Dutch Ships and Sailors Linked Data

Victor de Boer[1], Matthias van Rossum[2], Jurjen Leinenga[3], and Rik Hoekstra[3]

[1] Netherlands Institute for Sound and Vision, Hilversum, the Netherlands
Dept. of Computer Science, VU University Amsterdam, the Netherlands
v.de.boer@vu.nl
[2] Dept. of History, VU University Amsterdam, the Netherlands
m.van.rossum@vu.nl
[3] Huygens ING Institute for Dutch History, Den Haag, the Netherlands
leinenga@xs4all.nl, rik.hoekstra@huygens.knaw.nl

Abstract. We present the Dutch Ships and Sailors Linked Data Cloud. This heterogeneous dataset brings together four curated datasets on Dutch Maritime history as five-star linked data. The individual datasets use separate datamodels, designed in close collaboration with maritime historical researchers. The individual models are mapped to a common interoperability layer, allowing for analysis of the data on the general level. We present the datasets, modeling decisions, internal links and links to external data sources. We show ways of accessing the data and present a number of examples of how the dataset can be used for historical research. The Dutch Ships and Sailors Linked Data Cloud is a potential hub dataset for digital history research and a prime example of the benefits of Linked Data for this field.

Keywords: Digital History, Maritime Data, Heterogeneous Data Cloud

1 Introduction

As (digital) humanities researchers seek more (international and cross-domain) collaboration, integrating humanities datasets becomes more important to those researchers. One subdomain where this is very much prevalent is in (social) historical research. Often historical researchers collect data from historical archives for their specific research questions. However, these datasets are often not presented in sharable formats to other researchers. If they are shared at all, the datasets are published in a multitude of formats. To further the digital history agenda, it has been recognized that representing and sharing data is key [4, 10]. Using Linked Data principles and practices, we can integrate generic data with smaller datasets that have been created with a specific historical research goal. Linked Data allows us to publish these datasets using the modeling principles of the original datasets, while -through the use of (schema) links- still achieving a level of integration. In this paper, we present the Dutch Ships and Sailors (DSS) data cloud. This Linked Data cloud brings together four Dutch maritime historical datasets, each with its own datamodel. The data is available as five-star linked data making sharing and reuse possible. The data is integrated at

a meta-level through common vocabularies and linked to generic external data sources allowing for new types of queries and analysis.

As a sea-faring nation, a large portion of Dutch history is found on the water. The maritime industry has been central to regional and global economic, social and cultural exchange. It is also one of the best historically documented sectors of human activity. Many aspects of it have been recorded by shipping companies, governments, newspapers and other institutions. In the past few decades, much of the data in the preserved historical source material has been digitized. Among the most interesting data are those on shipping movement and crew members (cf. [15]). However, much of the digitized historical source material is still scattered across many databases and archives while still referring to common places, ships, persons and events. By linking the different available databases, the data can complement and amplify each other, and new research possibilities open up. The DSS datacloud bring together the rich maritime historical data preserved in four of these different databases. Two of these databases have been used extensively in historical research and by presenting them in this interoperable format, future reuse is likely to be easier.

The presented dataset is significant to the digital history community since it brings together seminal datasets on maritime history in a re-usable and integrated way. The complexity of the original data is retained and not 'dumbed down' to a specific data model for online presentation. At the same time, multiple enrichments have been performed and additional enrichments are possible at a later stage. The four datasets together integrated can serve as a pivot data cloud for international maritime historical datasets as well as for other (Dutch) historical datasets. The work here is also significant to the broader Linked Data community since it presents a prime example of how collaboration between historians and computer scientists can lead to high-quality digital history datasets that are actually trusted and used by the historians. Digital humanities is a rapidly growing field in which it is recognized that Linked Data presents interesting opportunities. Furthermore, this datacloud presents the results of a method where individual datasets are converted to RDF, maintaining their own datamodel but are integrated through RDF(S) links into a datacloud. This methodology can be re-used in other multi-part datasets.

2 General Approach

We here describe the general conversion pipeline and modeling principles. Section 3 describes the specific datamodels and conversion steps.

2.1 Conversion and modeling pipeline

The conversion and modeling pipeline is based on previous work described in [6] where more details about the methodology and tools can be found. We here give a brief overview. In a first step of the generic pipeline, we have data available in some XML format. For datasets, not available as XML, we use simple syntactic

export functions[4]. The output of the pipeline is linked RDF, corresponding to a specific datamodel. The pipeline is built on the ClioPatria semantic server (`http://cliopatria.swi-prolog.org`). ClioPatria is an RDF triple store that through a web interface provides feedback on the (intermediary) produced RDF, which is crucial for the interactivity of the conversion and modeling. We start by ingesting the XML into ClioPatria, which converts the XML tree into a raw RDF graph, assigning blank nodes to each node in the tree.

Graph restructuring The ClioPatria XMLRDF [5] package is a tool for restructuring an RDF graph using *graph rewrite rules*. In the second step, the crude RDF is rewritten to RDF adhering to a data model format, using handwritten rules which are interpreted by the XMLRDF tool. These rules are constructed in an iterative interactive process[6]. In this step, some blank nodes from the rough RDF graph are assigned URIs and resources and triples can be copied, merged, replaced or deleted. Depending on the datamodel, some literal values are consolidated to RDF resources. For each dataset, we also generated an RDFS schema which lists the produced classes and properties and relates them to the more generic DSS schema (see Section 3.5). ClioPatria provides support for this by presenting the user with a schema template based on the RDF data loaded.

