

Ροές Εργασίας στον Ψηφιακό Χώρο: Μελέτη Περίπτωσης Δεκαετίας '40

Βίκη Ντρίτσου
Μονάδα Ψηφιακής Επιμέλειας,
ΙΠΣΥ / Ερευνητικό Κέντρο Αθηνά
&
Τμήμα Πληροφορικής, Οικονομικό Πανεπιστήμιο
Αθηνών



Ροή εργασίας

- "A workflow consists of an orchestrated and repeatable pattern of activity, enabled by the systematic organization of resources into processes that transform materials, provide services, or process information." [Wikipedia]
 - Η ροή εργασιών μπορεί να θεωρηθεί ως η αφαιρετική αναπαράσταση μιας πραγματικής εργασίας.
- Αντικείμενό μας εδώ: Ροές εργασιών ψηφιακής επιμέλειας

Ψηφιακή Επιμέλεια

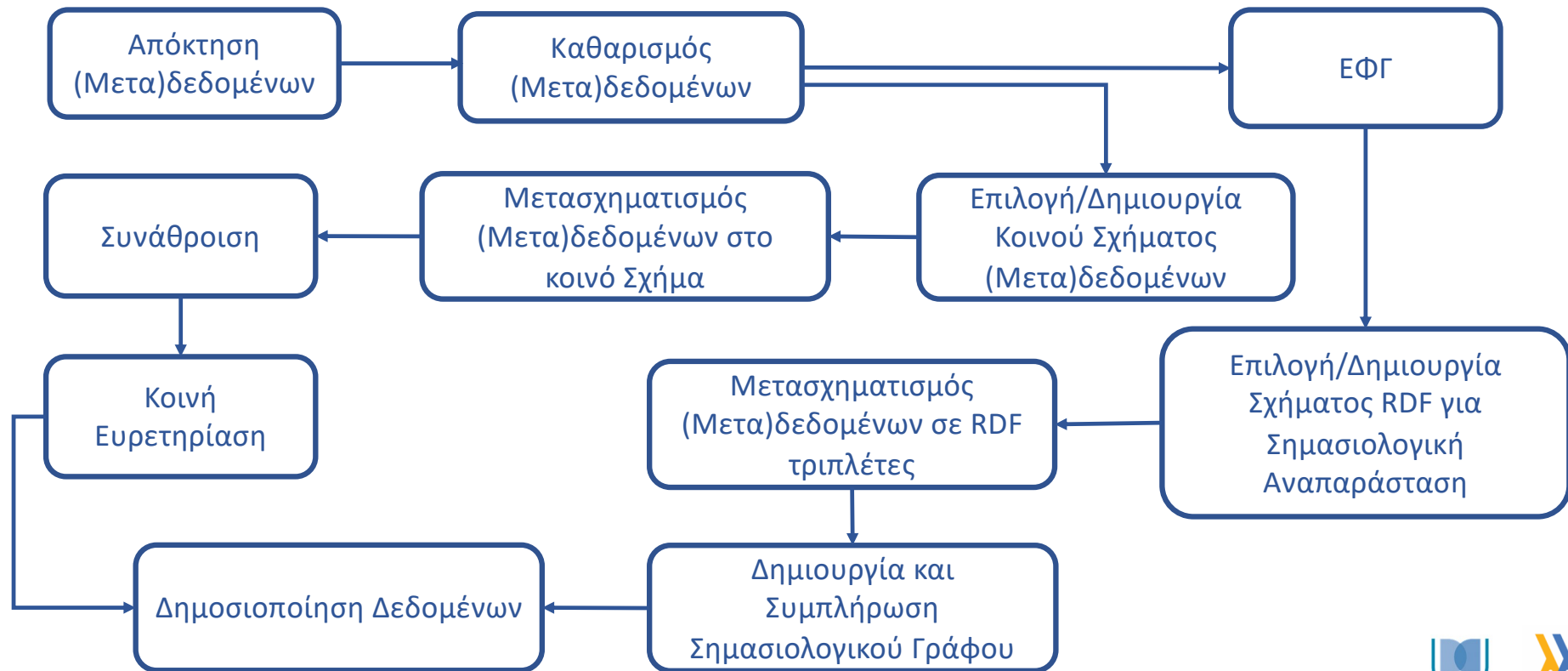
- Σκοπός της ψηφιακής επιμέλειας η υποστήριξη, η διατήρηση και η προσθήκη αξίας σε ψηφιακούς πόρους^[1].
- Διεργασίες κύκλου ζωής ψηφιακής επιμέλειας^[2]
"Appraisal, Ingest, Classification, indexing and cataloguing, Knowledge enhancement, Presentation, Publication and dissemination, User experience, Repository management, Preservation"

