Κεφάλαιο 5.

2.4.3 Η διαδικασία φιλτραρίσματος

Το τελικό δείγμα προέκυψε από τη διαδοχική εφαρμογή τριών φίλτρων στο OpenRefine, τα οποία περιγράφονται αναλυτικά στο Κεφάλαιο 3.

Πίνακας 2.3: Η διαδοχική εφαρμογή των τριών φίλτρων επιλογής

Βήμα	Φίλτρο	Κριτήριο	Εγγραφές
0	-	Πλήρες dataset	484.956
1	Text Facet	Department = Greek and Roman Art	33.726
2	Text Facet (invert)	Artist Display Name μη κενό	2.145
3	Custom Text Facet	Object ID από 710000 έως 820000	501

Το τρίτο φίλτρο απαιτεί σχολιασμό. Ο περιορισμός σε αριθμητικό εύρος Object ID δεν αντιστοιχεί σε κάποιο θεματικό ή ποιοτικό κριτήριο· λειτουργεί απλώς ως μηχανισμός περιορισμού του μεγέθους σε διαχειρίσιμο επίπεδο. Επειδή τα Object ID αποδίδονται σειριακά κατά την καταχώριση, το εύρος αντιστοιχεί σε συγκεκριμένη περίοδο ψηφιοποίησης - γεγονός που εξηγεί, όπως θα φανεί, την εντυπωσιακή ομοιογένεια του τελικού δείγματος.

Το τελικό δείγμα αριθμεί 501 εγγραφές, δηλαδή το 23,4% όλων των έργων του τμήματος που διαθέτουν επώνυμο δημιουργό. Πρόκειται επομένως για ουσιώδες ποσοστό του αξιοποιήσιμου υλικού και όχι για μικρό τυχαίο απόσπασμα.

2.5 Χαρακτηριστικά του Δείγματος

2.5.1 Σύνθεση και τυπολογία

Το δείγμα των 501 εγγραφών παρουσιάζει αξιοσημείωτη ομοιογένεια, γεγονός που επιβεβαιώνει την επιτυχία της θεματικής εστίασης και διευκολύνει την ερμηνεία των αποτελεσμάτων.

Πίνακας 2.4: Βασικά χαρακτηριστικά του δείγματος

Χαρακτηριστικό	Τιμή
Σύνολο εγγραφών	501
Στήλες μετά τον εμπλουτισμό	57
Τμήμα	Greek and Roman Art (100%)
Κατηγορία	Vases - 497 εγγραφές (99,2%)
Κυρίαρχος πολιτισμός	Greek, Attic - 479 εγγραφές (95,6%)
Κυρίαρχο υλικό	terracotta - 497 εγγραφές (99,2%)
Δεύτερο υλικό	plaster - 2 εγγραφές (0,4%)
Μοναδικά ονόματα δημιουργών	125
Χρονολογικό εύρος	750 π.Χ. – 1 π.Χ.
Κυρίαρχη περίοδος	Archaic (54%)
Έτος απόκτησης	2011 για 495 εγγραφές (98,8%)
Εύρος Object ID	710000 – 813903

Η τυπολογία επιβεβαιώνει τη συνοχή: 395 εγγραφές αφορούν θραύσματα κύλικας, δηλαδή αττικού αγγείου πόσεως, ενώ ακολουθούν θραύσματα σκύφου, κρατήρα και οinoχόης. Πρόκειται δηλαδή για συλλογή θραυσμάτων αρχαίας ελληνικής αγγειοπλαστικής, κατά κύριο λόγο αττικής παραγωγής της αρχαϊκής και πρώιμης κλασικής περιόδου.

Πίνακας 2.5: Οι συχνότεροι τύποι αντικειμένων

Τύπος αντικειμένου	Εγγραφές
Kylix fragment (θραύσμα κύλικας)	395
Kylix fragments (πολλαπλά θραύσματα)	27
Skyphos fragment (θραύσμα σκύφου)	14
Column-krater fragment (θραύσμα κρατήρα)	7
Oinochoe fragment (θραύσμα οinoχόης)	5
Λοιποί τύποι	53

2.5.2 Η ασυμμετρία στην κατανομή των δημιουργών

Η κατανομή των έργων ανά δημιουργό είναι έντονα ασύμμετρη και αποδεικνύεται καθοριστική για τα αποτελέσματα της συμφιλίωσης. Ενώ το δείγμα περιλαμβάνει 125 μοναδικά ονόματα, ένας μόνο αγγειογράφος - ο Makron - συγκεντρώνει 172 από τα 501 αντικείμενα, ποσοστό 34,3%.

Πίνακας 2.6: Οι δώδεκα δημιουργοί με τα περισσότερα έργα

Δημιουργός	Έργα	Ποσοστό	Τύπος ονόματος
Makron	172	34,3%	Πραγματικό (υπογραφή)
Euaion Painter	22	4,4%	Συμβατικό
Onesimos	21	4,2%	Πραγματικό (υπογραφή)
Douris	21	4,2%	Πραγματικό (υπογραφή)
Brygan	12	2,4%	Αβέβαιο
Brygos Painter	10	2,0%	Συμβατικό
Painter of the Paris Gigantomachy	10	2,0%	Συμβατικό
Veii Painter	10	2,0%	Συμβατικό
Orleans Painter	8	1,6%	Συμβατικό
Maler von London E 777	8	1,6%	Συμβατικό
Briseis Painter	7	1,4%	Συμβατικό
Boot Painter	7	1,4%	Συμβατικό

2.5.3 Η συλλογή von Bothmer

Η ομοιογένεια του δείγματος εξηγείται σε μεγάλο βαθμό από την κοινή προέλευσή του. Οι 495 από τις 501 εγγραφές αποκτήθηκαν από το μουσείο το 2011 ως δωρεά του Dietrich von Bothmer, ο οποίος υπηρέτησε επί δεκαετίες ως επιμελητής και στη συνέχεια ως Distinguished Research Curator του τμήματος Ελληνικής και Ρωμαϊκής Τέχνης.

Η δωρεά αποτελείται κατά κύριο λόγο από θραύσματα αττικών αγγείων. Τα θραύσματα, αν και στερούνται της οπτικής πληρότητας ενός ακέρατου αγγείου, έχουν ιδιαίτερη ερευνητική αξία: επιτρέπουν τη στυλιστική ανάλυση που αποτελεί τη βάση της απόδοσης, συχνά συμπληρώνουν αγγεία που βρίσκονται σε άλλες συλλογές, και τεκμηριώνουν την παραγωγή εργαστηρίων για τα οποία σώζονται λίγα ακέρατα δείγματα.

Για την παρούσα εργασία η κοινή προέλευση συνιστά μεθοδολογικό πλεονέκτημα. Τα αντικείμενα τεκμηριώθηκαν από τον ίδιο ειδικό με ενιαία κριτήρια και ορολογία, γεγονός που περιορίζει την ανομοιομορφία των μεταδεδομένων και επιτρέπει την απομόνωση των προβλημάτων που οφείλονται καθαρά στη διαδικασία συμφιλίωσης και όχι στην ποιότητα της πηγής.

2.6 Η Αττική Αγγειογραφία ως Αντικείμενο Τεκμηρίωσης

2.6.1 Η σπανιότητα των υπογραφών

Η σύνθεση του δείγματος επιβάλλει σύντομη εξέταση του τρόπου με τον οποίο η αρχαιολογική επιστήμη αποδίδει έργα σε δημιουργούς, καθώς η πρακτική αυτή επηρεάζει άμεσα τη δυνατότητα διασύνδεσης με βάσεις γνώσης.

Μόνο μικρό ποσοστό αρχαίων αγγείων φέρει υπογραφή. Ακόμη και όταν υπάρχει, η υπογραφή μπορεί να αφορά τον κεραμέα που κατασκεύασε το αγγείο, τον αγγειογράφο που το διακόσμησε, ή σπανιότερα και τους δύο. Η διάκριση δηλώνεται γραμματικά: το ρήμα έποίησεν παραπέμπει στην κατασκευή και το έγραψεν στη διακόσμηση.

Χαρακτηριστικό παράδειγμα από το ίδιο το δείγμα είναι η σχέση Makron και Ιέρωνα. Ο Ιέρων ήταν κεραμέας που υπέγραφε τα αγγεία του, ενώ ο Makron ήταν ο αγγειογράφος που τα διακοσμούσε. Η σύγχρονη έρευνα αποδίδει στον Makron εκατοντάδες αγγεία, παρότι ελάχιστα φέρουν τη δική του υπογραφή· η απόδοση στηρίζεται κυρίως στη στυλιστική ανάλυση και στη συστηματική συνύπαρξη με την υπογραφή του Ιέρωνα.

2.6.2 Η παράδοση Beazley και τα συμβατικά ονόματα

Το μεθοδολογικό πλαίσιο που κυριαρχεί στην απόδοση αττικών αγγείων διαμορφώθηκε από τον Βρετανό αρχαιολόγο Sir John Beazley στο πρώτο μισό του 20ού αιώνα. Ο Beazley ανέπτυξε συστηματική μέθοδο στυλιστικής ανάλυσης, εξετάζοντας λεπτομέρειες όπως η απόδοση του αυτιού, του ματιού ή των πτυχώσεων του ιματίου, και ταυτοποιώντας το χέρι μεμονωμένων τεχνιτών σε αγγεία χωρίς υπογραφή.

Για τους ανώνυμους τεχνίτες που εντόπισε, ο Beazley εισήγαγε την πρακτική των συμβατικών ονομάτων, γνωστών και ως notnames. Ένας τεχνίτης ονομάζεται από το χαρακτηριστικότερο έργο του, όπως ο Painter of the Paris Gigantomachy· από το μουσείο και τον αριθμό καταλόγου ενός αγγείου του, όπως ο Maler von London E 777· από εικονογραφικό μοτίβο, όπως ο Boot Painter· ή από τον τόπο εύρεσης, όπως ο Veii Painter.

Στο δείγμα συνυπάρχουν και οι δύο κατηγορίες. Τα ονόματα Makron, Onesimos, Douris και Hermonax αντιστοιχούν σε πραγματικές υπογραφές. Τα υπόλοιπα, που αποτελούν τη συντριπτική πλειονότητα των 125 μοναδικών τιμών, είναι συμβατικές αποδόσεις.

2.6.3 Οντολογικά ερωτήματα για τη διασύνδεση

Η διάκριση αυτή δεν είναι φιλολογική λεπτομέρεια αλλά θέτει ουσιώδη ερωτήματα για την αναπαράσταση σε βάση γνώσης. Ένα συμβατικό όνομα όπως Boot Painter δεν αναφέρεται σε ιστορικά τεκμηριωμένο πρόσωπο με γνωστό βίο, αλλά σε ερευνητική υπόθεση: το σύνολο των έργων που, κατά την κρίση των μελετητών, φιλοτεχνήθηκαν από το ίδιο χέρι.

Τα Wikidata αντιμετωπίζουν αυτές τις οντότητες ως instance of human, εφοδιάζοντάς τις με την ιδιότητα notname που δηλώνει ρητά τον συμβατικό χαρακτήρα της ονομασίας. Η επιλογή είναι πρακτική αλλά όχι αυτονόητη: ο Boot Painter είναι ταυτόχρονα υποθετικό πρόσωπο και κατηγορία έργων. Η αμφισημία ενδέχεται να συμβάλλει στη δυσκολία αυτόματης αντιστοίχισης που παρατηρήθηκε και αναλύεται στο Κεφάλαιο 4.

Πρόσθετο πρόβλημα δημιουργεί η μεταβλητότητα των συμβατικών ονομάτων στη διεθνή βιβλιογραφία. Ο ίδιος τεχνίτης μπορεί να αναφέρεται ως Maler von London E 777 στη γερμανική

παράδοση και ως Painter of London E 777 στην αγγλική. Το δείγμα περιέχει και τις δύο μορφές σε διαφορετικές εγγραφές, γεγονός που δυσχεραίνει τόσο τη συμφιλίωση όσο και την ενοποίηση.

Ιδιαίτερη περίπτωση αποτελεί η τιμή Brygan, με 12 εγγραφές. Το όνομα δεν αντιστοιχεί σε καθιερωμένη μορφή της βιβλιογραφίας και βρίσκεται εγγύτατα στο Brygos, τον κεραμέα του οποίου το εργαστήριο φιλοξένησε τον αποκαλούμενο Brygos Painter. Επειδή οι μελετητές θεωρούν σήμερα ότι κεραμέας και αγγειογράφος ήταν διαφορετικά πρόσωπα, και επειδή το δείγμα περιέχει ήδη ξεχωριστή τιμή Brygos Painter, η ασφαλής ταυτοποίηση της τιμής Brygan δεν είναι εφικτή. Ο χειρισμός της περίπτωσης συζητείται στο Κεφάλαιο 4.

2.7 Το Mapping προς τα Wikidata

Ο πίνακας που ακολουθεί συνοψίζει τη χαρτογράφηση των πεδίων του dataset προς τις ιδιότητες των Wikidata, όπως εφαρμόστηκε στην εργασία.

Πίνακας 2.7: Το mapping πεδίων dataset προς Wikidata properties

Πεδίο ή τιμή	Wikidata property	Κωδικός	Κατεύθυνση
Artist Display Name	creator	P170	Εισαγωγή στα Wikidata
Medium	material used	P186	Εισαγωγή στα Wikidata
Σταθερή τιμή Q1463658	instance of	P31	Εισαγωγή στα Wikidata
Title	label (en)	-	Εισαγωγή στα Wikidata
Σταθερή τιμή Q160236	stated in	P248	Τεκμηρίωση πηγής
Από τα Wikidata	country of citizenship	P27	Ανάκτηση στο dataset

Ο πίνακας αναδεικνύει τη διπλή κατεύθυνση της διασύνδεσης. Οι ιδιότητες P170, P186 και P31 αφορούν πληροφορία που μεταφέρεται από το dataset προς τα Wikidata. Η P248 τεκμηριώνει την προέλευση. Η P27, αντιθέτως, αφορά πληροφορία που ανακτάται από τα Wikidata και εμπλουτίζει το τοπικό dataset. Η αμφίδρομη αυτή ροή αποτελεί την ουσία της διασύνδεσης και τη διαφοροποιεί από μια απλή εξαγωγή δεδομένων.

Κεφάλαιο 3ο: Μεθοδολογία Διασύνδεσης με τα Wikidata

3.1 Εισαγωγή

Το παρόν κεφάλαιο αποτελεί τον πυρήνα της εργασίας. Παρουσιάζει αναλυτικά και βήμα προς βήμα τη μεθοδολογία που αναπτύχθηκε και εφαρμόστηκε για τη διασύνδεση του δείγματος με τα Wikidata. Η παρουσίαση είναι σκόπιμα λεπτομερής, ώστε να διασφαλίζεται η αναπαραγωγικότητα: κάθε ενέργεια, κάθε παράμετρος και κάθε απόφαση τεκμηριώνεται με τρόπο που επιτρέπει σε άλλον ερευνητή να επαναλάβει τη διαδικασία.

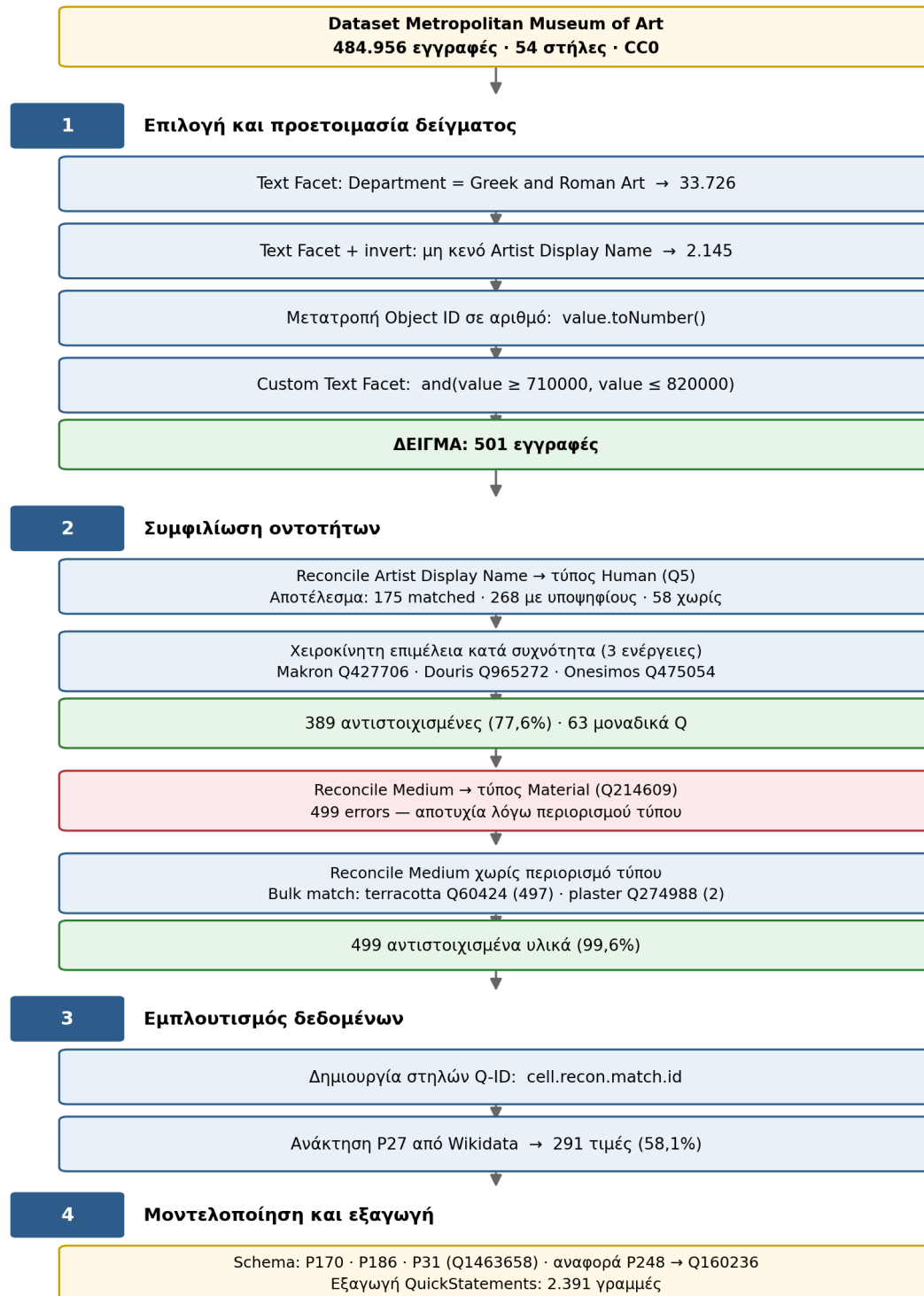
Η μεθοδολογία οργανώνεται σε τέσσερα διαδοχικά στάδια. Το πρώτο αφορά την επιλογή και προετοιμασία του δείγματος. Το δεύτερο, που αποτελεί την καρδιά της διαδικασίας, αφορά τη συμφιλίωση των οντοτήτων με τα αντίστοιχα items των Wikidata και συνδυάζει αυτόματη επεξεργασία με χειροκίνητη επιμέλεια. Το τρίτο αξιοποιεί τις συμφιλιωμένες οντότητες για τον εμπλουτισμό του dataset με πληροφορίες που δεν υπήρχαν στην πηγή. Το τέταρτο αφορά τη μοντελοποίηση των δεδομένων ως RDF και την παραγωγή εκτελέσιμων εντολών.

Το σύνολο των ενεργειών καταγράφηκε αυτόματα από το OpenRefine και εξήχθη ως αρχείο JSON, το οποίο αποτελεί ένα από τα τρία παραδοτέα και παρατίθεται πλήρες στο Παράρτημα Α. Το αρχείο περιέχει τις ενέργειες που εκτελέστηκαν, τόσο τις κύριες όσο και τις βοηθητικές. Ο πίνακας που ακολουθεί παρουσιάζει τις δώδεκα κύριες ενέργειες που συγκροτούν τη μεθοδολογία.

Πίνακας 3.1: Οι κύριες ενέργειες της μεθοδολογίας όπως καταγράφηκαν στο αρχείο JSON

#	Τύπος ενέργειας	Στήλη	Περιγραφή
1	core/text-transform	Object ID	Μετατροπή σε αριθμητικό τύπο
2	core/recon	Artist Display Name	Συμφιλίωση με τύπο Q5 (human)
3	core/recon-judge-similar-cells	Artist Display Name	Makron → Q427706 (172 κελιά)
4	core/recon-judge-similar-cells	Artist Display Name	Douris → Q965272 (21 κελιά)
5	core/recon-judge-similar-cells	Artist Display Name	Onesimos → Q475054 (21 κελιά)
6	core/column-addition	Artist Display Name	Δημιουργία Artist_Wikidata_ID
7	core/extend-reconciled-data	Artist Display Name	Ανάκτηση ιδιότητας P27
8	core/recon	Medium	Συμφιλίωση με τύπο Q214609
9	core/recon	Medium	Συμφιλίωση χωρίς περιορισμό τύπου
10	core/recon-judge-similar-cells	Medium	terracotta → Q60424 (497 κελιά)
11	core/recon-judge-similar-cells	Medium	plaster → Q274988 (2 κελιά)
12	core/column-addition	Medium	Δημιουργία Medium_Wikidata_ID

Το διάγραμμα που ακολουθεί αποτυπώνει συνολικά τη ροή εργασιών, από το αρχικό dataset έως το τελικό παραδοτέο, με τα ενδιάμεσα αποτελέσματα κάθε σταδίου.



Σχήμα 3.1: Η συνολική ροή εργασιών της μεθοδολογίας

3.1.1 Το περιβάλλον εργασίας

Η μεθοδολογία υλοποιήθηκε στο OpenRefine έκδοση 3.10.0, το οποίο εκτελείται τοπικά και προσπελάζεται μέσω φυλλομετρητή στη διεύθυνση 127.0.0.1:3333. Η αναφορά της έκδοσης δεν είναι διαδικαστική λεπτομέρεια: όπως θα φανεί στην ενότητα 3.5.4, η συγκεκριμένη έκδοση παρουσίασε σφάλμα στη λειτουργία εξαγωγής, γεγονός που επηρέασε τη διαδικασία.

Το OpenRefine λειτουργεί ως τοπικός διακομιστής, γεγονός που σημαίνει ότι τα δεδομένα παραμένουν στον υπολογιστή του χρήστη. Εξαίρεση αποτελούν τα αιτήματα συμφιλίωσης και ανάκτησης, που αποστέλλονται στο Wikidata Reconciliation Service και στο Wikidata API αντίστοιχα.

Αξίζει να σημειωθεί ένας πρακτικός περιορισμός που επηρέασε τη διαδικασία. Το OpenRefine εκτελείται πάνω στην εικονική μηχανή Java και εκκινεί με προκαθορισμένο ανώτατο όριο σωρού μνήμης (JVM maximum heap size) 1400 MB. Για dataset 484.956 γραμμών με 54 στήλες το όριο αυτό αποδείχθηκε ανεπαρκές: κατά την προσθήκη και αφαίρεση facets, τα οποία απαιτούν υπολογισμό μοναδικών τιμών σε ολόκληρο το σύνολο, το σύστημα εμφάνιζε παρατεταμένα διαστήματα αδράνειας.