Linking We establish links to external resources. This can be done using either the XMLRDF tool, for example when in dataset A there is an explicit reference to a unique identifier in dataset B. When linking requires more complex techniques, we employ the ClioPatria package Amalgame[7]. Amalgame is an iterative alignment platform that allows a user to mix-and-match multiple label- and structure-matching algorithms as well as filtering operations into an alignment workflow. The tool is used to establish identity or other semantic relations (e.g. broader/narrower) between concepts and instances.

2.2 Generic modeling decisions

Resources and URI schema RDF Resources can be either blank nodes or receive a URI. In general, we only use blank nodes to group properties. An example is given in Section 3.1, where statistics about specific crew membership are grouped. Any resource that is considered to be a meaningful 'thing' is assigned a URI. This includes resources that might be linked to from outside of the dataset. URIs are typically created from an identifier metadata field (such as the original database record ID). Within the DSS cloud, we have defined five namespaces: `http://purl.org/collections/nl/dss/` for DSS generic data and `http://`

[4] For example, for MS Excel files, we use the built-in export function.
[5] `http://semanticweb.cs.vu.nl/xmlrdf/`
[6] The XMLRDF scripts used for the DSS datacloud are found online at `https://github.com/biktorrr/dss/tree/master/script`
[7] `http://semanticweb.cs.vu.nl/amalgame/`

purl.org/collections/nl/dss/gzmvoc/, http://purl.org/collections/nl/dss/mdb/, http://purl.org/collections/nl/dss/das/ and http://purl.org/collections/nl/dss/vocopv/ for the four datasets. In this paper we abbreviate URIs with the respective CURIEs[8] dss:, gzmvoc:, mdb:, das: and vocopv:. We use PURL URIs that redirect to a ClioPatria instance, this allows for persistence of the URIs even beyond the life expectancy of the project or any specific institute.

Linked Data for Multilayered enrichment In some cases, new resources are created, where in the original metadata, there are only literal values. We do this specifically to group properties about things that are separately identifiable and that might reoccur in the datasets. Specifically, we do this for *persons, places, ships, ship types* and *ranks*. In most cases, the original literal values are retained and a new resource is created in a separate named graph with its own provenance information. An example of this is shown in Figure 2. By not 'hard coding' the enrichment but separating the enriched data from the original data, we can benefit from the latter, while still always being able to go back to the original data. This corresponds to an important requirement as put forward by the historical researchers.

Another important modeling decision that is partly specific to the domain is that for most types of resources, we assume that they are unique, even though they have a number of metadata fields in common. For example, two records (say from 1850 and 1851) might both refer to a person "Piet Janssen" who sailed on the ship "Alberdina". We do not assume that these are the same person, and therefore assign them separate URIs. This was an explicit modeling decision taken in collaboration with the historians, since many Dutch names are common and often fathers and sons with the same first and last name sailed on the same ships. Therefore, in the basic data, we assume that all persons and ships are unique and assigned separate URIs. At a later stage, automatic or manual methods can be used to establish identity links. In Section 3.2, we describe this effort for one of the datasets.

Mapping properties and classes to DSS interoperability layer We model the datasets using separate datamodels with their own properties and classes and do not use common classes or properties directly in the individual datasets. Rather we use subproperty and subclass relations to map our classes and properties to common ones (either in the DSS domain or to external schemas). This way we can retain the specificity of the dataset and the intended semantics of the model and still allow for reasoning and querying at the interoperability level (DSS). For example, the notion of a ship name is slightly different amongst the datasets even though they use the same field name. In some cases, some normalization process has taken place in the original archive data and in other cases it has not. These (sometimes subtle) differences are regarded as crucial by the

[8] http://www.w3.org/TR/2007/WD-curie-20070307/

historians and they need to be maintained in the converted datasets to ensure trust and usage. This example is shown in Figures 2 and 3. The DSS schema itself is mapped to often-used schema's. Other than RDF(S) these are: SKOS[9] to describe concepts schemes (ranks, ship types,...); FOAF[10] to describe person information; and Dublin Core terms[11] to describe record information (description, identifier,...). ClioPatria as well as many other triple stores supports RDFS entailment in its SPARQL interface and can therefore exploit these mappings.

2.3 The role of provenance and named graphs

Provenance plays an important role in historical research and specifically in archival research. The origin and history of archival data is crucial to estimate the scientific value of data [13]. This holds even truer for digital data, where in many cases its provenance is unknown or lost. For Linked Data, the provenance of resources can be modeled using the PROV-O ontology[7]. In the DSS cloud we model the provenance on the named graph level. Each named graph is a separate set of triples that come from one source. This can be either (a table in) an original data source, or the result of an enrichment or linking process. In the DSS cloud, each RDF named graph has a URI that is defined also as a `prov:Entity`. This URI is the subject and object for the provenance triples, including those listing the different conversion activities and the human and software agents involved in the conversion. We also refer to the original data sources and their web URIs as far as they are present. All the provenance triples are stored in a separate named graph[12].