^[1] <http://www.dcc.ac.uk/digital-curation/what-digital-curation>

^[2] Constantopoulos, P., Dallas, C., Androutsopoulos, I., Angelis, S., Deligiannakis, A., Gavrilis, D., Kotidis, Y., Papatheodorou, C. (2009). DCC&U: An Extended Digital Curation Lifecycle Model. International Journal of Digital Curation. 4. 10.2218/ijdc.v4i1.76.



Χάρτης εργασιών Ψηφιακής Επιμέλειας

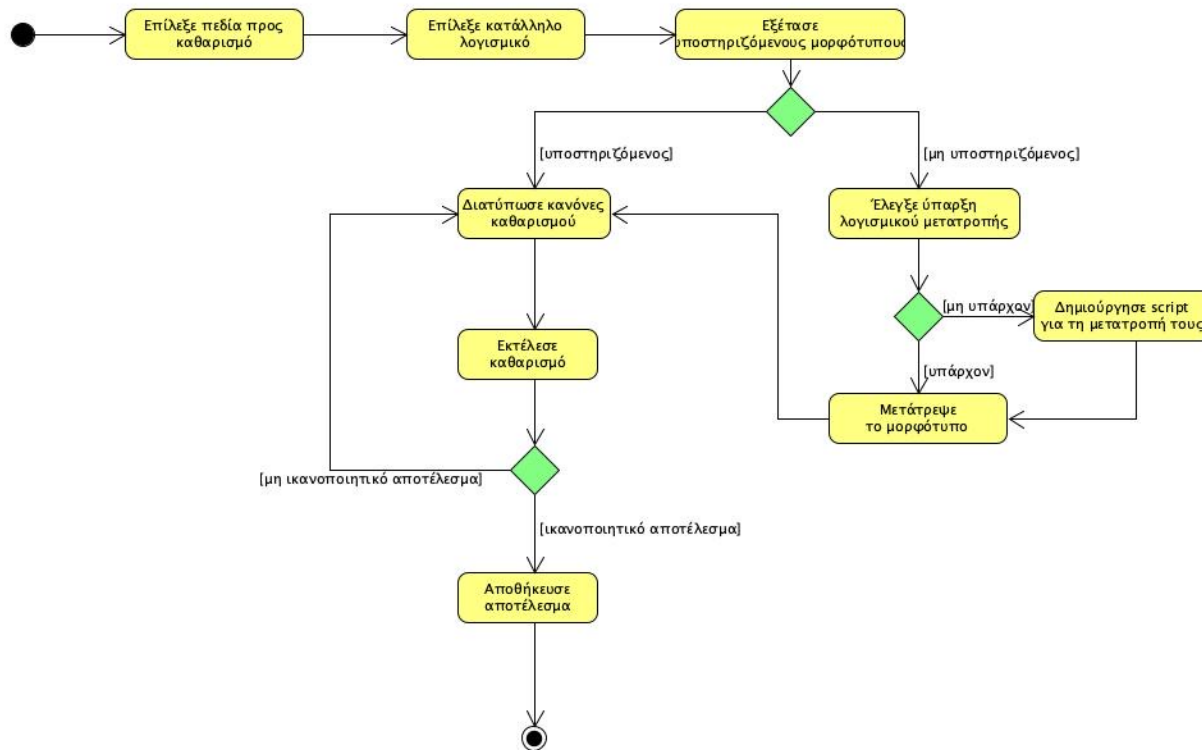


Απόκτηση αρχείων (Acquisition)

