This is a preprint of a paper accepted for publication at the 2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL)

FastCat Catalogues: Interactive Entity-based Exploratory Analysis of Archival Documents

Georgios Rinakakis
rinakakis1999@gmail.com
Information Systems Laboratory, ICS-FORTH &
Department of Computer Science, University of Crete
Heraklion, Greece

Kostas Petrakis
cpetrakis@ics.forth.gr
Information Systems Laboratory, ICS-FORTH
Heraklion, Greece

Yannis Tzitzikas
tzitzik@ics.forth.gr
Information Systems Laboratory, ICS-FORTH &
Department of Computer Science, University of Crete
Heraklion, Greece

Pavlos Fafalios
fafalios@ics.forth.gr
Information Systems Laboratory, ICS-FORTH
Heraklion, Greece

ABSTRACT

We describe FastCat Catalogues, a web application that supports researchers studying archival material, such as historians, in exploring and quantitatively analysing the data (transcripts) of archival documents. The application was designed based on real information needs provided by a large group of researchers, makes use of JSON technology, and is configurable for use over any type of archival documents whose contents have been transcribed and exported in JSON format. The supported functionalities include a) source- or record-specific entity browsing, b) source-independent entity browsing, c) data filtering, d) inspection of provenance information, e) data aggregation and visualisation in charts, f) table and chart data export for further (external) analysis. The application is provided as open source and is currently used by historians in maritime history research.

CCS CONCEPTS

• **Information systems → Digital libraries and archives**; **Search interfaces**.

KEYWORDS

archival research, archival data search, exploratory data analysis, archival data browsing, entity-based archival search

1 INTRODUCTION

Archival research is a type of research which involves investigating and extracting evidence from original archival material, such as historical documents [13]. These documents often have a repetitive structure, providing information about one or more *categories of entities*, such as persons or locations. Examples include logbooks, payrolls, censuses, civil registers, employment records, etc.

Researchers that study such documents usually start by first transcribing the important archival information in digital form. This enables them in performing exploratory analysis on the transcribed data, such as quantitative analysis of collective phenomena for drawing conclusions on possible impact factors [11].

For example, consider the real use case of the SeaLiT project (ERC project in the field of maritime history)[1], which studies the transition from sail to steam navigation and its effects on seafaring populations in the Mediterranean and the Black Sea between the 1850s and the 1920s [2]. Maritime historians have collected and transcribed a large and diverse set of archival documents (>600 records, 20 different types of sources, 5 languages), such as crew lists, payrolls, logbooks, and sailor registers, gathered from multiple authorities in different countries. These documents provide historical information for thousands of interconnected entities, including ships, captains, sailors, ship owners, departure ports, embarkation ports, residence locations, etc.

To effectively explore and analyse this amount of data, and extract useful information for their research, historians need interactive and intuitive-to-use interfaces that can support them in finding the desired information. To this end, in this paper we present FastCat Catalogues, a web-based system that allows the entity-based exploration and analysis of archival documents. The system has been designed by considering *real* information needs provided by a large group of historians. The aim is to provide a system that can support historians in finding answers to their information needs.

For a selected entity (e.g. a ship), the user can inspect its connection to other entities (e.g. crew members, voyages, departure ports, arrival ports), or directly visit the original transcripts that mention this entity for validation or retrieval of further (contextual) information. The user can also group a list of displayed entities by one of their characteristics and visualise the result in a chart (such as grouping the ship's crew members by their residence location).

Through a dedicated configuration model that allows defining the entities of interest and their relation to other entities per source type, the application can be configured for use over any type of archival documents whose transcripts are exported in JSON format.

