απολλωνίς

Integrating archival materials for the study of the turbulent Greek 40s

Vicky Dritsou[1,2], Maria Ilvanidou[1,2], Isidora Despotidou[2], Vicky Liakopoulou[1,2], Karmen Vourvachaki[1,2], Panos Constantopoulos[1,2]

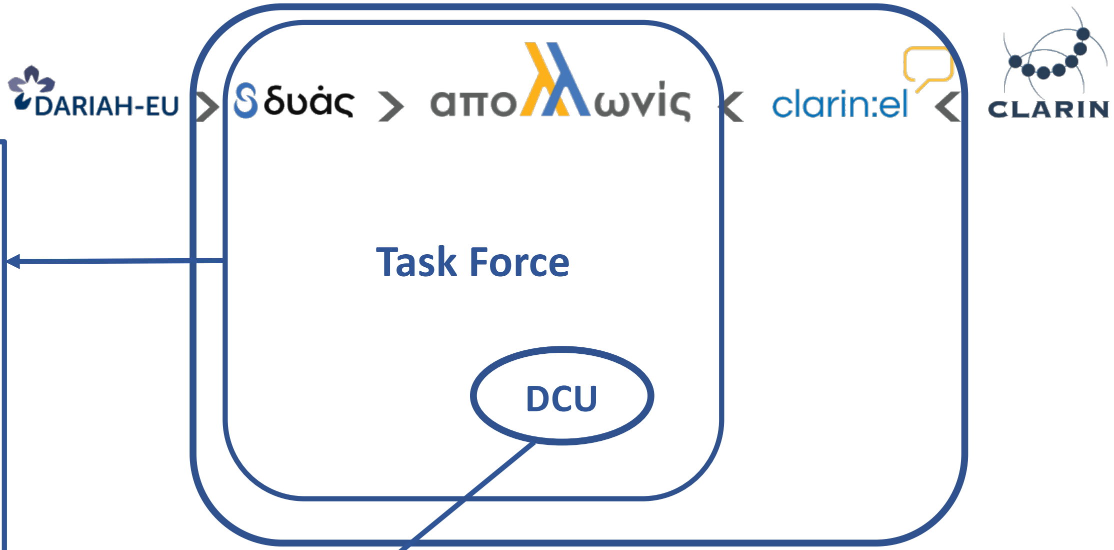
[1]Digital Curation Unit (DCU), IMSI, Athena Research Centre

[2]Department of Informatics, Athens University of Economics and Business

v.dritsou@dcu.gr

Digital Curation Unit
IMSI - Athena Research Centre

Who we are



Task Force
Academy of Athens
Athena Research Centre
Athens School of Fine Arts
Foundation of Research and Technology
National Technical University of Athens

The interdisciplinary DCU APOLLONIS team: Agiatis Benardou, Panos Constantopoulos, Isidora Despotidou, Vicky Dritsou, Georgina Dova, Maria Ilvanidou, Konstantinos Kravvaritis, Vicky Liakopoulou, Prokopis Prokopidis, Eirini Stergianeli, Karmen Vourvachaki.

Decade 1940: A micro-infrastructure

Why the Greek ‘40s:

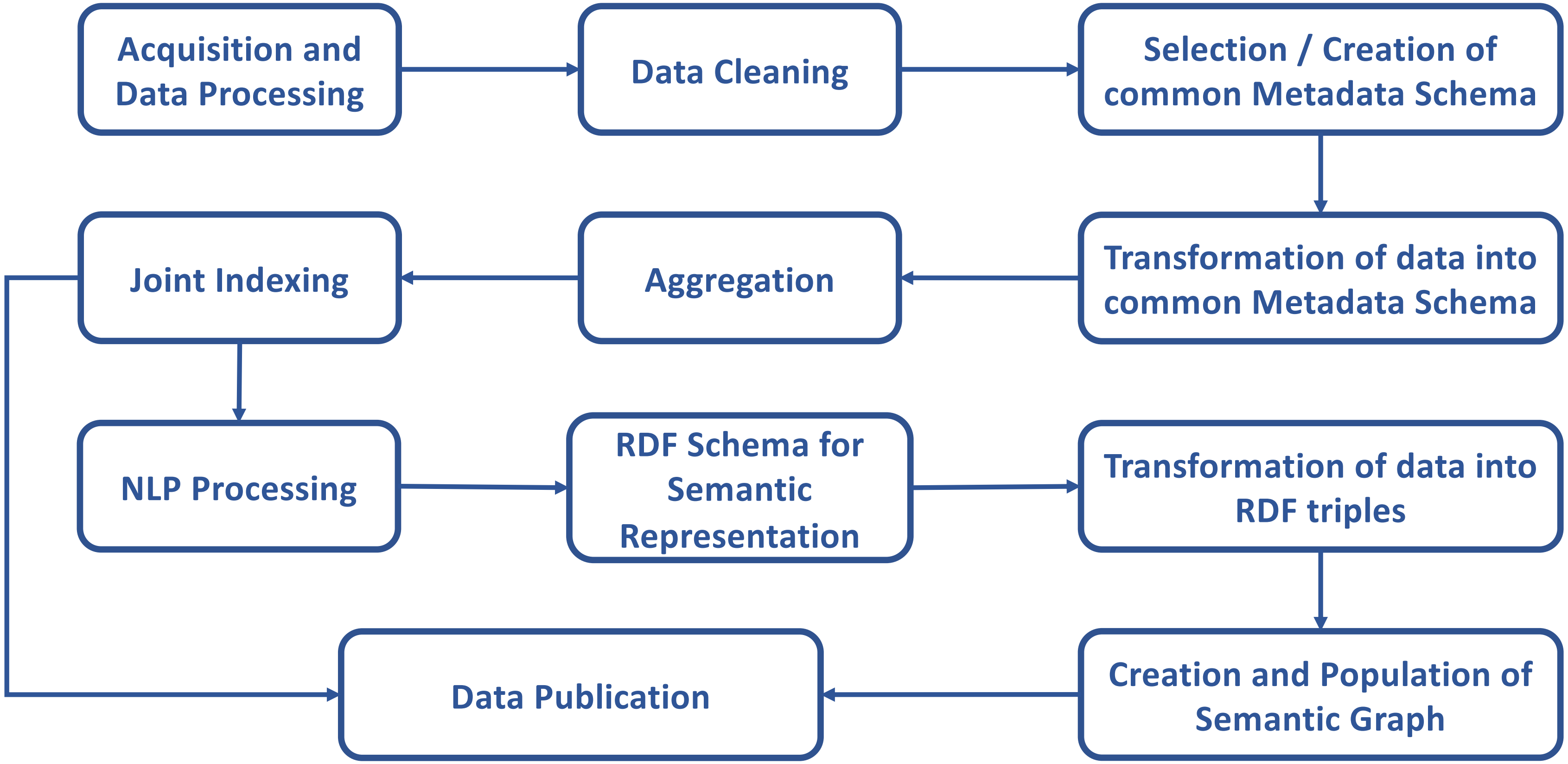
- WWII / Occupation / Resistance / Liberation / Civil War