Next to provenance information, for automatically derived data we list the content confidence[12]. This provenance information allows for SPARQL queries that include or exclude triples from specific named graphs because they are the result of an operation of a software agent or because they have a too low content confidence value. For a total of four link sets we performed a structured manual evaluation of random samples by the domain expert. For these named graphs we assign confidence levels based on the evaluation results.

3 The Datasets

In this section we describe the individual datasets. The first two are modeled and converted in close collaboration with the historical researchers responsible for the source datasets and we describe them in more detail. The third and fourth datasets are conversions of previously published historical datasets and are described less elaborately. They were converted with the help of the historians. We also list the main statistics in Section 3.6 as well as describe the interoperability

[9] http://www.w3.org/2004/02/skos/
[10] http://xmlns.com/foaf/0.1/
[11] http://dublincore.org/documents/dcmi-terms
[12] `http://www.dutchshipsandsailors.nl/data/browse/list_graph?graph=http://purl.org/collections/nl/dss/dss_provenance.ttl`

layer and links. Figure 1 gives an overview of the entire DSS data cloud and the internal and external links.

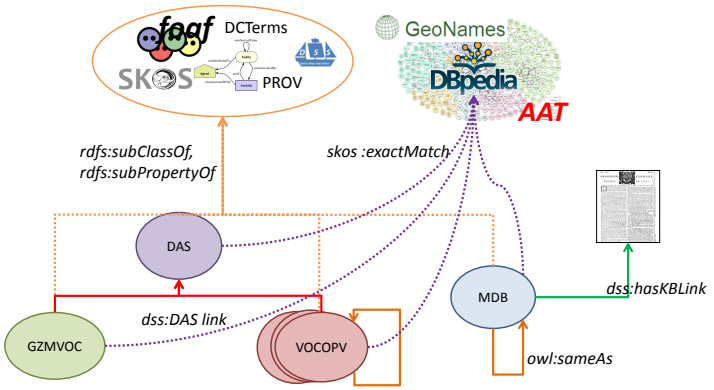

Fig. 1. The Dutch Ships and Sailors Linked Data cloud. The individual datasets are represented by ovals in the bottom half of the image. Internal links are represented by arrows. External links are represented by dotted arrows.

3.1 GZMVOC

Original data The "Generale Zeemonsterrollen VOC" (GZMVOC) (*en:* "General sea muster rolls VOC") is a dataset describing the crews of all ships of the Dutch East India Company (VOC)[13] from 1691–1791. The data was gathered by a Dutch social historian Matthias van Rossum (co-author of this paper) in the course of his research on labor situations for European and Asiatic crews on Dutch VOC ships. The data is based on archival records from the VOC itself and lists data of all ships that sailed between Europe and Asia. The data consists of the size of the crew as well as its composition (number of European and Asiatic sailors, soldiers and passengers). In a number of occasions the location of the ship on the moment of counting -the month of June of each year- as well as data on the name and type of ship. Where possible, details on the Asiatic crew members are listed, including wages, job descriptions, place of origin, categorization and hierarchical structure. For ships with a mixed European and Asiatic crew, often data about the captain and offices is listed. In this dataset, references to the Dutch Asiatic Shipping (DAS) dataset are present through numerical IDs (see Section 3.4). The original data was presented as a Microsoft Excel file, which we exported to XML.

[13] `http://en.wikipedia.org/wiki/Dutch_East_India_Company`

Data model and conversion An initial RDFS datamodel for GZMVOC was derived from the structure of the Excel sheet as well as documentation provided. After that, the model was corrected and refined in close collaboration with van Rossum. The primary citizens of this dataset are records (countings) which are the subjects of locations, registration numbers, etc. Counts of Asiatic and European crews are grouped using blank nodes, rather than linking numbers directly to individual records. Each record is connected to a ship resource, which groups information assumed to be persistent beyond the counting such as the ship name and type. For the captain, a resource is also created, with name and birthplace information. Several literal values are consolidated to resources, including ship types, ranks and places, to allow for later linking. The original triples with literals are always retained.

After ingestion and conversion to raw RDF. A total of 10 XMLRDF rules were created to restructure the graph to match the datamodel. The results were verified by van Rossum by inspecting a number of resources by hand. In total 110,986 triples are stored in the GZMVOC main data named graph`gzmvoc:gzmvoc_data.ttl` [14] (see Table 3.6 for all graphs and statistics). A further 591 triples make up the consolidated places and 166 triples make up a small vocabulary of ship types and ranks. This is the smallest dataset in the DSS datacloud. The figure below shows a small sample of the RDF graph for GZMVOC.