The application is available as open source[2] and has been deployed considering data of the SeaLiT project. The deployment is publicly available in the following link:

https://catalogues.sealitproject.eu/

The rest of this paper is organised as follows: Section 2 discusses related work. Section 3 describes the design requirements. Section 4 details the system's functionality, user interface (UI) and technology. Section 5 presents the system's configuration model. Section 6

[1] https://sealitproject.eu

[2] https://github.com/isl/FastCat-Catalogues

discusses its evaluation and use in a real context. Finally, Section 7 concludes the paper and outlines future work.

2 RELATED WORK AND INNOVATION

The majority of existing works supports searching archival documents based on metadata and/or textual search on their contents. Such systems are usually based on search engines like Elasticsearch[3] or Apache Solr[4], e.g. [12] and [7].

The innovation of FastCat Catalogues compared to these systems lies on the fact that it allows browsing and exploring information about a set of *highly-interrelated* entities (e.g. sailors, ships, ports, locations, etc., in the case of SeaLiT), where the type of the relation between the entities is highly dependant on the entities context in the archival documents (e.g. a location can be a departure port for a ship, a destination port, a port of call, etc.).

Another approach is to make use of semantic technologies for representing archival/historical data as a rich knowledge graph of linked data [4, 5, 8, 9]. Then, user-friendly interfaces aim at offering an intuitive way to the end users for accessing and exploring the data (e.g. [6, 10]). Such an approach allows specifying the semantic relations among the involved entities as well as linking the data to external datasets (if needed). Nevertheless, it requires extensive conceptual modeling work for representing the data ontologically, while offering a reliable, robust and efficient service over such semantic repositories is a challenging task. Semantifying the application is part of our future work, e.g. either by considering RDF triples as the input data, and/or by supporting the extraction of the displayed data to RDF.

3 REQUIREMENT ANALYSIS

We have collected a set of more than 100 information needs (*competency questions*) from a group of around twenty historians belonging to different research groups in five countries and working with archival documents in the context of the SeaLiT project. These information needs are directly related to the archival material, i.e. its analysis can provide answers to the information needs or important related evidence. Indicative examples are:[5]

- What are the places of construction of ships during a specific period?
- What are the most popular European destinations in different time periods of ships departed from ports in the Black Sea?
- How many ship owners per ship during a specific time period?

We tried to group the information needs in categories and used them as requirements for the design of the application. We identified the below four main categories: (i) finding information about a particular entity, such as the birth date and place of a person; (ii) retrieving a list of entities based on one or more properties of these entities, together with additional information about them, e.g. all persons having a specific residence location together with their birth date; (iii) grouping a list of retrieved entities based on some property or characteristic, e.g. grouping all retrieved persons by their profession; (iv) finding comparative information related to some entities, e.g. number of persons employed by the organisation in different time periods, or voyage duration per type of ship.

The aim is to design a system that can support humanities researchers studying archival material in finding answers to these categories of information needs.

Another important requirement highlighted by the historians is the ability to find the provenance of any piece of displayed information, by enabling end users to directly visit the original transcript that contains the information. This is very important for historians because it allows finding additional contextual data quickly, but also for verification purposes.

4 UI, FUNCTIONALITY AND TECHNOLOGY

Fig. 1 shows the home page of FastCat Catalogues, as deployed for the case of the SeaLiT project. There are two tabs for exploring the archival data: 'Explore by Source' (Fig. 1-A) and 'Explore all' (Fig. 1-B). The former (default option) shows all types of sources (and their number of records) grouped in categories (Fig. 1-C), allowing the user to select one and start exploring its entities. The latter shows a set of categories of entities (like persons, locations, etc.), allowing users to explore entities across sources (Fig. 3-A). The data in both cases is dynamically loaded based on the configuration made to the application (more below).

