CHAPTER 17

Building and Exploring a Semantic Network of Maritime History Data

Pavlos Fafalios, Georgios Samaritakis, Kostas Petrakis, Korina Doerr, Athina Kritsotaki, Anastasia Axaridou and Martin Doerr

1 Introduction

A vast area of research in historical science concerns the quantitative analysis of empirical facts extracted from historical sources, their description, and the interpretation of possible causes, influences, and evolutionary trends. This kind of research requires a holistic data management approach that supports historians in all the different activities involved in their research processes, from digitizing the (usually hand-written) archival sources, to curating the transcribed data and performing quantitative analysis and exploration.

Current practice nearly exclusively uses spreadsheets or simple relational databases to organize the data as rows with multiple columns of related parameters, such as prices for kinds of goods in certain harbours at certain times, for the case of maritime history. This form offers itself to direct quantitative analysis under varying parameters, and can consequently be used for the scholarly interpretation of causes and impacts. However, it makes difficult: i) the collaborative curation and analysis of the transcribed data; ii) the combination of data coming from different and diverse sources; iii) the documentation of information on provenance (important for verification and long-term validity); and in general iv) the exploitation of the transcribed data beyond the context of a particular research problem.

To cope with these problems, in this chapter we describe a data management approach and a set of tools that support historians in collaboratively digitizing, curating, and exploring their unique information sources. The proposed approach is focused on semantic interoperability,[1] making easy the integration of data coming from diverse data sources, the description of provenance information, and the storage and publication of all data in a form that makes their

1 Aris M. Ouksel and Amit Sheth, "Semantic interoperability in global information systems," *ACM Sigmod Record* 28, no. 1 (1999): 5–12.

future exploitation easier.[2] This is achieved by enabling the description of rich metadata about the data and the use of standard models for describing and storing the data.

For supporting historians in digitizing and curating their historical information sources, we have introduced the FAST CAT system. FAST CAT is a browser-based system that supports both online and offline data entry with the possibility of automated synchronization when the researcher gets online. In FAST CAT, data from different information sources can be transcribed as "records" belonging to specific "templates", where a "template" represents the structure of a single data source. A record organizes the data and metadata into tables (similar to spreadsheets), offering functionalities like nesting tables and selection of term from a vocabulary. The curation of the transcribed data can be performed through FAST CAT TEAM, a special environment of FAST CAT that allows the collaborative management of "entities" (persons, locations, ships, legal entities) and "vocabulary terms" (like ship type, profession, etc.) that appear in the transcribed records. An important characteristic here is that this curation activity does not alter the data in the records as transcribed from the original archival sources, which is very important for maintaining the transcribed data as close to the original sources as possible. FAST CAT is innovative in the sense that it supports features like nested tabular structures for data entry, embedded instance matching and vocabulary maintenance processes, as well as provenance-aware data curation. These are important features that are not currently supported by existing data management systems.

The transcribed and curated data, together with their metadata, are then transformed into a rich semantic network using a data model compatible to CIDOC-CRM, a reference ontology and standard for the representation of cultural heritage information. Finally, the semantic network is available for exploration by historians and other interested parties through a Web application that supports users in building complex queries and visualizing the results in various ways, like tables, charts, or in a map. For example, historians can inspect a bar chart showing "the distribution of the embarkation locations of seamen that were crew members at ships of type brigantine".

The context of this work is the SeaLiT project,[3] a European (ERC) project of maritime history which studies the transition from sail to steam navigation and its effects on seafaring populations in the Mediterranean and the Black

2 Martin Doerr and Dolores Iorizzo, "The dream of a global knowledge network—A new approach," *Journal on Computing and Cultural Heritage* (*JOCCH*) 1, no. 1 (2008): 1–23.

3 http://www.sealitproject.eu/ (accessed 11 October 2021).

Sea between the 1850s and the 1920s.[4] The project investigates the maritime labour market, the evolving relations among ship-owners, captain, crew, and local societies, and the development of new business strategies, trade routes, and navigation patterns, during the transitional period from sail to steam. The information sources used in SeaLiT range from hand written ship log books, crew lists, payrolls and student registers, to civil registers, business records, account books, and consulate reports, gathered from different authorities and written in different languages, including Spanish, Italian, French, Russian, and Greek. We showcase the use of the described tools using these data sources.

Below, we start by describing how the archival sources can be transcribed and curated using FAST CAT and FAST CAT TEAM, then we detail the modelling and transformation of the transcribed and curated data into a rich semantic network, and finally we describe the data exploration functionalities of the ResearchSpace system.

2 Data Digitization with FAST CAT

The FAST CAT system was designed with the objective to support historians in faithfully cataloguing historical data sources for use as a primary source for research and for long-term validity and reuse, beyond the objectives of a particular research problem or project. The focus is to enable historians to transcribe as much relevant information as possible and as exact and fast as possible.

In FAST CAT, the first step before starting the digitization process is the creation of the "templates", each one representing a distinct archival source. This is performed in a pre-processing step, in close collaboration between historians and data engineers. This collaboration is necessary for better designing the structure of the data entry forms in a template, in a way that enables historians to accurately and fast digitize the archival data. During the configuration of a template, we need to provide the structure of the tables that all together constitute the template. Each table has a title and consists of a set of columns. A column accepts values of different types, in particular: i) "entity" (the value is the name or an attribute of an entity, e.g., the name or birth date a person); ii) "vocabulary term" (the value is a term from a controlled vocabulary); iii) "literal" (the value is a literal, e.g., a free text, a number, or a date); or iv) "nested table" (the value is another table). Also, a set of columns can be configured to accept multiple values in a single table row (e.g., for providing

4 Apostolos Delis, "Seafaring Lives at the crossroads of Mediterranean maritime history," *International Journal of Maritime History* 32, no. 2 (2020): 464–478.

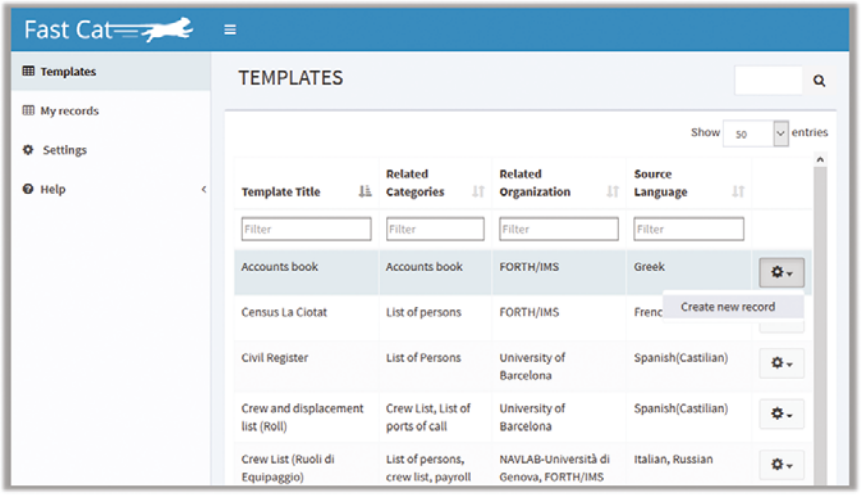

FIGURE 17.1 The home page of FAST CAT

multiple ship owners in a row describing information about a single ship). After having configured the templates, historians are able to start the digitization process by creating "records", each one corresponding to a particular template.

Fig. 17.1 shows the home page of FAST CAT, where the user is shown a table containing all the available templates. For each template, the table provides some basic information, in particular the "title" of the template, "related categories" (keywords that represent the type of the included information), "related organizations" (responsible for the data entry process), and the "source languages". The user can select a template and create a new record to start digitizing a particular archival source. For creating the record, the user first needs to provide some basic information by filling a small form. This information is different for each template. For instance, for a record of type "Logbook" the user must provide the following information: Type of Ship, Name of Ship, Date of Document (From/To), Author of Record, the latter being the name and surname of the historian performing the data entry.