Η αντιμετώπιση έγινε με αύξηση του ορίου σε 4096 MB. Στο macOS η παράμετρος ορίζεται στο αρχείο Info.plist που βρίσκεται εντός του πακέτου της εφαρμογής, στη διαδρομή OpenRefine.app/Contents/Info.plist. Η τιμή δηλώνεται ως όρισμα της εικονικής μηχανής στη μορφή -Xmx1400M και τροποποιήθηκε σε -Xmx4096M. Μετά την αλλαγή απαιτείται επανεκκίνηση της εφαρμογής, χωρίς απώλεια των αποθηκευμένων projects.

Η παρατήρηση είναι χρήσιμη για όποιον επιχειρήσει αντίστοιχη εργασία σε dataset παρόμοιου μεγέθους, καθώς το πρόβλημα δεν εκδηλώνεται ως σφάλμα αλλά ως φαινομενικό πάγωμα της εφαρμογής, γεγονός που δυσχεραίνει τη διάγνωσή του.

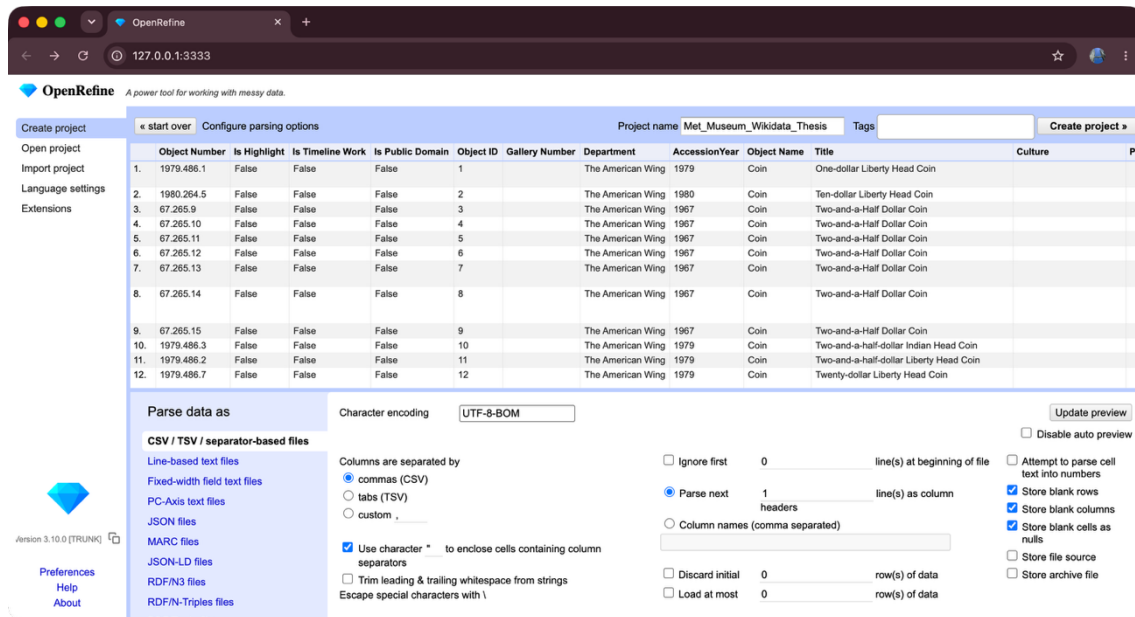
3.2 Στάδιο 1: Επιλογή και Προετοιμασία του Δείγματος

3.2.1 Εισαγωγή του dataset

Το πρώτο βήμα ήταν η εισαγωγή του CSV αρχείου στο OpenRefine μέσω της επιλογής Create Project. Το εργαλείο αναλύει αυτόματα το αρχείο και εμφανίζει προεπισκόπηση, επιτρέποντας την επιβεβαίωση των παραμέτρων πριν από την οριστική δημιουργία.

Η κωδικοποίηση αναγνωρίστηκε ως UTF-8-BOM. Το BOM, δηλαδή ο δείκτης σειράς byte, είναι ακολουθία χαρακτήρων που ορισμένα προγράμματα τοποθετούν στην αρχή του αρχείου για να δηλώσουν την κωδικοποίηση. Η ορθή αναγνώρισή του από το OpenRefine αποτρέπει την εμφάνιση παράσιτων χαρακτήρων στην πρώτη επικεφαλίδα στήλης, πρόβλημα συχνό όταν η κωδικοποίηση ορίζεται λανθασμένα ως σκέτο UTF-8.

Ως διαχωριστικό ορίστηκε το κόμμα και ενεργοποιήθηκε η ανάγνωση της πρώτης γραμμής ως επικεφαλίδες. Κρίσιμη υπήρξε η απενεργοποίηση της επιλογής «Attempt to parse cell text into numbers». Αν παρέμενε ενεργή, η στήλη Object ID θα εισαγόταν ήδη ως αριθμητική και το πρώτο βήμα της μεθοδολογίας θα ήταν περιττό· σημαντικότερα όμως, η αυτόματη μετατροπή σε datasets με ανομοιόμορφα δεδομένα οδηγεί συχνά σε σφάλματα που εντοπίζονται δύσκολα.



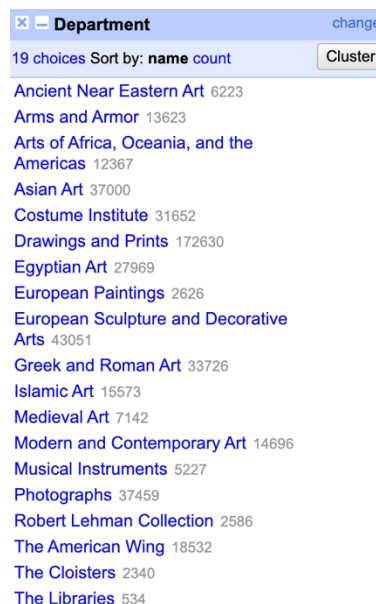
Σχήμα 3.2: Η οθόνη εισαγωγής με τις παραμέτρους κωδικοποίησης και διαχωριστικού

Μετά τη δημιουργία του project, το OpenRefine εμφάνισε 484.956 γραμμές και 54 στήλες.

3.2.2 Πρώτο φίλτρο: θεματική επιλογή

Εφαρμόστηκε Text Facet στη στήλη Department μέσω της διαδρομής βέλος στήλης, Facet, Text facet. Το OpenRefine εμφάνισε στο αριστερό πλαίσιο τη λίστα των δεκαεννέα τμημάτων με τα πλήη τους.

Επιλέχθηκε το Greek and Roman Art, το οποίο περιορίζει το σύνολο σε 33.726 εγγραφές. Το σκεπτικό της επιλογής αναλύθηκε στο Κεφάλαιο 2.



Σχήμα 3.3: Το Text Facet με τα δεκαεννέα τμήματα του μουσείου

Σχήμα 3.4: Το αποτέλεσμα του πρώτου φίλτρου - 33.726 από 484.956 εγγραφές

3.2.3 Δεύτερο φίλτρο: αποκλεισμός ανώνυμων έργων

Το δεύτερο φίλτρο εφαρμόστηκε στη στήλη Artist Display Name. Εφαρμόστηκε Text Facet και εντοπίστηκε η ειδική κατηγορία blank, η οποία αντιστοιχεί σε 31.581 εγγραφές.

Αντί για διαγραφή των εγγραφών αυτών, χρησιμοποιήθηκε η λειτουργία invert: επιλέχθηκε η κατηγορία blank και στη συνέχεια πατήθηκε το invert στην επάνω γραμμή του facet, ώστε να εμφανιστούν όλες οι εγγραφές εκτός των κενών.

Η επιλογή αυτή είναι μεθοδολογικά προτιμότερη από τη διαγραφή για δύο λόγους. Δεν αλλοιώνει τα δεδομένα του project, καθώς οι εγγραφές παραμένουν και απλώς αποκρύπτονται. Και παραμένει πλήρως αναστρέψιμη, χωρίς να επιβαρύνει το ιστορικό με μη αναιρέσιμη ενέργεια διαγραφής.

Μετά την εφαρμογή του φίλτρου απέμειναν 2.145 εγγραφές, δηλαδή το 6,4% του τμήματος.

Σχήμα 3.5: Το δεύτερο φίλτρο - 2.145 εγγραφές με επώνυμο δημιουργό

3.2.4 Μετατροπή του Object ID σε αριθμητικό τύπο

Η στήλη Object ID περιέχει μοναδικούς αριθμητικούς κωδικούς, αλλά μετά την εισαγωγή αναγνωρίζεται ως κείμενο. Για να καταστούν δυνατές οι αριθμητικές συγκρίσεις, μετατράπηκε μέσω της διαδρομής Edit cells, Common transforms, To number. Η ενέργεια καταγράφηκε ως η πρώτη του JSON με την έκφραση:

```
value.toNumber()
```

Το JSON αναφέρει ότι η μετατροπή εφαρμόστηκε σε 2.145 κελιά, δηλαδή μόνο στις ορατές μετά τα δύο φίλτρα εγγραφές. Αυτό αναδεικνύει μια σημαντική ιδιότητα του OpenRefine: κάθε ενέργεια εφαρμόζεται αποκλειστικά στο φιλτραρισμένο υποσύνολο, και η διαμόρφωση των φίλτρων καταγράφεται στο πεδίο engineConfig της κάθε ενέργειας.

Μετά τη μετατροπή, οι αριθμητικές τιμές εμφανίζονται με πράσινο χρώμα και δεξιά στοίχιση, σε αντίθεση με τις τιμές κειμένου που εμφανίζονται με μαύρο και αριστερή στοίχιση. Η οπτική αυτή διάκριση επιτρέπει άμεση επιβεβαίωση.

3.2.5 Τρίτο φίλτρο: αριθμητικός περιορισμός

Το τρίτο φίλτρο περιορίζει το δείγμα σε διαχειρίσιμο μέγεθος. Αρχικά επιχειρήθηκε μέσω Numeric Facet, το οποίο εμφανίζει ιστόγραμμα κατανομής με δύο συρόμενους δείκτες.

Η προσέγγιση αυτή παρουσίασε ουσιώδες πρόβλημα αναπαραγωγιμότητας. Ο ακριβής προσδιορισμός των ορίων μέσω συρόμενου δείκτη εξαρτάται από τη θέση του ποντικιού και είναι πρακτικά αδύνατο να επαναληφθεί. Κατά την επανάληψη της διαδικασίας για τους σκοπούς της τεκμηρίωσης διαπιστώθηκε απόκλιση στον αριθμό των επιλεγμένων εγγραφών σε σχέση με προηγούμενη εκτέλεση.

Για την αντιμετώπιση υιοθετήθηκε εναλλακτική προσέγγιση: αντί για Numeric Facet χρησιμοποιήθηκε Custom Text Facet με ρητή λογική συνθήκη. Η έκφραση GREL που ορίστηκε είναι:

```
and(value >= 710000, value <= 820000)
```

Η συνάρτηση and δέχεται δύο λογικές συνθήκες και επιστρέφει true μόνο όταν ισχύουν αμφότερες. Αξίζει να σημειωθεί ότι η GREL δεν υποστηρίζει τον τελεστή && που χρησιμοποιείται σε άλλες γλώσσες προγραμματισμού· η απόπειρα χρήσης του παράγει σφάλμα σύνταξης.

Το facet που προκύπτει εμφανίζει δύο μόνο επιλογές, true και false. Η επιλογή του true απομονώνει τις εγγραφές που πληρούν τη συνθήκη, με απόλυτη ακρίβεια και πλήρη αναπαραγωγιμότητα. Το αποτέλεσμα είναι 501 εγγραφές.

Η αντικατάσταση αυτή συνιστά μεθοδολογική βελτίωση που αξίζει γενίκευσης: όπου απαιτείται αριθμητικό φίλτρο με ακριβή όρια, το Custom Text Facet με ρητή συνθήκη είναι προτιμότερο του Numeric Facet.

OpenRefine Met_Museum_Wikidata_Thesis

501 matching rows (484,956 total)

Facet / Filter: Department, Artist Display Name, Object ID

Object Number	Is Highlight	Is Timeline Work	Is Public Domain	Object ID	Gallery Number	Department	AccessionYear	Object Name	Title
413989	False	False	True	710000		Greek and Roman Art	2011	Kylix fragment	Terracotta fragn (cup)
413990	False	False	True	710001		Greek and Roman Art	2011	Kylix fragment	Terracotta rim fr (drinking cup)
413991	False	False	True	710002		Greek and Roman Art	2011	Kylix fragment	Terracotta fragn (cup)
413992	False	False	True	710003		Greek and Roman Art	2011	Kylix fragment	Terracotta fragn (cup)
413993	False	False	True	710004		Greek and Roman Art	2011	Kylix fragment	Terracotta fragn (cup)
413994	False	False	True	710005		Greek and Roman Art	2011	Kylix fragment	Terracotta fragn (cup)
413995	False	False	True	710006		Greek and Roman Art	2011	Kylix fragment	Terracotta fragn (cup)
413996	False	False	True	710007		Greek and Roman Art	2011	Kylix fragment	Terracotta fragn (cup)
413997	False	False	True	710008		Greek and Roman Art	2011	Kylix fragment	Terracotta fragn (cup)
413998	False	False	True	710009		Greek and Roman Art	2011	Kylix fragment	Terracotta fragn (cup)

Σχήμα 3.6: Το τρίτο φίλτρο - 501 εγγραφές

Πίνακας 3.2: Η διαδοχική εφαρμογή των τριών φίλτρων

Φίλτρο	Τύπος	Κριτήριο	Εγγραφές	Μείωση
-	-	Πλήρες dataset	484.956	-
1	Text Facet	Greek and Roman Art	33.726	-93,0%
2	Text Facet invert	Μη κενό όνομα	2.145	-93,6%
3	Custom Text Facet	Object ID 710000–820000	501	-76,6%

3.3 Στάδιο 2: Συμφιλίωση Οντοτήτων

3.3.1 Η έννοια και ο μηχανισμός

Η συμφιλίωση αποτελεί τη διαδικασία αντιστοίχισης τιμών ενός τοπικού dataset με οντότητες εξωτερικής βάσης γνώσης. Πρόκειται για μη τετριμμένη διαδικασία, διότι δεν συνιστά απλή σύγκριση συμβολοσειρών: το ίδιο πρόσωπο μπορεί να αναφέρεται με διαφορετικά ονόματα, ενώ διαφορετικά πρόσωπα μπορεί να μοιράζονται το ίδιο όνομα.

Το Wikidata Reconciliation Service, προσβάσιμο στη διεύθυνση wikidata.reconci.link, λειτουργεί ως εξής. Για κάθε τιμή αναζητά items των οποίων η ετικέτα ή κάποιο ψευδώνυμο ταιριάζει. Για κάθε υποψήφιο υπολογίζεται βαθμός βεβαιότητας. Αν υπάρχει ένας μόνο υποψήφιος με πολύ υψηλό score και σημαντική διαφορά από τους υπόλοιπους, γίνεται αυτόματη αντιστοίχιση. Διαφορετικά το κελί παραμένει σε κατάσταση αναμονής.

Η ορολογία του judgment facet απαιτεί διευκρίνιση, καθώς συχνά παρανοείται. Διακρίνονται τρεις καταστάσεις: *matched*, όταν έχει επιλεγεί συγκεκριμένο item· *new*, όταν έχει σημειωθεί ότι πρέπει να δημιουργηθεί νέο· και *none*, όταν δεν έχει ληφθεί απόφαση. Η κατάσταση *none* δεν σημαίνει απαραίτητως ασάφεια - μπορεί να σημαίνει είτε ότι υπήρχαν πολλοί ισοδύναμοι υποψήφιοι, είτε ότι δεν βρέθηκε κανένας. Η διάκριση αυτή αποδείχθηκε ουσιώδης, όπως θα φανεί.

3.3.2 Αυτόματη συμφιλίωση των δημιουργών

Η διαδικασία ξεκίνησε από τη διαδρομή Reconcile, Start reconciling στη στήλη Artist Display Name. Ως υπηρεσία επιλέχθηκε το Wikidata endpoint και ως τύπος οντότητας ο Human (Q5).

Ο περιορισμός σε τύπο Q5 στοχεύει στον αποκλεισμό αντιστοιχίσεων με οντότητες που δεν είναι φυσικά πρόσωπα. Χωρίς αυτόν, ονόματα όπως Veii Painter θα μπορούσαν να αντιστοιχιστούν με τοπωνύμια. Όπως θα φανεί όμως στη συνέχεια, ο περιορισμός τύπου συνέβαλε και σε ανεπιτυχείς αντιστοιχίσεις.

Reconcile column Artist Display Name

Reconcile each cell to an entity of one of these types:

- ☒ human
Q5
- ☐ business
Q4830453
- ☐ bowl
Q153988
- ☐ glassworks
Q1179791
- ☐ scholarly article
Q13442814
- ☐ girandole
Q1326509
- ☐ vase
Q191851
- ☐ encyclopedia article
Q13433827

☐ Reconcile against type:

☐ Reconcile against no particular type

☒ Auto-match candidates with high confidence

Maximum number of candidates to return

Also use relevant details from other columns:

Column	As property
Object Number	
Is Highlight	
Is Timeline Work	
Is Public Domain	
Object ID	
Gallery Number	
Department	
AccessionYear	
Object Name	
Title	
Culture	
Period	

Back
Start reconciling... Cancel

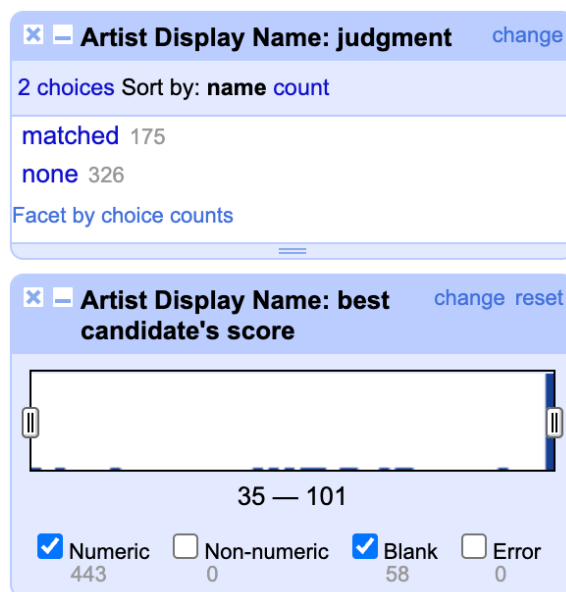
Σχήμα 3.7: Η ρύθμιση συμφιλίωσης με τύπο Human (Q5)

Μετά την ολοκλήρωση, το OpenRefine εμφάνισε δύο facets. Το judgment facet έδειξε 175 matched και 326 none. Το facet best candidate's score έδειξε 443 numeric και 58 blank.

Ο συνδυασμός των δύο facets επιτρέπει τριμερή ανάλυση που δεν θα ήταν εφικτή από το judgment facet μόνο. Οι 58 εγγραφές με κενό score είναι εκείνες για τις οποίες η υπηρεσία δεν επέστρεψε κανέναν υποψήφιο. Οι υπόλοιπες 443 είχαν τουλάχιστον έναν υποψήφιο· από αυτές οι 175 αντιστοιχίστηκαν αυτόματα και οι 268 παρέμειναν χωρίς απόφαση.

Πίνακας 3.3: Τριμερής ανάλυση των αποτελεσμάτων αυτόματης συμφιλίωσης

Κατηγορία	Εγγραφές	Ποσοστό	Ερμηνεία
Αυτόματη αντιστοίχιση	175	34,9%	Υψηλή βεβαιότητα
Υποψήφιοι χωρίς απόφαση	268	53,5%	Ανεπαρκής βεβαιότητα
Κανένας υποψήφιος	58	11,6%	Πιθανή απουσία από τα Wikidata
Σύνολο	501	100%	



Σχήμα 3.8: Τα δύο facets μετά την αυτόματη συμφιλίωση

Η διάκριση αυτή έχει πρακτική σημασία για τη στρατηγική επιμέλειας. Οι 268 εγγραφές με υποψηφίους είναι εκείνες όπου η ανθρώπινη κρίση μπορεί να προσθέσει αξία, καθώς η πληροφορία υπάρχει και απλώς χρειάζεται επιβεβαίωση. Οι 58 χωρίς υποψηφίους απαιτούν διαφορετική αντιμετώπιση, πιθανώς αναζήτηση με εναλλακτικές μορφές του ονόματος ή δημιουργία νέων items.

3.3.3 Στρατηγική χειροκίνητης επιμέλειας

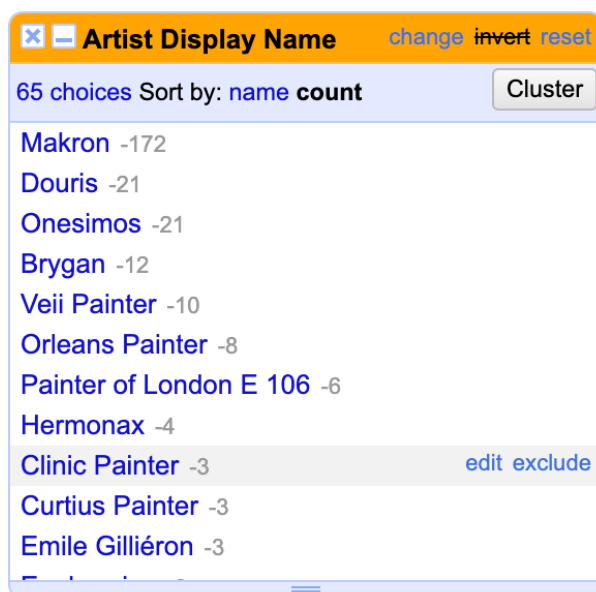
Η χειροκίνητη εξέταση και των 326 μη αντιστοιχισμένων εγγραφών θα ήταν χρονοβόρα. Αντ' αυτού υιοθετήθηκε στρατηγική προτεραιοποίησης βάσει συχνότητας, η οποία στηρίζεται σε μια απλή παρατήρηση: επειδή το OpenRefine επιτρέπει μαζική αντιστοίχιση όλων των κελιών με την ίδια τιμή, το όφελος κάθε ενέργειας είναι ανάλογο της συχνότητας του ονόματος.

Εφαρμόστηκε Text Facet στη στήλη Artist Display Name, περιορισμένο στις μη αντιστοιχισμένες εγγραφές, με ταξινόμηση κατά φθίνουσα συχνότητα. Το facet εμφάνισε 65 μοναδικά ονόματα με έντονα ασύμμετρη κατανομή.

Πίνακας 3.4: Τα δέκα συχνότερα μη αντιστοιχισμένα ονόματα και η σωρευτική κάλυψη

#	Όνομα	Εγγραφές	Σωρευτικά	Συνολικό ποσοστό
1	Makron	172	172	69,3%
2	Douris	21	193	73,5%
3	Onesimos	21	214	77,6%
4	Brygan	12	226	80,0%
5	Veii Painter	10	236	82,0%
6	Orleans Painter	8	244	83,6%
7	Painter of London E 106	6	250	84,8%
8	Hermonax	4	254	85,6%

9	Clinic Painter	3	257	86,2%
10	Curtius Painter	3	260	86,8%



Σχήμα 3.9: Η κατανομή των μη αντιστοιχισμένων ονομάτων

Η κατανομή δικαιολογεί πλήρως τη στρατηγική. Τα τρία πρώτα ονόματα καλύπτουν 214 από τις 326 εγγραφές, δηλαδή το 65,6%. Τα υπόλοιπα 62 ονόματα καλύπτουν συνολικά 112 εγγραφές, με μέση απόδοση 1,8 εγγραφές ανά ενέργεια έναντι 71 για τα τρία πρώτα.

