

Advanced IT Tools for Historical Research with Archival Material

Pavlos Fafalios

Centre for Cultural Informatics (CCI), Institute of Computer Science (ICS), FORTH
fafalios@ics.forth.gr



European
Research
Council



Seafaring Lives in Transition

Outline

- Introduction
- The tools
 - **FastCat** (data transcription)
 - **FastCat Catalogues** (data browsing)
 - **FastCat Team** (data curation)
 - **SeaLiT Ontology** (data integration)
 - **SeaLiT ResearchSpace** (data exploration & analysis)
- Conclusion

Introduction

Data Management in SeaLiT

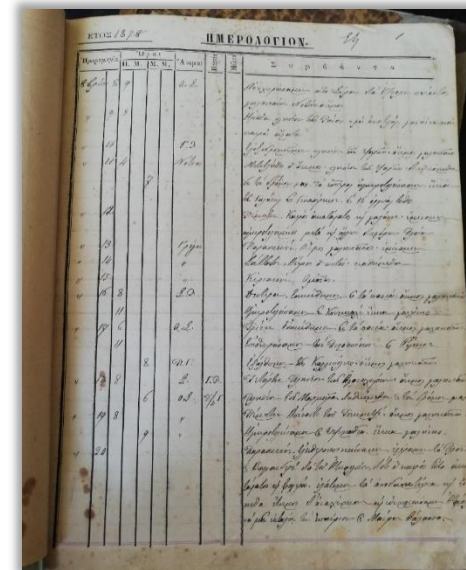
❑ The archival sources

- Crew and displacement lists
- Ship logbooks
- Payrolls
- Sailors registers
- Naval Ship registers
- Students registers
- Employment records
- Account books
- Censuses
- ...

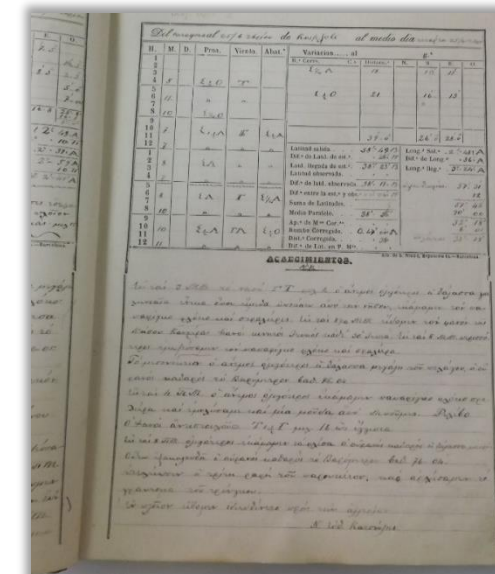
- Provided by **different authorities** in several countries (**5 languages**)
- Containing **interconnected information** about different types of **entities** (sailors, ships, locations, ...)

Description of the 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**

➤ **Problem 1:** Complementary information about the same ship may exist in different archival documents



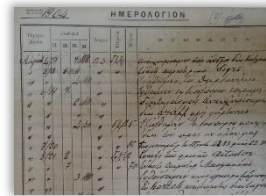
Account book:
Information about
the **ship's owners**



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

*Need of integrating
information coming
from different archival
documents!*

Data Management Challenges

- **Motivating Scenario:** Exploring information about a particular **ship**
 - **Problem 2:** The **same ship**, or the same ship-related **entity/term**, might appear under different representations in the archival documents, due to:
 - ❖ Typos (“Vindob**nn**a” vs “Vindob**o**na”, “Gaetano Schia**ff**ino” vs “Gaetano Schia**f**ino”)
 - ❖ Different languages (“Genoa” vs “Geno**v**a”, “brigantine” vs “brigant**i**no”)
 - ❖ Unrecognizable characters (“G**??**oa”)
 - ❖ Uncertainty (“**[**Genoa**]**”)
 - ❖ Use of abbreviation (“Gaetano Schia**ff**ino” vs “**G.** Schia**ff**ino”)

Need of curating the (transcribed) data:

- entity (instance) matching
- term alignment
- corrections

Data Management Challenges

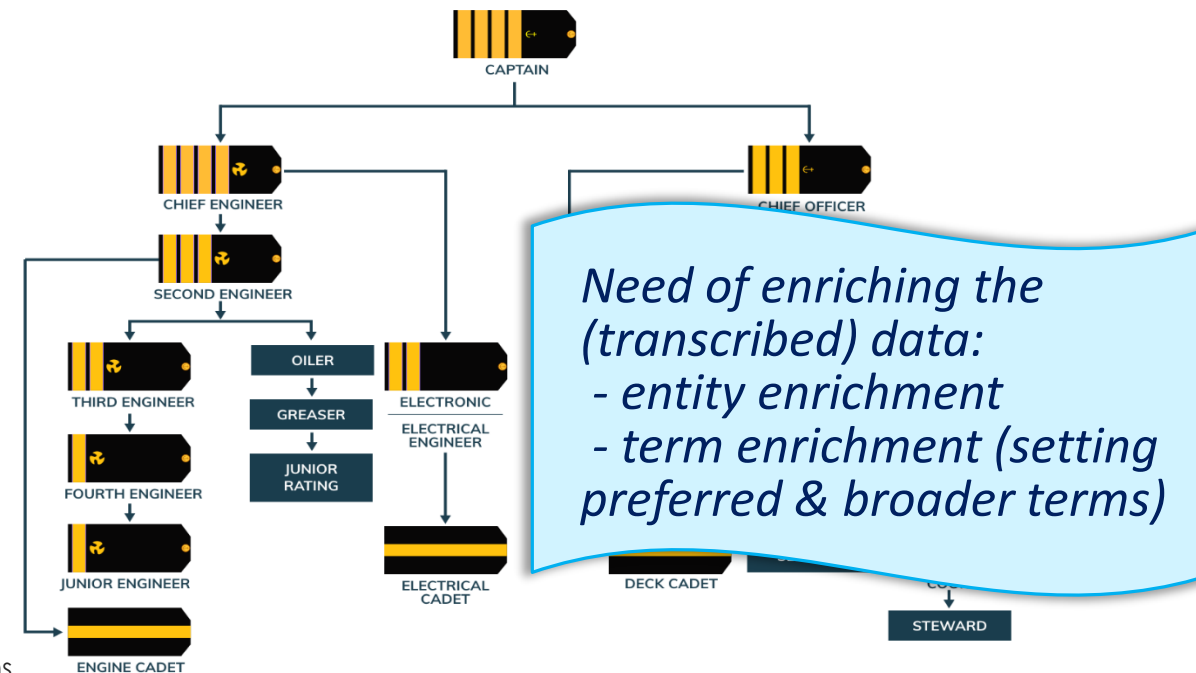
□ **Motivating Scenario:** Exploring information about a particular **ship**

➤ **Problem 3:** Data **enrichment** is needed for better analysis and exploration services

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



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



Need of enriching the (transcribed) data:

- entity enrichment
- term enrichment (setting preferred & broader terms)

Data Management Challenges

- **Motivating Scenario:** Exploring information about a particular **ship**
 - **Problem 4:** How to explore the (curated & integrated) data, both quantitatively and qualitatively, through user-friendly interfaces/visualizations?