After filling the form, the record is created and the user can start entering the data in the record tables. Fig. 17.2 shows an example of a record of template "Logbook". The first two tables of each record contain some basic metadata information about the record and the archival source. In the first table, entitled "FastCat Record Information", the user can provide the name and role of each historian performing the data entry, while the record ID, creation date, and last modification date are automatically filled by the system. In the second table, entitled "Source Identity", the user can provide information about the archival source, like the name and location of the archive or library, the collection title,

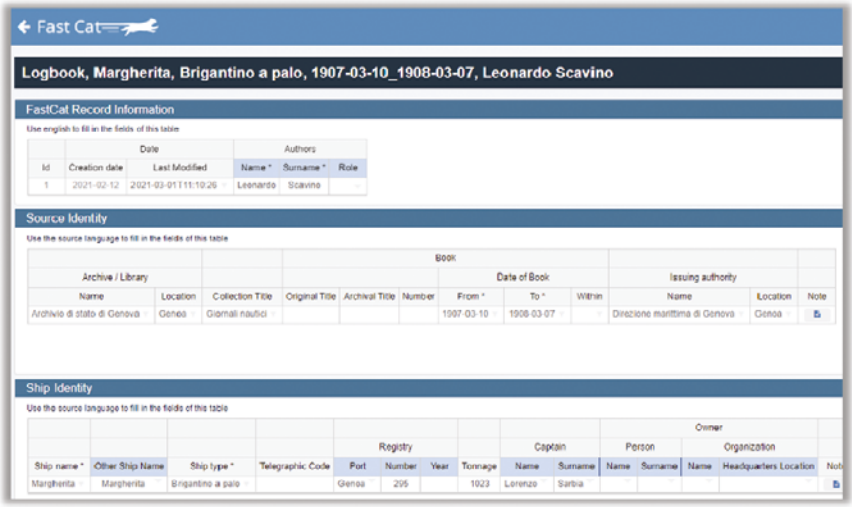

FIGURE 17.2 An example of a FAST CAT record belonging to the template "Logbook"

the book title and date, and the name and location of the issuing authority. The remaining record tables are different for each different archival source. For instance, for the case of logbooks, the template contains two additional tables: "Ship Identity" and "Voyage Calendar". In the "Ship Identity" table, the user can provide information like ship name, ship type, telegraphic code, registry port, tonnage, captain, owner, etc., while in the "Voyage Calendar" table, the user can provide information about the recorded voyages of the logbook, like the route and duration of each voyage or its analytical calendar. For the analytical calendar, the user can provide the data in a separate nested table, as shown in Fig. 17.3.

When the user has finished the data entry in a particular record, she/he can share it on FAST CAT TEAM and start data curation (more below). The user can also export the transcribed data of a record in Excel or XML format, e.g. for offline analysis.

FAST CAT is currently used in the SeaLiT project by around 30 historians from five organizations in five countries (Greece, Spain, Italy, France, Croatia). The number of configured templates is currently twenty, representing twenty different types of archival sources, while the total number of records is more than 600. Note here that some templates are shared by historians of different countries since the corresponding archival sources share a similar structure. Indicative examples of records are available online.[5]

5 https://www.ics.forth.gr/isl/fast-cat (accessed 11 October 2021).

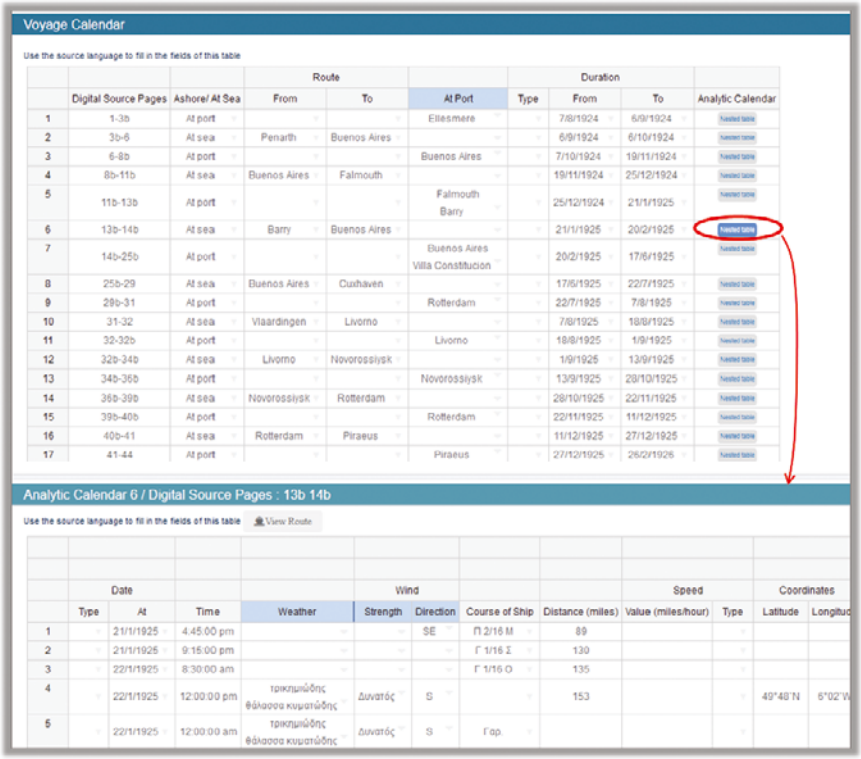

FIGURE 17.3 An example of a nested table in FAST CAT

3 Data Curation with FAST CAT TEAM

After finishing the digitization of the different archival sources, historians need to curate the transcribed data and sometimes enrich them with additional information, in order to integrate them into a common form from which historical research and quantitative analysis can be carried out correctly and efficiently. This involves several steps, including:

- applying corrections in entity names, like names of persons or locations
- adding missing entity information, or enriching with additional data, e.g., adding coordinates in the locations for enabling map visualizations
- maintaining vocabularies of terms for certain types of data that appear in the transcripts, e.g., a vocabulary containing preferred and broader terms for "ship types"
- dealing with varying entity identity assumptions; a problem known as *instance matching*

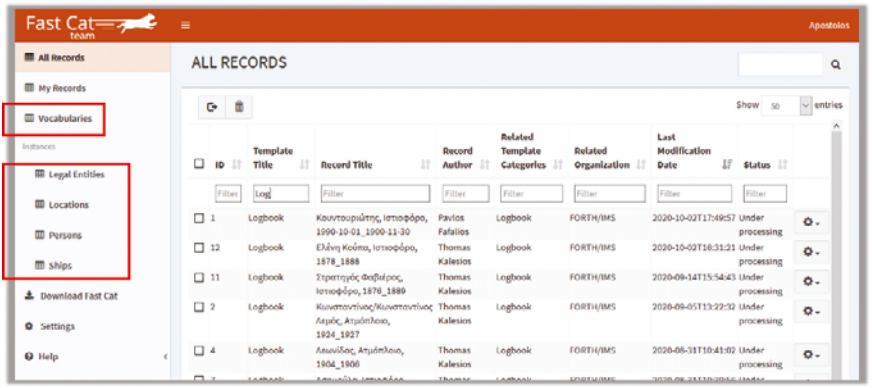

FIGURE 17.4 The home page of FAST CAT TEAM

An important requirement here is that all these curation steps must not alter the original transcribed data; historians must be able to go back at any time and check how a particular piece of information appears in the original transcripts. This is crucial for reliability of the research results as well as validation.

FAST CAT offers a special environment, called FAST CAT TEAM, through which the historians can collaboratively perform the above curation steps without changing the original transcribed data.

Fig 17.4 shows the home page of FAST CAT TEAM. In the home page, the user is shown a table containing all the publicly shared FAST CAT records whose data can be curated. The curation steps are organized into two categories, accessible through the left menu: i) management of vocabularies; and ii) management of entity instances (Legal Entities, Locations, Persons, and Ships, in our case).

When visiting the "Vocabularies" menu item, the user is shown a table with all vocabularies whose terms appear in the transcribed data (Fig. 17.5). For each vocabulary, the table shows its title, its source languages, the templates in which it is used, and the related organizations (responsible for the templates). Next to each vocabulary title, the table also shows the approximate number of vocabulary terms. The user can select a vocabulary from the table and edit or export it (in JSON format). By selecting to edit a vocabulary, the user is redirected to a web page showing a table with all terms of the selected vocabulary (Fig. 17.6). For each term, the table shows its value as it appears in the transcribed record(s), its preferred value (in English), and its broader term. The preferred term value and the broader term can be filled by the user by clicking on the "Edit term" option. The user can also directly inspect and visit the FAST CAT records in which the term appears by clicking on the information icon at

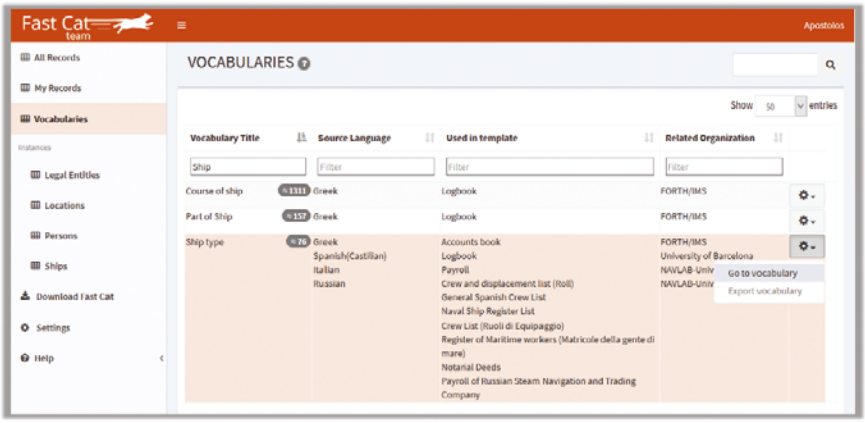

FIGURE 17.5 Management of vocabularies in FAST CAT TEAM

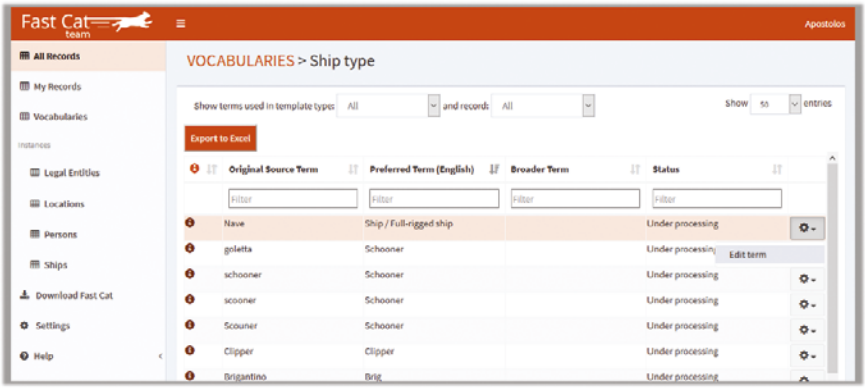

FIGURE 17.6 Inspecting the terms of a vocabulary in FAST CAT TEAM

the left of each term (Fig. 17.7), thereby getting additional context information that can help the user to disambiguate or better understand the term.

