

A Workflow Model for Holistic Data Management and Semantic Interoperability in Quantitative Archival Research

Pavlos Fafalios¹, Yannis Marketakis¹, Anastasia Axaridou¹,
Yannis Tzitzikas^{1,2}, and Martin Doerr¹

¹ Information Systems Laboratory, FORTH-ICS, Heraklion, Greece

² Computer Science Department, University of Crete, Heraklion, Greece
{fafalios, marketak, axaridou, tzitzik, martin}@ics.forth.gr

Abstract. Archival research is a complicated task that involves several diverse activities for the extraction of evidence and knowledge from a set of archival documents. The involved activities are usually unconnected, in terms of data connection and flow, making difficult their recursive revision and execution, as well as the inspection of provenance information at data element level. This paper proposes a workflow model for holistic data management in archival research; from transcribing and documenting a set of archival documents, to curating the transcribed data, integrating it to a rich semantic network (knowledge graph), and then exploring the integrated data quantitatively. The workflow is provenance-aware, highly-recursive and focuses on semantic interoperability, aiming at the production of sustainable data of high value and long-term validity. We provide implementation details for each step of the workflow and present its application in maritime history research. We also discuss relevant quality aspects and lessons learned from its application in a real context.

1 Introduction

Archival research is a type of research which involves investigating and extracting evidence from archival records usually held in libraries, museums or other organisations. In its most classic sense, archival research involves the study of historical documents, thus it lies at the heart of original historical research (Ventresca and Mohr, 2017).

A large body of research in the field concerns the study of archival documents that have a *repetitive* structure, such as registers, logbooks, payrolls, censuses, etc., and which provide information about one or more *types of entities*, such as persons, locations, objects, organisations, etc. Research in this case usually starts by first collecting a set of archival documents related to a domain of

interest, which are then transcribed and curated for enabling quantitative (but also qualitative) analysis of empirical facts, their description and interpretation of possible causes, influences and evolution trends (Petракis *et al.*, 2020).

Common data management problems in this context include: What data to transcribe and how? How to curate the transcribed data for enabling valid quantitative analysis and more effective exploration services? How to integrate the data under a common schema/model for supporting the investigation of information needs that require combining data from more than one source? How to support the long-term preservation and reuse of the data? How to maintain all data provenance information, which is important for the verification and the long-term validity of research findings that use the data?

Consider, for instance, the real use case of the SeaLiT project¹ (ERC Starting Grant in the field of *maritime history*), which studies the transition from sail to steam navigation and its effects on seafaring populations in the Mediterranean and the Black Sea (1850s-1920s) (Delis, 2020). Historians in this project have collected and studied a large number of archival documents of different types and languages, such as crew lists, payrolls, and sailor registers, gathered from multiple authorities in five countries. Complementary information about the same entity of interest, such as a ship or a sailor, may exist in different archival documents. For example, for the same ship, one source (*accounts book*) may provide information about its owners, another source (*naval ship register list*) may provide construction details and characteristics of the ship (length, tonnage, horsepower, etc.), while other sources (*crew lists*) may provide information about the ship’s voyages and crew. There might also be another source (*civil register*) that provides additional information about the crew members, such as their marital status and previous professions. Data integration is very important in this context, for supporting historians in finding answers to questions that require combining information from more than one source, such as “finding the nationality of sailors of large ships that arrived at a specific port”.

In addition, the name of the same entity (e.g. of a person) might be different in different sources due to typos, different language, unrecognisable characters, or use of abbreviation (e.g. ‘G. Schiaffino’, ‘Gaetano Schiaffino’, ‘Gaetano Schiaffino’). Moreover, the same term, such as a profession or a ship type, may appear under different names in different sources (e.g. ‘brigantine’, ‘brigantino’). Data curation, in particular entity (instance) matching and term alignment, is crucial in this context for enabling valid quantitative analysis (like grouping a list of retrieved sailors by profession). However, at the same time, such curation must not alter the original transcribed data since this is important for verification and thus the long-term validity of the research findings.

To cope with these problems, in this paper we describe a workflow model for holistic data management in archival research (depicted in Fig. 1). The workflow relies on the strong collaboration between researchers (domain experts) and data engineers (modeling experts), and focuses on *semantic interoperability*, the ability of computer systems to exchange data with unambiguous/shared meaning

¹ <https://sealitproject.eu/>

(Ouksel and Sheth, 1999), because such an approach supports the production of sustainable data of high value that can be extended and re-used beyond a particular research activity or project.

The workflow was designed based on real users' needs and is provenance-aware, in the sense that it retains the full provenance chain of each piece of data. It achieves this by decoupling data entry from data curation and integration. The researcher can go back to the transcript or the original source and inspect the initial form of a piece of information. It is also highly-recursive, supporting the revision of the transcription, curation and integration steps, e.g. due to new knowledge acquired in the course of research. In comparison to related work, we treat the relevant activities in an holistic manner, paying particular attention on maintaining the provenance information at micro (data element) level, which is important for reproducible research in the age of Open Science (Vicente-Saez and Martinez-Fuentes, 2018).

We showcase an implementation of the workflow model in a real use case in the field of maritime history and report empirical results from its application for satisfying real information needs of a large group of historians. We also discuss relevant data quality aspects and lessons learned.

The rest of this paper is organised as follows: Section 2 provides the required background and describes related work. Section 3 provides an overview and the main characteristics of the proposed workflow model. Section 4 details how each step of the workflow model can be realised. Section 5 provides information about the automation of the workflow. Section 6 describes a real use case. Section 7 discusses quality aspects and relevant lessons learned. Finally, Section 8 concludes the paper and outlines future work.

2 Background and Related Work

We first explain the basic notions about semantic technologies (Section 2.1) and review how such technologies are used in humanities research, a large part of which concerns archival research (Section 2.2). We then focus on the different data management activities towards semantic interoperability in archival research and present relevant works (Section 2.3). Finally, we position our work (Section 2.4).

2.1 Basic Notions

Semantic technologies aim at helping machines understanding data. RDF (Resource Description Framework)² and OWL (Web Ontology Language)³ are key semantic technologies that enable encoding the semantics of data, thus allowing to formally represent the meaning involved in information (Antoniou and Van Harmelen, 2004). This representation has the form of a *semantic network*

² <https://www.w3.org/TR/rdf11-concepts/>

³ <https://www.w3.org/TR/owl2-overview/>

(or *knowledge graph*) which stores interlinked descriptions of “entities” (objects, persons, events, concepts, etc.) in a graph structure in which vertices represent entities and edges represent semantic relations between the entities. Typical standardized semantic networks are expressed as RDF triples (statements of the form *subject-predicate-object*) stored in a semantic repository (RDF triplestore) (Ali *et al.*, 2021). Semantic technologies help achieving semantic interoperability, the ability of computer systems to exchange data with unambiguous/shared meaning, which is a requirement to enable machine computable logic, inferencing, knowledge discovery, and data federation between information systems (Ouksel and Sheth, 1999).

2.2 Semantic Technologies for Humanities Research

There is an increasing adoption of semantic technologies in the humanities field, with a main focus on how to produce and make publicly available interoperable *Linked Data* (Heath and Bizer, 2011) that can be easily queried and integrated with other datasets (Hyvönen, 2020; Hyvönen *et al.*, 2014; Hawkins, 2021; Beretta, 2021; Fafalios *et al.*, 2021a).

Oldman *et al.* (2015) provide a critical discussion on how semantic technologies and the idea of Linked Data are used in humanities research, and describe strategies for the wider adoption of these technologies for supporting high-quality digital humanities projects and the production of data that better represents human knowledge and better reflects the needs of humanities researchers. Hawkins (2021) examines how Linked Data about archives is beneficial for those engaged in digital humanities research and scholarship, considering some of the barriers that currently prevent digital humanists from being able to utilise digitised and born-digital archives.

We believe that the workflow model that we propose, in particular its provenance-awareness at data element level, is a first step towards tackling some of the major issues described in the aforementioned works, such as the ability “to trace the provenance of knowledge back to the source micro-level (with its original context and perspective intact)” (Oldman *et al.*, 2015, p.10), or “preventing the decontextualisation and loss of nuance of archives” (Hawkins, 2021, p. 11).