- ❖ Browsing the interconnected data

*Starting from a specific **ship**, check its **arrival ports**, then find information about the **sailors** or other **ships** that arrived at that port ...*

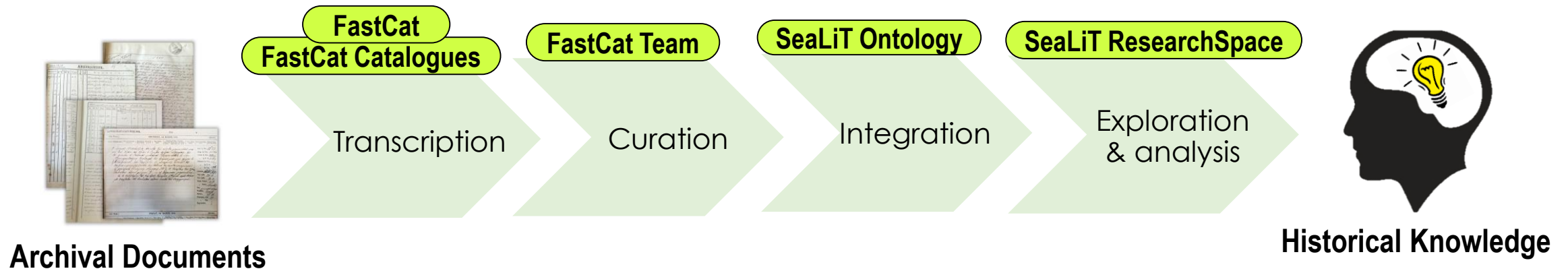
- ❖ Running complex questions / aggregating information

*“Find the **nationality** of **sailors** of **Greek ships** of a specific **type** (e.g., **Brig**) that arrived at a specific **port**”*

Need of tools/interfaces for supporting researchers in exploring and analyzing the (transcribed & curated) data

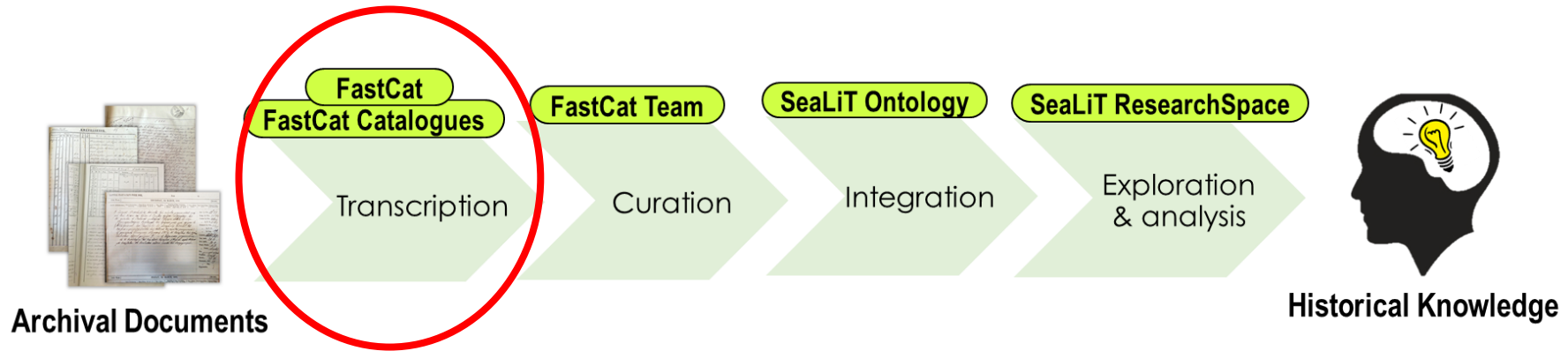
Our contribution

□ Workflow model and IT tools



Focus on maintaining the important **data provenance** information!

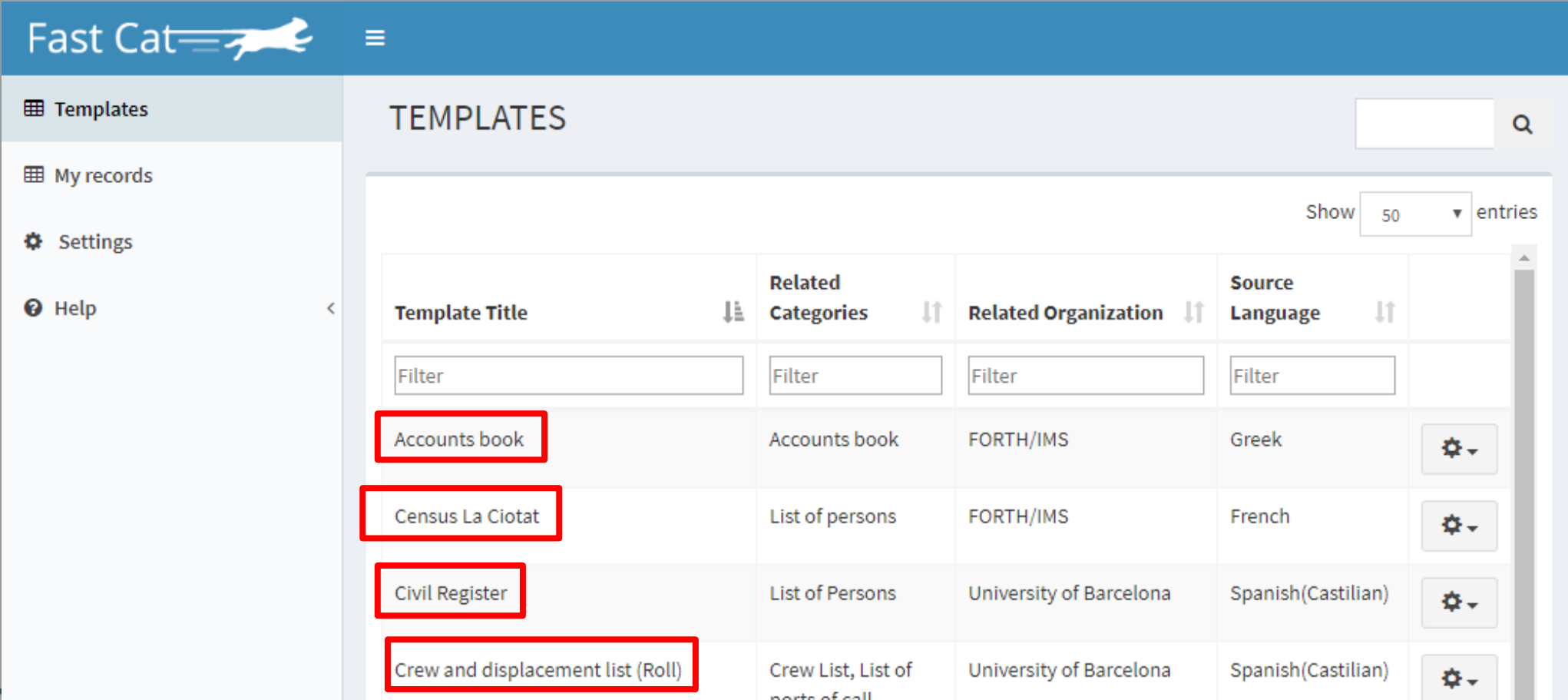
- Important for **verification** and the **long-term validity** of the research results