Archives	**Contributed records (92.286 in total)**
| Academy of Athens	4.868
| Contemporary Social History Archives – ASKI	9.766
| Athens School of Fine Arts	287
| Army History Directorate	76.979
| The Jewish Museum of Greece	13
| Historical Archive of the University of Athens	373

Our aim is twofold:

- To assemble and integrate the digitized archives (their metadata) and enable their joint indexing
  - Using 5 criteria: Actors, Places, Time, Topics, Events
- To identify the required digital curation activities, record their workflows and share them with researchers

Map of the digital curation activities



Acquisition and Data Processing

Data Cleaning

Selection / Creation of common Metadata Schema

Transformation of data into common Metadata Schema

Aggregation

Joint Indexing

NLP Processing

RDF Schema for Semantic Representation

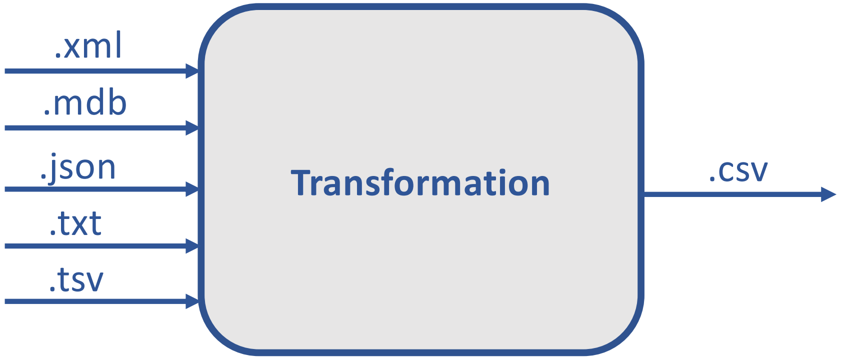
Transformation of data into RDF triples

Creation and Population of Semantic Graph

Data Publication

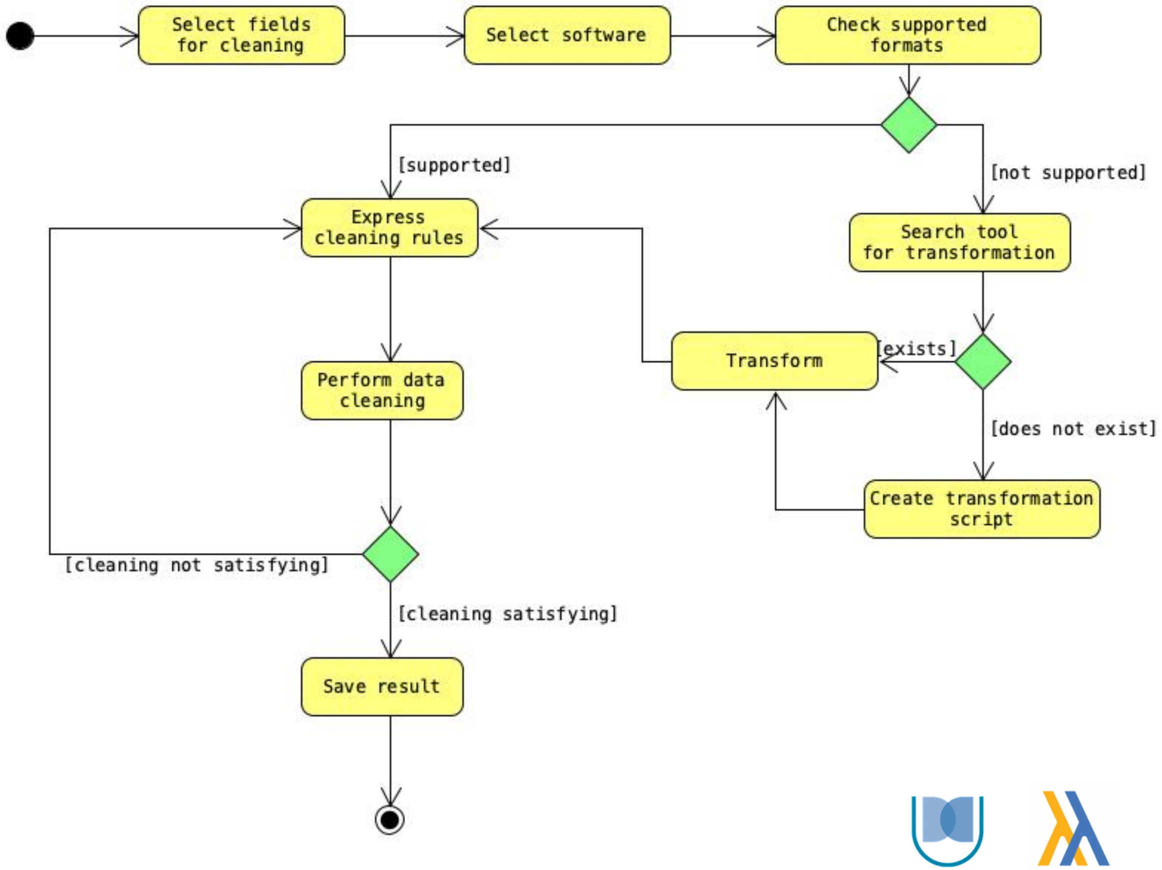
1. Acquisition and Data Processing

- Multiple file formats that required transformation



2. Data Cleaning

- Common problems faced: typos, abbreviations, upper case typing, metadata expressed in previous forms of modern Greek language
- Cleaning was performed using OpenRefine[1]



