Linking Maritime Datasets to Dutch Ships and Sailors Cloud - Case studies on Archangelvaart and Elbing

J.A. Entjes Vrije Universiteit Amsterdam
De Boelelaan 1105 1081 HV Amsterdam
J.A.Entjes@student.vu.nl

ABSTRACT

Although very different fields of science, History can profit a lot from Information Science in the digisation of historic information. Currently digitised datasets are limited by their own contents, but can offer new information and give rise to new research questions when converted to Linked Data standards. Once a dataset is converted to Linked Data, its contents and relations between different datasets are much easier to compare and explore than in traditional dataset standards. The Dutch Ships and Sailors project does just that, integrating different datasets of Dutch Maritime history as Linked Data. With the Dutch Ships and Sailors project completed, there are still datasets not yet digitised or turned into Linked Data.

This research explores how new digital datasets can effectively be linked to the Dutch Ships and Sailors datasets and if adding additional datasets can assist in answering existing research questions in the field of History. To this end, two new datasets are converted to RDF and integrated with the Dutch Ships and Sailors cloud for this research, based on requirements engineered with the aid of the historians who digitised these datasets in the first place. The datasets chosen are Archangel in northwest Russia and Elbing in the Baltic Sea. Trade in these regions played an important role in changing the Dutch economy and navy and helped give rise to the naval supremacy of the Dutch Golden Age. Each part of the process is evaluated to ensure the conversion does not deviate from its intended track. Next, visualisations are be made that access this data remotely, to show that different maritime datasets as Linked Data can be queried for information and represented in a meaningful way. These visualisations are also evaluated.

This research finds that digitised datasets can be converted without loss of information. By mapping concepts of these new datasets to existing Dutch Ships and Sailors concepts, the information stored in different datasets can be effectively linked to each other and makes recommendations on how to add datasets in the future. It also supplements the Dutch Ships and Sailors data with new concepts for future datasets. The visualisations show a new way of reviewing information stored in the datasets, but the total information is still too limited to form meaningful answers on existing questions.

Keywords

Linked Data, Digital Humanities, Semantic Web

1. INTRODUCTION AND RELATED WORK

Digital Humanities is a relatively new field of science, that uses the possibilities Computer Science offers for the traditional Humanities field of science (Schreibman et al., 2008). The importance of collaboration between the field of history and computer science was underlined by a meeting that took place in late June 2015. The participants were for a large part a mix of computer scientists, historians and representatives of historic institutions (such as one from the Huygens ING instititue). The topic of the day was digitising Dutch maritime history. During the meeting different presentations were held of current projects. There is quite some historic data that has been digitised, but during this meeting it became apparent that many historians were digitising data as a hobby, outside of office hours. It is these historians who have an interest in some data and want to form hypotheses about the data, or compare the data, but they often lack the tools to do so. Yet there are information scientists who have a good understanding of the digital field and the possibilities, but lack the historic insight of the historians to form hypotheses. Research on how semantic technologies can help historians has been performed before (Meroño-Peñuela et al., 2013), which states that: "One of the big claims of linked data is that, by linking datasets, relations established between nodes of these datasets highly enrich the information contained in them. That way, browsing datasets is not an isolated task anymore: by allowing users (and machines) to explore URI entities through their predicate links, data get new meanings, uncountable contexts and useful perspectives for historians".

Bringing these two fields of science together in digital humanities therefore could allow the technological prowess of a computer scientists to support the demand for knowledge of a historian, improving our understanding of history.

1.1 Linked Data

Linked data "is a set of best practices for publishing and interlinking structured data on the web" (Heath & Bizer,

2011). It is sometimes referred to as the Semantic Web, although not strictly the same. In this study, the term Linked Data will be used. The first step to turn data to Linked Data, is having it online. Naturally, simply having data online is no guarantee for it being easy to find or actually link to other data. Hence, the following four additional requirements[1] have been established, increasing the value of Linked Data as more are adherred to:

- Use Uniform Resource Identifiers (URI)[2] as names for things.
- Use HTTP URIs so that people can look up those names.
- When someone looks up a URI, provide useful information using the standards (RDF[3], SPARQL[4])
- Include links to other URIs, so people can discover more things.

URIs are a means of identifying resources. Although not strictly the same, most people are familiar with Uniform Resource Locators (URLs) as web addresses. A URL is a more specific version of a URI, one that points to a location. In the case of the world wide web, a website. URIs are a used by Linked Data (Heath & Bizer, 2011) because:

- They are a simple way to access unique data.
- Aside to being a name, they also allow an access to information and describe the resource.

RDF makes use of URIs to describe resources. RDF is built from triples, built from URIs. A triple is a data entry that contains a point of origin (the 'subject'), a relation (the 'predicate'), and an endpoint (the 'object'). Essentially, the subjects are resources whose properties are described by its relations. An example of a triple relation is given in Figure 1.

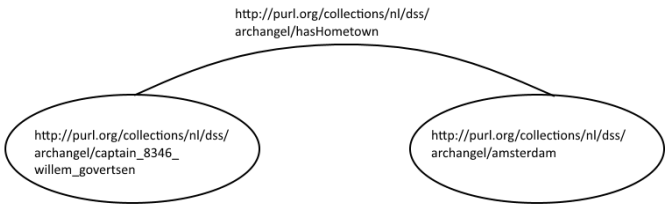

Figure 1: An example of a triple. It shows a relation between a subject (left) and the object (right). The subject, relation and object each have their own URI.

In RDF, it is possible to reuse existing vocabularies to define resources or concepts in a new dataset. In fact, it is good practice to reuse these when possible (Yu, 2011). This prevents several URIs pointing to the same concept. Reused concepts also allow them to be linked more easily.

RDF concepts can be defined in a schema. An RDF schema[5] is essentially a definition of all concepts and relations present in RDF data. For this data, it defines classes and properties. Using the schema, certain concepts can be mapped to other concepts. By mapping as many of the concepts to existing concepts, effectively existing vocabularies are reused.

To define RDF elements in this study, the 'Terse RDF Triple Language' (Turtle)[6] has been used. Using Turtle, RDF graphs can be written down in a compact textual form. It allows the use of prefixes. To store the turtle files, the ClioPatria[7] triplestore was used. A triplestore is a tool that can be installed for use as RDF database[8]. ClioPatria was used as it also features a SPARQL environment. SPARQL (Protocol and RDF Query Language) is the query language for RDF.

If machine readable data such as RDF is published and all these documents are connected to each other, a web of Linked Data will be created that can be processed by machines. This is the idea behind Linked Open Data (Yu, 2011). A graphical representation of what this currently looks like can be found in Figure 2, the Linked Open Data cloud.

