Digital Seafaring:
Digitising, Curating and Exploring Archival Sources of Maritime History

Pavlos Fafalios, Martin Doerr

Centre for Cultural Informatics,
Institute of Computer Science,
Foundation for Research & Technology (FORTH)

Heraklion, Greece

Apostolos Delis

Centre of Maritime History,
Institute of Mediterranean Studies,
Foundation for Research & Technology (FORTH)

Rethymno, Greece

Baltic Connections 2022
Pavlos Fafalios (fafalios@ics.forth.gr) | June 2022

Outline

- Introduction and Context
- Workflow Model
- IT tools
  - **FastCat** (data transcription)
  - **FastCat Team** (data curation)
  - **SeaLiT Ontology** (data integration)
  - **FastCat Catalogues** (data exploration & analysis)
  - **SeaLiT ResearchSpace** (data exploration & analysis)
  - **Ship Voyages** (data visualization on map)
- Conclusion and Future Work

Introduction and Context

Historical Science and Archival Research

- Historical science
  - A vast area of research concerns the study and analysis of **archival documents**
  - **Aim**: describe, examine and question a sequence of past events, and investigate patterns of causes and effects

- The case of **Maritime History**:
  - The archival documents often have a '**repetitive' structure** and provide information about a set of '**entities of interest**' (ships, sailors, etc.)
    - Logbooks, Crew Lists, Payrolls, Sailors Registers, Account Books, …

Logbook of Cargo steamer Leonidas (1904-1906)
Syrmas Archive, Hellenic Literary and Historical Archive [ELIA], Athens

Historical Science and Archival Research

- Common data management activities
  - **Transcription** & **documentation** of the archival documents
    - Not just metadata recording!
  - **Curation** & **consolidation**
    - For integrating and better analyzing data of different sources
  - **Data exploration** & **analysis** (both **quantitatively** and **qualitatively**)
    - Depending on the kind of data and the type of research questions that need to be answered

Archival Documents

Transcription & Documentation → Curation & consolidation → Data exploration & analysis

Historical Knowledge

The case of the SeaLiT project

- 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, IMS-FORTH
  - **Research groups** in Greece, Spain, Italy, France, Croatia

- The research focus
  - The **transition from sail to steam navigation** and its effects on seafaring populations
    - The effects of the technological innovation on the seafaring people and the maritime communities

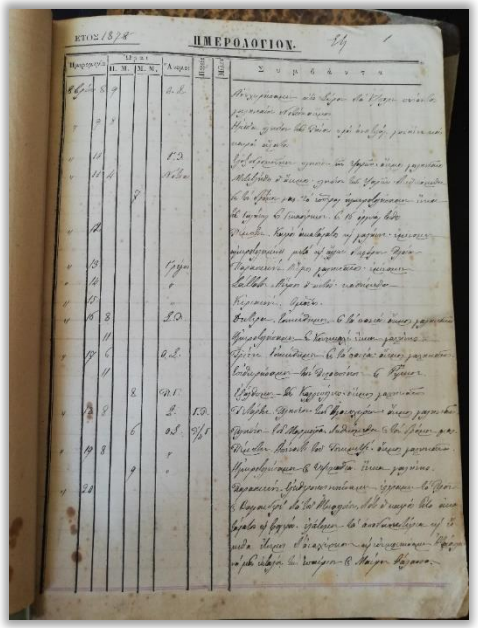
Seafaring Lives in Transition

The case of the SeaLiT project

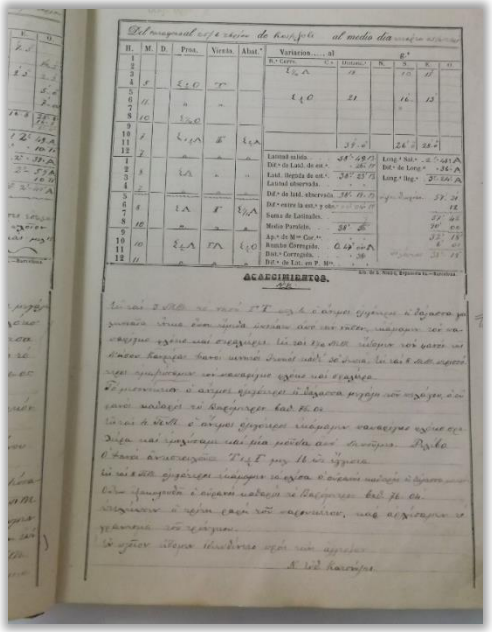
- The archival sources
  - Crew and displacement lists
  - Ship logbooks
  - Payrolls
  - Sailors registers
  - Naval Ship registers
  - Students registers
  - Employment records
  - Account books
  - Censuses
  - …

- *Handwritten, in 5 languages!*
- *Provided by different authorities in several countries!*

Full archival corpus:
https://sealitproject.eu/archival-corpus



Logbook (1878-1888)
Brig Eleni Koupa
Private Archive of Evangelos Rafalias, Hydra



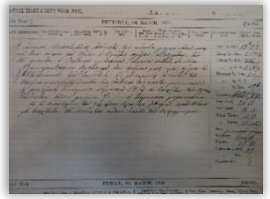
Logbook (1882-1885)
Barquentine Asimoula
Maritime Museum of Galaxidi

Data Management Challenges

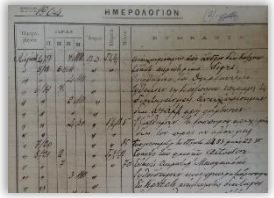
- **Motivating Scenario**: Exploring information about a particular **ship**
  - Complementary information about the same ship may exist in different archival documents

Account book:
Information about
the ship‘s owners

Naval ship register list:
Construction details
and ship characteristics



Civil register:
Information about
ship’s crew members



Crew and displacement list:
Ship’s voyages and crew
members

Data Management Challenges

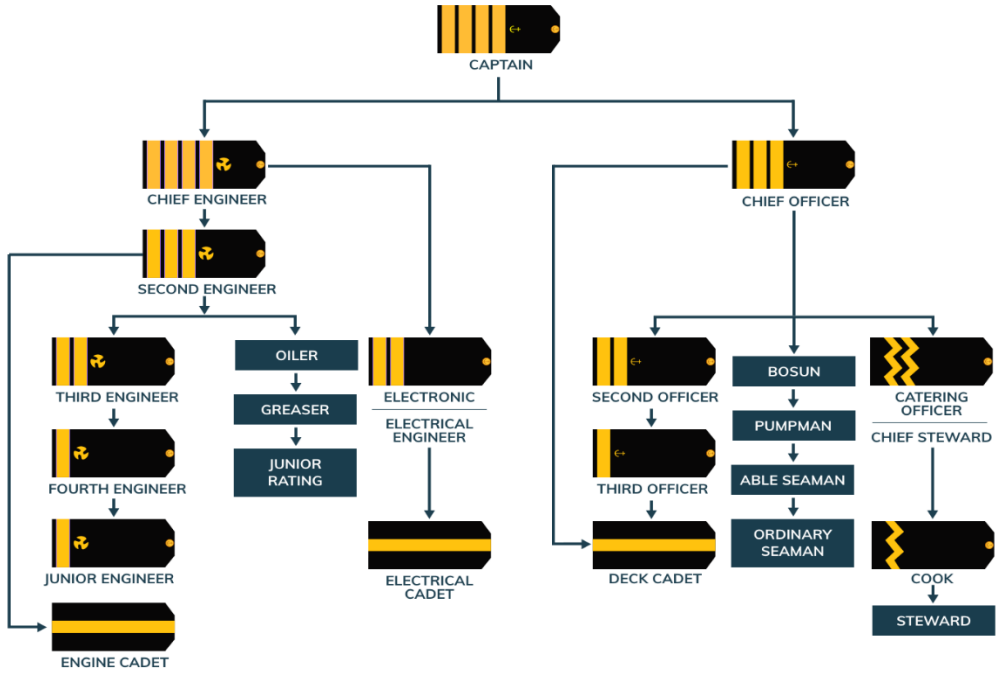
- **Motivating Scenario**: Exploring information about a particular **ship**
  - The **same ship**, or the same ship-related **entity/term,** might appear under different representations in the archival documents, due to:
    - Typos (**"Vindobonna"** vs **"Vindobona"**, "**Gaetano Schiaffino**" *vs* "**Gaetano Schiafino**")
    - Different languages ("**Genoa**" *vs* "**Genova**", "**brigantine**" *vs* "**brigantino**")
    - Unrecognizable characters ("**G??oa**")
    - Uncertainty ("**[Genoa]**")
    - Use of abbreviation ("**Gaetano Schiaffino**" *vs* "**G. Schiaffino**")
    - Difference in a property value (person's birth place: "**1880**" *vs* "**15-2-1880**")

Data Management Challenges

- **Motivating Scenario**: Exploring information about a particular **ship**
  - Data **enrichment** might be needed for better analysis and exploration services

Adding **coordinates** to locations
(for map visualization)

Creating **hierarchies** for terms
(e.g., for professions)



Data Management Challenges

- Need for an **holistic data management approach, supported by IT tools**
  - **Transcription** of the important historical information
  - **Curation** of the transcribed data
  - **Integration** of the transcribed and curated data

for enabling valid (qualitative & quantitative) analysis

- *"Find the **nationality** of **sailors** of large **ships** of a specific **type** (e.g., **Brig**) that arrived at a specific **port**"*

but without loosing the important provenance information!

- Important for **verification** and the **long-term validity** of the research results