With respect to historical research, for which archival research is a core part, Meroño-Peñuela *et al.* (2015) survey the joint work of historians and computer scientists in the use of semantic technologies. The article provides an extensive analysis on works and systems for knowledge modelling, text processing and mining, search and retrieval, and data integration. It also discusses aspects of semantic technologies that could be furtherly exploited in historical research. Such an aspect is the “non-destructive data transformations” (Meroño-Peñuela *et al.*, 2015, p. 22). Decoupling data entry from data curation and transformation, and maintaining a recursive workflow between these processes, are core characteristics of the proposed workflow model that help towards this direction.

2.3 Data Management for Semantic Interoperability in Archival Research

Common data management activities for enabling semantic interoperability in archival research include:

- digitization / transcription of archival documents (scanning of documents, text recognition, manual transcription)
- documentation / metadata recording (what is the origin of a document, what is the document about, who makes the transcription, etc.)
- data curation / preparing the data for statistical analysis (correction or normalisation of data values, instance matching, term alignment, etc.)
- data integration under a common representation language (ontology-based modeling, creation of mappings, data transformation)
- data publication (e.g. as Linked Data)
- data analysis and exploration (qualitative and/or quantitative analysis, query building, data visualisation, etc.)

There is a plethora of software tools and systems for each of these activities. Below we present relevant works that have a focus on humanities research.

Digitization/Transcription. One can either use text recognition software for automatically extracting text from historical documents, or manually perform the transcription process, each approach having its pros and cons. For example, the automated approach usually needs large amounts of training data and its effectiveness (quality of results) highly depends on the kind/quality of text to be extracted and the amount of training data. On the other hand, manual transcription provides high quality results but it requires a lot of effort. A mixed method is to combine automated extraction with manual correction and data entry. Regarding software tools, Transkribus (Kahle *et al.*, 2017) is a popular platform for the digitisation of historical documents, offering AI-powered text recognition. FastCat (Fafalios *et al.*, 2021b) is a web application for manual and collaborative transcription based on templates. It organises the data (and metadata) in tabular forms (tables), similar to spreadsheets, offering a fast and user-friendly way to data entry.

Documentation / metadata recording. There are two main approaches for documentation towards semantic interoperability: a) decoupling the documentation process from the ontology-based integration and the production of the semantic network, b) creating the semantic network from the very beginning, i.e. during the documentation process. Synthesis (Fafalios *et al.*, 2021a) is a web-based system that applies the first approach for the collaborative and scientific documentation of cultural entities (objects, events, persons, organisations, etc.), offering embedded processes for transforming the data to an ontology-based RDF dataset. ResearchSpace (Oldman and Tanase, 2018) and WissKi (Scholz and Goerz, 2012) are platforms that apply the second approach, supporting the direct ontological representation of (meta)data.

Spreadsheet software, such as Microsoft Excel, and relational database management systems (RDBMS), like Microsoft Access, are still popular (and probably the dominant) tools for (meta)data entry and analysis, and are extensively used for manual documentation and metadata recording. There are also RDBMS-based systems, such as HEURIST⁴ and nodegoat⁵, that are tailored to humanities researchers and which combine a set of functionalities for building and managing research datasets, without however focusing on semantic interoperability.

Data curation. This is an optional step which is usually undertaken when a quantitative (statistical) analysis of the transcribed data is needed. In such a case, curation is very important because data quality can affect the reliability of the analysis results. OpenRefine⁶ is a popular desktop application for data cleaning. It operates on rows of data which have cells under columns (similar to relational tables). Silk⁷ (Volz *et al.*, 2009) is an open source framework for finding links between related data items, e.g. for instance matching. It provides a declarative language for specifying linkage rules and support of RDF link generation, through *owl:sameAs* or other types of links. For fully-automated instance matching (entity resolution), there is a plethora of learning-based methods that require manually or automatically generated training data (Christophides *et al.*, 2020). Finally, the FastCat system (Fafalios *et al.*, 2021b) offers a web-based environment, called FastCat Team, which supports both automated (rule-based) and manual instance matching and vocabulary curation processes. The applied curation does not alter the original (transcribed) data and maintains links from the curated to the original data.

Data integration. The objective here is to semantically represent all data and metadata using a domain (formal) ontology (as the common representation language), in order to enable semantic interoperability and make the data exploitable beyond a particular research problem or project. This activity includes the *data modeling* and *data transformation* processes. Data modeling consists of defining or selecting the domain ontology and creating the schema mappings, while data transformation transforms the data based on the schema mappings and creates the semantic network of integrated data.

Regarding software systems, Protégé is a popular ontology editor which provides a graphic user interface to define ontologies. It can be used for creating a new ontology for a given domain in OWL, or for building an extension of an existing ontology. For the creation and execution of schema mappings, R2RML⁸ is a W3C standard for mapping relational databases into RDF, while Dimou *et al.* (2014) describe an extension called RML for mapping heterogeneous sources into RDF. Finally, the X3ML toolkit (Marketakis *et al.*, 2017) provides a declarative (XML-based) mapping definition language as well as a set of tools for the cre-

⁴ <http://heuristnetwork.org/>

⁵ <https://nodegoat.net/>

⁶ <https://openrefine.org/>

⁷ <http://silkframework.org/>

⁸ <https://www.w3.org/TR/r2rml/>

ation and maintenance of the schema mappings, and the actual transformation of the data to RDF.

Data publication The integrated data can be now imported in a semantic repository (RDF triplestore), either publicly available or private, which offers an Application Programming Interface (API) for accessing the data and running structured queries using the SPARQL⁹ protocol and language. Then, user-friendly applications can be built on top of this API for supporting end users in exploring and analysing the integrated data. The data can be also published as Linked Data, following the Linked Open Data (LOD) principles (Heath and Bizer, 2011). The Sampo model¹⁰ (Hyvönen *et al.*, 2014) provide a framework for collaborative publishing and using of LOD, which has been tested in several domains by building the so-called ‘Sampo portals’ (Hyvönen and others, 2020).

Data exploration and analysis. There are two main general methods that can be used for exploring the integrated data: (a) *free text search*: the user provides a set of keywords or a natural language question, as in ad-hoc information retrieval, (b) *interactive interface*: the user is supported by the system to express an information need, through a user-friendly interactive interface. In both cases the result is (usually) a ranked list of entities from which the user can start exploring relevant information, e.g. through browsing, faceted search, or different visualisations such as charts, maps, timelines, etc.

There is a plethora of different methods for implementing keyword search over RDF data, e.g. using a document-centric information retrieval system (Kadilierakis *et al.*, 2020), or by translating a keyword query to a structured (SPARQL) query (Izquierdo *et al.*, 2021). For the presentation of the keyword search results, Nikas *et al.* (2020) suggest a multi-perspective approach that offers multiple presentation methods (perspectives), allowing the user to easily switch between these perspectives and thus exploit the added value of each one. Regarding interactive interfaces, A-Qub (Kritsotakis *et al.*, 2018) and ResearchSpace (Oldman and Tanase, 2018) offer user-friendly environments which support end users in gradually building complex questions (corresponding to SPARQL queries) that associate different types of entities and information.

2.4 Positioning

To the best of our knowledge, there is no related work that approaches the data management part of archival research in an holistic manner, in the sense that the proposed workflow model enables the representation and efficient management of information, applies semantic data integration facilities in order to provide a rich knowledge graph of archival data, and at the same time it preserves the full provenance chain allowing researchers traverse from the final semantically integrated collection back to the original and transcribed manuscripts and vice versa.

⁹ <https://www.w3.org/TR/sparql11-overview/>

¹⁰ <https://seco.cs.aalto.fi/applications/sampo/>

3 Workflow Model: Overview and Main Characteristics

We first provide an overview of the workflow model (Section 3.1) and then highlight its distinctive characteristics (Section 3.2).

3.1 Roles, Input/Output and Processes

Fig. 1 depicts the proposed workflow model for supporting holistic data management in archival research.