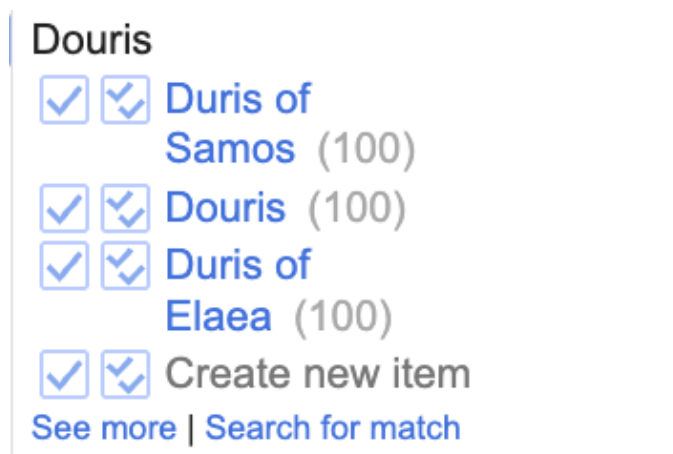
3.3.4 Η εκτέλεση της χειροκίνητης επιμέλειας

Για κάθε όνομα η διαδικασία ήταν η ίδια. Επιλογή του ονόματος στο facet ώστε να απομονωθούν οι γραμμές του, εξέταση των υποψηφίων που εμφανίζει το πρώτο κελί, επαλήθευση της καταλληλότητας, και εφαρμογή μέσω του εικονιδίου διπλού ελέγχου που αντιστοιχίζει όλα τα κελιά με την ίδια τιμή.

Τα κριτήρια επαλήθευσης που εφαρμόστηκαν συστηματικά ήταν τρία. Η επαγγελματική ιδιότητα του υποψήφιου item, με απαίτηση να σχετίζεται με την αγγειογραφία. Η χρονολογική συμβατότητα με την περίοδο του δείγματος. Και η απουσία ρητών προειδοποιήσεων σύγχυσης στο ίδιο το Wikidata item.

Πίνακας 3.5: Οι τρεις ενέργειες χειροκίνητης επιμέλειας

Όνομα	Wikidata item	Q-number	Εγγραφές
Makron	Makron	Q427706	172
Douris	Douris	Q965272	21
Onesimos	Onesimos	Q475054	21
Σύνολο			214



Σχήμα 3.10: Η διεπαφή Bulk Match με τους υποψηφίους

Ιδιαίτερη σημασία έχει η περίπτωση του Makron. Ο έλεγχος έδειξε ότι το ορθό item, Q427706, εμφανιζόταν κανονικά στη λίστα των υποψηφίων που είχε επιστρέψει η αυτόματη διαδικασία. Η υπηρεσία δηλαδή είχε εντοπίσει τη σωστή οντότητα αλλά δεν είχε επαρκή βεβαιότητα για αυτόματη επιλογή. Το εύρημα αναλύεται διεξοδικά στο Κεφάλαιο 4.

Στην περίπτωση του Douris, η αιτία της διστακτικότητας είναι ρητά τεκμηριωμένη μέσα στα ίδια τα Wikidata: το item Q965272 φέρει δήλωση ότι δεν πρέπει να συγχέεται με τον Duris, ιστορικό συγγραφέα με σχεδόν ταυτόσημο όνομα. Η ύπαρξη ομώνυμης οντότητας εξηγεί γιατί κανέναν υποψήφιο δεν συγκέντρωσε αρκετά υψηλό score.

3.3.5 Οι περιπτώσεις όπου η αντιστοίχιση δεν έγινε

Δύο περιπτώσεις από τη λίστα προτεραιότητας εξετάστηκαν αλλά δεν αντιστοιχίστηκαν, και η τεκμηρίωση των λόγων έχει μεθοδολογική αξία.

Η τιμή Brygan, με 12 εγγραφές, εμφάνιζε ως μοναδικό υποψήφιο τον Brygos Painter (Q322192). Η αντιστοίχιση απορρίφθηκε για δύο λόγους. Πρώτον, η σύγχρονη έρευνα διακρίνει τον Βρύγο, που ήταν κεραμέας και υπέγραφε τα αγγεία του, από τον αποκαλούμενο Brygos Painter, που ήταν ο αγγειογράφος του εργαστηρίου του· πρόκειται κατά πάσα πιθανότητα για διαφορετικά πρόσωπα. Δεύτερον, το δείγμα περιέχει ήδη ξεχωριστή τιμή Brygos Painter με 10 εγγραφές, η οποία είχε αντιστοιχιστεί αυτόματα στο Q322192. Η αντιστοίχιση και της τιμής Brygan στο ίδιο item θα συγχώνευε δύο διακριτές τιμές του dataset σε μία οντότητα, εισάγοντας πιθανό σφάλμα.

Η τιμή Veii Painter, με 10 εγγραφές, δεν εμφάνισε κανέναν υποψήφιο. Πρόκειται για συμβατικό όνομα που προφανώς δεν έχει καταχωριστεί στα Wikidata, τουλάχιστον όχι με αυτή τη μορφή.

Οι δύο περιπτώσεις αντιπροσωπεύουν διαφορετικούς τύπους ορίου. Η πρώτη είναι όριο γνώσης: τα δεδομένα δεν επαρκούν για ασφαλή απόφαση. Η δεύτερη είναι όριο κάλυψης: η οντότητα απλώς δεν υπάρχει στη βάση. Και στις δύο, η ορθή απόφαση ήταν η αποχή. Η χειροκίνητη επιμέλεια δεν σημαίνει μεγιστοποίηση των αντιστοιχίσεων αλλά ορθή κρίση, η οποία περιλαμβάνει και την άρνηση αντιστοίχισης.

3.3.6 Το τελικό αποτέλεσμα

Μετά τις τρεις ενέργειες χειροκίνητης επιμέλειας, το judgment facet έδειξε 389 matched και 112 none.

Πίνακας 3.6: Η επίδραση της χειροκίνητης επιμέλειας

Στάδιο	Αντιστοιχισμένες	Ποσοστό	Μεταβολή
Μετά την αυτόματη συμφιλίωση	175	34,9%	-
Μετά τη χειροκίνητη επιμέλεια	389	77,6%	+42,7 μονάδες

Οι 389 αντιστοιχισμένες εγγραφές αντιστοιχούν σε 63 μοναδικά Wikidata items. Το δείγμα περιλαμβάνει συνολικά 125 μοναδικά ονόματα, εκ των οποίων τα 63 αντιστοιχίστηκαν και τα 62 παρέμειναν χωρίς αντιστοίχιση.

3.3.7 Δημιουργία της στήλης Artist_Wikidata_ID

Μετά τη συμφιλίωση, οι αντιστοιχίσεις υπάρχουν ως μεταδεδομένα των κελιών και όχι ως ορατές τιμές. Για να αξιοποιηθούν και να εξαχθούν, μετατράπηκαν σε κανονική στήλη μέσω της λειτουργίας Add column based on this column με την έκφραση:

```
cell.recon.match.id
```

Η έκφραση προσπελαίνει το αντικείμενο recon του κελιού, στη συνέχεια το πεδίο match που αντιστοιχεί στην επιλεγμένη αντιστοίχιση, και τέλος το πεδίο id που περιέχει το Q-number. Για κελιά χωρίς αντιστοίχιση επιστρέφει κενή τιμή. Το JSON καταγράφει ότι η στήλη συμπληρώθηκε σε 389 γραμμές.

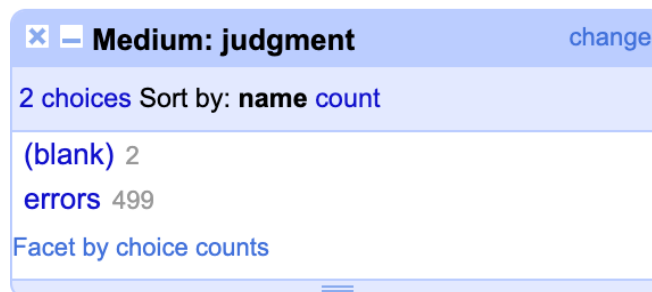
Αξίζει να διευκρινιστεί ο ρόλος αυτής της στήλης. Δεν χρησιμοποιείται από το Wikibase Schema, το οποίο αντλεί απευθείας από τα δεδομένα συμφιλίωσης της αρχικής στήλης. Η χρησιμότητά της είναι διαφορετική: καθιστά τα Q-numbers ορατά και εξαγωγίμα στο CSV, ώστε το παραδοτέο να είναι αυτοτελές και αναγνώσιμο χωρίς πρόσβαση στο OpenRefine.

3.4 Στάδιο 3: Συμφιλίωση Υλικών και Εμπλουτισμός

3.4.1 Η αποτυχία της πρώτης απόπειρας

Η συμφιλίωση της στήλης Medium ξεκίνησε με περιορισμό τύπου Material (Q214609), επιλογή που φαίνεται προφανής δεδομένου ότι το πεδίο περιγράφει υλικά κατασκευής.

Το αποτέλεσμα ήταν πλήρης αποτυχία. Το judgment facet εμφάνισε 499 errors και 2 blank. Δεν πρόκειται για αδυναμία εύρεσης αντιστοιχίσεων αλλά για σφάλμα εκτέλεσης: η υπηρεσία δεν μπόρεσε να επεξεργαστεί τα αιτήματα καθόλου.



Σχήμα 3.11: Η αποτυχία της πρώτης απόπειρας συμφιλίωσης

Το εύρημα είναι διδακτικό και αντιβαίνει στη διαίσθηση. Ο περιορισμός τύπου, που θεωρείται καλή πρακτική για τη βελτίωση της ακρίβειας, οδήγησε εδώ σε καθολική αποτυχία. Η πιθανή εξήγηση είναι ότι το reconciliation service, όταν του ζητείται να περιορίσει την αναζήτηση σε συγκεκριμένο τύπο,

εκτελεί πρόσθετο έλεγχο ο οποίος για τον τύπο Q214609 απέτυχε - είτε λόγω προσωρινού προβλήματος της υπηρεσίας είτε λόγω ιδιαιτερότητας του συγκεκριμένου τύπου.

3.4.2 Η δεύτερη απόπειρα

Η διαδικασία επαναλήφθηκε με την επιλογή `reconcile against no particular type`. Το judgment facet εμφάνισε τώρα 499 none και 2 blank.

Η αλλαγή από errors σε none είναι ουσιώδης. Σημαίνει ότι η υπηρεσία επεξεργάστηκε κανονικά τα αιτήματα και απλώς δεν έλαβε αυτόματες αποφάσεις. Η διαδικασία δηλαδή λειτούργησε, και το επόμενο βήμα της χειροκίνητης επιβεβαίωσης κατέστη εφικτό.



Σχήμα 3.12: Η δεύτερη απόπειρα χωρίς περιορισμό τύπου

Η αντιπαράθεση των δύο αποπειρών, με τα ίδια ακριβώς δεδομένα και μοναδική διαφορά τον περιορισμό τύπου, συνιστά ουσιαστικά ένα ελεγχόμενο πείραμα. Το συμπέρασμα είναι ότι οι περιορισμοί τύπου πρέπει να χρησιμοποιούνται με επιφύλαξη και να ελέγχεται πάντοτε το ενδεχόμενο να αποκλείουν έγκυρα αποτελέσματα.

3.4.3 Μαζική αντιστοίχιση των υλικών

Η ομοιογένεια του πεδίου Medium στο συγκεκριμένο δείγμα κατέστησε τη χειροκίνητη επιμέλεια εξαιρετικά αποδοτική. Το Text Facet αποκάλυψε μόλις δύο μοναδικές τιμές: terracotta με 497 εγγραφές και plaster με 2, συν 2 κενές.

Εκτελέστηκαν δύο ενέργειες μαζικής αντιστοίχισης. Η πρώτη αντιστόιχε το terracotta στο Q60424 και η δεύτερη το plaster στο Q274988. Το τελικό αποτέλεσμα ήταν 499 αντιστοιχισμένες εγγραφές, ποσοστό 99,6%.

Πίνακας 3.7: Αποτελέσματα συμπίλωσης της στήλης Medium

Υλικό	Εγγραφές	Ποσοστό	Wikidata item	Q-number
terracotta	497	99,2%	terracotta	Q60424
plaster	2	0,4%	plaster	Q274988
χωρίς τιμή	2	0,4%	-	-

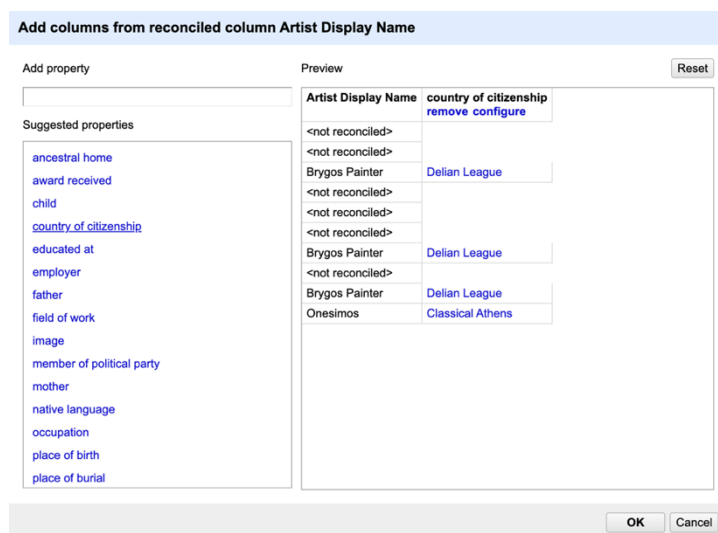
Παρατήρηση τεχνικού χαρακτήρα: το JSON καταγράφει ότι η μαζική αντιστοίχιση εφαρμόστηκε σε κελιά που περιέχουν Terracotta με κεφαλαίο αρχικό, ενώ οι τιμές στο dataset είναι πεζές. Το OpenRefine εκτελεί δηλαδή τη μαζική αντιστοίχιση χωρίς διάκριση πεζών και κεφαλαίων, στοιχείο χρήσιμο για όποιον σχεδιάζει αντίστοιχες διαδικασίες.

Στη συνέχεια δημιουργήθηκε η στήλη Medium_Wikidata_ID με την ίδια έκφραση `cell.recon.match.id`, η οποία συμπληρώθηκε σε 499 γραμμές.

3.4.4 Ανάκτηση της εθνικότητας

Το πλέον χαρακτηριστικό βήμα του σταδίου εμπλουτισμού είναι η ανάκτηση πληροφορίας που δεν υπάρχει στο αρχικό dataset. Η λειτουργία Add columns from reconciled values αξιοποιεί τις συμφιλιωμένες οντότητες για να ανακτήσει οποιαδήποτε ιδιότητά τους από τα Wikidata.

Επιλέχθηκε η ιδιότητα country of citizenship (P27). Το OpenRefine εκτέλεσε αιτήματα στο Wikidata API για κάθε συμφιλιωμένο item και τοποθέτησε τα αποτελέσματα σε νέα στήλη. Η ενέργεια καταγράφεται στο JSON ως core/extend-reconciled-data.



Σχήμα 3.13: Η ανάκτηση της ιδιότητας P27 από τα Wikidata

Η ανάκτηση απέδωσε τιμή σε 291 από τις 501 εγγραφές, ποσοστό 58,1%. Η σύγκριση με το ποσοστό των αντιστοιχισμένων εγγραφών είναι διαφωτιστική: από τις 389 με αντιστοιχισμένο δημιουργό, οι 291 απέκτησαν τιμή P27 και οι 98 όχι, επειδή τα αντίστοιχα Wikidata items δεν φέρουν τη συγκεκριμένη δήλωση.

Πίνακας 3.8: Οι τιμές της ιδιότητας P27 που ανακτήθηκαν

Τιμή	Εγγραφές	Ποσοστό	Τύπος οντότητας
Classical Athens	264	52,7%	Αρχαία πόλη-κράτος
Greece	13	2,6%	Σύγχρονο κράτος
Delian League	12	2,4%	Αρχαία συμμαχία
Sparta	1	0,2%	Αρχαία πόλη-κράτος
Apulia	1	0,2%	Γεωγραφική περιοχή
χωρίς τιμή	210	41,9%	-

Οι ίδιες οι τιμές που ανακτήθηκαν αποτελούν αντικείμενο κριτικής ανάλυσης στο Κεφάλαιο 4, καθώς αναδεικνύουν σημασιολογικό πρόβλημα στην εφαρμογή της ιδιότητας P27 σε πρόσωπα της αρχαιότητας.

3.5 Στάδιο 4: Μοντελοποίηση Σχήματος και Εξαγωγή

3.5.1 Η αρχιτεκτονική του Wikibase Schema

Το τελευταίο στάδιο αφορά τη μετατροπή του εμπλουτισμένου dataset σε μορφή εισαγώγιμη στα Wikidata. Το OpenRefine παρέχει την επέκταση Wikibase, η οποία προσπελαύνεται από το ομώνυμο μενού και ανοίγει ειδική καρτέλα Schema.

Η λογική του schema είναι δηλωτική. Ο χρήστης περιγράφει τη δομή των δηλώσεων που επιθυμεί, χρησιμοποιώντας τις στήλες ως μεταβλητές, και το OpenRefine εφαρμόζει τη δομή σε κάθε γραμμή. Η αρχιτεκτονική είναι ιεραρχική: στο ανώτερο επίπεδο ορίζεται το item, στο επόμενο οι δηλώσεις, και στο τρίτο οι qualifiers και οι references κάθε δήλωσης.

3.5.2 Ο υπό συνθήκη ορισμός του υποκειμένου

Το πρώτο και δυσκολότερο ζήτημα αφορά τον ορισμό του υποκειμένου, δηλαδή του item στο οποίο αναφέρονται οι δηλώσεις. Οι ιδιότητες P170 και P186 αφορούν τα ίδια τα αγγεία, όχι τους δημιουργούς ή τα υλικά.

Ο έλεγχος της στήλης Object Wikidata URL έδειξε ότι 15 από τις 501 εγγραφές αντιστοιχούν σε αντικείμενα που υπάρχουν ήδη ως Wikidata items, με κωδικούς στο εύρος Q116234941 έως Q116445024. Για τις υπόλοιπες 486 απαιτείται δημιουργία νέων.

Η διάκριση αυτή είναι ουσιώδης. Η δημιουργία νέου item για αντικείμενο που ήδη υπάρχει θα παρήγαγε διπλοεγγραφή, πρόβλημα που η κοινότητα των Wikidata αντιμετωπίζει με ιδιαίτερη αυστηρότητα καθώς η συγχώνευση διπλοτύπων απαιτεί χειροκίνητη παρέμβαση.

Για την αντιμετώπιση δημιουργήθηκε βοηθητική στήλη με ονομασία SubjectQ και την ακόλουθη έκφραση:

```
if(isBlank(cells["Object Wikidata URL"].value),
    "NEW_" + cells["Object ID"].value,
    cells["Object Wikidata URL"].value.split("/")[-1])
```

Η έκφραση εξάγει το Q-number από τον σύνδεσμο όπου αυτός υπάρχει, και παράγει μοναδικό δείκτη με πρόθεμα NEW_ στις υπόλοιπες περιπτώσεις. Το πρόθεμα συνδυάζεται με το Object ID ώστε κάθε τιμή να είναι μοναδική και να μη συγχωνευθούν εγγραφές.

Η στήλη υποβλήθηκε στη συνέχεια σε δύο διακριτές ενέργειες συμφιλίωσης. Για τα 15 κελιά που περιέχουν Q-number εφαρμόστηκε η λειτουργία Use values as identifiers, η οποία τα σημαίνει ως αντιστοιχισμένα στα υπάρχοντα items. Για τα 486 κελιά με πρόθεμα NEW_ εφαρμόστηκε η ενέργεια Create a new item for each cell από το υπομενού Actions.

Η στήλη SubjectQ ορίστηκε έπειτα ως υποκείμενο του σχήματος. Το OpenRefine αναγνωρίζει αυτόματα τη διαφορά: για τα συμφιλιωμένα κελιά χρησιμοποιεί το υπάρχον Q-number, ενώ για τα σημασμένα ως νέα παράγει εντολή CREATE.

3.5.3 Οι δηλώσεις, οι ιδιότητες ταυτοποίησης και οι αναφορές

Το τελικό σχήμα περιλαμβάνει μία ετικέτα και έξι δηλώσεις για κάθε εγγραφή.

Ετικέτα. Label στα αγγλικά με τιμή από τη στήλη Title. Ένα Wikidata item χωρίς ετικέτα δεν γίνεται αποδεκτό. Για τα 15 υπάρχοντα items η ετικέτα δηλώνεται με την επιλογή do not override, ώστε να μην αντικατασταθεί η υπάρχουσα.

Κεφάλαιο 3

P170 creator. Τιμή από τη συμφιλιωμένη στήλη Artist Display Name. Παράγεται για τις 389 εγγραφές με αντιστοίχιση.

P186 material used. Τιμή από τη συμφιλιωμένη στήλη Medium. Παράγεται για 499 εγγραφές.

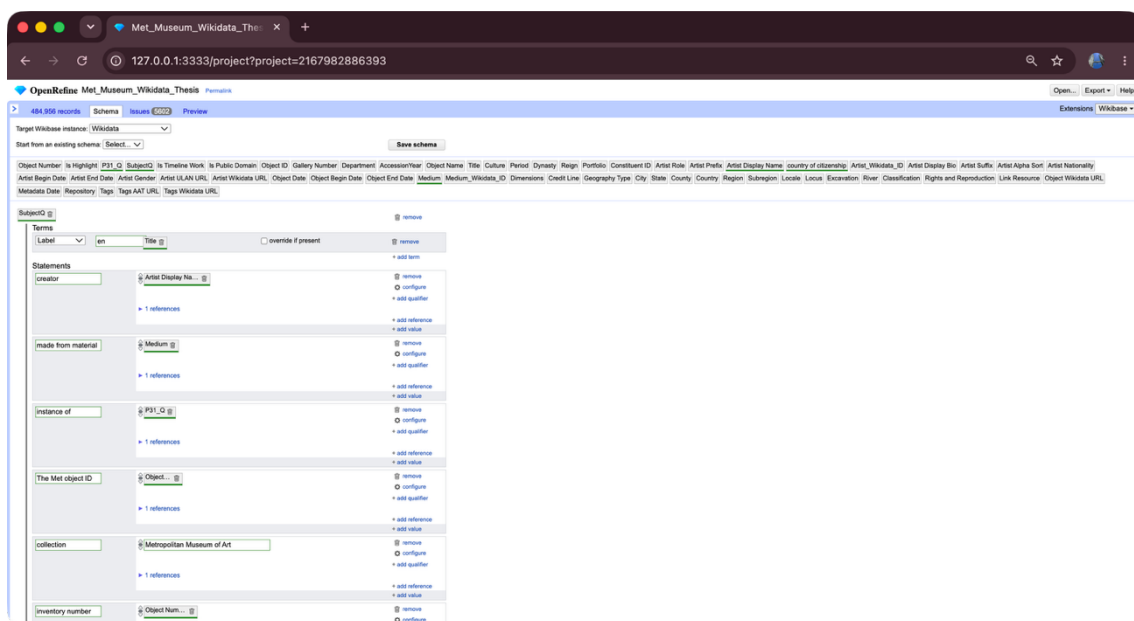
P31 instance of. Τιμή από τη στήλη P31_Q, η οποία περιγράφεται στη συνέχεια. Παράγεται για 493 εγγραφές.

P3634 The Met object ID. Τιμή από τη στήλη Object ID. Πρόκειται για εξωτερικό αναγνωριστικό που διατηρεί σαφή σύνδεση με την αρχική εγγραφή του μουσείου. Παράγεται για όλες τις 501 εγγραφές.