Our contribution

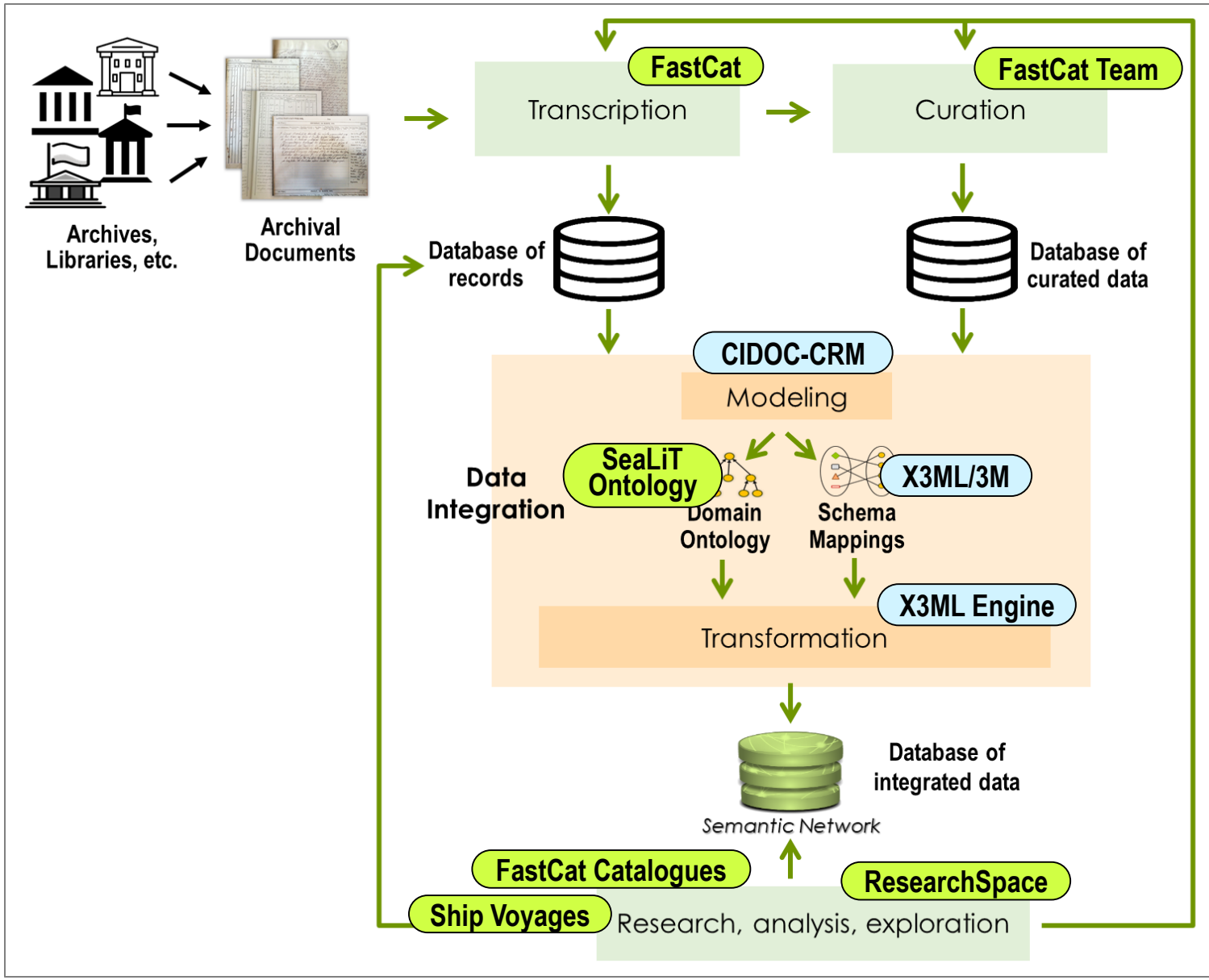
- A **workflow model** for holistic data management in historical-archival research
  - ***Transcription → Curation → Integration → Analysis*** *&* ***Exploration***
  - **Provenance-aware, following established documentation standards**
    - Aiming at the production of data with **high value** and **long-term validity**, that can be **re-used!**

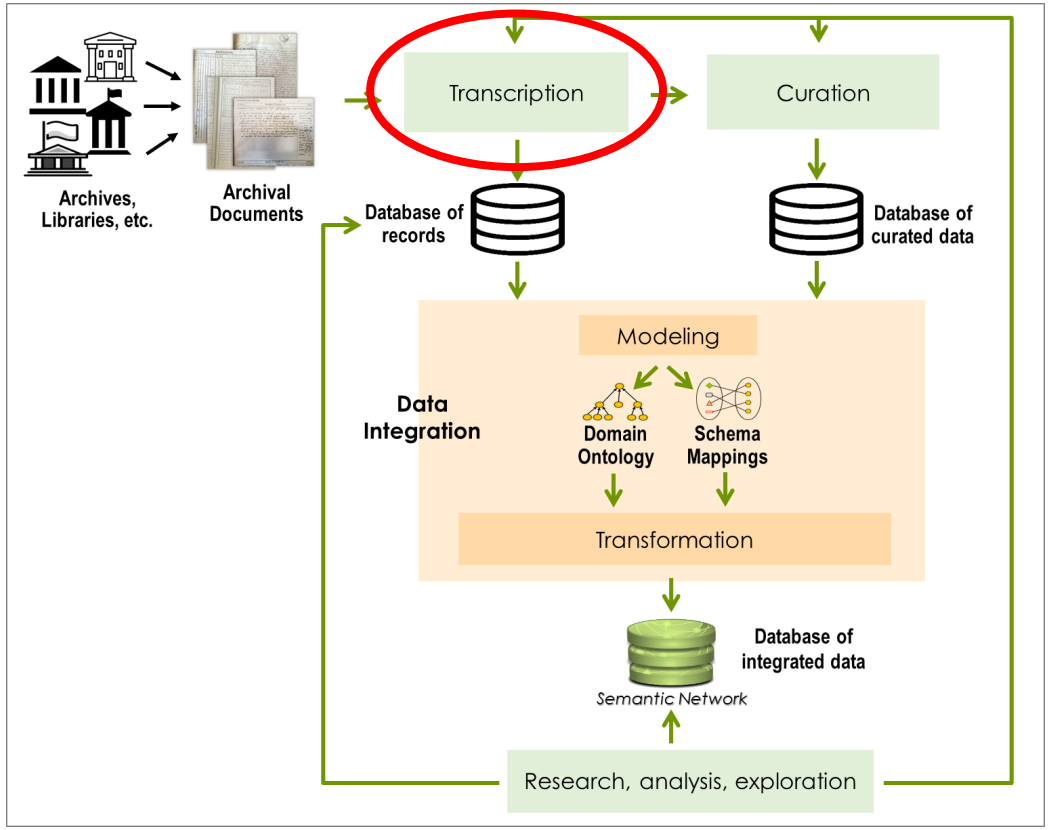
- **IT tools** for each step of the workflow
  - **FastCat** (*transcription*)
  - **FastCat Team** (*curation*)
  - **SeaLiT Ontology** (*integration*)
  - **FastCat Catalogues** (*analysis & exploration*)
  - **SeaLiT ResearchSpace** (*analysis & exploration*)
  - **Ship Voyages** (*analysis & exploration*)

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

Workflow Model & IT tools

Workflow Model

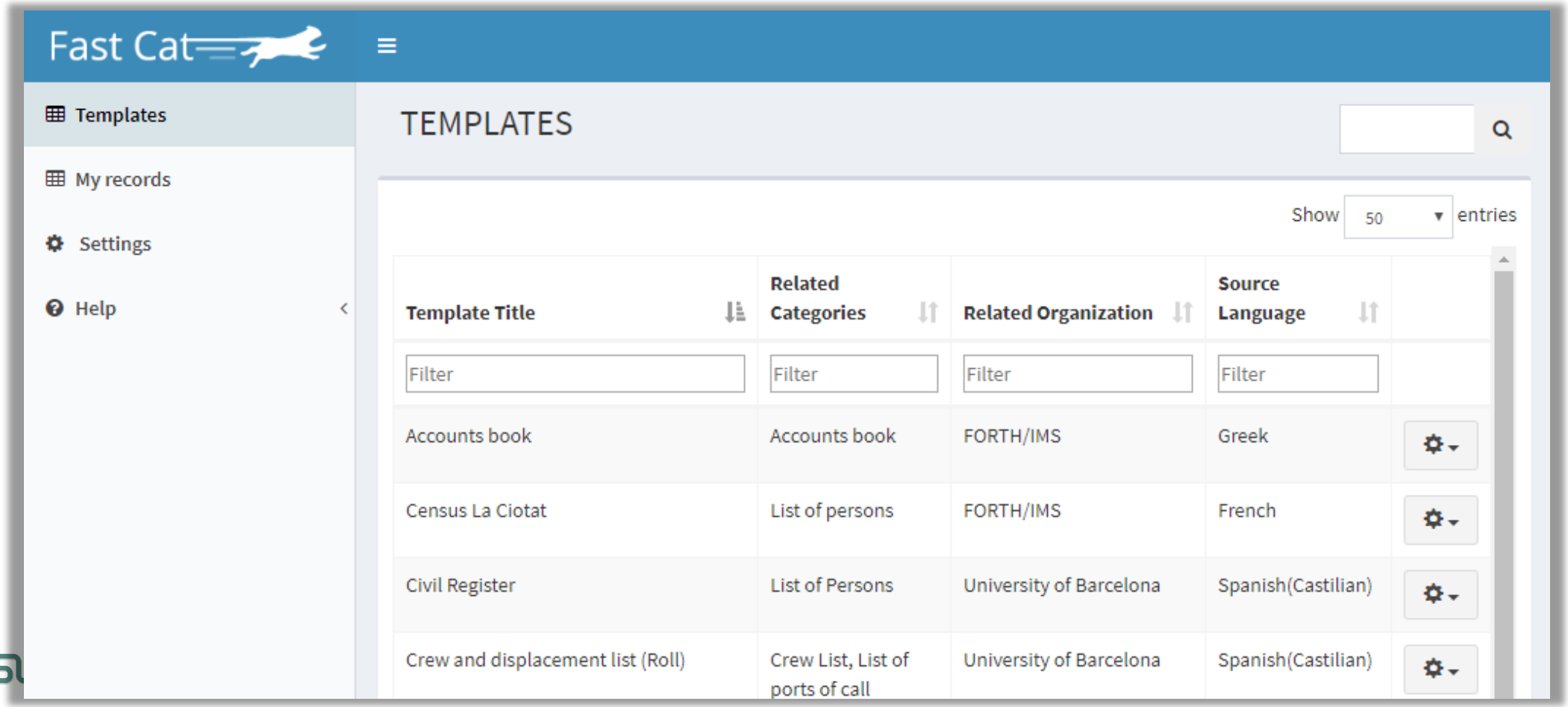




Data transcription with FastCat

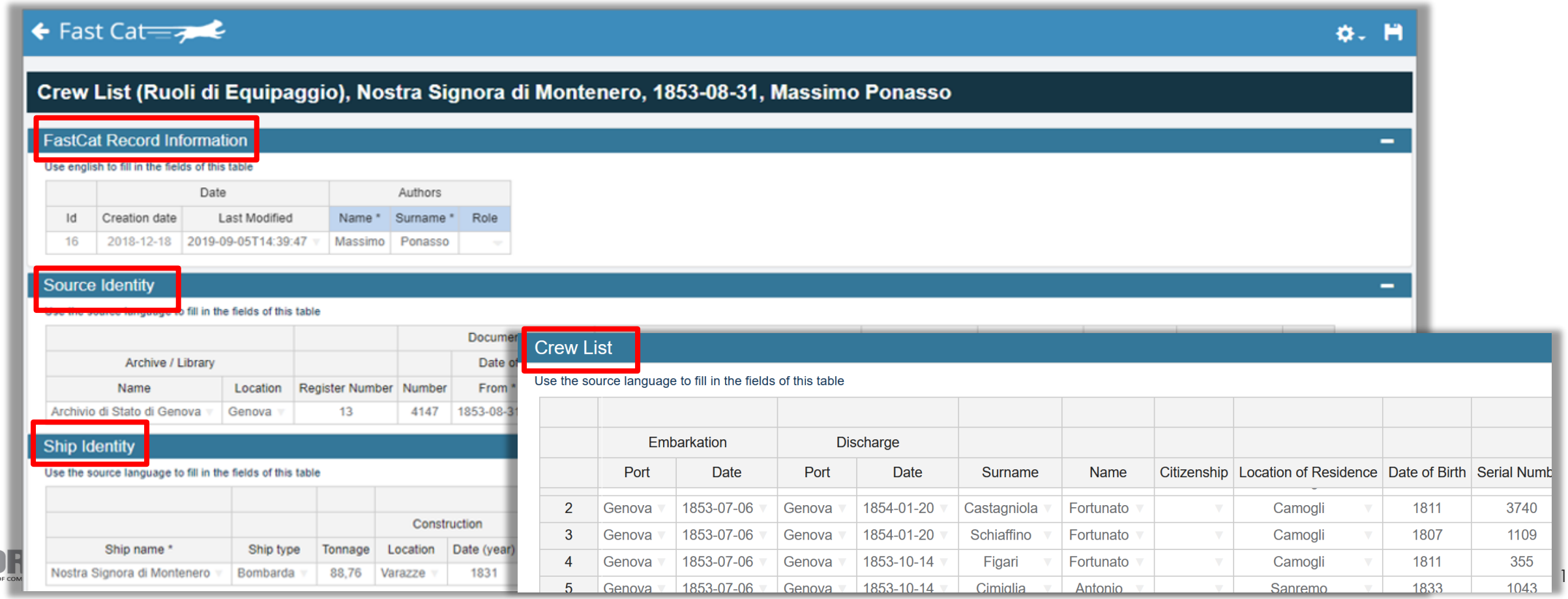
Data transcription with FastCat

- Data from different archival sources can be transcribed as '**records**' belonging to '**templates**'
  - A '**template**' represents the structure of a single data source