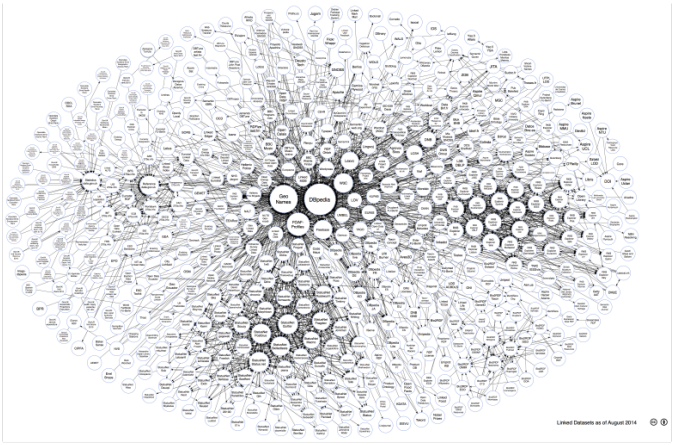

Figure 2: The Linked Open Data cloud[9]

1.2 Dutch Ships and Sailors

The Netherlands has a rich maritime history, in particular during the 17th century. The Dutch economy and fleet grew substantially during this time, in absolute levels as well as relative to other European powers (Van Zanden & Van Tielhof, 2009). Much of this history has been documented for

[1] http://www.w3.org/DesignIssues/LinkedData.html
[2] http://www.w3.org/Addressing/
[3] http://www.w3.org/RDF/
[4] http://www.w3.org/TR/rdf-sparql-query/
[5] http://www.w3.org/TR/rdf-schema/
[6] http://www.w3.org/TeamSubmission/turtle/
[7] http://cliopatria.swi-prolog.org/home
[8] https://www.w3.org/2001/sw/wiki/Category:Triple_Store
[9] Linking Open Data cloud diagram 2014, by Max Schmachtenberg, Christian Bizer, Anja Jentzsch and Richard Cyganiak. http://lod-cloud.net/

taxing, toll, trade contracts and more. Digitisation of these documents allows the processing power of machines to search through records at high speed, opening new research possibilities. Dutch Ships and Sailors[10](de Boer et al., 2014) is a project that aims to provide an infrastructure for maritime historical datasets, linking data through semantic web technology. The Dutch Ships and Sailors project brings together different individual datasets related to recruitment and shipping, so far mainly during the 18th century and in the shipping in the northern provinces of the Netherlands. This research aims to support new historic research by integrating two additional datasets with the Dutch Ships and Sailors cloud.

1.3 Integrating the datasets

Integrating new maritime datasets to the Dutch Ships and Sailors cloud, can provide historians with additional information regarding Dutch maritime activity. Before this can be done, the datasets need to be effectively linked to the existing cloud, in order to build an overview of these naval activities in the bigger picture of all the Dutch naval activities on the world's seas, without a loss of data.

2. RESEARCH QUESTIONS

As historians and information scientists can benefit from each other to create and answer new hypotheses in the field of history, the objective of this study is to investigate whether this applies to the two datasets chosen. As these datasets are already digitised, converting them to Linked Data should allow historians to gain new information from these datasets. To effectively link these datasets, the loss of data should be kept to a minimum and the dataset information should be integrated into the Dutch Ships and Sailors cloud. The research questions are:

1: How can additional datasets effectively be linked to those of the Dutch Ships and Sailors project?

2: How can the additional datasets assist in answering existing research questions in the field of History?

3. APPROACH AND METHODOLOGY

First and foremost, any converted file needs to be linkable to other datasets. Otherwise it is just Linked Data on its own, which is not richer than the existing XML dataset. This is another reason why two datasets instead of one have been converted and linked to the Dutch Ships and Sailors cloud, making these datasets part of the overall cloud of Linked Data.

The research done for this study consists of several parts. It builds on the structure used in earlier research on data conversion to Linked Data, by setting up requirements, using evaluation to make adjustments where needed and visualise the resulting conversion (Brandt & de Boer, 2013). Where possible, data will be converted and uploaded in a similar way as done for the original Dutch Ships and Sailors data (de Boer et al., 2014). The general approach in this study was to convert multiple datasets containing information of maritime historic significance, supported by historians with insight into these datasets. These historians have helped select appropriate datasets, set up conversion requirements, visualisation requirements and have helped by evaluating every step. The evaluation was considered essential to this research, as it helped the conversion to stay on track. To select the datasets, the Huygens ING Institute[11] was contacted. Using their help, the datasets of Archangel and Elbing were selected for conversion. More information about those datasets can be found in Section 3.1. After the data was selected, the research was done in four steps.

First, requirements engineering was performed to establish what kind of demands the end product of this research needed to meet. To this end, two historians from the Huygens ING Institute have been interviewed to gather research questions and demands for the data conversion. The interview was coded to select demands and research questions, which were evaluated with the historians. Second, converting data began. Upon completion of the conversion, it was evaluated with the historians again. Third, the data was mapped to the Dutch Ships and Sailors data and linked to the data cloud. The fourth and final part of the research was creating visual representations based on this data and evaluating these with historians, to find answers to the research questions. To be able to answer the research questions, the following outcomes are required:

- A theoretical conversion schema for new datasets, that links them the Dutch Ships and Sailors cloud. This is essentially a data model explaining how the data is to be converted.
- A practical implementation for the datasets. This is the actual conversion and implementation, from XML to RDF.
- Recommendations for converting XML datasets to RDF based on the research.
- Visualisations that use this data to meet requirements set.

3.1 The data

The two datasets for this project are the toll registry files of the city of Elbing, currently known as Elbląg[12], in Poland and hosted online by the Huygens ING Institute[13], and notarial documents kept by the city Archangel, currently known as Archangelsk[14], in Russia, also hosted by the Huygens ING Institute[15]. These two datasets are chosen, instead of just one, as converting a single dataset that somehow could not be linked to external resources, would not offer anything new to researchers. Furthermore, they serve the same generic trading region, as shown in Figure 3. This region is of interest as Baltic trade provided some of the foundations that lead to the Dutch World-trade Hegemony (Israel, 1989). Finally, these datasets overlap in time period, allowing their data to be compared during similar time frames.