[1] https://openrefine.org/

2. Data Cleaning

ΜΟΝΑΔΕΣ ΚΑΙ ΣΧΗΜΑΤΙΣΜΟΙ ΤΟΥ ΕΛΛΗΝΙΚΟΥ ΣΤΡΑΤΟΥ

ΜΟΝΑΔΕΣ ΚΑΙ ΣΧΗΜΑΤΙΣΜΟΙ ΤΟΥ ΕΛΛΗΝΙΚΟΥ ΣΤΡΑΤΟΥ

ΜΟΝΑΔΕΣ ΚΑΙ ΣΧΗΜΑΤΙΣΜΟΙ ΕΛΛΗΝΙΚΟΥ ΣΤΡΑΤΟΥ

ΜΟΝΑΔΕΣ ΚΑΙ ΣΧΗΜΑΤΙΣΜΟΙ ΤΟΥ ΕΛΛΗΝΙΚΟΥ ΣΤΡΑΤΟΥν

ΜΟΝΑΔΕΣ ΚΑΙ ΣΧΗΜΑΤΙΣΜΟΙ ΤΟΥ ΕΛΛΗΝΙΚΟΥ ΣΤΡΑΤΟΥΩ

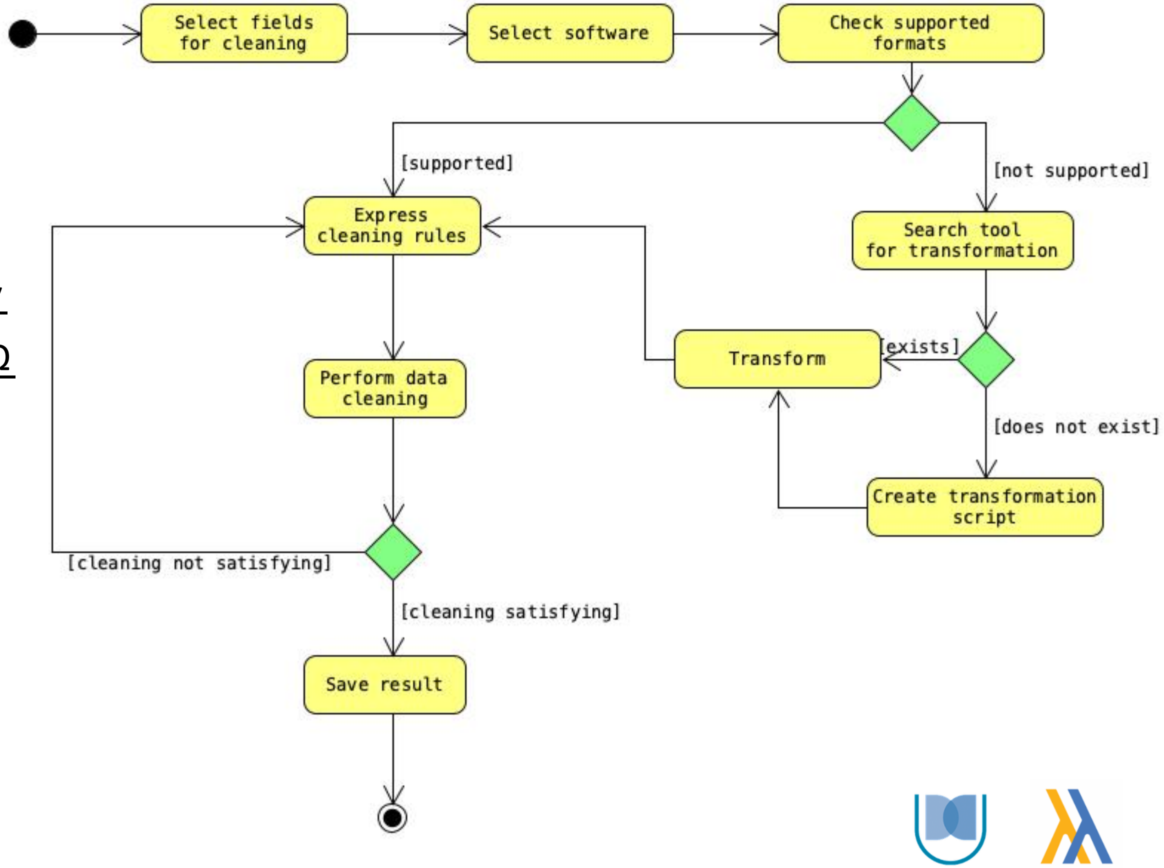