Data transcription with FastCat

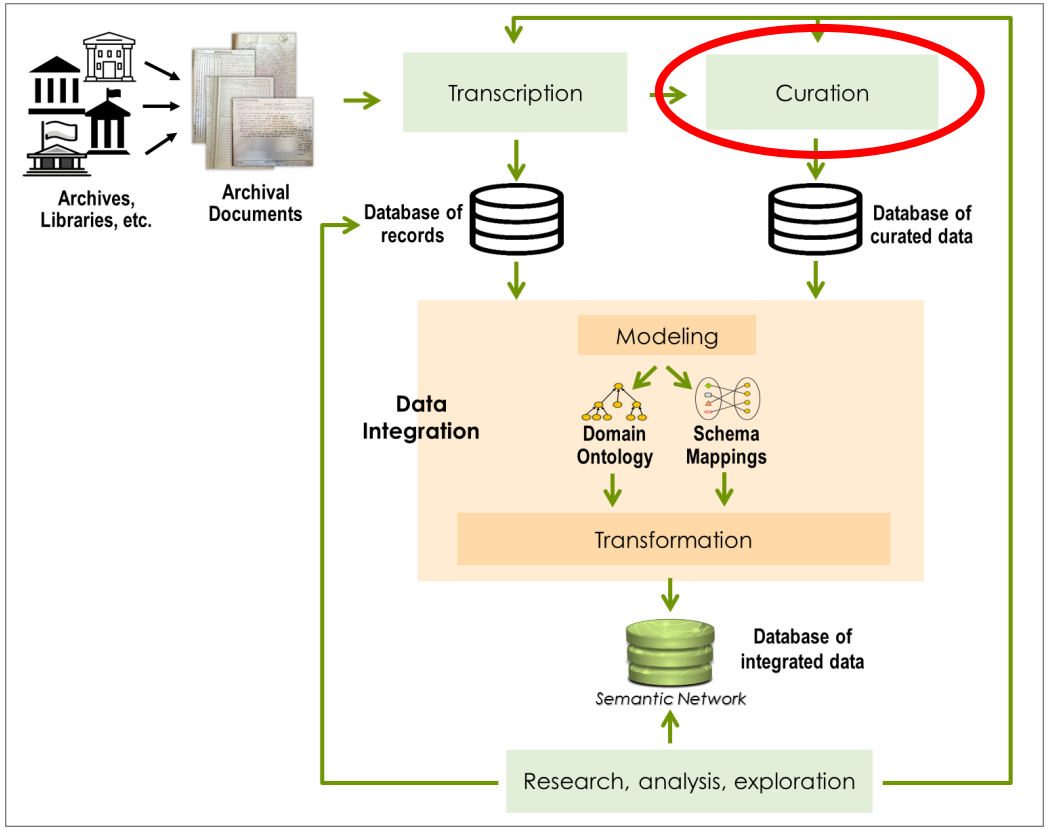
- Data from different archival sources can be transcribed as '**records**' belonging to '**templates**'
  - A '**record**' organizes the data and metadata in tabular form (tables)



Data transcription with FastCat

- Use in **SeaLiT**:
  - ~30 users
  - 20 templates
  - >600 records
  - 5 languages

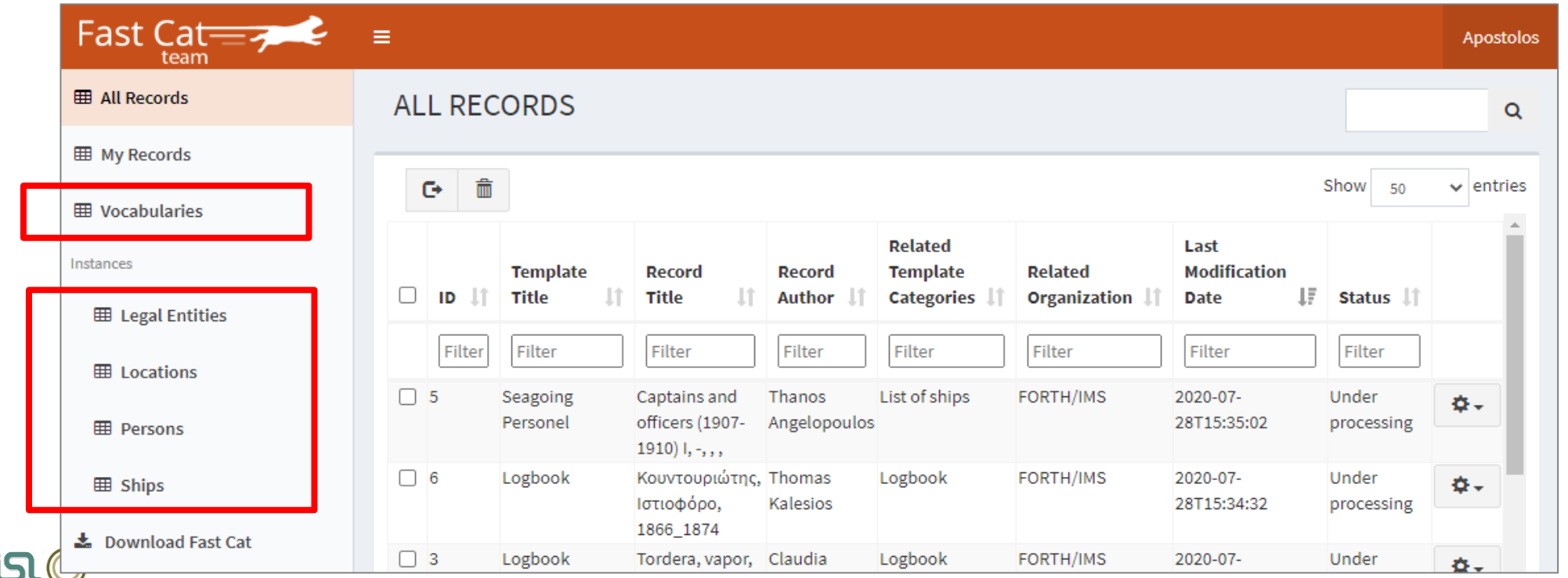
No	Template Name	# Records
1	Accounts book	14
2	Census La Ciotat	63
3	Civil Register	29
4	Crew and displacement list (Roll)	35
5	Crew List (Ruoli di Equipaggio)	98
6	Employment records, Shipyards of Messageries Maritimes, La Ciotat	50
7	First national all-Russian census of the Russian Empire	6
8	General Spanish Crew List	64
9	Maritime Register of the State for La Ciotat	1
10	List of ships	71
11	Logbook	17
12	Naval Ship Register List	2
13	Notarial Deeds	10
14	Payroll	7
15	Payroll of Russian Steam Navigation and Trading Company	14
16	Register of Maritime personel	4
17	Register of Maritime workers (Matricole della gente di mare)	6
18	Sailors register (Libro de registro de marineros)	52
19	Seagoing Personel	52
20	Students Register	10



Data curation with FastCat Team

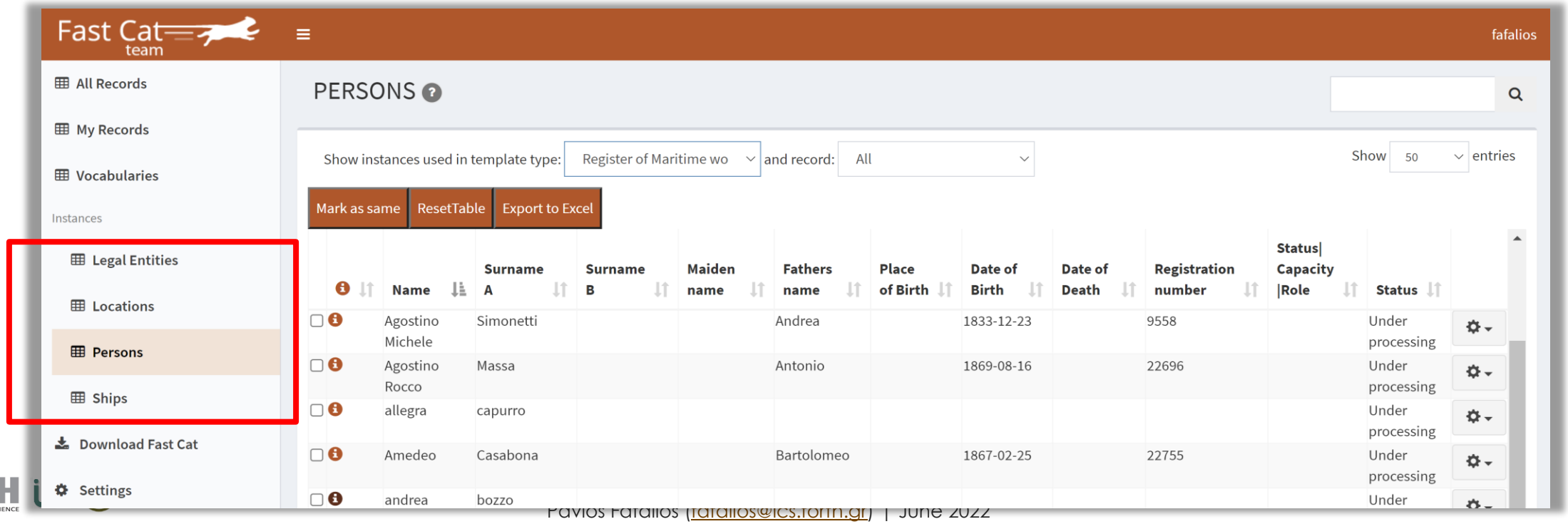
Data curation with FastCat Team

- **FastCat Team**:
  - A special environment within FAST CAT that allows the **collaborative curation** of the ***entities*** and ***vocabulary terms*** *that appear in the transcripts*
  - It **decouples** data curation from data transcription
    - We avoid 'spoiling' the original transcripts!