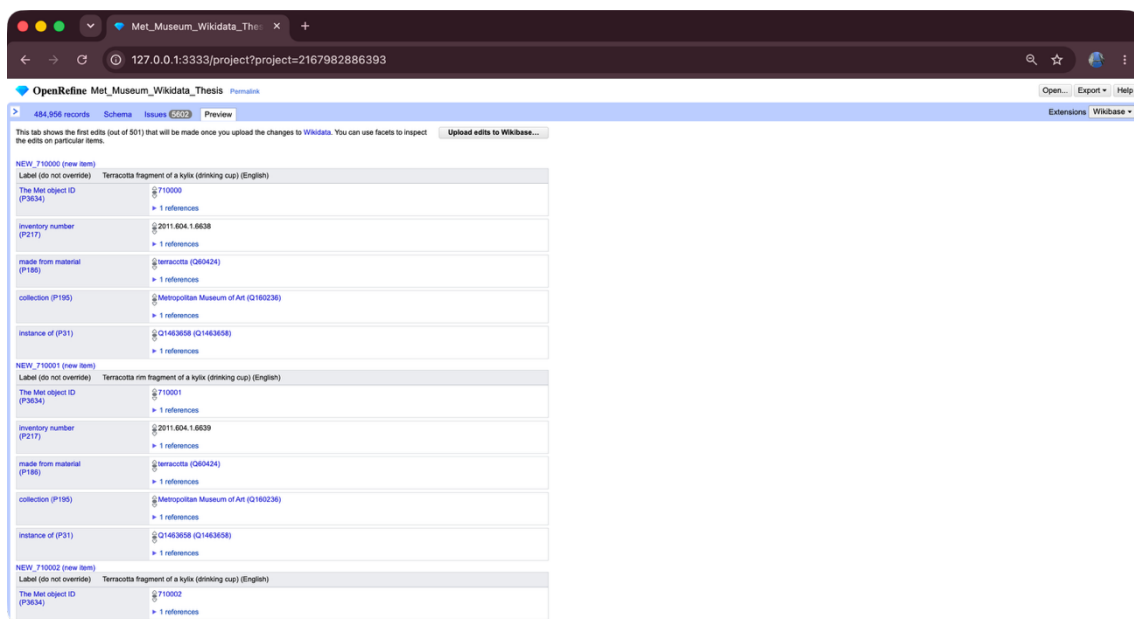
P195 collection. Σταθερή τιμή Q160236. Δηλώνει ότι το αντικείμενο ανήκει στη συλλογή του Metropolitan Museum of Art.

P217 inventory number. Τιμή από τη στήλη Object Number, δηλαδή τον αριθμό καταλόγου του μουσείου.

Οι τρεις τελευταίες ιδιότητες προστέθηκαν ώστε το μοντέλο να είναι πληρέστερο ως πραγματική αναπαράσταση μουσειακού αντικειμένου. Ιδιαίτερη σημασία έχει η P3634: η ύπαρξη σταθερού εξωτερικού αναγνωριστικού επιτρέπει την ταυτοποίηση της εγγραφής, διευκολύνει τη μελλοντική συγχρονισμένη ενημέρωση, και λειτουργεί ως μηχανισμός αποτροπής διπλοεγγραφών, καθώς επιτρέπει σε άλλους επεξεργαστές να διαπιστώσουν αν ένα αντικείμενο έχει ήδη καταχωριστεί.



Σχήμα 3.14: Το Wikibase Schema με τις επτά δηλώσεις



Σχήμα 3.15: Η προεπισκόπηση των παραγόμενων δηλώσεων

3.5.4 Ο υπό συνθήκη ορισμός της ιδιότητας P31

Η δήλωση instance of απαιτεί ιδιαίτερη προσοχή. Μια πρώτη προσέγγιση θα ήταν η απόδοση σταθερής τιμής σε όλες τις εγγραφές, δεδομένου ότι το δείγμα αποτελείται κατά κύριο λόγο από θραύσματα αγγείων.

Ο έλεγχος του πεδίου Object Name αποκάλυψε ωστόσο ότι οκτώ εγγραφές δεν είναι θραύσματα. Πρόκειται για ένα band-cup, τρεις εγγραφές με την ένδειξη undetermined closed shape, ένα cup, μία libation table και δύο σύνολα plaster impressions με 136 και 99 τεμάχια αντίστοιχα. Η απόδοση της τιμής shard σε αυτές θα ήταν σημασιολογικά εσφαλμένη.

Δημιουργήθηκε επομένως βοηθητική στήλη P31_Q με ρητή εξαίρεση των συγκεκριμένων αντικειμένων βάσει του αναγνωριστικού τους:

```
if(or(cells["Object ID"].value==738889,
      cells["Object ID"].value==768183,
      ... , cells["Object ID"].value==813903),
    null, "Q1463658")
```

Η επιλογή ρητής εξαίρεσης βάσει αναγνωριστικού, αντί για αναζήτηση μοτίβου στο κείμενο, έγινε συνειδητά. Οι δοκιμές με έλεγχο της συμβολοσειράς fragment στο πεδίο Object Name ή στον τίτλο απέδωσαν ασυνεπή αποτελέσματα, καθώς ορισμένοι τίτλοι περιέχουν τη λέξη ενώ το αντικείμενο δεν είναι θραύσμα. Η εξαίρεση βάσει ελεγμένης λίστας είναι πλήρως τεκμηριωμένη και αναπαραγώγιμη.

Η στήλη υποβλήθηκε σε Use values as identifiers. Τα 493 κελιά με τιμή Q1463658 σημάνθηκαν ως αντιστοιχισμένα, ενώ τα 8 κενά παρέμειναν ασυμφιλίωτα. Το OpenRefine παραλείπει δηλώσεις με κενή τιμή, οπότε για τα οκτώ αυτά αντικείμενα δεν παράγεται δήλωση P31.

Η επιλογή να μην αποδοθεί εναλλακτική τιμή στα οκτώ αντικείμενα αντί για παράλειψη αντανάκλα αρχή προσοχής: προτιμότερη είναι η απουσία δήλωσης από μια αβέβαιη δήλωση, ιδίως σε δημόσια βάση γνώσης όπου το σφάλμα διαδίδεται.

Σε κάθε δήλωση προστέθηκε αναφορά με ιδιότητα P248 και σταθερή τιμή Q160236.

3.5.5 Η εξαγωγή και το σφάλμα της έκδοσης

Η εξαγωγή του αρχείου QuickStatements επιχειρήθηκε αρχικά από το μενού Export. Η ενέργεια απέτυχε με σφάλμα HTTP 500 και μήνυμα `java.lang.UnsupportedOperationException`, το οποίο υποδεικνύει ότι η μέθοδος GET δεν υλοποιείται από τον συγκεκριμένο εξαγωγέα στην έκδοση 3.10.0.

Πρόκειται για σφάλμα του λογισμικού και όχι για πρόβλημα των δεδομένων ή του schema. Η έκδοση 3.10.0 φέρει την ένδειξη TRUNK, δηλαδή πρόκειται για αναπτυξιακή και όχι σταθερή έκδοση. Η εξαγωγή ολοκληρώθηκε τελικά μέσω εναλλακτικής διαδρομής από την καρτέλα Preview.

Η καταγραφή του περιστατικού έχει αξία για την αναπαραγωγιμότητα. Ένας ερευνητής που θα επιχειρήσει να επαναλάβει τη διαδικασία με την ίδια έκδοση θα αντιμετωπίσει το ίδιο εμπόδιο, και η γνώση της εναλλακτικής διαδρομής του εξοικονομεί χρόνο. Ευρύτερα, το περιστατικό υπενθυμίζει ότι η επιλογή σταθερής έκδοσης λογισμικού αποτελεί μέρος της μεθοδολογικής επιμέλειας.

3.5.6 Η δομή του παραγόμενου αρχείου

Το αρχείο QuickStatements ακολουθεί τη σύνταξη V1, στην οποία κάθε γραμμή αποτελείται από πεδία διαχωρισμένα με χαρακτήρα στηλοθέτη. Η λέξη CREATE σηματοδοτεί την έναρξη νέου item και η λέξη LAST αναφέρεται στο τελευταίο δημιουργηθέν item.

Ένα τυπικό μπλοκ έχει την ακόλουθη μορφή:

CREATE

```
LAST Len      "Terracotta fragment of a kylix (drinking cup) "
LAST P170     Q322192      S248  Q160236
LAST P186     Q60424       S248  Q160236
LAST P31      Q1463658     S248  Q160236
LAST P3634    "710002"     S248  Q160236
LAST P195     Q160236       S248  Q160236
LAST P217     "2011.604.1.6612" S248  Q160236
```

Ο κωδικός Len δηλώνει ετικέτα στα αγγλικά. Η σύμβαση S248 αντί για P248 στις αναφορές διακρίνει τις ιδιότητες που ανήκουν στην αναφορά από εκείνες της κύριας δήλωσης· το γράμμα S προέρχεται από τη λέξη source.

Το τελικό αρχείο περιέχει 3.871 γραμμές. Η επαλήθευση της δομής του παρουσιάζεται στο Κεφάλαιο 4.

```

CREATE
LAST Len "Terracotta fragment of a kylix (drinking cup)"
LAST P186 060424 S248 Q160236
LAST P31 01463658
CREATE
LAST Len "Terracotta rim fragment of a kylix (drinking cup)"
LAST P186 060424 S248 Q160236
LAST P31 01463658
CREATE
LAST Len "Terracotta fragment of a kylix (drinking cup)"
LAST P170 Q322192 S248 Q160236
LAST P186 060424 S248 Q160236
LAST P31 01463658
CREATE
LAST Len "Terracotta fragment of a kylix (drinking cup)"
LAST P186 060424 S248 Q160236
LAST P31 01463658
CREATE
LAST Len "Terracotta fragment of a kylix (drinking cup)"
LAST P186 060424 S248 Q160236
LAST P31 01463658
CREATE
LAST Len "Terracotta fragment of a kylix (drinking cup)"
LAST P186 060424 S248 Q160236
LAST P31 01463658
CREATE
LAST Len "Terracotta fragment of a kylix (drinking cup)"
LAST P170 Q322192 S248 Q160236
LAST P186 060424 S248 Q160236
LAST P31 01463658
CREATE
LAST Len "Terracotta fragment of a kylix (drinking cup)"
LAST P186 060424 S248 Q160236
LAST P31 01463658
CREATE
LAST Len "Terracotta fragment of a kylix (drinking cup)"
LAST P170 Q475054 S248 Q160236
LAST P186 060424 S248 Q160236
LAST P31 01463658
CREATE
LAST Len "Terracotta fragment of a kylix (drinking cup)"
LAST P170 Q475054 S248 Q160236
LAST P186 060424 S248 Q160236
LAST P31 01463658
CREATE
LAST Len "Terracotta fragment of a kylix (drinking cup)"
LAST P170 Q475054 S248 Q160236
LAST P186 060424 S248 Q160236
LAST P31 01463658

```

Σχήμα 3.16: Δείγμα του παραγόμενου αρχείου

3.6 Δεοντολογικές Παρατηρήσεις

Η καρτέλα Preview του OpenRefine περιλαμβάνει κουμπί Upload edits to Wikibase, το οποίο εκτελεί απευθείας τις δηλώσεις στα πραγματικά Wikidata. Η λειτουργία αυτή δεν χρησιμοποιήθηκε, και η απόφαση αξίζει αιτιολόγησης.

Η μαζική δημιουργία 501 νέων items σε δημόσια βάση γνώσης δεν είναι ουδέτερη ενέργεια. Επιβαρύνει την κοινότητα με το κόστος ελέγχου και ενδεχόμενης διόρθωσης, και για αντικείμενα όπως τα θραύσματα αγγείων εγείρει το ερώτημα της καταλληλότητας για συμπερίληψη. Οι κατευθυντήριες γραμμές των Wikidata προβλέπουν κριτήρια αξιοσημείωτου, και η μαζική εισαγωγή συνήθως προϋποθέτει προηγούμενη συνεννόηση με τη σχετική κοινότητα.

Επιπλέον, το ίδιο το σύστημα επικύρωσης επισήμανε ζητήματα ποιότητας: τα νέα items δημιουργούνται χωρίς περιγραφή, στοιχείο που δυσχεραίνει τη διάκρισή τους από ομώνυμα. Η προσθήκη περιγραφών θα ήταν προϋπόθεση για υπεύθυνη εισαγωγή.

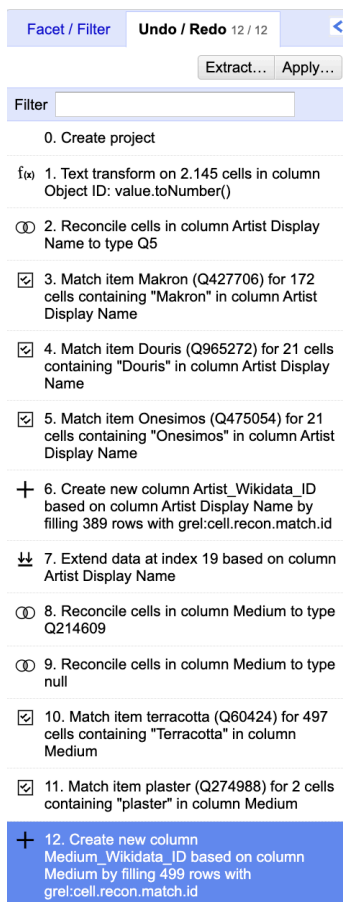
Για τους λόγους αυτούς, το αρχείο QuickStatements παράγεται, επαληθεύεται και παρουσιάζεται ως παραδοτέο, χωρίς να εκτελεστεί. Η επιλογή αυτή δεν αποδυναμώνει την εργασία· η μεθοδολογία αποδεικνύεται πλήρως λειτουργική και το αρχείο είναι έτοιμο προς χρήση από ίδρυμα που θα αναλάβει την ευθύνη της εισαγωγής μετά από την απαραίτητη συνεννόηση.

3.7 Η Τεκμηρίωση της Αναπαραγωγιμότητας

Το OpenRefine καταγράφει αυτόματα κάθε ενέργεια στην καρτέλα Undo/Redo, η οποία λειτουργεί ταυτόχρονα ως μηχανισμός αναίρεσης και ως πλήρες αρχείο καταγραφής της ερευνητικής διαδικασίας.

Μέσω του κουμπιού Extract το σύνολο των ενεργειών εξάγεται ως JSON. Το αρχείο περιέχει δομημένη περιγραφή κάθε μετασχηματισμού: τη λειτουργία, τη στήλη, τις παραμέτρους, τις GREL εκφράσεις και - κρίσιμα - την πλήρη διαμόρφωση των ενεργών φίλτρων στο πεδίο engineConfig. Ένας άλλος ερευνητής μπορεί να φορτώσει το αρχείο σε νέο project μέσω του κουμπιού Apply και να αναπαραγάγει την ίδια ακολουθία.

Η πρακτική αξία επιβεβαιώθηκε εμπειρικά. Η χειροκίνητη επανάληψη της διαδικασίας οδήγησε σε διαφορετικό αριθμό εγγραφών λόγω της αδυναμίας ακριβούς αναπαραγωγής του Numeric Facet. Η εφαρμογή του JSON, αντιθέτως, αναπαράγει το αποτέλεσμα με πλήρη πιστότητα. Το εύρημα καταδεικνύει ότι η αυτόματη καταγραφή δεν αποτελεί απλώς καλή πρακτική τεκμηρίωσης αλλά αναγκαία προϋπόθεση επιστημονικής εγκυρότητας για εργασίες που βασίζονται σε διαδραστική επεξεργασία δεδομένων.



Σχήμα 3.17: Το ιστορικό των ενεργειών στην καρτέλα Undo/Redo

Κεφάλαιο 4ο: Αποτελέσματα και Αξιολόγηση

4.1 Εισαγωγή

Το παρόν κεφάλαιο παρουσιάζει και αξιολογεί συστηματικά τα αποτελέσματα της μεθοδολογίας. Η αξιολόγηση οργανώνεται σε έξι άξονες: τα παραδοτέα, την ποιότητα της συμφιλίωσης των δημιουργών, την αντικειμενική μέτρηση έναντι χρυσού προτύπου, τη συμφιλίωση των υλικών, τον εμπλουτισμό, και τη βελτίωση της διαλειτουργικότητας.

Η προσέγγιση επιδιώκει να είναι κριτική και ισορροπημένη. Παρουσιάζει τόσο τα επιτεύγματα όσο και τα όρια, και επιχειρεί να εξηγήσει τα αποτελέσματα αντί απλώς να τα καταγράψει. Όπως θα φανεί, ορισμένα από τα φαινομενικά αρνητικά ευρήματα αποδεικνύονται από τα πιο διδακτικά στοιχεία της εργασίας.

4.2 Τα Παραδοτέα

4.2.1 Επισκόπηση

Η εφαρμογή της μεθοδολογίας παρήγαγε τρία παραδοτέα που λειτουργούν συμπληρωματικά: το εμπλουτισμένο dataset ως άμεσο προϊόν, το αρχείο μεθοδολογίας που τεκμηριώνει τη διαδικασία, και το αρχείο εντολών που επιτρέπει την εισαγωγή στα Wikidata.

Πίνακας 4.1: Τα τρία παραδοτέα της εργασίας

Παραδοτέο	Μορφή	Μέγεθος	Σκοπός
Εμπλουτισμένο dataset	CSV / XLSX	501 γραμμές, 57 στήλες	Ανάλυση και επαναχρησιμοποίηση
Αρχείο μεθοδολογίας	JSON	Πλήρες ιστορικό ενεργειών	Αναπαραγωγιμότητα
Αρχείο εντολών	TXT (QS V1)	3.871 γραμμές	Ενημέρωση Wikidata

4.2.2 Το εμπλουτισμένο dataset

Το κύριο παραδοτέο περιλαμβάνει 501 εγγραφές και 57 στήλες, τρεις περισσότερες από τις 54 του αρχικού αρχείου.

Artist_Wikidata_ID: Περιέχει τον κωδικό Q-number του Wikidata item κάθε δημιουργού. Η προσθήκη μετατρέπει το πεδίο Artist Display Name από απλή συμβολοσειρά σε αναγνωριστικό παγκόσμιας εμβέλειας. Ενώ η συμβολοσειρά Makron είναι κείμενο που γράφεται διαφορετικά σε κάθε γλώσσα, ο κωδικός Q427706 αναγνωρίζει μοναδικά το ίδιο πρόσωπο σε ολόκληρο το οικοσύστημα των Linked Open Data.

Medium_Wikidata_ID: Περιέχει τον κωδικό Q-number του υλικού κάθε αντικειμένου, συμπληρωμένο σε 499 από τις 501 εγγραφές.

country of citizenship: Περιέχει την ιδιότητα P27 όπως ανακτήθηκε αυτόματα. Πρόκειται για πληροφορία που δεν προέρχεται από το Met αλλά από τον παγκόσμιο γράφο γνώσης.

Πίνακας 4.2: Πληρότητα των νέων στηλών

Στήλη	Συμπληρωμένες	Ποσοστό	Μοναδικές τιμές
Artist_Wikidata_ID	389 / 501	77,6%	63
Medium_Wikidata_ID	499 / 501	99,6%	2

country of citizenship	291 / 501	58,1%	5
------------------------	-----------	-------	---

4.2.3 Το αρχείο μεθοδολογίας

Το αρχείο JSON περιέχει τις ενέργειες που εκτελέστηκαν, με πλήρη καταγραφή των παραμέτρων τους και παρατίθεται στο Παράρτημα Α. Η δομή κάθε καταγραφής περιλαμβάνει το πεδίο `op` που προσδιορίζει τον τύπο της λειτουργίας, το `columnName`, το `expression` όπου εφαρμόζεται, και το `engineConfig` με τη διαμόρφωση των ενεργών φίλτρων.

Η αξία του αρχείου δεν είναι μόνο τεκμηριωτική αλλά και λειτουργική: επιτρέπει την ακριβή αναπαραγωγή του αποτελέσματος, κάτι που η χειροκίνητη επανάληψη δεν εξασφαλίζει.

4.2.4 Το αρχείο QuickStatements και η επαλήθευσή του

Το τρίτο παραδοτέο υποβλήθηκε σε συστηματικό έλεγχο δομής, τα αποτελέσματα του οποίου παρουσιάζονται στον ακόλουθο πίνακα.

Πίνακας 4.3: Επαλήθευση της δομής του αρχείου QuickStatements

Στοιχείο	Πλήθος	Αναμενόμενο	Έλεγχος
Συνολικές γραμμές	3.871	—	—
Εντολές CREATE	486	501 – 15 υπάρχοντα	Ορθό
Υπάρχοντα items σε χρήση	15	15 με Object Wikidata URL	Ορθό
Ετικέτες (Len)	501	501 εγγραφές	Ορθό
Δηλώσεις P170	389	389 αντιστοιχίσεις	Ορθό
Δηλώσεις P186	499	499 αντιστοιχίσεις	Ορθό
Δηλώσεις P31	493	501 – 8 μη θραύσματα	Ορθό
Δηλώσεις P3634	501	501 εγγραφές	Ορθό
Δηλώσεις P195	501	501 εγγραφές	Ορθό
Δηλώσεις P217	501	501 εγγραφές	Ορθό
Αναφορές S248	2.884	Άθροισμα δηλώσεων	Ορθό
Τιμές που δεν είναι Q-number	0	0	Ορθό

Ο έλεγχος επιβεβαιώνει την πλήρη εσωτερική συνέπεια. Οι 486 εντολές CREATE αντιστοιχούν στις εγγραφές χωρίς υπάρχον Wikidata item, ενώ οι 15 που διαθέτουν item αναφέρονται σε αυτό με το Q-number τους. Οι 2.884 αναφορές ισούνται με το άθροισμα των δηλώσεων P170, P186, P31, P3634, P195 και P217, δηλαδή κάθε δήλωση φέρει τεκμηρίωση πηγής.

Πρέπει να διευκρινιστεί ότι ο έλεγχος αφορά τη συντακτική συνέπεια του αρχείου και όχι τη σημασιολογική ορθότητα των αντιστοιχίσεων. Το ότι μια δήλωση είναι συντακτικά άρτια δεν εγγυάται ότι το Q-number που περιέχει αντιστοιχεί όντως στη σωστή οντότητα. Η σημασιολογική διάσταση εξετάζεται χωριστά στην ενότητα 4.4.

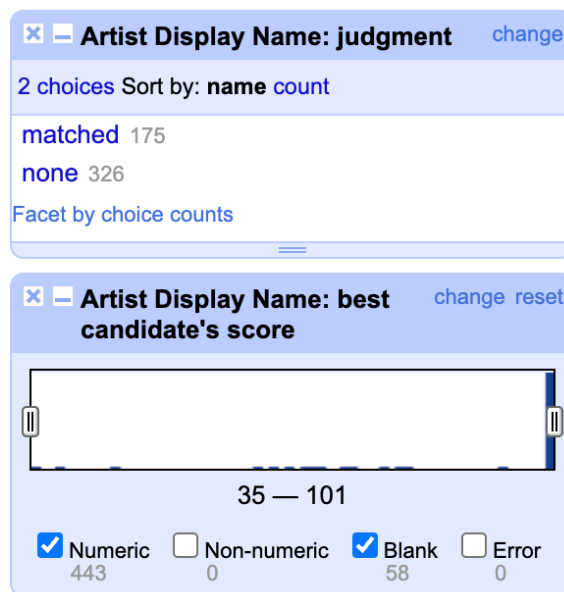
4.3 Αξιολόγηση της Συμφιλίωσης των Δημιουργών

4.3.1 Η τριμερής ανάλυση της αυτόματης φάσης

Η αυτόματη συμφιλίωση απέδωσε 175 αντιστοιχίσεις επί 501 εγγραφών. Ο συνδυασμός του judgment facet με το facet best candidate's score επιτρέπει ωστόσο πιο διαφοροποιημένη ανάγνωση.