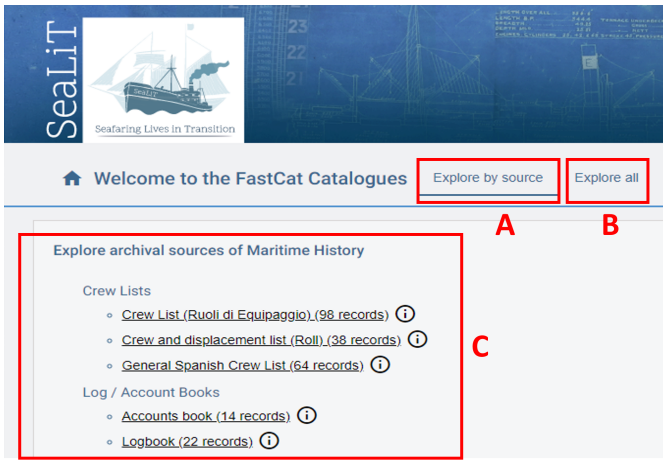

Figure 1: The home page of FastCat Catalogues as deployed for the case of SeaLiT.

4.1 Explore by source

By selecting a specific type of source in the 'Explore by Source' tab, the user first gets an overview of the categories of entities (and their number) that exist in all records of this source type (Fig. 2-A). The user can also filter the displayed information by selecting a specific record (Fig. 2-B). By selecting a category of entities, the user is shown with a table containing all instances of the selected category (Fig. 2-C). The user can filter the instances in the table by adding a filter in one or more of the table columns. Here there are different filtering options depending on the column type, e.g.

[3] https://elastic.co
[4] https://solr.apache.org
[5] The full list of gathered information needs is available at https://users.ics.forth.gr/~fafalios/SeaLiT_Competency_Questions_InfoNeeds.pdf