ΜΟΝΑΔΕΣ ΚΑΙ ΣΧΗΜΑΤΙΣΜΟΙ ΤΟΥ ΕΛΛΗΙΝΙΚΟΥ ΣΤΡΑΤΟΥ

ΜΟΝΑΔΕΣ ΚΑΙ ΣΧΗΜΑΤΙΣΜΟΙ ΤΟΥ ΕΛΗΝΙΚΟΥ ΣΤΡΑΤΟΥ

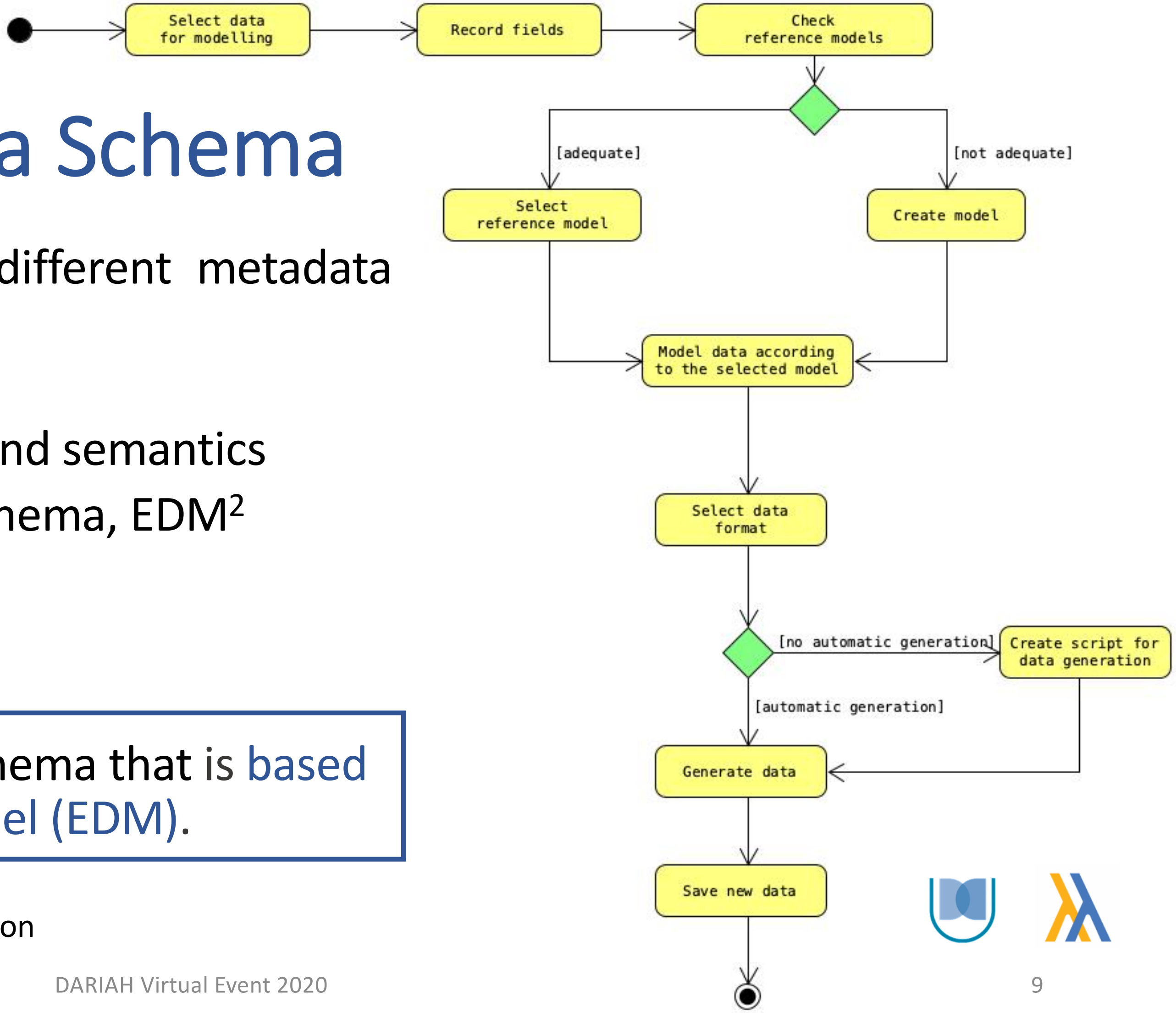
ΜΟΝΑΔΑ ΚΑΙ ΣΧΗΜΑΤΙΣΜΟΙ ΤΟΥ ΕΛΛΗΝΙΚΟΥ ΣΤΡΑΤΟΥ

ΜΟΝΑΔΕΣ ΚΑΙ ΣΧΗΜΑΤΑ ΤΟΥ ΕΛΛΗΝΙΚΟΥ ΣΤΡΑΤΟΥ

ΜΟΝΑΔΕΣ ΚΑΙ ΣΧΗΜΑΤΙΣΜΟΙ ΤΟΥ ΕΛΛΗΝΙΚΟΥ ΣΤΡΑΤΟΥ



3. Creation of Common Metadata Schema