Πίνακας 4.4: Τριμερής ανάλυση των αποτελεσμάτων αυτόματης συμφιλίωσης

Κατηγορία	Εγγραφές	Ποσοστό
Αντιστοιχίστηκαν αυτόματα	175	34,9%
Είχαν υποψηφίους χωρίς απόφαση	268	53,5%
Δεν είχαν κανέναν υποψήφιο	58	11,6%



Σχήμα 4.1: Ανάλυση των αποτελεσμάτων αυτόματης συμφιλίωσης

Η διάκριση είναι ουσιώδης για την ερμηνεία. Μόνο το 11,6% των εγγραφών αντιστοιχεί σε ονόματα για τα οποία η υπηρεσία δεν βρήκε τίποτα. Για το 53,5% η πληροφορία υπήρχε αλλά η βεβαιότητα ήταν ανεπαρκής. Η διαπίστωση αυτή προδιαγράφει τη στρατηγική: το μεγαλύτερο μέρος του κενού καλύπτεται με επιβεβαίωση υπαρχόντων υποψηφίων, όχι με αναζήτηση από το μηδέν.

4.3.2 Η επίδραση της χειροκίνητης επιμέλειας

Οι τρεις ενέργειες χειροκίνητης επιμέλειας που περιγράφηκαν στο Κεφάλαιο 3 μετέβαλαν δραστικά την εικόνα.

Πίνακας 4.5: Σύγκριση πριν και μετά τη χειροκίνητη επιμέλεια

Στάδιο	Αντιστοιχισμένες	Μη αντιστοιχισμένες	Ποσοστό επιτυχίας
Αυτόματη συμφιλίωση	175	326	34,9%
Μετά τρεις ενέργειες	389	112	77,6%

Μεταβολή	+214	-214	+42,7 μονάδες
----------	------	------	---------------



Σχήμα 4.2: Η επίδραση της χειροκίνητης επιμέλειας

Η απόδοση της παρέμβασης είναι εντυπωσιακή: τρεις ενέργειες κάλυψαν 214 εγγραφές, μέσος όρος 71 εγγραφές ανά ενέργεια. Για σύγκριση, τα υπόλοιπα 62 μη αντιστοιχισμένα ονόματα καλύπτουν συνολικά 112 εγγραφές, δηλαδή μέση απόδοση 1,8 ανά ενέργεια - σχεδόν σαράντα φορές χαμηλότερη.

Το εύρημα τεκμηριώνει τη στρατηγική προτεραιοποίησης κατά συχνότητα ως γενικεύσιμη αρχή. Σε datasets με ασύμμετρη κατανομή - που είναι ο κανόνας και όχι η εξαίρεση σε μουσειακές συλλογές - η ταξινόμηση των μη αντιστοιχισμένων τιμών κατά φθίνουσα συχνότητα και η εστίαση στις πρώτες αποδίδει δυσανάλογα μεγάλο όφελος με ελάχιστη προσπάθεια.

4.3.3 Το φαινόμενο Makron

Η ανάλυση της κατανομής αποκαλύπτει ότι το αρχικό ποσοστό του 34,9% δεν αντανakλά συστηματική αδυναμία της μεθοδολογίας αλλά καθορίζεται από την αποτυχία ελάχιστων, πολυπληθών σε έργα δημιουργών.

Στον αγγειογράφο Makron αποδίδονται 172 από τα 501 αντικείμενα, ποσοστό 34,3%. Μόνος του ευθύνεται για 172 από τις 326 αρχικές αποτυχίες, δηλαδή για το 52,8%. Προσθέτοντας τους Douris και Onesimos με 21 έργα ο καθένας, τρία ονόματα εξηγούν 214 αποτυχίες, ποσοστό 65,6%.

Το κρίσιμο ερώτημα είναι γιατί απέτυχε η αντιστοίχιση. Ο έλεγχος έδειξε ότι το ορθό item, Q427706, εμφανιζόταν κανονικά στη λίστα των υποψηφίων που είχε επιστρέψει η αυτόματη διαδικασία. Η υπηρεσία δηλαδή είχε εντοπίσει τη σωστή οντότητα.

Περαιτέρω έλεγχος του ίδιου του Wikidata item ενισχύει το εύρημα. Το Q427706 περιγράφεται ρητά ως Αττικός αγγειογράφος, χρονολογείται στη δεκαετία του 480 π.Χ., καταγράφει τη συνεργασία με τον κεραμέα Ιέρωνα, και περιλαμβάνει δηλώσεις για συλλογές που φυλάσσουν έργα του. Μεταξύ αυτών περιλαμβάνεται το ίδιο το Metropolitan Museum of Art, με αναφορά stated in Metropolitan Museum of Art συνοδευόμενη από The Met object ID.

Με άλλα λόγια, τα Wikidata γνωρίζουν όχι μόνο τον Makron αλλά και τη σχέση του με το συγκεκριμένο μουσείο. Η αποτυχία της αυτόματης συμφιλίωσης συνιστά επομένως ψευδώς αρνητικό αποτέλεσμα, όχι κενό στη βάση γνώσης.

Ανάλογη είναι η περίπτωση του Douris, όπου η αιτία της διστακτικότητας είναι ρητά τεκμηριωμένη: το item Q965272 φέρει δήλωση ότι δεν πρέπει να συγχέεται με τον Duris, ιστορικό συγγραφέα με σχεδόν ταυτόσημο όνομα. Η ύπαρξη ομώνυμης οντότητας εξηγεί γιατί κανένας υποψήφιος δεν συγκέντρωσε αρκετά υψηλό score.

4.3.4 Οι τεκμηριωμένες αρνήσεις

Δύο περιπτώσεις εξετάστηκαν και δεν αντιστοιχίστηκαν, και η τεκμηρίωσή τους έχει αυτοτελή αξία.

Η τιμή Brygan, με 12 εγγραφές, εμφάνιζε ως μοναδικό υποψήφιο τον Brygos Painter (Q322192). Η αντιστοίχιση απορρίφθηκε επειδή η σύγχρονη έρευνα διακρίνει τον Βρύγο ως κεραμέα από τον Brygos Painter ως αγγειογράφο του εργαστηρίου του, και επειδή το δείγμα περιέχει ήδη ξεχωριστή τιμή Brygos Painter αντιστοιχισμένη στο ίδιο item. Η αντιστοίχιση θα συγχώνευε δύο διακριτές τιμές του dataset σε μία οντότητα.

Η τιμή Veii Painter, με 10 εγγραφές, δεν εμφάνισε κανέναν υποψήφιο. Πρόκειται για συμβατικό όνομα που δεν έχει καταχωριστεί στα Wikidata με αυτή τη μορφή.

Οι δύο περιπτώσεις αντιπροσωπεύουν διαφορετικούς τύπους ορίου: όριο γνώσης η πρώτη, όριο κάλυψης η δεύτερη. Και στις δύο η ορθή απόφαση ήταν η αποχή. Η παρατήρηση αυτή έχει σημασία για τη μεθοδολογία της χειροκίνητης επιμέλειας: στόχος δεν είναι η μεγιστοποίηση των αντιστοιχίσεων αλλά η ορθότητά τους, και η άρνηση αντιστοίχισης αποτελεί έγκυρη απόφαση.

4.4 Αντικειμενική Αξιολόγηση έναντι Χρυσού Προτύπου

4.4.1 Το θεσμικό σύνολο αναφοράς

Η αξιολόγηση της ποιότητας μιας διαδικασίας συμφιλίωσης απαιτεί κάποιο ανεξάρτητο σημείο σύγκρισης. Στην πλειονότητα των αντίστοιχων εργασιών τέτοιο σημείο δεν είναι διαθέσιμο και η αξιολόγηση παραμένει ποιοτική.

Στην παρούσα εργασία η δυνατότητα υπήρχε εξ αρχής μέσα στα ίδια τα δεδομένα. Όπως επισημάνθηκε στο Κεφάλαιο 2, το dataset περιλαμβάνει τη στήλη Artist Wikidata URL με συνδέσμους που έχει καταχωρίσει το ίδιο το μουσείο. Στο δείγμα των 501 εγγραφών η στήλη είναι συμπληρωμένη σε 418 περιπτώσεις.

Οι καταχωρίσεις αυτές αποτελούν πολύτιμο σημείο αναφοράς για τρεις λόγους. Προέρχονται από το ίδρυμα που κατέχει τα αντικείμενα και διαθέτει την πληρέστερη γνώση για αυτά. Δημιουργήθηκαν ανεξάρτητα από τη διαδικασία της παρούσας εργασίας. Και αφορούν την ίδια ακριβώς σχέση που επιχειρεί να εντοπίσει η συμφιλίωση.

Θα ήταν ωστόσο μεθοδολογικά ανακριβές να χαρακτηριστούν χρυσό πρότυπο με την αυστηρή έννοια του όρου. Ένα χρυσό πρότυπο προϋποθέτει ότι οι απαντήσεις του είναι αντικειμενικά ορθές και ανεξάρτητα επικυρωμένες. Οι καταχωρίσεις του Met δεν έχουν υποβληθεί σε τέτοιο έλεγχο στο πλαίσιο της εργασίας, και όπως θα φανεί στην ενότητα 4.4.4, τουλάχιστον μία από αυτές αμφισβητείται από την ίδια την εργασία.

Στη συνέχεια του κειμένου η στήλη αναφέρεται επομένως ως θεσμικό σύνολο αναφοράς. Οι μετρικές που υπολογίζονται δεν εκφράζουν απόλυτη ορθότητα αλλά βαθμό συμφωνίας με τις υπάρχουσες αντιστοιχίες του μουσείου.

4.4.2 Οι μετρικές

Ως αληθώς θετικά ορίζονται οι εγγραφές όπου η συμφιλίωση απέδωσε Q-number ταυτόσημο με εκείνο του Met. Ως ψευδώς θετικά όπου απέδωσε διαφορετικό. Ως ψευδώς αρνητικά όπου το Met διαθέτε τιμή αλλά η συμφιλίωση δεν απέδωσε καμία.

Πίνακας 4.6: Σύγκριση με το θεσμικό σύνολο αναφοράς

Κατηγορία	Εγγραφές	Ερμηνεία
Αληθώς θετικά (TP)	375	Σωστή αντιστοίχιση

Ψευδώς θετικά (FP)	0	Λανθασμένη αντιστοίχιση
Ψευδώς αρνητικά (FN)	43	Χαμένη αντιστοίχιση
Επιπλέον ευρήματα	14	Το Met δεν διαθέτε τιμή
Μέγεθος χρυσού προτύπου	418	Εγγραφές με τιμή στο Met

Πίνακας 4.7: Μετρικές απόδοσης

Μετρική	Τιμή	Ορισμός
Precision (ακρίβεια)	100,0%	$TP / (TP + FP)$
Recall (ανάκληση)	89,7%	$TP / (TP + FN)$
F1-score	94,6%	Αρμονικός μέσος

4.4.3 Ερμηνεία των αποτελεσμάτων

Η απόλυτη συμφωνία στις επαληθεύσιμες περιπτώσεις είναι το σημαντικότερο ποσοτικό αποτέλεσμα. Καμία από τις 375 αντιστοιχίσεις που μπορούσαν να συγκριθούν δεν διέφερε από την καταχώριση του μουσείου, παρότι οι 214 από αυτές προέκυψαν από χειροκίνητη παρέμβαση.

Η διατύπωση απαιτεί ωστόσο προσοχή. Το ποσοστό 100% δεν σημαίνει ότι όλες οι αποφάσεις συμφιλίωσης είναι αντικειμενικά ορθές, αλλά ότι συμφωνούν με το θεσμικό σύνολο αναφοράς. Αν κάποια καταχώριση του Met είναι η ίδια εσφαλμένη, το σφάλμα αναπαράγεται χωρίς να ανιχνεύεται από τη μέτρηση. Το ποσοστό εκφράζει επομένως συνέπεια με μια αξιόπιστη πηγή, όχι επαληθευμένη αλήθεια.

Η ανάκληση του 89,7% αντιστοιχεί σε 43 εγγραφές που το Met γνώριζε αλλά η διαδικασία δεν εντόπισε. Δεδομένου ότι η χειροκίνητη επιμέλεια περιορίστηκε συνειδητά σε τρία ονόματα, το υπόλοιπο αυτό είναι αναμενόμενο και θα μπορούσε να καλυφθεί με περαιτέρω επιμέλεια.

Ιδιαίτερο ενδιαφέρον έχει η σύγκριση με το ενδιάμεσο στάδιο. Αν η αξιολόγηση είχε γίνει αμέσως μετά την αυτόματη συμφιλίωση, η ανάκληση θα ήταν 38,4% και το F1-score 55,5%. Η χειροκίνητη επιμέλεια τριών ονομάτων ανέβασε το F1 από 55,5% σε 94,6%, χωρίς καμία απώλεια ακρίβειας.

Πίνακας 4.8: Εξέλιξη των μετρικών

Στάδιο	Precision	Recall	F1-score
Μετά την αυτόματη συμφιλίωση	100,0%	38,4%	55,5%
Μετά τη χειροκίνητη επιμέλεια	100,0%	89,7%	94,6%

4.4.4 Ανάκληση σε επίπεδο ονομάτων

Η μέτρηση σε επίπεδο μοναδικών ονομάτων αντί για εγγραφές δίνει διαφορετική εικόνα. Από τα 81 διακριτά ονόματα για τα οποία το Met διαθέτε Wikidata URL, η διαδικασία εντόπισε σωστά τα 58, δηλαδή ανάκληση 71,6%.

Η απόκλιση από το 89,7% των εγγραφών οφείλεται στην ασυμμετρία της κατανομής. Ο Makron μετράει ως ένα όνομα αλλά ως 172 εγγραφές. Η διπλή μέτρηση είναι διαφωτιστική: σε επίπεδο εγγραφών η απόδοση είναι υψηλότερη επειδή η επιμέλεια στόχευσε τα πολυπληθέστερα ονόματα, ενώ σε επίπεδο ονομάτων αποτυπώνεται καλύτερα η κάλυψη του λεξιλογίου.

4.4.5 Τα όρια του συνόλου αναφοράς

Η ανάλυση των ψευδώς αρνητικών αποκάλυψε ένα εύρημα που φωτίζει τα όρια του ίδιου του συνόλου αναφοράς.

Από τις 43 εγγραφές όπου το Met διέθετε αντιστοίχιση και η διαδικασία δεν απέδωσε, οι 11 αφορούν την τιμή Brygan. Το μουσείο αποδίδει σε αυτές τον κωδικό Q322192, δηλαδή τον Brygos Painter. Η παρούσα εργασία απέρριψε συνειδητά αυτή την αντιστοίχιση, για τους λόγους που αναλύθηκαν στην ενότητα 4.3.4: η σύγχρονη έρευνα διακρίνει τον κεραμέα Βρύγο από τον ομώνυμο αγγειογράφο, και το δείγμα περιέχει ήδη ξεχωριστή τιμή Brygos Painter αντιστοιχισμένη στο ίδιο item.

Το περιστατικό έχει διπλή σημασία. Πρώτον, καταδεικνύει ότι το θεσμικό σύνολο αναφοράς δεν είναι αλάνθαστο και δεν μπορεί να λειτουργήσει ως απόλυτο κριτήριο ορθότητας. Δεύτερον, δείχνει ότι μέρος των ψευδώς αρνητικών δεν συνιστά αστοχία αλλά τεκμηριωμένη διαφωνία. Από τα 43, τα 11 είναι αποτέλεσμα κρίσης και όχι αδυναμίας.

Αν οι έντεκα αυτές εγγραφές εξαιρεθούν από τον υπολογισμό ως αμφισβητούμενες, η ανάκληση ανέρχεται σε 92,1%. Η παρουσίαση αμφοτέρων των τιμών είναι προτιμότερη από την επιλογή της ευνοϊκότερης, καθώς η διαφορά τους αποτυπώνει ακριβώς το εύρος της αβεβαιότητας που εισάγει η φύση του συνόλου αναφοράς.

4.4.6 Τα δεκατέσσερα επιπλέον ευρήματα

Σε 14 εγγραφές η συμφιλίωση απέδωσε Q-number ενώ το Met δεν διέθετε καμία τιμή. Οι περιπτώσεις αυτές δεν συνιστούν σφάλμα αλλά επέκταση του χρυσού προτύπου: η διαδικασία εντόπισε αντιστοιχίσεις που το μουσείο δεν είχε καταγράψει.

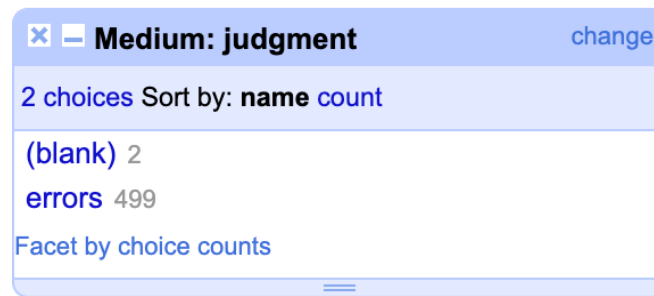
Το εύρημα ενισχύει το επιχείρημα υπέρ της αμφίδρομης αξίας. Η μεθοδολογία δεν λειτουργεί μόνο ως καταναλωτής της γνώσης των Wikidata αλλά και ως μηχανισμός εντοπισμού κενών στην τεκμηρίωση του ίδιου του ιδρύματος. Ένα μουσείο που εφαρμόζει την προσέγγιση λαμβάνει, ως παράπλευρο όφελος, λίστα εγγραφών προς συμπλήρωση.

4.5 Αξιολόγηση της Συμφιλίωσης των Υλικών

Η συμφιλίωση της στήλης Medium απέδωσε 499 από τις 501 εγγραφές, ποσοστό 99,6%. Οι δύο εξαιρέσεις είναι εγγραφές χωρίς τιμή στο πεδίο.

Πίνακας 4.9: Κατανομή και αντιστοίχιση των υλικών

Υλικό	Εγγραφές	Ποσοστό	Wikidata item	Q-number
terracotta	497	99,2%	terracotta	Q60424
plaster	2	0,4%	plaster	Q274988
χωρίς τιμή	2	0,4%	-	-



Σχήμα 4.3: Η ομοιογένεια των υλικών του δείγματος

Η επιτυχία οφείλεται στην ακραία ομοιογένεια του πεδίου, το οποίο περιέχει μόλις δύο μοναδικές τιμές. Η κυριαρχία του πηλού επιβεβαιώνει τη συνοχή του δείγματος και συμφωνεί με την κατηγοριοποίηση, όπου 497 αντικείμενα ταξινομούνται ως Vases.

Η διαδικασία ωστόσο δεν ήταν αυτονόητη. Η πρώτη απόπειρα με περιορισμό τύπου Q214609 απέτυχε καθολικά, με 499 errors. Μόνο η δεύτερη απόπειρα χωρίς περιορισμό επέτρεψε την ολοκλήρωση. Το γεγονός αυτό, σε συνδυασμό με την παρατήρηση για τον περιορισμό Q5 στους δημιουργούς, οδηγεί σε σαφές συμπέρασμα: οι περιορισμοί τύπου στο reconciliation service πρέπει να χρησιμοποιούνται με επιφύλαξη και να ελέγχεται πάντοτε το ενδεχόμενο ψευδώς αρνητικών.

4.6 Η Επίδραση των Περιορισμών Τύπου: Ελεγχόμενο Πείραμα

4.6.1 Το ερευνητικό ερώτημα

Η αποτυχία της πρώτης απόπειρας συμφιλίωσης του πεδίου Medium με περιορισμό τύπου Q214609 οδήγησε στην υπόθεση ότι οι περιορισμοί τύπου ενδέχεται γενικότερα να υπονομεύουν τη διαδικασία. Η υπόθεση επεκτάθηκε αρχικά και στους δημιουργούς, όπου είχε χρησιμοποιηθεί ο τύπος Human (Q5).

Η επέκταση αυτή δεν τεκμηριωνόταν όμως από κανένα άμεσο στοιχείο. Στην περίπτωση του Medium η αποτυχία ήταν καθολική και αδιαμφισβήτητη, με 499 errors. Στην περίπτωση των δημιουργών η διαδικασία εκτελέστηκε κανονικά και απέδωσε 175 αντιστοιχίσεις· το χαμηλό ποσοστό αποδόθηκε στον περιορισμό χωρίς αποδεικτικά στοιχεία.

Για τη διερεύνηση του ερωτήματος σχεδιάστηκε ελεγχόμενο πείραμα, το οποίο επαναλαμβάνει τη συμφιλίωση των δημιουργών με μοναδική μεταβλητή την παρουσία ή απουσία του περιορισμού τύπου.

4.6.2 Ο σχεδιασμός

Δημιουργήθηκε αντίγραφο της στήλης Artist Display Name με την ονομασία AB_test, ώστε η αρχική στήλη με τις χειροκίνητες αντιστοιχίσεις να παραμείνει ανέπαφη. Το αντίγραφο υποβλήθηκε σε συμφιλίωση με την ίδια υπηρεσία και τις ίδιες παραμέτρους, με μοναδική διαφορά την επιλογή reconcile against no particular type.

Η σύγκριση αφορά αποκλειστικά την αυτόματη φάση, χωρίς καμία χειροκίνητη παρέμβαση σε καμία από τις δύο εκτελέσεις. Το δείγμα, η υπηρεσία και η χρονική στιγμή είναι κοινά, οπότε η όποια διαφορά αποδίδεται στον περιορισμό τύπου.

4.6.3 Τα αποτελέσματα

Πίνακας 4.10: Σύγκριση συμφιλίωσης με και χωρίς περιορισμό τύπου

Κατηγορία	Με τύπο Q5	Χωρίς τύπο	Διαφορά
Αντιστοιχίστηκαν αυτόματα	175	178	+3
Είχαν τουλάχιστον έναν υποψήφιο	443	450	+7
Υποψήφιοι χωρίς απόφαση	268	272	+4
Κανένας υποψήφιος	58	51	-7
Ποσοστό αυτόματης συμφιλίωσης	34,9%	35,5%	+0,6 μον.

Το αποτέλεσμα αναιρεί την αρχική υπόθεση. Ο περιορισμός τύπου Q5 επηρεάζει την αυτόματη συμφιλίωση των δημιουργών κατά 0,6 ποσοστιαίες μονάδες, διαφορά που δεν δικαιολογεί τον χαρακτηρισμό του ως αιτίας του χαμηλού ποσοστού.

4.6.4 Ερμηνεία

Η ανάλυση των επιμέρους αριθμών είναι διαφωτιστική. Ο περιορισμός όντως απέκλειε οντότητες από την αναζήτηση: χωρίς αυτόν, επτά επιπλέον εγγραφές βρήκαν τουλάχιστον έναν υποψήφιο και η κατηγορία κανένας υποψήφιος μειώθηκε από 58 σε 51. Η επίδραση στη φάση της ανάκτησης είναι επομένως υπαρκτή.

Η επίδραση όμως δεν μεταφράστηκε σε αποφάσεις. Από τις επτά επιπλέον εγγραφές με υποψηφίους, μόνο τρεις έφτασαν σε αυτόματη αντιστοίχιση· οι υπόλοιπες τέσσερις πρόσθεσαν υποψηφίους χαμηλής βεβαιότητας που παρέμειναν χωρίς απόφαση.