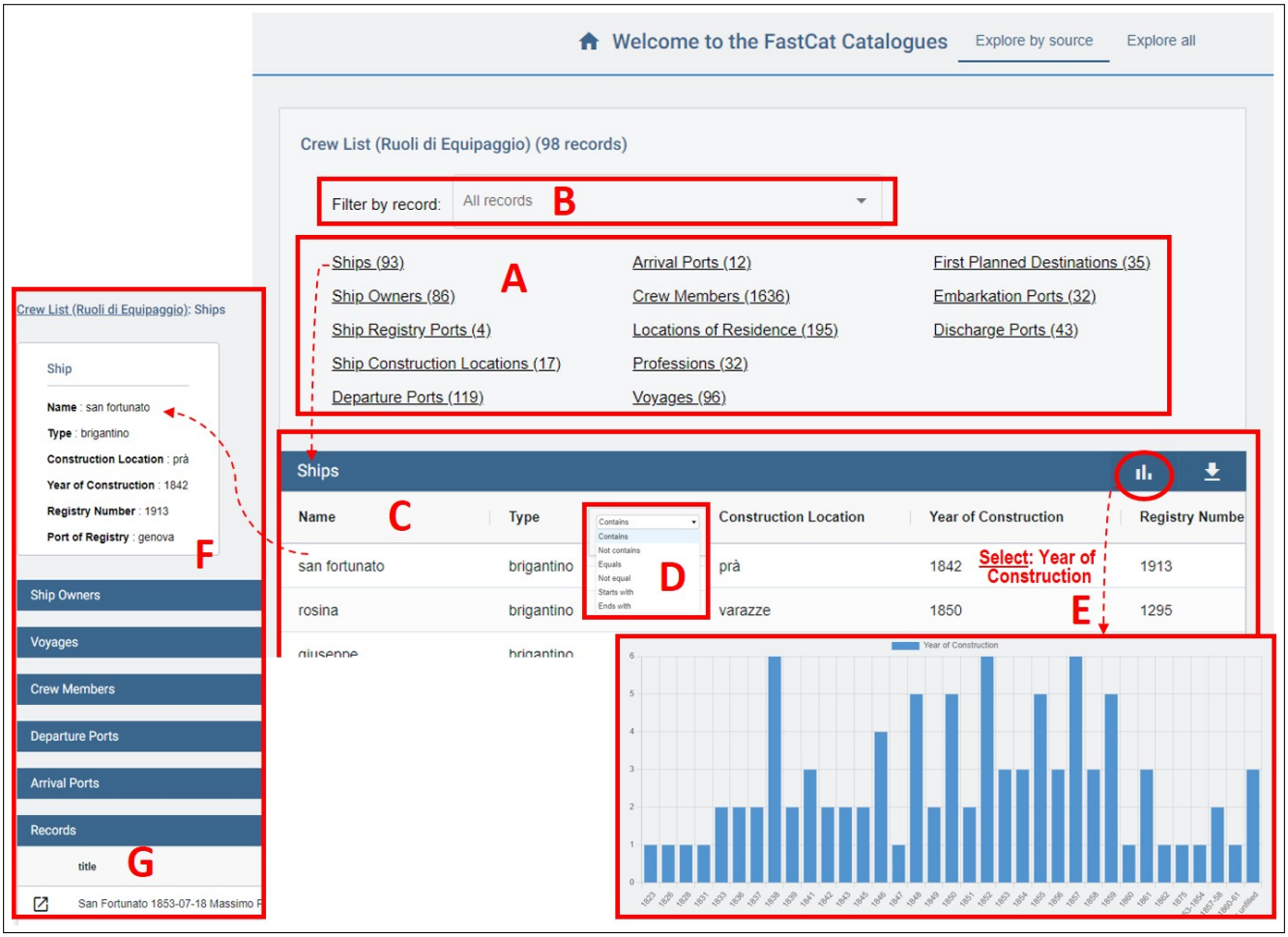

Figure 2: Exploring the entities of a specific type of source and visualising aggregated information in a chart.

contains, not contains, equals, not equal, starts with, ends with for string values (Fig. 2-D). Also, the entity instances shown in a table can be grouped by selecting a specific column (corresponding to an entity property) and visualised in a chart (Fig. 2-E).

We should mention here that, if the original record does not contain a value for an entity property (e.g. there is no construction location for a ship), we display the value *'None or unfilled'* in the tables or charts. This is very important for researchers for getting valid information and making safe conclusions.

By selecting an entity instance in the table of entities, the user can inspect the connections of the selected entity with other entities and start browsing the related information (Fig. 2-F). In the case of SeaLiT, for instance, for a ship of the source type 'Crew List', the user gets tables showing the ship's owners, voyages, crew members, departure ports, and arrival ports. Then, by selecting, for example, a departure port, the user can see all ships that have the same departure port together with their departure dates, and so on. In all these tables the user can apply filters, group the entities by one of its properties and show a chart, or export the data in CSV format for further (offline) analysis.

For a selected entity, the user also gets a table with all records (transcripts) that mention the entity with the ability to directly visit them (Fig. 2-G).

4.2 Explore Across Sources

The 'Explore all' tab allows users to explore the entities across all sources. The user is first shown a list with the different categories of entities mentioned in the transcripts of all sources (Fig. 3-A). By selecting a particular entity category (e.g. ships), the user is shown a table with all its instances (Fig. 3-B). The columns shown in the table for an entity category is the union of all the columns shown in the tables of the sources that contain the same entity category. If a source does not provide a value for a column, then we set 'n/a' in the corresponding cell.

By selecting an entity instance from the table, the user is shown with all sources that mention the entity (Fig. 3-C). Then, by selecting one of the displayed sources, the user is redirected to the corresponding 'Explore by source' page which shows the entity's connections with other entities in the selected source.