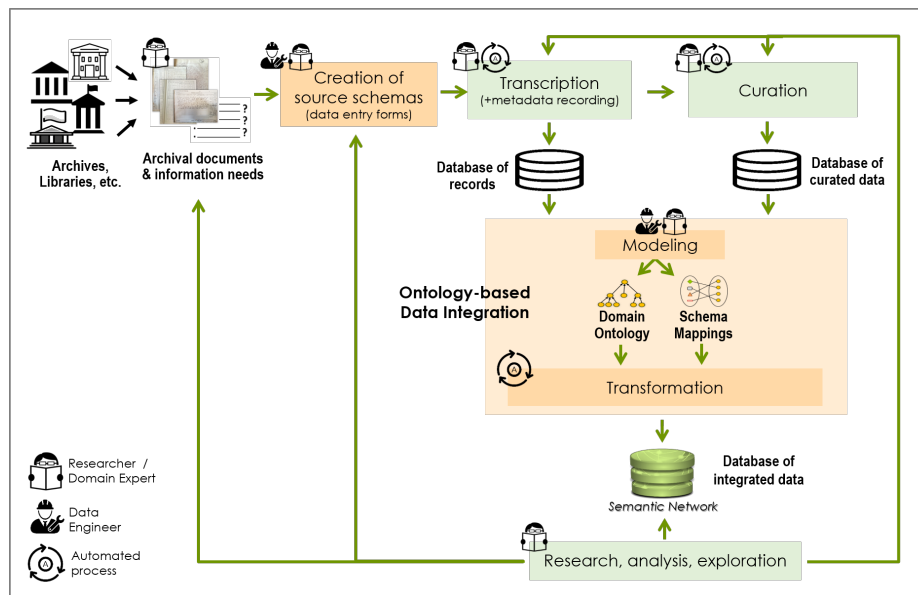


Fig. 1. Workflow model for holistic data management and semantic interoperability in archival research.

Roles. There are two main roles engaged in the workflow: a) the *researcher* (*domain expert / end user*), who collects and studies the archival material, provides domain knowledge, and defines requirements, and b) the *data engineer* (*modeling expert*), who designs and implements the different workflow processes.

Input/Output. The input of the workflow is a set of *archival documents* gathered from different authorities by one or more researchers, together with *information needs* provided by the researchers that are related to their research aims and for which the gathered archival material can provide important information (evidence). The gathering of information needs is very useful in this stage because it allows data engineers to better design and implement the next workflow processes. The output of the workflow is a rich semantic network (a *knowledge graph*) of integrated information, which is used by the researchers for

data analysis and exploration, as well as two distinct, intermediate databases: a database of records (original transcripts), and a database of curated data (curated entity instances and vocabulary terms).

Process 1: Creation of Source Schemas. Following the description logic based framework for information integration as introduced by Calvanese *et al.* (1998), we first need to create the source schemas, one for each different type of source, which provide the required data entry forms in a software system for the transcription and documentation of the original archival documents. This first step enables data curation and consolidation relative to source model semantics, as well as modeling and integration under a common ontology which can be modified in the course of research, without this affecting/delaying the transcription process. The close collaboration between the researchers and the data engineers is very important in this process for properly designing the schemas and avoiding mistakes during data entry that can cause difficulties/limitations in the next steps. An example of such a mistake is the use of a single data entry field for the recording of a measurement unit and value. This is very likely to cause issues to the end user when wanting to perform comparisons during data exploration.

The creation of a new source schema, or a modification/extension of an existing one, will be required if new archival documents of a different type of source are gathered by the researchers and need to be transcribed. This can happen at any stage of the overall pipeline and does not affect the other processes that can run in parallel for the existing gathered material.

Process 2: Transcription. After having created one or more source schemas for the gathered archival documents, the *transcription* of the documents can begin by the researchers using a software system that offers the required data entry forms. Apart from the transcription of the important document contents, this step includes the recording of metadata information for both the documents (archive/library, dating, etc.) and the transcription process (who makes the transcription, etc.). The result of the transcription process is a database of transcripts. This is a task solely performed by the group of researchers, but which can make use of software tools for facilitating/automating transcription, such as text recognition software.

Process 3: Curation. The next step of the workflow is the *curation* of the transcribed data. At this stage researchers need to harmonise the different data elements that appear in the transcripts and resolve identity ambiguities, so that different elements that co-refer to the same real-world entity/concept receive the same identifier, and false co-references are disassociated.

The data elements can be divided into two main categories: (a) *universals*; concept instances that belong to a specific vocabulary or thesaurus of terms, such as professions, object types, etc., and (b) *particulars*; entity instances that belong to specific categories and are accompanied by characteristics/properties, such as *persons* (first name, last name, birth date, etc.), *locations* (name, type, etc.), *organisations* (name, location, etc.). Curation can also include the provision of

corrected/preferred values (e.g. correcting the first name of a person instance) or the entity enrichment (e.g. adding coordinates to a location instance), tasks which are usually important for better data exploitation and visualisation by the external services that operate over the curated and integrated data.

The curation process is a task performed by the group of researchers and may include both manual and automated steps. For example, instance matching of entities, or alignment of vocabulary terms, can comprise both an automated step (based on rules) and a manual step (for validation of ambiguous cases). The result is a distinct database of curated data, with links to the original data elements, which means that the curation step does not alter the data as transcribed from the original sources.

Process 4: Ontology-based Data Integration. The next step is the ontology-based integration of the transcribed and curated data, which includes the *modeling* and *transformation* sub-processes.

For modeling, the good practice suggests to either use an established domain model (if such a model is available for the application domain), or create a new model (a specialised extension) that is compatible to an established upper ontology. This process usually requires extensive discussions between the domain experts, who know the data, and the data engineers, who build the domain ontology and create the mappings.

An important part of the modeling process is the creation of the *schema mappings* that describe how the input data (transcripts and curated data) are mapped to classes and properties of the domain ontology. In general, the creation of the schema mappings can be a time-consuming process when the source schemas are many and large/complex. Nevertheless, it needs to be done only once for each different type of source, while revisions may be required if there are changes in the schemas or the target ontology. The use of a declarative language for defining the mappings, such as X3ML (Marketakis *et al.*, 2017), is recommended because local changes in the sources require local changes in the mapping specifications that are easy to locate and perform.

The *transformation* process takes as input i) the databases (outputs of transcription and curation processes), ii) the domain ontology, and iii) the schema mappings, and produces a rich semantic network of integrated data. This step can be fully automated and can be repeated for any new data sources that are transcribed and curated, as long as there is no change in the transcription schemas.

Process 5: Research, analysis, exploration. The resulting semantic network of integrated data is exploited by the researchers through one or more services that operate over the semantic network and which offer user-friendly interfaces for data browsing, analysis, and exploration. Here it is important for the end users to be able to go back to the transcripts, or even the scans of the original sources, for inspecting the initial form of a piece of information (before its curation and transformation), or for gathering further contextual information. In addition, in the course of research, a user may identify that corrections are needed in the transcribed or curated data, thus researchers need to be able

to revisit the transcription and curation steps, make corrections, and then re-transform (automatically) the data for updating the semantic network. Likewise, new archival documents might be collected at any time, which means that one or more new source schemas and corresponding mappings might need to be created for enabling their transcription, curation and transformation.

3.2 Workflow Distinctive Characteristics

Below we highlight and motivate the distinctive design and methodological characteristics of the proposed workflow model:

- **Strong collaboration between researchers (domain experts) and data engineers (modeling experts).** Such a collaboration is required for a) better designing the source schemas (and the corresponding data entry forms), b) better defining/designing the target (domain) ontology and creating the schema mappings, and c) better creating/configuring the user interfaces of the data exploration service(s).
- **Decoupling data entry from data curation and maintaining links from the curated to the original data.** This is very important not only for maintaining the data provenance, verifying information, and thus validating the research findings that make use of the data, but also because data curation and consolidation may be ambiguous and require further research and repeated revision at any time in the future (by the same or other researchers).
- **Separating source schema creation from ontology modeling.** We aim at removing the bias of the initial research hypothesis from the target (integration) model, one of the most severe philosophical problems of unbiased research and at the core of the discussion about scientific realism (Turner, 2007; Chapman and Wylie, 2018). The target model (ontology) can be developed in parallel with the data entry process and can be re-adapted at any time to new insight from the sources, without invalidating the entered data and without this affecting (or delaying) the transcription and curation processes.
- **Separating the databases (of transcripts and curated data) from the semantic network.** Decoupling data entry and curation from the creation of the semantic network enables maintaining the semantics of the source model by keeping the transcripts as close to the original (archival) document as possible (trying to maintain their original structure), offering at the same time a familiar way to data entry that can highly speed up this time consuming process. In addition, this allows the straightforward production of different versions of the semantic network, considering different ontologies, or different versions of the same ontology (this only requires creating the schema mappings based on the desired target model).

