Challenges and solutions towards creating a semantic network of historical maritime data

Presenter: Dr. Pavlos Fafalios

Centre for Cultural Informatics (CCI), Institute of Computer Science (ICS), FORTH

fafalios@ics.forth.gr

The team (behind the presented work)

- **Athina Kritsotaki**, Data modeling engineer
- **Korina Doerr**, Systems design engineer
- **Anastasia Axaridou**, R&D engineer
- **Kostas Petrakis**, R&D engineer
- **Giorgos Samaritakis**, R&D engineer
- **George Bruseker,** Postdoctoral researcher
- **Pavlos Fafalios**, Postdoctoral researcher
- **Yannis Tzitzikas**, CCI Head
- **Martin Doerr**, CCI Honorary Head

Centre for Cultural Informatics (CCI)
Institute of Computer Science (ICS)
FORTH

Introduction

Historical research

- Historical science
  - A vast area of research concerns the analysis of **historical archival sources**
  - **Aim**: describe, examine and question a sequence of past events, and investigate patterns of causes and effects
- Common activities
  1. **Documentation / digitization / transcription**
  2. **Curation / consolidation**
  3. **Quantitative** and/or **qualitative analysis**

➢ Growing need for **holistic data management** approaches!

Our contribution

- A **workflow** and **tools** for holistic data management in **historical (archival) research**
  - **Provenance-aware**
  - **Highly-recursive**
  - Using established **documentation** and **publication standards**!
    - Towards **data sustainability** and **semantic interoperability!**

Applied in a large-scale research project of **Maritime History**

Seafaring Lives in Transition

Context, requirements and challenges

Context

- The project '**SeaLiT**' - http://www.sealitproject.eu/
  - **Full title**: Seafaring Lives in Transition. Mediterranean Maritime Labour and Shipping during Globalization, 1850s-1920s
  - **Project type**: ERC starting grant (Feb 2017–Jan 2023, ID: 714437)
  - **Field**: Maritime History
  - **Principal Investigator**: Dr. Apostolos Delis, Centre for Maritime History, Institute of Mediterranean Studies - FORTH

- Research groups in 5 countries
  - Greece, Spain, Italy, France, Croatia

SeaLiT – The research focus

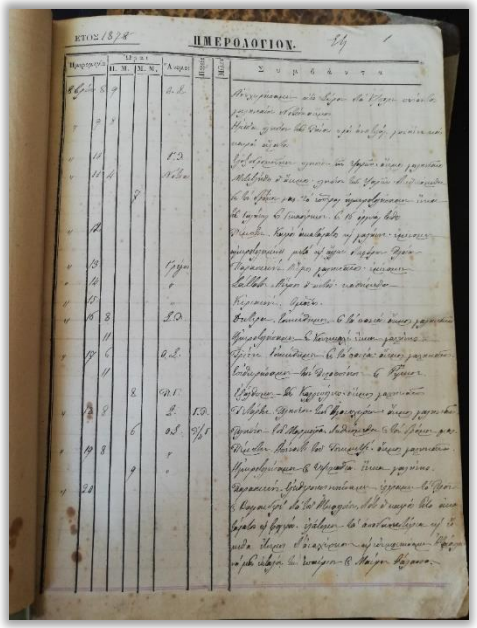
- The **transition from sail to steam navigation** and its effects on seafaring populations
  - The effects of technological innovation on seafaring people and maritime communities
  - The maritime labour market, the life on board and ashore, the evolving relations among ship owners, captains, crew and local societies

- Maritime regions and period of investigation:
  - **Mediterranean** and the **Black Sea**
  - **1850s-1920s**

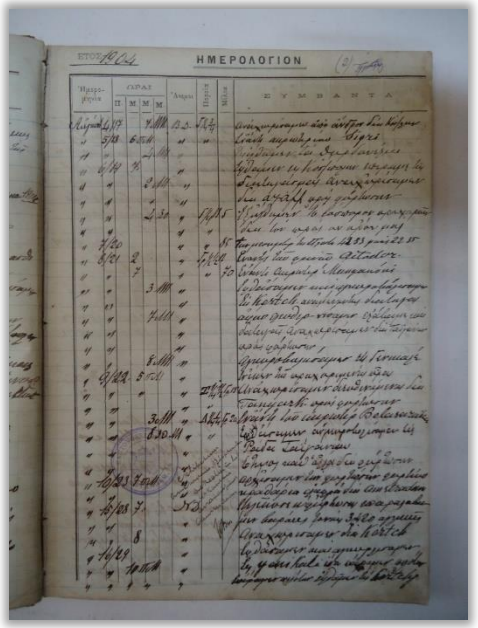
SeaLiT – The archival sources

- Crew lists
- Ship logbooks
- Payrolls
- Sailor / student registers
- Ship registers
- Employment records
- Account books
- Census data
- …