[10] http://dutchshipsandsailors.nl

[11] https://www.huygens.knaw.nl/

[12] https://www.google.nl/maps/place/Elblag,+Polen/

[13] http://resources.huygens.knaw.nl/pondtolregisterselbing

[14] https://www.google.nl/maps/place/Archangelsk,+Oblast+Archangelsk,+Rusland/

[15] http://resources.huygens.knaw.nl/archangel/app

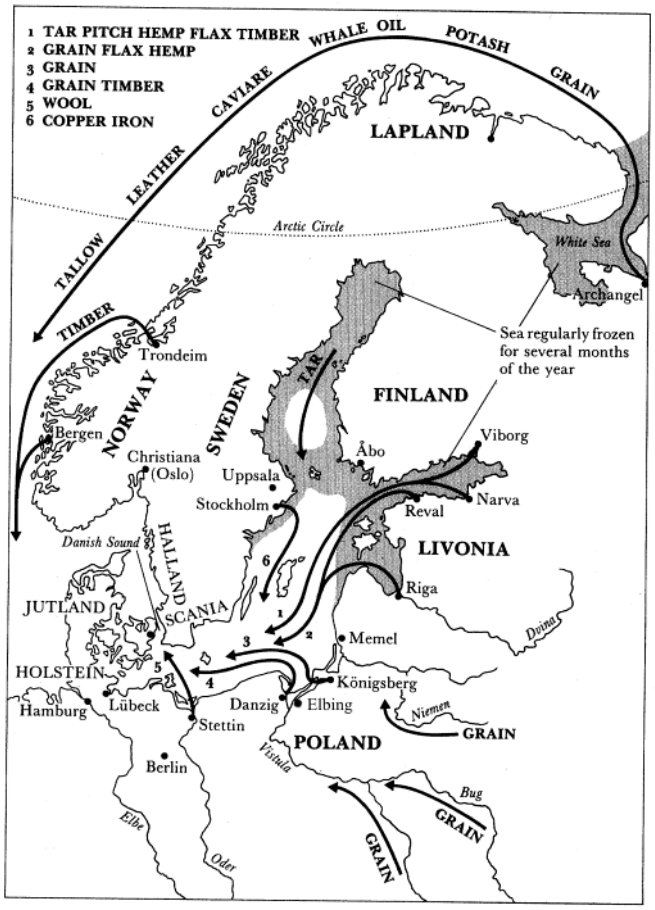

Figure 3: Dutch trade in the Baltic region, adapted from its original (Israel, 1989). Elbing is situated halfway between Danzig and Königsberg. Archangel is in the far northeast in the White Sea.

3.1.1 Elbing

The Elbing dataset contains toll registry information from voyages starting in 1585 until 1700. The data of Dutch shipping to Elbing has been collected from the complete Elbing registries from this time (Lindblad et al., 1995). In the creation of this dataset, all shipping that had goods headed towards the Netherlands, regardless of captain or ship nationality, and all shipping by Dutch captains, regardless of the destination of cargo, have been included. The database entries contains information about the ship, captain, cargo carried, the value of the cargo and the toll paid. An example of a database entry for Elbing is found in Figure 4.

3.1.2 Archangel

The data about Archangel contains entries of voyages to Archangel and other European ports from 1594 until 1724. This data was originally gathered by Piet de Buck (1931 - 1999) [16], who was a historian at the University of Leiden, in the Netherlands. Its sources are cargo contracts and other notarial acts from the Amsterdam city archives. It contains information about the ship, captain, the freight brokers and the ship's intended route and cargo pricing. From time to time, special notes have been made such as a shipwreck or a hijack. The dataset contains roughly 4700 acts, some of which are duplicates. Cargo carried is rarely disclosed. Therefore, this dataset is useful mostly for the history of naval commerce and the merchants involved in trade with Russia. An example of a database entry for Archangel is found in Figure 5.

[16] Piet de Buck (†), Sebastiaan Kerkvliet en Milja van Tielhof, Amsterdamse notariële akten over de Archangelvaart 1594-1724 (`http://resources.huygens.knaw.nl/archangel`)

1620 - #3

Op 14-4 heeft schipper Johan Peters uit Buiksloot met zijn schip, de Fortuna.

--»naar Elbing gebracht:				Tol per Product	Waarde per Product
HG	4.0	Last	Haring		
HG	0.0	None	Kramerijen		300.0 fl.
HG	102.0	Ohm	Rijnwijn		
HG	7.0	Pijp	Wijn		
Import Tol Totaal:					

«--vanuit Elbing naar Amsterdam meegenomen:				Tol per Product	Waarde per Product
BF	4.0	Last	Rogge		
EG	20.0	Last	Gerst	7.0 m. 13.0 gr. 6.0 d.	
EG	42.54	Centner	Koper	14.0 m. 17.0 gr. 9.0 d.	
EG	20.0	Last	Tarwe	15.0 m.	
EG	88.0	Steen	Wol	11.0 m.	
HG	31.0	Last	Rogge		
HG	10.0	Steen	Veren		
Export Tol Totaal:				47.0 m. 30.0 gr. 15.0 d.	

Figure 4: Example of front end representation of Elbing dataset. The first line explains generic journey information such as the captain's name, hometown, ship name and date. The first list shows what was transported to Elbing, the goods, intended recipient, quantities and toll or value, with a summary of the sum at the end. The second list shows what was transported to the ship's destination (Amsterdam in this case) and otherwise the same information as in the first list.

3.2 Requirements engineering and evaluation

The requirements of the conversion and eventual visualisation design, were engineered by interviewing two historians from the Huygens ING Institute. Requirements engineering is the act of systematically establishing what a product needs to be able to do, in order to fulfill its intended use (Ebert, 2011). There are a lot of techniques available for requirements engineering. Specific to software engineering (Runeson & Höst, 2009), case studies can serve as an empirical method to investigate phenomena in their context. The case study model should be based on containing five elements (Robson, 2002): what is the objective, what is studied as a case, what is the frame of reference, what are the research questions, what are the methods to collect data and where is this data searched for. Based on this model, in this study these five elements can be defined as follows:

Details van reis

Nummer	3325
Registratiedatum	1620-04-15
Bronverwijzing	SAA NA 162/88
Type akte	bevrachting
Bevrachters	Engelgraeff, Robert Valckenburch, Margr,Wed Vogelaer
Schipper	Govertsen, Willem
Herkomst schipper	Amsterdam
Naam schip	Engel
Naam schip (oorspronkelijk)	De Engel
Lastage	50.0
Gebruikte last	Russisch
Haven van vertrek	Amsterdam
Bestemmingshavens	Lapland Archangel Lapland Amsterdam
Vrachtprijs (totaal)	1670.0
Nummer De Buck	378

Figure 5: Example of front end representation of Archangel dataset. The fields from the top to bottom show the database number information, historic date of registry, source, type of source, freighters, captain, captain provenance, ship name, original ship name, information on goods, harbour of departure, destination harbours, total freight price and identifier number by De Buck.

- The objective is to gather research questions and demands for the data conversion and visualisation.
- As a case, the Elbing and Archangel datasets are studied.
- The frame of reference are data conversions and possibilities offered by them done in earlier research (de Boer et al., 2014).
- The research questions are defined in Section 2.
- Interviewing was considered as the most useful method of collecting data, as the historians at the Huygens ING Institute have a vast knowledge of the datasets and have worked on or with them in the past.
- This data is searched for in the expert knowledge of the Huygens ING Institute historians.

3.2.1 Interviewing historians

The interview was designed to make the historians aware of the possibilities that linking the datasets of Elbing and Archangel to each other and the Dutch Ships and Sailors dataset could offer, after which they were asked to come up with research questions that this could offer them. The interview strategy was to hold semi-structured interviews, in order to be able to inform the interviewees about the possibilities of the project, while not directing their research questions. A semi structured interview is performed as follows: *"The interviewer has an interview guide that serves as a checklist of topics to be covered and a default wording and order for the questions, but the wording and order are often substantially modified based on the flow of the interview, and additional unplanned questions are asked to follow up on what the interviewee says"* (Robson, 2011). It was also important to keep the interview going long enough that a sufficient amount of research questions could be formulated. As this project has a limited scope, oriented on the use and possibilities Linked Data has to offer, rather than solving some great unknown in the whole of maritime history, a sweet spot needed to be found. The research questions selected needed to be within the scope of this project, but interesting and diverse enough that they can be answered by linking the datasets to others. The interview questions are referenced in the appendix.