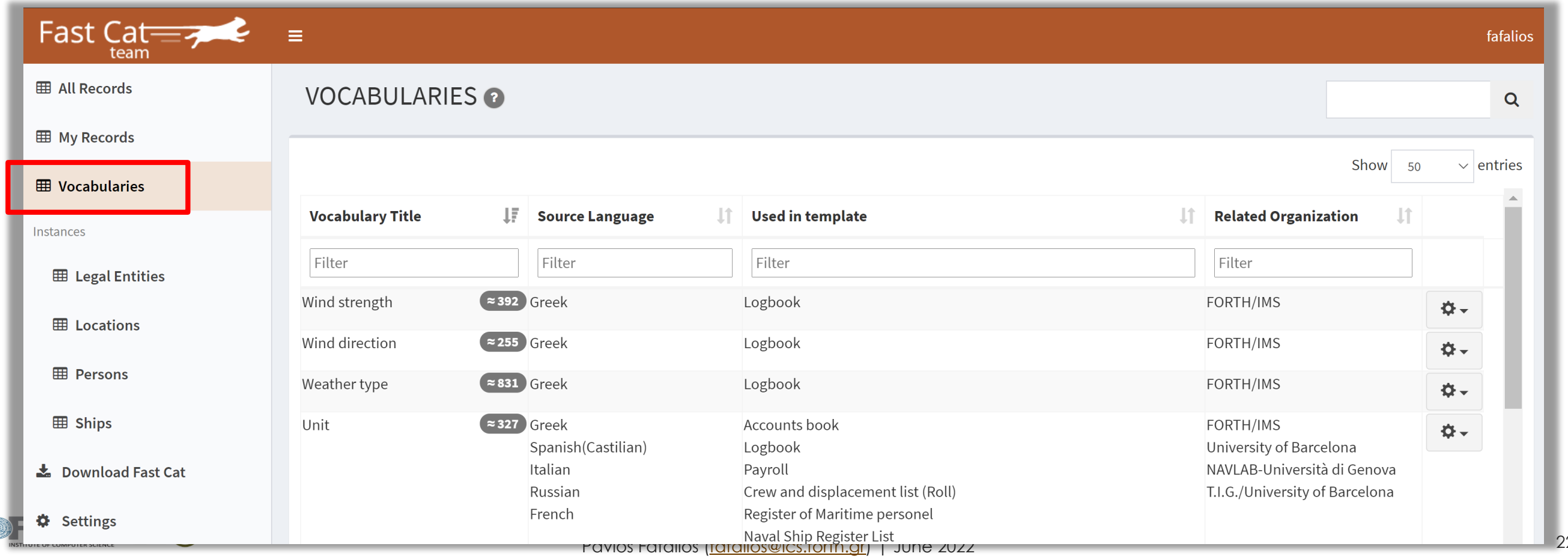
Data curation with FastCat Team

- Management of **entity instances**
  - Inspection of **records** in which the entity appears
  - **Correction** of entity names or other entity properties
  - Instance **matching** or **unmatching**
  - Entity **enrichment**



Data curation with FastCat Team

- Management of **vocabularies**
  - Inspection of **records** in which a vocabulary term appears
  - Setting a **preferred** and/or a **broader** term for a vocabulary term

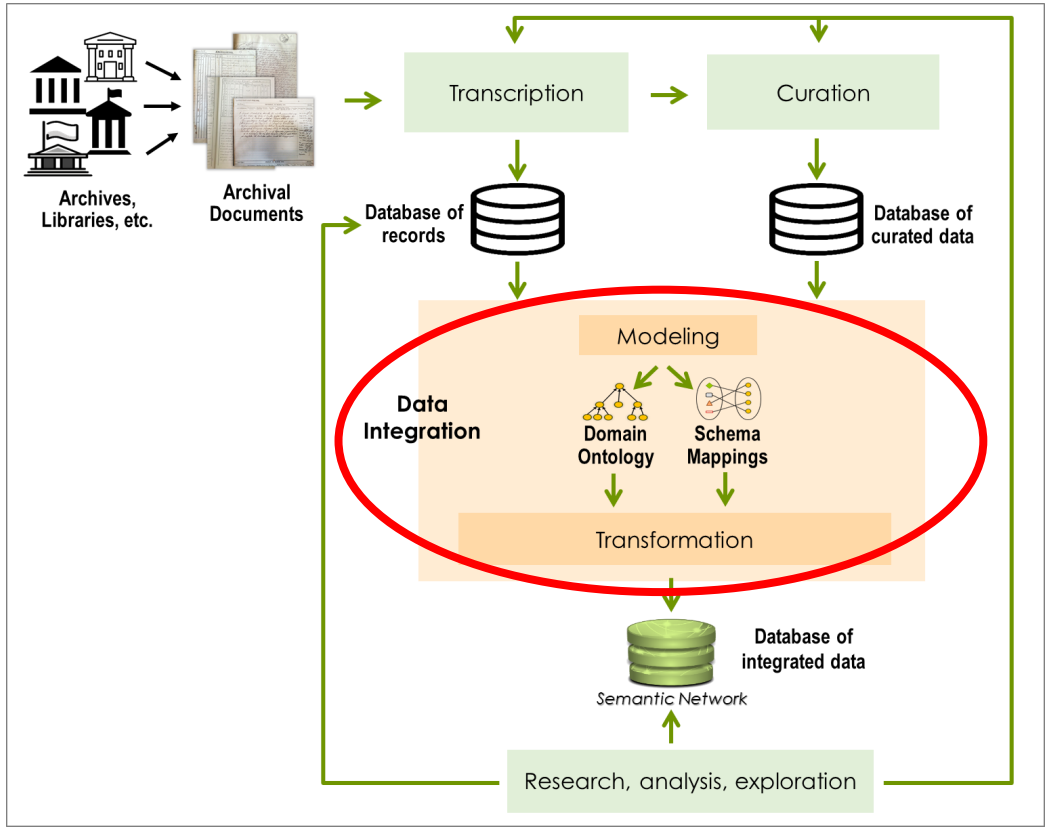


Data curation with FastCat Team

- Use in **SeaLiT**:
  - ~50 vocabularies *(ship type, profession, marital status, religion, …)*
  - >90K person instances
  - >9.8K location instances
  - >2.4K ship instances
  - >1.1K legal entity instances

More about FastCat and FastCat Team:

- P. Fafalios et al. (2021). FAST CAT: collaborative data entry and curation for semantic interoperability in digital humanities. *Journal on Computing and Cultural Heritage (JOCCH)*, *14*(4), 1-20. http://users.ics.forth.gr/~fafalios/files/pubs/fafaliosJOCCH2021.pdf
- **https://www.ics.forth.gr/isl/fast-cat**



Data Integration with SeaLiT Ontology

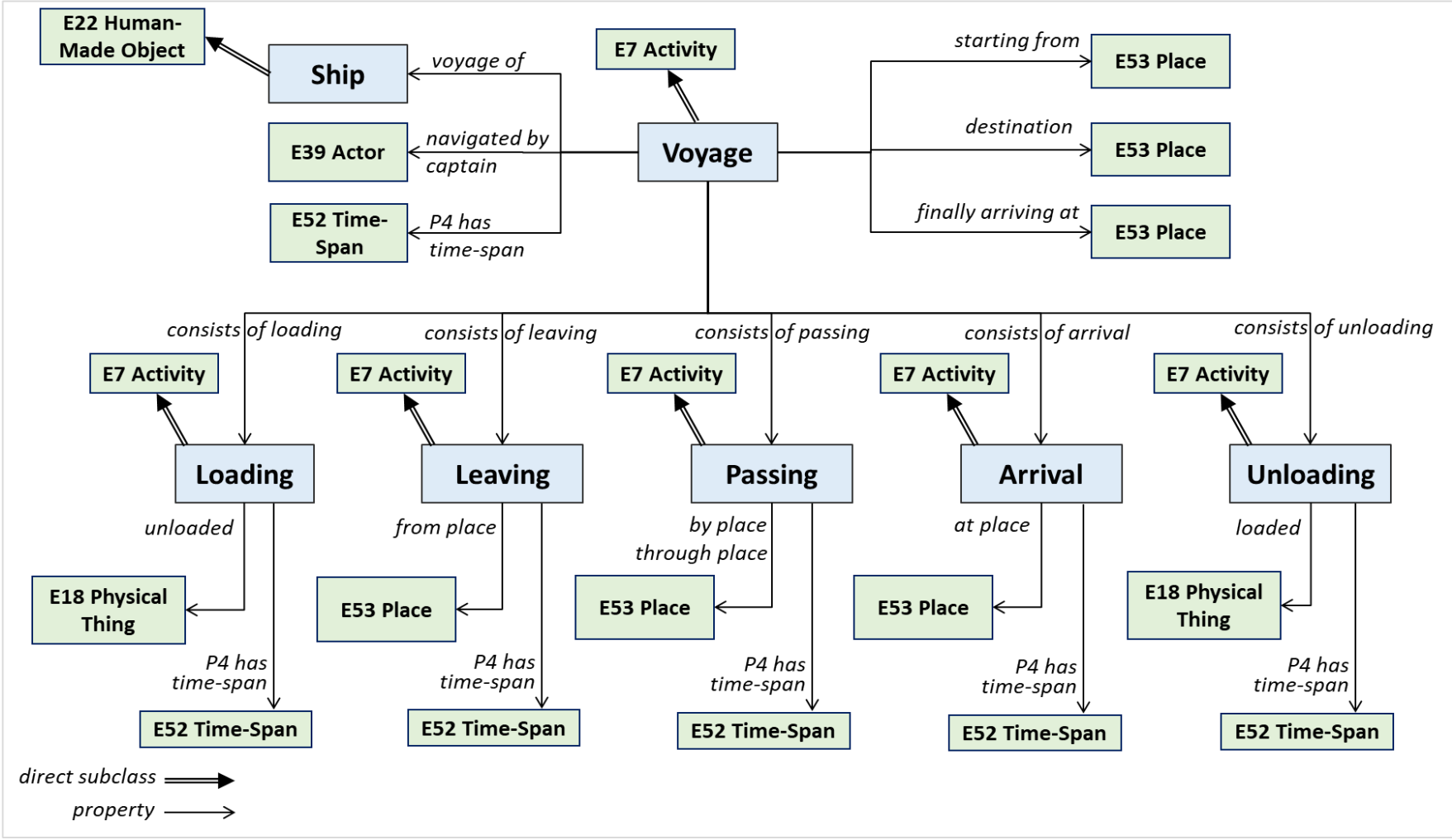
The SeaLiT Ontology

- It offers a **common** and **formal** '**language**' for describing information about Maritime History in a **structured** way
- Created as an extension of **CIDOC-CRM** (Conceptual Reference Model)
  - A high-level ontology (**ISO standard 21127:2014**) of human activity, things and events happening in spacetime
  - It provides a **formal structure** (classes and properties) for describing concepts and relationships used in cultural heritage documentation (and beyond)

- Why using CIDOC-CRM?
  - It gives '**meaning**' to the data **(semantic interoperability**, **data sustainability)**
  - It facilitates **data integration** and **re-use**