- *Handwritten!*
- *In different languages !*
- *From different issuing authorities!*



Brig Eleni Koupa
(Private Archive of Evangelos Rafalias, Hydra)



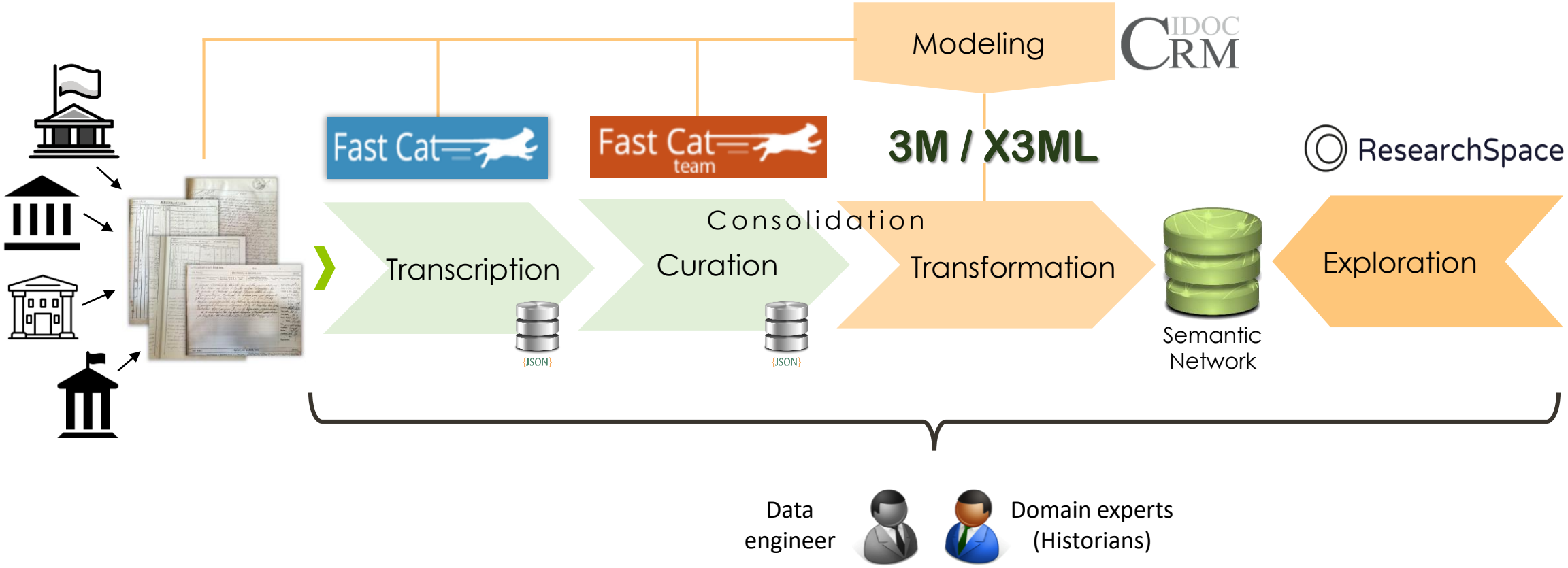
Cargo steamer Leonidas
(Syrmas Archive, Hellenic Literary and Historical Archive [ELIA], Athens)

SeaLiT – Data management requirements and challenges

- Data digitization
  - Faithful transcription of the original sources
  - **Not just metadata recording!**
- Data consolidation
  - Correction, normalization, matching, enrichment, integration, …
  - **Important: preservation of provenance information!**
- Data analysis and exploration
  - Analysis and visualization of collective social phenomena
  - **Use the consolidated data as a primary source for historical research!**