The interview was then coded in a very simple manner. Because there were only two interviews held and the length of these interviews was just short of an hour each, it was enough to simply summarise both interviews based on concepts handled in the questions and make particular note of research questions encountered during the interview. The interview coding can be found in the appendix.

The interview code lead to a list of research questions, which have been combined with an estimate of how they can be answered. This list can be found in Figure 6. After the interviews, the demands for the data conversion were set up. The historians from the Huygens ING Institute were asked what they considered vital about the data and what the meaning of different data fields was. This was analysed to create an understanding of the data. The decision was then made to make sure that the first conversion of data would keep the data as close as possible to its origins. Any enrichment would only be made with later conversions or additional files. The reason for this is that if the data was tailored to the research questions devised by the historians, it would not show the possibilities of emerging research questions and Linked Data, but only that it is possible to convert data in a labour intensive way to make it useful to answer new predetermined research questions.

The research questions chosen are:

- How big was shipping on Elbing/Archangel in total Dutch Shipping?
- How did wars influence shipping?
- How can climate/weather be linked to shipping?
- Can economic growth be linked to shipping?

Exhaustive list of research topics or questions	How to represent after conversion
How big was shipping on Elbing/Archangel in total Dutch shipping	Compare datasets to DSS
How did wars influence shipping*	War times from external data, compare to volumes of shipping
How can climate/weather be linked to shipping	Weather data, compare to shipping volumes
Search for specific information in large datasets, such as certain names	Can be resolved with Query
How do the price of goods change over time	Can be resolved with Query
Do captains appear in multiple datasets*	Compare datasets to DSS
Can economic growth be linked to shipping*	Compare shipping volumes to economic data from external resource
Do ships appear in multiple datasets*	Can be resolved with Query
Compare load data to DSS to see transfer of goods	Query and compare to DSS
European shipping was usually only Dutch captains, considered more privilliged	Query to verify, compare to DSS
Traders who load ships can be compared, maybe see how finances were organised	Does not appear in both dataset
Baltic trade was less profitable but lower risk than other shipping, compare this	Query and compare to DSS

Figure 6: The research questions derived from the interview on the left, questions suggested in both interviews are marked with an asterisk. On the right, the estimate of how these questions can be represented has been shown. Based on these estimates and questions, the research questions were chosen.

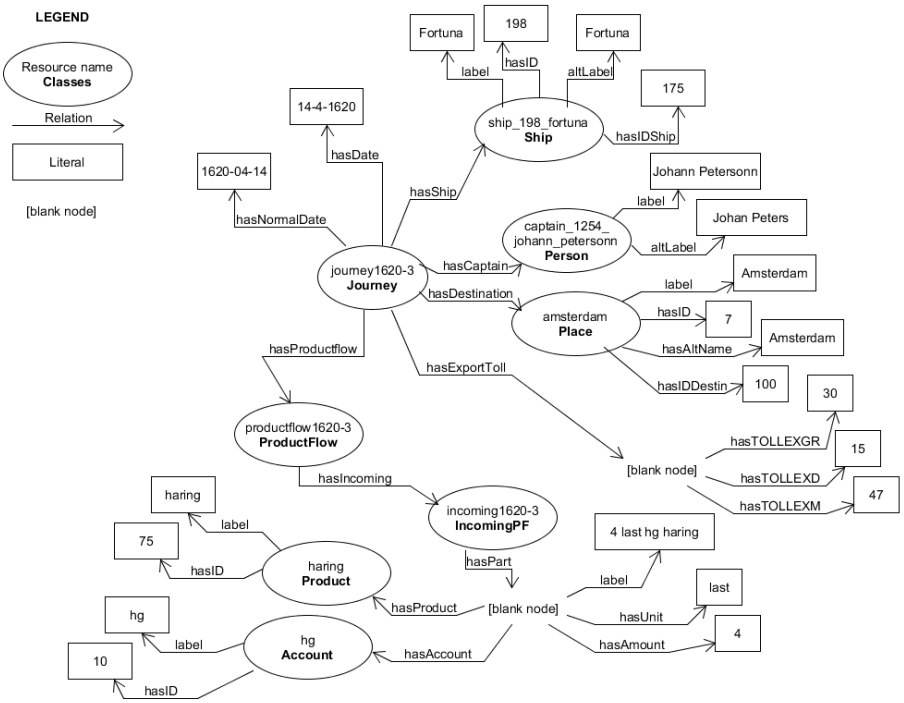

Figure 7: Graphical representation of Elbing journey1620-3. Namespaces have been excluded.

These questions were chosen as they all could visualised in the same way, by combining the data of multiple datasets into one graph. These would include the Elbing Archangel datasets, as well as information on total dutch shipping, possibly from DSS, economic data, wartime data and weather data.

After the requirements analysis with the historians had been completed, they were asked to evaluate the research questions taken from them. Their approval was needed to help decide on how this research would go about answering them. After all, there is no use answering a question that was not asked. Once the research questions were approved of and the requirements for the project known, data conversion began.

3.3 Data conversion

The original data was stored in a relational database. These were provided as a dump in Extensible Markup Language (XML)[17] format. The journey class in XML looks like this:

```
<table name="journey">
    <column name="YEAR">1620</column>
    <column name="NR">39</column>
    <column name="DAY">30</column>
    <column name="MONTH">7</column>
    <column name="SHIP_ID">42</column>
    <column name="CAPTAIN_ID">16</column>
    <column name="DESTIN_ID">7</column>
    <column name="TOLLIMFL">0</column>
    <column name="TOLLIMM">0</column>
    <column name="TOLLIMGR">0</column>
    <column name="TOLLIMSCH">0</column>
    <column name="TOLLIMD">0</column>
    <column name="TOLLEXFL">0</column>
    <column name="TOLLEXM">0</column>
    <column name="TOLLEXGR">0</column>
    <column name="TOLLEXSCH">0</column>
    <column name="TOLLEXD">0</column>
</table>
```

To make these datasets compatible with the Linked Data principles, the XML dump needed to be converted to RDF. As explained in Section 1.1, RDF is based upon so called triples; a concept - relation - object. One advantage of RDF is that types of relations can be described, whereas XML only indicates that there is some kind of relation.

First we look at how the data was originally organised for both datasets. The Elbing dataset has been highly structured in a manner typical for a relational database. It features a different table per concept, all of which feature a unique numerical 'key' that is used to relate to other tables. In the example XML code, this is shown by a journey having a ship called "SHIP_ID 198". There is only one ship with 198 as an ID, which is used as a key. The advantage of such a database layout is that if something needs to be added to the ship with key 198, only one table needs to be updated.