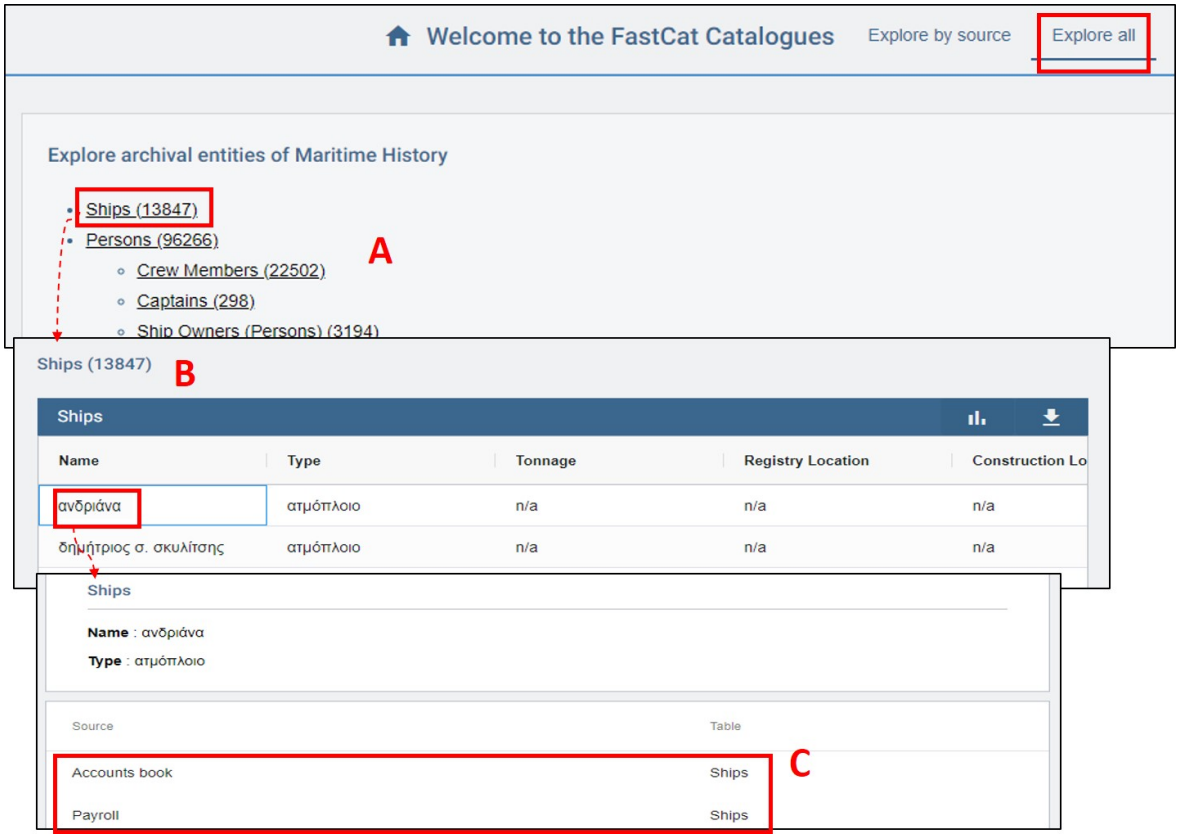

Figure 3: Exploring an entity across sources.

4.3 Technologies

The front-end has been implemented in Angular[6] and the back-end in Express.js[7], a Node.js framework for creating APIs. For displaying the data in tables with filtering and ranking capabilities we use ag-grid[8], while for visualising the table data in chart we use Chart.js[9].

5 CONFIGURATION MODEL

The application can be configured for use over any type (structure) of transcripts. It operates over a set of JSON files, each one containing the transcribed data of a single archival document. The JSON files are organised in folders, where each folder contains files of the same JSON structure (representing a specific archival source type).

For configuring the system, we first need to define the *templates.json* file. In this file, we provide for each different type of archival source: i) a category name (used for grouping the different types of sources), ii) an ID (used for creating the links to the original transcripts), iii) a name (shown in the UI), iv) a description (shown in the UI), and v) the name of its configuration file.

In the configuration file of each source, we define the entity categories (e.g. persons, ships, etc.) that appear in records of this source type and which will be available for exploration. For each entity category, we define the JSON fields that provide entity-related information like properties of the entity or its relations to other entities.