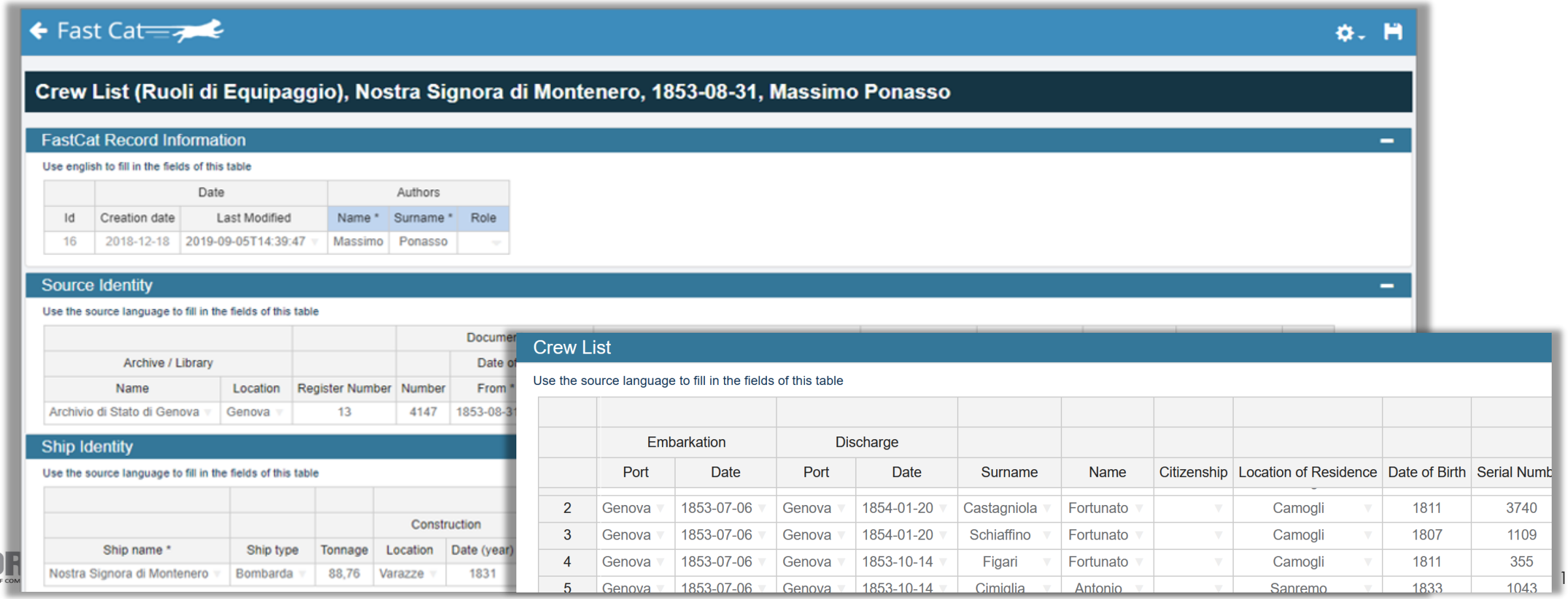
The workflow and tools

Workflow overview



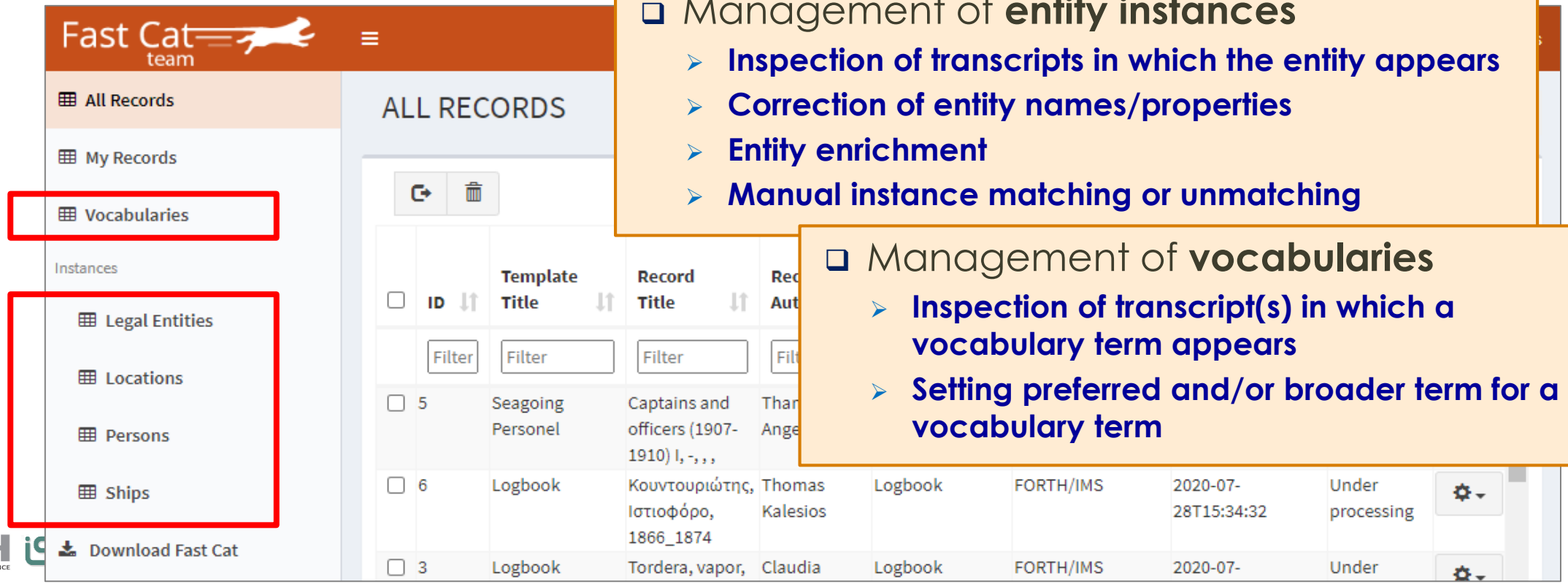
Data transcription with FAST CAT

- Documents are transcribed as '**records**' belonging to '**templates**'
  - A '**template**' represents the structure of a single data source
  - It allows organizing the data and metadata in tabular form (tables)



Data curation with FAST CAT TEAM

- FAST CAT TEAM:
  - A special environment that allows the **curation** of ***entities*** and ***vocabulary terms***
  - **It decouples data curation from data transcription**
    - Avoids 'spoiling' the original transcripts / Important for verification & long-term validity



- Management of **entity instances**