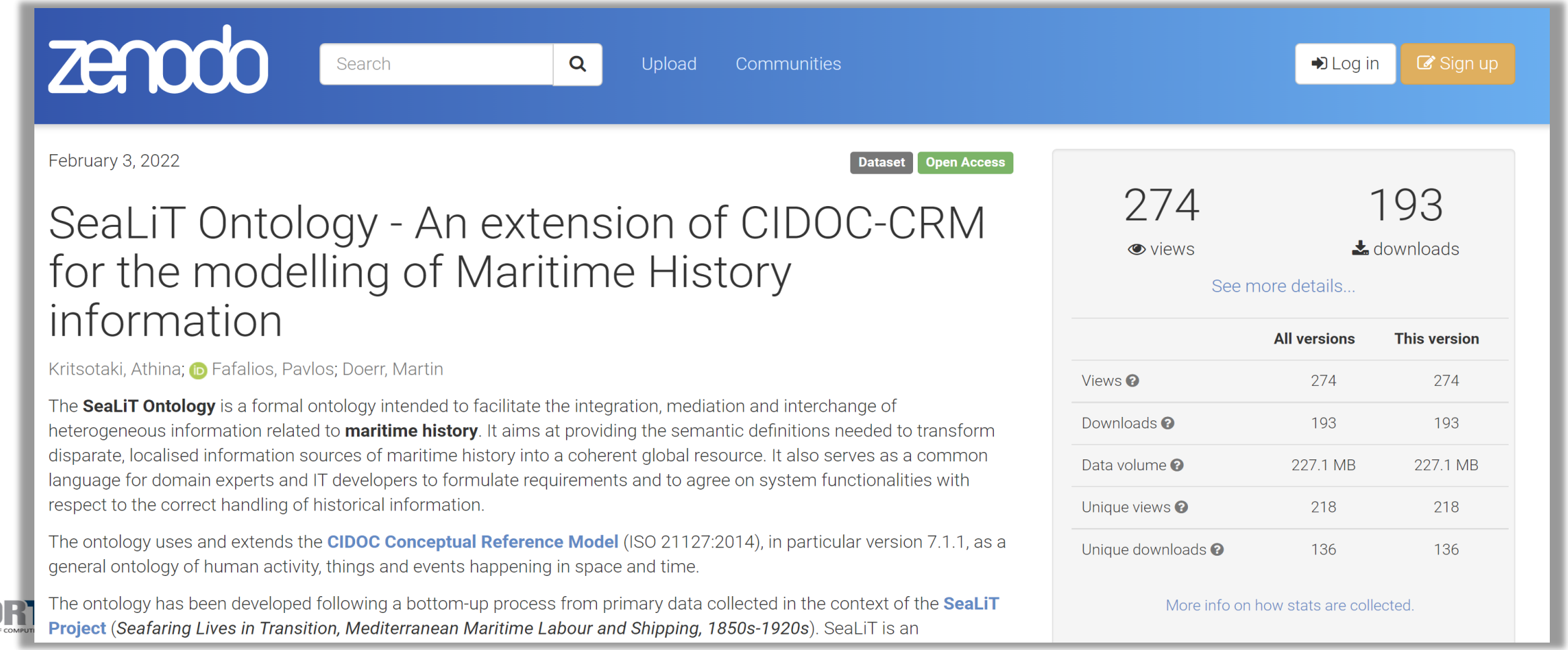
The SeaLiT Ontology

- A (small) part of the ontology



The SeaLiT Ontology

- It currently contains **46 classes** and **79 properties**
  - Available at: **https://zenodo.org/record/5964240**



zenodo

Search

Upload Communities

Log in Sign up

February 3, 2022

Dataset Open Access

SeaLiT Ontology - An extension of CIDOC-CRM for the modelling of Maritime History information

Kritsotaki, Athina; Fafalios, Pavlos; Doerr, Martin

The **SeaLiT Ontology** is a formal ontology intended to facilitate the integration, mediation and interchange of heterogeneous information related to **maritime history**. It aims at providing the semantic definitions needed to transform disparate, localised information sources of maritime history into a coherent global resource. It also serves as a common language for domain experts and IT developers to formulate requirements and to agree on system functionalities with respect to the correct handling of historical information.

The ontology uses and extends the **CIDOC Conceptual Reference Model** (ISO 21127:2014), in particular version 7.1.1, as a general ontology of human activity, things and events happening in space and time.

The ontology has been developed following a bottom-up process from primary data collected in the context of the **SeaLiT Project** (*Seafaring Lives in Transition, Mediterranean Maritime Labour and Shipping, 1850s-1920s*). SeaLiT is an

274 views

193 downloads

See more details...

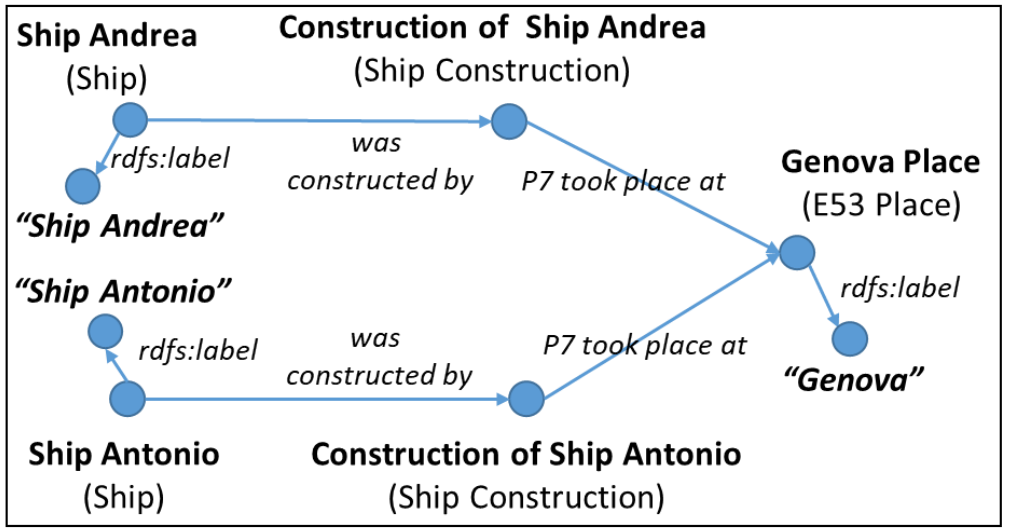
	All versions	This version
Views	274	274
Downloads	193	193
Data volume	227.1 MB	227.1 MB
Unique views	218	218
Unique downloads	136	136

More info on how stats are collected.

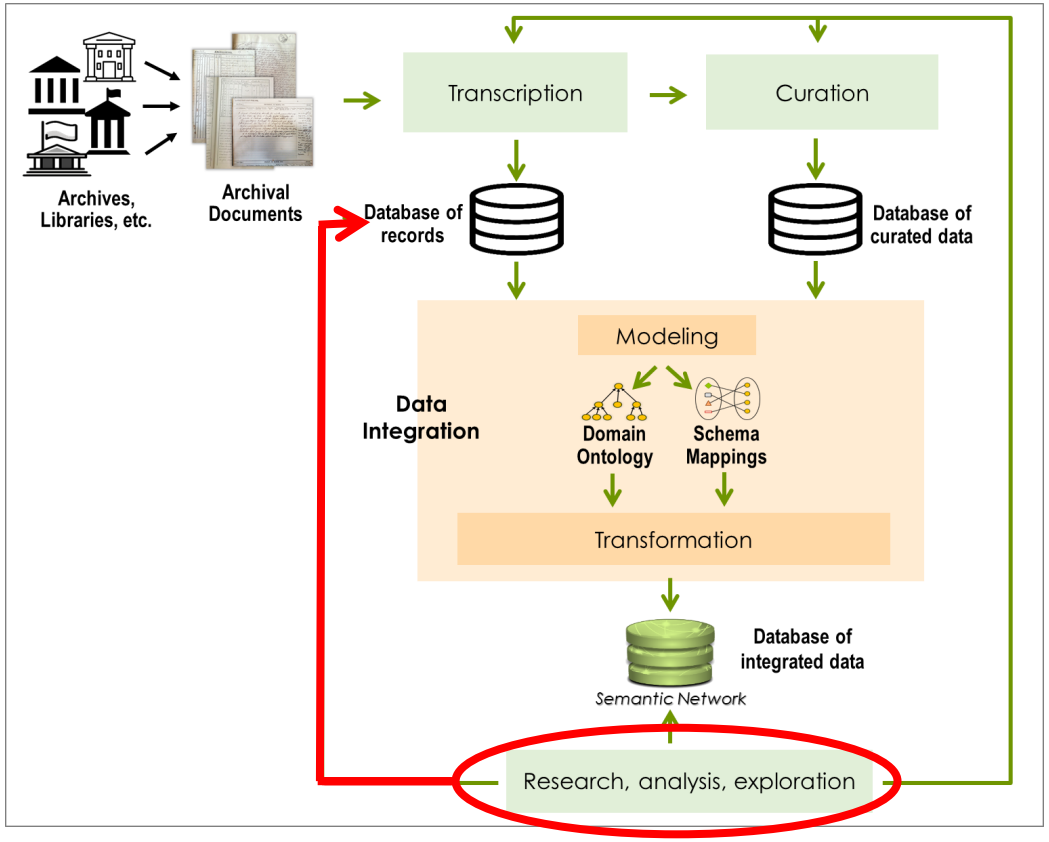
Creation of the Semantic Network

- Data transformation to a rich **semantic network** based on the **SeaLiT Ontology**
  - Using the **X3ML toolkit** (language, 3M user interface, transformation engine)
  - A (very small) part of the derived semantic network:

https://www.ics.forth.gr/isl/x3ml-toolkit



  - The full network contains >**18 million** edges!
  - Available at: **https://zenodo.org/record/6460841**



Exploring the FastCat data with FastCat Catalogues

FastCat Catalogues

- A Web application for **exploring** and **analyzing (quantitively)** the data in all FastCat records
  - **Configurable** for use over other types of sources!