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



The screenshot displays the FastCat web application interface. The top navigation bar includes the 'Fast Cat' logo and a search icon. A left sidebar contains menu items: 'Templates', 'My records', 'Settings', and 'Help'. The main content area is titled 'TEMPLATES' and features a search bar and a 'Show 50 entries' dropdown. Below this is a table with columns: 'Template Title', 'Related Categories', 'Related Organization', and 'Source Language'. Each row includes a gear icon for settings. Four rows are highlighted with red boxes: 'Accounts book', 'Census La Ciotat', 'Civil Register', and 'Crew and displacement list (Roll)'.

Template Title	Related Categories	Related Organization	Source Language
Filter	Filter	Filter	Filter
Accounts book	Accounts book	FORTH/IMS	Greek
Census La Ciotat	List of persons	FORTH/IMS	French
Civil Register	List of Persons	University of Barcelona	Spanish(Castilian)
Crew and displacement list (Roll)	Crew List, List of ports of call	University of Barcelona	Spanish(Castilian)

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)

The screenshot displays the FastCat interface for a specific record. The title bar reads "Crew List (Ruoli di Equipaggio), Nostra Signora di Montenero, 1853-08-31, Massimo Ponasso".

FastCat Record Information (highlighted in red):

Use english to fill in the fields of this table

Id	Date		Authors		
	Creation date	Last Modified	Name *	Surname *	Role
16	2018-12-18	2019-09-05T14:39:47	Massimo	Ponasso	

Source Identity (highlighted in red):

Use the source language to fill in the fields of this table

Archive / Library				Document
Name	Location	Register Number	Number	Date of
Archivio di Stato di Genova	Genova	13	4147	1853-08-31

Ship Identity (highlighted in red):

Use the source language to fill in the fields of this table

Ship name *	Ship type	Tonnage	Construction	
			Location	Date (year)
Nostra Signora di Montenero	Bombarda	88,76	Varazze	1831

Crew List (highlighted in red):

Use the source language to fill in the fields of this table

	Embarkation		Discharge		Surname	Name	Citizenship	Location of Residence	Date of Birth	Serial Number
	Port	Date	Port	Date						
2	Genova	1853-07-06	Genova	1854-01-20	Castagniola	Fortunato		Camogli	1811	3740
3	Genova	1853-07-06	Genova	1854-01-20	Schiaffino	Fortunato		Camogli	1807	1109
4	Genova	1853-07-06	Genova	1853-10-14	Figari	Fortunato		Camogli	1811	355
5	Genova	1853-07-06	Genova	1853-10-14	Cimidlia	Antonio		Sanremo	1833	1043

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

Exploring the FastCat records with FastCat Catalogues

□ FastCat Catalogues

- A Web application for **browsing** and **analyzing (quantitatively)** 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/>

SeaLiT
Seafaring Lives in Transition

↑ Welcome to the FastCat Catalogues Explore by source Explore all

Explore archival sources of Maritime History

- Log / Account Books
 - Accounts book (14 records) ⓘ
 - Logbook (26 records) ⓘ
- Censuses
 - Census La Ciotat (53 records) ⓘ
 - First national all-Russian census of the Russian Empire (6 records) ⓘ
- Registers / Lists
 - Civil Register (32 records) ⓘ
 - Inscription Maritime - Maritime Register of the State for La Ciotat (1 records) ⓘ
 - List of ships (75 records) ⓘ

FastCat Catalogues

<https://catalogues.sealitproject.eu/>

Data
browsing
and
exploration

Welcome to the FastCat Catalogues [Explore by source](#) [Explore all](#)

Crew List (Ruoli di Equipaggio) (98 records)