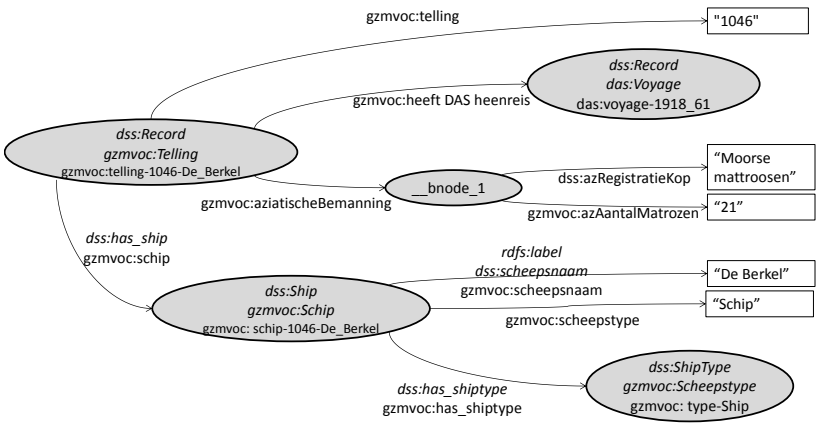

Fig. 2. Small sample of the RDF graph for GZMVOC showing a counting, a linked ship and detailed counting information connected to a blank node. Resources are represented using ovals, with the URI at the bottom line under italicized superclasses above, properties are represented by arrows, with property URIs next to them and their superproperties italicized. Literals are represented using boxes.

[14] `mdb:mdb_data.ttl`

Links The referenced identifiers of the DAS dataset are used to establish RDF links to resources in that dataset using a simple lookup script. There are two types of properties linking GZMVOC and DAS: one representing outgoing journeys (`gzmvoc:has_das_link_heen`) and one representing homebound journeys (`gzmvoc:has_das_link_terug`). Those link triples are stored in a separate named graph, to enable listing separate provenance information. A total of 5,303 link triples are stored.

3.2 MDB

Original data The "Noordelijke Monsterollen Databases" (MDB) (*en:* "Northern muster rolls databases") is a dataset describing mustering information found in mustering archives in the three northern Dutch provinces (Groningen, Friesland, Drenthe)[15] in the period 1803–1937. The original Noordelijke Monsterollen Databases (MDB) was provided as a SQL dump file by the original maker of the data, historian Jurjen Leinenga (also co-author of this paper). The database consists of two tables, one with records of ship muster rolls and one with records of person-contracts, related to those muster rolls. The SQL dump was loaded into a MySQL database and exported to XML. This resulted in two XML files, one for the ship records and one for the person records.

Data model and conversion The datamodel was developed interactively in collaboration with the historian, based on the original SQL data model and extensive written documentation. In this model, the two main classes are a "Person Contract", and "Mustering". A Person Contract holds information that is subject to change, including ranks, wages and time stamps. The Person resource is used for persistent information such as names, birth place etc. The same choices are made for "Mustering" which holds specific information about a mustering of a ship on a specific date. It is related to exactly one ship resource, which holds persistent information about that ship (name, type, ...). Figure 3 shows an example graph snippet. The complete RDFS datamodel is found in the named graph `mdb:mdb_schema.ttl`[16]. The main data graph `mdb:mdb_data.ttl` has 1,296,641 triples, with 27 predicates and 8 classes.

The conversion script for the MDB dataset is composed of 20 rewrite rules and can be found at `https://github.com/biktorrr/dss/blob/master/script/rewrite_mdb.pl`. To ensure unique URIs "Mustering" URIs are constructed using internal identifiers plus a code for the archive it originates from (this archive is also a resource in the dataset itself). For ship, person and URIs, we add expand this URI with the name of the ship, person etc. Places, ranks and shiptypes are consolidated to place resources.

[15] `http://en.wikipedia.org/wiki/Provinces_of_the_Netherlands`

[16] For brevity, we shorten graph URIs with CURIEs. The expanded URIs are dereferenceable

Internal links In the MDB dataset many ships occur multiple times, however it is initially unknown which ships are which. We therefore assume that all ships are unique and only at a later state attempt to identify recurring ships. For this enrichment, multiple algorithms were designed and implemented. A sample of the results was evaluated by Jur Leinenga and a subset with an acceptable precision was found (0.95). More details about the linking are found in [14]. The links are stored in a separate named graph (`mdb:mdb_ship_sameas.ttl`) with appropriate provenance and content confidence metadata. A total of 33,435 sameAs links are established.

External links One of the more interesting external links are those from DSS records to digital historical newspaper articles from the Dutch Royal Library (KB)[17]. The linking algorithm uses a number of features such as ship names, captain names, time constraints and automatically derived indicator phrases for maritime events (such as "left port", "sailing for" etc.) to establish likely links between MDB records and KB articles. Multiple versions of the algorithm were developed, focusing either more on precision or on recall. For each version, random samples of the results were evaluated manually by Jurjen Leinenga. More details about the linking can be found in [1]). In the end, it was decided that the results of a high-precision version (precision here is 0.90) of the algorithm were consolidated and added to the datacloud as a separate named graph (`mdb:mdb_2_kb.ttl`) with appropriate provenance and content confidence metadata. Links are manifested as RDF links between MDB musterings and external KB paragraph URIs. Figure 3 shows such a link. Note that the KB as of yet does not provide RDF after dereferencing, rather an XML snippet with the text of the newspaper article is returned. In total 179,120 `dss:has_kb_link` triples are stored.

3.3 VOCOPV

The original dataset "VOC Opvarenden"[17] is the result of a manual digitization of the personnel data of the VOC in the 18th Century. The original data consists of three separate parts (*en:* 'voyagers', 'salary books' and 'beneficiaries') and was downloaded as a CSV file from DANS Easy website[18]. It was converted to an XML version using a simple python script.