- Απόκτηση αρχείων από φορείς
 - Μορφότυπος
 - Ακαδημία Αθηνών: βάση δεδομένων Microsoft Access, 4866 εγγραφές
 - Αρχεία Σύγχρονης Κοινωνικής Ιστορίας: xml, 9767 εγγραφές
 - Ανωτάτη Σχολή Καλών Τεχνών: json, 287 εγγραφές
 - Διεύθυνση Ιστορίας Στρατού: xml, 76979 εγγραφές
 - Εβραϊκό Μουσείο Ελλάδος: html, 13 εγγραφές
 - Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών: json, 373 εγγραφές
- Μελέτη μορφής και μετατροπή μορφότυπου
 - ΑΑ: Εξαγωγή δεδομένων σε csv μέσω εκτέλεσης κατάλληλων SQL επερωτήσεων
 - ΑΣΚΙ: Ενοποίηση πεδίων ανά εγγραφή – μη αυτόματη διεργασία, δημιουργία script
 - ΑΣΚΤ, ΔΙΣ, Εβραϊκό Μουσείο, ΕΚΠΑ : Αυτόματη μετατροπή δεδομένων σε csv μέσω λογισμικού



Καθαρισμός δεδομένων



- Είσοδος: αρχεία συλλογών (xml, xls, tsv, json)
- Έξοδος: «καθαρισμένα» αρχεία συλλογών

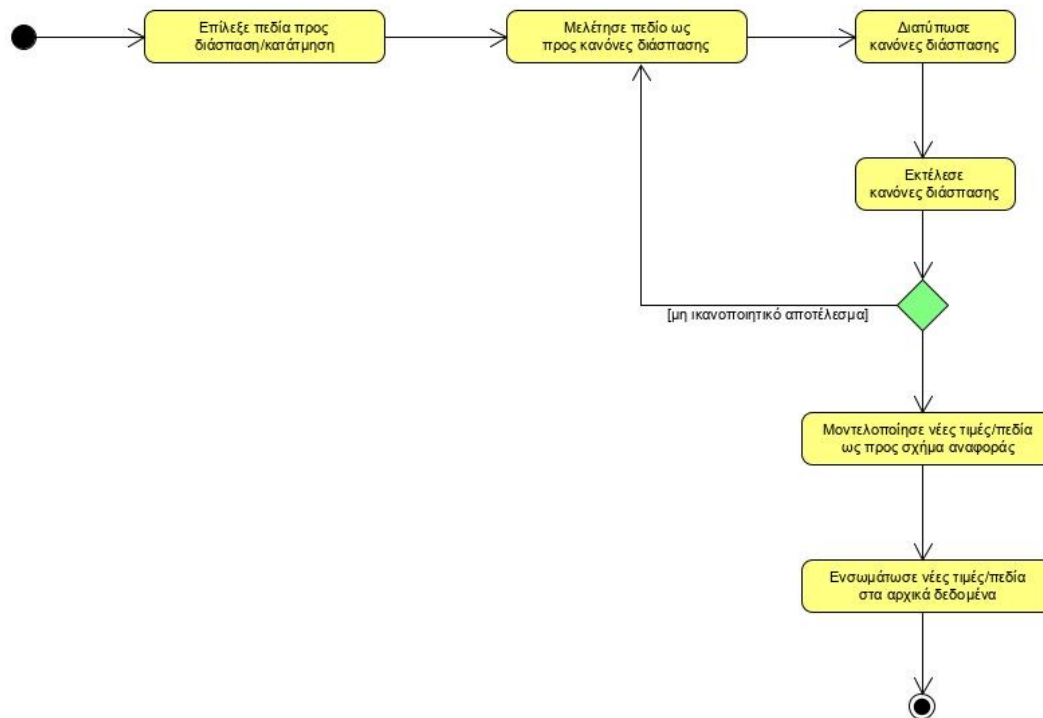
Επαναλαμβανόμενα πεδία - τεμαχισμός

ID	Subject
AA_1207	Ε.Λ.Α.Σ., Α. Βελουχιώτης (Αρης) Καπετάνιος Γενικού Στρατηγείου Ε.Λ.Α.Σ., Myers 5/42 Σύνταγμα Ευζώνων Ε.Κ.Κ.Α.