Filter by record: All records

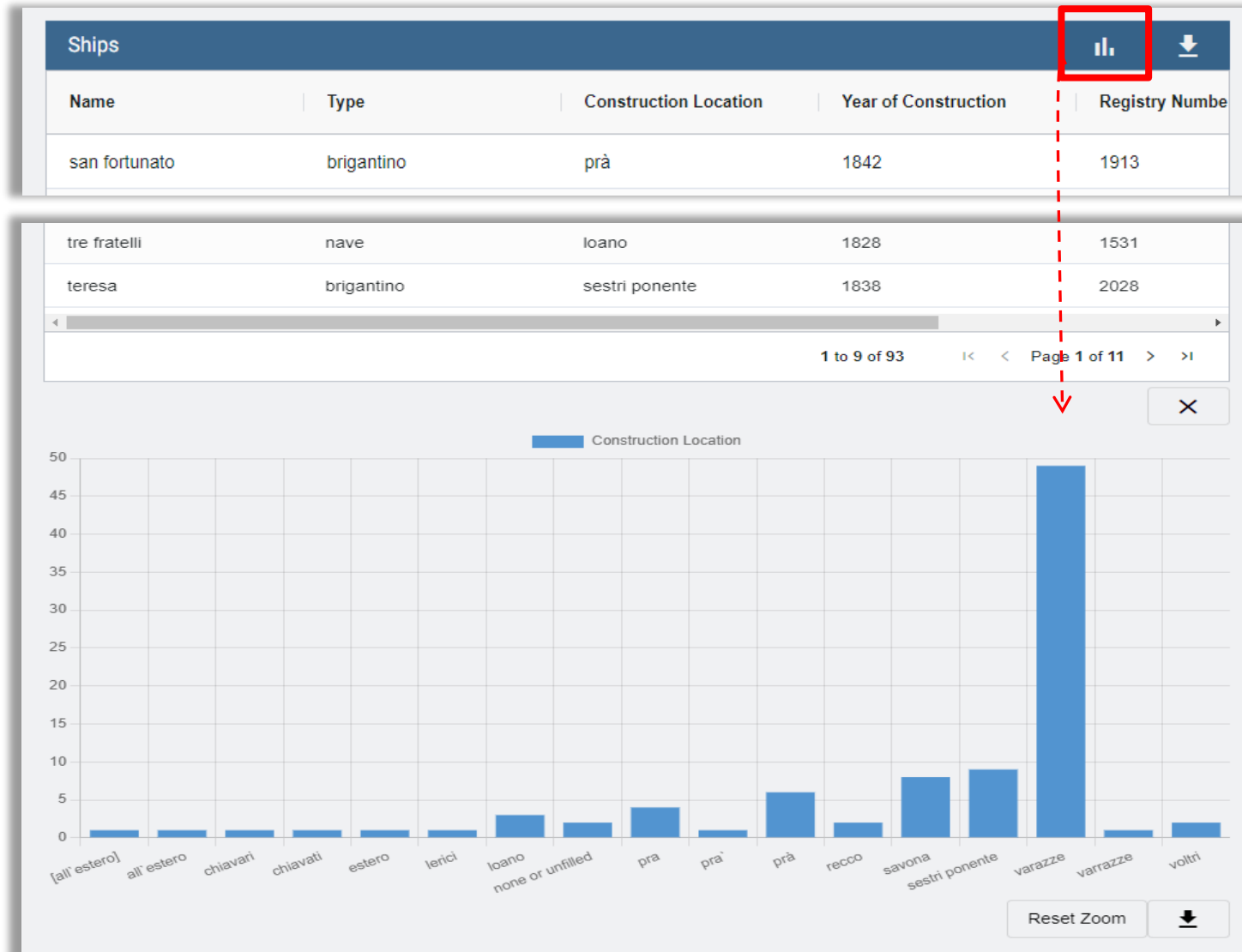
- Ships (93)**
- Ship Owners (86)
- Ship Registry Ports (4)
- Ship Construction Locations (17)
- Departure Ports (119)
- Arrival Ports (12)
- Crew Members (1636)
- Locations of Residence (195)
- Professions (32)
- Voyages (96)
- First Planned Destinations (35)
- Embarkation Ports (32)
- Discharge Ports (43)

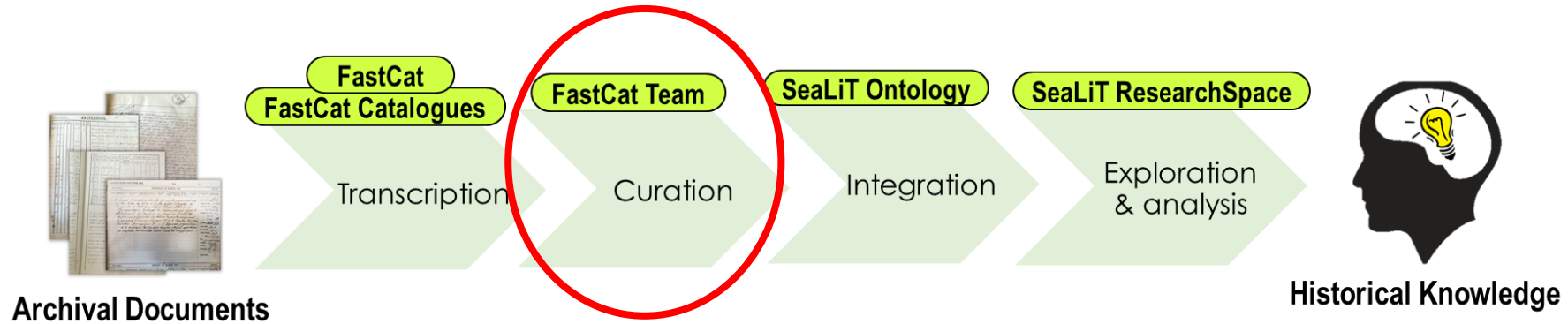
Ships				
Name	Type	Construction Location	Year of Construction	Registry Number
san fortunato	brigantino	prà	1842	1913
rosina	brigantino	varazze	1850	1295
giuseppe	brigantino	savona	1850	1319

FastCat Catalogues

<https://catalogues.sealitproject.eu/>

Data
visualization



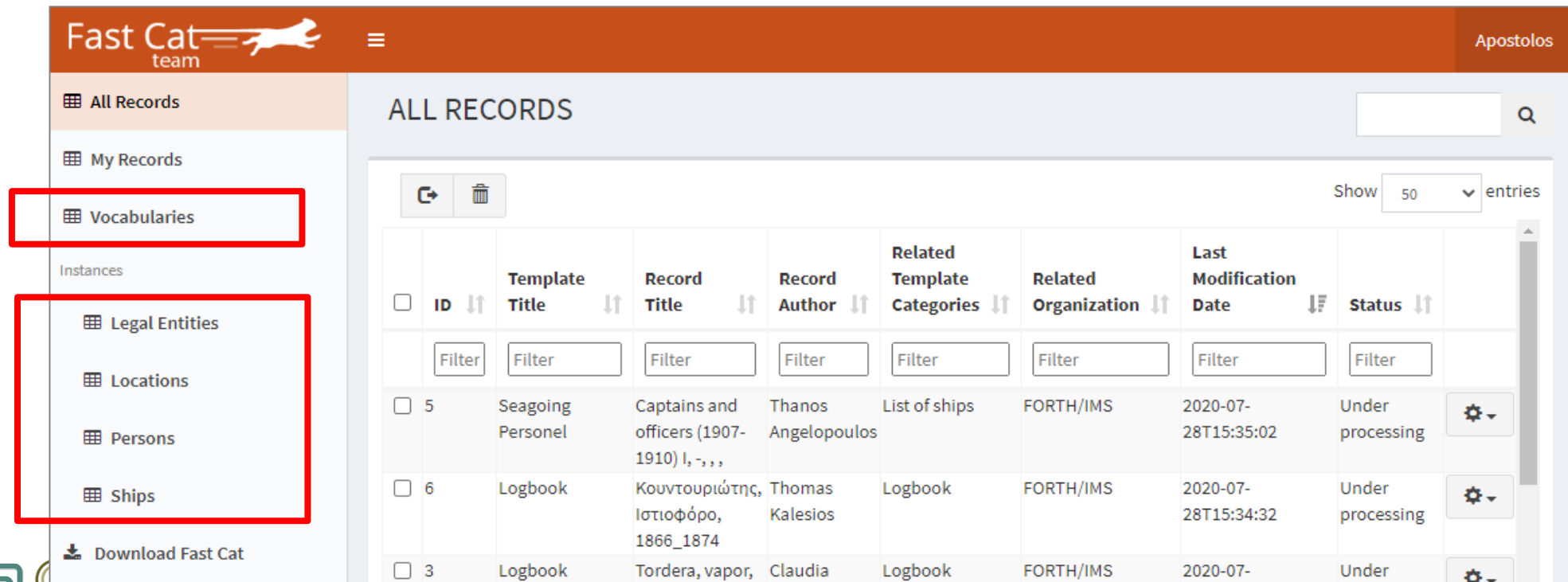


Data curation with FastCat Team

Data curation with FastCat Team

□ FastCat Team:

- A special environment within FastCat 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!