The curation of vocabularies is very important for supporting more accurate analysis services and more effective data exploration services. For example, by knowing that the preferred terms "captain", "steward", "guardian", and "able seaman" of the vocabulary "profession"—that might have been written in different languages or in different forms in the original sources—have the term "sailor" as a broader term, we can use the term "sailor" for handling collectively all these narrower terms, e.g., in a search service ("show me all crew members of profession sailor"), or when performing quantitative analysis ("show me the number of crew members per profession").

For the curation of the entity instances that appear in the transcribed data (in our case: Legal Entities, Locations, Persons, and Ships), users can select

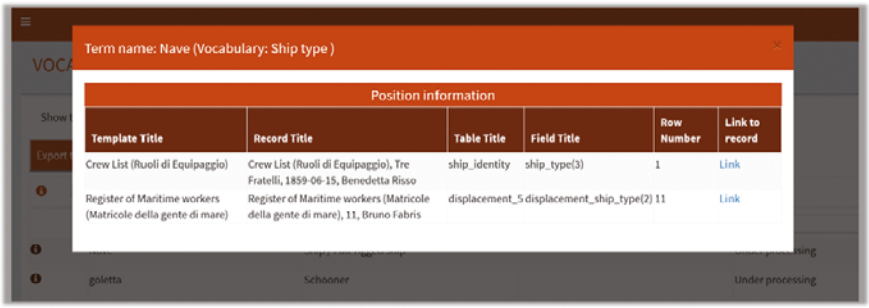

FIGURE 17.7 Inspecting the FAST CAT records in which a vocabulary term appears

the corresponding menu item (see Fig 17.3) and inspect a table with all the instances of the selected entity type. For each different entity type, the table displays different information and the user has different curation options.

For Persons, the table shows the following attributes: name, surname A, surname B, maiden name, father's name, place of birth, date of birth, date of death, registration number, status/capacity/role. Note here that some of these attributes might be empty if there is no such information in the transcribed records. The user can select a person and change the value of one of its attributes, select two or more persons and indicate that they correspond to the same person (manual instance matching), also allowing the indication of the preferred value if there is a conflict in the name or one of the attributes, or mark as different one or more already matched person instances (Fig. 17.8). Giving the same identity to different person references (like "A.M. Micheletti" and "Achille Marius Michelletti") for which we learn (through historical research) that they correspond to the same individual, is very important when performing historical quantitative analysis, like computing the number of sailors per birth place (since in this way we avoid misleading statistics and errors in the analysis results), while it is also very important in qualitative terms, e.g., when constructing biographical trajectories. For the same reasons, it is important to give different identities to similar person references that we learn correspond to different individuals (like two different persons with the same name).

For "Ships", the user is shown with the following attributes: name, previous name, type, call signal, construction location, construction date, telegraphic code, flag, owner company, registration list, registration number, registration location. Similar to the case of persons, the user can correct one of the ship attributes, select two or more ships and indicate that they correspond to the same ship, or mark as different one or more already matched instances. For Legal Entities, the system only shows the name of the legal entity as it appears in the record(s), allowing the user to change its value and set a preferred

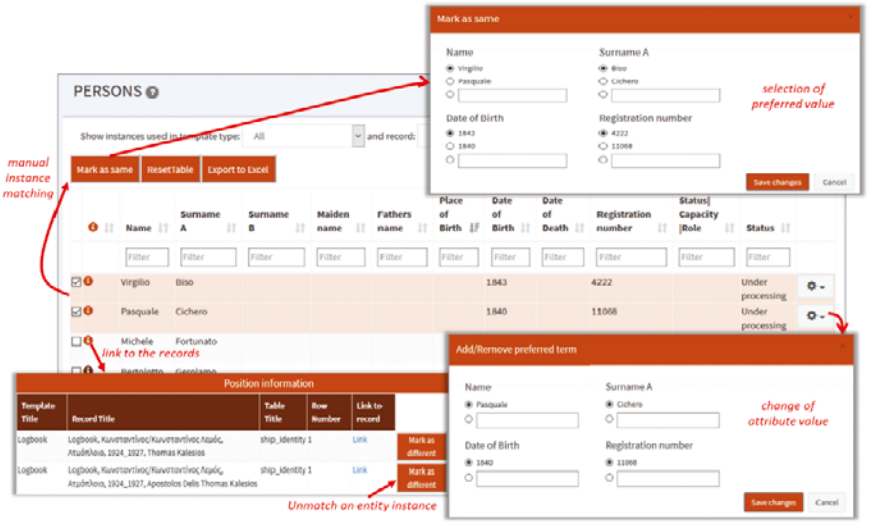

FIGURE 17.8 Curation of "Person" instances in FAST CAT TEAM

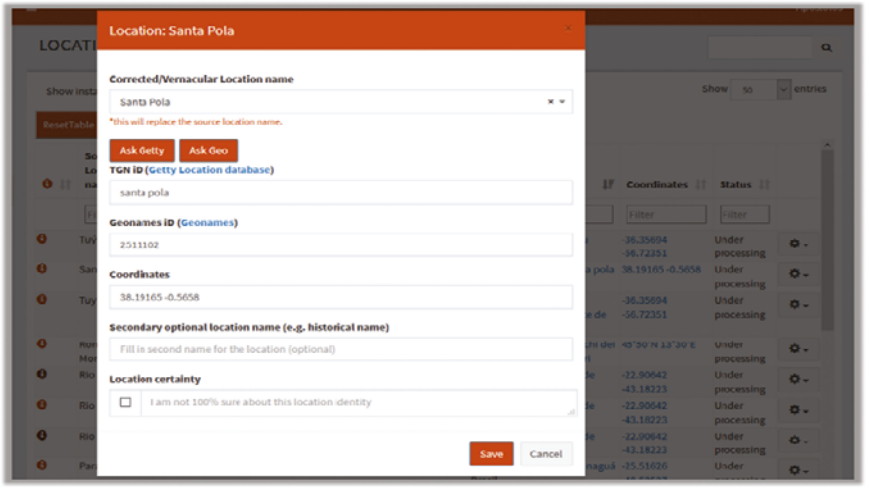

FIGURE 17.9 Curation of "Location" instances in FAST CAT TEAM

one. The user can also select two or more legal entities and indicate that they correspond to the same entity, or mark as different one or more already matched instances.

Finally, for "Locations" the system shows the following data: source location name, corrected/vernacular location name, other location name, location type, broader location name, ID (TGN or Geonames ID), coordinates. The user can select a location and correct or provide one of its attributes, as shown in

Fig. 17.9. Here the system offers the capability to directly query external geolocation services, in particular Getty Thesaurus of Geographic Names (TGN)[6] and Geonames,[7] and get the unique ID of the location as well as its coordinates. When querying these services, the system retrieves and shows a list with the relevant locations, allowing the user to select the correct one. When selecting one of the retrieved locations, the system shows the location on a map, helping the user to directly check if the location is correct, and the coordinates are automatically filled. If the user is not sure about the exact location, she/he can tick the checkbox "Location certainty" and optionally include a comment. Linking the locations to coordinates is very important for enabling the visualization of the transcribed data on a map.

In the context of the SeaLiT project, the total number of vocabularies available for curation is currently 52. Examples include: ship type, flag, marital status, religion, military service status, nationality, profession, reason of death, wind direction, and wind strength. As regards the number of entity instances, there are currently 72,093 person instances, 7,638 location instances, 1,876 ship instances, and 1,137 legal entity instances. Since the investigation process is still open and new records are created, and also the manual instance matching process made by the historians is still in progress, the number of distinct instances in each entity type is expected to be different at the end of the data entry and curation processes.