The XML version was then converted to RDF with an XMLRDF rewriting script. The model was developed in a collaboration between the authors, based on the original data model, expert knowledge and documentation available. There are three main classes: "Voyager"; "Salary Book", which links to ships and "Beneficiary". Links are present between instances of each of these classes. The complete RDFS datamodel is found in the named graph `vocopv:vocopv_schema.ttl`. With more than 22 Million triples, this is the largest dataset in the DSS cloud.

[17] `http://kranten.delpher.nl`
[18] `https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:33602`

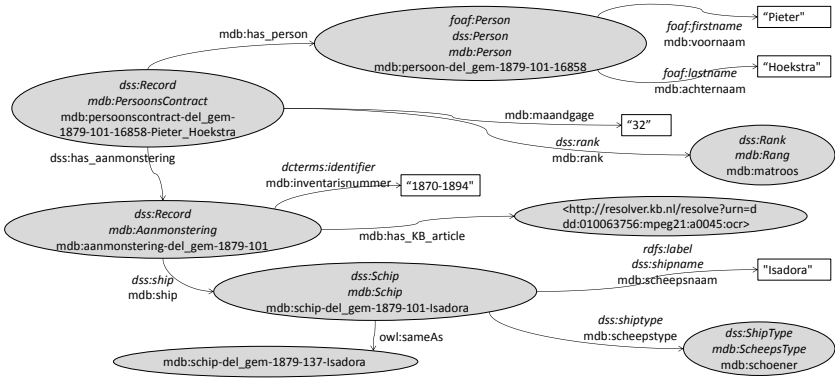

Fig. 3. Small sample of the RDF graph for MDB showing a person-contract and linked person; the counting (mustering) and a linked ship. Also shown are an internal owl:sameAs and an external link to a KB newspaper resource. For a number of properties, we list the DSS-superproperties in italics.

The original VOC Opvarenden dataset uses explicit references to DAS voyages. These were used to generate explicit links between VOC Opvarenden and DAS. We use three different RDF properties, which correspond to the original metadata fields. All links are stored in a separate named graph (`vocopv:vocopv_2_das.ttl.gz`). In total 1,128,416 links are established.

3.4 DAS

The Dutch Asiatic Shipping (DAS) dataset contains data regarding outward and homeward voyages of more than 4,700 ships that sailed under the flag of the (VOC) between 1595 and 1795. The dataset is a conversion of a previously digitized DAS dataset hosted at Huygens ING [3] at `http://resources.huygens.knaw.nl/das/index_html_en`. Between 1595 and 1795 the Dutch East India Company (VOC) and its predecessors before 1602 equipped more than 4,700 ships to sail from the shores of the Netherlands bound for Asia. More than 3,400 ships made the return voyage home. The reference work Dutch-Asiatic Shipping has classified these voyages on which Dutch trade between Europe and Asia was founded in a systematic survey.

The original Dutch Asiatic Shipping data was downloaded as a CSV file. That data was converted to an XML version using a simple python script (available on Github). The XML version was then converted to RDF with an XMLRDF rewriting script. The model was developed in a collaboration between the authors and is based on the original data model, expert knowledge and documentation available. Here the main class is "Voyage" detailing a specific voyage of a VOC ship either from the Netherlands to Asia or back. The complete RDFS datamodel is found in the named graph `das:das_schema.ttl`. The main data graph `das:das_data.ttl` has 149,357 triples, with 21 predicates and 6 classes.

3.5 Generic DSS data

As was described in Section 2.2, places, ship types and ranks were consolidated to resources so that they can be linked to internal or external data sources. For a number of ranks and ship types, a manually constructed SKOS thesaurus was created by the historians. In a DSS schema (`dss:dss_schema.ttl`), manually defined 7 classes that are common to the four datasets (Ship, Chamber, Sailor etc.) as well as three DSS specific properties (eg. ship name), found in `dss:dss_schema.ttl`. The other properties and classes in the interoperability layer are from SKOS, FOAF or DCTERMS. We use `rdfs:subPropertyOf` and `rdfs:subClassOf` triples to relate the properties and classes to this layer.

Identity Links Although *GeoNames*[19] does not provide historical place information, it still is a very usable source of information, providing lat/long coordinates, hierarchical information and place names in other languages. It is a much linked-to data source on the Web of Data, thereby increasing the reusability of the DSS data. Place names from all four datasets are aligned with the GeoNames dataset, but only for the subset of Dutch places[20]. For this, we used the Amalgame toolkit using simple label matching algorithms. The links are stored in a separate named graph (`dss:al_all_place_2_geonames.ttl`). We used `skos:exactMatch` properties to link DSS place names to GeoNames resources. In total 2,510 links are established. These place names are spread over the four datasets. *The Getty Art and Architecture Thesaurus (AAT)*[21] lists many concepts that are relevant for our dataset, for example ship types and ranks. We use a version of the AAT that has Dutch language labels making it possible to semi-automatically link DSS ranks and ship types to AAT. The mappings were based on matching labels and performed by Amalgame. The links are stored in the named graph (`mdb:ranks_and_shiptypes_1.ttl`) A total of 75 concepts were matched. We finally link ranks and ship types to *DBPedia*[22] again using the Amalgame alignment tool. A total of 123 links are established and stored in `dss:dbpedia_links.ttl`

3.6 Statistics

In Table 3.6, we list the named graphs that make up the DSS datacloud. For each named graph we list the URI, the number of triples and the dataset it belongs to[23]. A number of linked external data sources are also loaded to allow for single access-point SPARQL querying.