  - **Inspection of transcripts in which the entity appears**
  - **Correction of entity names/properties**
  - **Entity enrichment**
  - **Manual instance matching or unmatching**

- Management of **vocabularies**
  - **Inspection of transcript(s) in which a vocabulary term appears**
  - **Setting preferred and/or broader term for a vocabulary term**

FAST CAT and FAST CAT TEAM

Use in SeaLiT

- 30 users in 5 countries
- 20 templates
- >600 records
- 5 languages

- 52 vocabularies
- >76K person instances
- >8.8K location instances
- >2.3K ship instances
- >1.1K legal entity instances

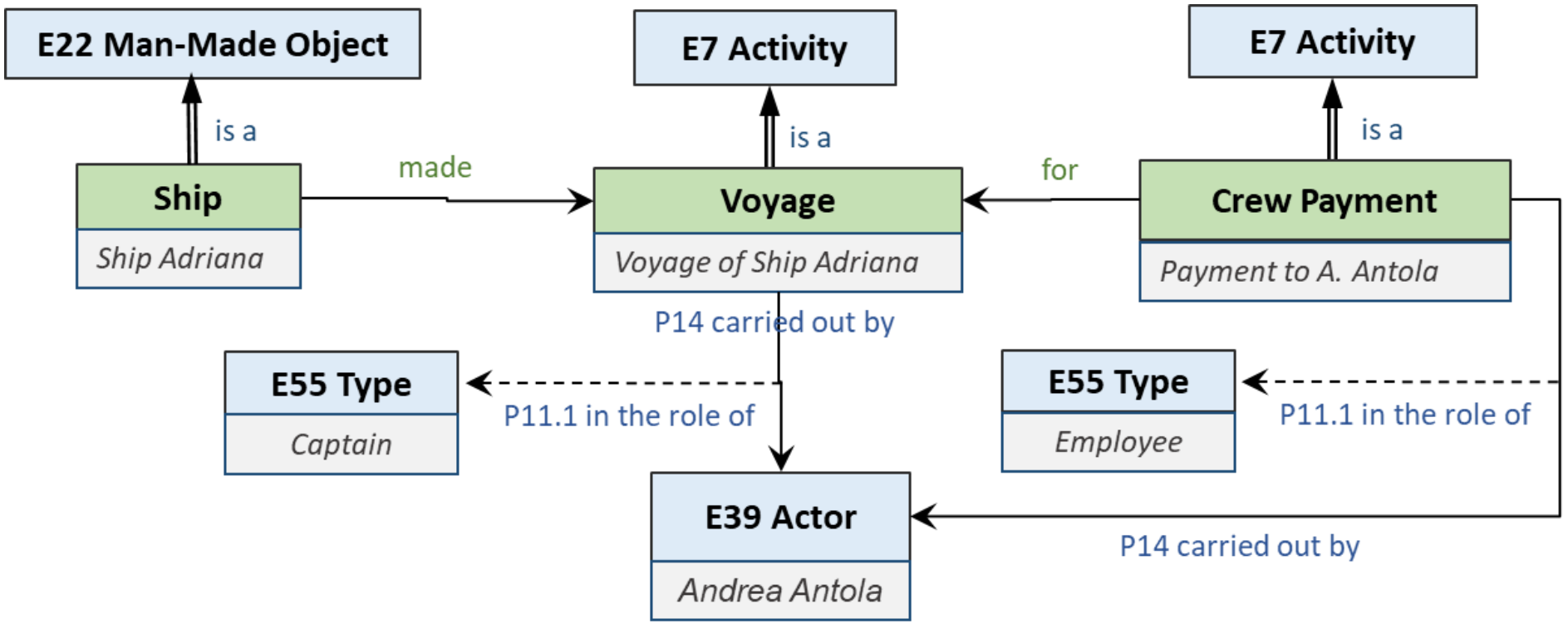
Ontology-based Data Integration

(and creation of a rich semantic network)