4 How to Realise the Workflow

We now provide implementation details for realising the workflow.

4.1 Faithful, Fast and Collaborative Data Transcription

Common requirements that a data transcription system should satisfy, include:

- Supporting the *faithful* and *structured* transcription of information from the archival documents (as exact to the original information as possible), as well as the recording of *metadata* information.
- Supporting *fast* data entry through an intuitive user interface that researchers are familiar with or can quickly get familiar with.
- Supporting the *collaborative* transcription by more than one researcher, making use of the same structures (source schemas) for data entry.

These characteristics can highly affect the usability of the data entry system and thus its acceptance by the end users (researchers).

For enabling the next *data curation* process, we first need to identify what are the main entity categories (like persons, locations, objects, etc.) and the main vocabularies or hierarchies of terms that appear in the transcribed data and need curation. To this end, we need to define the fields in the data entry forms that provide entity or term related information. For example, the data entry fields *first name* and *last name* provide information for a person instance, while the field *profession* provides a vocabulary term. The values of these fields must be copied (ideally, automatically) to a new environment that allows for their curation without altering the original data as it appears in the transcripts. We then only need to provide a link from the curated to the original data and/or position information (e.g. record name, table name, row number), in order to retain the provenance information.

4.2 Provenance-aware Data Curation

Data curation activities that need to be supported by a dedicated software system include:

- *Correcting* the name of an entity or the value of one of its properties (by setting a preferred label).
- *Instance matching*: matching two or more entity instances that refer to the same real-world entity, which means that they must receive the same identity.
- *Instance unmatching*: unmatching a specific entity instance from a set of automatically matched instances, which means that the instance will receive a different identity.
- *Enrichment*: complementing an entity instance with additional information, like adding coordinates to a location.
- Providing a *preferred term* for a vocabulary term (e.g. a term from a fixed thesaurus, or a term in English for a term in another language).

- Providing a *broader term* for a vocabulary term (thereby creating an hierarchy of terms).

Instance matching in this context can be multi-stage. A first automated step can assign the same identity to a set of entity instances having some common characteristics, e.g. common first name, last name, and birth date, in the case of person instances (rule-based approach), or make use of machine learning techniques (supervised or semi-supervised approach) (Christophides *et al.*, 2020). Then, a second manual step (performed by the researchers) can match additional entity instances that the automated step did not manage to match, or unmatch an entity instance that was incorrectly matched to other instances by the automated step.

The instance matching/unmatching activities and the provision of preferred terms for vocabulary terms are of key importance for valid quantitative (statistical) analysis over the integrated data. Consider, for example, that a researcher who studies archival documents related to maritime history (like crew lists) wants to find the birth place of sailors that arrived at a specific port, or group them by their profession. Providing the same identity to all sailor instances that represent the same real-world person, as well as providing the same ‘preferred’ term for all different professions that correspond to the same profession, ensures that the generated aggregated information is correct.

4.3 Ontology-based Integration

The ontology-based integration of the transcribed and curated data consists of the below tasks:

1. Data modeling using a domain ontology.
2. Creation of schema mappings and definition of how to generate the entity identifiers (URIs).
3. Running the transformations for producing the semantic network of integrated data.

Data modeling. CIDOC-CRM¹¹ (Doerr, 2003) is a high-level, ISO standard ontology (ISO 21127:2014)¹² of human activities, things and events happening in space and time, thus it can be used for modeling the transcribed data and supporting semantic interoperability and long-term data preservation. Depending on the application domain, an extension of CIDOC-CRM might be required for specialising particular notions of interest. For instance, in our use case we created the SeaLiT Ontology, an extension of CIDOC-CRM for the modeling and integration of data related to maritime history (more in Section 6). For semantic data management using CIDOC-CRM, Tzitzikas *et al.* (2022) analyse the relevant processes and tasks, and review the literature on applying machine

¹¹ <http://www.cidoc-crm.org/>

¹² <https://www.iso.org/standard/57832.html>

learning techniques for reducing the costs related to compliance and interoperability based on CIDOC-CRM.

Mapping & Generation of Identifiers. This step defines how the transcribed and curated datasets will be transformed so that they will eventually construct the semantic network. The challenge is to preserve the full provenance chain, from the curated data to the original data of the transcript of the source, so that researchers can easily validate, further improve, or seek for further information. The first part of this step is the definition of the schema mappings, that identify which parts from the input schema (e.g. a particular table column) will be mapped to concrete classes and properties of the domain ontology, ensuring that the semantics of the original data are well-defined, non-ambiguous, and no data is lost. The second part defines how resource URIs and labels will be generated. At this point URIs will be used as the ‘glue’ connecting relevant pieces of information.

Fig. 2 shows an indicative example on how URIs are used for establishing such connections. In this example there are two different transcription records, each one of them describing various persons. In one of them there is a person called ‘Agostino B??ndi’ (i.e. the question marks reveal that the characters could not be recognised from the original source), and in another one there is a person called ‘A Brondi’. For these persons two different URIs are created, since their names do not match and also they are found in different records. However in the curated dataset, historians agreed that these references point to the same person. Therefore a new person instance is created, with a new URI, linked to the previous ones. This new instance is called ‘master’, while the linked instances are considered ‘local’.

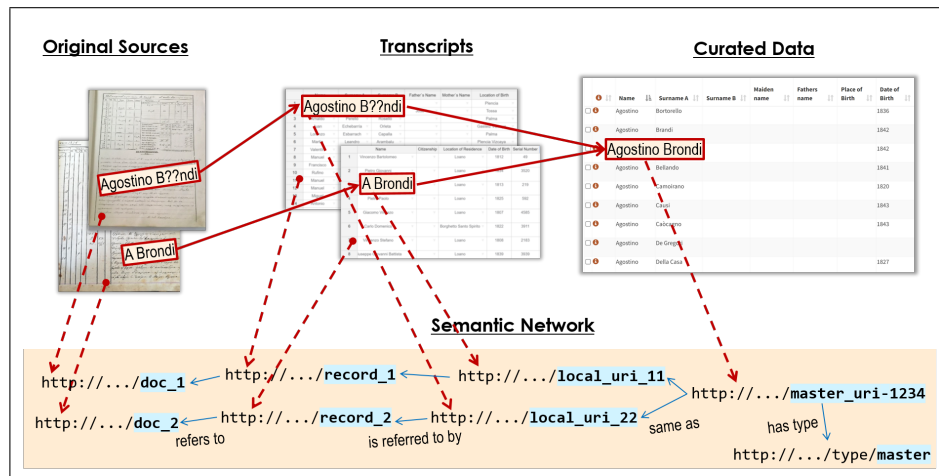


Fig. 2. Identity (URI) management and provenance chain.

Each URI consists of three parts: (a) the URI prefix which is common for all the resources, (b) the type or hierarchy of the resource, (c) the actual or hashed content of the resource. An indicative URI is: *https://rs.sealitproject.eu/kb/location/sardinia*. We should also mention, that there are cases where the aforementioned strategy is not applied. An indicative case is the construction of intermediate nodes in the semantic network, for which a URI is not required (e.g. the ‘E67 Birth’ event). In such cases a random UUID is assigned for them.

Transformation. This step takes as input (a) the transcribed and curated datasets and (b) the definitions of the schema mappings and URI generators, and produces the ontological instances (RDF triples) with respect to the domain ontology, that are the core contents of the semantic network. This step does not require any human intervention and can be fully automated. One apparent advantage of this automation is that the semantic network can be fully or partially refreshed as soon as new data have been transcribed and/or researchers have curated more data.

A good practice for managing the semantic data in terms of updating and versioning flexibility is the use of *named graphs* (Carroll *et al.*, 2005), one for each source record. When there is a new version of a record, or of its mapping definition file, the record output produced with a new workflow cycle can be easily integrated in the semantic repository by replacing the RDF data in the corresponding named graph. Also, the hierarchies of terms and locations can be effectively managed and updated in distinct named graphs, as well as the result of the *materialisation* process for semantically inferred statements (the production of new RDF triples as shortcuts that represent long paths, for improving query performance).