Finally, for configuring the 'Explore all' functionality, we first need to define the names of all the supported entity categories and their grouping (in the *explore_all.json* configuration file). Each of the entity categories can then be configured by defining the sources and the tables in each source that provide instances (in the *explore_all_conf.json* file). All other information needed for creating the entity tables is read from the source-specific configuration files.

This type of configuration allows the setup of the application for use over any JSON structure, making it independent to the system used for data transcription. Details on how to prepare the configuration are available at the system's GitHub repository (Footnote 2).

6 EVALUATION AND USE

6.1 Requirements Satisfaction

We evaluated the application in terms of how it can support finding answers to the four categories of real information needs described in Section 3, as well as its ability to satisfy the requirement about the provenance of the displayed data. Specifically:

(i) Finding information about a particular entity. The user can first select the category of the searching entity (Fig. 2-A or Fig. 3-A) and

[6]https://angular.io/
[7]https://expressjs.com/
[8]https://ag-grid.com
[9]https://chartjs.org

then detect the desired entity by using the table filtering options (Fig. 2-D). By selecting the desired entity, the user can explore information about it (properties, related entities, etc.) (Fig. 2-F).

(ii) Retrieving a list of entities based on one or more properties of these entities, together with additional information about them. The user can select a category of entity and then apply filters in the columns of the displayed table for defining the desired entity properties (Fig. 2-D).

(iii) Grouping a list of retrieved entities based on some property or characteristic. As in (ii), the user can select a category of entity and then apply filters in the columns of the displayed table for defining the desired entity properties. Then, the entities can be grouped and visualised in a chart by selecting a specific entity property (Fig. 2-E).

(iv) Finding comparative information related to some entities. The aggregated information displayed in a chart considers any applied filters. In this way, the user can inspect charts using different filters each time and thus find comparative information through such multiple interactions. For example, we can filter the list of a ship's crew members based on different residence locations and each time group the persons by their birth year (to check for any difference in the age of crew members for different residence locations).

(v) Finding the provenance of any piece of displayed information. When the user selects an entity from the table of entities (Fig. 2-F), the last table called "Records" shows all records containing the displayed data of that particular entity (Fig. 2-G). The user can directly select a row and be redirected to the corresponding record. Similarly, when the user explores entities across sources and selects an entity (Fig. 3-A,B), she/he is shown with all sources containing it (Fig. 3-C). By selecting one of the sources, the user is redirected in the corresponding 'Explore by source' page which shows the entity's connections with other entities and the corresponding records table.

6.2 Use in Maritime History Research

We have deployed the application in a real context (SeaLiT project) for supporting a large group of historians in exploring transcripts of their archival material. The archival documents studied by historians of SeaLiT consist of crew and displacement lists, logbooks, payrolls, account books, censuses, employment records, notarial deeds, and registers of different types such as sailors registers and naval ship registers. Archival documents covering these types of sources have been transcribed in tabular form by the historians (using the FastCat system [3]) and stored in a JSON database. The total number of transcripts is more than 600, providing information for about 100K persons (sailors, etc.), 2.4K ships, 9.8K locations, and 1.1K legal entities (organisations). The application is publicly accessible for use.[10]

7 CONCLUSION AND FUTURE WORK

We have presented FastCat Catalogues, a web application that supports researchers and domain experts working with archival material in exploring and quantitatively analysing the transcribed data. The application is configurable, provenance-aware and goes beyond searching archival documents based on metadata or textual search on their full contents, by exploiting the interrelations of the entities mentioned in the documents.

The system is currently used in maritime history research, for supporting historians in exploring the data of more than 600 records belonging to 20 different types of archival sources and which provide interrelated historical information for more than 100K entities (persons, ships, locations, etc.).