The Archangel data is actually very different. All the information is stored in one giant table and that is it. Each table simply contains 53 columns.

[17] http://www.w3.org/XML/

To decide on how the conversion should take place, the relations between all concepts was explored and defined. The entry point into the Elbing XML dataset is the 'journey'. Nothing has a relation to a journey in the XML data, whereas the journey has multiple relations to other rows. Beginning at journey, all concepts were represented in Figure 7. The same was done for Archangel, though this only lead to one concept and some 50 relations pointing to literals. Therefore, the Archangel data has been changed slightly in that additional structure was added in the conversion. This was done, for instance, by making Captains a unique resource, something that was not done in the original table. This has no effect on the output of the data, however, and therefore does not conflict with the requirement of keeping the original data intact.

We considered it good practice that any object of a relation would not be a literal until there were no further relations possible. At this point the final relation to a concept's literal was added. A literal can be seen as a label to call a resource by. As shown in the conversion example later, a relation between the concepts 'journey' elb:journey1620-3 and 'person' elb:captain_1254_johann_petersonn is established. The concept 'person' has a relation to his label (name), which is a literal as indicated by the quotation marks. The relation 'person' label 'literal' is thus defined. Had the concept 'journey' had a relation to a 'literal' "Johann Petersonn", there would be no relations possible to inform about this "Johann Petersonn".

Sometimes, a blank node is used. Blank nodes are unnamed unique resources. In this conversion, blank nodes have been used to group multiple relations to another relation. As you can not make a 'quadruple' in RDF, the blank node is added that functions as the object of one relation, and the subject of another. In the example of Elbing journey1620-3, all export tolls were grouped under the relation 'hasExportToll'. For most journeys, tolls were paid in multiple currencies. By using the blank node, we can attach multiple toll objects to one ExportToll.

The 'person' and 'ship' resources received a unique number. This number is an identifier added to ensure that this entry remains unique. In the original data, all entries were unique and keeping in line with the requirements, this is maintained after conversion. An optional conversion schema has been added that renders persons and ships with the same name the same entity, which is what would have happened if the identifier had not been added. This is based on the assumption that entries with the same names are in fact the same people, something that is not expected to be true in this database.

All town or city names found in the data, generally taken from destinations and captain hometowns, were grouped by their name. These did not receive an identifier. The reason for this is that in the Elbing dataset, these locations were searched for by their identifiers, so two resources named 'Amsterdam' in the conversion had the same identifier in the original data. However, in the Archangel dataset only names were referenced. To group these concepts using unique identifiers would create thousands of different 'Amsterdam' entries, that are likely almost all the same. Grouping them

all as one 'Amsterdam' entry, means that if there is another town with that name, that information is lost. With no information about this in the original dataset, the choice was made to group all towns with the same name together in the same concept.

The only deviation from the original data is that values of 'null' and some instances of values of '0' were ignored. It is common for database systems to automatically enter a null value into a field that has received no entry. To a database, this means 'nothing'. This was also the case for some toll information, shown in the example XML code. In these cases, an amount of '0' was paid and as such, these values could be ignored. In relational databases, the database table is generally predefined. This means that it expects an entry in a field, for instance "TOLLEXFL". In RDF, it is possible to simply omit this relation. The conversion of the XML code example given earlier, leads to the following turtle syntax:

```
@prefix elb: <http://purl.org/collections/nl/dss/
    elbing/> .

elb:journey1620-3
a elb:Journey  ;
elb:hasDate "14-4-1620"  ;
elb:hasNormalDate "1620-04-14"^^xsd:date ;
elb:hasShip elb:ship_198_fortuna  ;
elb:hasCaptain elb:captain_1254_johann_petersonn  ;
elb:hasDestination elb:amsterdam  ;
elb:hasExportToll
[a elb:Toll  ;
elb:hasTOLLEXM "47"  ;
elb:hasTOLLEXGR "30"  ;
elb:hasTOLLEXD "15"  ];
elb:hasProductflow elb:productflow1620-3  .
```

The turtle syntax has a prefix declared at the top, which is appended to the relations. The URI for 'elb:journey1620-3' is in reality: "`http://purl.org/collections/nl/dss/elbing/journey1620-3`".

3.4 Tools available

There are tools available to convert data from XML to RDF. Their uses and possibilities have been briefly reviewed for this research.

The Dutch Ships and Sailors project used ClioPatria for its conversion (de Boer et al., 2014). XML data was inserted into ClioPatra, that converts it to RDF and assigns blank nodes to each node in the tree. Afterwards a tool called 'XMLRDF'[18] was used to rewrite RDF to the data model format. This methodology was not used as the manual rewriting was extensive enough that we deemed it not to offer a time benefit over creating a manual script for these two new datasets.

XSLT is a language used to transform XML documents into other XML documents[19]. A tool built upon this language is Astro Grid [20]. This tool, however, avoids using blank nodes entirely and converts data in such a way that it can be converted back to XML. This proves a problem to the conversion of this research as blank nodes are used by design in the concept graph. Furthermore, conversion to the original XML data might prove problematic as certain concepts are added, for example to Archangel.

This proved to be a problem with other XSLT-based conversion tools. A more specific tool for adjustment is OpenRefine[21]. Originally Google Refine, OpenRefine lets users adjust data from a database. This could be used to change the syntax of the XML file to turtle. It could have been used for conversion, but as with XMLRDF, was considered too extensive to offer any real benefit over making a manual conversion. A plugin specific to RDF conversions exists for this tool, called 'RDF Refine'[22]. However, this tool had limitations similar to automatic XML converters, that the specific graph layout of Figure 7 could not efficiently be modelled.

No other tools were found that can convert the data close to the way envisioned in the concept graph of Figure 7. As such, manual conversion commenced.

The data was converted by writing several conversion scripts in the Java programming language[23]. The scripts are available for download[24]. Each dataset has its own script. For both datasets, the script first creates a temporary file from which all XML syntax is removed. Next it creates a Turtle file with customisable name and prints a predefined list of prefixes.

The Elbing conversion scripts uses a switch-statement to find different table names in sequence. This is possible due to Elbing being a highly structured dataset. Based on the found tablename, another Class is called, which is tailored to the properties of that XML table. Since all tables with the same name share the same structure, the data is converted line by line into RDF. The Archangel conversion script only really has to process one table and thus just converts in sequence.

After the conversion is done, a number of print statements enter the information into the Turtle-file, after which the next table is converted and printed, until the end of the document is reached. Originally an all purpose script was written that did a quick conversion of XML to RDF generally according to the following idea:

```
<table name>  has<column name> <column contents>
```

However, such a conversion did not allow for selection between which values would be literals and which would be concepts. Moreover, the design choice was made to have meaningful names for resources. This means that a captain would be referred to not by his CAPTAIN_ID, but by his

[18] `http://semanticweb.cs.vu.nl/xmlrdf/`
[19] `http://www.w3.org/TR/xslt20/#what-is-xslt`
[20] `http://www.gac-grid.org/project-products/Software/XML2RDF.html`
[21] `http://openrefine.org`
[22] `http://refine.deri.ie`
[23] `http://www.oracle.com/nl/java/overview/index.html`
[24] `http://www.entjes.nl/jeroen/thesis/java`

first and last name. Hence a more customised script was desired.