The screenshot displays the FastCat Team web interface. The top navigation bar includes the 'Fast Cat team' logo and the user name 'Apostolos'. The left sidebar contains a menu with 'All Records', 'My Records', 'Vocabularies', and 'Instances'. Under 'Instances', there are sub-items for 'Legal Entities', 'Locations', 'Persons', and 'Ships'. The main content area is titled 'ALL RECORDS' and features a search bar, a 'Show 50 entries' dropdown, and a table of records. The table has columns for ID, Template Title, Record Title, Record Author, Related Template Categories, Related Organization, Last Modification Date, and Status. Three records are visible, each with a checkbox, filter buttons, and a settings icon.

ID	Template Title	Record Title	Record Author	Related Template Categories	Related Organization	Last Modification Date	Status
5	Seagoing Personal	Captains and officers (1907-1910) I, -, , ,	Thanos Angelopoulos	List of ships	FORTH/IMS	2020-07-28T15:35:02	Under processing
6	Logbook	Κουντουριώτης, Ιστιοφόρο, 1866_1874	Thomas Kalesios	Logbook	FORTH/IMS	2020-07-28T15:34:32	Under processing
3	Logbook	Tordera, vapor,	Claudia	Logbook	FORTH/IMS	2020-07-	Under

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**

Fast Cat team fafalios

☰ All Records
☰ My Records
☰ Vocabularies

Instances

- ☰ Legal Entities
- ☰ Locations
- ☰ **Persons**
- ☰ Ships

Download Fast Cat
Settings

PERSONS ?

Show instances used in template type: Register of Maritime wo and record: All Show 50 entries

Mark as same ResetTable Export to Excel

	Name	Surname A	Surname B	Maiden name	Fathers name	Place of Birth	Date of Birth	Date of Death	Registration number	Status Capacity Role	Status
<input type="checkbox"/>	Agostino Michele	Simonetti			Andrea		1833-12-23		9558		Under processing
<input type="checkbox"/>	Agostino Rocco	Massa			Antonio		1869-08-16		22696		Under processing
<input type="checkbox"/>	allegra	capurro									Under processing
<input type="checkbox"/>	Amedeo	Casabona			Bartolomeo		1867-02-25		22755		Under processing
<input type="checkbox"/>	andrea	bozzo									Under

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

Fast Cat team fafalios

☰

All Records

My Records

Vocabularies

Instances

Legal Entities

Locations

Persons

Ships

Download Fast Cat

Settings

VOCABULARIES ?

Show 50 entries

Vocabulary Title	Source Language	Used in template	Related Organization
Filter	Filter	Filter	Filter
Wind strength	~ 392 Greek	Logbook	FORTH/IMS
Wind direction	~ 255 Greek	Logbook	FORTH/IMS
Weather type	~ 831 Greek	Logbook	FORTH/IMS
Unit	~ 327 Greek Spanish(Castilian) Italian Russian French	Accounts book Logbook Payroll Crew and displacement list (Roll) Register of Maritime personel Naval Ship Register List	FORTH/IMS University of Barcelona NAVLAB-Università di Genova T.I.G./University of Barcelona

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

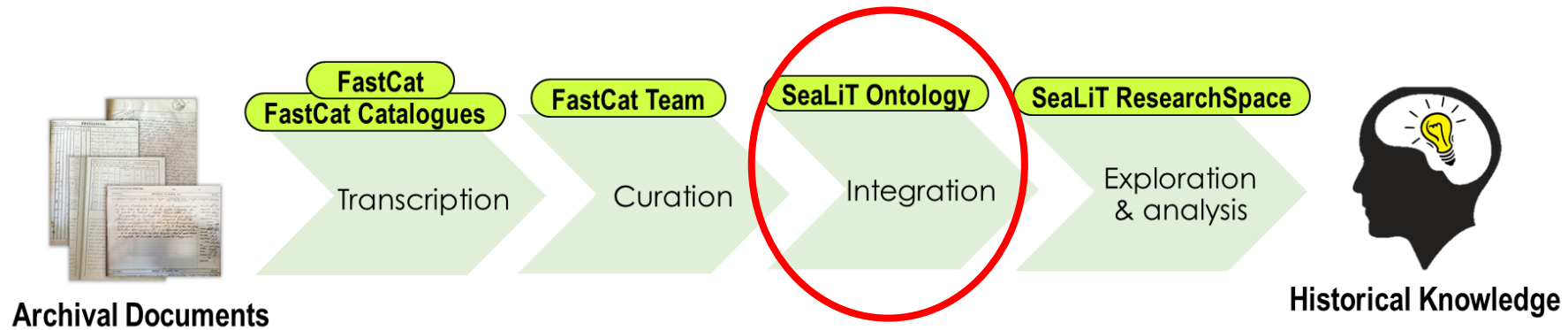
FastCat & FastCat Team

□ Available as open-source:

- <https://github.com/isl/FastCat>

□ Papers:

- 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>
- K. Petrakis et al. (2021), “Digitizing, Curating and Visualizing Archival Sources of Maritime History: the case of ship logbooks of the nineteenth and twentieth centuries”. *Drassana, No. 28*, 60–87.
<https://doi.org/10.51829/Drassana.28.649>