4 Data Modelling and Transformation to a Semantic Network

The transcribed and curated data is now ready for transformation into a rich semantic network of "Linked Data"[8] that will allow for advanced exploration and analysis of the historical information that appears in the archival sources. The objective is to model and semantically represent the curated data and metadata using established and standard data models, for supporting data exchange, interoperability, and long-term validity and reuse. For this, we need:

a) to decide on the domain schema (ontology) to use for representing the data of the FAST CAT records
b) to create the schema mappings for each template in FAST CAT, which map the table columns of a template to classes and properties in the domain ontology

6 http://www.getty.edu/research/tools/vocabularies/tgn/ (accessed 11 October 2021).
7 https://www.geonames.org/ (accessed 11 October 2021).
8 Tom Heath and Christian Bizer, "Linked data: Evolving the web into a global data space," *Synthesis lectures on the semantic web: theory and technology* 1, no. 1 (2011): 1–136.

c) to run the transformations considering the designed schema mappings and the domain ontology, and then ingest the produced semantic data (RDF triples[9]) to a semantic repository (RDF database), from where data analysis and exploration can be initiated

4.1 *The Domain Ontology*

As the domain ontology, we created a data model compatible with the CIDOC Conceptual Reference Model (CRM).[10] CIDOC-CRM is a high-level, event-centric ontology (ISO standard 21127:2014)[11] of human activity, things and events happening in spacetime, providing definitions and a formal structure for describing the implicit and explicit concepts and relationships used in cultural heritage documentation.[12] It is intended to be used as a common language for domain experts and implementers to formulate requirements for information systems, providing a way to integrate cultural heritage information of different sources. CIDOC-CRM has been used in a plethora of projects and data management activities related to (mainly) cultural heritage, history, and archaeology.[13]

The data model, which we call "SeaLiT Ontology", is under constant evaluation as long as new archival sources are analysed. It currently allows describing information about "Ships", "Ship voyages", "Crew payments" and "Actors" (like persons or organizations).

Fig. 17.*10* shows how these main concepts are related. A "Ship" made a "Voyage" which was carried out by "Actors" in particular roles (e.g., captain), while a "Crew Payment" was made for a particular "Voyage" and was carried out by "Actors" in specific roles (e.g., employer and employee). The classes "Ship", "Voyage", and "Crew Payment" are newly introduced classes of the SeaLiT Ontology, extending classes of CIDOC-CRM.

Fig. 17.11 depicts how information about a ship is modelled in the SeaLiT Ontology. First, a "Ship" is the result of a "Ship Construction" that took place at a particular "Place" and in a particular "Time-Span", and which gave the "Ship Name" to the ship. A "Ship" also has some characteristics, like "Horsepower", "Tonnage" and "Crew Number Capacity", and is registered during a "Ship Registration" activity by a "Port of Registry", with a ship flag of a particular "Country" and with a particular "Ship ID". The "Ship Registration" is

9 https://www.w3.org/TR/rdf-concepts/ (accessed 11 October 2021).
10 http://www.cidoc-crm.org/ (accessed 11 October 2021).
11 https://www.iso.org/standard/57832.html (accessed 11 October 2021).
12 Martin Doerr, "The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata," *AI magazine* 24, no. 3 (2003): 75–75.
13 http://www.cidoc-crm.org/useCasesPage (accessed 11 October 2021).

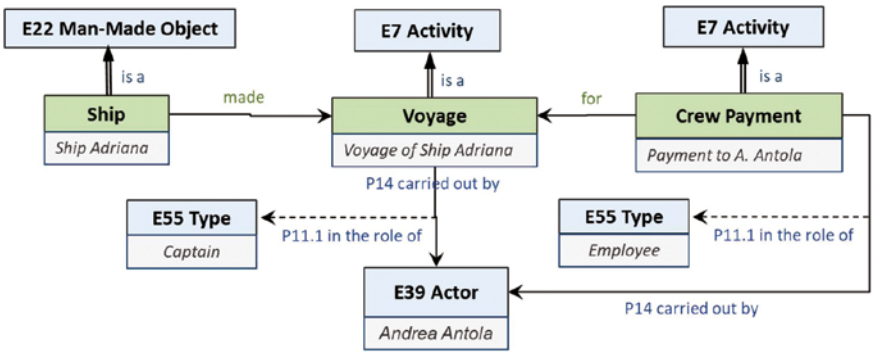

FIGURE 17.10 The main concepts of the SeaLiT Ontology

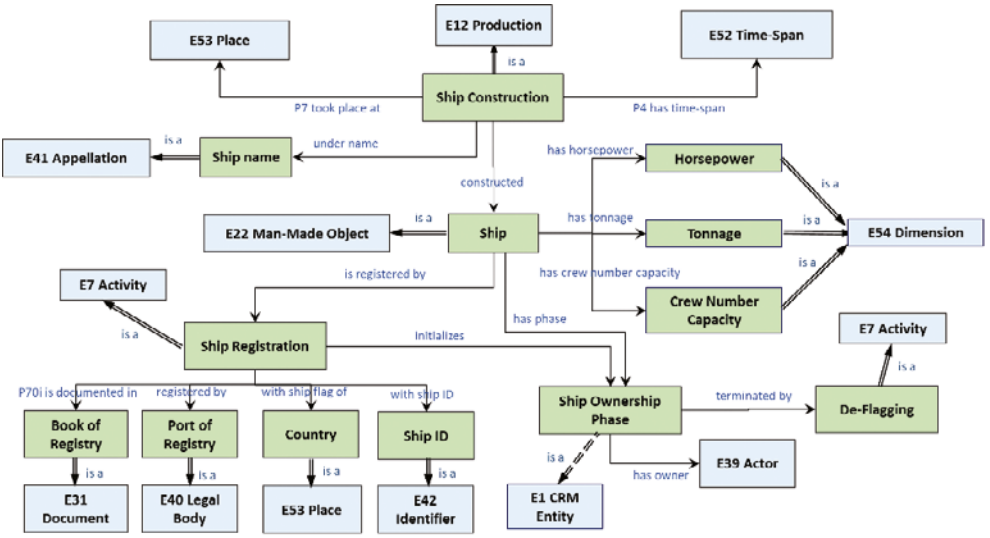

FIGURE 17.11 Modelling information about a ship

documented in a "Book of Registry". Finally, a "Ship" has one or more "Ship Ownership Phases" which can be terminated by a "De-Flagging" activity.

Fig. 17.12 shows all different activities related to a "Ship" that have been modelled in the ontology. This includes "Ship Construction", "Ship Registration", "Ship Ownership Phase", "Voyage", "Ship Repair", "Ship Destruction", and "De-Flagging". De-Flagging can be the result of a "Ship Destruction", "Purchase or Transfer of Custody". Note that each of these classes allows for the describing of further information about the corresponding activity, like time-span, locations, involved actors, identifiers, etc.