To add use to the data, dates have been standardised as well[25]. The Elbing data was searched through by a script and all date values were added to the file as a relation to a voyage in standardised form. Since this data was being converted anyway, it was added to the Elbing conversion file. However, it could also have been added in a separate file, as was done with Captain names. Having these in a separate file has the advantage of being possibly excluded. The Archangel had its data standardised in the same way.

Once the data conversion was completed, it was evaluated with historians as well. They were guided step by step through the RDF files, as the syntax was new to them. Once the file conversion was approved of, the visualisations could be made. After the conversion had been completed, using ClioPatria a schema for each conversion was automatically generated.

3.5 Mapping data to Dutch Ships and Sailors

With the conversions complete, linking the data to the DSS cloud began. This is done by relating concepts defined in the dataset schema to concepts in another dataset. As an example, in the Archangel data, a journey is referred to as a voyage. In Dutch Ships and Sailors, journeys are also defined as voyages, but in Elbing, these are defined as journeys. They all refer to the same concept, thus making the mapping between these concepts vital to linking these datasets to each other. As in the Dutch Ships and Sailors cloud a lot of concepts have been defined, such as a 'Voyage', a 'Captain' and a 'Ship', by relating the concepts of Archangel and Elbing to the ones in DSS, any shared resources and properties between the two datasets could potentially be seen as the same. This next piece of code shows how concepts of the Elbing dataset are defined as subClasses of DSS concepts. This does not include all concepts, it is only a representation of three classes. The full list can be found in the appendix.

```
@prefix skos: <http://www.w3.org/2004/02/skos/
    core#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix rdfs: <http://www.w3.org/2000/01/
    rdf-schema#> .
@prefix ns2: <http://purl.org/collections/nl/dss/
    elbing/> .
@prefix dss: <http://purl.org/collections/nl/dss/> .

ns2:Journey
a rdfs:Class ;
rdfs:subClassOf dss:Record, dss:Voyage ;
rdfs:label "Journey" .

ns2:Person
a rdfs:Class ;
rdfs:subClassOf foaf:Person, dss:Person;
rdfs:label "Person" .

ns2:Place
a rdfs:Class ;
rdfs:subClassOf dss:Place, skos:Concept;
rdfs:label "Place" .
```

[25] http://www.w3.org/TR/xmlschema11-2/#date

Immediately, a context graph that was devised to work with entries for Dutch Ships and Sailors also provided image results for the Elbing entries, as seen in Figure 8. As the Elbing entries have no concepts such as 'chamber', it does not show this. However, as 'master' was mapped to 'hasCaptain' in DSS, these concepts are considered equal. The coloured shapes represent different concept types. A blue ellipse is a captain or master, a brown trapezoid is a ship and the violet hexagon is a journey or record. This visualisation was originally made in ClioPatria as a way to visualise the structure of Dutch Ships and Sailors concepts. Just mapping concepts in the datasets of this research to those of DSS was enough to visualise them in this way.

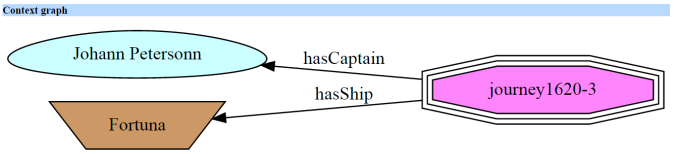

Figure 8: Elbing journeys considered as Dutch Ships and Sailors entries.

3.6 Conversion results

All data has been converted and was approved of by the historians of the Huygens ING Institute. A schema was made that maps all the classes and most of the properties to those of Dutch Ships and Sailors. On top of this, a number of new classes were added to Dutch Ships and Sailors. The practical implementation of the conversion can be found online in the triplestore[26] and the files can be downloaded from GitHub.

The schema mappings have all been published online[27] on GitHub for reviewing, some noteworthy parts will be explained here.

3.6.1 *Elbing noteworthy conversions*

The class "Account" has not been mapped. Account represents a two-letter abbreviated direction for individual goods. These directions include organisations, cities, but also individuals. It is not present in DSS yet, is very specific to this dataset and unlikely to return in a similar way, so it was not added as a new class.

The classes 'Toll', 'Value' and 'Productflow' were newly added to DSS. These were not present yet, but similar concepts are expected to return in future datasets.

3.6.2 *Archangel noteworthy conversions*

The class 'Freighter' was newly added to DSS. DSS has a slightly similar class called 'Chamber', but Freighters can be people or organisations, and is expected to possibly return in other datasets.

[26] http://semanticweb.cs.vu.nl/dss/browse/list_graphs

[27] https://github.com/biktorrr/dss_oostzee

3.7 Visualisation design

Once the data of Elbing and Archangel was linked to the Dutch Ships and Sailors cloud, truly external linking could begin. For this, the GeoNames[28] data was used. GeoNames has provided their data in RDF online. A version of this data with only Dutch places[29] was used to map to from the voyages in Archangel and Elbing. We considered just mapping to Dutch names a good proof of concept of Linked Data can be used to enrich the two datasets converted in this research, and mapping to all places in the Europe or the World was considered inefficient due to the size of those files. Moreover, mapping to GeoNames was not part of any research question derived from the interviews.

By mapping the place names to GeoNames, the datasets of this conversion gained access to coordinates stored in the GeoNames dataset. These coordinates were inserted into a Google Maps heatmap, to show the frequency some places were visited. The heatmap can be found in Figure 9. To answer the research questions derived from historians, a query to the triplestore was constructed that results in a table of shipping volume per year of the datasets of Archangel, Elbing and the sum of both. Although such a table on its own is interesting enough, the results have been mapped into a Google Chart[30]. The point is to provide historians with a visualisation of data, that is completely independent of the triplestore, instead just processing information retrieved from a search query. This is important as it shows that an visualisation can be built upon Linked Data only by manipulating a SPARQL query. Moreover, two different datasets are queried, while the results are added to one graph. The search query results are also kept very basic, meaning that any dataset that provides its entries with a standardised date can be added to the results. Any dataset that has a date entry in a non standardised form can be supplemented with a file that maps the dates to standardised dates, as was done for the datasets of Elbing. The resulting graph output for the Archangel and Elbing datasets, is shown in Figure 10.