[19] `http://www.geonames.org`

[20] We are planning on expanding the links and adding those as separate named graphs to the data. Initial experiments linking to Indonesian locations have been performed.

[21] `http://www.getty.edu/research/tools/vocabularies/aat/`

[22] `http://dbpedia.org/`

[23] A live version of these statistics can be seen at `http://www.dutchshipsandsailors.nl/data/browse/list_graphs`

RDF Graph	Triples	Dataset
vocopv:vocopv_opv.ttl.gz	19,104,514	VOC Opvarenden
vocopv:vocopv_sol.ttl.gz	2,231,367	VOC Opvarenden
mdb:mdb_data.ttl.gz	1,296,641	MDB
vocopv:vocopv_2_das.ttl.gz	1,128,416	VOC Opvarenden
vocopv:vocopv_beg.ttl.gz	636,333	VOC Opvarenden
http://sws.geonames.org/geonames-NL.ttl	309,678	External
http://e-culture.multimedian.nl/ns/rkd/aatned/aatned.rdf	264,968	External
mdb:mdb_2_kb.ttl	179,120	MDB
das:das_data.ttl	149,357	DAS
gzmvoc:gzmvoc_data.ttl	110,986	GZMVOC
http://sws.geonames.org/geonames_nl_as_skos.ttl	42,811	External
mdb:mdb_ship_sameas.ttl	33,435	MDB
vocopv:vocopv_gen_thes.ttl	12,851	VOC Opvarenden
das:das_thes_gen.ttl	7,034	DAS
dss:dbpedia_links.ttl	5,449	External
gzmvoc:gzmvoc_2_das.ttl	5,303	GZMVOC
http://sws.geonames.org/ontology_v2.2.1.rdf	2,895	External
dss:al_all_place_2_geonames.ttl	2,528	DSS (all)
mdb:mdb_thes_places.ttl	2,273	MDB
gzmvoc:gzmvoc_thes_gen_places.ttl	591	GZMVOC
mdb:mdb_thes_rangen.ttl	585	MDB
vocopv:vocopv_schema.ttl	337	VOC Opvarenden
dss:dss_provenance.ttl	273	DSS (all)
mdb:ranks_and_shiptypes_1.ttl	245	MDB
gzmvoc:gzmvoc_schema.ttl	232	GZMVOC
mdb:mdb_thes_generated.ttl	196	MDB
file:///data/cliopatria/ClioPatria/rdf/base/rdfs.rdfs	190	External
gzmvoc:gzmvoc_thes_gen.ttl	166	GZMVOC
mdb:mdb_schema.ttl	149	MDB
das:das_schema.ttl	98	DAS
dss:dss_schema.ttl	59	DSS (all)
http://e-culture.multimedian.nl/ns/rkd/aatned/aatned.rdfs	27	External
Total no. triples:	25,529,107	

4 Accessing the data

Web interface The data is accessible through two live ClioPatria triple store instances. A 'stable version' is published at `http://dutchshipsandsailors.nl/data` with a development version online at `http://semanticweb.cs.vu.nl/dss`. The stable version is especially interesting since it is hosted and maintained at the Huygens ING institute for historical research as part of their digital history infrastructure, rather than through a university server. This ensures stability and sustainability of the dataset beyond the research project. The ClioPatria web interface allows for browsing the data. The graphs can be browsed or downloaded and basic statistics are provided[24]. Local views of resources are also provided[25]. A search functionality, which includes autocompletion, is available. The provenance can be visualized using the PROV-O-Viz tool[26], which is integrated with the triple store at `http://dutchshipsandsailors.nl/data/provoviz`.

SPARQL endpoint A SPARQL 1.1 compliant endpoint is provided at `http://dutchshipasandsailors.nl/data/sparql/`, with a number of interactive inter-

[24] `http://www.dutchshipsandsailors.nl/data/browse/list_graphs`

[25] For example `http://www.dutchshipsandsailors.nl/data/browse/list_resource?r=http://purl.org/collections/nl/dss/vocopv/opvarenden-344716`

[26] `http://provoviz.org/`

faces provided, such as the YASGUI interface at `http://dutchshipasandsailors.nl/data/dss/yasgui/`. A number of editable example SPARQL queries are also presented at `http://www.dutchshipsandsailors.nl/data/dss_queries`.

Linked Data The PURL URIs redirect to the specific resources on the stable server which will respond through content negotiation. In the case of an RDF value for the HTTP accept header the server returns RDF triples concerning the resource. In the current setup the *symmetric concise bounded description* of a resource is returned, which is made up by all triples that have that resource either as a subject or as an object. This conforms to the Linked Data principles [2]. ClioPatria can respond with RDF in XML, ntriples, turtle or JSON-LD serialization. In the case of a HTML request, the HTML local view is returned

Raw Data Finally, the raw RDF data is available i) through the web interface, where individual graphs can be downloaded as RDF/Turtle or RDF/XML; ii) through a public repository at `https://www.github.com/biktorrr/dss`; iii) as archived humanities datasets at the EASY online archiving system of Data Archiving and Networking Services (DANS)[27]. Here the four datasets as well as the interoperability layer are available as RDF/XML files with persistent identifiers. Here they are ensured sustainability beyond the life expectancy of the live versions.