4.4 Semantic Network Exploitation

The integrated data of the semantic network can be now exploited as a primary source for archival research. This includes finding answers to complex information needs and analytical queries that require combining information from different sources, as well as visualising the results in different forms, such as tables, charts, timelines, or maps, for direct use in research.

The actual information needs depend on the application domain and the type of exploration or analysis needed by the end users. The challenge here is to provide researchers with user-friendly and intuitive-to-use interfaces that they can trust for expressing their information needs and findings relevant information. Thus, the key success factors of such data exploration services are usability and trustworthiness. The latter can be achieved by enabling users to directly inspect the provenance of the displayed information, by allowing them to directly visit the transcript containing the information, or even a scan of the original archival document.

Some general categories of information needs include: (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 (e.g.

all persons having a specific residence location); (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).

Finally, a strategy on how to handle missing values in the data, which is very common for certain types of archival documents, is very important in order to get valid aggregated information and make safe conclusions. For example, the residence location for some persons might be empty in the original document. When grouping a set of persons by their residence location, there must be an ‘unknown’ value for this missing information.

5 Workflow Automation

The systems used in the transcription and curation processes need to intercommunicate for automating the copy of the data elements (entities, terms) that need curation from the transcription system to the curation system. Then, an important part of the workflow can be fully automated as long as the modeling process has been completed and the mappings for all different source schemas have been created (tasks that need to be done *once* for each different type of archival documents). In this case, new transcribed and curated data can be automatically transformed and imported in the semantic repository of integrated data, and thus directly be explored by the end users through the data exploration application.

Specifically, the workflow scenario is the following: a group of researchers have collected a first set of archival documents and the data entry forms have been created in a dedicated system for each different type of source. The researchers start the transcription process. When transcription has been completed for the collected set of archival documents, the data elements that need curation are automatically copied to the curation environment and researchers start curating them. At the same time, data engineers, with the support (domain knowledge) of the researchers and by studying the available material evidence and the experts’ requirements, define the target (domain) ontology and create the schema mappings for each different type of source. When both the transcription and curation processes have been completed for all (or a large set) of the archival documents, and the corresponding schema mappings have been created, researchers can ‘publish’ the data, which means that the transformation process is executed and the semantic network is created and ingested in a semantic repository. Researchers can then start exploring the integrated data through the user-friendly interface of an application that operates over the semantic repository.

At any time, researchers can transcribe and curate new archival documents, or make corrections in the existing (curated) data due to new knowledge acquired in the course of research, and then re-execute the transformation process and update the semantic repository automatically. The changes in the seman-

tic repository are directly (and automatically) reflected in the data exploration application.

The entire set of archival documents to be considered by the researchers does not need to be known from the beginning, meaning that new documents might be collected for transcription at any time. In this case, creation of new source schemas (data entry forms) is needed if such new documents belong to a new type of source which is different from the existing ones. Accordingly, changes in an existing data entry form might be needed (e.g. addition of a new column) in order to enable the transcription of a new important type of information that was not originally planned or known for an existing type of source. In both cases, revision/extension of the domain ontology might be needed, as well as creating new schema mappings or applying changes in the existing ones.

Note here that, even if there are changes in the transcription schemas and the integration model, which actually occur during the course of a project, such changes are independent of the other transcription and curation processes performed (in parallel) by the researchers (thus, they do not affect or delay them). Moreover, the full automation of the data transformation step reduces the overhead for the researchers to the absolute minimum.

The two steps of the workflow that are the most time consuming are the *transcription* and *curation* processes. As already stated, several sub-tasks in these two processes can be automated or semi-automated, e.g. using state-of-the-art text recognition software (Kahle *et al.*, 2017), or applying automated instance matching / entity resolution (Christophides *et al.*, 2020). Here the challenge is to find the best trade-off between fully automating the tasks and having results of high accuracy for enabling valid data analysis. We suggest semi-automated solutions that consider human-in-the-loop for ensuring high quality (Wu *et al.*, 2022; Gurajada *et al.*, 2019).

6 Use Case in Maritime History Research

The workflow has been fully implemented in a real use case for supporting a large number of historians in managing a diverse set of archival sources related to *maritime history*. The context is the project SeaLiT¹³, in which maritime historians study the transition from sail to steam navigation and its effects on seafaring populations in the Mediterranean and the Black Sea (1850s-1920s).

Below we provide details on how each process of the workflow was implemented and illustrate an example on how a real information need provided by the historians is satisfied by exploiting the integrated data.

Archival material and information needs. The archival material studied in SeaLiT covers a variety of sources in five languages (Spanish, Italian, French, Russian, Greek), including crew and displacement lists, registers of different types (sailors, naval ships, students, etc.), logbooks, payrolls, account books,

¹³ <https://sealitproject.eu/>

employments records, and censuses. Details about the full archival corpus and its origin is available in the project’s web site.¹⁴

Our first task was to gather a set of information needs from the historians of SeaLiT, related to their research aims and for which the studied archival material can provide important information. This is fundamental for better designing the source schemas (data entry forms), the integration model, as well as the data exploration services. We collected around 100 information needs. Indicative examples are:¹⁵

- What are the places of construction of ships during a specific period?
- What are the most popular European destinations (under a chronological perspective) of the ships from the Black Sea?
- How many people that arrived at a specific place (e.g. Barcelona) have place of birth more than X miles away?
- How many ship owners per ship during a specific period?

Creation of source schemas and transcription. The FastCat system (Fafalios *et al.*, 2021b), which is available as open source software¹⁶, was used for the creation of the source schemas and the transcription of the archival documents by around 30 users in 5 countries (historians of SeaLiT). In FastCat, users can transcribe documents and provide metadata information by creating ‘records’ belonging to specific ‘templates’. A ‘record’ organises the data and metadata of an archival document in a set of tables, while a ‘template’ represents the structure of a distinct data source, i.e. it defines the data entry tables, their columns as well as the type of each column (for denoting columns that provide vocabulary terms or entity-related information, whose values will be curated after transcription). For the case of SeaLiT, twenty templates were created, one for each different type of archival source. Table 1 provides the templates as well as an overview of the information that can be recorded in each template.

The total number of records transcribed by the historians of SeaLiT is currently more than 620. Fig. 3 shows a part of a real record belonging to the template *Crew List (Ruoli di Equipaggio)*¹⁷ (there are totally 98 records belonging to this template). This template consists of six tables, enabling historians to provide/transcribe information about: i) the record itself (creation date, last modification date, transcriber); ii) the source (archive/library, location, register number, issuing authority, etc.); iii) the ship (name, type, tonnage, construction location, etc.); iv) the crew list (embarkation port and date, discharge port and date, surname, name, residence location, profession, payment information, etc.); v) the documented navigation (date, duration, first planned destination, total crew number); vi) the route (departure port and date, arrival port and date). In the record of Fig. 3, for instance, the transcriber has provided data for twenty

¹⁴ <https://sealitproject.eu/archival-corpus>

¹⁵ The full list of gathered information needs is available at https://users.ics.forth.gr/~fafalios/SeaLiT_Competyency_Questions_InfoNeeds.pdf

¹⁶ <https://github.com/isl/FastCat>

¹⁷ The full record is accessible at: <https://tinyurl.com/2u35frya>

Table 1. Considered archival sources and overview of recorded information.