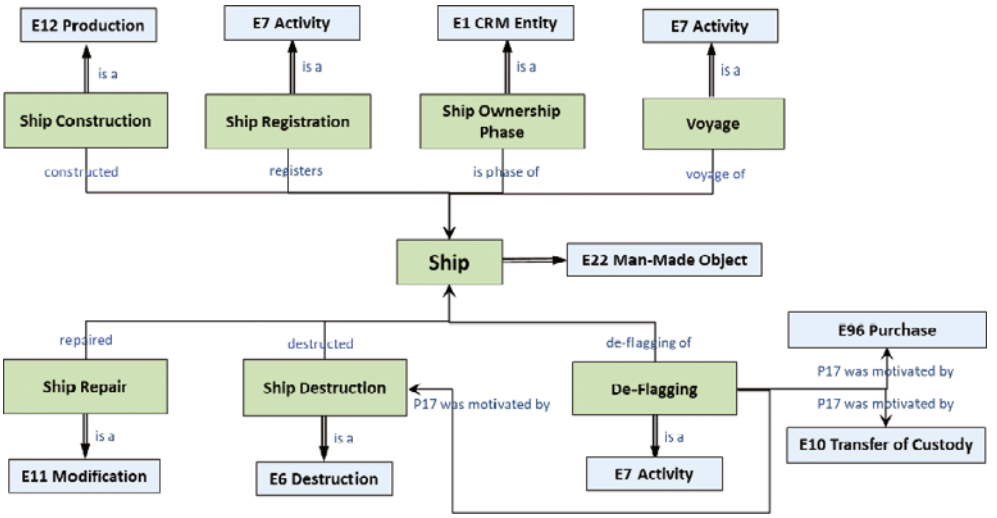

FIGURE 17.12 Activities related to a ship

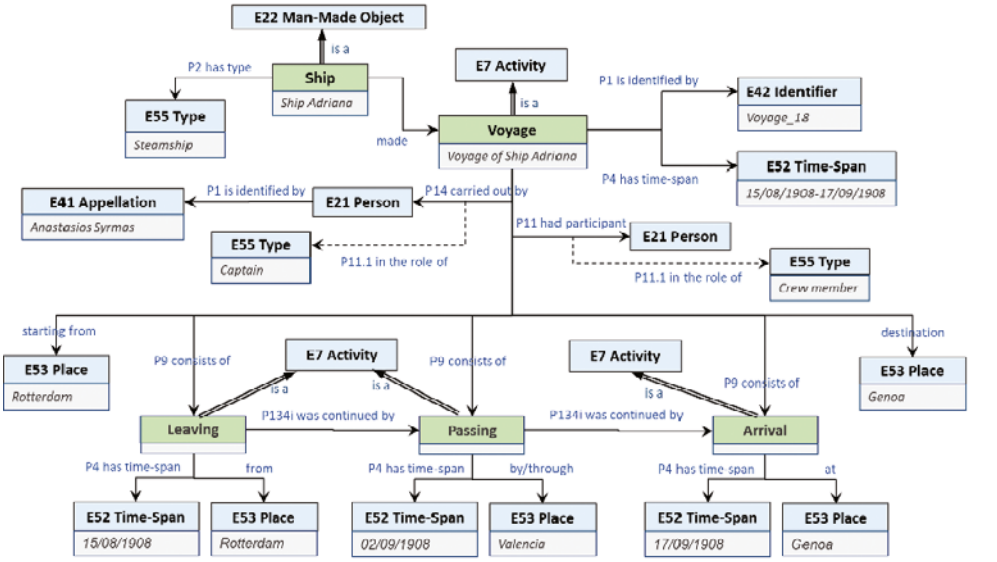

FIGURE 17.13 Modelling information about a ship voyage

Fig. 17.13 depicts how we model information about a ship voyage. A "Voyage" refers to a particular "Ship" and has a particular "Time-Span". It has been carried out by some "Persons" and has some other "Persons" as participants. It has a starting "Place", a destination "Place", and also consists of one or more sailing activities, like "Leaving" a place, "Passing" by a place, or "Arriving" at a place in a particular "Time-Span". These activities can be connected through

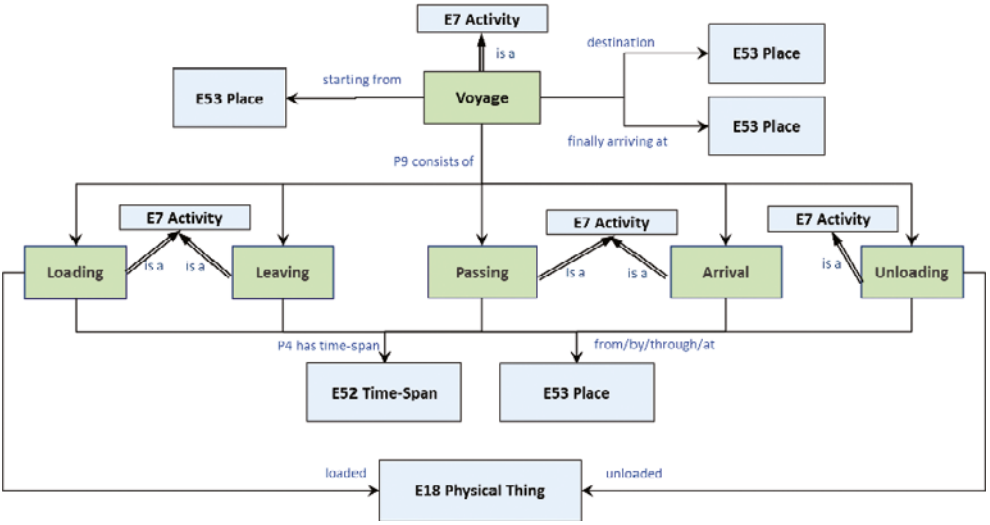

FIGURE 17.14 Activities related to a ship voyage

the CIDOC-CRM property "P134i was continued by", thereby enabling modelling their sequence (since we may know the sequence of the activities but not the time-spans of all of them). Fig. 17.14 shows all different activities and properties related to a ship voyage. This includes "Loading" and "Unloading" things, "Leaving" a place, "Passing" by or through a place, and "Arrival" at a place. Note also that, apart from the "starting from" and "destination" places, the voyage can be connected to a "finally arriving at" place, since arriving at a different place from the one originally planned is quite common in historical voyages.

Fig. 17.*15* shows how we model information about payments and employments. A "Crew Payment" activity concerns a particular "Voyage", has a particular "Monetary Amount", has been carried out by some "Actors" (like the person who received the payment and the person who gave the payment), took place within a particular "Time-Span", and has been agreed in a "Payment Contract". A "Crew Payment" is also connected to particular "Ship Crew Employment", carried out by some "Actors" (like the employer and the employee), started by a "Recruitment" activity and ended by a "Discharge" activity, each one linked to a particular "Time-Span". Each of these classes allows describing additional information, like connected places, persons, or other activities.

Fig. 17.16 shows how we model information about a person. A "Person" is identified by an "Appellation" (like first name and last name), was born at a particular "Place" and within a particular "Time-Span", can be a current or former member of a "Legal Body", and has some attributes, like "Social Status", "Language Capacity", "Literacy Status", "Sex Type", and "Religion Status". Each of these attributes can receive values from a dedicated vocabulary or thesaurus of

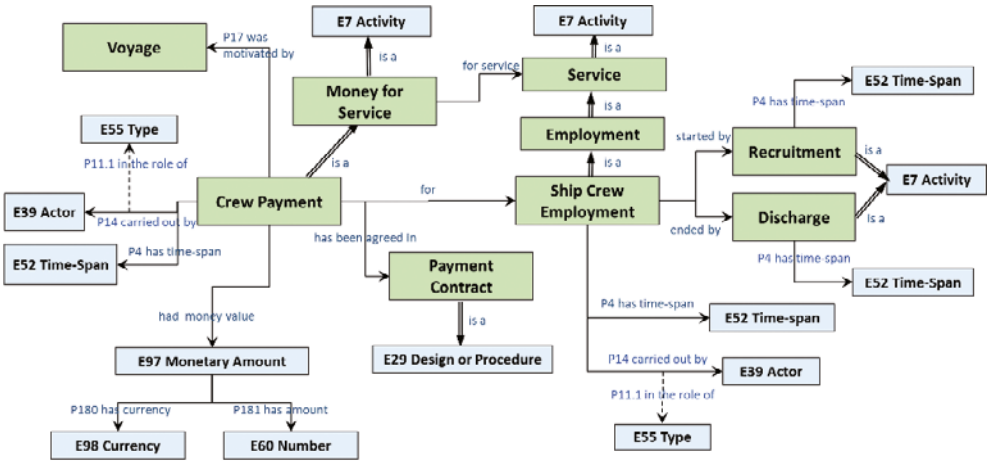

FIGURE 17.15 Modelling information about a crew payment

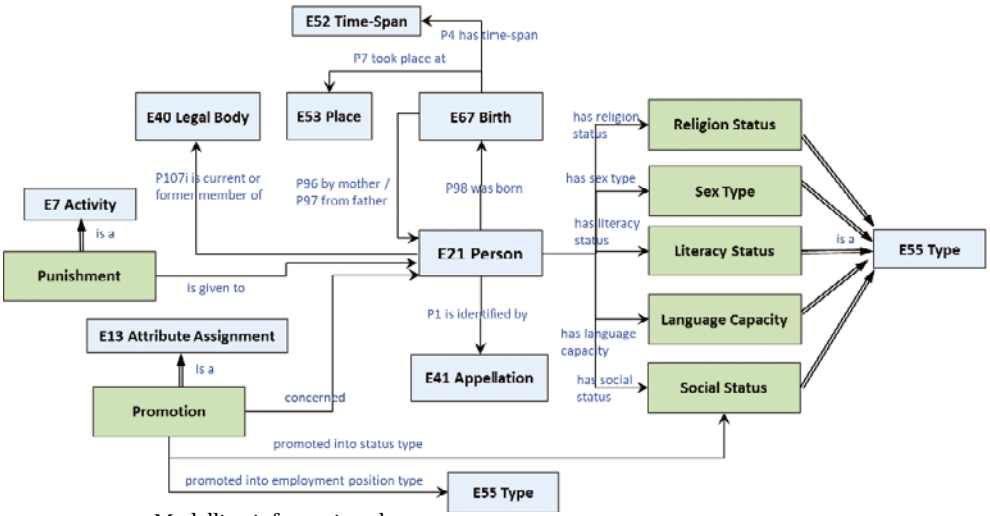

FIGURE 17.16 Modelling information about a person

terms in an information management system (as we do in FAST CAT). Finally, a "Person" can be connected to a "Promotion" (i.e., into a new social status or employment position) or "Punishment" (which is common for crew members during voyages).

4.2 *Data Transformation*

After defining the domain ontology to use for representing the data in the FAST CAT records as a semantic network, we need to create the schema mappings that associate equivalent concepts and relationships from the source schemata

FIGURE 17.17 The home page of 3M editor

(FAST CAT templates, in our case) to the target schema (the SeaLiT ontology). For this, we make use of the X3ML framework and the X3ML mapping definition language, a declarative, XML-based and human readable language that supports the cognitive process of schema mapping definition.[14] X3ML separates schema mappings from the generation of proper resource identifiers (URIs), so it distinguishes between activities carried out by the domain experts and data engineers, who know the data, from activities carried out by the IT experts who implement data transformation.