Figure 9: Heatmap of Dutch Ports in datasets

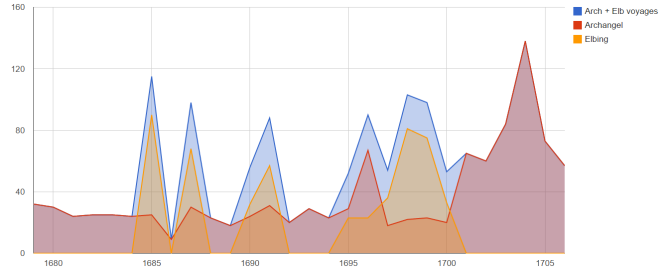

Figure 10: Graph showing voyages per year between 1675 and 1710

3.8 Results of Visualisation design

Two visualisations have been created: A Google Maps heatmap of destinations in the Netherlands and a Google Chart that shows frequency of voyages per year of each individual dataset and the sum of ships[31].

The visualisations are intentionally hosted from a different location than where the data is stored, to show that the visualisations actually retrieve current data from an external resource and that no adaptation of the data is made to fit the visualisations. In essence, both visualisations send a query via a URL request in php to the URL of the SPARQL endpoint where the data is hosted. The query is present in the code of the visualisations and can be found in the appendix. The php script is returned a string that contains all data, which is automatically converted to XML and entered into the Google Chart or map as arrayData. Script files can be found online[32].

The map has had the values for Amsterdam normalised as the heatmap would otherwise be relatively coloured and Amsterdam as a destination outnumbers others so vastly in these datasets, that no other place would be considered 'hot'. Aside from Amsterdam, only a few other cities received more than 100 visits and none more than 200, so after trying a range of limits, 100 gave the desired result. In other words, a location on the map can not get 'hotter' than 100. In comparison, Amsterdam was the destination over 1000 times.

The chart timeline can be manipulated to get a zoomed view of voyage volume. This can be compared to historic information such as war times, weather data, economic data and population growth. The graph does not add the data itself but receives this from the SPARQL query. This was done to ensure that no secondary enrichment of the data was done, the graph is merely a visual representation of the query results.

[28] http://www.geonames.org
[29] http://www.entjes.nl/jeroen/thesis/geonames.txt
[30] https://developers.google.com/chart/?hl=nl
[31] http://www.entjes.nl/jeroen/thesis
[32] www.entjes.nl/jeroen/thesis/scripts

4. RESULT IN NUMBERS

The conversion results and visualisation results in numbers are summarised in Table 1. This table shows the file and number of triples.

File description	Triples
Archangel triples	154,031
Elbing triples	128,509
Archangel same as links	35,351
Elbing same as links	14,705
Elbing schema mappings	241
Elbing schema	215
Archangel schema	210
Elbing maps to GeoNames	197
Archangel schema mappings	195
Archangel maps to GeoNames	162
New Dutch Ships and Sailors schema entries	36

Table 1: The results in numbers.

The Java files accumulate to two separate programs (one for each conversion) and a total of 26 classes, 8 for the Archangel conversion and 18 for the Elbing conversion, each ranging from 100 to 500 lines of code in length. The visualisations use 4 php files and 2 html files, visualised in the online environment through an inline frame.

5. EVALUATION

A short list of questions has been set up for this research and used to evaluate the data. The questions are tailored to be used in conjunction with the research questions that have been requirement engineered, while evaluating the visualisations. The goal was to get the opinion of the historians on these data visualisations and the questions are to serve as a guideline. The list is:

1. Does this representation offer enough information to answer the research questions?
   (a) For which questions does it and for which does it not?
   (b) Any remarks as to why it does or does not?
2. These graphs have been made possible by converting the original data. Do you believe it enriches the data?
3. Do you feel these datasets provide a meaningful contribution to the existing DSS data?
   (a) Any remarks as to why it does or does not?

These questions were given to the historians interviewed earlier, their remarks and answers have been combined and summarised below:

1. Does this representation offer enough information to answer the research questions?
   (a) For which questions does it and for which does it not?
   (b) Any remarks as to why it does or does not?

   The chart can spur questions about fluctuations in volume of travels. Maybe this can be related to economic growth or wars. It can possibly say something about the history of Dutch shipping, but it all depends on how complete these datasets really are. The ability to see a combination of data is very nice, for example, it shows that almost always there were more ships headed towards Elbing than to Archangel. The way the chart allows selection of date ranges also works beautifully. This is an obvious contribution over the data representations already in our posession.

   The graphs give rise to a lot of questions. Why does the Elbing data spike so heavily? Yet Archangel does not? My first thought is that the source material could be lacking, but that can not be derived from the representation. For a good comparison, these peaks should be normalised, for example by taking a 5-year average.

   Another option is to visualise the data by creating dots instead of a line.
2. These graphs have been made possible by converting the original data. Do you believe it enriches the data?

   Yes, but mind the remarks. It is a big plus that with relatively little effort, previously separate data can be related to each other.
3. Do you feel these datasets provide a meaningful contribution to the existing DSS data?

   Yes as the goal of Dutch Ships and Sailors is to create a representation of Dutch shipping that is as complete as possible. Moreover, these sets contain data not previously contained in DSS and therefore increase its depth.

   (a) Any remarks as to why it does or does not?
   See above.

Further remarks made:

A lot of the research questions require accurate data to compare the representations with. To link weather data or economic growth to shipping, an accurate representation of that data needs to be provided, which is not the case. On second thought the use of the representation is not limited to answering questions, but also asking enabling you to **ask** *questions.*

6. RECOMMENDATIONS BASED ON RESEARCH

Based on this research, some recommendations can be made for future conversions to Linked Data of maritime historic data to the Dutch Ships and Sailors cloud.

- Set up data requirements.

  For this research this was done with historian interviews, however, the requirements depend entirely on the dataset to be converted, and as such also the method in which they are established. Setting them up early will offer the possibility to stay on track during the research. It allows reviewing of tools that could save

time and it helps getting an understanding of the current data structure as well as the structure it will have once converted.

- Predefine RDF structure based on requirements.

  By taking an example from the data and converting this, problems can be uncovered early.

- Convert tables based on RDF structure.

  Using a specific conversion script per table allows each table's specifics to be modelled correctly. If it is possible to convert all data with one script, do it. However, in both datasets used for this research, specific scripts were needed.

- Map to existing concepts for data enrichment.

  Your Linked Data will be more meaningful if it reuses concepts from other datasets. This can easily be done by using mappings that define whether a resource in your data is an instance of other data, or entirely the same as other data. Mapping can be done in the schema, and on a data level.

7. DISCUSSION

The possibilities of converting digital data only go as far as the integrity of the source. The Elbing dataset lacks records due to a fire in the Elbing town hall in the 18th century (Lindblad et al., 1995) and are therefore not likely to be complete. The Archangel dataset has largely been the work of one historian, who used his own logic to denote information. There is not much documentation available on the conversion, as opposed to Elbing. Moreover, because of the structure of the Archangel database, it is more error prone than the Elbing database.

The main intention for the visualisations was to provide this newly converted Linked Data with graphical support and to show how Linked Data offers the possibility to combine information from different datasets in a meaningful way. However, there are some unresolved issues with the visualisation modelling: The Archangel dataset does not discern between cities, provinces or countries when it comes to directions. This results in some problems for the heatmap in Figure 9. For example, in the middle of the map, just above the word 'Nederland', a dot can be found close to the coast. This actually refers to the Dutch village of 'Frankrijk', which is also the Dutch name for the country France. In the dataset, it most likely refers to the country.