Data Integration with SeaLiT Ontology

The SeaLiT Ontology

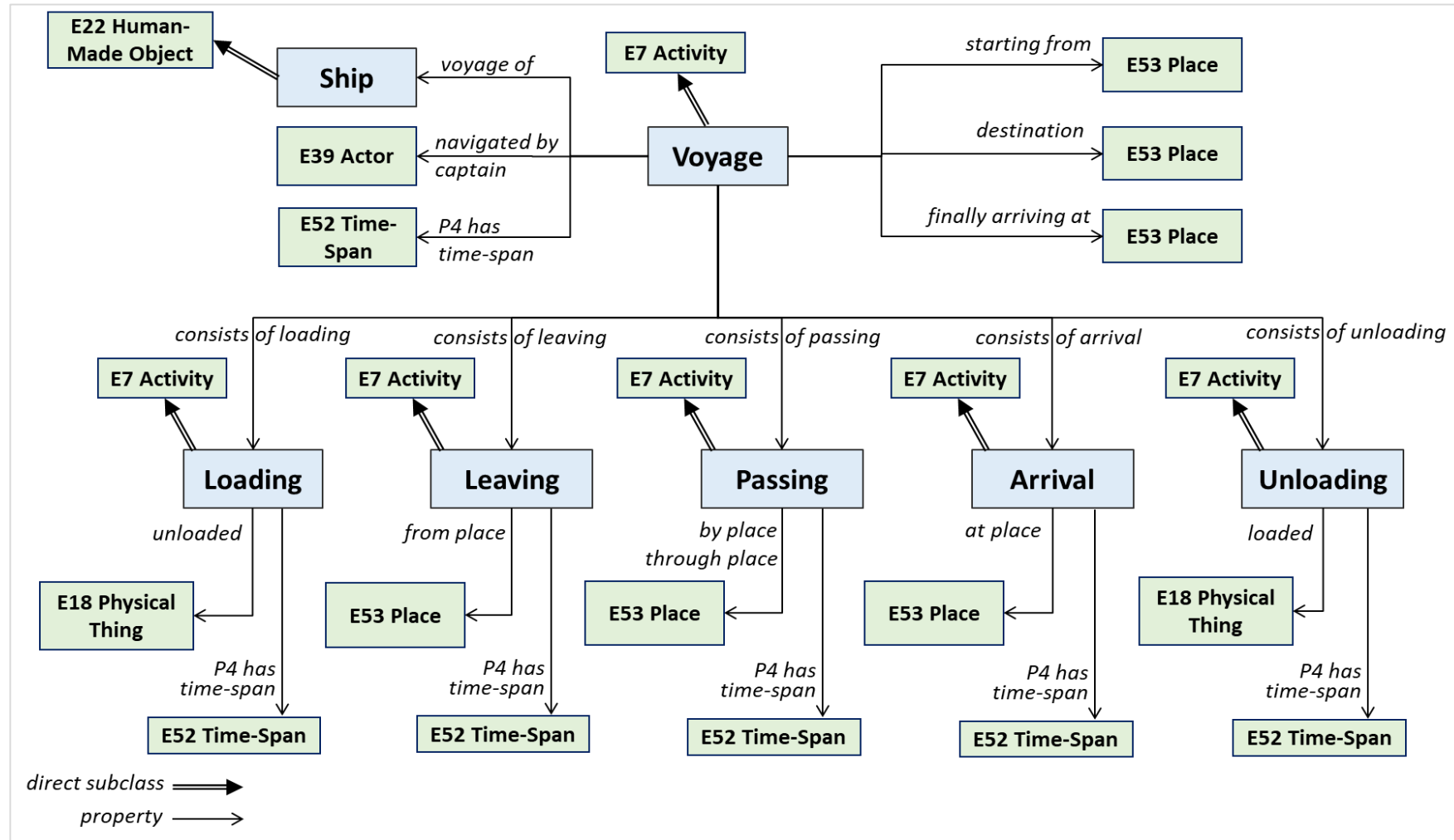
- ❑ It offers a **common** and **formal** 'language' for describing information about Maritime History in a **structured** way
- ❑ Created as an extension of **CIDOC-CRM** (Conceptual Reference Model)
 - A high-level ontology ([ISO standard 21127:2014](#)) for describing concepts and relationships used 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>

The screenshot shows the Zenodo record page for the SeaLiT Ontology dataset. The page features a blue header with the Zenodo logo, a search bar, and navigation links for 'Upload' and 'Communities'. On the right side of the header, there are 'Log in' and 'Sign up' buttons. Below the header, the record is dated 'February 3, 2022' and is labeled as a 'Dataset' with 'Open Access' status. The main title is 'SeaLiT Ontology - An extension of CIDOC-CRM for the modelling of Maritime History information'. The authors listed are Kritsotaki, Athina; Fafalios, Pavlos; and Doerr, Martin. A descriptive paragraph explains that the ontology is a formal ontology intended to facilitate the integration, mediation, and interchange of heterogeneous information related to maritime history. It aims to provide semantic definitions needed to transform disparate, localized information sources into a coherent global resource. 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. A table on the right side of the page displays statistics for the dataset, comparing 'All versions' and 'This version'. The statistics include Views (274), Downloads (193), Data volume (227.1 MB), Unique views (218), and Unique downloads (136). A link for 'More info on how stats are collected' is provided at the bottom of the statistics table. The footer of the page includes the FORI logo and the text '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'.

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

	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.

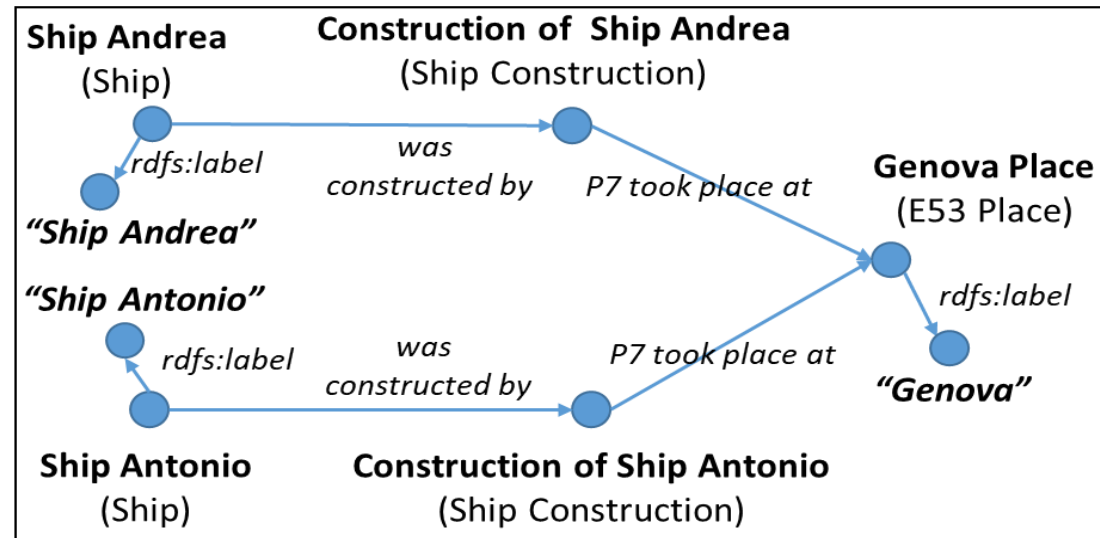
Creation of the Semantic Network

□ Data transformation to a rich **semantic network** based on the **SeaLiT Ontology**

➤ Using the **X3ML toolkit** (language, 3M user interface, transformation engine)