The definition of the mappings is a time-consuming process that can last for several weeks and can require many revisions as long as the data engineer better understands the data or changes are made to the FAST CAT templates. This process is supported by 3M Editor, an X3ML mapping management system suitable for creating and handling the mapping files. It offers a user interface and a variety of actions that help experts manage their schema mappings collaboratively. Fig.17.17 shows the home page of 3M Editor. The user can search for an existing mapping, create a new mapping, select a mapping from the list and view it, or edit one of the available mappings. When creating a new mapping, the user needs to upload a part of the source data (an XML file of the record data that can be obtained through FAST CAT) as well as the target schema (i.e., the SeaLiT Ontology).

Fig. 17.18 shows a screen dump of a mapping definition for the FAST CAT template "Crew List (Ruoli di Equipaggio)". We see that the "ship name" field of the source schema (first column of the table "Ship Identity"; see Fig. 17.2)

14 Yannis Marketakis, Nikos Minadakis, Haridimos Kondylakis, Konstantina Konsolaki, Georgios Samaritakis, Maria Theodoridou, Giorgos Flouris, and Martin Doerr, "X3ML mapping framework for information integration in cultural heritage and beyond," *International Journal on Digital Libraries* 18, no. 4 (2017): 301–319.

FIGURE 17.18 A part of a mapping definition in 3M editor for the template crew list (*Ruoli di Equipaggio*)

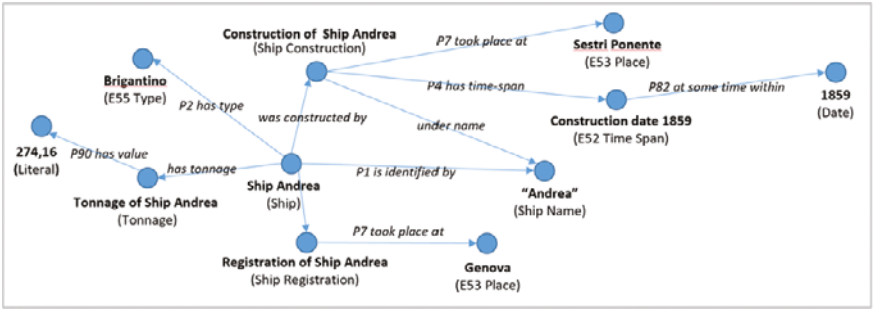

FIGURE 17.19 A small part of the derived semantic network

is mapped to the paths "Ship—was constructed by—Ship Construction—under name—Ship Name" and "P1 is identified by—Ship Name" on the SeaLiT Ontology. Likewise, the "ship type" field of the source schema is mapped to the path "Ship—P2 has type—E55 Type", the "ship tonnage" field to the path "Ship—has tonnage—Tonnage—P90 has value—Literal", the "construction location" field to the path "Ship—was constructed by—Ship Construction—P7 took place at—E53 Place", the "construction date" field to the path "Ship—was constructed by—Ship Construction—P4 has time-span—E52 Time-Span—P82 at some time within—Date", and the "registry port" field to the path "Ship—is registered by—Ship Registration—P7 took place at—E53 Place". Fig. 17.19 shows the corresponding part of the derived semantic network.

5 Semantic Network Exploration

When the transformation process is completed, the derived semantic network is loaded to a semantic repository (database of RDF data) and is made available for analysis and exploration by other applications. Here, the user-friendly exploration of a semantic repository can be performed through two main general access methods:

i) Keyword Search: the user submits a free text query and gets back a ranked list of results that are relevant to the query terms[15,16]

ii) Interactive Access: the user explores the data through intuitive interactions with a data access system, e.g., through a data browsing system,[17] a faceted search interface,[18] or an assistive query building system[19]

In the context of the SeaLiT project, we make use of ResearchSpace,[20] an open source platform built on top of the metaphactory platform.[21] ResearchSpace combines a variety of access methods, including data browsing and assistive query building, as well as a variety of visualization methods, such as tables, charts, or visualization on a map.

---

15 Giorgos Kadilierakis Pavlos Fafalios, Panagiotis Papadakos, and Yannis Tzitzikas "Keyword search over RDF using document-centric information retrieval systems," in *The Semantic Web, 17th International Conference, ESWC 2020, Heraklion, Crete, Greece, May 31–June 4, 2020, Proceedings* eds. Andreas Harth, Sabrina Kirrane, Axel-Cyrille Ngonga Ngomo, Heiko Paulheim, Anisa Rula, Anna Lisa Gentile, Peter Haase, and Michael Cochez (Cham: Springer, 2020), 121–137.

16 Nikas, Christos, Giorgos Kadilierakis, Pavlos Fafalios, and Yannis Tzitzikas, "Keyword Search over RDF: Is a Single Perspective Enough?," *Big Data and Cognitive Computing* 4, no. 3 (2020): 22.

17 Aidan Hogan, Andreas Harth, Jürgen Umbrich, Sheila Kinsella, Axel Polleres, and Stefan Decker, "Searching and browsing linked data with swse: The semantic web search engine," *Journal of web semantics* 9, no. 4 (2011): 365–401.

18 Yannis Tzitzikas, Nikos Manolis, and Panagiotis Papadakos, "Faceted exploration of RDF/S datasets: a survey," *Journal of Intelligent Information Systems* 48, no. 2 (2017): 329–364.

19 Vangelis Kritsotakis, Yannis Roussakis, Theodore Patkos, and Maria Theodoridou, "Assistive Query Building for Semantic Data," in Proceedings of the Posters and Demos Track of the 14th International Conference on Semantic Systems *co-located with the 14th International Conference on Semantic Systems* (*SEMANTiCS 2018*), eds. Ali Khalili and Maria Koutraki, 2018, http://ceur-ws.org/Vol-2198/ (accessed 20 December 2021).

20 Dominic Oldman and Diana Tanase, "Reshaping the Knowledge Graph by connecting researchers, data and practices in ResearchSpace," in *The Semantic Web-ISWC 2018. 17th International Semantic Web Conference, Monterrey, CA, USA, October 8–12, 2018, Proceedings*, eds. Denny Vrandečić, Kalina Bontcheva, Mari Carmen Suárez-Figueroa, Valentina Presutti, Irene Celino, Marta Sabou, Lucie-Aimée Kaffee and Elena Simperl (Cham: Springer, 2020), 325–340.

21 Haase, Peter, Daniel M. Herzig, Artem Kozlov, Andriy Nikolov, and Johannes Trame, "Metaphactory: A platform for knowledge graph management," *Semantic Web* 10, no. 6 (2019): 1109–1125.

5.1 *Data Browsing*

Fig. 17.20 shows how the user can start browsing the data of the semantic network. The interface displays the main types of entities in a left menu (Ships, Voyages, Persons, Organizations, Places, Crew Employments, and Payments). By selecting to view one of these entity types, the user is shown a table with all instances together with additional information about each instance. In particular, for a ship, the table shows the following information: ship type, tonnage, registration port, registration number, construction location, construction date, owner. For a voyage, it shows the name of the ship, the destination, the starting and ending date, its duration and the name of the corresponding data source. For persons, it shows: first name, last name, father name, serial number, birth date, and residence place. For an organization, it shows its name and location. For a place, it shows the type (e.g., region, port, etc.) and its coordinates, while for crew employments and payments it shows: employee name, serial number, recruitment place, starting date, discharge place, ending date, profession, ship, pension fund, wage, duration, and data source name. For places, the user can also inspect a map of all places referred to in the archival sources, thereby allowing them to get an overview of the covered areas. In SeaLiT, for instance, almost all places are in Europe.

The user can select one of the entity instances shown in the table and start exploring the associated data, taking advantage of the Linked Data concept. Fig. 17.21 shows an example where the user has selected to browse the data of the voyage of the ship *Mardocheo*. We can inspect information like the timespan of the voyage and the crew members, while, by selecting to explore further one of the displayed entities (e.g., one of the crew members), the user can start browsing information about that entity. Since the ship voyages are connected to places (like embarkation ports), the user can also inspect a map with all these places, as shown in Fig. 17.22. For example, we see that the voyage of the

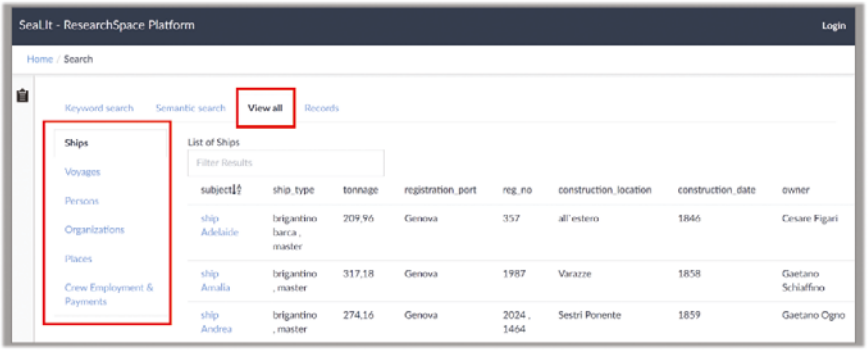

FIGURE 17.20 Browsing the main entities that appear in the transcribed and curated data in ResearchSpace