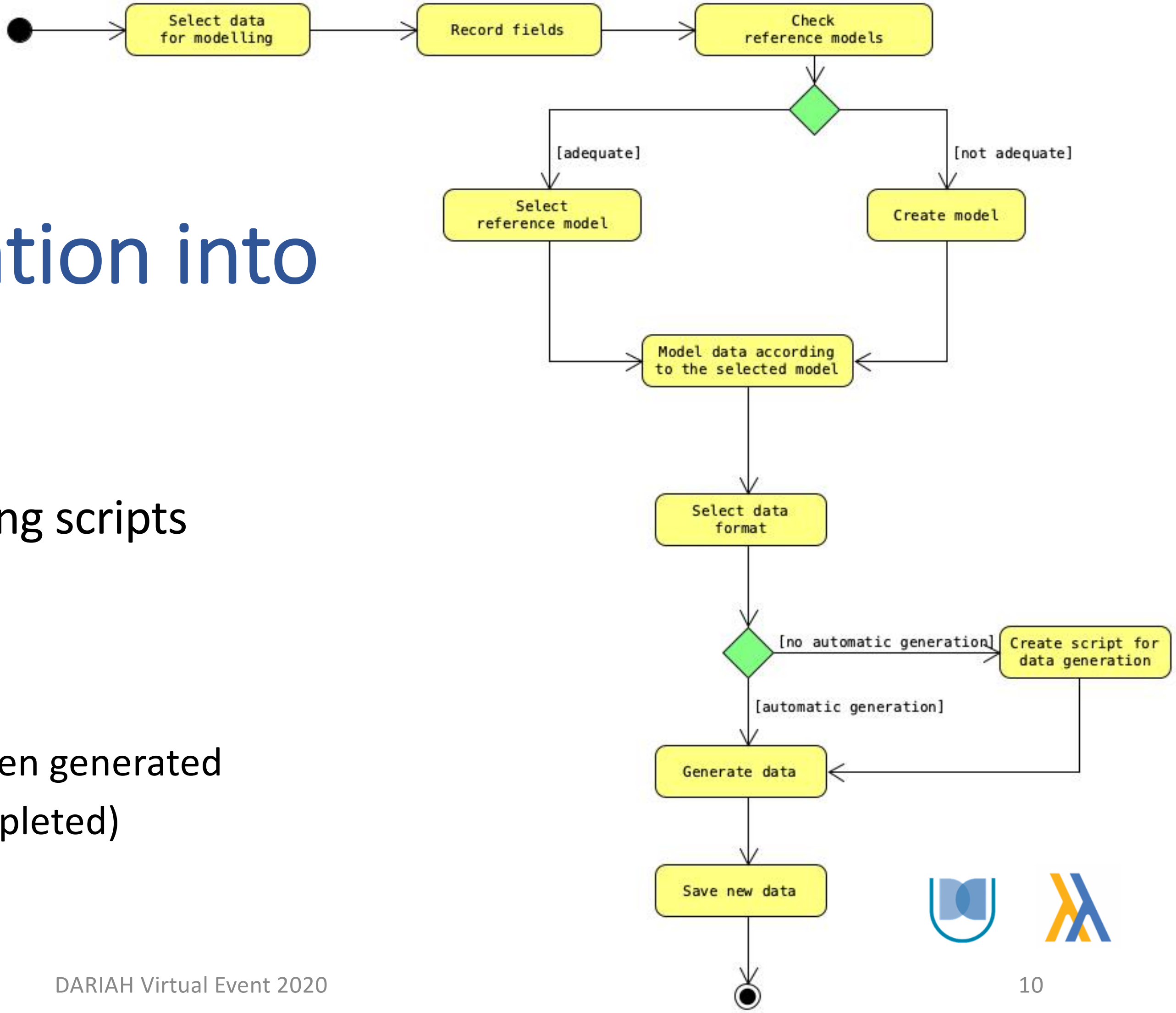
- 6 different archives – many different metadata schemata
  - Even within the same collection
- Heterogeneity w.r.t. structure and semantics
- Need to adopt of a common schema, EDM[2]
  - Homogeneity
  - Disambiguation

We have adopted a metadata schema that is based on Europeana Data Model (EDM).