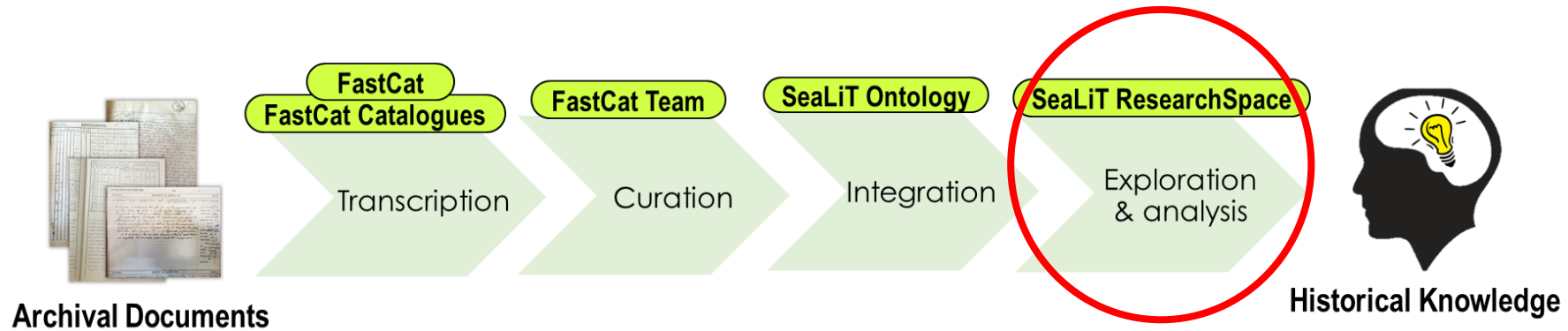
➤ A (very small) part of the derived semantic network:

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



➤ The full network contains >**18 million** edges!

▪ Available at: <https://zenodo.org/record/6460841>



Data exploration and analysis with SeaLiT ResearchSpace

SeaLiT ResearchSpace

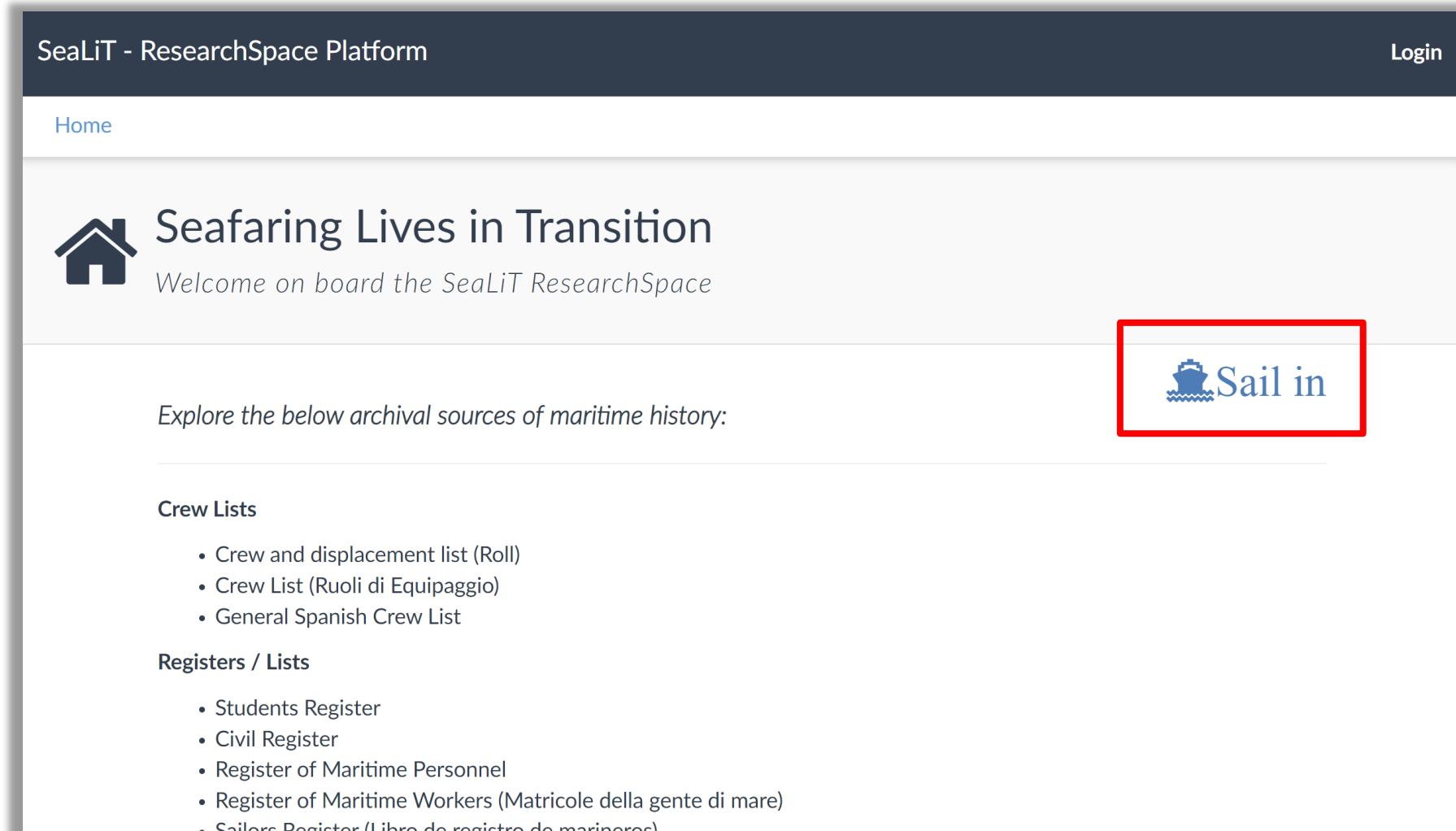
- ❑ **ResearchSpace**: open source platform, developed by the British Museum
- ❑ Offers a variety of functionalities:
 - Semantic search (through an assistive query building interface)
 - Data browsing
 - Variety of visualization methods (tables, charts, map)
 - Filtering (based on entity properties/relations)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



<http://rs.sealitproject.eu/>

SeaLiT ResearchSpace


<http://rs.sealitproject.eu/>




The screenshot shows the homepage of the SeaLiT ResearchSpace Platform. At the top, there is a dark blue header with the text "SeaLiT - ResearchSpace Platform" on the left and "Login" on the right. Below the header, there is a white navigation bar with the word "Home" in blue. The main content area has a light gray background. On the left, there is a dark blue house icon. To its right, the text "Seafaring Lives in Transition" is displayed in a large, dark font, followed by the subtitle "Welcome on board the SeaLiT ResearchSpace" in a smaller, italicized font. Below this, there is a section titled "Explore the below archival sources of maritime history:" followed by a horizontal line. To the right of this section, there is a red-bordered box containing a blue icon of a ship and the text "Sail in". Below the horizontal line, there are two sections: "Crew Lists" and "Registers / Lists". Each section contains a bulleted list of links to various archival sources.