Archival source	Overview of recorded information
Crew and displacement list (Roll)	Information about ships, crew members, ports.
Crew List (Ruoli di Equipaggio)	Information about ships, voyages, crew members, ports.
General Spanish Crew List	Information about ships, ship owners, crew members, voyages, ports.
Sailors Register (Libro de registro de marineros)	Information about sailors (including profession and military service organisation locations)
Register of Maritime Personnel	Information about persons (including residence location, marital status, previous profession, military service organisation locations).
Register of Maritime Workers	Information about maritime workers, ships, captains, ports.
Seagoing Personnel	Information about persons (including marital status, profession, end of service reasons), ships, destinations.
Naval Ship Register List	Information about ships (including tonnage, length, construction location, registration location) and ship owners.
List of Ships	Information about ships (including previous names, registry port and year, construction place and year, tonnage, engine characteristics, owners).
Civil Register	Information about persons (including profession, origin location, marital status, death location and reason).
Maritime Register, La Ciotat	Information about persons, embarkation and disembarkation locations, ships, captains, patrons.
Students Register	Information about students and courses.
Census La Ciotat	Information about occupants (including nationality, marital status, religion, profession, working organisation, household role).
Census of the Russian Empire	Information about occupants (including marital status, estate, religion, native language, household role, occupation).
Payroll (of Greek Ships)	Information about ships, captains, voyages, persons, employments (including wages).
Payroll (of Russian Steam Navigation and Trading Company)	Information about ships, persons, recruitments (including salary per month).
Employment records (Shipyards of Messageries Maritimes, La Ciotat)	Information about workers (including marital status, profession, status of service in company).
Logbook	Information about ships, captains, ports, route movements, voyage events.
Accounts Book	Information about ships, voyages, captains, ports, transactions.
Notarial deeds	Information about deeds, notaries, witnesses, contracting parties, ships.

six sailors and thirteen route ports that concern the navigation of the the ship *Pallade* (type *Brigantino*) from 11-01-1861 to 26-02-1862.

The creation and configuration of the templates in FastCat was not an ‘one shot’ process. New templates were created periodically based on new archival material gathered from the historians, or existing templates were changed several times even after the creation of records (e.g. by including additional columns in a table), for incorporating new (and important) type of information provided by particular archival documents.

Crew List (Ruoli di Equipaggio), Pallade, 1861-01-11, Carolina Gaggero

FastCat Record Information

Use english to fill in the fields of this table

Data			Authors		
Id	Creation date	Last Modified	Name *	Surname *	Role
13	2018-03-20	2020-09-22T16:16:02	Carolina	Gaggero	

Source Identity

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

Archive / Library				Document		Issuing Authority			Fond		
Name	Location	Register Number	Number	From *	To	Name	Location	Source Type Name	Number	Title	Series Number
Archivio di Stato di Genova	Genova	14	6740	1861-01-11		Direzione marittima di Genova	Genova				

Ship Identity

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

Construction				Ammunition			Registry		Owner					
Ship name *	Ship type	Tonnage	Location	Date (year)	Type	Value	Unit	Port	Number	Organization Name	Name	Surname	Father's Name	Note
Pallade	Brigantino	160,51	Varazze	1852				Genova	1439		Nicolò	Schiaffino	Giobatta	

Crew List

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	Profession/Rank
	Port	Date	Port	Date							
1	Genova	1861-01-11	Castellammare	1861-05-04	Mortola	Giuseppe		Camogli	1826	5054	Capitano di seconda cl
2	Genova	1861-01-11	Genova	1862-02-26	Dellacasa	Emanuele		Camogli	1830	5779	Secondo
3	Genova	1861-01-11	Castellammare	1861-04-29	Lardone	Biaggio		Camogli	1827	1694	Dispensiere
4	Genova	1861-01-11	Castellammare	1861-04-29	Agnese	Franco		Camogli	1834	14272	Marinaio
5	Genova	1861-01-11	Castellammare	1861-04-29	Olivari	Pasquale		Camogli	1828	17624	Marinaio
6	Genova	1861-01-11	Castellammare	1861-04-29	Ansaldo	Giuseppe		Camogli	1835	8499	Marinaio
7	Genova	1861-01-11			Massa	Filippo		Camogli	1841	10980	Marinaio
8	Genova	1861-01-11	Castellammare	1861-04-29	Bertelli	Franco		Camogli	1843	12666	Mozzo
9	Genova	1861-01-11	Genova	1862-02-26	Mortola	Franco		Camogli	1848	15489	Mozzo
10	Genova	1861-01-11	Castellammare	1861-05-04	Mortola	Franco		Camogli	1853		Mozzo

Fig. 3. An example of a real FastCat record belonging to the template ‘Crew List (Ruoli di Equipaggio)’.

Curation. The curation of the transcribed data (vocabulary terms and entity instances) is performed through a dedicated environment within FastCat, called FastCat Team. Specifically, when a historian has completed the transcription of one or more documents (records), the record(s) can be ‘published’, which means

that all data concerning vocabulary terms and entity instances are copied to FastCat Team for enabling their curation.

In the case of SeaLiT, the current number of vocabularies is fifty two (examples include: *ship type*, *engine type*, *profession*, *marital status*), while the types (and current number) of entities that can be curated are *ships* (about 2,400), *persons* (about 99,200), *locations* (about 9,800), *legal entities* (about 1,100). For each term in a vocabulary, the user can provide a preferred term (in English) and a broader term, or inspect the records in which the term appears. For the curation of the entity instances, the user can correct values, select two or more instances for matching them (indicating that they represent the same real-world entity), unmatch a particular instance from a set of automatically-matched instances, or inspect the records in which the entity instance appears. In the case of locations, the user is able to add an identifier (TGN/Geonames ID), as well as coordinates or a secondary location name (e.g. a historical name).

Fig. 4 shows the user interface of FastCat Team, in particular the page that allows the curation of ship instances. For more information about FastCat (and FastCat Team), the reader can refer to Fafalios *et al.* (2021b).

	Name	Previous name	Type	Call signal	Construction location	Construction date
<input type="checkbox"/>	Ulisse		brigantino		Varazze	1855
<input type="checkbox"/>	Ugo Bassi		Brigantino		Varazze	1853
<input type="checkbox"/>	Telemaco		Brigantino		Varazze	1841
<input type="checkbox"/>	Santa Caterina di Genova		Brigantino		Varazze	1837
<input type="checkbox"/>	San Rocco		Brigantino		Varazze	1850
<input type="checkbox"/>	S. Prospero		Brigantino		Varazze	1856
<input type="checkbox"/>	S. GioBattista		Brigantino		Varazze	1843

Fig. 4. Curation of ship instances in FastCat Team.

Ontology-based integration and transformation. For data integration we created a data model compatible with CIDOC-CRM, called ‘SeaLiT Ontology’¹⁸. The current version of the ontology (v1.1) contains forty six classes and seventy nine properties, allowing the description of information about ships, voyages,

¹⁸ <https://zenodo.org/record/6797750>

employments, payments, seafaring people, teaching courses, and other relevant activities. For creating the schema mappings and transforming the data to RDF we make use of the X3ML framework (Marketakis *et al.*, 2017). In particular, one mapping definition file has been created for each template in FastCat, as well as one for each of the four categories of entities in FastCat Team and one for all the vocabularies.

The derived semantic network contains more than 18.5M RDF triples and is currently exploited by the data exploration application (ResearchSpace; more below) for supporting historians in finding answers to their information needs. The full RDF datasets are publicly available¹⁹. The network contains interconnected information for thousands of sailors, ships, locations, organisations, voyages, and many other relevant activities, as well as connections with publicly available resources (Geonames, Getty TGN).

Semantic network exploration. For enabling historians of SeaLiT and other interested parties to explore the integrated data and find answers to their information needs, we make use of ResearchSpace (Oldman and Tanase, 2018). ResearchSpace is a configurable, open source platform which operates over a semantic network accessible through an RDF triplestore. It offers a variety of functionalities, including a *query building* interface that supports users in gradually building and running complex queries through a user-friendly interface. The results can then be browsed and analysed quantitatively through different visualisations, such as bar charts.

The platform was configured for the case of SeaLiT data, offering three main data exploration functionalities: a) keyword search, b) semantic search (through its assistive query building interface), and c) entities browsing (per type of archival source). Fig. 5 shows a screen dump of the semantic search functionality. The user inspects the “construction location of ships that were constructed between 1830 and 1840”. The user first searched for ships constructed between 1830 and 1840 (Fig.5-A), and then selected to group the retrieved ships by their construction location (Fig.5-B) and visualise the results in a bar chart (Fig.5-C). This query corresponds to a real information need as provided by the historians of SeaLiT, and the answer is shown to the user instantly (in less than one second). If the construction location is unknown for a ship, this missing information is displayed in the chart (see ‘Unknown’ bar, Fig.5-D). The user can also start browsing information about the retrieved ships (e.g. inspecting the owners of a ship and then other ships owned by the same owner), visit the FastCat transcripts that provide the corresponding information (for validation, or inspection of additional contextual information), or download the results in CSV format for further (external) analysis.