Καθοριστική είναι η περίπτωση του Makron. Ο αγγειογράφος, στον οποίο αποδίδονται 172 από τα 501 αντικείμενα, παραμένει στην κατηγορία none ακόμη και χωρίς τον περιορισμό τύπου. Η αποτυχία της αυτόματης αντιστοίχισής του δεν σχετίζεται επομένως καθόλου με τη ρύθμιση: ο σωστός υποψήφιος εμφανίζεται και στις δύο εκτελέσεις, αλλά ο βαθμός βεβαιότητας δεν φτάνει το κατώφλι αυτόματης αποδοχής σε καμία.

4.6.5 Συμπεράσματα του πειράματος

Το πείραμα οδηγεί σε τρία συμπεράσματα με πρακτική αξία.

Πρώτον, δεν υπάρχει γενικός κανόνας για τους περιορισμούς τύπου. Η αντιπαράθεση των δύο περιπτώσεων είναι εντυπωσιακή: ο τύπος Q214609 στο Medium προκάλεσε καθολική αποτυχία με 499 errors, ενώ ο τύπος Q5 στους δημιουργούς είχε επίδραση μικρότερη της μιας ποσοστιαίας μονάδας. Το πρόβλημα δεν είναι οι περιορισμοί καθαυτοί αλλά η αλληλεπίδραση συγκεκριμένου τύπου με τη συγκεκριμένη υπηρεσία.

Δεύτερον, η πραγματική αιτία του χαμηλού ποσοστού αυτόματης συμφιλίωσης εντοπίζεται αλλού: στο συντηρητικό κατώφλι απόφασης της υπηρεσίας, σε συνδυασμό με τη φύση των συμβατικών αρχαιολογικών ονομάτων. Η διαπίστωση αυτή ενισχύει το επιχείρημα για την αναγκαιότητα της χειροκίνητης επιμέλειας, καθώς δείχνει ότι καμία ρύθμιση των παραμέτρων δεν θα υποκαθιστούσε την ανθρώπινη κρίση.

Τρίτον, η μεθοδολογική αξία του ίδιου του πειράματος. Η αρχική υπόθεση ήταν εύλογη και συνεπής με το ένα από τα δύο παρατηρούμενα περιστατικά, αλλά αποδείχθηκε εσφαλμένη όταν ελέγχθηκε. Η καταγραφή της διάψευσης, αντί της αποσιώπησής της, αποτελεί στοιχείο επιστημονικής εγκυρότητας.

4.7 Αξιολόγηση του Εμπλουτισμού

4.7.1 Ποσοτικά αποτελέσματα

Η ανάκτηση της ιδιότητας P27 απέδωσε τιμή σε 291 από τις 501 εγγραφές, ποσοστό 58,1%. Το αποτέλεσμα είναι συνάρτηση δύο διαδοχικών περιορισμών: μόνο 389 εγγραφές διέθεταν συμφιλωμένο δημιουργό, και από αυτές μόνο ένα υποσύνολο αντιστοιχεί σε items που φέρουν δήλωση P27.

Πίνακας 4.11: Οι τιμές της ιδιότητας P27

Τιμή	Εγγραφές	Ποσοστό	Τύπος οντότητας
Classical Athens	264	52,7%	Αρχαία πόλη-κράτος
Greece	13	2,6%	Σύγχρονο κράτος
Delian League	12	2,4%	Αρχαία συμμαχία
Sparta	1	0,2%	Αρχαία πόλη-κράτος
Apulia	1	0,2%	Γεωγραφική περιοχή
χωρίς τιμή	210	41,9%	-



Σχήμα 4.4: Οι ανακτηθείσες τιμές εθνικότητας

Η αύξηση σε σχέση με το ενδιαμέσο στάδιο είναι αξιοσημείωτη. Πριν από τη χειροκίνητη επιμέλεια η ανάκτηση κάλυπτε 15,4% των εγγραφών· μετά, 58,1%. Η βελτίωση των 42,7 μονάδων ταυτίζεται ακριβώς με εκείνη της συμφιλίωσης, γεγονός που καταδεικνύει ότι κάθε επιπλέον αντιστοίχιση μεταφράζεται άμεσα σε επιπλέον ανακτημένη πληροφορία.

Από τις 389 αντιστοιχισμένες εγγραφές, οι 98 δεν απέκτησαν τιμή P27. Το κενό αυτό δεν οφείλεται στη μεθοδολογία αλλά στην ελλιπή τεκμηρίωση των ίδιων των Wikidata items, τα οποία δεν φέρουν όλα τη συγκεκριμένη δήλωση.

4.7.2 Η σημασιολογική ανεπάρκεια της ιδιότητας P27

Πέρα από το ποσοτικό αποτέλεσμα, οι ίδιες οι τιμές αποτελούν σημαντικό μεθοδολογικό εύρημα. Οι πέντε διακριτές τιμές ανήκουν σε τέσσερις διαφορετικές οντολογικές κατηγορίες: αρχαίες πόλεις-κράτη όπως η Classical Athens και η Sparta, αρχαία στρατιωτική συμμαχία όπως η Delian League, σύγχρονο κράτος όπως η Greece, και γεωγραφική περιοχή όπως η Apulia.

Η ανάμειξη δεν οφείλεται σε σφάλμα της μεθοδολογίας αλλά αντανακλά πραγματική ασυνέπεια στη μοντελοποίηση των Wikidata. Η ιδιότητα P27 ορίζεται ως country of citizenship και σχεδιάστηκε για τη σύγχρονη νομική έννοια της υπηκοότητας, η οποία προϋποθέτει εθνικό κράτος με θεσμοθετημένη ιθαγένεια. Η έννοια δεν έχει αντίστοιχο νόημα για έναν Αθηναίο αγγειογράφο του 5ου αιώνα π.Χ.

Το αποτέλεσμα είναι ότι διαφορετικοί επεξεργαστές συμπλήρωσαν την ιδιότητα με διαφορετική λογική: άλλοι με την πόλη-κράτος όπου έδρασε το πρόσωπο, άλλοι με το σύγχρονο κράτος στο έδαφος του οποίου βρίσκεται η αρχαία πόλη, άλλοι με την πολιτική συμμαχία της εποχής. Η ασυνέπεια καθιστά την ανακτηθείσα πληροφορία δύσκολα αξιοποιήσιμη για συγκριτική ανάλυση.

Το εύρημα οδηγεί σε συγκεκριμένη σύσταση για εφαρμογές σε αρχαιολογικές συλλογές: αντί για την P27, καταλληλότερες θα ήταν ιδιότητες όπως η P937 (work location), που δηλώνει τον τόπο δραστηριότητας χωρίς νομικές συνδηλώσεις, ή η P2596 (culture), που δηλώνει το πολιτισμικό πλαίσιο. Η επιλογή της κατάλληλης ιδιότητας δεν αποτελεί τεχνική λεπτομέρεια αλλά ουσιώδη σημασιολογική απόφαση.

4.8 Αξιολόγηση της Διαλειτουργικότητας

4.8.1 Η μετάβαση στο μοντέλο 5 αστεριών

Το μοντέλο 5 αστεριών του Berners-Lee παρέχει αντικειμενικό πλαίσιο αξιολόγησης της ωριμότητας ανοιχτών δεδομένων.

Πίνακας 4.12: Αξιολόγηση βάσει του μοντέλου 5 αστεριών

Κριτήριο	Πριν	Μετά	Παρατήρηση
Ανοιχτή άδεια	Ναι	Ναι	CC0 εξ αρχής
Μηχαναγνώσιμο	Ναι	Ναι	Μορφή CSV
Μη ιδιοκτησιακή μορφή	Ναι	Ναι	CSV, όχι XLS
URIs για οντότητες	Όχι	Μερικώς	Q-numbers
Διασύνδεση με LOD	Όχι	Δυνητική	Μέσω Wikidata
Επίπεδο	3 αστέρια	Προετοιμασία για 5 αστέρια	Μερική αναβάθμιση

Ο χαρακτηρισμός αυτός απαιτεί ωστόσο ουσιώδη επιφύλαξη. Το τέταρτο αστέρι, δηλαδή η χρήση URIs, επιτυγχάνεται μερικώς: οι οντότητες των δημιουργών και των υλικών απέκτησαν παγκόσμια αναγνωριστικά, αλλά το ίδιο το dataset παραμένει αρχείο CSV και δεν δημοσιεύεται σε μορφή RDF ούτε διαθέτει αποεπιλύσιμα URIs για τα αντικείμενα.

Το πέμπτο αστέρι, δηλαδή η πραγματική διασύνδεση με άλλα σύνολα, προϋποθέτει ότι τα δεδομένα είναι προσβάσιμα και συνδεδεμένα εντός του οικοσυστήματος. Το αρχείο QuickStatements παρήχθη και επαληθεύτηκε, αλλά δεν εκτελέστηκε στα Wikidata για τους δεοντολογικούς λόγους που εκτίθενται στην ενότητα 3.6. Χωρίς την εκτέλεση, η διασύνδεση παραμένει δυνητική.

Η ακριβέστερη διατύπωση είναι επομένως ότι η εργασία δεν ολοκλήρωσε τη μετάβαση σε πέντε αστέρια, αλλά προετοίμασε πλήρως το έδαφος για αυτήν. Τα δεδομένα βρίσκονται σε κατάσταση όπου η μετάβαση απαιτεί μόνο μία ενέργεια εκτέλεσης, η οποία εναπόκειται στο ίδρυμα που θα αναλάβει την ευθύνη. Για ολοκληρωμένη τεκμηρίωση πέντε αστεριών θα απαιτούνταν είτε η εκτέλεση της εισαγωγής είτε αυτοτελής δημοσίευση σε RDF με αποεπιλύσιμα URIs.

4.8.2 Πρακτικές συνέπειες

Η βελτίωση εκδηλώνεται σε τρία επίπεδα. Στην ανακαλυψιμότητα, καθώς τα δεδομένα γίνονται εντοπίσιμα μέσω ερωτημάτων SPARQL. Στη γλωσσική προσβασιμότητα, καθώς τα Wikidata items διαθέτουν ετικέτες σε εκατοντάδες γλώσσες και ένας Έλληνας ερευνητής μπορεί να εργαστεί στη γλώσσα του. Και στην τεχνική ενσωμάτωση, καθώς οι κωδικοί Q επιτρέπουν αυτόματη σύζευξη με οποιοδήποτε άλλο dataset χρησιμοποιεί τους ίδιους κωδικούς.

Το τελευταίο έχει ιδιαίτερη αξία για τη συγκεκριμένη συλλογή. Έργα του Makron και άλλων Αττικών αγγειογράφων φυλάσσονται σε δεκάδες μουσεία παγκοσμίως. Το ίδιο το item Q427706 καταγράφει συλλογές όπως το J. Paul Getty Museum, το Musée Saint-Raymond και η Antikensammlung Berlin. Η διασύνδεση επιτρέπει σε έναν ερευνητή να εντοπίσει με ένα ερώτημα το σύνολο των γνωστών έργων ενός αγγειογράφου, ανεξαρτήτως μουσείου.

Ιδιαίτερο ενδιαφέρον παρουσιάζει και η σύνδεση με εξειδικευμένους θησαυρούς. Η βάση Kerameikos.org, αφιερωμένη στην αρχαία ελληνική κεραμική, χρησιμοποιεί τα ίδια Wikidata identifiers. Η διασύνδεση της συλλογής του Met με τα Wikidata τη συνδέει επομένως έμμεσα και με τον εξειδικευμένο αυτό πόρο.

4.9 Συνολική Αποτίμηση - Συζήτηση

Η συνολική αποτίμηση οδηγεί σε πέντε συμπεράσματα.

Πρώτον, η μεθοδολογία λειτούργησε και παρήγαγε τα προβλεπόμενα παραδοτέα, με επαληθευμένη δομή και πλήρη εσωτερική συνέπεια.

Δεύτερον, η ποιότητα των αποτελεσμάτων είναι υψηλή και μετρήσιμη. Η ακρίβεια 100% και το F1-score 94,6% έναντι ανεξάρτητου χρυσού προτύπου συνιστούν ισχυρή τεκμηρίωση, σπάνια διαθέσιμη σε αντίστοιχες εργασίες.

Τρίτον, η ανάλυση των αποτυχιών αποδείχθηκε πιο διδακτική από την ανάλυση των επιτυχιών. Η διαπίστωση ότι ο Makron υπάρχει στα Wikidata με πλήρη τεκμηρίωση και ρητή σύνδεση με το Met, αλλά δεν αντιστοιχίστηκε αυτόματα, αναδεικνύει πραγματικό όριο των αυτόματων εργαλείων και τεκμηριώνει την αναγκαιότητα ανθρώπινης εποπτείας.

Τέταρτον, οι περιορισμοί τύπου στη συμφιλίωση αποδείχθηκαν προβληματικοί σε δύο ανεξάρτητες περιπτώσεις - στους δημιουργούς με τον τύπο Q5 και στα υλικά με τον Q214609, όπου προκάλεσαν καθολική αποτυχία. Η παρατήρηση αυτή αποτελεί πρακτική σύσταση προς όσους εφαρμόσουν παρόμοια μεθοδολογία.

Πέμπτον, η εφαρμογή σε αρχαιολογική συλλογή ανέδειξε σημασιολογικά ζητήματα που δεν θα εμφανίζονταν σε συλλογή νεότερης τέχνης: την ακαταλληλότητα της P27 για πρόσωπα της αρχαιότητας, τη δυσκολία αντιστοίχισης συμβατικών αρχαιολογικών ονομάτων, και τη συντριπτική κυριαρχία της ανωνυμίας που περιορίζει την εφαρμοσιμότητα της P170. Τα ζητήματα αυτά συνιστούν πρωτότυπη συνεισφορά.

Κεφάλαιο 5ο: Συμπεράσματα και Προτάσεις για Μελλοντική Έρευνα

5.1 Εισαγωγή

Το παρόν κεφάλαιο αποτελεί τον επίλογο της εργασίας. Συνοψίζει τα συμπεράσματα που προκύπτουν από τη συνδυαστική εξέταση του θεωρητικού πλαισίου, της μελέτης της ψηφιακής συλλογής, της μεθοδολογίας διασύνδεσης και της αξιολόγησης των αποτελεσμάτων. Συζητά με κριτική ειλικρίνεια τους περιορισμούς, και παρουσιάζει τεκμηριωμένες προτάσεις για μελλοντική έρευνα.

Η δομή ακολουθεί τη λογική της επιστημονικής εργασίας: απαντήσεις στα ερευνητικά ερωτήματα, συστηματοποίηση των συμπερασμάτων ανά άξονα, αυτοκριτική εξέταση των ορίων, και πλαίσιο μελλοντικής έρευνας με ιδιαίτερη έμφαση στην ελληνική πολιτιστική κληρονομιά.

5.2 Απαντήσεις στα Ερευνητικά Ερωτήματα

5.2.1 Είναι εφικτή η διασύνδεση με ανοιχτά εργαλεία;

Το θεμελιώδες ερώτημα της εργασίας ήταν αν η διασύνδεση μιας μουσειακής ψηφιακής συλλογής με τα Wikidata είναι εφικτή χρησιμοποιώντας αποκλειστικά ανοιχτά εργαλεία, χωρίς εξειδικευμένο προγραμματιστικό υπόβαθρο και χωρίς οικονομικό κόστος. Η απάντηση είναι καταφατική, με επιφυλάξεις που αξίζει να διατυπωθούν.

Η διασύνδεση 501 αντικειμένων επιτεύχθηκε με τρία εργαλεία - OpenRefine, Wikidata Reconciliation Service και QuickStatements - όλα διαθέσιμα δωρεάν. Δεν απαιτήθηκε συγγραφή κώδικα με την παραδοσιακή έννοια, παρά μόνο η χρήση τριών απλών εκφράσεων GREL. Δεν απαιτήθηκε εξειδικευμένος εξοπλισμός.

Οι επιφυλάξεις αφορούν τρία σημεία. Το OpenRefine απαιτήσε αύξηση του ορίου μνήμης για να χειριστεί dataset 484.956 γραμμών, παρέμβαση που προϋποθέτει εξοικείωση με τη δομή αρχείων της εφαρμογής. Η επέκταση Wikibase παρουσίασε σφάλμα εξαγωγής στην έκδοση 3.10.0, το οποίο απαίτησε εναλλακτική διαδρομή. Και η δήλωση νέων items στο schema δεν είναι προφανής στη διεπαφή, γεγονός που οδήγησε σε αρχικό λάθος. Η μεθοδολογία είναι επομένως προσβάσιμη, αλλά όχι χωρίς τριβή.

5.2.2 Ποια εμπόδια εμφανίζονται και πώς αντιμετωπίζονται;

Η εργασία εντόπισε πέντε κατηγορίες εμποδίων, καθεμία με διακριτή αντιμετώπιση.

Πρώτο εμπόδιο — οι περιορισμοί τύπου στη συμφιλίωση. Η πρώτη απόπειρα συμφιλίωσης του πεδίου Medium με περιορισμό τύπου Material προκάλεσε καθολική αποτυχία, με 499 σφάλματα εκτέλεσης, και χρειάστηκε επανάληψη χωρίς περιορισμό. Το ελεγχόμενο πείραμα που παρουσιάζεται στην ενότητα 4.6 έδειξε ωστόσο ότι ο αντίστοιχος περιορισμός στους δημιουργούς είχε επίδραση μόλις 0,6 ποσοστιαίων μονάδων. Το συμπέρασμα είναι ότι δεν υπάρχει γενικός κανόνας: η επίδραση εξαρτάται από τον συγκεκριμένο τύπο και πρέπει να ελέγχεται εμπειρικά σε κάθε περίπτωση.

Η αναπαραγωγικότητα του φιλτραρίσματος. Το Numeric Facet με συρόμενο δείκτη δεν επιτρέπει ακριβή επανάληψη. Η αντικατάστασή του με Custom Text Facet και ρητή λογική συνθήκη έλυσε το πρόβλημα οριστικά.

Τα συμβατικά αρχαιολογικά ονόματα. Ονόματα όπως Veii Painter ή Boot Painter αντιστοιχούν σε ερευνητικές υποθέσεις και όχι σε τεκμηριωμένα πρόσωπα, με ασταθή απόδοση στη διεθνή

βιβλιογραφία. Η αντιμετώπιση ήταν η κατά περίπτωση εξέταση και, όπου χρειάστηκε, η αποχή από την αντιστοίχιση.

Η ασυμμετρία της κατανομής. Ένας δημιουργός συγκέντρωνε το ένα τρίτο του δείγματος. Αντί για εμπόδιο, η ασυμμετρία αξιοποιήθηκε: η στρατηγική προτεραιοποίησης κατά συχνότητα μετέτρεψε το χαρακτηριστικό αυτό σε πλεονέκτημα.

Η απαίτηση τεκμηρίωσης. Κάθε δήλωση συνοδεύτηκε από αναφορά P248 προς Q160236, εξασφαλίζοντας συμμόρφωση με τις αρχές επαληθευσιμότητας των Wikidata.

5.2.3 Ποια είναι η αξία της διασύνδεσης;

Η αξία αποδείχθηκε αμφίδρομη και μετρήσιμη.

Για τη συλλογή, τα οφέλη είναι τέσσερα. Εμπλουτισμός με 291 τιμές εθνικότητας που δεν υπήρχαν στο αρχικό αρχείο. Διαλειτουργικότητα μέσω 63 μοναδικών Q-numbers δημιουργών και 2 υλικών. Ανακαλυψιμότητα μέσω ένταξης στο οικοσύστημα LOD. Και εντοπισμός 14 αντιστοιχίσεων που το ίδιο το μουσείο δεν είχε καταγράψει.

Για τα Wikidata, η εισαγωγή τεκμηριωμένων δηλώσεων από αξιόπιστη πηγή θα εμπλούτιζε τη βάση με 501 νέα items και 1.389 δηλώσεις. Η εργασία επέλεξε συνειδητά να μην προχωρήσει στην εκτέλεση, για λόγους που αναλύονται στην ενότητα 5.4.4.

5.3 Συστηματοποίηση των Συμπερασμάτων

5.3.1 Μεθοδολογικά συμπεράσματα

Η βασικότερη μεθοδολογική συνεισφορά είναι η ανάπτυξη ενός τετρασταδιακού πλαισίου που είναι ταυτόχρονα αποτελεσματικό και πλήρως αναπαραγώγιμο. Η αναπαραγωγιμότητα δεν αποτελεί δευτερεύον χαρακτηριστικό αλλά κεντρικό επίτευγμα, τεκμηριωμένο μέσω του αρχείου JSON του ιστορικού ενεργειών.

Η πρακτική αξία επιβεβαιώθηκε εμπειρικά. Η χειροκίνητη επανάληψη της διαδικασίας δεν αναπαρήγαγε πιστά το αποτέλεσμα, καθώς ο ακριβής προσδιορισμός αριθμητικού εύρους μέσω συρόμενου δείκτη είναι πρακτικά ανέφικτος. Το εύρημα καταδεικνύει ότι η αυτόματη καταγραφή δεν αποτελεί απλώς καλή πρακτική τεκμηρίωσης αλλά αναγκαία προϋπόθεση επιστημονικής εγκυρότητας για εργασίες που βασίζονται σε διαδραστική επεξεργασία δεδομένων.

Δεύτερο μεθοδολογικό συμπέρασμα αφορά τη στρατηγική της χειροκίνητης επιμέλειας. Η προτεραιοποίηση κατά συχνότητα απέδωσε 71 εγγραφές ανά ενέργεια στα τρία πρώτα ονόματα, έναντι 1,8 στα υπόλοιπα εξήντα δύο. Η αρχή είναι γενικεύσιμη σε κάθε dataset με ασύμμετρη κατανομή, που αποτελεί τον κανόνα σε μουσειακές συλλογές.

Τρίτο συμπέρασμα αφορά τον ρόλο της ανθρώπινης κρίσης. Η επιμέλεια δεν συνίσταται στη μεγιστοποίηση των αντιστοιχίσεων αλλά στην ορθότητά τους. Η τεκμηριωμένη άρνηση αντιστοίχισης της τιμής Brygan συνέβαλε στη διατήρηση της τέλει ακρίβειας εξίσου με τις τρεις θετικές αποφάσεις.

5.3.2 Εμπειρικά συμπεράσματα

Σε εμπειρικό επίπεδο η εργασία παρήγαγε συγκεκριμένα και επαληθευμένα αποτελέσματα. Η συμφιλίωση των δημιουργών απέδωσε 389 αντιστοιχίσεις επί 501 εγγραφών, που αντιστοιχούν σε 63 μοναδικά Wikidata items. Η συμφιλίωση των υλικών κάλυψε 499 εγγραφές μέσω δύο μόνο ενεργειών. Ο εμπλουτισμός απέδωσε 291 τιμές εθνικότητας.

Σημαντικό εμπειρικό εύρημα προέκυψε από την αξιολόγηση έναντι του θεσμικού συνόλου αναφοράς. Η αξιοποίηση της στήλης Artist Wikidata URL ως ανεξάρτητου σημείου σύγκρισης επέτρεψε τον υπολογισμό αντικειμενικών μετρικών: πλήρης συμφωνία στις 375 επαληθεύσιμες περιπτώσεις, ανάκληση 89,7% και F1-score 94,6%. Οι τιμές αυτές εκφράζουν βαθμό συμφωνίας με τις καταχωρίσεις του μουσείου και όχι απόλυτη ορθότητα, όπως αναλύεται στην ενότητα 4.4.