In the east of the Netherlands on the heatmap, a blue dot on the German border can be found. This refers to the village of 'Zeeland', also the name of the southwestern Dutch province that was a major shipping hub.

In the chart in Figure 10, the data is reliant on correct input. The SPARQL query can only process dates that follow a standardised date syntax. Although not frequently appearing in these datasets, sometimes a date is unknown or rounded to a year. The interactive chart provided online claims all dates to be at 'January 1st' of a year, as this is appended in the chart data. The input for the charts is years but a date scale requires the month and day as well.

The heatmap visualisation uses GeoData information that is based on coordinates of current cities. The information stored in the datasets were the names given by who entered the data. When spelling changes over time, some of these might not correspond anymore. Furthermore, matches with the GeoData file are sometimes incorrect, as some cities carry the same name, or even the names of countries. No heuristics or optimisation were applied on the data, which means that although it would generally be easy for a human to know whether the country 'Frankrijk'(France) is meant or the village, to this conversion they were the same.

The information of GeoData is based on the geographical center of cities currently, which might not be in the same place as in the past.

No ready to use RDF information about any of the research questions set up with the historians was found. For example RDF data of Dutch economic growth during this time could not be found. The original intent for the visualisations was to combine such data with these datasets. However, if some resources were found providing some information, the chart can be compared to this.

Although these accounts come from official sources, it is possible that unknown amounts of data have been lost if they were never recorded in the first place. Toll collectors could possibly be prone to bribery, not recording some information in exchange of some personal gain.

Finally, datasets converted only represent a part of total Dutch shipping in the baltic trade, so generalised historical conclusions on this data can not be made.

8. CONCLUSION

New technological possibilities can be used by historians in support of their research. Linked Data is one such possibility: a means of accessing information in the world online through URIs. With historic data being digitised, the Dutch Ships and Sailors project aims to have as much as possible of the Dutch maritime history available as Linked Data. With the formal completion of the Dutch Ships and Sailors project, its initial goals did not change. More maritime data can still be added to the Dutch Ships and Sailors cloud. This study has researched how additional datasets can effectively be linked to those of the Dutch Ships and Sailors project and how additional datasets can assist in answering existing questions in the field of History.

Following Linked Data standards, and by use of guidelines set up with requirements engineering, it has been shown that a dataset can be converted to Linked Data without loss of information. After the digitisation has been completed, by use of concept mapping, the information stored in one dataset can be linked to information stored in another, completely separate dataset, without changes to the data. Of the datasets converted, the Elbing dataset has had all its original structure and data kept intact. Essentially, only the way the information is stored (originally as XML in a table, now as RDF in a triplestore) has been changed and then the concepts have been mapped, using a separate file. The Archangel dataset has had no changes to its content, but some to its structure. Without any information

loss, however, this upheld all conditions set for an effective link. As such, we consider the dataset effectively linked to the Dutch Ships and Sailors cloud and have provided recommendations on how to do this in future work.

After the conversion was completed, two simple visualisations were designed that would access the data from a remote server. They performed a data request in the form of a SPARQL query and processed the returned information in their own local environment. These graphical representations have been evaluated by historians to enrich the data by offering a relatively low effort ways to compare separate datasets to each other. Furthermore, the representations immediately sparked new questions about the possible causes for data fluctuations. However, just these datasets are not enough to assist in answering existing questions in the field of History. More digitised information that can be linked to each other would provide some much desired answers, and hopefully even more questions.

9. FUTURE WORK

Future work includes the continued digitisation of naval records, converting them to RDF and placing them online. This will help build future visualisations and compare data more detailed, as the datasets present in Dutch Ships and Sailors including these two new ones are simply too small to do research on. On these two datasets in particular, additional conversions can be performed that discern destination ports from each other. Some work can be done on providing standardisation formats for DSS conversions and recurring concepts could be compared with each other. In this research, captains are estimated to be the same when they share first and last name, but a more thorough comparison can likely be worked on for both captains and ships.

There are also many aspects of these datasets left unexplored by this research. Elbing offers toll and good value information, that have been converted but not used. Archangel also offers information on goods transported and even the intended price they were supposed to fetch. However, as this data was unique to each dataset, it had little use to this particular research.

Additional work can be done on these datasets by comparing them and existing datasets in the Dutch Ships and Sailors cloud to external resources. For this research, no ready to use RDF information was found but many historical sources provide estimates (Israel, 1989; Van Zanden & Van Tielhof, 2009). In future research, these datasets can be compared to such sources.

Acknowledgements

Vital to this research has been Victor de Boer as a supervisor, but also by providing support in the form of technical and theoretical knowledge, assistance in creating the SPARQL queries and patience.
This research could also not have been made possible without the historians at the Huygens ING Institute who have digitised these datasets and provided their feedback and time. In particular, dr. Rik Hoekstra and dr. Milja van Tielhof have provided a great deal of assistance.

References

Brandt, K., & de Boer, V. (2013). *Linked data for iati* (Unpublished doctoral dissertation). MSc Thesis, Vrije Universiteit Amsterdam.

de Boer, V., van Rossum, M., Leinenga, J., & Hoekstra, R. (2014). Dutch ships and sailors linked data. In *The semantic web–iswc 2014* (pp. 229–244). Springer.

Ebert, C. (2011). Requirements engineering. *Global Software and IT: A Guide to Distributed Development, Projects, and Outsourcing*, 37–44.

Heath, T., & 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.

Israel, J. I. (1989). *Dutch primacy in world trade, 1585-1740*. Oxford University Press.

Lindblad, J. T., Dufour-Briët, F., & DeCoursey, R. (1995). *Dutch entries in the pound-toll registers of elbing: 1585-1700*. Instituut voor Nederlandse Geschiedenis.

Meroño-Peñuela, A., Ashkpour, A., Erp, M., Mandemakers, K., & Breure, L. (2013). Semantic technologies for historical research: A survey. *Semantic Web Journal*.

Robson, C. (2002). *Real world research: A resource for social scientists and practitioner-researchers* (Vol. 2). Blackwell Oxford.

Robson, C. (2011). *Real world research: a resource for users of social research methods in applied settings*. John Wiley & Sons.

Runeson, P., & Höst, M. (2009). Guidelines for conducting and reporting case study research in software engineering. *Empirical software engineering*, *14*(2), 131–164.

Schreibman, S., Siemens, R., & Unsworth, J. (2008). *A companion to digital humanities*. John Wiley & Sons.

Van Zanden, J. L., & Van Tielhof, M. (2009). Roots of growth and productivity change in dutch shipping industry, 1500–1800. *Explorations in Economic History*, *46*(4), 389–403.

Yu, L. (2011). Linked open data. In *A developerâĂŹs guide to the semantic web* (pp. 409–466). Springer.

APPENDIX

The online Appendix can be found at: `http://www.entjes.nl/jeroen/thesis/appendix`