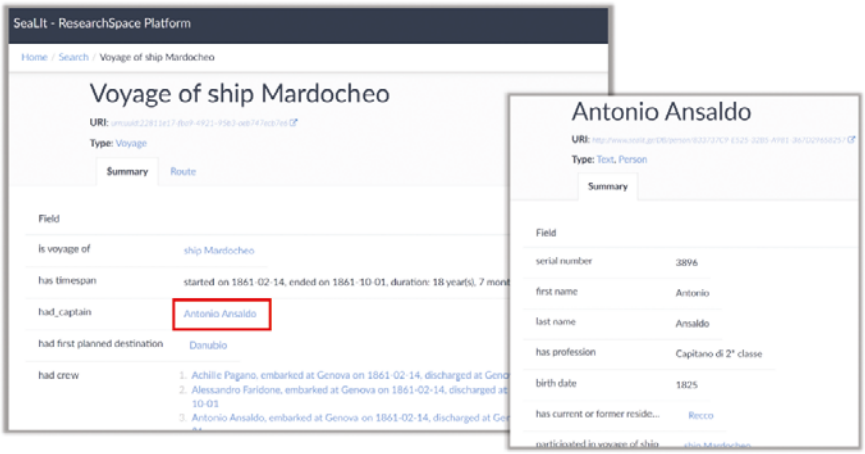

FIGURE 17.21 Browsing the data of a ship voyage in ResearchSpace

ship *Mardocheo* is connected to four places (Spezia, Malta, Constantinople, and Sulina). Each place spot on the map provides information in the context of the current voyage, like the departure and/or the arrival date, as well as useful links for further exploration of the place.

5.2 *Assistive Query Building*

The second data exploration functionality offered by ResearchSpace is an assistive query building interface, called "Semantic Search". The principal idea behind this mechanism is the definition of the fundamental entity categories and relationships among these categories.[22] A relationship between two entity categories is defined as a path that connects the entities in the semantic network. Based on that, the user can define a complex information need (query) that involves different types of entities and associations. Specifically, the user is guided to build the query through an intuitive user interface and then can inspect the results through a variety of visualization methods, including table, charts, or in a map.

Fig. 17.23 displays an example of how we can build a semantic search query. In this example, the user first selects the entity type "Person" as the target searching entity ("I am looking for persons"). The user is then shown a list of entity types that are associated to persons, and selects "Ship" and the

22 Katerina Tzompanaki and Martin Doerr, "Fundamental Categories and Relationships for intuitive querying CIDOC-CRM based repositories," *ICS-FORTH Technical Report 429* (2012).

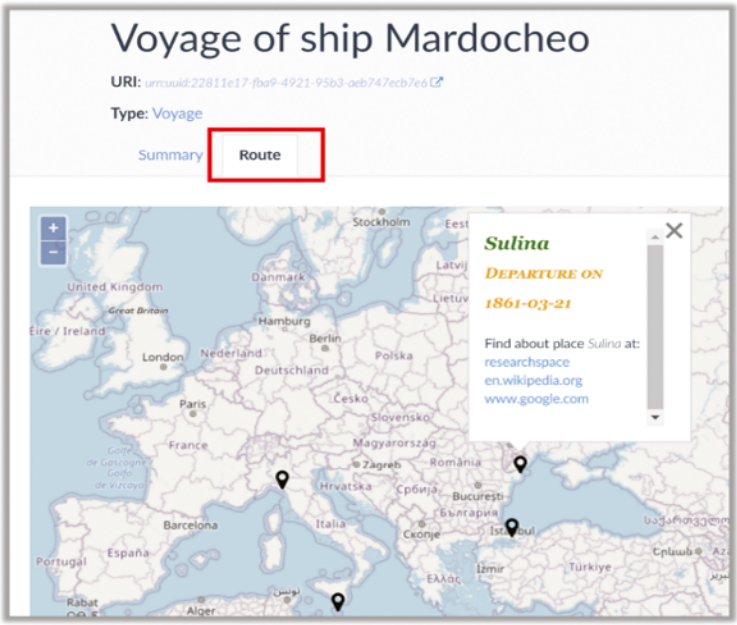

FIGURE 17.22 Inspecting the places connected to a ship voyage on a map in ResearchSpace

relationship "was crew at" ("I am looking for persons that were crew at ships"). Then, the user starts typing the ship name "*Andrea*" and selects the instance "Ship *Andrea*" from the list ("I am looking for persons that were crew at the ship *Andrea*"). In a similar way, the user can further connect ships to an arrival place or other types of entities related to ships, or connect the initial target entity persons to additional types of entities (e.g., birth place or birth date). If the entity type is a location, i.e. it has coordinates, the user can also set the area of interest directly on the map (Fig. 17.24).

Fig. 17.*25* shows an example of a semantic search query where the user has selected to search for "persons that were crew at ships that arrived at Livorno". The results (21 person instances) are displayed in the same page, while the user has different options on how to visualize them. By default, the user is shown a list of the matched instances, together with additional information for each instance (Fig. 17.*25*; lower part). If we are searching for persons, for example, the results show the following information for each person instance: first name, last name, father name, serial number, birth date, and residence location. Here the user can start browsing the data of a particular instance and inspect further information related to the selected instance, as previously described.

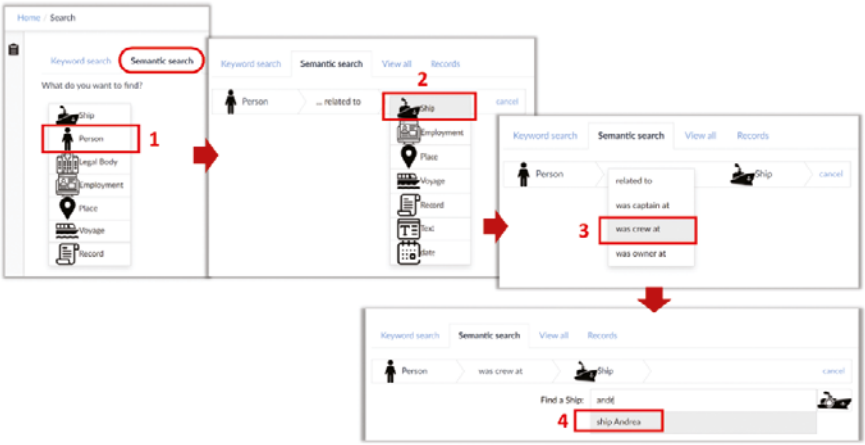

FIGURE 17.23 An example of how the user can build a semantic search query in ResearchSpace

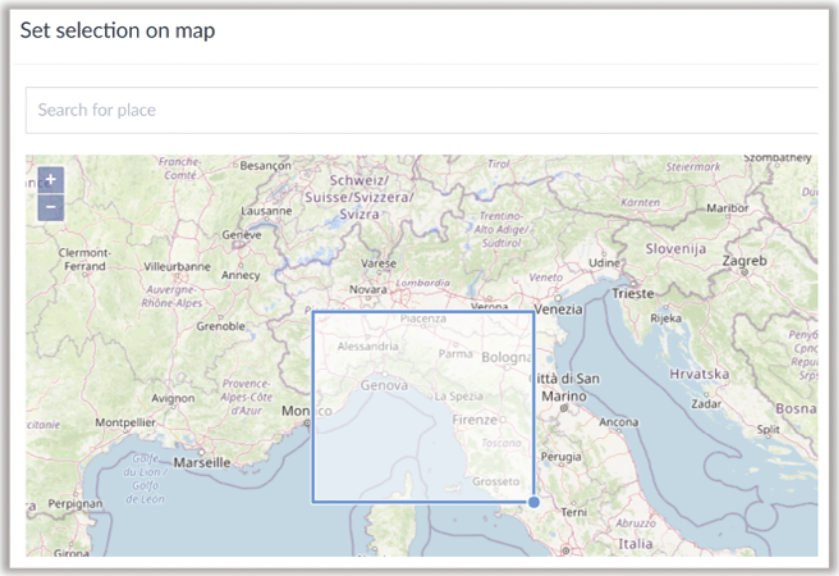

FIGURE 17.24 Selecting an area of interest directly on the map in ResearchSpace

The user can also select to visualize the results in a chart, like a bar chart or a pie chart, by first selecting the visualization context to consider aggregating the results. The options here are different for each type of entity. For example, if the search results are persons, we can select visualization contexts like

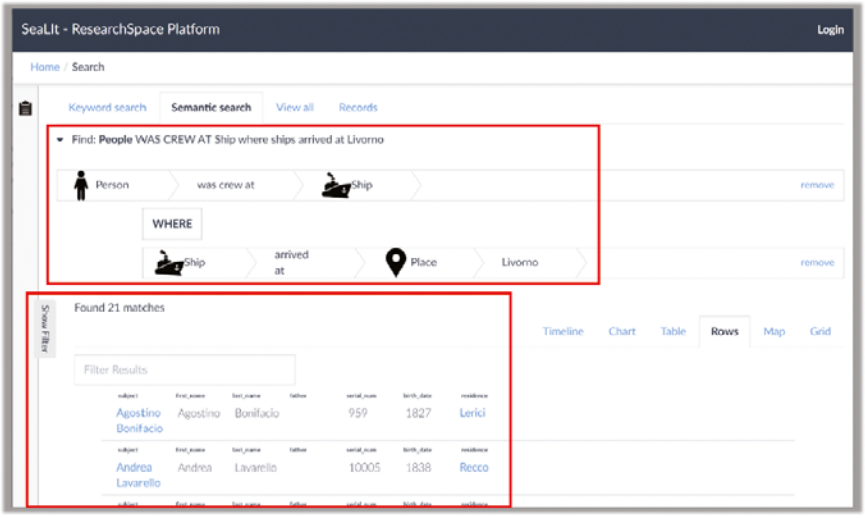

FIGURE 17.25 An example of a semantic search query and its results in ResearchSpace

"embarked at (Place)", "embarked in (Date)", "discharged at (Place)", "location of residence (Place)", "was crew at (Ship)", "was captain in (Ship)", etc. Fig. 17.26 depicts an example in which the user selected the visualization context "embarked at (Place)". We see that, in this example, the embarkation places are four (Genova, Livorno, Constantinople, Marsiglia), while the majority of crew members embarked at Genova.