ID	Subject
AA_1207	Ε.Λ.Α.Σ.
AA_1207	Α. Βελουχιώτης (Αρης) Καπετάνιος Γενικού Στρατηγείου Ε.Λ.Α.Σ.
AA_1207	Myers
AA_1207	5/42 Σύνταγμα Ευζώνων Ε.Κ.Κ.Α.

Τεμαχισμός πεδίων

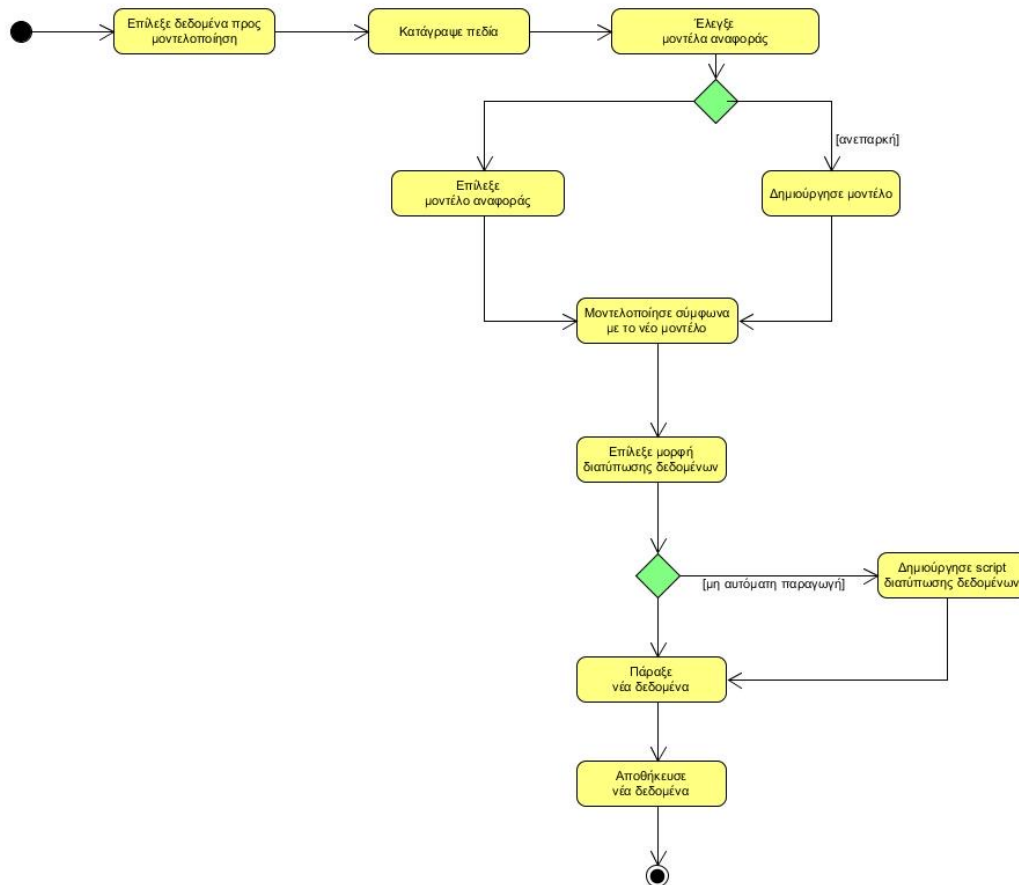


- Είσοδος: αρχείο συλλογής
- Έξοδος: νέο αρχείο συλλογής με πρόσθετα πεδία

Μελέτη σχήματος

- Κάθε αρχείο ακολουθεί το δικό του σχήμα μεταδεδομένων
 - Διαφορετικά σχήματα σε συλλογές του ίδιου φορέα
 - Συντακτικές και σημασιολογικές αμφισημίες
- Στόχος: κοινή ευρετηρίαση των αρχείων
- Ανάγκη εφαρμογής ενιαίου σχήματος
 - Ευρετήρια παραγόμενα από συγκεκριμένα πεδία
 - Στο σενάριο Δεκαετίας '40: Πρόσωπο, Τόπος, Χρόνος, Γεγονός, Θεματική ενότητα