- Steps:
  1. Decide on the **domain ontology**
  2. Create **schema mappings**
  3. Run the **transformations**

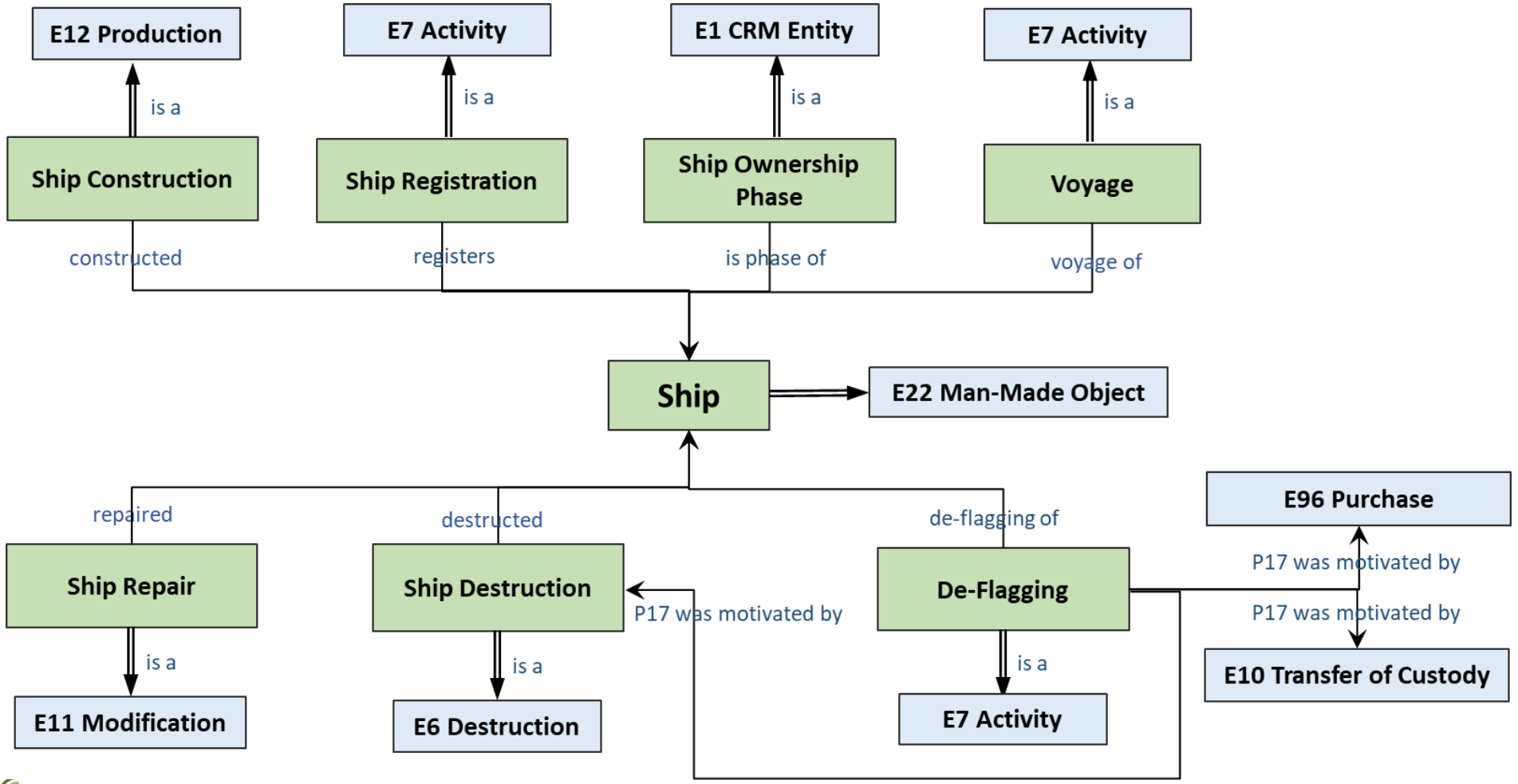
The domain ontology

- Data model compatible with **CIDOC-CRM**



The domain ontology