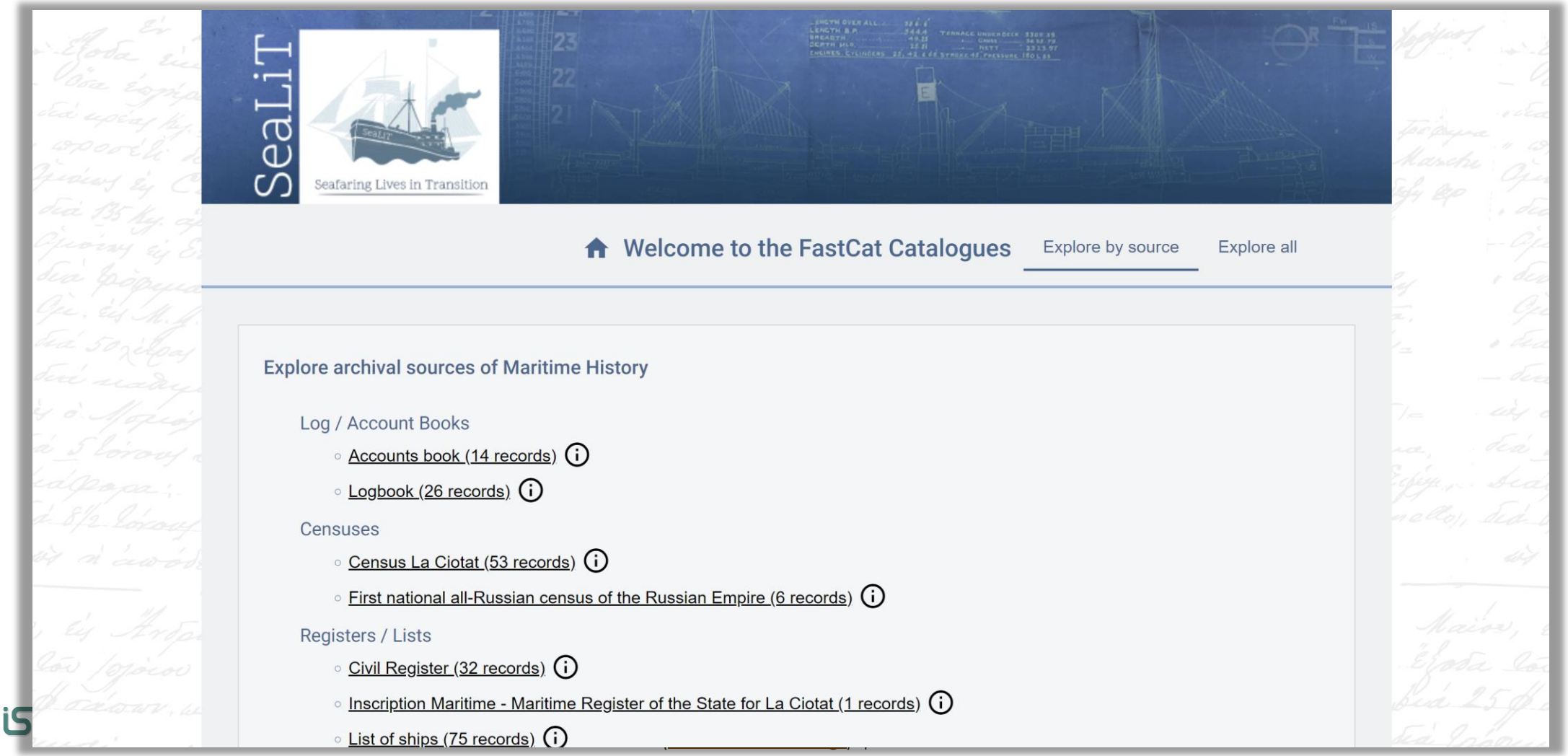
- Functionality:
  - **browsing** "entities of interest"(source-specific, source-independent)
  - data **ranking** and **filtering**
  - inspection of entity **provenance** information
  - data **aggregation** and **visualization** in **charts**
  - data **export** for further (external) analysis