Εξίσου σημαντικό είναι το αποτέλεσμα του ελεγχόμενου πειράματος για τους περιορισμούς τύπου. Η αρχική υπόθεση ότι ο περιορισμός Q5 ευθυνόταν για το χαμηλό ποσοστό αυτόματης συμπίλωσης διαψεύστηκε: η διαφορά με και χωρίς περιορισμό ανέρχεται σε 0,6 ποσοστιαίες μονάδες, ενώ ο πολυπληθέστερος δημιουργός παραμένει ασυμφιλίωτος και στις δύο εκτελέσεις. Το εύρημα μετατοπίζει την αιτία στο κατώφλι απόφασης της υπηρεσίας και ενισχύει το επιχείρημα υπέρ της χειροκίνητης επιμέλειας.

Η δυνατότητα αυτή σπανίως υπάρχει σε αντίστοιχες εργασίες και αποτελεί μεθοδολογική συνεισφορά από μόνη της. Το γεγονός ότι πολλά σύγχρονα μουσειακά datasets περιλαμβάνουν πλέον συνδέσμους προς εξωτερικά authority files σημαίνει ότι η προσέγγιση μπορεί να γενικευτεί.

5.3.3 Συμπεράσματα για τις αρχαιολογικές συλλογές

Η εφαρμογή σε συλλογή ελληνικής αρχαιότητας ανέδειξε τρία ζητήματα που δεν εμφανίζονται σε συλλογές νεότερης τέχνης και συνιστούν πρωτότυπη συνεισφορά.

Πρώτον, η κυριαρχία της ανωνυμίας. Το 93,6% των αντικειμένων του τμήματος στερείται επώνυμου δημιουργού, γεγονός που περιορίζει δραστικά την εφαρμοσιμότητα της ιδιότητας P170. Για αρχαιολογικό υλικό, ιδιότητες όπως η P2596 (culture), η P571 (inception) ή η P189 (location of discovery) θα είχαν πολύ ευρύτερη κάλυψη.

Δεύτερον, η φύση των συμβατικών ονομάτων. Η παράδοση Beazley απέδωσε ονόματα σε ανώνυμους τεχνίτες βάσει στυλιστικής ανάλυσης. Οι οντότητες αυτές είναι ταυτόχρονα υποθετικά πρόσωπα και κατηγορίες έργων, αμφισημία που δυσχεραίνει την αυτόματη αντιστοίχιση και θέτει οντολογικά ερωτήματα για την αναπαράστασή τους.

Τρίτον, η σημασιολογική ανεπάρκεια της P27. Οι πέντε τιμές που ανακτήθηκαν ανήκουν σε τέσσερις διαφορετικές οντολογικές κατηγορίες, αντανακλώντας το γεγονός ότι η έννοια της υπηκοότητας δεν έχει νόημα για πρόσωπα της αρχαιότητας. Η παρατήρηση έχει άμεση πρακτική συνέπεια για όποιον σχεδιάσει αντίστοιχο mapping.

5.3.4 Το εύρος ισχύος των ευρημάτων

Τα ποσοτικά αποτελέσματα της εργασίας απαιτούν προσεκτική οριοθέτηση ως προς την εμβέλειά τους.

Η ακρίβεια 100% δεν σημαίνει ότι η μεθοδολογία είναι αλάνθαστη, αλλά ότι καμία λανθασμένη αντιστοίχιση δεν εντοπίστηκε εντός του διαθέσιμου συνόλου ελέγχου. Το θεσμικό σύνολο αναφοράς κάλυπτε 418 από τις 501 εγγραφές, και από αυτές μόνο οι 375 είχαν αντίστοιχη πρόβλεψη. Η μέτρηση αφορά επομένως αυτό το υποσύνολο και όχι το δείγμα στο σύνολό του, πολύ περισσότερο δε ολόκληρη τη συλλογή.

Το δείγμα των 501 αντικειμένων δεν συνιστά μεγάλο σύνολο δεδομένων με τα σύγχρονα μέτρα. Αντιστοιχεί στο 1,5% του τμήματος Ελληνικής και Ρωμαϊκής Τέχνης και σε ποσοστό μικρότερο του 0,2% του συνολικού dataset. Επιπλέον, χαρακτηρίζεται από ακραία ομοιογένεια: το 99,2% των αντικειμένων είναι θραύσματα αττικών αγγείων από πηλό, με κοινή προέλευση από μία δωρεά.

Η ομοιογένεια αυτή επηρεάζει άμεσα τα αποτελέσματα. Το ποσοστό 99,6% στη συμφιλίωση των υλικών οφείλεται στο γεγονός ότι το πεδίο Medium περιείχε μόλις δύο μοναδικές τιμές. Οποιαδήποτε εικασία ότι ένα διαφορετικό σύνολο δεδομένων, όπως για παράδειγμα μια συλλογή νομισμάτων με ποικιλία μετάλλων, κραμάτων και τεχνικών κοπής, θα απέδιδε αντίστοιχες μετρικές δεν τεκμηριώνεται από την παρούσα εργασία και θα απαιτούσε ανεξάρτητη δοκιμή.

Οι μετρικές που παρουσιάστηκαν αφορούν επομένως τα συγκεκριμένα δεδομένα υπό τις συγκεκριμένες συνθήκες. Η αξία τους έγκειται στο ότι τεκμηριώνουν τη λειτουργικότητα της μεθοδολογίας και αποκαλύπτουν συστηματικές συμπεριφορές των εργαλείων, όχι στο ότι προβλέπουν την απόδοση σε άλλα σύνολα.

5.3.5 Η αναγκαιότητα της χειροκίνητης επιμέλειας

Το σημαντικότερο πρακτικό συμπέρασμα της εργασίας είναι ότι η αυτόματη συμφιλίωση δεν επαρκεί από μόνη της και η ανθρώπινη παρέμβαση παραμένει αναγκαία.

Η αυτόματη διαδικασία κάλυψε μόλις το 34,9% των εγγραφών. Η χειροκίνητη επιμέλεια τριών μόνο ονομάτων ανέβασε το ποσοστό στο 77,6% και βελτίωσε το F1-score από 55,5% σε 94,6%, χωρίς καμία απώλεια ακρίβειας. Η συνεισφορά της ανθρώπινης κρίσης δηλαδή δεν ήταν συμπληρωματική αλλά καθοριστική.

Η αναγκαιότητα αυτή δεν οφείλεται σε ατέλεια των εργαλείων αλλά στη φύση του προβλήματος. Το reconciliation service λειτουργεί συντηρητικά: αποφεύγει να αποφασίσει όταν δεν διαθέτει επαρκή βεβαιότητα, επιλογή που εξασφαλίζει την ακρίβεια αλλά αφήνει μεγάλο μέρος της εργασίας ανολοκλήρωτο. Η περίπτωση του Makron το καταδεικνύει με σαφήνεια: η σωστή οντότητα είχε εντοπιστεί και προταθεί, αλλά η τελική απόφαση απαιτούσε ανθρώπινη κρίση.

Εξίσου σημαντική είναι η αντίστροφη διάσταση. Η χειροκίνητη επιμέλεια δεν συνίσταται μόνο στην προσθήκη αντιστοιχίσεων αλλά και στην τεκμηριωμένη άρνησή τους. Η απόφαση να μην αντιστοιχιστεί η τιμή Brygan, επειδή τα δεδομένα δεν επέτρεπαν ασφαλή ταυτοποίηση, συνέβαλε στη διατήρηση της τέλει ακρίβειας εξίσου με τις τρεις θετικές αποφάσεις.

Το συμπέρασμα για την πρακτική εφαρμογή είναι ότι κάθε έργο διασύνδεσης πρέπει να προβλέπει ρητά χρόνο και εξειδίκευση για τη φάση της επιμέλειας. Η αντιμετώπισή της ως προαιρετικού σταδίου βελτίωσης οδηγεί σε αποτελέσματα σημαντικά κατώτερα των δυνατοτήτων της μεθοδολογίας.

5.3.6 Συμβολή στη βιβλιογραφία

Η εργασία συνεισφέρει με τέσσερις τρόπους. Παρέχει πλήρως τεκμηριωμένη μεθοδολογία με λεπτομερή περιγραφή κάθε βήματος, χρήσιμη ως πρακτικός οδηγός. Επαληθεύει σε νέο πλαίσιο ευρήματα διεθνών ερευνών σχετικά με την προσβασιμότητα του OpenRefine και τη σημασία των αναφορών πηγής. Εισάγει τη χρήση υπαρχόντων συνδέσμων authority ως χρυσό προτύπου για αντικειμενική αξιολόγηση. Και τεκμηριώνει τις ιδιαιτερότητες των αρχαιολογικών συλλογών, τομέα υποεκπροσωπούμενο στη σχετική βιβλιογραφία που εστιάζει κυρίως σε συλλογές νεότερης τέχνης.

Επιπλέον, αποτελεί μία από τις πρώτες ελληνόγλωσσες εργασίες που πραγματεύονται συστηματικά το θέμα, καλύπτοντας κενό στην ελληνική βιβλιογραφία και προτείνοντας ελληνική ορολογία για έννοιες που μέχρι σήμερα συζητούνται κυρίως στα αγγλικά.

5.4 Περιορισμοί της Εργασίας

5.4.1 Περιορισμοί δείγματος

Το δείγμα των 501 εγγραφών αντιστοιχεί στο 1,5% του τμήματος Greek and Roman Art και στο 0,1% του συνολικού dataset. Αν και αποτελεί το 23,4% των έργων του τμήματος με επώνυμο δημιουργό, δεν είναι στατιστικά τυχαίο.

Η μέθοδος επιλογής - συνεχές εύρος Object ID - παρήγαγε δείγμα με εξαιρετική ομοιογένεια: 99,2% θραύσματα αττικών αγγείων από πηλό, με 98,8% κοινή προέλευση από τη δωρεά von Bothmer. Η ομοιογένεια διευκόλυνε την ανάλυση αλλά περιορίζει τη γενικευσιμότητα. Σε δείγμα με ποικιλία υλικών, τα αποτελέσματα της συμφιλίωσης του πεδίου Medium θα ήταν πιθανότατα ουσιωδώς χαμηλότερα.

5.4.2 Περιορισμοί του mapping

Η εργασία εστίασε σε τέσσερις ιδιότητες: P170, P186, P31 και P248, με ανάκτηση της P27. Το dataset περιλαμβάνει 54 στήλες, πολλές από τις οποίες φέρουν αξιοποιήσιμη πληροφορία που παρέμεινε αδιασύνδετη - χρονολόγηση, πολιτισμός, περίοδος, τύπος αντικειμένου, διαστάσεις.

Ιδιαίτερα αισθητή είναι η απουσία της χρονολόγησης. Τα πεδία Object Begin Date και Object End Date είναι συμπληρωμένα σχεδόν καθολικά και θα μπορούσαν να αντιστοιχιστούν στην ιδιότητα P571, προσφέροντας πληροφορία πολύ ευρύτερης κάλυψης από την P170.

5.4.3 Περιορισμοί της αξιολόγησης

Το θεσμικό σύνολο αναφοράς θεωρήθηκε έγκυρο χωρίς ανεξάρτητο έλεγχο. Αν και η υπόθεση είναι εύλογη δεδομένης της προέλευσής του, αυστηρότερη αξιολόγηση θα απαιτούσε δειγματοληπτική επαλήθευση των ίδιων των καταχωρίσεων του μουσείου.

Επιπλέον, το πρότυπο καλύπτει 418 από τις 501 εγγραφές. Για τις υπόλοιπες 83 δεν υπάρχει σημείο αναφοράς, οπότε οι μετρικές αφορούν υποσύνολο και όχι το πλήρες δείγμα.

Τέλος, δεν πραγματοποιήθηκε μέτρηση συμφωνίας μεταξύ κριτών. Οι αποφάσεις της χειροκίνητης επιμέλειας λήφθηκαν από έναν ερευνητή, και η επανάληψη από δεύτερο θα επέτρεπε την εκτίμηση της υποκειμενικότητας.

5.4.4 Ο περιορισμός της μη εκτέλεσης

Το αρχείο QuickStatements παρήχθη και επαληθεύτηκε αλλά δεν εκτελέστηκε. Η επιλογή ήταν συνειδητή και στηρίχθηκε σε τρεις λόγους.

Η μαζική δημιουργία 501 νέων items σε δημόσια βάση γνώσης επιβαρύνει την κοινότητα με το κόστος ελέγχου. Οι κατευθυντήριες γραμμές των Wikidata προβλέπουν κριτήρια αξιοσημείωτου, και η καταλληλότητα θραυσμάτων αγγείων για αυτοτελή καταχώριση δεν είναι δεδομένη. Και το ίδιο το σύστημα επικύρωσης επισήμανε ότι τα items θα δημιουργούνταν χωρίς περιγραφές, στοιχείο που δυσχεραίνει τη διάκρισή τους.

Η επιλογή αυτή περιορίζει την εργασία υπό μία έννοια: η μεθοδολογία δεν δοκιμάστηκε στο τελευταίο της βήμα. Ταυτόχρονα όμως αντανακλά ώριμη στάση απέναντι σε κοινό αγαθό, και το αρχείο παραμένει διαθέσιμο για ίδρυμα που θα αναλάβει την ευθύνη μετά την απαραίτητη συνεννόηση.

5.5 Προτάσεις για Μελλοντική Έρευνα

5.5.1 Βελτιστοποίηση των παραμέτρων συμφιλίωσης

Η πλέον άμεση κατεύθυνση προκύπτει από τα ίδια τα ευρήματα. Η εργασία διαπίστωσε ότι οι περιορισμοί τύπου προκάλεσαν προβλήματα σε δύο ανεξάρτητες περιπτώσεις, χωρίς όμως να ποσοτικοποιήσει την επίδρασή τους.

Δεδομένου ότι το θεσμικό σύνολο αναφοράς είναι διαθέσιμο, μια συγκριτική μελέτη θα μπορούσε να μετρήσει με ακρίβεια τη διαφορά. Θα εκτελούσε τη συμφιλίωση στο ίδιο δείγμα με διαφορετικές ρυθμίσεις - με και χωρίς περιορισμό τύπου, με και χωρίς βοηθητικές ιδιότητες όπως η χρονολογία γέννησης - και θα συνέκρινε precision και recall σε κάθε περίπτωση.

Η παρούσα εργασία υλοποίησε ήδη ένα σκέλος αυτής της κατεύθυνσης, μετρώντας την επίδραση του περιορισμού τύπου. Απομένει η διερεύνηση των βοηθητικών ιδιοτήτων: το OpenRefine επιτρέπει τον ορισμό πρόσθετων στηλών ως reconciliation properties, ώστε η υπηρεσία να αξιοποιεί επιπλέον κριτήρια αποσαφήνισης. Η σύνδεση της στήλης Artist Begin Date με την ιδιότητα P569 θα παρείχε χρονολογικό κριτήριο, το οποίο για συμβατικά αρχαιολογικά ονόματα θα μπορούσε να αποδειχθεί καθοριστικό. Δεδομένου ότι το θεσμικό σύνολο αναφοράς είναι διαθέσιμο, η μέτρηση της επίδρασης θα ήταν άμεση.

5.5.2 Επέκταση του mapping

Η δεύτερη κατεύθυνση αφορά τη διεύρυνση των ιδιοτήτων. Ο πίνακας που ακολουθεί παρουσιάζει συγκεκριμένες προτάσεις με εκτίμηση της κάλυψης και των προκλήσεων.

Πίνακας 5.1: Προτεινόμενες επεκτάσεις του mapping

Πεδίο	Wikidata property	Κωδικός	Κάλυψη	Πρόκληση
Object Begin/End Date	inception	P571	Σχεδόν καθολική	Εύρη και π.Χ. χρονολογίες
Culture	culture	P2596	95,6%	Ελεγχόμενο λεξιλόγιο
Object Name	instance of	P31	Καθολική	Ιεραρχία τύπων αγγείων
Period	period	P2348	Υψηλή	Αντιστοίχιση περιόδων
Dimensions	dimensions	P2048	Υψηλή	Ανάλυση μονάδων
Credit Line	donated by	P1028	98,8%	Ταυτοποίηση δωρητή

Ιδιαίτερη αξία θα είχε η αντιστοίχιση της χρονολόγησης. Σε αντίθεση με τον δημιουργό, που είναι γνωστός μόνο για το 6,4% της συλλογής, η χρονολόγηση υπάρχει σχεδόν παντού. Η πρόκληση είναι τεχνική: οι τιμές είναι αρνητικές, δηλώνοντας π.Χ. χρονολογίες, και συχνά εκφράζουν εύρη αντί για συγκεκριμένα έτη.

Ενδιαφέρουσα είναι και η πρόταση για την P1028. Δεδομένου ότι το 98,8% του δείγματος προέρχεται από τη δωρεά von Bothmer, η αντιστοίχιση θα συνέδεε τα αντικείμενα με τον δωρητή τους, πληροφορία με ιστορική και ερευνητική αξία.

5.5.3 Εφαρμογή σε ελληνικές μουσειακές συλλογές

Η σημαντικότερη κατεύθυνση για τον ελληνικό χώρο είναι η εφαρμογή της μεθοδολογίας σε εγχώριες συλλογές. Ιδρύματα όπως το Εθνικό Αρχαιολογικό Μουσείο, το Μουσείο της Ακρόπολης, το Βυζαντινό

και Χριστιανικό Μουσείο, η Εθνική Πινακοθήκη και το Μουσείο Μπενάκη διαθέτουν συλλογές διεθνούς σημασίας που παραμένουν σε μεγάλο βαθμό εκτός του οικοσυστήματος LOD.

Η παρούσα εργασία προσφέρει ιδιαίτερα χρήσιμο υπόβαθρο για τέτοιο εγχείρημα, καθώς το δείγμα της αφορά ακριβώς το ίδιο είδος υλικού: αρχαία ελληνική κεραμική. Τα ευρήματα για την ανωνυμία, τα συμβατικά ονόματα και την ακαταλληλότητα της P27 μεταφέρονται αυτούσια.

Θα προέκυπταν ωστόσο και πρόσθετες προκλήσεις. Η γραφή ονομάτων σε ελληνικό αλφάβητο απαιτεί προσοχή, καθώς τα Wikidata items ενδέχεται να φέρουν ελληνική ετικέτα, λατινική μεταγραφή ή και τα δύο. Πολλές ελληνικές οντότητες - τοπωνύμια, τεχνικές, τύποι αγγείων - ενδέχεται να απουσιάζουν και να απαιτούν δημιουργία νέων items. Και η ελληνική βιβλιογραφική παράδοση χρησιμοποιεί ενίοτε διαφορετικές συμβατικές ονομασίες από τη διεθνή.

Το όφελος θα ήταν αντίστοιχα μεγάλο. Κάθε νέο item για ελληνικό καλλιτέχνη, τεχνική ή τοποθεσία εμπλουτίζει τον παγκόσμιο γράφο γνώσης με ελληνικό πολιτιστικό περιεχόμενο. Ιδιαίτερα σημαντική θα ήταν η σύνδεση με εξειδικευμένους θησαυρούς όπως το Kerameikos.org, το οποίο ήδη χρησιμοποιεί Wikidata identifiers και αποτελεί φυσικό εταίρο για ελληνικά ιδρύματα.

5.5.4 Ανάπτυξη τυποποιημένης διαδικασίας αξιολόγησης

Η παρούσα εργασία κατέδειξε ότι η αντικειμενική αξιολόγηση είναι εφικτή όταν υπάρχουν υπάρχοντες σύνδεσμοι authority. Η προσέγγιση αυτή θα μπορούσε να συστηματοποιηθεί.

Πολλά μουσειακά datasets περιλαμβάνουν πλέον συνδέσμους προς Wikidata, VIAF, Getty ULAN ή άλλα authority files. Η ανάπτυξη τυποποιημένης διαδικασίας που αξιοποιεί αυτούς τους συνδέσμους ως θεσμικό σύνολο αναφοράς θα επέτρεπε τη συγκρίσιμη αποτίμηση διαφορετικών στρατηγικών συμφιλίωσης σε διαφορετικά ιδρύματα.

Μια τέτοια διαδικασία θα μπορούσε να υλοποιηθεί ως επέκταση του OpenRefine ή ως αυτόνομο εργαλείο, δεχόμενη ως είσοδο ένα dataset με στήλη authority και παράγοντας αυτόματα τις μετρικές precision, recall και F1.

5.5.5 Αξιοποίηση μεγάλων γλωσσικών μοντέλων

Η πέμπτη κατεύθυνση αφορά την αυτοματοποίηση της αποσαφήνισης μέσω Μεγάλων Γλωσσικών Μοντέλων. Τα ευρήματα της εργασίας υποδεικνύουν ότι η προοπτική είναι ρεαλιστική για συγκεκριμένο τύπο περιπτώσεων.

Στις 268 εγγραφές που είχαν υποψηφίους αλλά δεν έλαβαν αυτόματη απόφαση, το ζητούμενο δεν είναι η ανακάλυψη άγνωστης πληροφορίας αλλά η επιλογή μεταξύ γνωστών εναλλακτικών βάσει συμφραζομένων. Πρόκειται ακριβώς για τον τύπο εργασίας στον οποίο τα γλωσσικά μοντέλα αποδίδουν.

Ένα υποθετικό σενάριο θα παρείχε στο μοντέλο το όνομα από το dataset, τα συμφραζόμενα - χρονολογία, υλικό, τύπος αντικειμένου, τμήμα - και τη λίστα υποψηφίων με τα χαρακτηριστικά τους, ζητώντας αξιολόγηση με αιτιολόγηση. Ο ερευνητής θα επικύρωνε ή θα απέρριπτε την πρόταση.

Η προσέγγιση δεν θα εξάλειφε την ανθρώπινη εποπτεία αλλά θα τη μετέτρεπε από εξαντλητική εξέταση σε επικύρωση προτάσεων. Κρίσιμο ερευνητικό ερώτημα παραμένει η μέτρηση του ποσοστού σφαλμάτων που θα εισήγαγε - ιδίως δεδομένου ότι η παρούσα εργασία πέτυχε ακρίβεια 100% με ανθρώπινη κρίση, και οποιαδήποτε αυτοματοποίηση θα έπρεπε να διατηρήσει αντίστοιχο επίπεδο.

5.5.6 Μελέτη της αξιοσημείωτης καταχώρισης αρχαιολογικού υλικού

Η έκτη πρόταση προκύπτει από το δεοντολογικό ζήτημα που ανέκυψε. Το ερώτημα αν μεμονωμένα θραύσματα αγγείων δικαιολογούν αυτοτελή καταχώριση στα Wikidata δεν έχει προφανή απάντηση και δεν έχει συζητηθεί συστηματικά.

Μια μελέτη που θα εξέταζε τις υπάρχουσες πρακτικές - πόσα αρχαιολογικά αντικείμενα έχουν ήδη items, με ποια κριτήρια, από ποια ιδρύματα - και θα διατύπωνε προτάσεις κατευθυντήριων γραμμών, θα είχε άμεση πρακτική αξία για την κοινότητα των GLAM.

5.5.7 Εκπαιδευτικές εφαρμογές

Τέλος, η αναλυτική τεκμηρίωση της εργασίας, σε συνδυασμό με τη χρήση ελεύθερων εργαλείων και ανοιχτών δεδομένων, την καθιστά κατάλληλη βάση για εκπαιδευτικά εργαστήρια στον χώρο των ψηφιακών ανθρωπιστικών επιστημών.