- Modeling activities related to a **Ship**

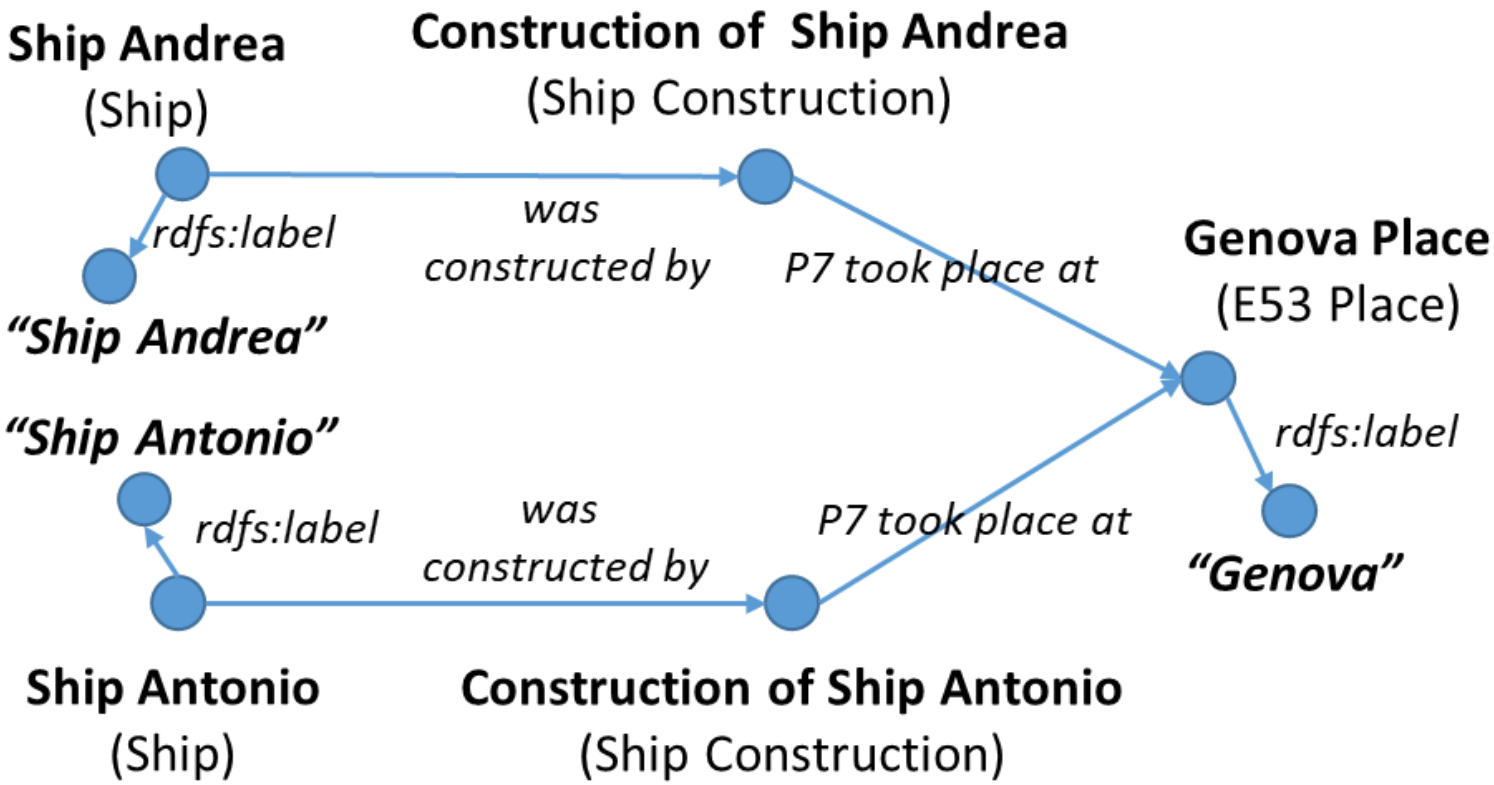


Definition of schema mappings and transformation

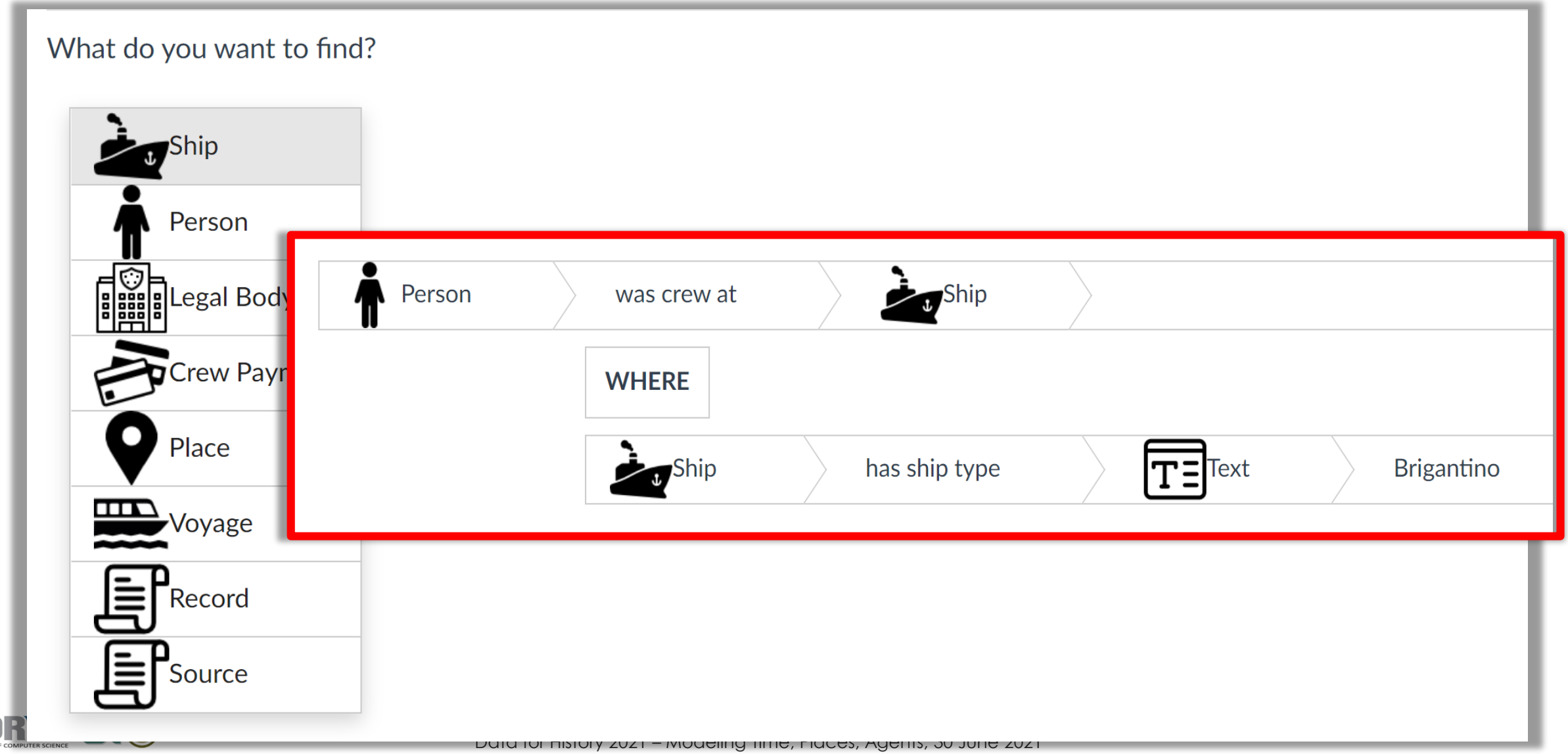
- **X3ML toolkit** https://www.ics.forth.gr/isl/x3ml-toolkit
  - X3ML mapping definition language