https://catalogues.sealitproject.eu/

Available as open source:
https://github.com/isl/FastCat-Catalogues

FastCat Catalogues

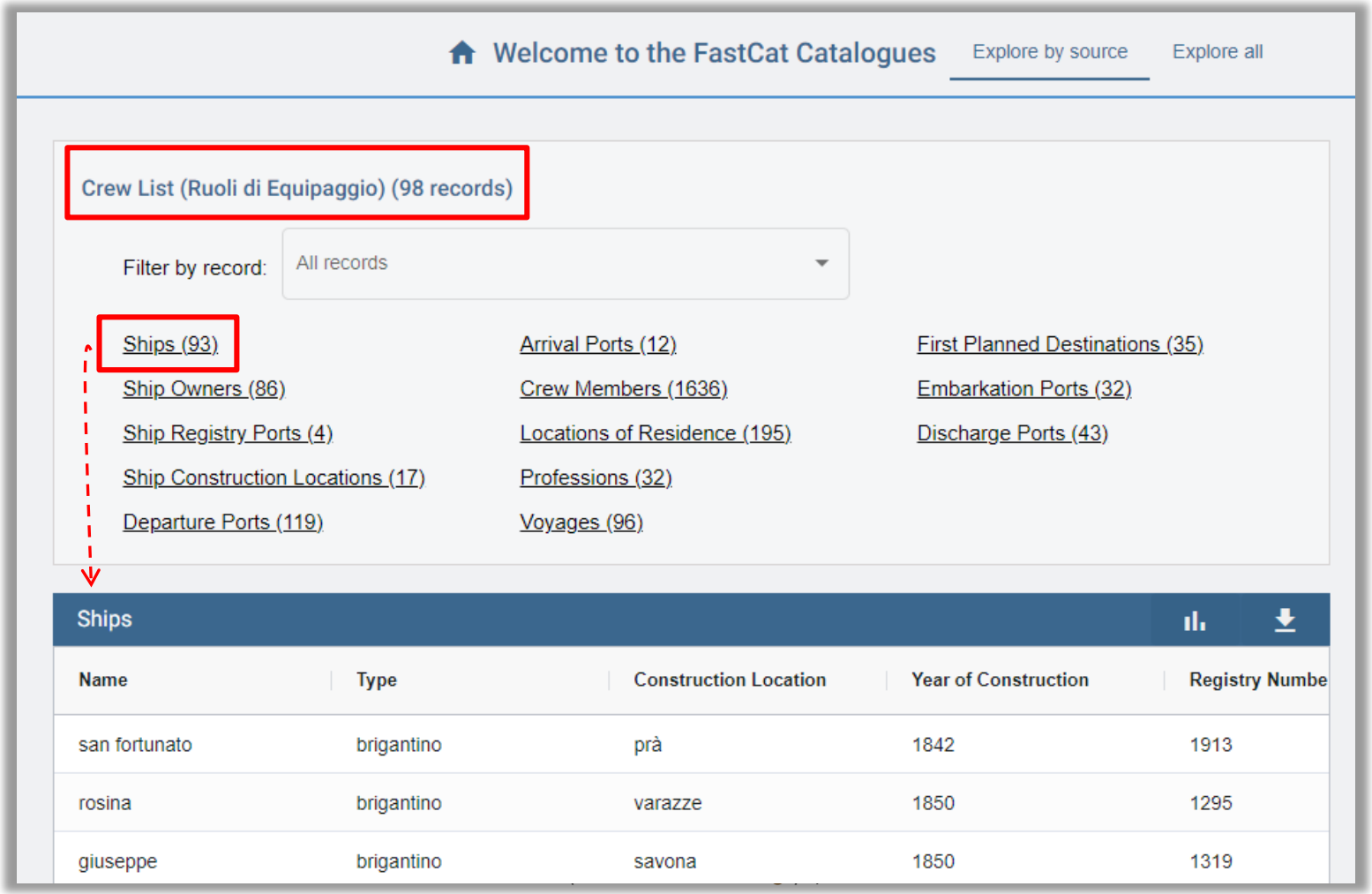
https://catalogues.sealitproject.eu/



FastCat Catalogues

https://catalogues.sealitproject.eu/

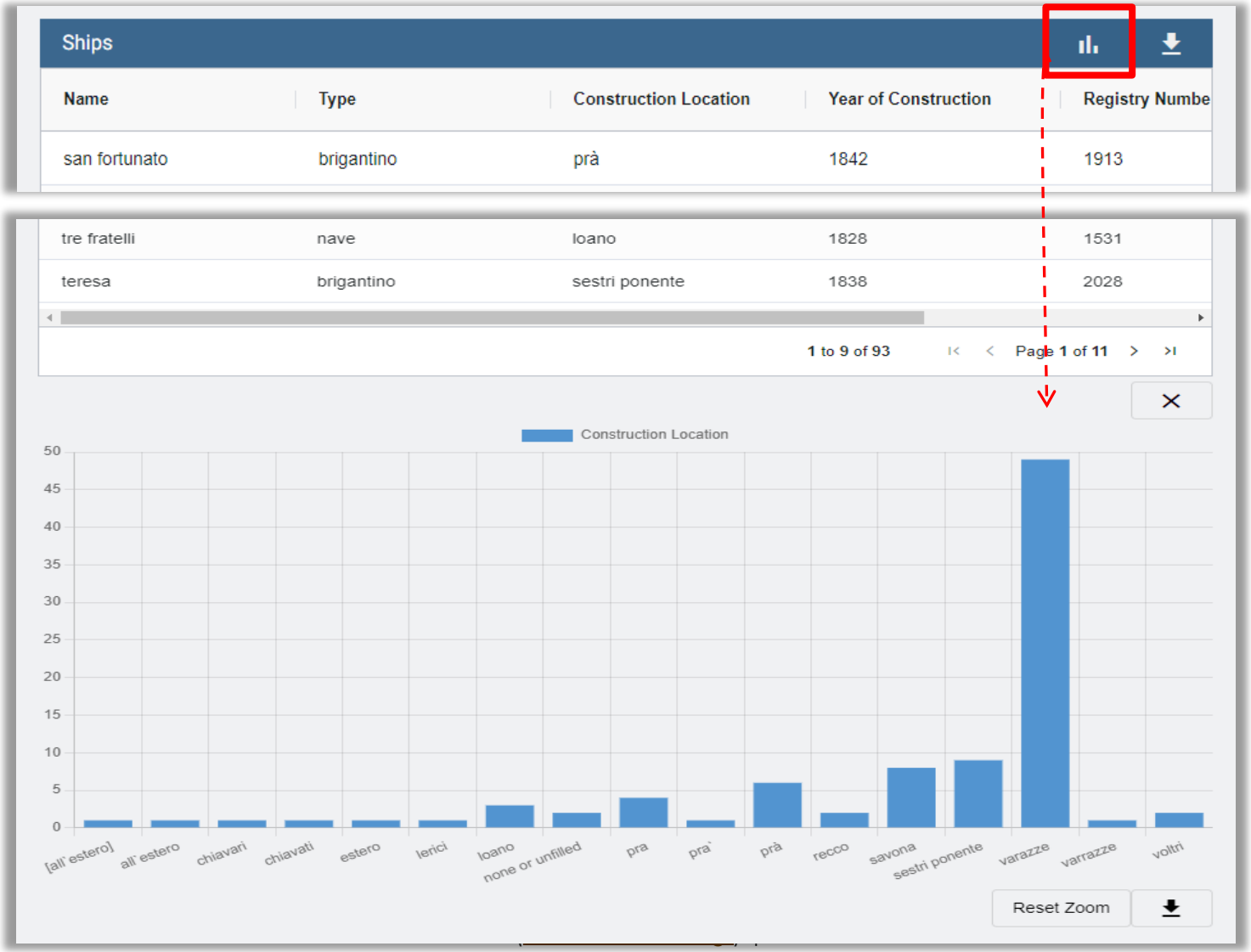
Data browsing and exploration

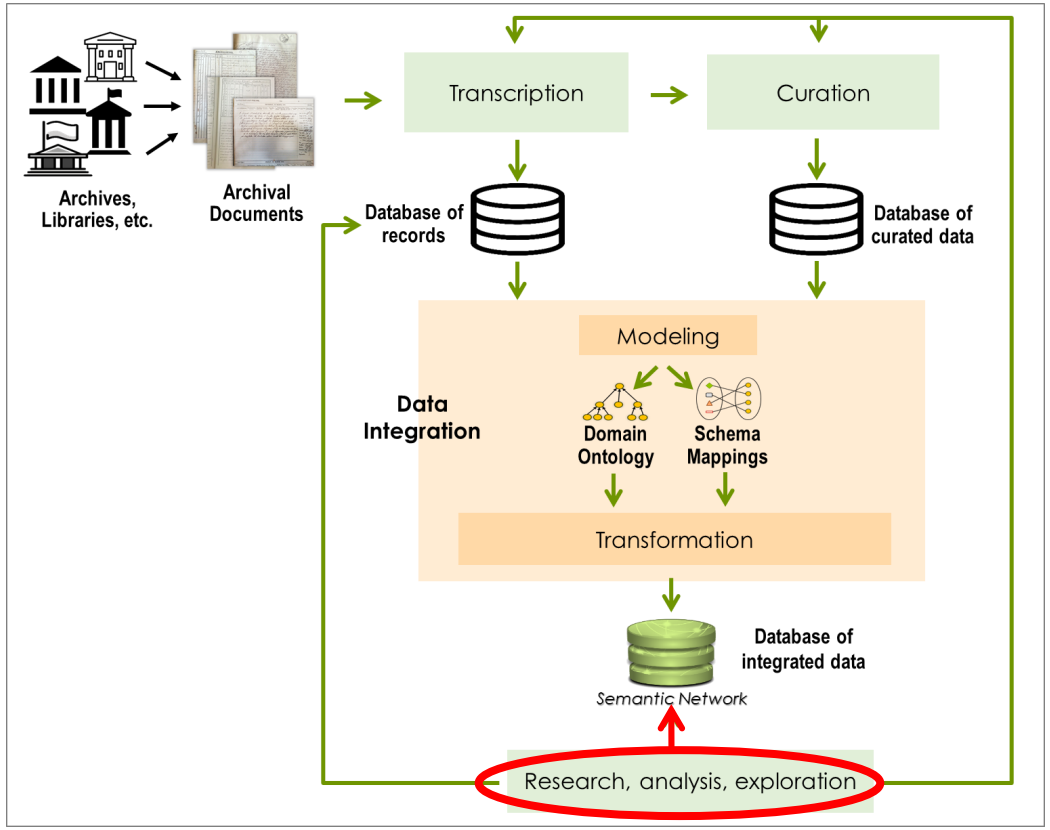


FastCat Catalogues

https://catalogues.sealitproject.eu/

Data visualization





Exploring the integrated data with ResearchSpace

SeaLiT ResearchSpace

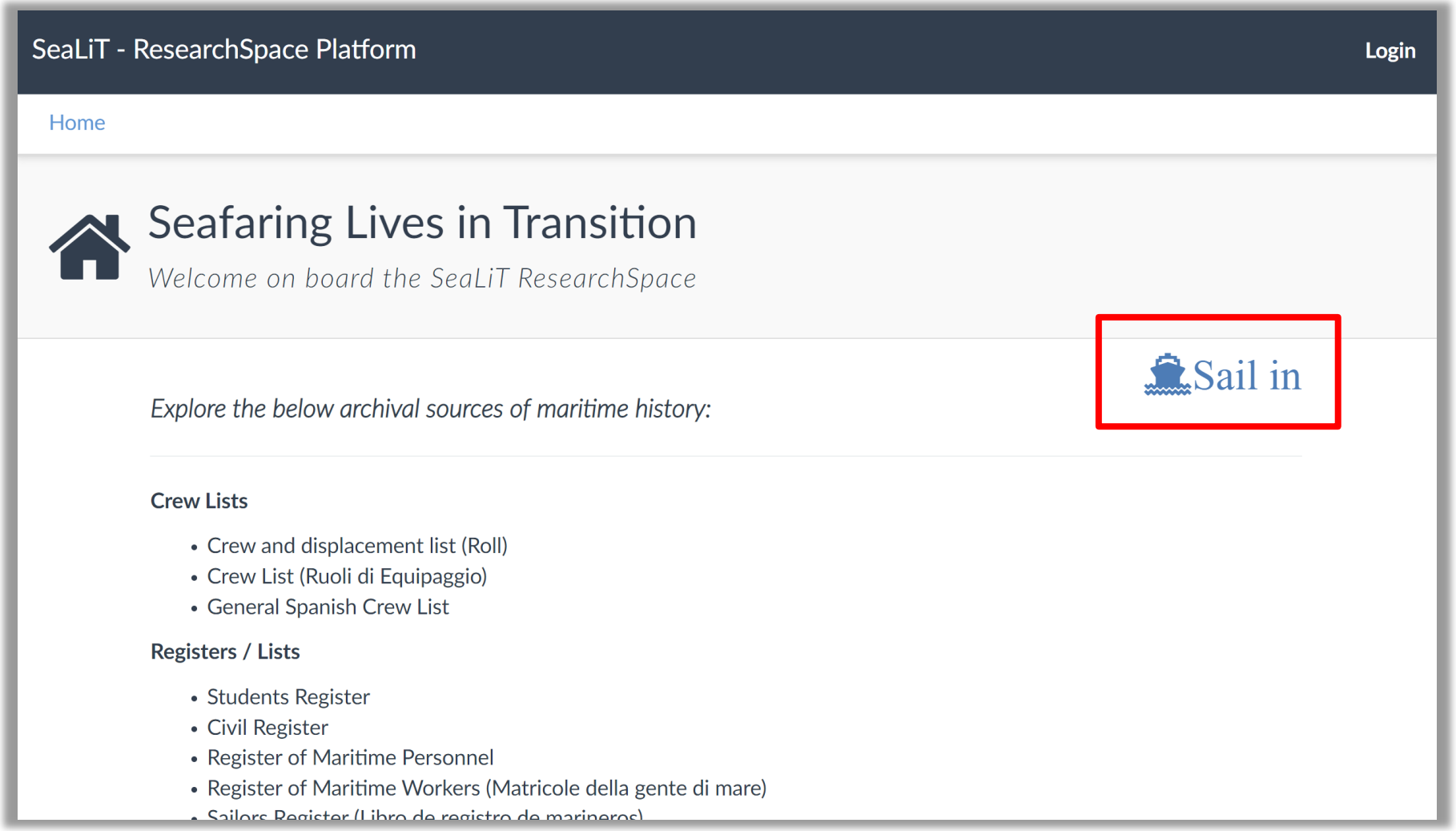
- **ResearchSpace**: open source platform, developed by the British Museum
- Offers a variety of functionalities:
  - Semantic search (assistive query building)
  - Data browsing
  - Variety of visualization methods (tables, charts, map)
  - Filtering (based on entity properties/relations)
- **Configured** for the case of the SeaLiT data

http://rs.sealitproject.eu/

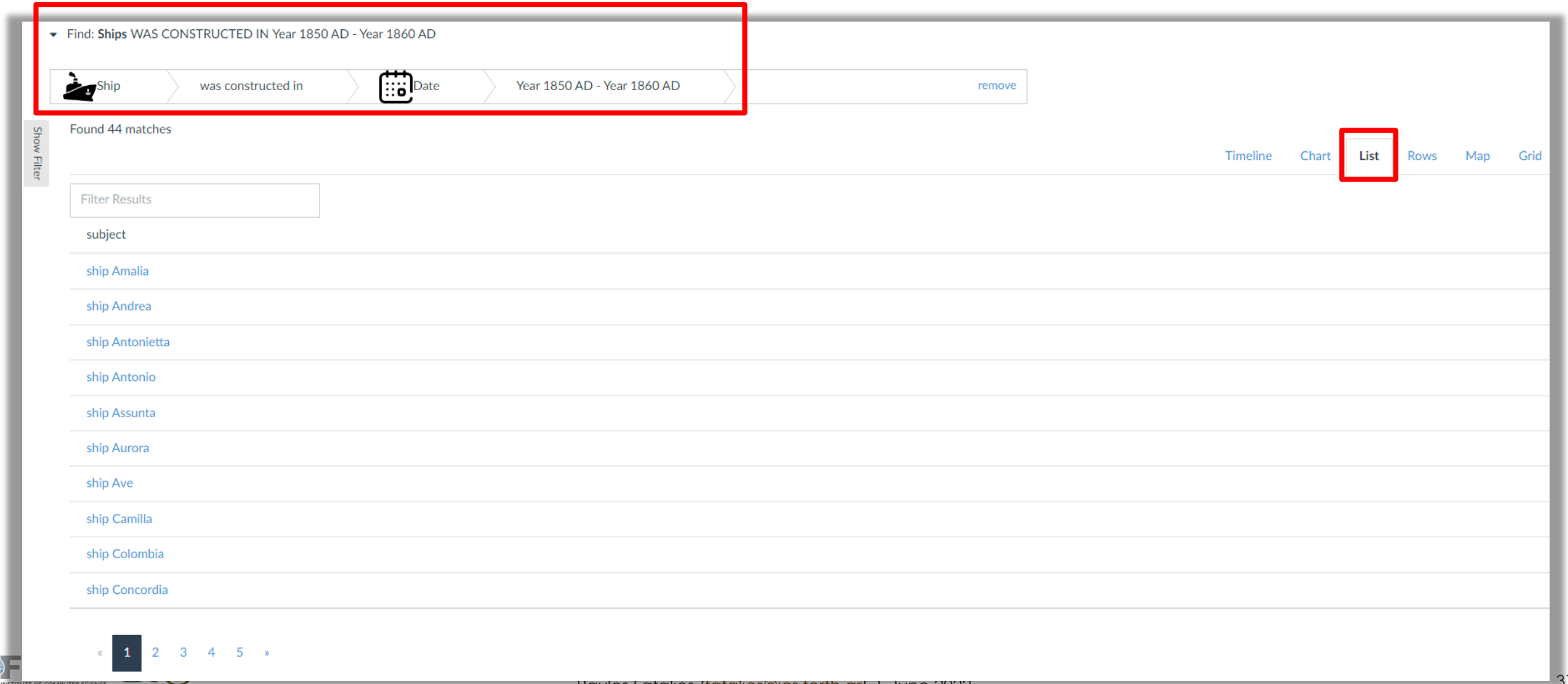
ResearchSpace
https://www.researchspace.org/