Εννοιολογική μοντελοποίηση δεδομένων



- Μη αυτόματη διεργασία
 - Απαιτεί νοητική εργασία
- Δεκαετία '40: EDM^[4] σχήμα
 - Εμπλουτισμένο
- Είσοδος: πολλαπλά xml σχήματα
- Έξοδος: ενιαίο xml σχήμα + αντιστοιχίσεις

^[4] <https://pro.europeana.eu/resources/standardization-tools/edm-documentation>

Μελέτη τιμών πεδίων

- Ενιαίο μοντέλο επιτρέπει τη μελέτη/σύγκριση τιμών ανάμεσα στις εγγραφές
- Ανάδειξη προβλημάτων
 - Ορθογραφικά σφάλματα
 - Συντακτικές τακτικές γραφής (κύριων ονομάτων, τόπων, κλπ.)
 - Πρωτεύοντες και εναλλακτικοί όροι
 - Σύμπλοκα πεδία

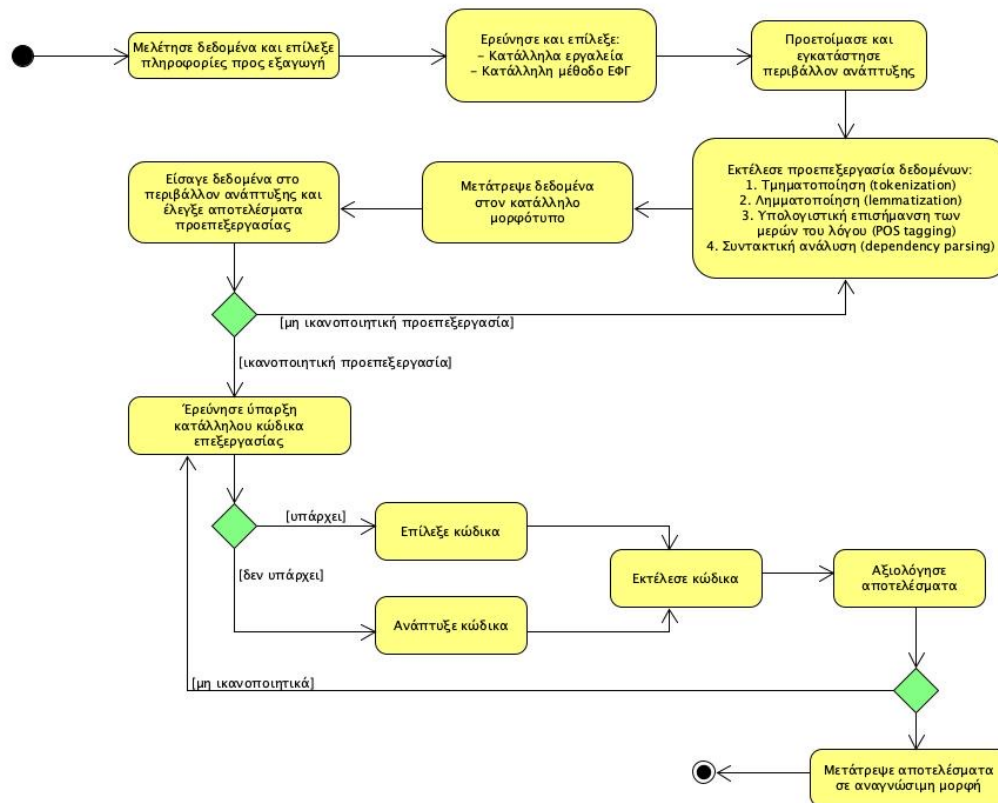
Ενοποίηση συλλογών

- Τελικός στόχος: δημιουργία ενιαίας συλλογής και δημοσίευσή της
- Είσοδος: επιμέρους συλλογές με επεξεργασμένα πεδία (καθαρισμού, διάσπασης τιμών - xml, xls, tsv), σχήμα δεδομένων (xsd)
- Διεργασία: δημιουργία και εκτέλεση σεναρίου (script) μετατροπής δεδομένων
 - Έξοδος: ΜΙΑ ενιαία δομή/συλλογή (π.χ. xml, rdf)