  - 3M Editor (user interface supporting the creation of mappings)
  - X3ML Engine (for executing the transformations)

Semantic Network

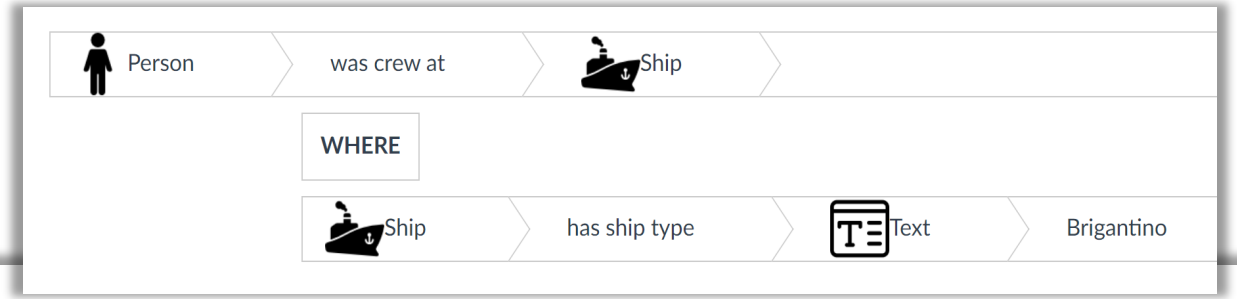
- A (very small) part of the derived semantic network