SeaLiT ResearchSpace

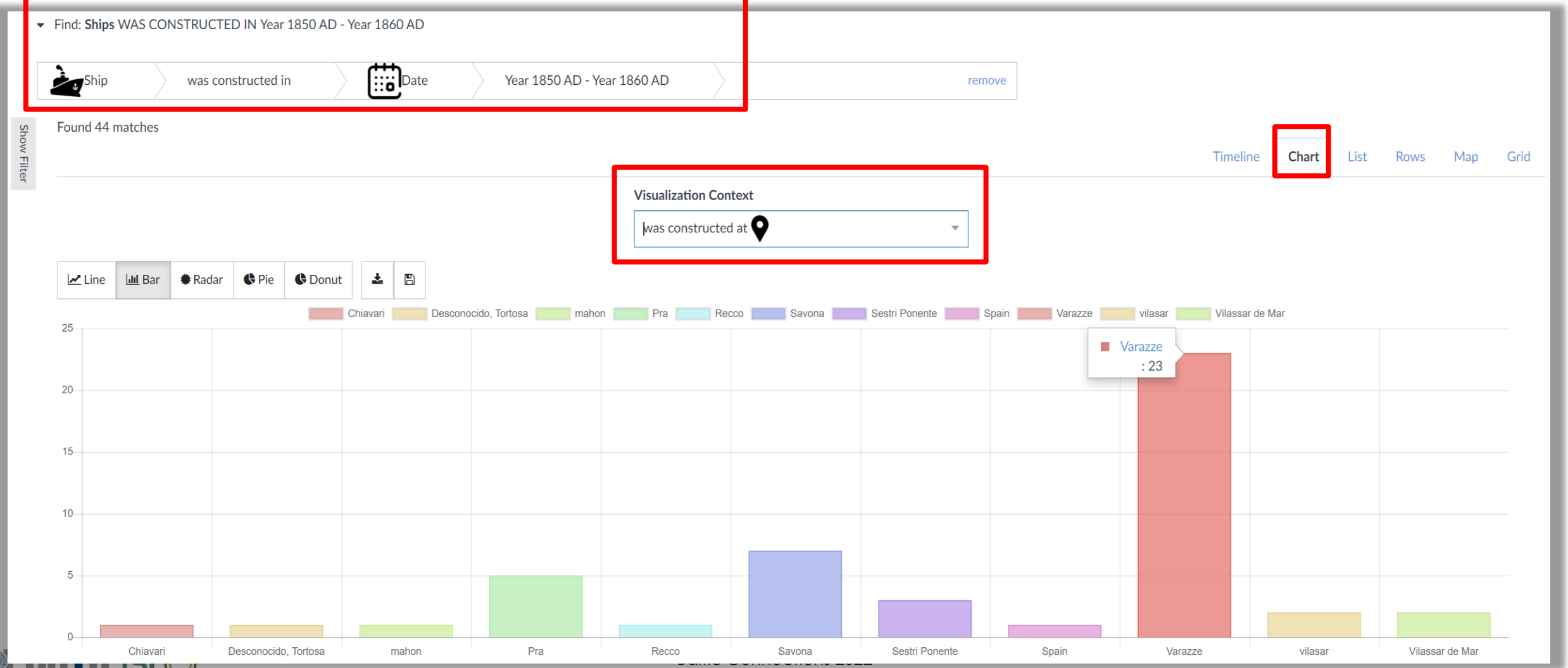
http://rs.sealitproject.eu/

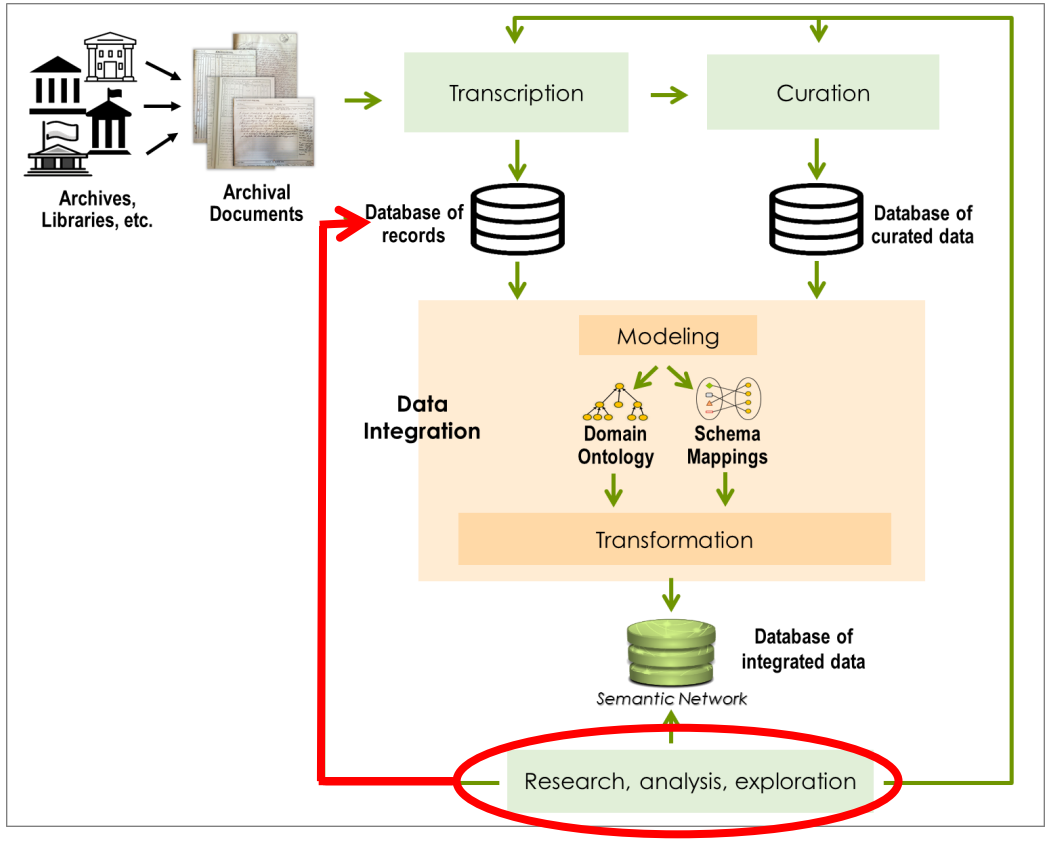


Assistive query building in ResearchSpace



Assistive query building in ResearchSpace



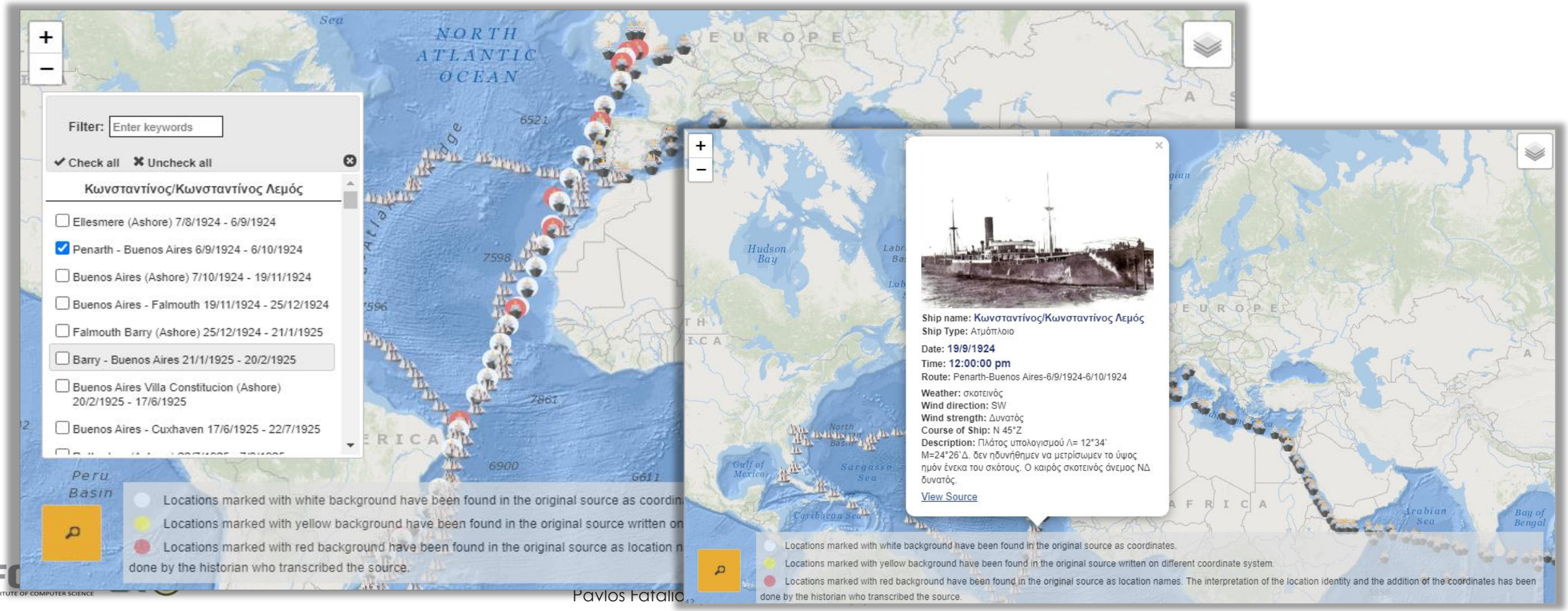


Data visualization in map with Ship Voyages

Map application ‘Ship Voyages’

- Exploiting **logbooks** of 17 merchant ships (1860s-1920s)

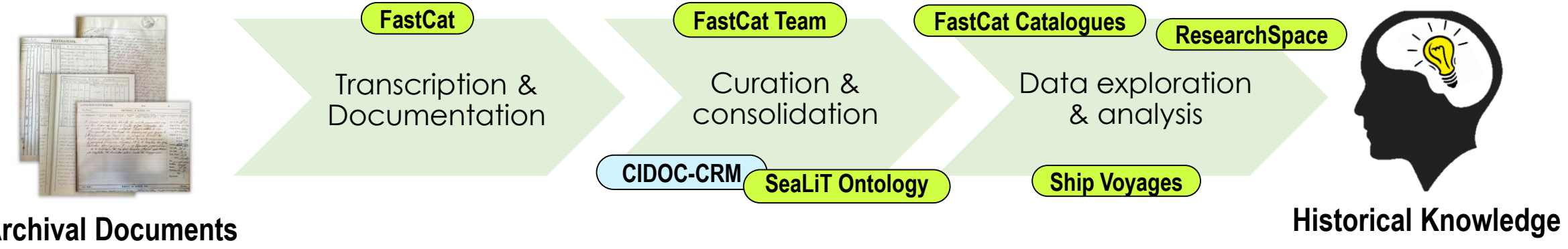
https://sealitproject.eu/digital-seafaring



Conclusion and Future Work

Conclusion

- **Workflow Model** for holistic data management in historical archival research
  - Supported by **IT tools**
  - **Provenance-aware**
  - Following established **documentation standards** (CIDOC-CRM)



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

Limitations / Future Work

- **Data entry errors** are common [FastCat]
  - Accidentally filling the wrong column, putting data in wrong place due to misunderstanding, …
  - **We need mechanisms that support users in detecting or avoiding such errors!**
- **Manual instance mapping** and **vocabulary curation** [FastCat Team]
  - Very **important**, but at the same time very **laborious** and **time consuming**
  - **We need tools that can automate these processes as much as possible, without significantly affecting quality!**
- **Comparing values** (**dates**, **dimensions**, **quantities, etc.**) is sometimes difficult/impossible [FastCat / FastCat Team]
  - Due to use of different reference points or units across sources
  - **We need mechanisms to align such values!**

Try the applications:
https://sealitproject.eu/digital-seafaring

Thank you!

Pavlos Fafalios
fafalios@ics.forth.gr

The research and development team:

- Anastasia Axaridou (R&D engineer)
- Korina Doerr (Systems design engineer)
- Athina Kritsotaki (Data modeling engineer)
- Yannis Marketaks (R&D engineer)
- Kostas Petrakis (R&D engineer)
- Georgios Samaritakis (R&D engineer)
- Maria Theodoridou (R&D engineer)
- Pavlos Fafalios (Postdoctoral researcher)
- Martin Doerr (CCI Honorary Head)

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

ReKnow

European Research Council

Seafaring Lives in Transition