A current issue is the fact that the same entity may appear under different representations in the archival documents. This can happen due to several reasons, like different language, unrecognisable characters, or difference in an entity's property. Our current work is concerned with the implementation of an effective and efficient solution to this *entity matching/resolution* problem [1]. Another direction for future work is the 'semantification' of the application through the use of semantic technologies for the ontological representation of the data and their linking to external relevant resources.

[10] https://catalogues.sealitproject.eu/

ACKNOWLEDGEMENTS

This work has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No. 890861 (Project ReKnow), and the European Research Council (ERC) grant agreement No. 714437 (Project SeaLiT).

REFERENCES

[1] Vassilis Christophides, Vasilis Efthymiou, Themis Palpanas, George Papadakis, and Kostas Stefanidis. 2020. An overview of end-to-end entity resolution for big data. *ACM Computing Surveys (CSUR)* 53, 6 (2020), 1–42.

[2] Apostolos Delis. 2020. Seafaring Lives at the crossroads of Mediterranean maritime history. *International Journal of Maritime History* 32, 2 (2020), 464–478.

[3] Pavlos Fafalios, Kostas Petrakis, Georgios Samaritakis, Korina Doerr, Athina Kritsotaki, Yannis Tzitzikas, and Martin Doerr. 2021. FAST CAT: collaborative data entry and curation for semantic interoperability in digital humanities. *Journal on Computing and Cultural Heritage (JOCCH)* 14, 4 (2021), 1–20.

[4] Pavlos Fafalios, Georgios Samaritakis, Kostas Petrakis, Korina Doerr, Athina Kritsotaki, Anastasia Axaridou, and Martin Doerr. 2022. Building and Exploring a Semantic Network of Maritime History Data. In *Mediterranean Seafarers in Transition*. Brill, 509–535.

[5] Ashleigh Hawkins. 2022. Archives, linked data and the digital humanities: increasing access to digitised and born-digital archives via the semantic web. *Archival Science* 22, 3 (2022), 319–344.

[6] Vangelis Kritsotakis, Yannis Roussakis, Theodore Patkos, and Maria Theodoridou. 2018. Assistive Query Building for Semantic Data.. In *SEMANTICS Posters&Demos*.

[7] Richard Marciano, Victoria Lemieux, Mark Hedges, Maria Esteva, William Underwood, Michael Kurtz, and Mark Conrad. 2018. Archival records and training in the age of big data. In *Re-Envisioning the MLS: Perspectives on the future of library and information science education*. Emerald Publishing Limited.

[8] Albert Meroño-Peñuela, Ashkan Ashkpour, Marieke Van Erp, Kees Mandemakers, Leen Breure, Andrea Scharnhorst, Stefan Schlobach, and Frank Van Harmelen. 2015. Semantic technologies for historical research: A survey. *Semantic Web* 6, 6 (2015), 539–564.

[9] Dominic Oldman, Martin Doerr, and Stefan Gradmann. 2016. Zen and the art of Linked Data: new strategies for a Semantic Web of humanist knowledge. (2016).

[10] Dominic Oldman and Diana Tanase. 2018. Reshaping the knowledge graph by connecting researchers, data and practices in ResearchSpace. In *International Semantic Web Conference*. Springer, 325–340.

[11] Kostas Petrakis, Georgios Samaritakis, Thomas Kalesios, Enric Garcia i Domingo, Apostolos Delis, Yannis Tzitzikas, Martin Doerr, and Pavlos Fafalios. 2020. Digitizing, Curating and Visualizing Archival Sources of Maritime History: the case of ship logbooks of the nineteenth and twentieth centuries. *Drassana: revista del Museu Marítim* 28 (2020), 60–87.

[12] Kim Pham, Fernando Reyes, and Jeff Rynhart. 2020. Building a Library Search Infrastructure with Elasticsearch. *Code4Lib Journal* 48 (2020).

[13] Marc J Ventresca and John W Mohr. 2017. Archival research methods. *The Blackwell companion to organizations* (2017), 805–828.