A deployment of the application is publicly accessible.²⁰

¹⁹ <https://zenodo.org/record/6460841>

²⁰ <http://rs.sealitproject.eu/>

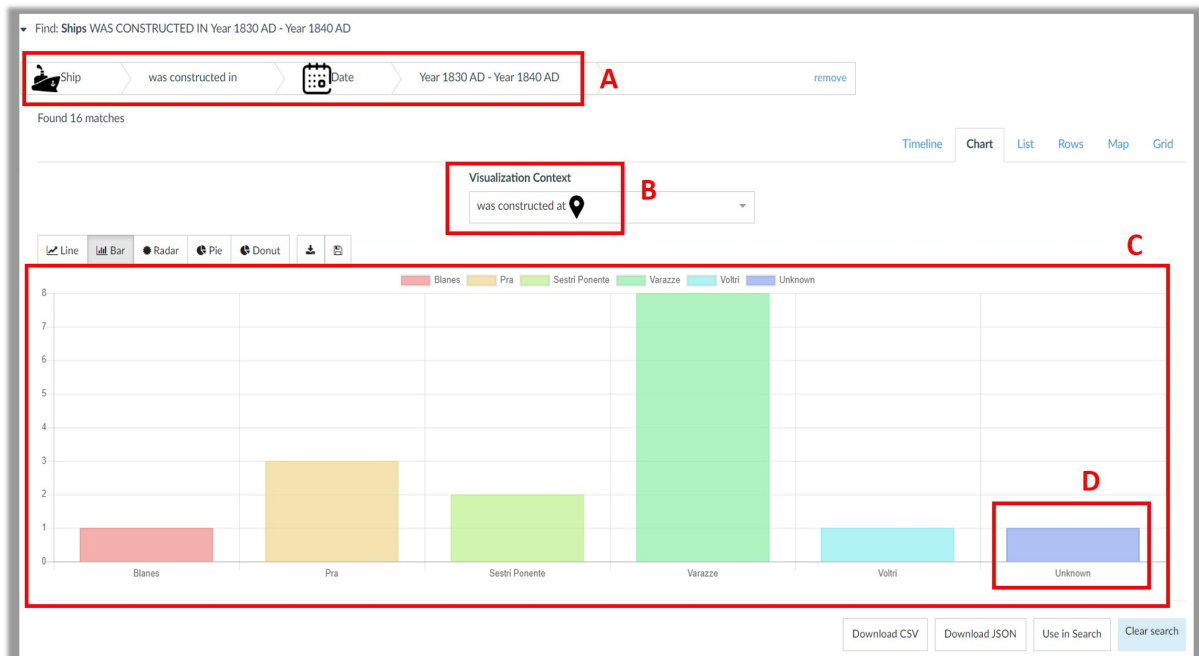


Fig. 5. Semantic search and results visualisation in ResearchSpace.

7 Quality Aspects and Lessons Learned

We discuss data quality aspects as well as relevant lessons learned from the application of the proposed workflow model in maritime history research.

7.1 Quality Aspects

Every workflow cycle ends up with semantic data that in some cases may suffer low quality characteristics, making the data practically difficult to be exploited for the needs of research. In literature, data quality is commonly considered as “fitness for use” as well as an indicator of data usability (Pipino *et al.*, 2002; Wang and Strong, 1996), and several dimensions and metrics for measuring data quality have been proposed (Pipino *et al.*, 2002; Zaveri *et al.*, 2016). Although studying quality factors in detail is out of the scope of this paper, below we focus on three main quality dimensions of the semantic data that can significantly affect the quantitative analysis process: completeness, consistency, conciseness.

Data completeness. A quality dimension that can be easily assessed in the context of a schema/ontology or the particular use case scenario (Zaveri *et al.*, 2016). The lack of essential information, like missing dates and locations of events, or names and professions of actors of a registry, may affect the research analysis and the evidence for making a decision about a historical subject.

Data consistency. This dimension can be viewed from a number of perspectives (Zaveri *et al.*, 2016; Hassenstein and Vanella, 2022). Our perspective comprises the *schema-based* and the *value-based* (or *representational*) consistency. Schema-based consistency can be evaluated against a particular schema/ontology. It prevents modeling issues, like the incompatible attribution/interlinking of the entities, and averts potential reasoning malfunction. For example, assigning ‘tonnage’ to a person (instead of a ship) makes no sense, and under particular reasoning premises it may produce inaccurate inference that people were used for the transportation of goods. Value-based consistency concerns the format and the structure of comparative values (numbers, dates, measurement values) to enable comparability. Magnitudes, dimensions, quantities, time-spans, dates, places’ coordinates, etc., to be effectively compared, they have to align their reference points or units of measurement.

Data conciseness. This quality dimension comprises two perspectives: *schema-level* conciseness and *instance-level* conciseness (Zaveri *et al.*, 2016; Mendes *et al.*, 2012). Schema-level conciseness means that the data does not contain equivalent attributes with different names (responsibility of the data modeling engineer), while instance-level conciseness means that the data does not contain equivalent objects with different identifiers (highly-dependant on the quality of the curation process).

7.2 Lessons Learned

Next we present issues related to data quality that we faced while implementing the workflow and which should be taken into account.

Missing information. Missing values are very common and an important-to-know information for researchers because they can affect the accuracy of quantitative (statistical) analysis. This is related to the *completeness* quality aspect described above. When a piece of information is not provided in the original source, the corresponding cell in the data entry system is left empty. The data exploration system must consider such empty values while aggregating and showing information.

Data entry errors. Errors in the transcripts during data entry are common, such as accidentally filling the wrong column in a table, or putting the information in the wrong place due to misunderstanding. This is related to the *schema-based consistency* quality aspect described above. Such errors are directly reflected in the data exploration interfaces and can spoil user experience. Thus, it is important to allow researchers visit the original transcripts for validation or making corrections. Moreover, offering mechanisms in the user interface that support users to avoid such errors during data entry can limit the problem.

Non-consistent comparative values. It is very common that comparative values, such as dates, dimensions, quantities, location coordinates, are not consistent across archival sources of different types, because of different reference points or units of measurement, making difficult their use in comparisons, filtering, etc. This is related to the *value-based consistency* quality aspect described above. An additional (automated, semi-automated or manual) step is needed

for aligning such values, however without changing the values as they appear in the original source. This can happen either during data curation or during data transformation.

Costly data curation. Low-quality data curation can reduce user satisfaction and produce invalid analysis results. This is related to the *instance-level conciseness* quality aspect described above. The cost of manual data curation is relative to the size of the data that need curation (number of entity instances, number of vocabulary terms). The process can be very time consuming for researchers in cases such as SeaLiT where the number of entities and vocabularies is high. Thus, there is a need for tools that automate as much as possible curation without significantly affecting quality, e.g. through semi-automatic processes, supervised algorithms, or application-specific machine learning.

8 Conclusion

We presented a workflow model for holistic data management in archival research: from transcribing and documenting a set of archival documents, to curating the transcribed data, integrating it to a rich semantic network, and then exploring and analysing the integrated data. The merits of the approach is that it speeds up data entry, it is provenance-aware decoupling data entry from data curation and integration, it is interactive as well as appropriate for semantic interoperability, aiming at the production of sustainable data of high value and long-term validity.

We have showcased the feasibility and effectiveness of the model in maritime history research, and we have reported empirical results from its application (about thirty users, twenty types of archival documents, more than 600 records, more than fifty vocabularies, more than 110,000 entity instances, more than 18.5 million triples of integrated information).

Issues that are worth further research include: (a) semi-automated methods to speedup data curation, (b) investigate the evolution requirements of the semantic network, as proposed by Marketakis *et al.* (2021), (c) methods and interfaces to support researchers in defining and updating the source schemas by themselves.

Acknowledgements

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

References

- Ali, W., Saleem, M., Yao, B., Hogan, A., and Ngomo, A.-C. N. (2021). A survey of RDF stores & SPARQL engines for querying knowledge graphs. *The VLDB Journal*, pages 1–26.