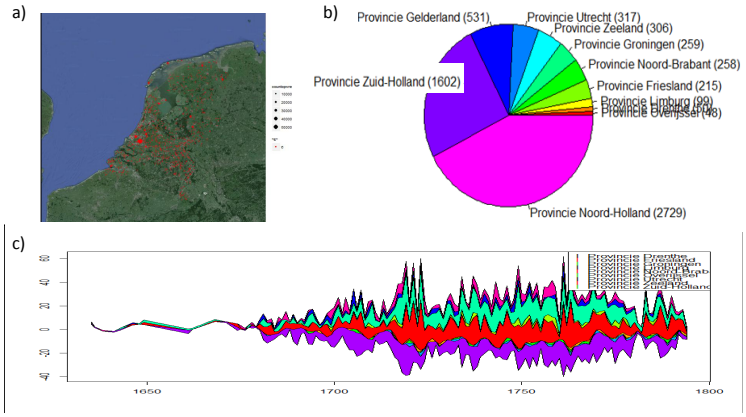

Fig. 4. Three visualizations of VOC data made possible through GeoNames links a) shows a plot of birth places on a map; b) shows aggregation by provinces of sailors in one year (1750) and c) shows a stream plot of the sailors per province over all the years for which we have data. These visualizations are made through a simple SPARQL query on the datacloud and visualizing the results using R.

[27] `https://easy.dans.knaw.nl/ui/home`

5 Digital History Examples

In this section, we present three example uses developed in collaboration with the historians associated with the project. For the sake of brevity, we omit the complete SPARQL queries used in these use cases here, but they are reproduced on the semantic server at `http://dutchshipsandsailors.nl/data/dss_queries`.

Because many dataset-specific properties are mapped to DSS properties, we can use RDFS reasoning to search for resources across the different datasets. It is not hard to define a search query that retrieves all ships with the ship name "Johanna" or that have some person with the rank of Captain that has "Veldman" as a last name. This allows for search and comparison between the datasets and for example to research correlations between variables (rank and wages?) using data from more than one dataset.

Analysis of the types of persons that sailed on the VOC ships can give insight into the socio-economic realities of the 18th Century. The datasets lists the birthplaces of many of those embarked on (VOC) ships. Through the links with GeoNames, we can get more information about those places of origin. One of these uses is to use the GeoNames geo-coordinates to plot information on a map. Figure 4a) shows such a plot. We can also use the GeoNames geographical hierarchy to -for example- analyze the provinces of origin of the voyagers, giving insight at an aggregated level. We used the SPARQL package for the statistical analysis tool R to provide a quantitative analysis and visualizations of the results[28]. Figure 4b)shows the birth provinces of sailors for one year (1750) and Figure 4c) shows a stream plot of the birth provinces of sailors over multiple years. These visualizations are made possible through the links with an external dataset, they can easily be done for one or multiple DSS datasets and give an insight into the geographical origins of sailors. These visualizations can be used to detect anomalies, formulate hypotheses and to make the work of the quantitative historian more effective and efficient.

In their research, historians combine analysis of data with their expert knowledge as well as common-sense knowledge. Through the link with AAT and DBPedia, we can use the formalized common sense and expert knowledge to automatically analyze the data. For example, the ship type hierarchy from AAT can be used to analyze features of specific ship types. One of the example queries lists persons that embarked on coastal ships (which has a number of subtypes such as "kof" or "tjalk"). Without the explicit links, a very complex conjunctive query would have to be formulated.

6 Related Work

This work builds on previous research that resulted in the Amsterdam Museum Linked Dataset as well as the Verrijkt Koninkrijk Linked dataset[6, 5]. The latter effort also was done in close collaboration with historians, using specific digital

[28] `http://cran.r-project.org/web/packages/SPARQL/`

history research goals. In this case, multiple datasets are combined into one datacloud, which makes new types of analysis possible. Some tools and methods are re-used for this paper. Our work has a similar relation to other efforts that attempt to link historical data to the Web of data [8, 16]. In fact there are multiple examples of datasets that are the result of collaborations between computer scientists and historians[11]. However, in most cases, this concerns a single dataset, published using a single metadata model. In our approach, we work with historians from different backgrounds, who are responsible for their own data and datamodel. This results in a datacloud of multiple datasets rather than one monolithic dataset. In the related cultural heritage domain, publishing of metadata as linked data is gaining ground. Examples include Europeana [9] which uses the Linked Data architecture to provide access to Europe's cultural heritage metadata from multiple collection metadata providers.