[2] https://pro.europeana.eu/page/edm-documentation

4. Data transformation into EDM

- Transformations performed using scripts
  - One script for each collection
  - Time-consuming task
- Output format: XML
  - Till now, 91.986 xml files have been generated (99,67% of acquired records completed)

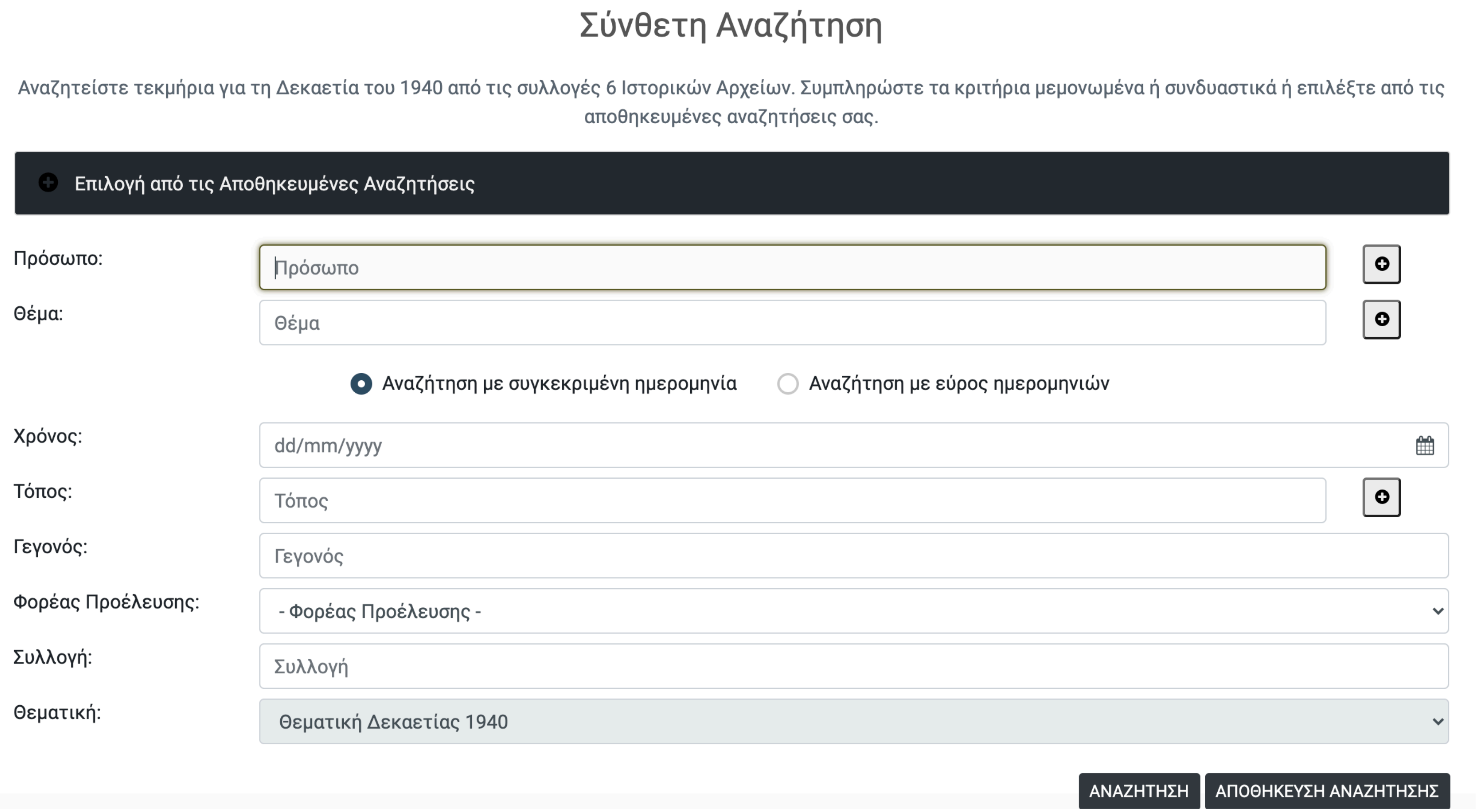


5-6. Aggregation and Joint Indexing

- The xml files were ingested into the MoRE[3] aggregator
- Indices were built using Elasticsearch[4] based on the five basic criteria (99,67% completed for 4 out of 5 indices)
    1. Actors: Persons & Groups (645.440 entries – 84.590 distinct data values)
    2. Places (188.317 entries – 82.654 distinct values)
    3. Time (all expressions – numeric, alphanumeric, textual – were also programmatically transformed into date ranges)
    4. Topics (448.160 entries – 82.004 distinct values)
    5. Events (not addressed yet)

[3] http://more.dcu.gr/

[4] https://www.elastic.co/elasticsearch



Σύνθετη Αναζήτηση

Αναζητείστε τεκμήρια για τη Δεκαετία του 1940 από τις συλλογές 6 Ιστορικών Αρχείων. Συμπληρώστε τα κριτήρια μεμονωμένα ή συνδυαστικά ή επιλέξτε από τις αποθηκευμένες αναζητήσεις σας.