Data exploration using ResearchSpace



Data exploration using ResearchSpace



Timeline | Chart | Table | Rows | Map | Gric

Visualization Context

embarked at

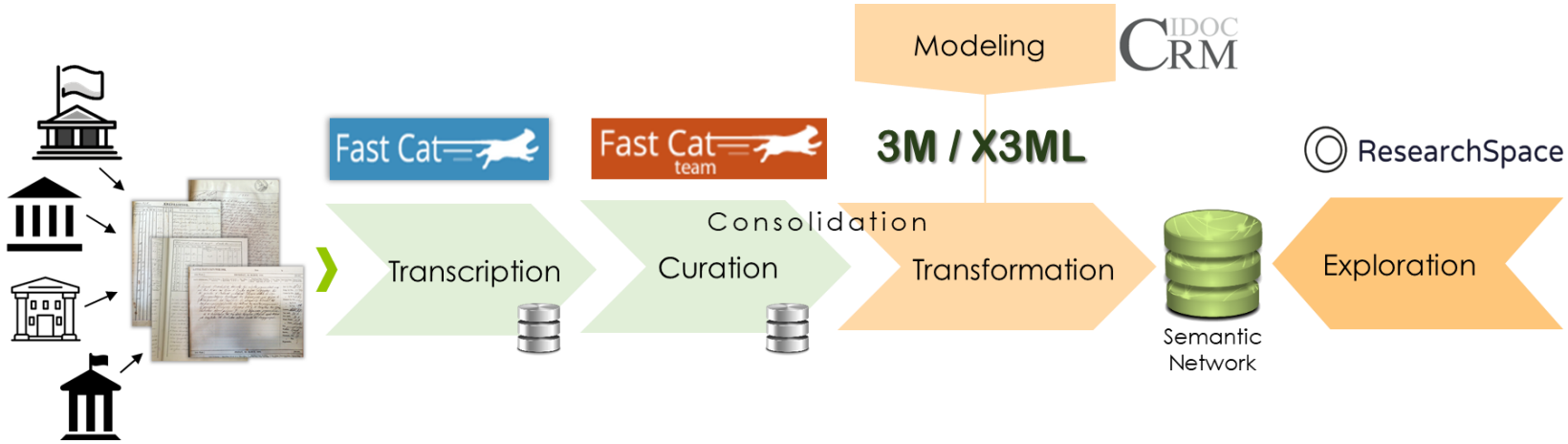
Line | Bar | Radar | Pie | Donut

Amsterdam, Cóbh, Cagliari, Camogli, Cardiff, Castellammare di Stabia, Falmouth, Galaţi, Genova, Gibraltar, Strait of, Hartlepool, İstanbul, Livorno, Malta, Mariupol', Marseille, Messina, Napoli, Nice, Odessa, Portoferraio, Portofino, Santa Margherita Ligure, Savona, Spezia, Swansea

Conclusion, Limitations and Future Work

Conclusion

- **Workflow** for holistic data management in historical archival research
  - Supported by **tools**
  - **Provenance-aware**
  - Following **standards** (CIDOC-CRM, RDF)



- Successful **application** in a large scale **history research project** (SeaLiT)

Limitations and Challenging issues

- **Configuration/update** of the FAST CAT templates
  - Cannot (currently) be done by historians
- **Instance mapping** and **vocabulary curation** in FAST CAT TEAM
  - Very time consuming and laborious
- **Modeling** and **creation of mappings**
  - Difficult and time consuming task / Needs expertise in CIDOC-CRM

Current / Future Work

- Improved **configurability** of FAST CAT templates
- Full automation of **recursive workflow**
- Semi-automated **instance matching**

Related Publications

Petrakis et al. **“Digitizing, Curating and Visualizing Archival Sources of Maritime History: the case of ship logbooks of the nineteenth and twentieth centuries”**. *Drassana: revista del Museu Marítim*, (28), pp.60-87. 2020. https://doi.org/10.51829/Drassana.28.649

[Available as open access: https://raco.cat/index.php/Drassana/article/view/385133/478237]

Fafalios et al. **“FAST CAT: Collaborative Data Entry and Curation for Semantic Interoperability in Digital Humanities”**. ACM Journal on Computing and Cultural Heritage, 2021. https://doi.org/10.1145/3461460

[Preprint: https://arxiv.org/pdf/2105.13733.pdf]

Thank you!

Pavlos Fafalios

fafalios@ics.forth.gr