7 Conclusions and Future work

We presented the Dutch Ships and Sailors Linked Data cloud, developed in collaboration with the historical researchers responsible for those datasets. We make four separate and important maritime digital history datasets available as linked data to researchers and the public. Beyond these four datasets, this paper shows how Linked Data principles and technologies serve to integrate different datasets in a flexible way. In the case of these relatively "small" datasets, close collaboration between data experts and the converting party ensures that the richness of the original data is not lost, and interoperability is gained up to a level where it can be used for further historical research. It is an example of how Linked Data can benefit humanities research -more specifically digital history. The datacloud can serve as a hub dataset for international maritime historical datasets as well as for other (Dutch) historical datasets. We identified a total of 25 maritime historical datasets that can be added to the datacloud[29]. Links to more datasets are currently being established. For example, part of the Dutch historical census data made available through the CEDAR project[10] is already partly linked available in the development version. This presents opportunities for even more elaborate types of analysis beyond the maritime context.

We are also experimenting with more user-friendly interfaces for specific types of historical research questions. For the MDB dataset, we will make the digital scans available and link these to the MDB records, deepening the provenance information. This enables tracing results of (SPARQL) queries back to the original data even more than is currently possible, ensuring further trust and usability in the historical research context.

Acknowlegdgements

This work was supported by CLARIN-NL (`http://www.clarin.nl`) under project name DSS. We would like to thank Robin Ponstein and Andrea Bravo Balado.

[29] A list can be found at `http://dutchshipsandsailors.nl`

References

1. A. Bravo Balado. Information extraction on newspaper archives for historical research. a dutch maritime history case study. M.Sc. thesis VU University Amsterdam (forthcoming), 2014.
2. Tim Berners-Lee. Linked data - design issues. http://www.w3.org/DesignIssues/LinkedData.html, 2006.
3. J. R. Bruijn, F. S. Gaastra, and I. Schffer. Dutch-asiatic shipping in the 17th and 18th centuries, i, ii and iii. Rijks Geschiedkundige Publication Grote Serie 165, 166, 167; Den Haag: Martinus Nijhoff, 1987, 1979 and 1979.
4. Daniel J. Cohen et al. Interchange: The promise of digital history. *Special issue, Journal of American History*, 95, no.2, 2008.
5. V. de Boer, J. van Doornik, L. Buitinck, M. Marx, T. Veken, and K. Ribbens. Linking the kingdom: Enriched access to a historiographical text. In *Proceedings of KCAP 2013, Banff, Canada, 23-26 June 2013*, 2013.
6. V. de Boer, J. Wielemaker, J. van Gent, M. Hildebrand, A. Isaac, J. R. van Ossenbruggen, and G. Schreiber. Supporting Linked Data Production For Cultural Heritage Institutes: The Amsterdam Museum Case Study. In *Proceedings of European Semantic Web Conference (ESWC)*, 2012.
7. Paul Groth and Luc Moreau (eds.). PROV-Overview. An Overview of the PROV Family of Documents. W3C Working Group Note NOTE-prov-overview-20130430, World Wide Web Consortium, April 2013.
8. E. Hyvönen, T. Lindquist, J. Törnroos, and E. Mäkelä. History on the semantic web as linked data – an event gazetteer and timeline for the world war i. In *Proc. of CIDOC 2012 - Enriching Cultural Heritage, Helsinki, Finland*, June 2012.
9. A. Isaac and B. Haslhofer. Linked open data - data.europeana.eu. Semantic Web 4(3): 291-297 (2013), 2012.
10. A. Meroño-Peñuela, A. Ashkpour, L. Rietveld, and R. Hoekstra. Linked humanities data: The next frontier? a case-study in historical census data. *Proceedings of the 2nd International Workshop on Linked Science 2012*, 951, 2012.
11. A. Meroño-Peñuela, A. Ashkpour, M. van Erp, . Mandemakers, L. Breure, A. Scharnhorst, S. Schlobach, and F. van Harmelen. Semantic technologies for historical research: A survey. *Semantic Web Journal [to appear].*, 2012.
12. Tom De Nies, Sam Coppens, Erik Mannens, and Rik Van de Walle. Modeling uncertain provenance and provenance of uncertainty in w3c prov. In *Proceedings of WWW 2013*, pages 167–168, 2013.
13. N. Ockeloen, A. Fokkens, S. ter Braake, P. Vossen, V. de Boer, G. Schreiber, and S. Legne. Biographynet: Managing provenance at multiple levels and from different perspectives. In *Proceedings of the Workshop on Linked Science (LiSC) at ISWC 2013, Sydney, Australia, October 2013*, 2013.
14. R. Ponstein. Reconciling dutch ships and sailors. M.Sc. thesis VU University Amsterdam (forthcoming), 2014.
15. M. van Rossum. De intra-aziatische vaart. schepen, de aziatische zeeman en ondergang van de voc. *Tijdschrift voor Sociale en Economische Geschiedenis*, 8, nr. 3:32–69, 2011.
16. M. van Erp, J. Oomen, R. Segers, C. van de Akker, L. Aroyo, G. Jacobs, S Legne, L. van der Meij, J. R. van Ossenbruggen, and G. Schreiber. Automatic Heritage Metadata Enrichment With Historic Events. In *Proc. of Int. Conference for Culture and Heritage On-line-Museums and the Web 2011.* Archimuse, April 2011.
17. A.J.M van Velzen and F.S. Gaastra. Thematische collectie: Voc opvarenden; voc sea voyagers. urn:nbn:nl:ui:13-v73-sq8, 2000–2010.