⊕ Επιλογή από τις Αποθηκευμένες Αναζητήσεις

Πρόσωπο: Πρόσωπο ⊕

Θέμα: Θέμα ⊕

● Αναζήτηση με συγκεκριμένη ημερομηνία ○ Αναζήτηση με εύρος ημερομηνιών

Χρόνος: dd/mm/yyyy

Τόπος: Τόπος ⊕

Γεγονός: Γεγονός

Φορέας Προέλευσης: - Φορέας Προέλευσης -

Συλλογή: Συλλογή

Θεματική: Θεματική Δεκαετίας 1940

7. Natural Language Processing (NLP)

- Important information contained within free text metadata fields
- In collaboration with the Institute of Language and Speech Processing (ILSP) we have applied NLP methods
  - Using the Apache UIMA Ruta[5] rule language
- Successful identification of:
  - Actors – 15.723 entries (Persons, Groups and Armed Units)
  - Topics – 3.851 entries
  - Dates – 4.545 entries
  - Places – 2.125 entries
  - Document types – 9.949 entries
- The results were then treated as ‘enhanced knowledge’

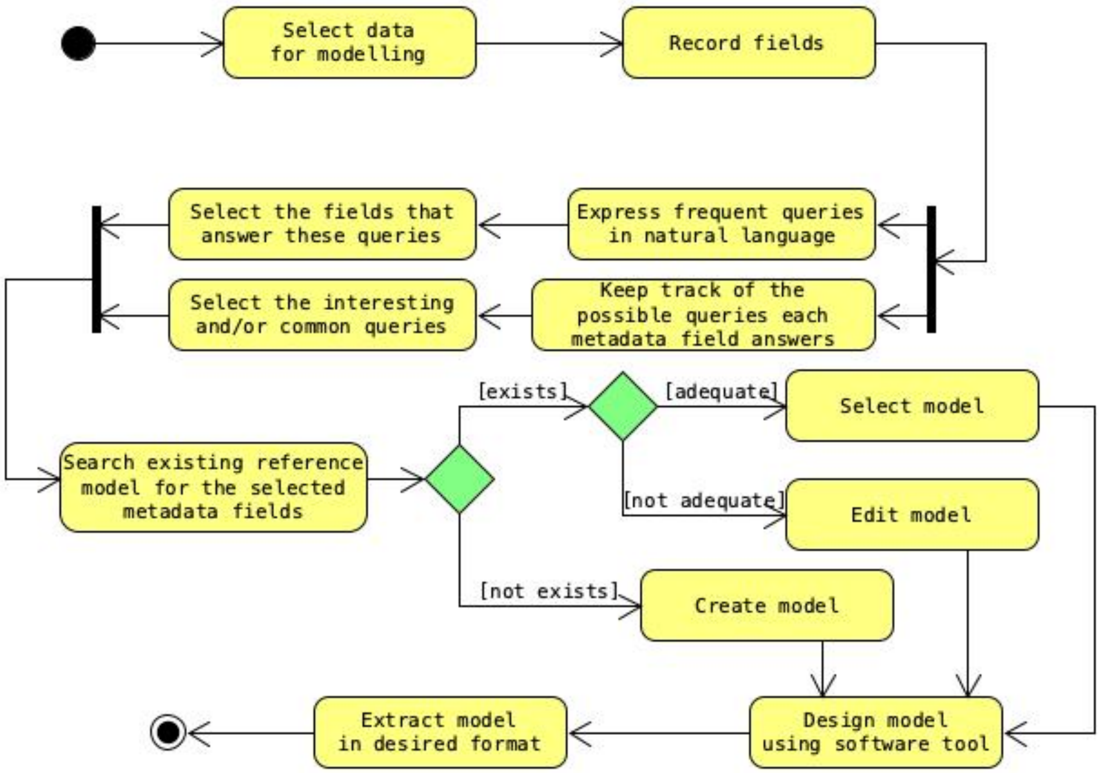
[5] https://uima.apache.org/ruta.html

Semantic Graph for Enhanced Information

- We’ve chosen to represent the primary data of all sources together with the enhanced information in a semantic graph
  - Joint indices contain only original data
  - The semantic graph also contains the additional information extracted by the NLP methods
  - Thus, exploring semantic relations between data coming from different sources

8. RDF Schema for Semantic Representation

- An RDF Schema was created based on the CIDOC CRM[6] standard
- All data (original+NLP) was programmatically expressed into RDF triples
- The outcome was ingested into a Neo4j[7] graph database