Fig. 17.27 shows another example in which the user is searching for places. In particular, the query searches for "places that were the construction place of ships of type brigantino". In this case, the results can also be visualized on a map, since the target entities are locations and the coordinates are available.

The results of a semantic search query, as well as the data of a chart diagram, can be downloaded in a CSV format, which is compatible for use in a spreadsheet application like Excel (see the download icon and the "Download CSV" button in Fig. 17.26). This means that the user can download the data, open it in an application like Excel, and make modifications, further analysis, or create other charts.

Another functionality offered by ResearchSpace is the "filter" option which allows filtering of the displayed results based on one or more entity properties or associations. For instance, if the displayed results are persons, the user can select to show only persons who were crew members in a particular ship, or who had a particular place as residence. Fig. 17.28 shows an example in which the user has set the filter "was born in 1830–1838 AD". By setting this filter, the

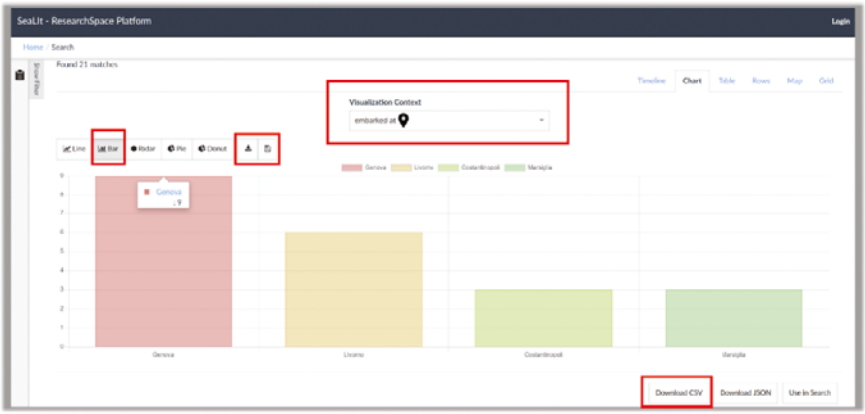

FIGURE 17.26 Inspecting a bar chart of the results in ResearchSpace by selecting a particular visualization context

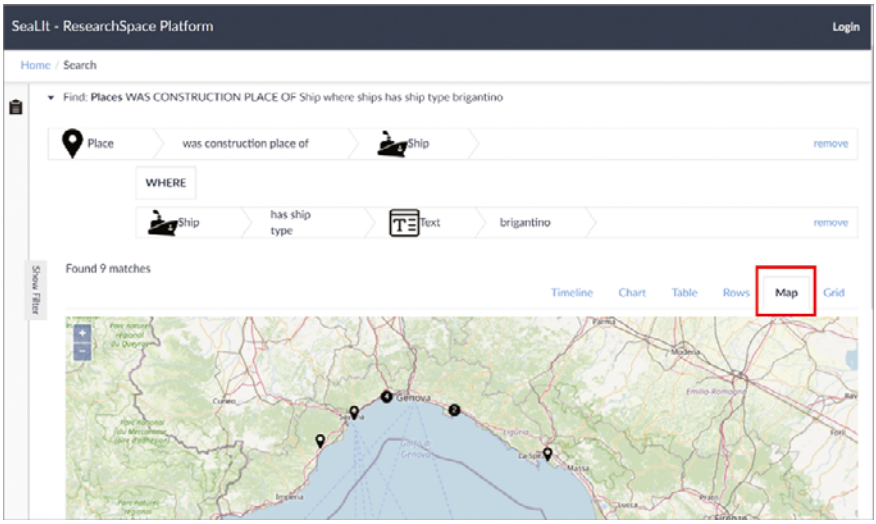

FIGURE 17.27 Visualizing the results on a map in ResearchSpace

displayed persons (returned by the semantic search query) are filtered to only those with a birth date between 1830–38 (nine in total). The user can add more filters on demand, or remove an already applied filter. After any such action, the displayed results are instantly updated to reflect the selected criterion. The filters can also help the user to better understand the distribution of the results in the different filter values. For example, when inspecting the values of the filter "had location of residence" (Fig. 17.29), we notice that most of the persons

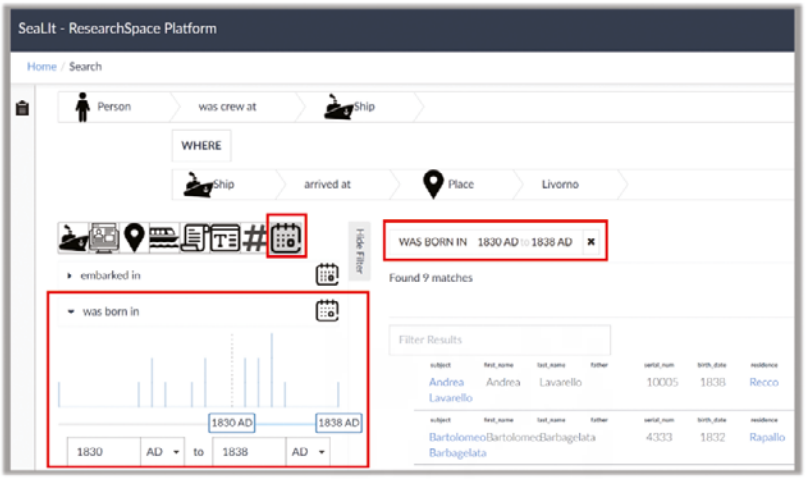

FIGURE 17.28 Filtering the displayed results based on an entity property in ResearchSpace

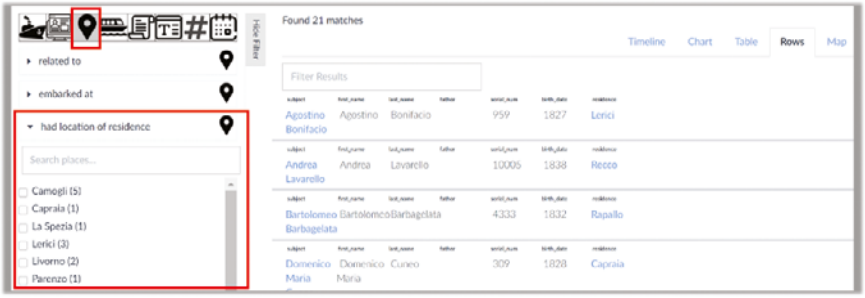

FIGURE 17.29 Inspecting the values of the filter "had location of residence" in ResearchSpace

are from Camogli, while we also see the name of the other cities together with the corresponding number of persons.

6 Conclusions

We have presented a process and a set of Web-based tools for the holistic management of archival sources of maritime history. First, the historical sources can be transcribed and curated collaboratively by historians using the FAST CAT system. FAST CAT supports innovative features, like nested tabular structures for data entry, embedded instance matching and vocabulary

maintenance processes, as well as provenance-aware data curation and enrichment. The transcribed and curated data are then transformed and integrated into a rich semantic network (RDF graph) using the X3ML data transformation framework and a specially-designed ontology, called "SeaLiT Ontology", which is compatible with existing standards for cultural heritage documentation, in particular CIDOC-CRM. Finally, the derived semantic network can be explored using the ResearchSpace platform. ResearchSpace offers a user-friendly interface for browsing the integrated data of the semantic network, as well as building complex queries and visualizing the results in different forms, like tables, charts, or in a map. All described tools are configurable, in the sense that they can be used for digitizing, curating, and exploring other data sources, beyond the case of maritime history.

Acknowledgements

This work has received funding from the European Union's Horizon 2020 research and innovation programme under: i) the European Research Council (ERC) grant agreement No. 714437 (Project "SeaLiT—Seafaring Lives in Transition. Mediterranean Maritime Labour and Shipping During Globalization, 1850s–1920s"); and ii) the Marie Sklodowska-Curie grant agreement No. 890861 (Individual Fellowship, Project "ReKnow—Research Knowledge Documentation, Analysis and Exploration in Empirical and Descriptive Sciences").