Επεξεργασία Φυσικής Γλώσσας

- Σύμπλοκα πεδία (υπό το πρίσμα του περιεχομένου τους)
 - Περιέχουν τιμές άλλων πεδίων
 - Παραδείγματα
 - Πεδίο 'Τίτλος' της ΔΙΣ, πεδίο 'Περιγραφή' της ΑΑ
 - Εξαγωγή πληροφορίας με χρήση Επεξεργασίας Φυσικής Γλώσσας (Natural Language Processing – NLP)
 - Η εξαγόμενη πληροφορία χρήζει μοντελοποίησης εκ νέου

Επεξεργασία Φυσικής Γλώσσας (NLP)



Αποτελέσματα ΕΦΓ στη Συλλογή ΔΙΣ

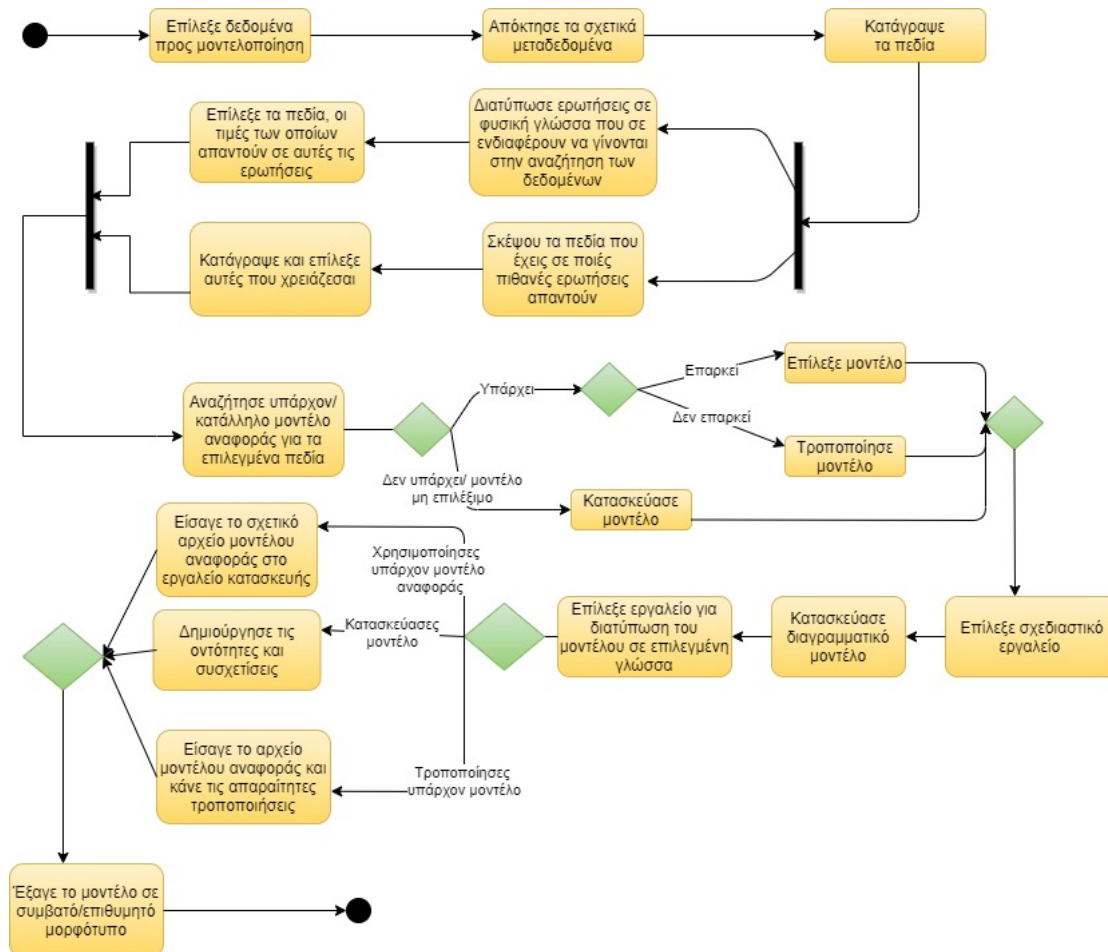
- Επιτυχής αναγνώριση των οντοτήτων:
 1. Δράστες – 15.723 επισημειώσεις (Πρόσωπα, Ομάδες και Στρατιωτικές Μονάδες)
 2. Θέματα – 3.851 επισημειώσεις
 3. Ημερομηνίες – 4.545 επισημειώσεις
 4. Τόποι – 2.125 επισημειώσεις
 5. Τύποι εγγράφων– 9.949 επισημειώσεις
- Οι επισημειώσεις αυτές θεωρήθηκαν πρόσθετη γνώση
 - Επιλέχθηκε να οργανωθεί σε ξεχωριστή δομή
- Για την οργάνωση και αναπαράσταση της πρόσθετης γνώσης επιλέχθηκε η δομή σημασιολογικού γραφου
 - Βάση γνώσης, όπου η πληροφορία είναι οργανωμένη κάτω από κατάλληλο σχήμα (οντολογία)