SeaLiT - ResearchSpace Platform Login

Home

 **Seafaring Lives in Transition**
Welcome on board the SeaLiT ResearchSpace

Explore the below archival sources of maritime history:



Crew Lists

- Crew and displacement list (Roll)
- Crew List (Ruoli di Equipaggio)
- General Spanish Crew List

Registers / Lists

- Students Register
- Civil Register
- Register of Maritime Personnel
- Register of Maritime Workers (Matricole della gente di mare)
- Sailors Register (Libro de registro de marineros)

Assistive query building in ResearchSpace

Find: Ships WAS CONSTRUCTED IN Year 1850 AD - Year 1860 AD

Ship was constructed in Date Year 1850 AD - Year 1860 AD [remove](#)

Found 44 matches

Timeline Chart **List** Rows Map Grid

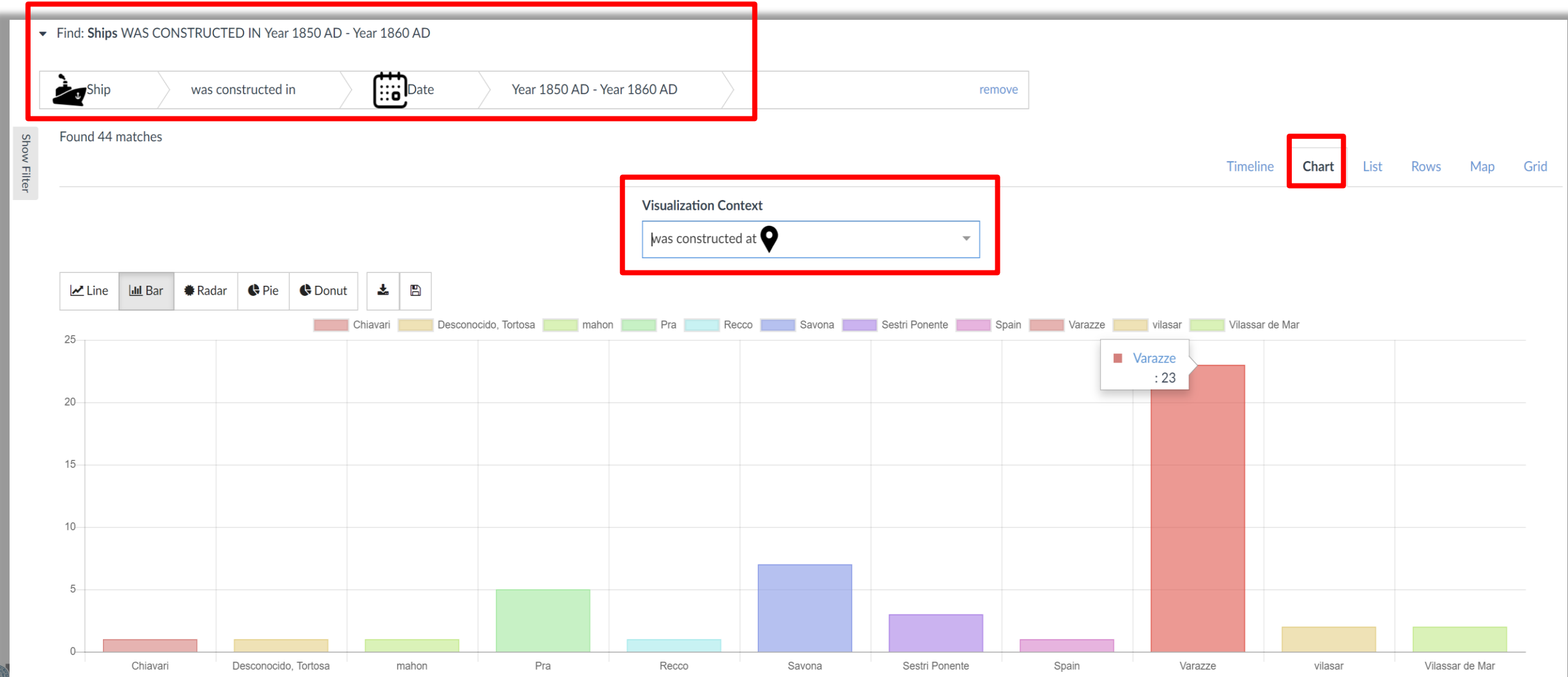
Filter Results

subject

- ship Amalia
- ship Andrea
- ship Antonietta
- ship Antonio
- ship Assunta
- ship Aurora
- ship Ave
- ship Camilla
- ship Colombia
- ship Concordia

« 1 2 3 4 5 »

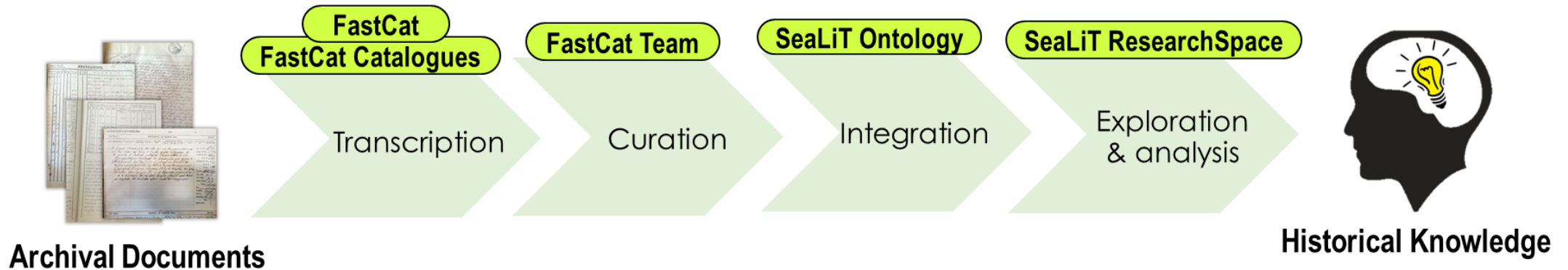
Assistive query building in ResearchSpace



Conclusion

Conclusion

□ IT tools for holistic data management in historical-archival research



- **Data provenance-aware**, following established **documentation standards (CIDOC-CRM)**
- Successful **application** in a large scale research project of maritime history (**SeaLiT**)
- **Configurable** for use over other types of sources

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, quantities, etc.)** is sometimes difficult or impossible [FastCat / FastCat Team]
 - Due to use of different reference points or units across sources
 - **We need mechanisms to align such values!**

Try the data exploration applications:
<https://sealitproject.eu/digital-seafaring>

Thank you!

Pavlos Fafalios
fafalios@ics.forth.gr

The research & 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



European
Research
Council