- Antoniou, G.** and **Van Harmelen, F.** (2004). *A semantic web primer*. MIT press.
- Beretta, F.** (2021). A challenge for historical research: making data FAIR using a collaborative ontology management environment (OntoME). *Semantic Web*, 12(2):279–294. Publisher: IOS Press.
- Calvanese, D., De Giacomo, G., Lenzerini, M., Nardi, D., and Rosati, R.** (1998). Description logic framework for information integration. In *KR*, pages 2–13.
- Carroll, J. J., Bizer, C., Hayes, P., and Stickler, P.** (2005). Named graphs. *Journal of Web Semantics*, 3(4):247–267. Publisher: Elsevier.
- Chapman, R.** and **Wylie, A.** (2018). *Evidential reasoning in archaeology*. Bloomsbury Publishing.
- Christophides, V., Eftymiou, V., Palpanas, T., Papadakis, G., and Stefanidis, K.** (2020). An overview of end-to-end entity resolution for big data. *ACM Computing Surveys (CSUR)*, 53(6):1–42. Publisher: ACM New York, NY, USA.
- Delis, A.** (2020). Seafaring Lives at the crossroads of Mediterranean maritime history. *International Journal of Maritime History*, 32(2):464–478. Publisher: SAGE Publications Sage UK: London, England.
- Dimou, A., Vander Sande, M., Colpaert, P., Verborgh, R., Mannens, E., and Van de Walle, R.** (2014). RML: a generic language for integrated RDF mappings of heterogeneous data. In *Ldow*.
- Doerr, M.** (2003). The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata. *AI magazine*, 24(3):75–75.
- Fafalios, P., Konsolaki, K., Charami, L., Petrakis, K., Paterakis, M., Angelakis, D., Tzitzikas, Y., Bekiari, C., and Doerr, M.** (2021a). Towards Semantic Interoperability in Historical Research: Documenting Research Data and Knowledge with Synthesis. In *International Semantic Web Conference*, pages 682–698. Springer.
- Fafalios, P., Petrakis, K., Samaritakis, G., Doerr, K., Kritsotaki, A., Tzitzikas, Y., and Doerr, M.** (2021b). FAST CAT: collaborative data entry and curation for semantic interoperability in digital humanities. *Journal on Computing and Cultural Heritage (JOCCH)*, 14(4):1–20. Publisher: ACM New York, NY, USA.
- Gurajada, S., Popa, L., Qian, K., and Sen, P.** (2019). Learning-based methods with human-in-the-loop for entity resolution. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 2969–2970.
- Hassenstein, M. J.** and **Vanella, P.** (2022). Data Quality—Concepts and Problems. *Encyclopedia*, 2(1):498–510. Publisher: MDPI.
- Hawkins, A.** (2021). Archives, linked data and the digital humanities: increasing access to digitised and born-digital archives via the semantic web. *Archival Science*, pages 1–26. Publisher: Springer.
- Heath, T.** and **Bizer, C.** (2011). Linked data: Evolving the web into a global data space. *Synthesis lectures on the semantic web: theory and technology*, 1(1):1–136. Publisher: Morgan & Claypool Publishers.

- Hyvönen, E.** (2020). Using the Semantic Web in digital humanities: Shift from data publishing to data-analysis and serendipitous knowledge discovery. *Semantic Web*, 11(1):187–193. Publisher: IOS Press.
- Hyvönen, E. and others** (2020). “Sampo” Model and Semantic Portals for Digital Humanities on the Semantic Web. In *Proceedings of the Digital Humanities in the Nordic Countries 5th Conference (DHN 2020)*. CEUR-WS.org.
- Hyvönen, E., Tuominen, J., Alonen, M., and Mäkelä, E.** (2014). Linked Data Finland: A 7-star model and platform for publishing and re-using linked datasets. In *European Semantic Web Conference*, pages 226–230. Springer.
- Izquierdo, Y. T., García, G. M., Menendez, E., Leme, L. A. P., Neves, A., Lemos, M., Finamore, A. C., Oliveira, C., and Casanova, M. A.** (2021). Keyword search over schema-less RDF datasets by SPARQL query compilation. *Information Systems*, 102:101814. Publisher: Elsevier.
- Kadilierakis, G., Fafalios, P., Papadakos, P., and Tzitzikas, Y.** (2020). Keyword search over RDF using document-centric information retrieval systems. In *European Semantic Web Conference*, pages 121–137. Springer.
- Kahle, P., Colutto, S., Hackl, G., and Mühlberger, G.** (2017). Transkribus—a service platform for transcription, recognition and retrieval of historical documents. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 4, pages 19–24. IEEE.
- Kritsotakis, V., Roussakis, Y., Patkos, T., and Theodoridou, M.** (2018). 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)*.
- Marketakis, Y., Minadakis, N., Kondylakis, H., Konsolaki, K., Samaritakis, G., Theodoridou, M., Flouris, G., and Doerr, M.** (2017). X3ML mapping framework for information integration in cultural heritage and beyond. *International Journal on Digital Libraries*, 18(4):301–319. Publisher: Springer.
- Marketakis, Y., Tzitzikas, Y., Gentile, A., Niekerk, B. V., and Taconet, M.** (2021). A workflow for supporting the evolution requirements of RDF-based semantic warehouses. *International Journal of Metadata, Semantics and Ontologies*, 15(3):220–232.
- Mendes, P. N., Mühleisen, H., and Bizer, C.** (2012). Sieve: linked data quality assessment and fusion. In *Proceedings of the 2012 joint EDBT/ICDT workshops*, pages 116–123.
- Meroño-Peñuela, A., Ashkpour, A., Van Erp, M., Mandemakers, K., Breure, L., Scharnhorst, A., Schlobach, S., and Van Harmelen, F.** (2015). Semantic technologies for historical research: A survey. *Semantic Web*, 6(6):539–564. Publisher: IOS Press.
- Nikas, C., Kadilierakis, G., Fafalios, P., and Tzitzikas, Y.** (2020). Keyword search over RDF: is a single perspective enough? *Big Data and Cognitive Computing*, 4(3):22. Publisher: Multidisciplinary Digital Publishing Institute.

- Oldman, D., Doerr, M., and Gradmann, S.** (2015). Zen and the Art of Linked Data - New Strategies for a Semantic Web of Humanist Knowledge. In *A New Companion to Digital Humanities*, pages 251–273. John Wiley & Sons, Ltd. Section: 18.
- Oldman, D. and Tanase, D.** (2018). Reshaping the knowledge graph by connecting researchers, data and practices in ResearchSpace. In *International Semantic Web Conference*, pages 325–340. Springer.
- Ouksel, A. M. and Sheth, A.** (1999). Semantic interoperability in global information systems. *ACM Sigmod Record*, 28(1):5–12. Publisher: ACM New York, NY, USA.
- Petrakis, K., Samaritakis, G., Kalesios, T., i Domingo, E. G., Delis, A., Tzitzikas, Y., Doerr, M., and Fafalios, P.** (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):60–87.
- Pipino, L. L., Lee, Y. W., and Wang, R. Y.** (2002). Data quality assessment. *Communications of the ACM*, 45(4):211–218. Publisher: ACM New York, NY, USA.
- Scholz, M. and Goerz, G.** (2012). WissKI: a virtual research environment for cultural heritage. In *ECAI 2012*, pages 1017–1018. IOS Press.
- Turner, D.** (2007). *Making prehistory: Historical science and the scientific realism debate*. Cambridge University Press.
- Tzitzikas, Y., Mountantonakis, M., Fafalios, P., and Marketakis, Y.** (2022). CIDOC-CRM and Machine Learning: A Survey and Future Research. *Heritage*, 5(3):1612–1636.
- Ventresca, M. J. and Mohr, J. W.** (2017). Archival Research Methods. In *The Blackwell Companion to Organizations*, pages 805–828. John Wiley & Sons, Ltd. Section: 35.
- Vicente-Saez, R. and Martinez-Fuentes, C.** (2018). Open science now: A systematic literature review for an integrated definition. *Journal of business research*, 88:428–436.
- Volz, J., Bizer, C., Gaedke, M., and Kobilarov, G.** (2009). Silk-a link discovery framework for the web of data. In *Proceedings of the WWW2009 Workshop on Linked Data on the Web*.
- Wang, R. Y. and Strong, D. M.** (1996). Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*, 12(4):5–33. Publisher: Taylor & Francis.
- Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., and He, L.** (2022). A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, 135:364–381.
- Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., and Auer, S.** (2016). Quality assessment for linked data: A survey. *Semantic Web*, 7(1):63–93. Publisher: IOS Press.