Ανάπτυξη Σχήματος Βάσης Γνώσης

- Εννοιολογική μοντελοποίηση
 - πρωτογενούς γνώσης
 - πρόσθετης γνώσης
- Σκοπός: δημιουργία ενιαίου σχήματος για την αναπαράσταση των πληροφοριακών πόρων
- Μελέτη μοντέλων αναφοράς
- Επιλέχθηκε το πρότυπο CIDOC CRM^[2]
 - Πρότυπο ολοκλήρωσης πληροφορίας στο πεδίο της πολιτιστικής κληρονομιάς
- 13 πρόσθετες οντότητες
 - Αναγκαίες για την αναπαράσταση ενδιαφέρουσας (από ιστορική άποψη) πληροφορίας

^[2]<http://www.cidoc-crm.org/>

Εννοιολογική Μοντελοποίηση Σχήματος Βάσης Γνώσης



Παραγωγή RDF δεδομένων

- Για τη διατύπωση της πληροφορίας χρησιμοποιήθηκε το W3C πρότυπο RDF (Resource Description Framework)^[2]
 - Δηλώσεις με δομή τριπλέτας (RDF triples): subject – predicate – object
- Διατύπωση σχήματος σε RDF
- Προγραμματιστική μετατροπή της πρωτογενούς και εξαγόμενης πληροφορίας σε RDF τριπλέτες
- Δημιουργία βάσης επαυξημένης γνώσης

^[1] <https://www.w3.org/RDF/>

Συμπεράσματα & παρατηρήσεις (I)

- Καταγραφή και αποτύπωση των εργασιών που εκτελέστηκαν σε διαγράμματα ροής
 - Διαμοιρασμός τους με την ερευνητική κοινότητα και ενδιαφερόμενους φορείς
 - Επαναχρησιμοποιήσιμες
 - Εξυπηρετούν ως οδηγοί για την εκτέλεση εργασιών επιμέλειας από ενδιαφερόμενους φορείς
- Η ακολουθία των βημάτων δεν είναι απόλυτα αυστηρή
 - Εξαρτάται από τον επιμελητή και το ίδιο το περιεχόμενο
 - Πλην κομβικών βημάτων, π.χ. απόκτηση, δημιουργία τελικής συλλογής

Συμπεράσματα & παρατηρήσεις (II)

- Ενδιάμεσα αποτελέσματα μπορεί να απαιτούν την επανεκτέλεση προηγούμενων βημάτων
 - Διάσπαση τιμών και καθαρισμός
 - Διάσπαση τιμών και μοντελοποίηση
 - NLP και μοντελοποίηση
- Τελικά βήματα πριν τη δημοσιοποίηση απαιτούν προγραμματιστικές ικανότητες
 - Εισαγωγή σε σχεσιακή βάση ή συσσωρευτή xml, ευρετηρίαση, ανάπτυξη διεπαφής ή
 - Παραγωγή σημασιολογικού γράφου rdf, εκτέλεση πολύπλοκων ερωτημάτων (π.χ. SPARQL)