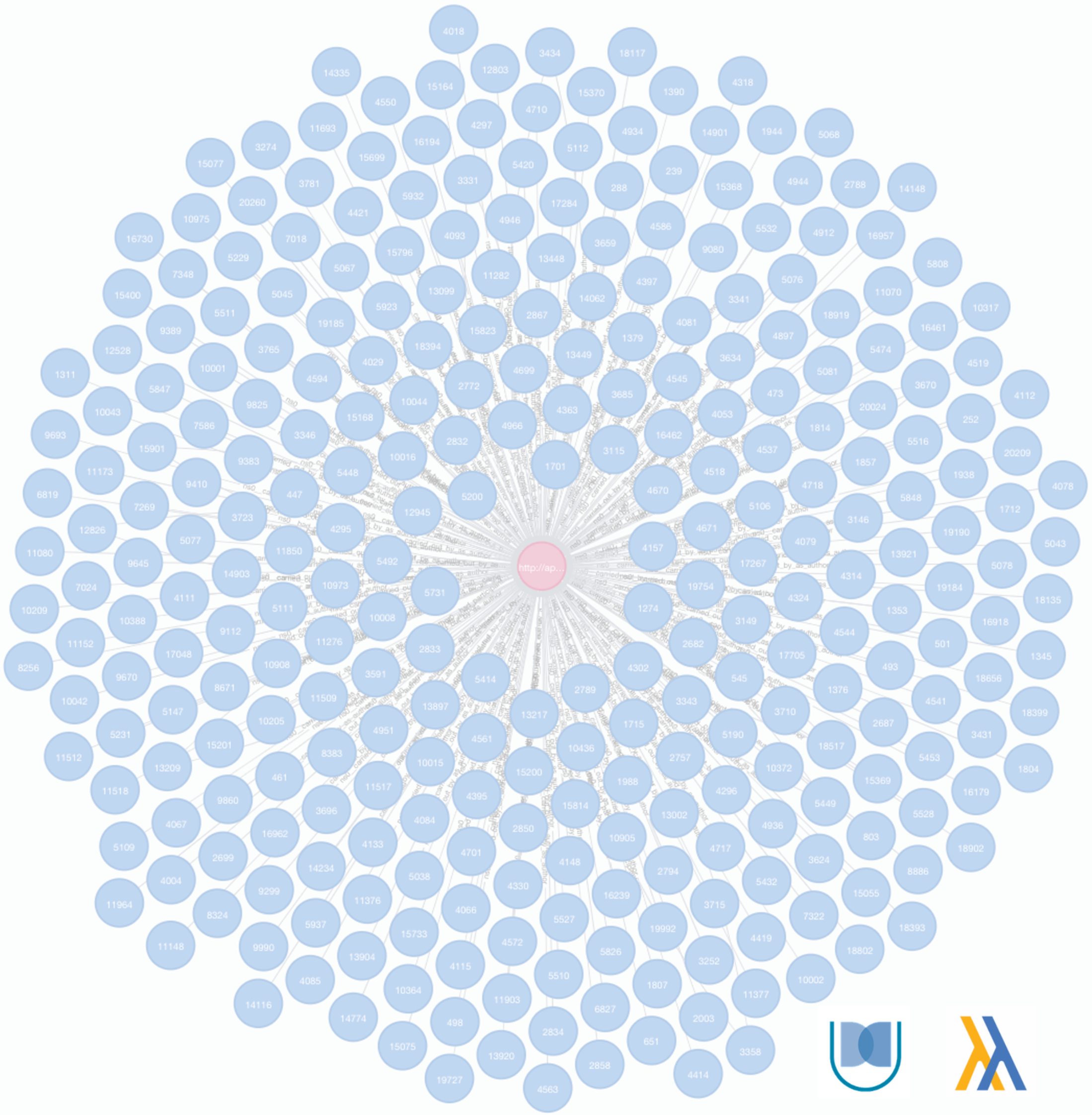
Select data for modelling
Record fields
Express frequent queries in natural language
Select the fields that answer these queries
Keep track of the possible queries each metadata field answers
Select the interesting and/or common queries
Search existing reference model for the selected metadata fields
[exists]
[adequate]
Select model
[not adequate]
Edit model
[not exists]
Create model
Design model using software tool
Extract model in desired format

[6] http://www.cidoc-crm.org/

[7] https://neo4j.com/

Semantic graph

- 228.848 RDF triples loaded so far (ongoing)
- The semantic graph serves for performing queries of higher complexity
- However, it requires some experience in RDF query languages



```
PREFIX apl: <https://services.apollonis-infrastructure.gr/decade1940#>
PREFIX cidoc: <http://www.cidoc-crm.org/cidoc-crm/>
SELECT ?identifier ?number ?location ?title
WHERE {
    ?ar cidoc:P1_is_identified_by ?identifier.
    ?ar cidoc:P55_has_current_location ?number.
    ?number cidoc:P89_falls_within ?location.
    ?ar cidoc:P102_has_title ?title.
    ?ar cidoc:P128_carries ?text.
    ?text cidoc:P67_refers_to ?activity.
    ?activity cidoc:P117_occurs_during    apl:Περίοδος_Β_Παγκοσμίου_Πολέμου.
}
```

```
PREFIX apl: <https://services.apollonis-infrastructure.gr/decade1940#>
PREFIX cidoc: <http://www.cidoc-crm.org/cidoc-crm/>
SELECT ?identifier ?number ?location ?period ?title
WHERE {
    ?ar cidoc:P1_is_identified_by ?identifier.
    ?ar cidoc:P55_has_current_location ?number.
    ?number cidoc:P89_falls_within ?location.
    ?ar cidoc:P102_has_title ?title.
    ?ar cidoc:P128_carries ?text.
    ?text cidoc:P67_refers_to ?period.
    ?period cidoc:P120_occurs_before apl:Περίοδος_Εμφυλίου_Πολέμου
}
```

- To support non-experienced users, we have expressed 14 frequent research queries initially in SPARQL
- Currently being expressed also in Cypher

Frequent queries (in natural language)

Search for archives:

1. of a specific medium type (e.g. photograph)
2. that are digitized or born digital or analogue
3. owned by a specific institution
4. of a specific type (e.g. army documents, thesis)
5. produced by a specific producer or by specific types of institutions
6. created in a specific time period
7. having a specific title
8. containing text that is written by a specific actor
9. of a specific subject
10. written in a specific language
11. that carry information regarding a specific actor
12. that carry information about a specific place
13. that carry information about activities of a specific type (e.g. military operations)
14. that carry information about a specific time period

Visit our alpha version of ‘Decade 1940’ here (in Greek):

https://services.apollonis-infrastructure.gr/Apol/forties_searchIndex.html

As our work is still ongoing, we are open to collaborations!

Co-financed by Greece and the European Union

Thank you!