Ιδιαίτερη αξία θα είχε η ανάπτυξη υλικού στα ελληνικά, καθώς η υπάρχουσα τεκμηρίωση του OpenRefine και των Wikidata είναι σχεδόν αποκλειστικά αγγλόφωνη. Η εμπειρία της κοινότητας Wikimedia με τα GLAM-Wiki workshops δείχνει ότι τέτοιες δράσεις είναι αποτελεσματικές τόσο ως εκπαιδευτικά εργαλεία όσο και ως μηχανισμοί παραγωγής πραγματικού περιεχομένου.

5.6 Τελικά Σχόλια

Η παρούσα διπλωματική εργασία εκκίνησε από ένα σαφές ερώτημα: είναι δυνατή η διασύνδεση μιας μουσειακής ψηφιακής συλλογής με τα Wikidata χρησιμοποιώντας ανοιχτά εργαλεία, χωρίς εξειδικευμένο τεχνικό υπόβαθρο και χωρίς οικονομικό κόστος; Η απάντηση, που προέκυψε από τη συστηματική εφαρμογή μεθοδολογίας σε 501 αντικείμενα ελληνικής αρχαιότητας, είναι καταφατική - με τις επιφυλάξεις και τους περιορισμούς που καταγράφηκαν ειλικρινώς.

Πέρα από το τεχνικό επίτευγμα, η εργασία αναδεικνύει μια ευρύτερη μετατόπιση στον τρόπο που αντιλαμβανόμαστε τα πολιτιστικά δεδομένα. Ένα θραύσμα αττικής κύλικας δεν είναι πλέον απομονωμένη εγγραφή σε τοπικό κατάλογο, αλλά κόμβος σε παγκόσμιο δίκτυο γνώσης - συνδεδεμένος με τον αγγειογράφο του, το υλικό του, την εποχή του, και μέσω αυτών με χιλιάδες άλλα αντικείμενα σε συλλογές ανά τον κόσμο.

Η περίπτωση του Makron είναι εύλωτη. Ο αγγειογράφος που εργάστηκε στην Αθήνα τη δεκαετία του 480 π.Χ. έχει σήμερα Wikidata item που καταγράφει έργα του σε μουσεία τεσσάρων ηπείρων. Η διασύνδεση της συλλογής του Met προσθέτει 172 ακόμη ψηφίδες σε αυτό το παγκόσμιο ψηφιδωτό. Πρόκειται για μορφή επανένωσης ενός διασκορπισμένου έργου που θα ήταν αδύνατη με φυσικά μέσα.

Τα Linked Open Data και τα Wikidata δεν αποτελούν πλέον εξειδικευμένο τεχνολογικό πεδίο προσβάσιμο μόνο σε ειδικούς. Εργαλεία όπως το OpenRefine έχουν καταστήσει τη διασύνδεση προσιτή στους ίδιους τους επαγγελματίες των GLAM ιδρυμάτων. Αυτή η δημοκρατικοποίηση αποτελεί ίσως τη σημαντικότερη εξέλιξη στον χώρο της ψηφιακής πολιτιστικής κληρονομιάς της τελευταίας δεκαετίας.

Στον ελληνικό χώρο η πορεία μόλις αρχίζει. Η εργασία φιλοδοξεί να αποτελέσει χρήσιμη αφετηρία για όσους θελήσουν να τη συνεχίσουν - ερευνητές, φοιτητές, επαγγελματίες μουσείων και ιδρύματα που επιθυμούν να καταστήσουν τη γνώση τους ανοιχτή, διασυνδεδεμένη και προσβάσιμη στον κόσμο.

Αναφορές

Οι αναφορές παρατίθενται με τη σειρά που εμφανίζονται στο κείμενο, σύμφωνα με το πρότυπο IEEE. Κάθε αναφορά φέρει αριθμό εντός αγκύλης, ο οποίος χρησιμοποιείται για την παραπομπή εντός του κειμένου.

- [1] T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web," *Scientific American*, vol. 284, no. 5, pp. 34–43, May 2001.
- [2] W3C, "RDF 1.1 Concepts and Abstract Syntax," W3C Recommendation, Feb. 2014. [Online]. Available: <https://www.w3.org/TR/rdf11-concepts/>
- [3] T. Berners-Lee, "Linked Data," W3C Design Issues, 2006 (updated 2009). [Online]. Available: <https://www.w3.org/DesignIssues/LinkedData.html>
- [4] Wikimedia Foundation, "Wikidata: Introduction," [wikidata.org](https://www.wikidata.org/wiki/Wikidata:Introduction). [Online]. Available: <https://www.wikidata.org/wiki/Wikidata:Introduction>
- [5] A. Fagerving, A. Andersson, and M. Berggren, "Wikidata for authority control: sharing museum knowledge with the world," *Journal of Open Humanities Data*, vol. 9, 2023. doi: 10.5334/johd.138
- [6] G. Candela, N. Pearson, and S. Chambers, "A Systematic Review of Wikidata in GLAM Institutions: a Labs Approach," in *Proc. 28th Int. Conf. on Theory and Practice of Digital Libraries (TPDL 2024)*, Ljubljana, Slovenia, Sep. 2024. doi: 10.1007/978-3-031-72440-4_4
- [7] Dublin Core Metadata Initiative, "Dublin Core Metadata Element Set, Version 1.1," DCMI Recommendation. [Online]. Available: <https://www.dublincore.org/specifications/dublin-core/dces/>
- [8] CIDOC CRM Special Interest Group, "Definition of the CIDOC Conceptual Reference Model," Version 7.1.3, ISO 21127:2023. [Online]. Available: <https://cidoc-crm.org/>
- [9] G. Candela, M. Dobрева, and P. Escobar, "Open Data Practices of Art Museums in Wikidata: A Compliance Assessment," *Journal of Open Humanities Data*, Dec. 2025. doi: 10.5334/johd.438
- [10] M. Rossenova, L. Tolomelli, and F. Ciotti, "Empowering Italian GLAMs: A Workflow for Building a Comprehensive Open Dataset of Museums on Wikidata," *Journal of Open Humanities Data*, Jan. 2026. doi: 10.5334/johd.477
- [11] C. Nakasone, J. Bullard, and A. Rasmussen, "Towards a Linked Open Index: Reconciling Museum Records to Wikidata for Index of American Design Constituents," *Journal of Open Humanities Data*, Jan. 2026. doi: 10.5334/johd.426
- [12] Metropolitan Museum of Art, "The Metropolitan Museum of Art's Open Access Initiative," [metmuseum.org](https://www.metmuseum.org/about-the-met/policies-and-documents/open-access). [Online]. Available: <https://www.metmuseum.org/about-the-met/policies-and-documents/open-access>
- [13] Metropolitan Museum of Art, "Open Access CSV Dataset," GitHub. [Online]. Available: <https://github.com/metmuseum/openaccess>
- [14] OpenRefine Community, "OpenRefine User Manual: Reconciling," openrefine.org. [Online]. Available: <https://openrefine.org/docs/manual/reconciling>
- [15] OpenRefine Community, "GREL Functions Reference," openrefine.org. [Online]. Available: <https://openrefine.org/docs/manual/grelfunctions>

- [16] OpenRefine Community, "Wikibase: Uploading data to Wikidata," openrefine.org. [Online]. Available: <https://openrefine.org/docs/manual/wikibase/>
- [17] J. D. Beazley, Attic Red-Figure Vase-Painters, 2nd ed. Oxford: Clarendon Press, 1963.
- [18] D. Williams, "Brygos and the Brygos Painter," in Greek Vases in the J. Paul Getty Museum, Malibu, CA: J. Paul Getty Museum, 1992.
- [19] Wikidata, "Makron (Q427706)." [Online]. Available: <https://www.wikidata.org/wiki/Q427706>
- [20] Wikidata, "Douris (Q965272)." [Online]. Available: <https://www.wikidata.org/wiki/Q965272>
- [21] Kerameikos.org, "Linked Open Data for Ancient Greek Pottery." [Online]. Available: <https://kerameikos.org/>
- [22] Wikidata, "Wikidata:Notability." [Online]. Available: <https://www.wikidata.org/wiki/Wikidata:Notability>

Βιβλιογραφία

Η ενότητα περιλαμβάνει το σύνολο των πηγών που μελετήθηκαν κατά την εκπόνηση, συμπεριλαμβανομένων και εκείνων που δεν παρατίθενται ρητά ως αναφορές. Οι πηγές οργανώνονται θεματικά.

Semantic Web και Linked Open Data

Berners-Lee, T., Hendler, J., and Lassila, O., "The Semantic Web," Scientific American, vol. 284, no. 5, pp. 34–43, 2001.

Berners-Lee, T., "Linked Data," W3C Design Issues, 2006 (updated 2009).

Bizer, C., Heath, T., and Berners-Lee, T., "Linked Data - The Story So Far," International Journal on Semantic Web and Information Systems, vol. 5, no. 3, pp. 1–22, 2009.

Heath, T. and Bizer, C., Linked Data: Evolving the Web into a Global Data Space. San Rafael, CA: Morgan & Claypool, 2011.

W3C, "RDF 1.1 Concepts and Abstract Syntax," W3C Recommendation, 2014.

W3C, "SPARQL 1.1 Query Language," W3C Recommendation, 2013.

W3C, "OWL 2 Web Ontology Language Document Overview," W3C Recommendation, 2012.

Wikidata

Vrandečić, D. and Krötzsch, M., "Wikidata: A Free Collaborative Knowledgebase," Communications of the ACM, vol. 57, no. 10, pp. 78–85, 2014.

Wikimedia Foundation, "Wikidata: Introduction," [wikidata.org](https://www.wikidata.org/).

Wikimedia Foundation, "Wikidata:WikiProject Visual arts," [wikidata.org](https://www.wikidata.org/).

Wikimedia Foundation, "Help:Sources," [wikidata.org](https://www.wikidata.org/).

Wikimedia Foundation, "Wikidata:Notability," [wikidata.org](https://www.wikidata.org/).

Πρότυπα μεταδεδομένων

Dublin Core Metadata Initiative, "Dublin Core Metadata Element Set, Version 1.1," DCMI Recommendation.

CIDOC CRM Special Interest Group, "Definition of the CIDOC Conceptual Reference Model," v.7.1.3, ISO 21127:2023.

Doerr, M., "The CIDOC Conceptual Reference Module: An Ontological Approach to Semantic Interoperability of Metadata," AI Magazine, vol. 24, no. 3, pp. 75–92, 2003.

Europeana Foundation, "Europeana Data Model - Definition of the Europeana Data Model," v5.2.8, 2017.

GLAM και ψηφιακές συλλογές

Candela, G., Pearson, N., and Chambers, S., "A Systematic Review of Wikidata in GLAM Institutions: a Labs Approach," TPD L 2024.

Candela, G., Dobрева, M., and Escobar, P., "Open Data Practices of Art Museums in Wikidata: A Compliance Assessment," J. Open Humanities Data, 2025.

Fagervig, A., Andersson, A., and Berggren, M., "Wikidata for authority control," J. Open Humanities Data, vol. 9, 2023.

Nakasone, C., Bullard, J., and Rasmussen, A., "Towards a Linked Open Index," J. Open Humanities Data, 2026.

Rossenova, M., Tolomelli, L., and Ciotti, F., "Empowering Italian GLAMs," J. Open Humanities Data, 2026.

Metropolitan Museum of Art, "Open Access Initiative," [metmuseum.org](https://www.metmuseum.org).

Metropolitan Museum of Art, "Open Access CSV Dataset," GitHub.

Αρχαία ελληνική κεραμική

Beazley, J. D., Attic Red-Figure Vase-Painters, 2nd ed. Oxford: Clarendon Press, 1963.

Beazley, J. D., Attic Black-Figure Vase-Painters. Oxford: Clarendon Press, 1956.

Boardman, J., Athenian Red Figure Vases: The Archaic Period. London: Thames & Hudson, 1975.

Kunisch, N., Makron. Mainz: Philipp von Zabern, 1997.

Williams, D., "Brygos and the Brygos Painter," in Greek Vases in the J. Paul Getty Museum, 1992.

Kerameikos.org, "Linked Open Data for Ancient Greek Pottery."

Εργαλεία και τεχνική τεκμηρίωση

OpenRefine Community, "OpenRefine User Manual," openrefine.org.

OpenRefine Community, "Reconciling," OpenRefine User Manual.

OpenRefine Community, "GREL Functions Reference," OpenRefine User Manual.

OpenRefine Community, "Wikibase: Uploading data to Wikidata," OpenRefine User Manual.

Manske, M., "QuickStatements," Wikimedia Toolforge.

Wikimedia Foundation, "Wikidata Query Service," query.wikidata.org.

Παράρτημα Α: Το Αρχείο Μεθοδολογίας

Το αρχείο JSON που ακολουθεί εξήχθη από την καρτέλα Undo/Redo του OpenRefine και περιέχει την πλήρη καταγραφή των ενεργειών που εκτελέστηκαν. Η φόρτωσή του σε νέο project μέσω της λειτουργίας Apply αναπαράγει πιστά ολόκληρη τη διαδικασία.

Για λόγους οικονομίας χώρου, παρατίθεται η δομή κάθε ενέργειας χωρίς τα εκτενή τμήματα engineConfig, τα οποία περιέχουν τη διαμόρφωση των φίλτρων. Το πλήρες αρχείο συνοδεύει την εργασία ως ψηφιακό παραδοτέο.

Α.1 Επισκόπηση των κύριων ενεργειών

Το ιστορικό περιλαμβάνει τόσο τις κύριες ενέργειες που συγκροτούν τη μεθοδολογία όσο και βοηθητικές, όπως δημιουργίες και διαγραφές ενδιάμεσων στηλών. Ο πίνακας που ακολουθεί παρουσιάζει τις κύριες, με τη σειρά εκτέλεσης.

Πίνακας Α.1: Οι κύριες ενέργειες με τις παραμέτρους τους

#	Λειτουργία	Στήλη	Παράμετρος
1	text-transform	Object ID	value.toNumber()
2	recon	Artist Display Name	τύπος Q5
3	recon-judge-similar-cells	Artist Display Name	Makron → Q427706
4	recon-judge-similar-cells	Artist Display Name	Douris → Q965272
5	recon-judge-similar-cells	Artist Display Name	Onesimos → Q475054
6	column-addition	Artist_Wikidata_ID	cell.recon.match.id
7	extend-reconciled-data	Artist Display Name	ιδιότητα P27
8	recon	Medium	τύπος Q214609 (απέτυχε)
9	recon	Medium	χωρίς περιορισμό τύπου
10	recon-judge-similar-cells	Medium	terracotta → Q60424
11	recon-judge-similar-cells	Medium	plaster → Q274988
12	column-addition	Medium_Wikidata_ID	cell.recon.match.id
13	column-addition	SubjectQ	υπό συνθήκη υποκείμενο
14	recon-use-values-as-identifiers	SubjectQ	15 υπάρχοντα items
15	recon-mark-new-topics	SubjectQ	486 νέα items
16	column-addition	P31_Q	υπό συνθήκη instance of
17	recon-use-values-as-identifiers	P31_Q	493 → Q1463658
18	column-addition	AB_test	αντίγραφο για το πείραμα
19	recon	AB_test	χωρίς περιορισμό τύπου
20	column-removal	AB_test	αφαίρεση μετά τη μέτρηση

Οι ενέργειες 13 έως 17 αφορούν τη διαμόρφωση του υπό συνθήκη υποκειμένου και της υπό συνθήκη ιδιότητας P31, όπως περιγράφεται στις ενότητες 3.5.2 και 3.5.4. Οι ενέργειες 18 έως 20 αφορούν το ελεγχόμενο πείραμα της ενότητας 4.6· η βοηθητική στήλη αφαιρέθηκε μετά τη λήψη της μέτρησης, ώστε να μην επιβαρύνει τα παραδοτέα.

A.2 Ενδεικτικό απόσπασμα

Το ακόλουθο απόσπασμα αντιστοιχεί στην πρώτη ενέργεια, τη μετατροπή της στήλης Object ID σε αριθμητικό τύπο.

```
{
  "op": "core/text-transform",
  "engineConfig": {
    "facets": [ ... ],
    "mode": "row-based"
  },
  "columnName": "Object ID",
  "expression": "value.toNumber()",
  "onError": "keep-original",
  "repeat": false,
  "repeatCount": 10,
  "description": "Text transform on 2145 cells in column Object ID"
}
```

Το πεδίο engineConfig, που παρατίθεται συνοπτικά, περιέχει τη διαμόρφωση των τριών ενεργών φίλτρων: του Text Facet στη στήλη Department με επιλεγμένη την τιμή Greek and Roman Art, του Text Facet στη στήλη Artist Display Name με ενεργή τη λειτουργία invert, και του Custom Text Facet στη στήλη Object ID με τη λογική συνθήκη του αριθμητικού εύρους.

Παράρτημα Β: Δείγμα του Αρχείου QuickStatements

Το αρχείο QuickStatements αποτελείται από 3.871 γραμμές που αντιστοιχούν σε 501 αντικείμενα. Από αυτά, τα 486 δεν διαθέτουν Wikidata item και δημιουργούνται με την εντολή CREATE, ενώ τα 15 που διαθέτουν ήδη item αναφέρονται απευθείας με το Q-number τους.

Κάθε αντικείμενο συνοδεύεται από ετικέτα και έως έξι δηλώσεις: creator, material used, instance of, The Met object ID, collection και inventory number. Κάθε δήλωση φέρει αναφορά πηγής S248 προς το Q160236.

Ο κωδικός Len δηλώνει ετικέτα στα αγγλικά. Η λέξη LAST αναφέρεται στο τελευταίο δημιουργηθέν item. Η σύμβαση S248 αντί για P248 διακρίνει τις ιδιότητες της αναφοράς από εκείνες της κύριας δήλωσης· το γράμμα S προέρχεται από τη λέξη source.

B.1 Δημιουργία νέων items

Τα τρία πρώτα μπλοκ του αρχείου. Παρατηρείται ότι το πρώτο και το δεύτερο δεν περιέχουν δήλωση P170, καθώς οι αντίστοιχες εγγραφές δεν διέθεταν αντιστοιχισμένο δημιουργό.

```
CREATE
LAST    Len      "Terracotta fragment of a kylix (drinking cup)"
LAST    P186     Q60424          S248   Q160236
LAST    P31      Q1463658       S248   Q160236
LAST    P3634    "710000"         S248   Q160236
LAST    P195     Q160236        S248   Q160236
LAST    P217     "2011.604.1.6638"  S248   Q160236
CREATE
LAST    Len      "Terracotta rim fragment of a kylix (drinking cup)"
LAST    P186     Q60424          S248   Q160236
LAST    P31      Q1463658       S248   Q160236
LAST    P3634    "710001"         S248   Q160236
LAST    P195     Q160236        S248   Q160236
LAST    P217     "2011.604.1.6639"  S248   Q160236
CREATE
LAST    Len      "Terracotta fragment of a kylix (drinking cup)"
LAST    P170     Q322192         S248   Q160236
LAST    P186     Q60424          S248   Q160236
LAST    P31      Q1463658       S248   Q160236
LAST    P3634    "710002"         S248   Q160236
LAST    P195     Q160236        S248   Q160236
LAST    P217     "2011.604.1.6612"  S248   Q160236
```

B.2 Εμπλουτισμός υπάρχοντος item

Το ακόλουθο μπλοκ αφορά αντικείμενο που διαθέτει ήδη Wikidata item. Δεν εκδίδεται εντολή CREATE· οι δηλώσεις προστίθενται απευθείας στο υπάρχον Q116444622. Η προσέγγιση αυτή αποτρέπει τη δημιουργία διπλοεγγραφών.

```
Q116444622  Len      "Terracotta fragment of a kylix (drinking cup)"
Q116444622  P170     Q7686622       S248   Q160236
Q116444622  P186     Q60424          S248   Q160236
Q116444622  P31      Q1463658       S248   Q160236
```

Αναφορές και Παραρτήματα

Q116444622	P3634	"713828"	S248	Q160236
Q116444622	P195	Q160236	S248	Q160236
Q116444622	P217	"2011.604.1.6788"	S248	Q160236

Αξίζει να σημειωθεί ότι η ετικέτα δηλώνεται με την επιλογή `do not override`, ώστε να μην αντικατασταθεί τυχόν υπάρχουσα ετικέτα του item.

Παράρτημα Γ: Πίνακας Αντιστοιχισμένων Δημιουργών

Ο πίνακας που ακολουθεί παραθέτει το σύνολο των 63 μοναδικών Wikidata items στα οποία αντιστοιχίστηκαν οι δημιουργοί του δείγματος, με το πλήθος των έργων που αποδίδονται σε καθένα. Το άθροισμα των έργων ανέρχεται σε 389, όσες και οι αντιστοιχισμένες εγγραφές.

Πίνακας Γ.1: Οι 63 αντιστοιχισμένοι δημιουργοί

Όνομα στο dataset	Wikidata ID	Έργα
Makron	Q427706	172
Euaion Painter	Q37630169	22
Douris	Q965272	21
Onesimos	Q475054	21
Painter of the Paris Gigantomachy	Q1887164	10
Brygos Painter	Q322192	10
Maler von London E 777	Q1887200	8
Briseis Painter	Q18511839	7
Boot Painter	Q136530475	7
Euergides Painter	Q89520234	7
Triptolemos Painter	Q713053	6
Cage Painter	Q44199115	5
Colmar Painter	Q18508104	4
Sabouroff Painter	Q18697269	4
Eretria Painter	Q386045	4
Foundry Painter	Q1366158	3
Stieglitz Painter	Q26206528	3
Painter of London D 12	Q26203783	3
Wedding Painter	Q328060	3
Penthesilea Painter	Q71976	3
Bowdoin-Eye Painter	Q89061387	3
Altamura Painter	Q4736137	3
Lewis Painter	Q134124245	3
Antiphon Painter	Q116616	3
Rider Painter	Q1449974	3
Michigan Painter	Q20745835	2

Lysippides Painter	Q1878978	2
Florence Painter	Q14906666	2
Psiax	Q1387942	2
Oltos	Q1360826	2
Telephos Painter	Q133877584	2
Thalia Painter	Q90070025	2
Berlin Painter	Q821647	2
Heidelberg Painter	Q628012	2
Tarquinius Painter	Q7686622	2
Gorgon Painter	Q256026	2
Villa Giulia Painter	Q26203793	2
Nikosthenes Painter	Q18508558	2
Aberdeen painter	Q18670007	1
Kleophrades Painter	Q1347940	1
Orchard Painter	Q136530482	1
Skythes	Q1297344	1
Painter of the Nicosia Olpe	Q1347300	1
Rote-Linie-Maler	Q1459078	1
Carlsruhe Painter	Q18507999	1
Phineus Painter	Q18710601	1
Hunt Painter	Q18711122	1
Theseus Painter	Q2420238	1
Meleager Painter	Q1919255	1
Painter of Vatican 365	Q1887210	1
Boreads painter	Q18785939	1
Pisticci Painter	Q1851255	1
Phiale Painter	Q353443	1
Tyszkiewicz Painter	Q26250784	1
Epeleios Painter	Q26214499	1
Penelope Painter	Q37631991	1
Painter of Bologna 417	Q37632501	1
Tarporley Painter	Q353664	1

Syleus Painter	Q37630048	1
Pistoxenos Painter	Q711763	1
Leningrad Painter	Q46001881	1
Falmouth Painter	Q86997044	1
Borden Wood Painter	Q893403	1

Παρατηρείται η έντονη ασυμμετρία που αναλύθηκε στα Κεφάλαια 2 και 4: ο πρώτος δημιουργός συγκεντρώνει 172 έργα, ενώ η πλειονότητα των υπολοίπων εκπροσωπείται από ένα έως τρία.